Wireless Communications

Andrea Goldsmith

Draft of Second Edition
Chapters 1-9
Feb. 18, 2020

Contents

Chapter 1

Overview of Wireless Communications

Wireless communication is one of the most impactful technologies in history, drastically affecting the way we live, work, play, and interact with people and the world. There are billions of cellphone subscribers worldwide, and a wide range of devices in addition to phones use cellular technology for their connectivity. Wireless network technology using the Wi-Fi standard has been incorporated into billions of devices as well, including smartphones, computers, cars, drones, kitchen appliances, watches, and tennis shoes. Satellite communication systems support video, voice, and data applications for receivers on earth, in the air, and in space. Revenue across all areas of wireless technology and services is trillions of dollars annually. The insatiable demand for wireless data along with new and compelling wireless applications indicate a bright future for wireless systems. However, many technical challenges remain in designing wireless networks and devices that deliver the performance necessary to support existing and emerging applications. In this introductory chapter we will briefly review the history of wireless communications, from the smoke signals of antiquity to the rise of radio communication that underlies the Wi-Fi, cellular, and satellite networks of today. We then discuss the most prevalent wireless communication systems in operation today. The impact of spectrum properties and regulation as well as standards on the design and success of wireless systems is also illuminated. We close the chapter by presenting a vision for the wireless communication systems of the future, including the technical challenges that must be overcome to make this vision a reality. Techniques to address many of these challenges are covered in subsequent chapters. The huge gap between the capabilities of current systems and the vision for future systems indicates that much research and development in wireless communications remains to be done.

1.1 History of Wireless Communications

1.1.1 Origins of Radio Technology

The first wireless networks were developed in antiquity. These systems transmitted information visually 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.

The origins of radio communications began around 1820 with experiments by Oersted demonstrating that an electric field could move a compass needle, thereby establishing a connection between electricity and magnetism. Work by Ampére, Gauss, Henry, Faraday, and others further advanced knowledge about electromagnetic waves, culminating in Maxwell's theory of electromagnetism published in 1865. The first transmission of electromagnetic

waves was performed by Hertz in the late 1880s, after which he famously declared to his students that these waves would be "of no use whatsoever." He was proved wrong in 1895 when Marconi demonstrated the first radio transmission across his father's estate in Bologna. That transmission is considered the birth of radio communications, a term coined in the early 1900s. Marconi moved to England to continue his experiments over increasingly large transmission ranges, culminating in the first trans-Atlantic radio transmission in 1901. In 1900 Fessenden became the first person to send a speech signal over radio waves, and six years later he made the first public radio broadcast. From these early beginnings, described in more detail in [1], 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.

1.1.2 From Analog to Digital

Radio systems designed prior to the invention of the transistor, including AM/FM radio, analog television, and amateur radio, transmitted analog signals. Most modern radio systems transmit digital signals generated by digital modulation of a bit stream. The bit stream may represent binary data (e.g., a computer file, digital photo, or digital video stream) or it may be obtained by digitizing an analog signal (e.g., by sampling the analog signal and then quantizing each sample). 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. When packet radios transmit continuous data such as voice and video, the delay between received packets must not exceed the delay constraint of the data. The first wireless network based on packet radio, ALOHAnet, was developed at the University of Hawaii and began operation 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 saw great potential for communication systems exploiting the combination of packet data and broadcast radio inherent to ALOHAnet. Throughout the 70's and early 80's the Defense Advanced Research Projects Agency (DARPA) invested significant resources to develop networks using packet radios for communications in the battlefield. The nodes in these packet radio networks had the ability to configure (or reconfigure) into a network without the aid of any established infrastructure. Self-configuring wireless networks without any infrastructure were later coined *ad hoc* wireless networks.

DARPA's investment in packet radio networks peaked in the mid 1980's, but these networks fell far short of expectations in terms of speed and performance. This was due in part to the limited capabilities of the radios and in part to the lack of robust and efficient access and routing protocols. Packet radio networks also found commercial application in supporting wide-area wireless data services. These services, first introduced in the early 1990's, enabled wireless data access (including email, file transfer, and web browsing) at fairly low speeds, on the order of 20 Kbps. The market for these wide-area wireless data services did not take off due mainly to their low data rates, high cost, and lack of "killer applications". All of these services eventually folded, spurred in part by the introduction of wireless data in 2G cellular services [2], which marked the dawn of the wireless data revolution.

1.1.3 Evolution of Wireless Systems and Standards

The ubiquity of wireless communications has been enabled by the growth and success of Wireless Local Area Networks (WLANs), standardized through the family of IEEE 802.11 (Wi-Fi) protocols, as well as cellular networks. Satellite systems also play an important role in the wireless ecosystem. The evolution of these systems and their corresponding standards is traced out in this subsection.

Wi-Fi Systems

The success story of Wi-Fi systems, as illuminated in [3], began as an evolution of the Ethernet (802.3) standard for wired local area networks (LANs). Ethernet technology, developed at Xerox Parc in the 1970s and standardized in 1983, was widely adopted throughout the 1980s to connect computers, servers, and printers within office buildings. WLANs were envisioned as Ethernet LANs with cables replaced by radio links. In 1985 the Federal Communications Commission (FCC) enabled the commercial development of WLANs by authorizing for unlicensed use three of the Industrial, Scientific, and Medical (ISM) frequency bands: the 900 MHz band spanning 902-928 MHz, the 2.4 GHz band spanning 2.4-2.4835 GHz, and the 5.8 GHz band spanning 5.725-5.875 GHz. Up until then, these frequency bands had been reserved internationally for radio equipment associated with industrial, scientific and medical purposes other than telecommunications. Similar rulings followed shortly thereafter from the spectrum regulatory bodies in other countries. The new rulings allowed unlicensed use of these ISM bands by any radio following a certain set of restrictions to avoid compromising the performance of the primary band users. Such radios were also subject to interference from these primary users. The opening of these ISM bands to "free" use by unlicensed wireless devices unleashed a flurry of research and commercial wireless system development, particularly for WLANs.

The first WLAN product for the ISM band, called WaveLAN, was launched in 1988 by the NCR corporation and cost several thousand dollars. These WLANs had data rates up to 2 Mbps and operated in the 900 MHz and 2.4 GHz ISM bands. Dozens of WLAN companies and products appeared over the ensuing few years, mostly operating in the 900 MHz ISM band using direct-sequence spread spectrum, with data rates on the order of 1-2 Mbps. The lack of standardization for these products led to high development costs, poor reliability, and lack of interoperability between systems. Moreover, Ethernet's 10 Mbps data rate and high reliability far exceeded the capabilities of these early WLAN products. Since companies were willing to run cables within and between their facilities to get this better performance, the WLAN market remained small for its first decade of existence. The initial WLAN products were phased out as the 802.11 standards-based WLAN products hit the market in the late 1990s. In Europe a WLAN standard called HIPERLAN was finalized in 1996. An improved version was launched in 2000, but HIPERLAN systems never gained much traction.

The first 802.11 standard, finalized in 1997, was born from a working group within the IEEE 802 LAN/MAN Standards Committee that was formed in 1990. This standard, called 802.11-1997, operated in the 2.4 GHz ISM band with data rates up to 2 Mbps. The standard used channel sensing with collision avoidance for medium access as well as frequency hopping or direct-sequence spread spectrum to mitigate the main sources of interference in that band which, at the time, consisted primarily of microwave ovens, cordless phones, and baby monitors. Also in 1997, the FCC authorized 200 MHz of additional unlicensed spectrum in the 5 GHz band, with certain constraints to avoid interfering with the primary users, mostly radar systems, operating in this band. Two years later, in 1999, the standard that would ignite the Wi-Fi revolution, 802.11b, was finalized. This standard increased data rates to 11 Mbps and eliminated the frequency-hop option of its predecessor. A plethora of 802.11b products soon appeared. Their interoperability, coupled with dramatically lower costs relative to earlier WLAN products, led to widespread use of this new technology. Millions of 802.11b products were shipped in 2000, just one year after the standard was finalized, and these shipments grew tenfold by 2003. The year 1999 marked two other important milestones for WLANs. That year the 802.11a standard for the 5GHz ISM frequency band was finalized to capitalize on the new unlicensed spectrum in that band. This standard enabled 54 Mbps data rates in 20 MHz channels and introduced orthogonal-frequency-division-multiplexing (OFDM) coupled with adaptive modulation as a new physical layer design. Also in 1999 the Wireless Ethernet Compatibility Alliance was formed to facilitate interoperability and certification of WLAN products. The name of this group was later abbreviated to the Wi-Fi Alliance, thereby coining the widely used moniker for WLAN technology and standards today.

The Wi-Fi standard has evolved rapidly since its early days, with new versions developed about every five years, as described in more detail in Appendix D.1. The 802.11g standard, introduced in 2003, is essentially the same in its physical layer and multiple access design as 802.11a, but operates in the 2.4 GHz ISM frequency band. Later standards provided improvements over these first OFDM Wi-Fi systems, including wider channels, multiple transmit and receive antennas to enable multiple spatial streams and improved robustness through beamforming, larger signal constellations, improved error-correction codes, and coordinated multiple access. These improvements have led to highly-reliable Wi-Fi products that are capable of 10 Gbps data rates within a building or outdoor area. Wi-Fi has also moved into the unregulated 60 GHz frequency band through the 802.11ad standard. Even its name has evolved; In 2018 the 802.11 standards body abandoned using letter suffixes for new generations of the standard. Instead it coined the sixth generation of the WiFi standard, originally named 802.11ax, as Wi-Fi 6. Today Wi-Fi technology is pervasive indoors and out even in remote corners of the world. In addition to its pervasiveness, Wi-Fi has experienced an explosion of applications beyond its original use of connecting computers to each other and their peripheral devices. In addition to computers, smartphones, and tablets, many electronic devices today, from medical devices to refrigerators to cars, are equipped with Wi-Fi, allowing them to download new software, exchange data with other devices, and take advantage of cloud-based storage and computation.

Cellular Systems

Cellular systems are another exceedingly successful wireless technology. The convergence of radio and telephony began in 1915, when wireless voice transmission between New York and San Francisco was first established. The first analog mobile telephone system was deployed in St. Louis Missouri in 1946, launching AT&T's Mobile Telephone Service (MTS). Within two years AT&T had deployed MTS over approximately 100 cities and highway corridors. Only six channels were allocated by the FCC for the service and, due to their close spacing in frequency, only three were usable at any given time. Hence only three people within a city could make a call simultaneously. The monthly service and per-call cost was very high, and the equipment bulky and heavy. Evolution of the system was slow; while the equipment improved and spectrum to support up to 12 channels was added, the system capacity remained extremely limited.

Ironically, about the same time MTS was first being deployed, a solution to this capacity problem had already emerged from researchers at AT&T Bell Laboratories: the notion of cellular systems. The cellular system concept, articulated in a 1947 Bell Laboratories Technical Memo by D. H. Ring [4], exploited the fact that the power of a transmitted signal falls off with distance. Thus, channels using the same frequency can be allocated to users at spatially-separate locations with minimal interference between the users. To exploit this principle of frequency reuse, a cellular system partitions a geographical area into non-overlapping *cells*, as shown in Fig. 1.3 below. Sets of channels are assigned to each cell, and cells that are assigned the same channel set are spaced far enough apart so that interference between the users in these cells is small. In early cellular systems the distance between cells using the same channel set was relatively large, but today sophisticated interference mitigation techniques allow channels to be reused in every cell. As a user moves between adjacent cells, its call is *handed off* to a channel associated with the new cell. Frequency reuse enables much more efficient use of spectrum as the number of simultaneous users is no longer limited to the number of available channels. Indeed, while many aspects of cellular system technology have changed over time, frequency reuse remains at the heart of cellular system design.

Although the cellular concept was introduced in the late 1940s, it was not implemented for several decades, as technology was not yet ripe to realize the system in practice. In the mid-1960s engineers at Bell Laboratories began work on a design and feasibility study for a metropolitan analog cellular system. The details of the design and analysis, along with successful system tests in Newark and Philadelphia, formed the basis of an AT&T FCC proposal in 1971 to approve cellular service and allocate spectrum for it [5]. The FCC approved experimental cellular licenses to telephone companies in 1974, which led to the construction of several cellular systems. Fol-

lowing a long and convoluted process, in 1981 the FCC finalized its ruling for the issuance of commercial cellular licenses, instituting a duopoly in every metropolitan area whereby one license would be granted to a traditional phone company, and the other to a non-traditional operator. The first generation cellular system for U.S. deployment, called the advanced mobile phone system (AMPS) and described in [6], was launched in Chicago in 1983. A similar service had been launched in Tokyo in 1979 and in Scandinavia in 1981, although those systems were incompatible with AMPS. Cellular service in many other countries launched in the ensuing years, mostly with incompatible standards, which precluded cellular roaming between countries. These first generation systems are referred to as 1G systems.

Like Wi-Fi, the exponential growth of the cellular industry exceeded all expectations, increasing by at least an order of magnitude every decade, from a modest 100,000 subscribers in 1984 to tens of millions in 1994, hundreds of millions in 2004, and billions in 2014. To meet the growing demand for wireless data along with a diverse requirements for different types of wireless devices, a new generation of cellular systems and standards has emerged approximately every decade. In fact, that timeline was compressed for 1G systems, since the analog cellular system deployed in Chicago in 1983 was already saturated by 1984. At that point the FCC increased the cellular spectral allocation from 40 MHz to 50 MHz. As cellular systems throughout more and more cities became saturated with demand, the development of digital cellular technology for increased capacity and better performance became essential. These enhanced system requirements launched the process to create a second generation (2G) cellular standard in the late 1980s with deployments in the early 1990s. In addition to voice communication, the move to digital technology paved the way for these systems to support low-rate data as well, in particular short texts, voice mail, and paging services. Unfortunately, the great market potential for cellular phones led to a proliferation of 2G cellular standards, with three different standards in the U.S. alone. One of these matched the European standard; Japan adopted a separate standard. Hence global roaming required a multi-mode phone. This deficiency was corrected for the third generation of cellular standards (3G), for which seven telecommunications standards bodies across Asia, Europe and the United States formed the third-generation partnership project (3GPP) to develop a single worldwide standard. The 3G cellular systems based on the 3GPP standard, whose deployments began in the early 2000s, provided an order of magnitude higher peak data rates than 2G systems. Indeed, it was the capabilities of 3G systems that transitioned the "killer application" of cell phones from voice to wireless data. The proliferation of smart phone technology in the mid-2000s, which were designed to consume vast amounts of wireless data, greatly stressed the capacity of 3G networks. Moreover, the 3G networks had far lower data rates than Wi-Fi, which became the access mode of choice for high-speed data. These developments paved the way for the next-generation 4G "long term evolution (LTE)" cellular standard. These systems, deployed in the early 2010s, supported an order-of-magnitude peak data rate increase over 3G systems. The 5G cellular standard supports higher data rates than 4G systems, as well as lower latency and better energy efficiency. System deployments for 5G began in 2019, continuing the trend of new cellular systems and standards every decade. Starting with the 3G systems, each generation of cellular standard has been developed to meet a set of International Mobile Telecommunications (IMT) requirements outlined by the International Telecommunications Union (ITU). The ITU certifies which cellular standards meet its IMT requirements; the 3GPP standard is designed to meet these requirements, as are alternate standards developed by specific countries or other standards organizations. More details on current cellular technology will be given in Chapter 15, with the evolution of cellular standards described in Appendix D.2.

Satellite Systems

Satellite communication systems are another major component of today's wireless communications infrastructure. Commercial satellite systems can provide broadcast services over very wide areas, fill the coverage gap in locations without cellular service, and provide connectivity for aerial systems such as airplane Wi-Fi. Satellite systems are typically categorized by the height of the satellites' orbit: low-earth orbit (LEO) satellites operate at

roughly 2000 km altitude, medium-earth orbit (MEO) satellites at roughly 9000 km altitude, and geosynchronous orbit (GEO) satellites at roughly 40,000 km altitude. GEO satellites with geostationary orbits are seen as stationary from the earth, whereas GEO satellites with other orbits (such as elliptical) have their coverage area change over time. The disadvantage of high altitude orbits is that it takes a great deal of power to reach the satellite, and the propagation delay is typically too large for two-way delay-constrained applications like voice, video, and gaming. However, satellites at these orbits tend to have larger coverage areas, so fewer satellites are necessary to provide wide-area or global coverage. In addition to communication services, satellites are used for many other applications including weather monitoring, earth observation, surveillance, imaging, navigation, and localization.

The concept of using GEO satellites for communications was first suggested by Herman Potocnik in 1928 and later popularized through a 1945 article in *Wireless World* written by the science fiction writer Arthur C. Clarke. 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. Following these launches, in 1962 the Communications Satellite Corporation (Comsat) was formed in the United States to develop commercial communication satellite systems. Two years later, the International Telecommunications Satellite Consortium (IntelSat) emerged as a public-private consortium of 18 countries with the goal of enabling global telecommunications connectivity. SYNCOM 2, the first communication satellite to successfully reach geosynchronous orbit, was launched in 1963. The following year SYNCOM 3 was launched into geostationary orbit, providing a two-way 10 MHz communication channel at a carrier frequency of 1.8 GHz for the satellite-to-earth link (downlink) and at 7.3 GHz for the reverse link (uplink). Shortly after its launch, SYNCOM 3 was used to provide live television coverage to US viewers of the 1964 Summer Olympics in Tokyo. The IntelSat consortium launched a number of satellites during the late 1960s in order to reach near-global coverage, culminating in the broadcasting of the 1969 landing of the first human on the moon to 600 million viewers. In addition to television broadcasting, these early systems also supported voice, teletype, and paging.

GEO satellite technology has evolved continuously since its early days. The C band of the radio spectrum (4-8 GHz) was used in early GEO systems, but subsequent systems have taken advantage of the larger bandwidths available at the higher frequency X (8-12.5 GHz), Ku (12.5-18 GHz), and Ka (26.5-40 GHz) bands in order to provide higher data rates. Antenna system technology for GEOs has also evolved to focus the transmission energy more precisely, leading to higher signal quality and hence higher data rates. There are hundreds of GEO satellites deployed today, with some capable of a total data rate in excess of one hundred Gbps. For individual users, typical data rates are on the order of ten Mbps in the downlink and several Mbps in the uplink [7]. Due to their large coverage regions, geosynchronous satellites are the primary communications mechanism for vessels at sea and in the air. They are also well-suited for broadcast entertainment where round-trip delay is not a consideration; they support hundreds of channels of digital television and radio at very high quality.

As cellular systems were rolled out in the late 1980s, the focus for satellite technology turned to building LEO systems that might compete with them. This led to the launch of several LEO communication systems in the late 1990s, including Globalstar and Iridium [8]. These LEOs provided global coverage but the link rates remained low due to power and bandwidth constraints. The handsets for these systems were also much larger and heavier than their cellular counterparts, primarily due to their large batteries. The LEO satellite systems deployed in the 1990s did not experience significant commercial success as, for most users, they provided worse performance and coverage than the competing 2G cellular phone technology at a higher cost. While these satellite systems provided coverage to remote areas without access to cellular service, the cost was prohibitive for the large majority of people in these areas. As a result, pretty much all the initial LEO companies, including Globalstar and Iridium, filed for bankruptcy protection within a few years of launching their service. Teledesic, which was planning a very ambitious LEO system with more than 800 satellites, declared bankrupcy before a single launch. After several iterations of restructuring, Iridium and Globalstar emerged as profitable services and today provide global Internet connectivity, albeit for a relatively small number of subscribers compared to that of cellular service. Interest in

LEO satellites resurfaced as 4G cellular systems rolled out due to increased demand for connectivity in remote locations not served by cellular, better technology for both satellites and ground transceivers, and reduced launch costs. Significant commercial development of LEO satellites emerged concurrent with 5G cellular rollouts as a means to provide broadband Internet service to the billions of people worldwide that do not have access to high-speed connectivity. It is anticipated that such systems will each deploy hundreds to thousands of satellites.

Today we take for granted the magic of wireless communications, which allows us and our devices to be connected anywhere in the world. This magic has been enabled by the pioneers that contributed to advances in wireless technology underlying the powerful systems ubiquitous today.

1.2 Current Systems

This section provides a design overview of the most prevalent wireless systems in operation today. The design details of these systems are constantly evolving, incorporating new technologies and innovations. This section will focus mainly on the high-level design aspects of these systems. More details on these systems and their underlying technologies will be provided in specific sections of the book. A summary of Wi-Fi, cellular, and short-range networking standards can be found in Appendix D with more complete treatments of recent standards in [9, 10, 11, 12, 13].

1.2.1 Wireless Local Area Networks

WLANs support data transmissions for multiple users within a "local" coverage area whose size depends on the radio design, operating frequency, propagation characteristics, and antenna capabilities. A WLAN consists of one or more access points (APs) connected to the Internet that serve one or more WLAN devices, also called clients. The basic WLAN architecture is shown in Figure 1.1. In this architecture clients within the coverage area of an AP connect to it via single-hop radio transmission. For better performance and coverage, some WLANs use relay nodes to enable multi-hop transmissions between clients and APs, as shown in Figure 1.2. WLANs today follow the 802.11 Wi-Fi family of protocols. Handoff of a moving client between APs is not supported by the 802.11 protocol, however some WLAN networks, particularly corporate and campus systems, incorporate this feature into their overall design.

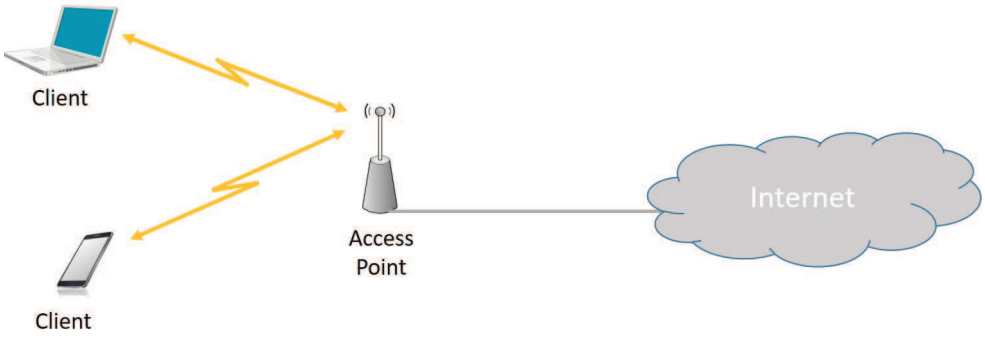

Figure 1.1: Basic wireless LAN architecture

While early WLANs were exclusively indoor systems, current WLANs operate both indoors and outdoors. Indoor WLANs are prevalent in most homes, offices, and other indoor locations where people congregate. Outdoor systems are generally deployed in areas with a high density of users such as corporate and academic campuses, sports stadiums, and downtown areas. The range of an indoor WLAN is typically less than 50 meters and can

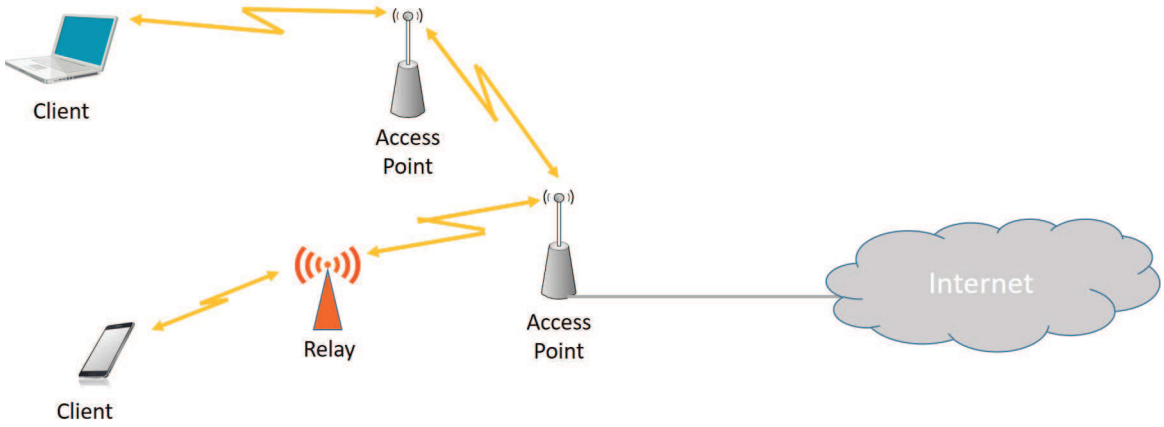

Figure 1.2: Wireless LAN architecture with multihop transmissions

be confined to a single room for those operating at 60 GHz. Outdoor systems have a bigger range than indoor systems due to their higher power, better antennas, and lower density of obstructions between the transmitter and receiver. WLANs generally operate in unlicensed frequency bands, hence share these bands with other unlicensed devices, such as cordless phones, baby monitors, security systems, and Bluetooth radios. Interference between WLAN devices is controlled by the WLAN access protocol, whereas interference between WLAN devices and other unlicensed devices is mitigated by a limit on the power per unit bandwidth for such devices.

WLANs use packet data transmission for better sharing of the network resources. Hence, data files are segmented into packets. Sharing of the available bandwidth between different APs and, for most systems, between clients accessing the same AP is typically controlled in a distributed manner using the carrier sense multiple access with collision avoidance (CSMA/CA) protocol. In CSMA/CA, energy from other transmissions on a given channel is sensed prior to a transmission and, if the energy is sensed above a given threshold, the transmitter waits a random *backoff* time before again attempting a transmission. A transmitted packet is acknowledged by the receiver once received. If a transmitted packet is not acknowledged, it is assumed lost and hence retransmitted. As discussed in more detail in Chapter 14, the CSMA/CA access protocol is very inefficient, which leads to poor performance of WLANs under moderate to heavy traffic loads. As a result, some WLANs use more sophisticated access protocols, similar to those in cellular systems, that provide centralized scheduling and resource allocation either to all clients served by a given AP or, more generally, to all APs and clients within a given system.

1.2.2 Cellular Systems

Cellular systems provide connectivity to a wide range of devices, both indoors and out, with a plethora of features and applications including text messages, voice calls, data and video transfer, Internet access, and mobile "apps" (application software tailored to run on mobile devices). Cellular systems typically operate in licensed frequency bands, whereby the cellular operator must purchase or lease the spectrum in which their system operates. These licenses typically grant the operator exclusive use of the licensed spectrum. Starting in 2017 regulators in several countries began allowing cellular systems to operate in the 5 GHz unlicensed band in addition to within their licensed spectrum. Given the large amount of spectrum in these unlicensed bands compared to what is available in the licensed bands, this development has the potential to drastically increase cellular system capacity. However, it will also increase interference in the unlicensed bands, particularly for Wi-Fi systems.

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, a cellular

system consists of multiple cells, where each cell is assigned one or more channels to serve its users. Each channel may be reused in another cell some distance away, and the interference between cells operating on the same channel 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, the intercell interference increases owing to the smaller propagation distance between interfering cells. Early cellular systems had reuse distances greater than one, but current systems typically reuse each channel in every cell while managing the resulting interference through sophisticated mitigation techniques.

The most basic cellular system architecture consists of tessellating cells, as shown in Figure 1.3, where C_i denotes the channel set assigned to cell i. The two-dimensional cell shapes that tessellate (cover a region without gaps or overlap) are hexagons, squares, and triangles. Of these, the hexagon best approximates an idealized omni-directional transmitter's circular coverage area. Early cellular systems followed this basic architecture using a relatively small number of cells to cover an entire city or region. The cell base stations in these systems 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*. The first few generations of macrocell base stations used single-antenna omni-directional transmitters, so a mobile moving in a circle around these base station had approximately constant average received power unless the signal was blocked by an attenuating object. In deployed systems cell coverage areas overlap or have gaps as signal propagation from a set of base stations never creates tessellating shapes in practice, even if they can be optimally placed. Optimal base station placement is also impractical, as zoning restrictions, rooftop availability, site cost, backhaul and power availability, as well as other considerations influence this placement.

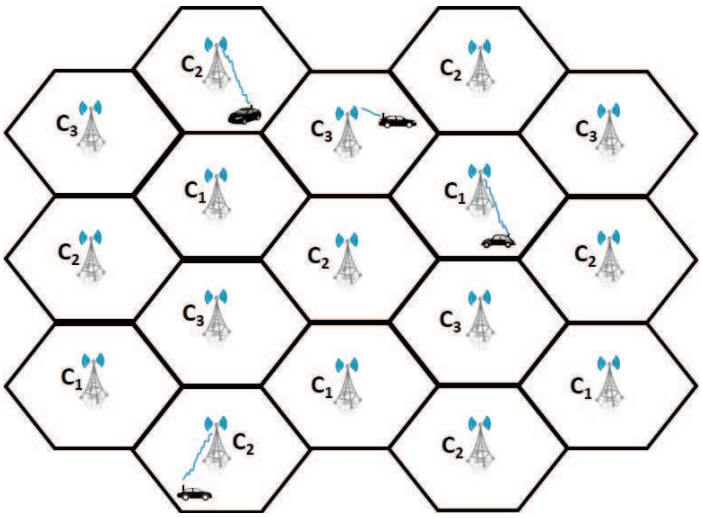

Figure 1.3: Cellular network architecture (homogeneous cell size).

Macrocells have the benefit of wide area coverage, but they can often become overloaded when they contain more users than channels. This phenomenon has led to hierarchical cellular system architectures, with macrocells that provide wide area coverage and small cells embedded within these larger cells to provide high capacity, as shown in Figure 1.4. Not only do the small cells provide increased capacity over a macrocell-only network, but they also reduce the transmit power required at both the base station and mobile terminal, since the maximum transmission distance within the small cell is much less than in the macrocell. Cellular systems with heterogeneous cells sizes are referred to as *Hetnets*. In current cellular systems, channels are typically assigned dynamically based on interference conditions. Another feature of many current cellular system designs is for base stations in

adjacent macrocells to operate on the same frequency, utilizing power control, adaptive modulation and coding, as well as interference mitigation techniques to ensure the interference between users does not preclude acceptable performance.

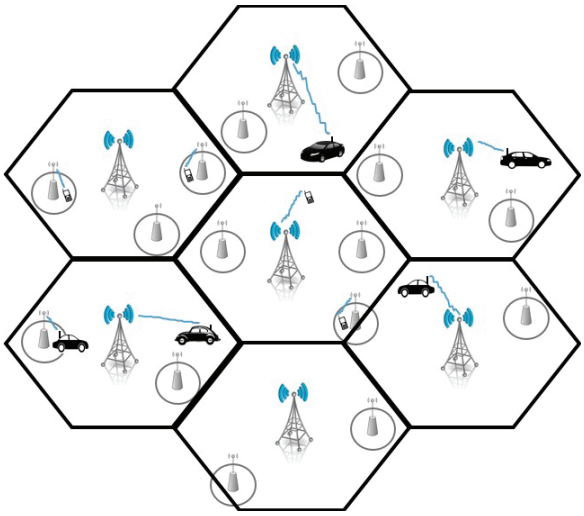

Figure 1.4: Hierarchical cellular network architecture (a Hetnet with macrocells and small cells).

All base stations in a given geographical area, both macrocells and small cells, are connected via a high-speed communications link to a mobile telephone switching office (MTSO), as shown in Figure 1.5. The MTSO acts as a central controller for the cellular system, 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 calls to mobile users in other geographic regions via the local MTSO in that region or to landline users through the public switched telephone network (PSTN). In addition, the MTSO can provide connectivity to the Internet.

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 dropped. A call will also be dropped if the signal strength between a mobile and its base station falls below the minimum threshold needed for communication due to signal propagation effects such as path loss, blockage, or multipath fading. These propagation characteristics are described in Chapters 2-3.

Spectral sharing in cellular systems, also called multiple access, is done by dividing the signaling dimensions along the time, frequency, code, and/or the spatial dimensions. Current cellular standards use a combination of time and frequency division for spectral sharing within a cell. However, there are still 2G and 3G systems in operation that use code-division multiple access based on direct sequence spread spectrum. Channels assigned to users within a cell may be orthogonal or non-orthogonal. In the latter case the channels of two users within a cell will overlap in time, frequency, code or spatial dimensions, which is referred to as intracell interference. More details on multiple access techniques and their performance analysis will be given in Chapters 13 and 14.

Efficient cellular system designs are *interference limited* – that is, interference from within and outside a cell dominates the noise floor, since otherwise more users could be added to the system. As a result, any tech-

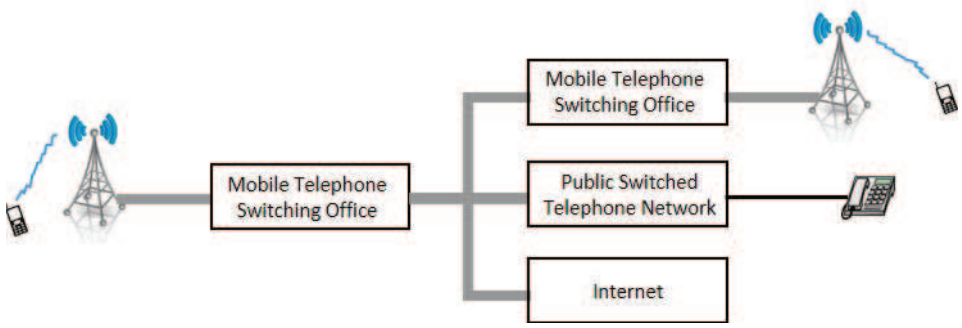

Figure 1.5: Cellular network architecture evolution.

nique to reduce interference in cellular systems leads directly to an increase in system capacity and performance. Some methods for interference reduction in cellular systems include cell sectorization, power control, directional antennas and antenna-array signal processing, multiuser detection and interference cancellation, base station cooperation, and user scheduling. Details of these techniques will be given in Chapters 14 and 15.

1.2.3 Satellite Systems

Satellite systems are another major component of the wireless communications infrastructure [14, 15]. There is a plethora of satellite communications services available today, including mobile service to airplanes, ships, vehicles, and hand-held terminals, fixed satellite service to earth stations, as well as broadcast radio and television service. Fig. 1.6 illustrates the satellite architecture supporting all such services. One of the biggest advantages of satellites over terrestrial wireless systems is their ability to provide coverage in remote areas.

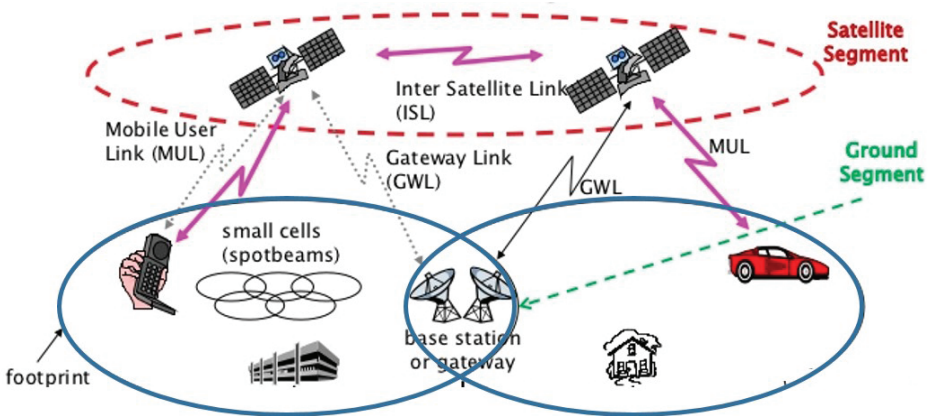

Figure 1.6: Satellite System Architecture

Satellite communication systems consist of one or more satellites communicating with stations on earth or in the air. Communication satellites typically serve as relays between stations on the ground and in some cases provide direct links to other satellites as well. Ground stations may be at fixed locations, in which case their antennas may be large to maximize received signal strength. Mobile stations and indoor fixed stations have smaller antennas, typically on the order of 1-3 meters. The coverage area or *footprint* of a satellite depends on its orbit. A GEO satellite can cover one or more continents, hence only a handful are needed to provide coverage for the entire globe. However, the transmission latency between the earth and a GEO satellite is high, on the order of 300 ms.

MEO satellites have a coverage area of around ten thousand square kilometers and a latency of around 80 ms. A LEO coverage area is about a thousand square kilometers with a latency of around 10 ms.

A GEO satellite provides continuous coverage to any station within its footprint, whereas a satellite with a different orbit will have a fixed station within its footprint for only part of a 24 hour period. GEO satellites are heavy, sophisticated, and costly to build and launch. Portable stations or handsets for GEOs tend to be large and bulky due to the power required to reach the satellite. In addition, the large round-trip propagation delay to a GEO satellite is quite noticeable in two-way voice communication. The most common service offered by GEO satellites is radio and TV broadcasting since delay is not a constraint and the satellite construction, launch, and operational costs are amortized over the many users within the GEO footprint. GEO satellites are also commonly used as a backbone link for terrestrial networks, for airplane and maritime communications, and for connectivity in remote locations.

LEO satellites are much lighter and lower in cost than GEOs to manufacture, launch, and operate. In addition, the stations and handsets in a LEO system have smaller size, transmit power, and latency than GEO systems due to the closer proximity of LEO satellites to the earth. Hence, most mobile services use LEO satellites, typically in a constellation of dozens to hundreds of satellites whose total footprint covers the locations around the globe supported by the system. Continuous coverage of a given location in a LEO system requires handoff between satellites as a fixed location on earth is within the footprint of a given LEO satellite for only 10-40 minutes. Hence, when the footprint of a LEO satellite moves away from a given station or handset, its connection is handed off to another LEO satellite so as to maintain continuous connectivity. Sophisticated antennas on LEO satellites can create very small spotbeams for mobile terminals that focus that transmission energy within a small footprint. These spotbeams allow for frequency reuse similar to that of cellular systems.

For the reasons outlined in the previous paragraph, LEOs tend to support the highest data rates, best performance, and lowest cost among satellite services. As expected, MEO satellite systems provide a compromise between the system requirements, performance, and costs compared with those of GEO and LEO systems. MEO satellite services mainly compete with GEO systems by offering better performance at a lower cost for users in relatively remote locations that are not well served by LEO or terrestrial systems.

1.2.4 Fixed Wireless Access

Fixed wireless access (FWA) systems support wireless communications between a fixed access point and multiple terminals. The FWA system architecture is shown in Fig 1.7. The access point transmits to the receivers of multiple terminals in the downlink direction, and receives signals from the transmitters of multiple terminals in the uplink direction. Different multiple access techniques can be used to share the system spectrum between the multiple terminals.

FWA systems provide an alternative to the wired broadband options of DSL, cable and fiber, whose availability may be limited in rural areas. In the United States, two frequency bands were set aside for these systems in the late 1990s: part of the 28 GHz spectrum for local distribution systems (local multipoint distribution service, LMDS) and a band in the 2 GHz spectrum for metropolitan distribution service (multichannel multipoint distribution services, MMDS). MMDS systems were never widely deployed, and are used today in sparsely populated rural areas, where laying cables is not economically viable. The initial LMDS systems were expensive and performed poorly due to the challenges of providing service at such high frequencies. As a result these systems did not succeed commercially and were discontinued. The LMDS spectrum is just below the millimeter wave band of 30-300 GHz. Recent interest in utilizing the large amount of unregulated millimeter wave spectrum has led to advances in system, circuit, and antenna design at these high frequencies, which in turn has renewed interest in FWA systems using the LMDS and millimeter wave bands.

One of the main growth drivers of FWA was the IEEE 802.16 (WiMAX) standard, finalized in 2001. The first WiMAX standard defined operation between 2 GHz and 11 GHz for non-line-of-sight links and between

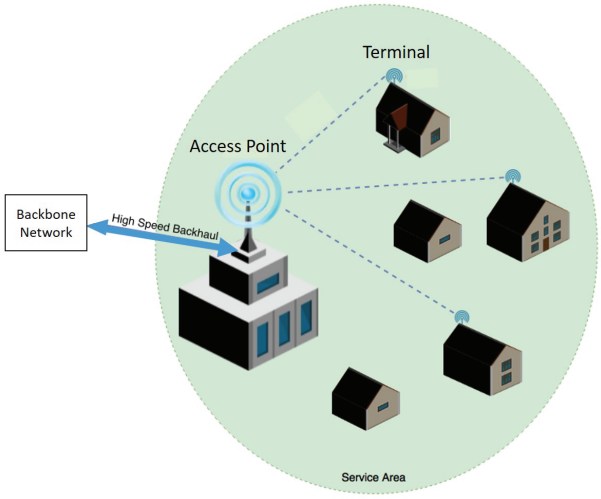

Figure 1.7: Fixed Wireless Access System

10 GHz and 66 GHz for line-of-sight links, however WiMAX systems generally operate below 6 GHz. The first commercial WiMAX systems followed the 802.16d standard with data rates of around 40 Mbps. A later version, 802.16e, was developed in the mid-2000s to support mobile users with a system design similar to that of cellular. The 802.16e systems offered 15 Mbps data rates, much higher than that of 3G cellular systems. The significantly higher data rates of 802.16e over 3G led to speculation that it would be adopted for 4G cellular. However, in the end 4G cellular adopted the LTE standard, which ultimately pushed WiMAX systems out of most mobile service markets. WiMAX is still used to support mobile services in areas where LTE has not been deployed, as well as in industrial markets such as aviation, utilities, and transportation. FWA systems based on the LTE and 5G cellular standards have been deployed to provide high-speed connectivity for homes, apartments, and office buildings.

1.2.5 Short Range Radios with Multihop Routing

As radios decrease their cost and power consumption, it becomes feasible to embed them into more types of electronic devices, which enables applications such as smart homes, sensor networks, vehicular networks, and other interconnected systems. The most common radio standards that have emerged to support this trend are Bluetooth, ZigBee, and Z-Wave. All of these radios support a *multihop routing protocol*, whereby a given radio can communicate with any other radio in its transmission range. If the destination radio is not within this range, intermediate radios relay the message to this destination. Radios with multihop routing form an ad hoc wireless network since they can reconfigure and have no established infrastructure. Such ad hoc wireless networks are commonly used by the military as well as for emergency response. In principle a multihop routing protocol can support hundreds or even thousands of nodes, but network performance generally degrades as the number of nodes increases. The short range radios described in this section have not yet demonstrated that their multihop routing protocols are feasible in practice for large numbers of nodes. In addition to Bluetooth, Zigbie, and Z-Wave, proprietary radios have been used in a number of devices, products, and military systems to support their communication requirements.

The Bluetooth standard is based on a small radio transceiver microchip built into digital devices.[1] The standard and device certification is managed by the Bluetooth Special Interest Group. In most applications the Bluetooth

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

radio takes the place of a connecting cable for electronic devices such as cell phones, tablets, headsets, audio equipment, cameras, watches, and smart meters. Bluetooth has also been integrated into larger devices such as cars and medical equipment. Bluetooth is mainly for short-range communications – for example, 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 2.5-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, so it can be used worldwide without any licensing issues. The Bluetooth standard provides one data channel at 723.2 kbps. In this mode, there is a reverse channel with a data rate of 57.6 kbps. The specification also allows up to three additional channels each at a rate of 64 kbps, which are primarily used for voice connections with headsets. These different modes result in an aggregate bit rate of approximately 1 Mbps. An enhanced data rate mode provides up to 3 Mbps, while a low-energy mode significantly reduces power consumption at the expense of range or data rate.

Bluetooth uses frequency hopping for multiple access with a carrier spacing of 1 MHz. Typically, up to eighty different frequencies are used for a total bandwidth of 80 MHz. At any given time, the bandwidth available is 1 MHz. Bluetooth radios form small ad hoc networks of up to eight devices sharing the same logical channel (same hop sequence), which is called a *piconet*. Different channels (different hopping sequences) can simultaneously share the same 80 MHz bandwidth. Collisions will occur when devices in different piconets that are 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 original Bluetooth standard was developed jointly by 3 Com, Ericsson, Intel, IBM, Lucent, Microsoft, Motorola, Nokia, and Toshiba. Many additional manufactures have contributed to each new generation of the standard. Bluetooth is integrated into a wide range of electronic devices with several billion Bluetooth-enabled devices shipped annually.

The ZigBee[2] radio specification is designed for lower cost and power consumption than Bluetooth. It follows the IEEE 802.15.4 standard with device certification managed by the Zigbie Alliance. The radio operates in the same 2.4 GHz ISM band as Bluetooth. Zigbee radios support 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. Zigbee also operates in a "green" mode whereby the radio is powered through energy harvesting of its environment, reducing or in some cases completely elimiating the need for battery power. The goal of ZigBee is to provide radio operation for months or years without recharging, thereby targeting devices such as smart tags, meters, lights, and thermostats, as well as those used for sensing and automation.

Z-Wave radios are designed primarily for smart home applications with operation in the 900 MHz ISM band. Since power falls off more slowly in this band than at the higher 2.4 GHz spectrum, Z-Wave radios have a higher range than standard Zigbee or Bluetooth radios, on the order of 100 m. In addition, there is no interference between Z-Wave radios and those operating in the crowded 2.4 GHz band, including Wi-Fi as well as Bluetooth and Zigbee. On the downside, Z-Wave has significantly lower data rates than either Bluetooth or Zigbee, ranging from 10 to 100 Kbps. The Z-Wave standard and device certification is managed by the Z-Wave Alliance.

1.3 Wireless Spectrum

1.3.1 Regulation

Most countries have government agencies responsible for allocating and controlling use of the radio spectrum. In the United States, spectrum is allocated by the Federal Communications Commission (FCC) for commercial use and by the Office of Spectral Management (OSM) under the auspices of the National Telecommunications and Information Administration (NTIA) for government use. Countries throughout the world have similar regula-

[2]ZigBee takes its name from the dance that honey bees use to communicate information about newly found food sources to other members of the colony.

tory agencies to regulate spectrum within their borders. Certain regions of the world have a common regulatory agency for spectrum allocation, e.g. commercial spectral allocation across Europe is governed by the European Telecommunications Standards Institute (ETSI). Satellite systems cover large areas spanning many countries and sometimes the globe. Globally, spectrum is allocated by the International Telecommunications Union Radio Communications group (ITU-R) through its World Radiocommunication Conferences (WRC). The standards arm of this body, ITU-T, adopts telecommunication standards for global systems that must interoperate across national boundaries. Regulatory agencies typically have many competing considerations in deciding how to allocate any given block of spectrum, including whether to allocate it for commercial, military, or shared use. These decisions are generally driven by a broad mandate to regulate spectrum "in the public interest," as stated in the U.S. Communications Act of 1934 establishing the FCC.

Historically government agencies 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 spectrum around 850 MHz for analog cellular phone service, in particular 824-849 MHz for the downlink (base station to mobile) and 869-894 MHz in the uplink (mobile to base station). Spectral licenses were provided to two operators in each geographical area based on a number of criteria. While this method of licensed spectral allocation is still used in some circumstances, a fundamental shift occurred worldwide in the early 1990s for licensed spectrum to be auctioned to the highest bidder, with some restrictions in place to ensure fairness and competitive use of the spectrum. The basis for the shift was the market-based reasoning that auctions provide the fairest and most efficient way for governments to allocate the limited spectral resource and, moreover, this method provides significant revenue to the government. However, auctions are not universally supported for spectral allocation based on the contention that they can stifle innovation, limit competition, and hurt 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, sometimes indefinitely, the ability to invest in infrastructure for system rollout. Finally, high spectral cost is usually passed on to the end user. The early 3G spectral auctions, with suspected collusion between bidders, low bids, and several auction winners that ultimately defaulted, provided ammunition to the opponents of spectral auctions. Lessons learned from these early auctions were adopted in the design of subsequent auctions, which generally went smoothly and raised up to tens of billions of dollars. In addition, reverse or incentive auctions were initiated starting in 2016, whereby license holders could sell back their spectrum to regulatory bodies for future auctioning. A comprehensive treatment of spectrum regulation and its allocation through auctions can be found in [16].

In addition to spectral auctions, spectrum can be set aside in specific frequency bands, called unlicensed bands, that are free to use without a license according to a specific set of rules. The rules may correspond to a specific access protocol to ensure fairness, restrictions on power levels, and so forth. The purpose of these *unlicensed bands* is to encourage innovation and low-cost implementation. Wi-Fi is often associated with the unlicensed frequency bands, however it is just one of the hundreds of successful unlicensed systems, which include standardized short-range radio systems as well as proprietary radios in cordless phones, wireless home security systems, baby monitors, medical equipment, inventory systems, smart meters, and keyless automobile entry systems. Indeed, it is estimated that unlicensed wireless devices contribute tens of billions of dollars annually to the US economy alone. A major difficulty of unlicensed bands is that they can be killed by their own success. If many unlicensed devices on the same channel are used in close proximity then they interfere with each other, which can make the band unusable. Cellular systems can also operate in the unlicensed bands by following the unlicensed spectrum rules. However, since these systems can use their licensed bands for control, they have an advantage over systems such as Wi-Fi that use the unlicensed bands for both control and data transmission and can crowd out Wi-Fi users as a result [17]. Much of the spectrum above 30 GHz is unregulated or lightly regulated, and hence can be used by unlicensed users with little to no restriction.

Underlay systems are another alternative for allocating spectrum. An underlay system operates as a secondary

user in a licensed frequency band simultaneous with the licensed users in a manner such that the licensed users experience minimal interference from them. This is usually accomplished by spreading the signal over a very wide bandwidth, typically more than 500 MHz, and restricting its power per Hertz. The first underlay standard approved for operation was ultrawideband (UWB) communications [18]. Specifically, in 2002 the FCC approved 7500 MHz of spectrum for the use of UWB devices, with the very stringent power restriction of no more than 75 nW/MHz. This sparked regulatory activities in countries throughout Europe and Asia to also enable UWB, albeit with different restrictions than those in the US. Indeed, regulatory approval of UWB proved to be quite controversial given the complexity of characterizing how interference affects the primary band users, and the fact the UWB transmissions span many licensed users across both commercial and government domains. The regulatory challenges coupled with the severe power constraints on UWB systems ultimately proved insurmountable for most commercial systems, hence the technology failed to achieve much success [19]. The interference constraint for underlay users may alternatively be met without restricting power per Hertz by using multiple-antenna techniques to guide the underlay signals away from the spatial dimensions occupied by licensed users [20].

Following the introduction of underlay systems, regulatory bodies began exploring other innovative technologies that could make spectrum utilization in the licensed bands more flexible and efficient. This push for innovation was long overdue; other than spectral auctions and underlay systems, the basic mechanisms for licensed spectral allocation had not changed much since the inception of regulatory bodies in the early to mid-1900s. Many of the compelling ideas for exploiting technology to better utilize licensed spectrum fall under the notion of a *cognitive radio*. A cognitive radio utilizes advanced radio and signal processing technology along with novel spectrum allocation policies to support unlicensed users operating in the existing licensed spectrum, without degrading the performance of the licensed users. In particular, a cognitive radio "learns" about coexisting licensed users within its spectrum and then uses this information to utilize the spectrum without degrading the transmissions of these users [21]. Based on the nature of the coexisting user information the cognitive radio can collect, as well as a priori rules about spectrum usage, a cognitive radio seeks to *overlay* or *interweave* its signal with the transmissions of licensed nodes. Spatial underlay systems that dynamically avoid the spatial dimensions of licensed users also fall within the paradigm of cognitive radios. Cognitive radio techniques can also be applied in the unlicensed bands to reduce interference between users and thus improve spectral efficiency in these bands as well.

Interweave cognitive radios utilize unused parts of the licensed spectrum. The idea of interweave radios came about after studies conducted by the FCC and industry showed the existence of space-time-frequency voids, referred to as *spectrum holes*, in both the licensed and unlicensed bands that are not in constant use. These spectrum holes, which can change with time and geographic location, can be exploited by interweave radios to support their communication. In some cases the spectrum holes are permanent, either because a certain block of licensed spectrum is unused, or because "guard channels" between occupied channels are needed in the overall design to reduce interference between the licensed users. Such guard channels, or "white spaces," were specified in the analog television broadcast standard. As analog television bands transitioned to digital, arguments were made that such guard bands were no longer needed, and hence these white spaces could be freed up for other uses. Despite resistance from the television broadcasters, unlicensed use of television white spaces was approved by regulatory bodies in multiple countries starting in 2010. White space devices must generally consult a database of available spectrum holes in their geographical area before using them, and such databases are relatively static. A more sophisticated interweave radio looks for dynamic spectrum holes by periodically monitoring the radio spectrum, detecting spectrum holes as they occur in time, space, and/or frequency, and then using such holes opportunistically to communicate over them. Such opportunistic use of white spaces, which was the original motivation behind the notion of cognitive radio [22], has yet to be approved for spectrum allocation.

In overlay systems the cognitive transmitter has some knowledge about the transmissions of noncognitive users in the band. This knowledge can be exploited in two ways: to mitigate the interference caused to licensed users and to improve the performance of licensed users by amplifying (relaying) their signals. In particular, an

overlay user with knowledge of the licensed user's transmitted signal can use a special type of signal encoding to completely cancel the interference caused by this licensed user at its own receiver. While this coded signal will cause interference to the licensed user, by also using part of its power to amplify the licensed user's signal, the impact of this interference can be completely mitigated and, in fact, the licensed user might even experience better performance than without the existence of the overlay user. The overlay cognitive radio paradigm was originally proposed in [23] and capacity limits as well as practical implementations of overlay systems have been extensively studied since then. However, regulatory agencies have not yet considered allowing overlay systems to operate in the licensed spectrum.

Overlay, underlay, and interweave radio innovations could make spectrum utilization far more efficient than it is today, in both the licensed and unlicensed bands, thereby enabling new wireless systems, products, and applications. More details on the technologies behind these cognitive radio paradigms will be discussed in Chapter 16.7. However, even once these and other cognitive radio technologies are fully developed, changing regulatory policies to include these novel mechanisms will likely be fraught with conflicting opinions from government, industry, and the end users about what best serves the public interest in the allocation of spectrum.

1.3.2 Properties and Existing Allocations

Most wireless communication systems operate in the radio spectrum between 30 MHz and 30 GHz, with some in the millimeter wave frequency band (30 GHz-300 GHz) as well. For communication systems at frequencies below 800 MHz, antenna sizes are too large for small devices, and at frequencies above 3 GHz, signal attenuation with distance precludes long-range transmission. Given these tradeoffs, the primary frequency bands that fueled the extraordinary grown of cellular, Wi-Fi, and short-range radios like Bluetooth were in the 0.8-3 GHz range. As those frequency bands became congested, these systems moved to adjacent bands (0.6-0.8 for cellular and 3-5 GHz for cellular and Wi-Fi). Cellular and Wi-Fi systems started to expand into the millimeter wave bands as well due to the plentiful spectrum there. A similar evolution into higher frequency bands occured in satellite systems. In particular, early satellite systems used the 4-8 GHz C band, but over time moved into the X (8-12.5 GHz), Ku (12-18 GHz), K (18-26 GHz) and Ka (26-40 GHz) bands. The K bands were assigned to terrestrial fixed wireless services as well, including LMDS. Note that the required antenna size for good reception is inversely proportional to the 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 high-frequency signals.

As discussed previously, 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). Often different countries try to match their frequency bands for licensed and unlicensed use so that technology developed for that spectrum is compatible worldwide, however that isn't possible if a country has allocated one of these frequency bands to another use. Figure 1.8 shows the unlicensed spectrum allocations in the United States. In most cases there are similar frequency allocations in Europe and Asia. ISM Band I at 900 MHz has limited spectrum availability as well as licensed users transmitting at high power who interfere with the unlicensed users. Since performance in this band is somewhat poor, it is not heavily utilized by unlicensed devices. The U-NII bands have a total of 555 MHz of spectrum in four separate bands, with different power restrictions. In addition, some countries have imposed restrictions on parts of the U-NII bands so that unlicensed systems must avoid interference with radar systems licensed to operate in those bands.

Figure 1.9 shows the frequency bands allocated to current and legacy cellular systems in the United States along with their band name abbreviations[3]. Note from this table that 4G (LTE) systems operate in the 3.5 GHz

[3]Specialized Mobile Radio (SMR); Personal Communication System (PCS); Advanced Wireless Services (AWS); Digital Dividend (DD); Wireless Communications Services (WCS); Broadband Radio Service (BRS); Citizen's Broadband Radio Service (CBRS).

Band Name	Primary Use	Frequency Range (GHz)	Total Bandwidth (MHz)	Transmit Power Restriction (W)
ISM band I	Cordless phones, Z-Wave, First-generation WLANs	0.902–0.928	26	1
ISM band II	Cordless phones, Bluetooth, Zigbie, 802.11b/g	2.4–2.4835	83.5	1
ISM band III	Wi-Fi 6, 802.11a/n/ac	5.725-5.875	150	1
U-NII band I	Wi-Fi 6, 802.11a/n/ac	5.15–5.25 GHz	100	0.05
U-NII band II	Wi-Fi 6, 802.11a/n/ac	5.25–5.35 GHz	100	0.25
U-NII band III	Wi-Fi 6, 802.11a/n/ac	5.47–5.725 GHz	255	0.25
U-NII band IV	Wi-Fi 6, 802.11a/n/ac	5.725–5.825 GHz	100	1

Figure 1.8: Unlicensed Frequency Bands

CBRS band and the 5.2 GHz U-NII band, which are unlicensed bands, and hence must be shared with other unlicensed users. Prior to 4G, cellular systems operated only in dedicated licensed spectrum and hence interference was centrally managed. Under this new paradigm of spectrum sharing, the licensed and unlicensed users using the same band must manage shared access and interference between the different systems.

	2G Systems			3G Systems			4G Systems									5G Systems			
Approximate Carrier Frequency (MHz)	800	850	1900	850	1700 2100	1900	600	700	850	1700 2100	1900	2300	2500	3500	5200	600	2500	28000	39000
Band Acronym	SMR	Cellular	PCS	Cellular	AWS	PCS	DD	SMH	Cellular	AWS	PCS	WCS	BRS	CBRS	U-NII-1	DD	BRS	K	Ka
Total Bandwidth (MHz)	14	70	200	70	90	200	70	88	70	90	200	30	70	150	100	70	140	850	3000

Figure 1.9: Cellular Frequency Band Allocations in the US and their Band Abbreviations: 2G-5G Systems.

The licensed and unlicensed frequencies allocated to different wireless systems are constantly evolving to meet their capacity demands. For example, once the unlicensed spectrum in the 900 MHz and 2.5 GHz spectrum opened up, many new products and services were launched to exploit this spectrum. This led to the opening of the 3 U-NII bands for unlicensed use. TV broadcasters that were allocated spectrum in the 700 MHz band back in the 20th century were gradually moved out of these bands to open up this spectrum for cellular systems. Cellular systems are now operating in several unlicensed bands and may move into more of them. Millimeter wave spectrum above the 39 GHz Ka band is also being considered for 5G cellular systems.

1.4 Communication Standards

Communication systems that interact with each other require standardization. Standards are typically decided on by national or international committees; in the United States this role is played by the Telecommunications Industry Association (TIA) while ETSI plays this role in Europe. The IEEE is the major player for WLAN standards development through its 802.11 working group. Other IEEE 802 working groups develop standards for wireless networks, such as short-range or fixed-wireless access networks, to complement WLANs. Cellular system standards are primarily developed by 3GPP. Standards groups typically follow a lengthy process for standards development that entails input from companies and other interested parties as well as a long and detailed review process. The standards process is a large time investment, but companies participate because incorporating their ideas into the standard gives them an advantage in developing the resulting system. In general, standards do not include all the details of the system design, rather only those needed for interoperability. This allows companies to innovate and differentiate their products from other standardized systems. The main goal of standardization is

enabling systems to interoperate.

In addition to ensuring interoperability, standards also allow economies of scale and pressure prices lower. For example, WLANs typically operate in the unlicensed spectral bands, so they are not required to follow a specific standard. The first generation of WLANs were not standardized and so specialized components were needed for many systems, leading to excessively high cost that, when coupled with poor performance, led to limited adoption. This experience resulted in a strong push to standardize the next WLAN generation, which yielded the highly successful IEEE 802.11 family of standards. Adherence of products to the 802.11 standard is certified by the Wi-Fi Alliance.

There are, of course, disadvantages to standardization. The standards process is not perfect, as company participants often have their own agendas, which do not always coincide with the best technology or the best interests of consumers. In addition, the standards process must be completed at some point, after which it becomes more difficult to add new innovations and improvements to an existing standard. Finally, the standards process can become quite politicized. This happened in Europe with 1st generation cellular systems, where each country had its own standard, and with the second generation of cellular phones in the United States which ultimately adopted three different standards. The formation of the 3GPP standards body to create a single unified cellular standard throughout much of the world was a response to these pitfalls in earlier cellular standards, and is largely responsible for the massive growth in and success of cellular technology. Hence, despite its flaws, standardization is often an essential component of wireless system design and operation in order to ensure its success.

1.5 Wireless Vision

"It is always wise to look ahead, but difficult to look further than you can see." - Winston Churchill

Wireless communication is ubiquitious in the world we live in, enabling vast connectivity among people and devices as well as rapid access to information. Wireless technology impacts every aspect of modern life: culture, business, politics, economics, health, entertainment, and education. So what might the future bring?

Demand for higher data rates seems unending, hence future wireless systems could support peak speeds of hundreds or perhaps thousands of Gigabits per second. There are billions of people and locations throughout the world today without wireless (or wired) connectivity, so perhaps future systems will fill in these coverage holes so that no person or place on the planet lacks wireless connectivity. Wireless devices might shrink to such small sizes that they can be deployed within any object or living being. Some wireless devices may be powered from a very small battery or even self-power through energy harvesting or wireless charging, eliminating the need to ever plug in. Wireless technology might evolve to support the "five-nines" standard in reliability for tradition telephone service, meaning that the service is reliable 99.999% of the time in any location, indoors and out. Finally, wireless systems must be extremely secure against natural impairments as well as eavesdroppers, attackers and spoofers.

If this vision of wireless technology comes to pass, what will it enable? In addition to providing people with voice, high-speed data, and broadcast entertainment, future wireless networks will also support machine-to-machine communications for tens of billions of devices. In the home these networks will enable a new class of intelligent electronic devices that can interact with each other and with the Internet. Such "smart" homes will drastically improve energy efficiency, security, emergency response, as well as help the elderly and disabled with assisted living. Other applications of these networks include sensing and data collection in the power grid to improve robustness and efficiency, "smart cities" that provide services such as trash collection and road maintenance when the need is detected, and in-body communications for medical devices, biosensors, and targeted drug delivery. Wireless video and virtual reality will permeate the home and any place that people congregate with entertainment, and also enable remote classrooms, remote training facilities, and remote hospitals anywhere in the world. Wireless sensor networks will improve monitoring of and response to fire hazards, toxic waste sites,

stress and strain in buildings and bridges, carbon dioxide movement, and the spread of chemicals and gases at disaster sites. Finally, wireless networks with very low latency will enable distributed control systems with remote devices, sensors, and actuators linked together via wireless communication channels. Such systems will in turn enable intelligent transportation systems including self-driving vehicles, mobile robots and drones, as well as easily reconfigurable industrial automation.

The exponential growth of smartphone use and wireless Internet access has 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 the first generation of wireless LANs and LEO satellite systems, as well as wide area data services, and fixed wireless access 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 – ensure a bright future for wireless technology.

1.6 Technical Challenges

This section provides an overview of the many technical challenges that must be addressed to make the wireless vision a reality. These challenges extend across all aspects of the system, including hardware design, channel characterization, physical layer and multiple access techniques as well as networking protocols and architectures. Techniques to address many of these challenges are described in subsequent chapters of the book.

The design of wireless systems begins with a model for the underlying channel through which the signals will propagate. In all such channels, signal power decreases with distance due to the physics of propagation as well as attenuation from blocking objects. These signal propagation models are developed in Chapter 2. If the transmitter, receiver, or surrounding objects are moving, the channel changes randomly with time due to changing reflections and attenuation. These random channel variations, whose statistical models are developed in Chapter 3, make it difficult to design reliable systems with guaranteed performance. Channel characteristics, including signal attenuation, also depend on the frequency of operation. In particular, received power generally decreases with the carrier frequency and, in the case of free space propagation with omnidirectional antennas, it is inversely proportional to the square of this frequency. Thus, most wireless systems today operate at carrier frequencies below 5 GHz to ensure good coverage, leading to a spectrum shortage in this range of frequencies. Moving to higher carrier frequencies, such as millimeter wave (30-300 GHz) or terahertz (.3-3 THz), provides much more spectrum than what is available in the lower frequency bands. However, these higher frequencies of operation reduce range unless energy is directionally focused using multiple or directional antenna techniques. Signal propagation characteristics at these frequencies create challenges in designing reliable communication links, as described in Chapter 2.9.5. In addition, hardware components are expensive and power hungry.

The maximum data rate that can be reliably sent over a wireless (or wireline) channel is its Shannon capacity, which is derived in Chapter 4. This capacity is directly proportional to the channel bandwidth, i.e. the amount of spectrum allocated to the channel. This rate also depends on the number of antennas at the transmitter and receiver, as multiple-input multiple-output (MIMO) techniques allow for independent data streams to be transmitted along the independent spatial dimensions these multiple antennas create, as described in Chapter 10. Achievable data rates for a given wireless system also depend on signal propagation and interference characteristics. Due to these challenges, data rates for both cellular and Wi-Fi systems are several orders of magnitude lower than for a fiber optic cable, but that could change as millimeter wave systems are deployed.

In terms of hardware challenges, as the size of wireless devices shrink, breakthroughs are needed to make both the analog and digital circuitry significantly smaller, cheaper, and more energy efficient. Wireless systems operating at millimeter wave and higher frequencies require cheap and reliable RF components, which remains a

significant challenge. The size, power consumption, and precision of analog-to-digital converters is also becoming a bottleneck as systems move to larger bandwidths. Finally, large antenna arrays that improve signal propagation and mitigate interference require hardware innovations, including hybrid analog and digital processing, to reduce their footprint, power consumption, and cost.

The physical layer design of wireless systems today is quite sophisticated, with dense-constellation modulation (Chapters 5-6), diversity and adaptive modulation techniques to combat random signal variations (Chapters 7 and 9), powerful error-correction coding (Chapter 8), as well as techniques to combat intersymbol interference caused by delayed signal components that arise from channel reflections (Chapters 11-13). Systems operating in rapidly changing environments require new modulation, coding, detection, and multi-antenna techniques that are robust to such changes when adaptation to them is infeasible. Machine learning may play a role in improving physical layer techniques for channel estimation, signal detection, decoding, and equalizaiton, particularly when channels are hard to model, hard to estimate, or rapidly varying [24, 25]. In addition, algorithms that best exploit the many spatial degrees of freedom offered by large antenna arrays are needed. For energy-constrained systems, particularly those running off non-rechargeable batteries, communication schemes must be developed that can meet performance requirements while minimizing total system power consumption (for signal transmission, reception, and processing).

Multiple access techniques, developed in Chapter 14, allow users to share the same system bandwidth. This bandwidth sharing is done either through coordinated access, as in cellular systems and current Wi-Fi standards, or through distributed techniques, as used in early generations of Wi-Fi systems. Most wireless systems use access schemes that assign orthogonal time and frequency slices of the total system bandwidth to different users. For MIMO systems, independent spatial dimensions can be used as separate channels as well. Non-orthogonal access designs improve spectral efficiency by overlapping channels in the time, frequency, code, or spatial dimensions while managing the interference that results. Channel assignment is done by the base stations in cellular systems and by access points in Wi-Fi systems. Wireless access techniques are ripe for innovation, in the centralized and distributed mechanisms used to assign channels as well as in slicing up channels in a non-orthogonal manner to increases spectral efficiency with minimal interference. Machine learning is also being applied to multiple access, resource allocation, and scheduling [24].

The network architectures and protocols for future wireless systems must support a much wider range of devices and applications than exist today. For infrastructure-based systems like cellular networks (Chapter 15), this will require a more flexible architecture, with a range of cell sizes, greater densification of cells, as well as significant data processing, dynamic optimization, and resource allocation that may be centralized, decentralized, or centralized within a subsets of nodes (coined neighborhood or "fog-based" optimization). Rethinking cellular system architectures to take into account novel forms of cooperation across base stations and users to exploit rather than mitigate interference may lead to significantly higher capacity and robustness. Reduction of latency and overhead for applications with short delay-constrained messages is also needed, as are innovations in energy-efficient architectures, e.g. for systems whose backbone is powered by batteries, solar, other forms of energy harvesting, or wireless power transfer. In ad hoc wireless networks (Chapter 16), advances in cooperative techniques for transmission, reception, and relaying have the potential to significantly improve coverage, reliability, and data rates.

Seamless handoff between different wireless networks, such as Wi-Fi and cellular, is precluded by existing protocols and addressing mechanisms, which must evolve to allow a device to continuously connect on any and all available networks. How best to utilize a multiplicity of available networks for a given device or application is an open challenge. Many wireless systems consist of predominantly wired links, with wireless as the last hop only. Hence, challenges in wired networking, such as latency and bandwidth shortages, impact wireless networks as well. Deployment of caching and computation capabilities in base stations and Wi-Fi access points give rise to new design challenges; should applications utilize these edge capabilities or rely on more powerful capabilities situated farther from the network edge, which entail more latency to access. Security is also more difficult to implement for

signals traveling over wireless channels, since the airwaves are susceptible to snooping and jamming by anyone with an RF antenna.

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 or drone and driving or flying 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 $P_b = 1/2\overline{\gamma}$ 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. Because of this extremely 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 $3 \cdot 10^8$ m/s. If the maximum acceptable delay for a voice system is 30 ms, which of these satellite systems would be acceptable for two-way voice communication?

5. What applications might significantly increase the demand for wireless data?

6. This problem illustrates some of the economic issues facing service providers for mixed-media systems. Suppose you are a service provider with 120 kHz of bandwidth that you must allocate between voice and data users. The voice users require 20 kHz of bandwidth and the data users require 60 kHz of bandwidth. So, for example, you could allocate all of your bandwidth to voice users, resulting in six voice channels, or you could divide the bandwidth into 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 .2 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 \$.50 if his call is processed. How should you allocate your bandwidth to maximize your expected revenue?

7. Describe three disadvantages of using a fixed wireless access systems instead of DSL or cable. Describe three scenarios where the disadvantages override the advantages.

8. Cellular systems have migrated to Hetnets consisting of a mix of macrocells and small cells in order to increase system capacity and energy efficiency. Name at least three design issues that are complicated by this trend.

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

10. This problem demonstrates the capacity increase associated with a decrease in cell size. Consider a square city of 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 and 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 km^2?

(b) What cell size would you use if your system had to 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 P.M. 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 so that subscribers need tolerate no greater than a 2% blocking probability during this peak hour. (Blocking probability is computed using the Erlang B model: $P_b = (A^C/C!)/\left(\sum_{k=0}^{c} A^k/k!\right)$, 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 per user, and H the average duration of a call [5, Chap. 3.6].

(c) How many total subscribers can be supported in the macrocell system (1-km^2 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 the monthly user fee in each system is $50, what will be the monthly revenue in each case? How long will it take to recoup the infrastructure (base station) cost for each system?

Bibliography

[1] G.R.M. Garratt, The Early History of Radio: From Faraday to Marconi, IEEE History of Technology Series, 1994.

[2] S. M. Cherry, "What went wrong at Ricochet?," *IEEE Spectrum*, Vol. 39, No. 3, pp. 60-61, March 2002.

[3] *The Innovation Journey of Wi-Fi: The Road to Global Success,* Edited by W. Lemstra, V. Hayes and J. Groenewegen, Cambridge University Press, 2011.

[4] D. H. Ring, "Mobile Telephony - Wide Area Coverage," *Bell System Tech. Memo*, Dec. 1947.

[5] "High Capacity Mobile Telephone System Feasibility Studies and System Plan" J. Engel, R. Frenkiel, and P. Porter, FCC Proposal, 1971.

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

[7] P. Nelson, "A new breed of broadband satellites could have you living on a desert island", *Network World*, July 2014.

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

[9] M. Sauter *From GSM to LTE-Advanced Pro and 5G: An Introduction to Mobile Networks and Mobile Broadband*, 3rd ed., Wiley, 2017.

[10] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, Academic Press, 2018.

[11] E. Perahia and R. Stacey, Next Generation Wireless LANs: 802.11n and 802.11ac, 2nd ed., Cambridge University Press, 2013.

[12] C. Wang, T. Jiang, and Q Zhang, ZigBee Network Protocols and Applications, CRC Press, 2014.

[13] K. Townsend, C. Cufi, Akiba, and R. Davidson, Getting Started with Bluetooth Low Energy: Tools and Techniques for Low-Power Networking, O'Reilly Media, 2014.

[14] J. Pelton, S. Madry, and S. Camacho-Lara (eds), *Handbook of Satellite Applications*. Springer, Cham, 2017.

[15] S. K. Sharma, S. Chatzinotas, and P.-D. Arapoglou (eds.), *Satellite Communications in the 5G Era Year,* IET Telecommunication Series, 2018

[16] M. Cave and W. Webb, *Spectrum Management - Using the Airwaves for Maximum Social and Economic Benefit,* Cambridge University Press, 2016.

[17] F. M. Abinader et al., "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 54-61, Nov. 2014.

[18] S. Emami, *UWB Communication Systems: Conventional and 60 GHz: Principles, Design and Standards,* Springer, 2013.

[19] L. Frenzel, "What ever happened to ultrawideband wireless," *Electronic Design,* Sep. 2014.

[20] R. Zhang and Y. C. Liang, Exploiting multi-antennas for opportunistic spectrum sharing in cognitive radio networks, *IEEE J. Sel. Topics Signal Proc.*, vol. 2, no. 1, pp. 88-102, Feb. 2008.

[21] E. Biglieri, A. J. Goldsmith, L. J. Greenstein, N. B. Mandayam, and H. V. Poor, *Principles of Cognitive Radio,* Cambridge University Press, 2012.

[22] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Commun.*, Vol. 6, No. 4, pp. 13-18, Aug. 1999.

[23] N. Devroye, P. Mitran, and V. Tarokh, Achievable Rates in Cognitive Radio Channels, *Proc. 39th Annual Conf. Info. Sci. and Sys.*, Mar. 2005.

[24] D. Gündüz, Paul de Kerret, Nicholas D. Sidiropoulos, David Gesbert, Chandra Murthy, Mihaela van der Schaar, "Machine Learning in the Air," Arxiv: https://arxiv.org/pdf/1904.12385.pdf.

[25] O. Simeone, A very brief introduction to Machine Learning with applications to communication systems, *IEEE Trans. on Cognitive Commun. and Networking*, vol. 4, no. 4, pp. 648664, Dec. 2018.

Chapter 2

Path Loss, Shadowing, and Multipath

The wireless radio channel poses a severe challenge as a medium for reliable high-speed communication. Not only is it susceptible to noise, interference, and other channel impediments, but these impediments change over time in unpredictable ways as a result of user movement and environment dynamics. In this chapter we characterize the primary phenomena that affect signal propagation: path loss and shadowing, as well as signal reflection, diffraction, and scattering. Path loss characterizes how a signal's received power decreases with transmit-receive distance. It is caused by dissipation of the power radiated by the transmitter as well as by effects of the propagation channel. Path-loss models assume that path loss is the same at a given transmit–receive distance (assuming that the path-loss model does not include shadowing effects). Shadowing is the attenuation caused by obstacles between the transmitter and receiver that absorb the transmitted signal. When the attenuation is strong, the signal is blocked. The number and type of objects that cause shadowing at any given receiver location is typically unknown. Hence attenuation due to shadowing is modeled as a random parameter. Unlike path loss, shadowing does not depend on the transmit-receive distance itself but rather on the objects between the transmitter and receiver. Reflection, diffraction, and scattering are caused by a transmitted signal interacting with objects in the environment around the transmitter or receiver. The signal components that arise due to these objects are called *multipath* components. Different multipath components arrive at the receiver with different time delays and phase shifts. When the phase shifts are aligned, the multipath components add constructively; when they are not aligned, they add destructively. This constructive and destructive addition of multipath components leads to significant variations in the received signal power.

Received power variation due to path loss occurs over long distances (100-1000 m), whereas variation due to shadowing occurs over distances that are proportional to the length of the obstructing object (10-100 m in outdoor environments and less in indoor environments). Since variations in received power due to path loss and shadowing occur over relatively large distances, these variations are sometimes referred to as *large-scale propagation effects*. The received power variations due to constructive and destructive addition of multipath components occur over very short distances, on the order of the signal wavelength, since each component's phase rotates 360 degrees over that distance. Hence, power variations due to multipath are sometimes referred to as *small-scale propagation effects*. Figure 2.1 shows an example of the received-to-transmit power ratio in decibels[1] (dB) versus $\log$ distance for the combined effects of path loss, shadowing, and multipath. As indicated in the figure, received power due to path loss alone is generally modeled as decreasing linearly with respect to the log of the transmit-receive distance, with additional slow variations due to shadowing and fast variations due to multipath.

After a brief introduction to propagation and a description of our signal model, we present the simplest model for signal propagation: free-space path loss. A signal propagating between two points with no attenuation or reflection follows the free-space propagation law. We then describe the two-ray multipath model, which augments

[1] The decibel value of x is $10 \log_{10} x$.

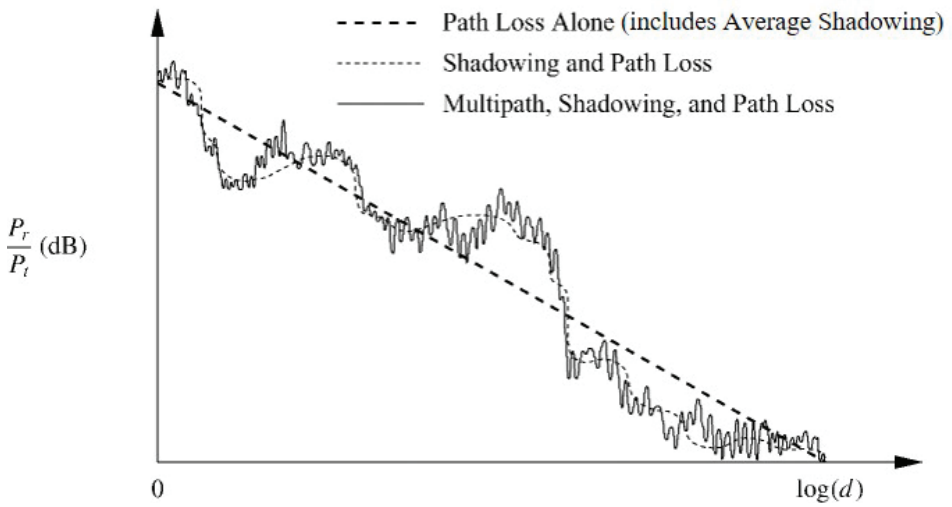

Figure 2.1: Effects of path loss, shadowing, and multipath on received power as a function of distance.

the free-space model with a single reflected ray. The two-ray model introduces the notion of a variable path-loss exponent, which gives rise to more general path loss exponent models. After introducing these models, we discuss the commonly-used log-normal model for shadowing. General ray tracing is then introduced to model the multipath components that arise due to signal reflections, diffraction, and scattering. These models approximate signal propagation according to Maxwell's equations and depend heavily on the geometry and dielectric properties of the region through which the signal propagates. If the number of multipath components is large or if the geometry and dielectric properties of the propagation environment are unknown, then statistical multipath models must be used instead of ray tracing. These statistical multipath models will be described in Chapter 3. We close the chapter by describing empirical channel models with parameters for path loss and shadowing based on measurements for both indoor and outdoor channels.

Although this chapter gives a brief overview of channel models for path loss, shadowing, and multipath, comprehensive coverage of channel and propagation models at different frequencies of interest merits a book in its own right, and in fact there are many excellent references on this topic including [1, 2, 3, 4, 10, 6]. Models specialized to multiple antenna, ultrawideband, and millimeter wave channels can be found in [7], [8], and [9], respectively.

2.1 Radio Wave Propagation

The initial understanding of radio wave propagation goes back to the pioneering work of James Clerk Maxwell, who in 1864 formulated a theory of electromagnetic propagation that predicted the existence of radio waves. In 1887, the physical existence of these waves was demonstrated by Heinrich Hertz. However, Hertz saw no practical use for radio waves, reasoning that since audio frequencies were low, where propagation was poor, radio waves could never carry voice. In 1894 Oliver Lodge used these principles to build the first wireless communication system, though its transmission distance was limited to 150 meters. By 1897 the entrepreneur Guglielmo Marconi had managed to send a radio signal from the Isle of Wight to a tugboat eighteen miles away, and in 1901 Marconi's wireless system could traverse the Atlantic ocean. These early systems used telegraph signals for communicating information. The first transmission of voice and music was made by Reginald Fessenden in 1906 using a form of amplitude modulation, which circumvented the propagation limitations at low frequencies observed by Hertz by translating signals to a higher frequency, as is done in all wireless systems today.

Electromagnetic waves propagate through environments where they are reflected, scattered, and diffracted by walls, terrain, buildings, and other objects. The ultimate details of this propagation can be obtained by solving Maxwell's equations with boundary conditions that express the physical characteristics of these obstructing objects. This often requires the calculation of the radar cross-section (RCS) of large and complex structures. Since these calculations are difficult and since the necessary parameters are often not available, approximations have been developed to characterize signal propagation without resorting to Maxwell's equations.

The most common signal propagation approximations use ray-tracing techniques based on ray-optic theory [10]. Ray-tracing approximates the propagation of electromagnetic waves by representing the wavefronts as discrete narrow beams or rays. This approximation determines the reflection and refraction effects on the wavefront but ignores the more complex scattering phenomenon predicted by Maxwell's coupled differential equations. The ray-tracing model most accurately approximates Maxwell's equations when the wavelength of the signal is much less than the size of the objects off of which it is reflected, refracted, or scattered. The simplest ray-tracing model is the two-ray model, which accurately describes signal propagation when there is one direct path between the transmitter and receiver and one reflected path. The reflected path typically bounces off the ground, and the two-ray model is a good approximation for propagation along highways or rural roads and over water. We will analyze the two-ray model in detail, as well as more complex models with additional reflected, scattered, or diffracted components. Many propagation environments are not accurately characterized by ray-tracing models. In these cases it is common to develop analytical models based on empirical measurements, and we will discuss several of the most common of these empirical models.

Often the complexity and variability of the radio channel make it difficult to obtain an accurate deterministic channel model. For these cases, statistical models are often used. The attenuation caused by signal path obstructions such as buildings or other objects is typically characterized statistically, as described in Section 2.7. Statistical models are also used to characterize the constructive and destructive interference for a large number of multipath components, as described in Chapter 3. Indoor environments tend to be less regular than outdoor environments, since the geometric and dielectric characteristics change dramatically depending on whether the indoor environment is an open factory, cubicled office, or metal machine shop. For these environments computer-aided modeling tools are available to predict signal propagation characteristics [11].

2.2 Transmit and Receive Signal Models

The transmitted and received signals in any wireless system are real-valued. The channel introduces an amplitude and phase change at each frequency of the transmitted signal so that the received signal is also real-valued. Real modulated and demodulated signals are often represented as the real part of a complex signal in order to facilitate analysis. This model gives rise to the equivalent lowpass representation of bandpass signals, which we use for our transmitted and received signals. More details on the equivalent lowpass representation of bandpass signals and systems can be found in Appendix A.

We model the transmitted signal at carrier frequency f_c as

$$\begin{aligned} s(t) &= \mathrm{Re}\{u(t)e^{j2\pi f_c t}\} \\ &= \mathrm{Re}\{u(t)\}\cos(2\pi f_c t) - \mathrm{Im}\{u(t)\}\sin(2\pi f_c t) \\ &= s_I(t)\cos(2\pi f_c t) - s_Q(t)\sin(2\pi f_c t), \end{aligned} \tag{2.1}$$

where $u(t) = s_I(t) + j s_Q(t)$ is a complex baseband signal with in-phase component $s_I(t) = \mathrm{Re}\{u(t)\}$, quadrature component $s_Q(t) = \mathrm{Im}\{u(t)\}$, bandwidth B_u, and power P_u. The signal $u(t)$ is called the *complex envelope* or *equivalent lowpass signal* of $s(t)$. We call $u(t)$ the complex envelope of $s(t)$ because the magnitude of $u(t)$ is the magnitude of $s(t)$. The phase of $u(t)$ includes any carrier phase offset. The equivalent lowpass representation of

band-pass signals with bandwidth $B \ll f_c$ allows signal manipulation via $u(t)$ irrespective of the carrier frequency. The power in the transmitted signal $s(t)$ is $P_t = P_u/2$.

For time-invariant channels, the received signal is the convolution of $s(t)$ with the channel impulse response $h(t)$ plus an additional noise component $n(t)$ introduced by the channel: $r(t) = s(t) * h(t) + n(t)$. It can be written in a similar form as the transmitted signal as:

$$r(t) = \mathrm{Re}\{v(t)e^{j2\pi f_c t}\} + n(t), \tag{2.2}$$

where $v(t)$ is the equivalent lowpass signal for the received signal without the noise. This equivalent signal depends on $h(t)$. In particular, as discussed in Appendix A, $v(t) = u(t) * c(t)$, where $c(t)$ is the equivalent lowpass channel impulse response for $h(t)$. Time-varying channels will be treated in Chapter 3.

The received signal in (2.2) consists of two terms, the first term corresponding to the transmitted signal after propagation through the channel, and the second term corresponding to the noise added by the channel. The signal-to-noise power ratio (SNR) of the received signal is defined as the power of the first term divided by the power of the second term. In this chapter (and in Chapter 3) we will neglect the random noise component $n(t)$ in our analysis, since these chapters focus on signal propagation, which is not affected by noise. However, noise will play a prominent role in the capacity and performance of wireless systems studied in later chapters.

When the transmitter or receiver is moving, the received signal will have a Doppler shift of $f_D = \frac{v}{\lambda}\cos\theta$ associated with it, where θ is the arrival angle of the received signal relative to the direction of motion, v is the receiver velocity toward the transmitter in the direction of motion, and $\lambda = c/f_c$ is the signal wavelength ($c = 3 \cdot 10^8$ m/s is the speed of light). The geometry associated with the Doppler shift is shown in Figure 2.2. The Doppler shift results from the fact that transmitter or receiver movement over a short time interval Δt causes a slight change in distance $\Delta d = v\Delta t \cos\theta$ that the transmitted signal needs to travel to the receiver. The phase change due to this path-length difference is $\Delta\phi = \frac{2\pi v}{\lambda}\Delta t \cos\theta$. The Doppler frequency is then obtained from the relationship between signal frequency and phase:

$$f_D = \frac{1}{2\pi}\frac{\Delta\phi}{\Delta t} = \frac{v}{\lambda}\cos\theta. \tag{2.3}$$

If the receiver is moving toward the transmitter (i.e., if $-\pi/2 \leq \theta \leq \pi/2$) , then the Doppler frequency is positive; otherwise, it is negative. We will ignore the Doppler term in the path loss models of this chapter, since for typical vehicle speeds (75 km/hr) and frequencies (about 10 GHz) it is small, on the order of 1 KHz [1]. However, we will include Doppler effects in Chapter 3 on statistical fading models where it is used to characterize a random channel's rate of change.

Suppose a signal $s(t)$ of power P_t is transmitted through a given channel with corresponding received signal $r(t)$ of power P_r, where P_r is averaged over any random variations due to shadowing. We define the *linear path loss* of the channel as the ratio of transmit power to receive power:

$$P_L = \frac{P_t}{P_r}. \tag{2.4}$$

We define the *path loss* of the channel as the value of the linear path loss in decibels or, equivalently, the difference in dB between the transmitted and received signal power:

$$P_L \text{ dB} = 10\log_{10}\left(\frac{P_t}{P_r}\right) \text{ dB}. \tag{2.5}$$

The dB path loss is a nonnegative number since the channel does not contain active elements, and thus it can only attenuate the signal. The dB *path gain* is defined as the negative of the dB path loss: $P_G = -P_L = 10\log_{10}(P_r/P_t)$ dB. Due to the laws of physics underlying signal propagation, the dB path loss generally increases

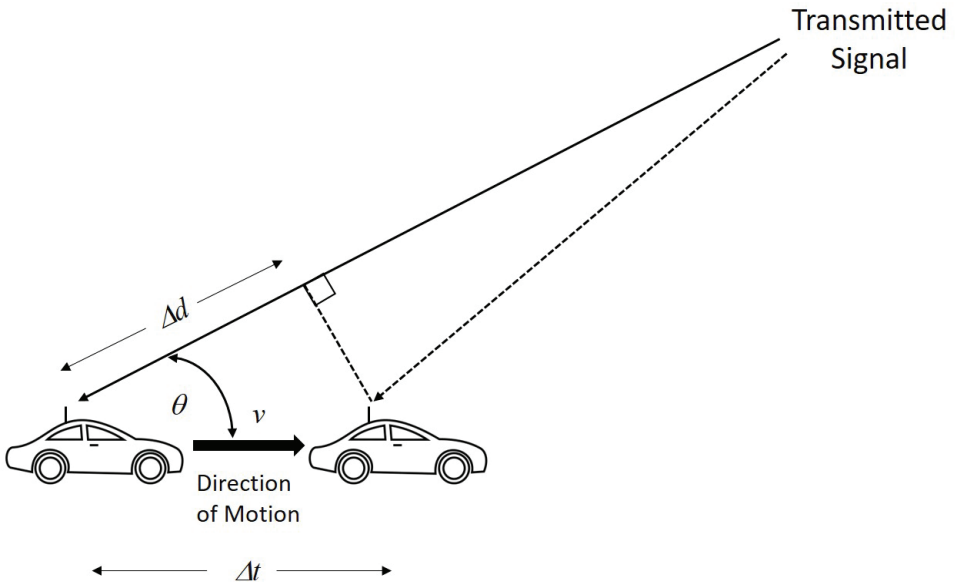

Figure 2.2: Geometry associated with Doppler shift.

with distance, as illustrated in Figure 2.1 (note that this figure shows a decreasing path gain or, equivalently, an increasing path loss). The path loss generally increases with distance, which means the received power in dB decreases with distance, as illustrated in Figure 2.1. With shadowing, also illusrated in Figure 2.1, the received power is random owing to random blockage from objects, as we discuss in Section 2.6.

2.3 Free-Space Path Loss

Consider a signal transmitted through free space to a receiver located at distance d from the transmitter. Assume there are no obstructions between the transmitter and receiver and that the signal propagates along a straight line between the two. The channel model associated with this transmission is called a line-of-sight (LOS) channel, and the corresponding received signal is called the LOS signal or ray. Under free-space path loss the received signal is given by [2]:

$$r(t) = \mathrm{Re}\left\{\left[\frac{\lambda\sqrt{G_t G_r}u(t-\tau_l)e^{-j2\pi d/\lambda}}{4\pi d}\right]e^{j2\pi f_c t}\right\}, \tag{2.6}$$

where G_t and G_r are, respectively, the transmit and receive antenna power gains in the LOS direction relative to a unity gain isotropic antenna, $\tau_l = d/c$ is the signal propagation delay of the LOS signal traveling the distance d, and the phase shift $e^{-j2\pi d/\lambda}$ is due to the distance d that the wave travels[2]. Transmit and receive directional antennas have gains G_t and G_r greater than unity in one or more angular directions relative to the idealized isotropic antenna gains $G_t = G_r = 1$. Directional antenna gains can range from 2.15 dB for a half-wavelength dipole to tens of dB in horn or dish antennas. More details on directional antenna designs and their gains can be found in [12].

[2]When the transmit and receive antennas are at the same height, the distance d equals the horizontal separation distance between the transmitter and receiver. When the transmitter and receiver are at different heights, the distance the wave travels exceeds this separation distance. These different distances will be characterized in the two-ray model (Section 2.4).

The power in the transmitted signal $s(t)$ is P_t, so from (2.6), the linear path loss for free-space propagation, a formula first introduced by Friis [13], is

$$\frac{P_r}{P_t} = G_t G_r \left[\frac{\lambda}{4\pi d}\right]^2. \tag{2.7}$$

The Friis formula (2.7) dictates that the receive power P_r equals the transmit power per unit area under free space propagation that is incident on the receive antenna, given by $P_t G_t/(4\pi d^2)$, multiplied by that antenna's *effective area* A_r, a quantity that determines how effectively the receive antenna captures this incident power. From (2.7) this effective area equals

$$A_r = \frac{\lambda^2 G_r}{4\pi}. \tag{2.8}$$

The dependence of A_r on λ^2 is due to the resonance of a signal with a given wavelength on a linear conductor half that size. Since $\lambda = c/f_c$, the Friis formulae indicates that, as the carrier frequency increases, the received power decreases as the square of the increase. Antenna arrays, discussed in Chapter 10, create a different effective area than single-element antennas, which can mitigate or remove the dependence of received power on λ [14]. The Friis formula also indicates that the received signal power falls off in inverse proportion to the square of the distance d between the transmit and receive antennas. We will see in the next section that, for other signal propagation environments, the received signal power falls off more quickly relative to this distance.

The received power can be expressed in dBm as[3]

$$P_r\ (\text{dBm}) = P_t\ (\text{dBm}) + 10\log_{10}(G_t G_r) + 20\log_{10}(\lambda) - 20\log_{10}(4\pi) - 20\log_{10}(d). \tag{2.9}$$

Equation (2.9) is an example of a *link budget equation*, which expresses the received power of a signal transmitted through a given channel (or link) as a function of the transmit power and all the losses the signal experiences on that link. *Free-space path loss* is defined as the path loss of the free-space model:

$$P_L\ (\text{dB}) = 10\log_{10}\left(\frac{P_t}{P_r}\right) = -10\log_{10}\left(G_t G_r \left[\frac{\lambda}{4\pi d}\right]^2\right). \tag{2.10}$$

The *free-space path gain* is thus

$$P_G = -P_L = 10\log_{10}\left(G_t G_r \left[\frac{\lambda}{4\pi d}\right]^2\right). \tag{2.11}$$

Example 2.1: Consider an outdoor small cell with $f_c = 2.5$ GHz, cells of radius 10 m, and isotropic antennas. Under the free-space path loss model, what transmit power is required at the base station in order for all terminals within the cell to receive a minimum power of $0.1\ \mu$W? How does this change if the system frequency is 5 GHz?

Solution: We must find a transmit power such that the terminals at the cell boundary receive the minimum required power. We obtain a formula for the required transmit power by inverting (2.7) to obtain:

$$P_t = P_r \left[\frac{4\pi d}{\sqrt{G_t G_r}\lambda}\right]^2.$$

[3]The dBm value of a power x is its dB value relative to a milliwatt: $10\log_{10}(x/.001)$. The dBW value of a power x is its dB value relative to a watt: $10\log_{10} x$, so 1W corresponds to 30 dBm.

Substituting in $G_t = G_r = 1$ (isotropic antennas), $\lambda = c/f_c = 0.12$ m, $d = 10$ m, and $P_r = 0.1$ μW yields $P_t = .1097$ W. At 5 GHz only $\lambda = .06$ changes, so $P_t = .4388$ W. We see that doubling the carrier frequency leads to a requirement for four times more transmit power, illustrating the power consumption challenges in moving wireless systems to high-frequency spectral bands.

2.4 Two-Ray Multipath Model

The two-ray model is used when a single ground (or other) reflection dominates the multipath effect, as illustrated in Figure 2.3. The received signal consists of two components: the LOS component or ray, which corresponds to the transmitted signal propagating through free space, and a reflected component or ray, which is the transmitted signal reflected off the ground. The received LOS ray is given by the free-space propagation loss formula (2.6) with the distance the signal travels set to d_0. The reflected ray shown in Figure 2.3 travels distance $d_1 = d_{11} + d_{12}$. Since the two signal components in this model travel different distances, we use d in this model and all subsequent propagation models in this chapter to denote the horizontal distance between the transmitter and receiver. The distances d_0 and d_1 then depend on d as well as the transmitter and receiver heights h_t and h_r. We see from Figure 2.3 that when $h_t = h_r$, $d = d_0$, and when d is very large relative to h_t and h_r, $d \approx d_0 \approx d_1$.

If we ignore the effect of surface wave attenuation[4] then, by superposition, the received signal for the two-ray model is

$$r_{\text{2-ray}}(t) = \text{Re}\left\{\frac{\lambda}{4\pi}\left[\frac{\sqrt{G_0}u(t-\tau_0)e^{-j2\pi d_0/\lambda}}{d_0} + \frac{R\sqrt{G_1}u(t-\tau_1)e^{-j2\pi d_1/\lambda}}{d_1}\right]e^{j2\pi f_c t}\right\}, \tag{2.12}$$

where $\tau_0 = d_0/c$ is the time delay of the LOS ray, $\tau_1 = d_1/c$ is the time delay of the ground reflection ray, $\sqrt{G_0} = \sqrt{G_{t_0}G_{r_0}}$ is the product of the transmit and receive antenna field radiation patterns in the LOS direction, R is the ground reflection coefficient, and $\sqrt{G_1} = \sqrt{G_{t_1}G_{r_1}}$ is the product of the transmit and receive antenna field radiation patterns corresponding to the ground reflection. The *delay spread* of the two-ray model equals the difference between the delay of the LOS ray and that of the reflected ray: $(d_1 - d_0)/c$.

If the transmitted signal is narrowband relative to the delay spread such that $(\tau_1 - \tau_0 \ll B_u^{-1})$ then $u(t-\tau_1) \approx u(t-\tau_0)$. With this approximation, the received power of the two-ray model for narrowband transmission is

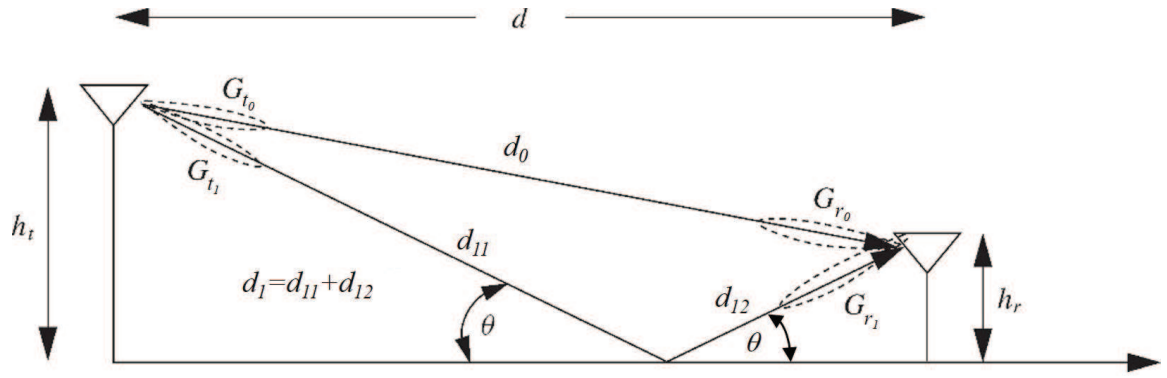

Figure 2.3: Two-ray model.

[4]This is a valid approximation for antennas located more than a few wavelengths from the ground.

$$P_r = P_t \left[\frac{\lambda}{4\pi}\right]^2 \left|\frac{\sqrt{G_0}}{d_0} + \frac{R\sqrt{G_1}e^{-j\Delta\phi}}{d_1}\right|^2, \tag{2.13}$$

where $\Delta\phi = 2\pi(d_1 - d_0)/\lambda$ is the phase difference between the two received signal components. Equation (2.13) has been shown [15] to agree closely with empirical data. From the geometry of the two-ray model, $d_1 = \sqrt{(h_t + h_r)^2 + d^2}$ and $d_0 = \sqrt{(h_t - h_r)^2 + d^2}$. Hence

$$d_1 - d_0 = \sqrt{(h_t + h_r)^2 + d^2} - \sqrt{(h_t - h_r)^2 + d^2}. \tag{2.14}$$

When d is very large compared to $h_t + h_r$, we can use a Taylor series approximation in (2.14) to get

$$\Delta\phi = \frac{2\pi}{\lambda}\left(\sqrt{(h_t + h_r)^2 + d^2} - \sqrt{(h_t - h_r)^2 + d^2}\right) \approx \frac{4\pi h_t h_r}{\lambda d}. \tag{2.15}$$

The ground reflection coefficient is given by [1, 16]

$$R = \frac{\sin\theta - Z}{\sin\theta + Z}, \tag{2.16}$$

where

$$Z = \begin{cases} \sqrt{\varepsilon_r - \cos^2\theta}/\varepsilon_r & \text{for vertical polarization,} \\ \sqrt{\varepsilon_r - \cos^2\theta} & \text{for horizontal polarization,} \end{cases} \tag{2.17}$$

and ε_r is the dielectric constant of the ground. For earth or road surfaces this dielectric constant is approximately that of a pure dielectric (for which ε_r is real with a value of about 15).

Since for asymptotically large d, $d_1 \approx d_0 \approx d$, $G_1 \approx G_0$ where we denote this approximate antenna gain as G, and $\theta \approx 0$ which by (2.16) then implies that $R \approx -1$. Substituting these approximations into (2.13) yields

$$P_r \approx P_t \left[\frac{\lambda\sqrt{G}}{4\pi d}\right]^2 |1 - e^{-j\Delta\phi}|^2. \tag{2.18}$$

Using the approximation for $\Delta\phi$ in (2.15) at large d we get the final approximation for received power at large distance d as

$$P_r \approx P_t \left[\frac{\lambda\sqrt{G}}{4\pi d}\right]^2 \left[\frac{4\pi h_t h_r}{\lambda d}\right]^2 = P_t \left[\frac{\sqrt{G}h_t h_r}{d^2}\right]^2, \tag{2.19}$$

or, in dB,

$$P_r\,\text{dBm} = P_t\,\text{dBm} + 10\,\log_{10}(G) + 20\,\log_{10}(h_t h_r) - 40\,\log_{10}(d). \tag{2.20}$$

Thus, in the limit of asymptotically large d, the received power falls off inversely with the fourth power of d and is independent of the wavelength λ. This is in contrast to free-space propagation where power is inversely proportion to d^2 and proportional to λ^2. The received signal power becomes independent of λ because directional antenna arrays have a received power that does not necessarily decrease with λ, and combining the direct path and reflected signal at the receiver effectively forms an antenna array.

A plot of (2.13) as a function of the transmit-receive separation distance d is shown in Figure 2.4 for $f_c = 900$ MHz, $R = -1$, $h_t = 50$ m, $h_r = 2$ m, $G_0 = G_1 = 1$, and received power normalized to 0 dB at a reference distance of 1 m. This plot reflects the case where the transmit antenna height h_t is much larger than the receiver

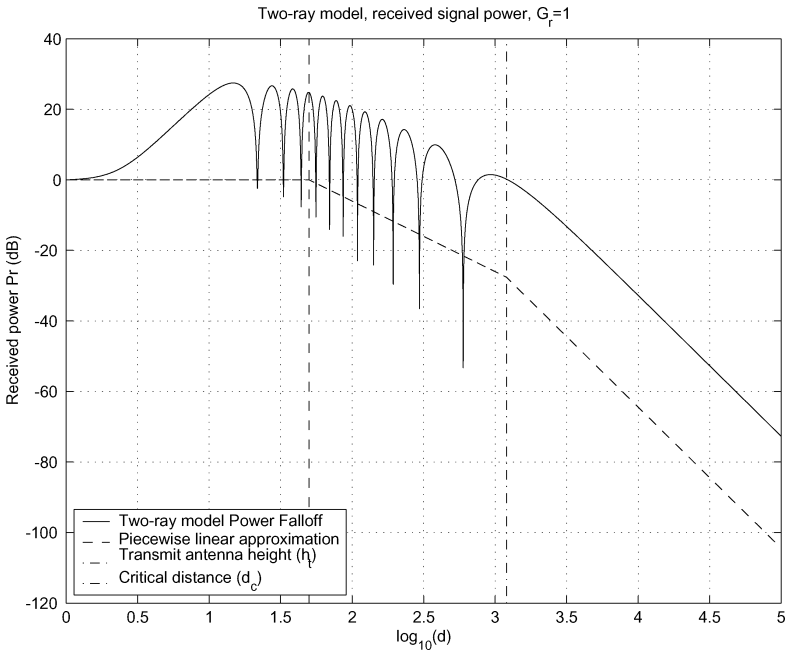

Figure 2.4: Received power versus distance for two-ray model. The received power is normalized to 0 dB at a reference distance $d_r = 1$ m (so $10\log(d/d_r) = 0$ dB for $d = 1$ m).

antenna height h_r, as is typical in an urban cellular system where macrocell base stations are located on top of tall buildings with mobile devices near street level. The plot can be separated into three segments. For small distances $(d < h_t)$ the two rays add constructively as d increases from zero. The local maximum over this segment occurs for d such that $\Delta\phi = \pi$ in (2.15). At that distance the phases of the two multipath components are perfectly aligned $(Re^{-j\Delta\phi} = 1)$. With this phase alignment the path loss over this segment is roughly proportional to $1/(d^2 + h_t^2)$ since both multipath components travel roughly the same distance $\sqrt{d^2 + h_t^2}$. That is because, for $h_r << h_t$,

$$d_1 = \sqrt{d^2 + (h_t + h_r)^2} \approx d_0 = \sqrt{d^2 + (h_t - h_r)^2} \approx \sqrt{d^2 + h_t^2}. \tag{2.21}$$

After this first local maximum and up to a certain critical distance d_c, the wave experiences constructive and destructive interference of the two rays as $\Delta\phi$ changes with d, resulting in a wave pattern with a sequence of maxima (when $|\Delta\phi|$ is an odd integer multiple of π) and minima (when $|\Delta\phi|$ is an even integer multiple of π). These maxima and minima are also referred to as *multipath fading*, discussed in more detail in the next chapter. At the critical distance d_c, the final maximum is reached, after which the signal power falls off proportionally with d^{-4}. This rapid falloff with distance is due to the fact that, for $d > d_c$, the signal components only combine destructively since $\Delta\phi \to 0$ as $d \to \infty$. Since the final maximum is achieved for d such that $\Delta\phi = \pi$, an approximation for d_c can be obtained by setting $\Delta\phi = \pi$ in (2.15), obtaining

$$d_c = 4h_t h_r/\lambda. \tag{2.22}$$

For the parameters used to generate Figure 2.4, $d_c = 1.2$ Km ($\log_{10} d_c = 3.08$), which is also shown in the figure.

The power falloff with distance in the two-ray model can be approximated by averaging out its local maxima and minima. This results in a piecewise linear model with three segments, which is also shown in Figure 2.4 slightly offset from the actual power falloff curve for illustration purposes. In the first segment, in order to get a

constant power falloff we assume $d_0 \approx d_1 \approx h_t^2$, leading to a power falloff proportional to $1/h_t^2$ that is independent of d; for distances between h_t and d_c, power falls off at -20 dB/decade; and at distances greater than d_c, power falls off at –40 dB/decade.

The critical distance d_c is proportional to the transmit antenna height h_t and to the carrier frequency $f_c = c/\lambda$. Thus, for small cells and indoor systems where h_t is ten meters or less, for f_c on the order of 1 GHz, d_c is hundreds of meters, as illustrated in Example 2.2. On the other hand, for millimeter wave systems where f_c is tens of GHz, d_c typically well exceeds a kilometer. The value of d_c can be used for system design. For example, if propagation in a cellular system obeys the two-ray model then the critical distance would be a natural size for the cell radius, since the path loss associated with interference outside the cell would be much larger than path loss for desired signals inside the cell. However, setting the cell radius to d_c could result in very large cells, as illustrated in Figure 2.4 and in the next example. Since smaller cells are more desirable – both to increase capacity and reduce transmit power – cell radii are typically much smaller than d_c. Thus, with a two-ray propagation model, power falloff within these relatively small cells goes as distance squared. Moreover, propagation in cellular systems rarely follows a two-ray model, since cancellation by reflected rays rarely occurs in all directions.

Example 2.2: Determine the critical distance for the two-ray model in an outdoor cell with $h_t = 10\ \mathrm{m},\ h_r = 3\ \mathrm{m}$ and in an indoor cell with $h_t = 3\ \mathrm{m},\ h_r = 2\ \mathrm{m}$ for $f_c = 2$ GHz.

Solution: $d_c = 4h_t h_r/\lambda = 800\ \mathrm{m}$ for the outdoor cell and 160 m for the indoor cell. A cell radius of 800 m in an outdoor cell is typical for a macrocell but a bit large for a small cell which today are on the order of 10-100 m to maintain large capacity. However, for a macrocell with an 800 m radius under these system parameters, signal power would fall off as d^2 inside the cell, while interference from neighboring cells would fall off as d^4 and thus would be greatly reduced. Similarly, 160 m is quite large for the cell radius of an indoor system, as there would typically be many walls the signal would have to penetrate for an indoor cell radius of that size. Hence an indoor system would typically have a smaller cell radius: on the order of 10–20 m, the size of one room or a few adjacent rooms in a typical building.

2.5 Path Loss Exponent Models

2.5.1 Single-Slope

The complexity of signal propagation makes it difficult to obtain a single model that characterizes path loss accurately across a range of different environments and frequencies. Accurate path loss models can be obtained from complex analytical models or empirical measurements when tight system specifications must be met or the best locations for base stations or access-point layouts must be determined. However, for general trade-off analysis of various system designs it is sometimes best to use a simple model that captures the essence of signal propagation without resorting to complicated path loss models, which are only approximations to the real channel anyway. The following single-slope model for path loss as a function of distance is a simple model that captures several important propagation characteristics.

$$P_r = P_t K \left[\frac{d_r}{d}\right]^{\gamma}, \tag{2.23}$$

Table 2.1: Typical path loss exponents

Environment	γ **range**
Urban macrocells	3.7–6.5
Urban microcells	2.7–3.5
Office building (same floor)	1.6–3.5
Office building (multiple floors)	2–6
Store	1.8–2.2
Factory	1.6–3.3
Home	3

where d_r is a reference distance for the antenna far field, γ is the path-loss exponent, and K is a unitless constant equal to the path gain P_r/P_t at distance $d = d_r$. The dB attenuation is thus

$$P_r\,(\text{dBm}) = P_t\,(\text{dBm}) + K\ \text{dB} - 10\gamma \log_{10}(d/d_r)\,, \tag{2.24}$$

and the path loss consists of two terms, where all environment and antenna parameters are captured by the first term, and the path loss due only to distance is captured by the second term:

$$P_L\,(\text{dB}) = 10\log_{10}(P_t/P_r) = -10\log_{10} K + 10\gamma \log_{10}(d/d_r). \tag{2.25}$$

The values for K, d_r, and γ can be obtained to approximate either an analytical or empirical model. In particular, the free-space path-loss model and some of the empirical models described in Section 2.7 are all of the same form as (2.24). Because of antenna near field effects at transmit distances on the order of a signal wavelength, the model (2.24) is generally valid only at transmission distances $d > d_r$, where d_r is typically assumed to be $1\ \text{m}$ for systems transmitting over distances greater than this nominal value.

When the single-slope model is used to approximate empirical measurements, the value of $K < 1$ is sometimes set to the free-space path gain at distance d_r assuming isotropic antennas:

$$K\ \text{dB} = 20\log_{10}\frac{\lambda}{4\pi d_r}, \tag{2.26}$$

and this assumption is supported by empirical data for free-space path loss at a transmission distance of 100 m [17]. Alternatively, K can be determined by measurement at d_r or optimized (alone or together with γ) to minimize the mean-square error (MSE) between the model and the empirical measurements [17]. The value of γ depends on the propagation environment: for propagation that approximately follows a free-space or two-ray model, γ is set to 2 or 4 respectively. The value of γ for more complex environments can be obtained via a minimum mean-square error (MMSE) fit to empirical measurements, as illustrated in Example 2.3. Alternatively, γ can be obtained from an empirically based model that takes into account frequency and antenna height [17]. Table 2.6 summarizes values of the path loss exponent γ for different environments based on data from [11, 17, 18, 19, 20, 21, 52]. Empirical measurements indicate that path loss exponents at higher frequencies tend to be higher [19, 23, 21, 22, 57, 59] whereas path loss exponents at higher antenna heights tend to be lower [17, 57, 59]. The more complex empirical models described below in Section 2.7, particularly the widely used 3GPP and WINNER II channel models [57, 59], have additional terms that explicitly capture the dependence of path loss on frequency and antenna height. Note that the wide range of empirical path-loss exponents for indoor propagation may be due to attenuation caused by floors, objects, and partitions (see Section 2.5.5).

Example 2.3: Consider the set of empirical measurements of P_r/P_t given in Table 2.2 for an indoor system at 900

MHz. Find the path loss exponent γ that minimizes the MSE between the single-slope model (2.24) and the empirical dB power measurements, assuming that $d_r = 1\,\mathrm{m}$ and K is determined from the free-space path-gain formula at this d_r. Find the received power at 150 m for the single-slope path loss model with this path loss exponent and a transmit power of 1 mW (0 dBm). Note that, since the path loss model in dB is linear, minimizing the MSE of the empirical data in dB (versus in linear units) relative to this model is a simple linear regression.

Solution: We first set up the MMSE error equation for the dB power measurements as

$$F(\gamma) = \sum_{i=1}^{5} [M_{\text{measured}}(d_i) - M_{\text{model}}(d_i)]^2,$$

where $M_{\text{measured}}(d_i)$ is the path loss measurement in Table 2.2 at distance d_i and where $M_{\text{model}}(d_i) = K - 10\gamma \log_{10}(d_i)$ is the path loss at d_i based on (2.24) since $d_r = 1$. Now using the free-space path gain formula yields $K = 20\log_{10}(.3333/(4\pi)) = -31.53$ dB. Thus

Table 2.2: path loss measurements

Distance from transmitter	$M = P_r/P_\mathrm{t}$
10 m	–70 dB
20 m	–75 dB
50 m	–90 dB
100 m	–110 dB
300 m	–125 dB

$$\begin{aligned} F(\gamma) &= (-70 + 31.53 + 10\gamma)^2 + (-75 + 31.53 + 13.01\gamma)^2 \\ &\quad + (-90 + 31.53 + 16.99\gamma)^2 + (-110 + 31.53 + 20\gamma)^2 \\ &\quad + (-125 + 31.53 + 24.77\gamma)^2 \\ &= 21682.50 - 11656.60\gamma + 1571.47\gamma^2. \end{aligned} \tag{2.27}$$

Differentiating $F(\gamma)$ relative to γ and setting it to zero yields

$$\frac{\partial F(\gamma)}{\partial \gamma} = -11656.60 + 3142.94\gamma = 0 \Rightarrow \gamma = 3.71.$$

For the received power at 150 m under the single-slope path loss model with $K = -31.53, \gamma = 3.71$, and $P_t = 0$ dBm, we have $P_r = P_t + K - 10\gamma \log_{10}(d/d_r) = 0 - 31.53 - 10 \cdot 3.71 \log_{10}(150) = -112.26$ dBm. Clearly the measurements deviate from the single-slope path loss model; this variation can be attributed to shadow fading, described in Section 2.7.

2.5.2 Multi-Slope

The multi-slope path loss model, which is piecewise linear in dB, generalizes the single-slope model by allowing for different path loss exponents at different distances. Given this greater flexibility, the multi-slope model is commonly used to analytically model empirical measurements. The multi-slope path loss model and a set of

measurement data on which it would be based is illustrated in Figure 2.5 for dB attenuation versus log distance, where the dots represent hypothetical measurements and the multi-slope model is the analytical approximation to these measurements. A multi-slope model with N segments must specify $N-1$ breakpoints $d_1, \ldots, d_{N-1}$ as well as the slopes corresponding to each segment $s_1, \ldots, s_N$. The slope of the ith segment is $s_i = -10\gamma_i$ for γ_i the path loss exponent on that segment. Different methods can be used to determine the number and location of breakpoints to be used in modeling the empirical data. Once these are fixed, the slopes corresponding to each segment can be obtained by linear regression on the data. The transition to a new slope might be instantaneous at a given breakpoint distance or entail some smoothing. Multi-slope models capture the analytical two-ray propagation model of Section 2.4. They also fit well to empirical measurements in propagation environments where the path loss exponent increases with distance [52], where the LOS path loss differs from that of non-LOS path loss [98], and in heterogeneous cellular networks where users may connect to a relatively distant macrocell base station while experiencing interference from a nearby small cell [99].

A special case of the multi-slope model is the dual-slope model. The dual-slope model is characterized by a constant path loss factor K and a path loss exponent γ_1 above some reference distance d_r and up to some breakpoint distance $d_{\rm BP}$, after which power falls off with path loss exponent γ_2. The dB path loss is thus given by:

$$P_r(d)\,(\text{dB}) = \begin{cases} P_t + K - 10\gamma_1 \log_{10}(d/d_r), & d_r \le d \le d_{\rm BP}, \\ P_t + K - 10\gamma_1 \log_{10}(d_{\rm BP}/d_r) - 10\gamma_2 \log_{10}(d/d_{\rm BP}), & d > d_{\rm BP}. \end{cases} \tag{2.28}$$

The path loss exponents, K, and $d_{\rm BP}$ are typically obtained via a regression fit to empirical data [17, 92]. The two-ray model described in Section 2.4 for $d > h_t$ can be approximated by the dual-slope model, with the breakpoint distance $d_{\rm BP} = d_c$ given by (2.22) and path loss exponents $\gamma_1 = 2$ and $\gamma_2 = 4$. Many of the measurement-based models described in Section 2.9 use the dual-slope model for path loss.

The transition between the multiple equations in the dual-slope model (2.28) can be smoothed by the following model [34, 93]:

$$P_r = \frac{P_t K}{L(d)}, \tag{2.29}$$

where

$$L(d) \triangleq \left(\frac{d}{d_r}\right)^{\gamma_1} \left(1 + \left(\frac{d}{d_{\rm BP}}\right)^{(\gamma_1 - \gamma_2)q}\right)^{1/q}. \tag{2.30}$$

In this expression, q is a parameter that determines the smoothness of the path loss at the transition region close to the breakpoint distance $d_{\rm BP}$. This smoothing in this model can be extended to more than two regions [91].

2.6 Shadowing

A signal transmitted through a wireless channel will typically experience random variation due to blockage from objects in the signal path, giving rise to random variations of the received power at a given distance. Such variations are also caused by changes in reflecting surfaces and scattering objects. Thus, a model for the random attenuation due to these effects is also needed. The location, size, and dielectric properties of the blocking objects – as well as the changes in reflecting surfaces and scattering objects that cause the random attenuation – are generally unknown, so statistical models must be used to characterize this attenuation.

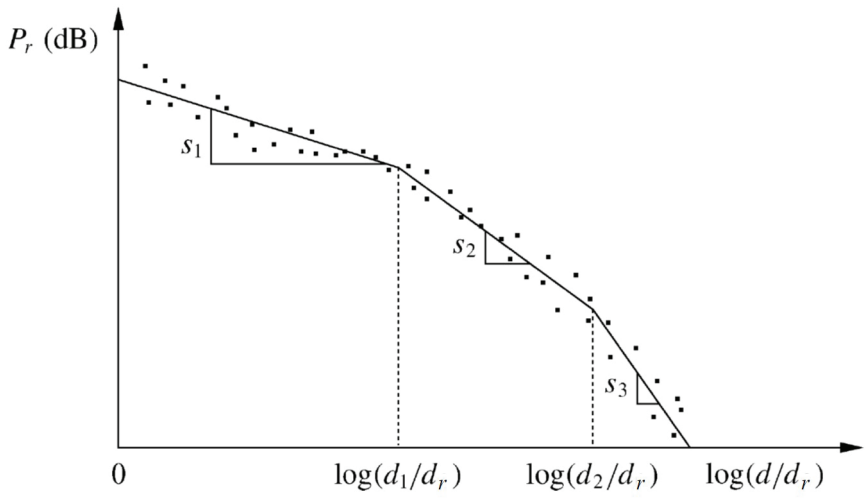

Figure 2.5: Multi-slope model for path loss.

The most common model for this additional attenuation is log-normal shadowing; under this model the dB value of the random attenuation is Gauss-distributed. The log-normal shadowing model has been empirically confirmed to model accurately the variation in received power in both outdoor and indoor radio propagation environments (see, e.g., [17, 102]). In the log-normal shadowing model, the ratio of transmit-to-receive power $\psi = P_t/P_r$ is assumed to be random with a log-normal distribution given by

$$p(\psi) = \frac{\xi}{\sqrt{2\pi}\sigma_{\psi_{\text{dB}}}\psi} \exp\left[-\frac{(10\log_{10}\psi - \mu_{\psi_{\text{dB}}})^2}{2\sigma^2_{\psi_{\text{dB}}}}\right], \ \psi > 0, \tag{2.31}$$

where $\xi = 10/\ln 10$, $\mu_{\psi_{\text{dB}}}$ is the mean of $\psi_{\text{dB}} = 10\log_{10}\psi$ in decibels, and $\sigma_{\psi_{\text{dB}}}$ is the standard deviation of ψ_{dB} (also in dB). The mean can be based on an analytical model or empirical measurements. For empirical measurements $\mu_{\psi_{\text{dB}}}$ equals the empirical path loss, since average attenuation from shadowing is already incorporated into the measurements. For analytical models, $\mu_{\psi_{\text{dB}}}$ must incorporate both the path loss (e.g., from a free-space or ray-tracing model) as well as average attenuation from blockage. Alternatively, path loss can be treated separately from shadowing, as described in the next section.

A random variable with a log-normal distribution is called a *log-normal random variable*. Note that if ψ is log-normal then the received power and received SNR will also be log-normal, since these are just constant multiples of ψ. For received SNR the mean and standard deviation of this log-normal random variable are also in decibels. For log-normal received power the random variable has units of power, so its mean and standard deviation will be in dBm or dBW instead of dB. The mean of ψ (the linear average path gain) can be obtained from (2.31) as

$$\mu_\psi = \mathrm{E}[\psi] = \exp\left[\frac{\mu_{\psi_{\text{dB}}}}{\xi} + \frac{\sigma^2_{\psi_{\text{dB}}}}{2\xi^2}\right]. \tag{2.32}$$

The conversion from the linear mean (in dB) to the log mean (in dB) is derived from (2.32) as

$$10\log_{10}\mu_\psi = \mu_{\psi_{\text{dB}}} + \frac{\sigma^2_{\psi_{\text{dB}}}}{2\xi}. \tag{2.33}$$

Performance in log-normal shadowing is typically parameterized by the log mean $\mu_{\psi_{\text{dB}}}$, which is referred to as the *average dB path loss* and is given in units of dB. We can compute the distribution of the dB value of ψ, $\psi_{\text{dB}} = 10\log_{10}\psi$ from the distribution of ψ given in (2.31). This computation reveals that ψ_{dB} is Gauss-distributed with mean $\mu_{\psi_{\text{dB}}}$ and standard deviation $\sigma_{\psi\text{dB}}$:

$$p(\psi_{\text{dB}}) = \frac{1}{\sqrt{2\pi}\sigma_{\psi_{\text{dB}}}} \exp\left[-\frac{(\psi_{\text{dB}} - \mu_{\psi_{\text{dB}}})^2}{2\sigma^2_{\psi_{\text{dB}}}}\right]. \tag{2.34}$$

The log-normal distribution of ψ or, equivalently, the Gaussian distribution of ψ_{dB} is defined by two parameters: $\mu_{\psi_{\text{dB}}}$ and $\sigma_{\psi_{\text{dB}}}$. Given that ψ_{dB} is based on the ratio of transmit and receive powers, it can be expressed as a difference of these powers in either dBW or dBm, as follows:

$$\psi_{\text{dB}} = 10\log_{10}(P_t/P_r) = P_t \text{ dBW} - P_r \text{ dBW} = P_t \text{ dBm} - P_r \text{ dBm}. \tag{2.35}$$

Since ψ_{dB} is Gauss-distributed, by (2.35) and properties of Gaussian random variables we have that P_r dBW is also Gauss-distributed with mean P_t dBW $- \mu_{\psi_{\text{dB}}}$ and standard deviation $\sigma_{\psi\text{dB}}$, i.e. the same standard deviation as ψ_{dB}. Similarly, P_r dBm is Gauss-distributed with mean P_t dBm $- \mu_{\psi_{\text{dB}}}$ and standard deviation $\sigma_{\psi\text{dB}}$.

The log-normal distribution is an imperfect approximation to the effect of shadowing because it has a non-zero probability that the received power exceeds the transmit power, which violates the laws of physics. In particular, since the transmit power should always exceed the received power, $\psi = P_t/P_r$ should always be greater than unity. Thus shadowing models set $\mu_{\psi_{\text{dB}}}$ to be greater than zero. Note, however, that the log-normal distribution (2.31) takes values for $0 \leq \psi \leq \infty$. Hence, for $\psi < 1$ we have $P_r > P_t$, which is physically impossible. However, this probability will be very small when $\mu_{\psi_{\text{dB}}}$ is large and positive. Thus, the log-normal model captures the underlying physical model for shadowing most accurately when $\mu_{\psi_{\text{dB}}} \gg 0$.

If the mean and standard deviation for the shadowing model are based on empirical measurements, then the question arises as to whether they should be obtained by taking averages of the linear or rather the dB values of the empirical measurements. In other words, given empirical (linear) path loss measurements $\{p_i\}_{i=1}^N$, should the mean path loss be determined as $\mu_\psi = (1/N)\sum_{i=1}^N p_i$ or as $\mu_{\psi_{\text{dB}}} = (1/N)\sum_{i=1}^N 10\log_{10} p_i$? A similar question arises for computing the empirical variance. In practice it is more common to determine mean path loss and variance based on averaging the dB values of the empirical measurements for several reasons. First, as we shall see, the mathematical justification for the log-normal model is based on dB measurements. In addition, the literature shows that obtaining empirical averages based on dB path loss measurements leads to a better approximation for the physical channel being modeled [103]. Finally, as we saw in Section 2.5.4, power falloff with distance models are often obtained by a piecewise linear approximation to empirical measurements of dB power versus the log of distance [11].

Most empirical studies for outdoor channels support a standard deviation $\sigma_{\psi_{\text{dB}}}$ ranging from 4 dB to 13 dB [1, 34, 104, 105, 106]. Larger deviations are associated with environments that contain a high density of blocking objects such as buildings and foliage outdoors, or walls and furniture indoors. Moreover, since attenuation by objects is more severe and more variable at high frequencies, and there is more variance in the number of attenuating objects at large distances, $\sigma_{\psi_{\text{dB}}}$ generally increases with both frequency and distance [72].The mean power $\mu_{\psi_{\text{dB}}}$ depends on the path loss and building properties in the area under consideration. The mean power $\mu_{\psi_{\text{dB}}}$ varies with distance; this is due to path loss and to the fact that average attenuation from objects increases with distance owing to the potential for a larger number of attenuating objects.

The Gaussian model for the distribution of the mean received signal in dB can be justified by the following attenuation model when shadowing is dominated by the attenuation from blocking objects. The attenuation of a signal as it travels through an object of depth d is approximately equal to

$$s(d) = ce^{-\alpha d}, \tag{2.36}$$

where α is an attenuation constant that depends on the object's interior materials and associated dielectric properties while c is a fixed attenuation constant that doesn't depend on the object's depth (e.g. the penetration loss the signal experiences in entering and exiting the object). If we assume that α is approximately equal for all blocking objects and that the ith blocking object has a random depth d_i, then the attenuation of a signal as it propagates through this region is

$$s(d_t) = ce^{-\alpha \Sigma_i d_i} = ce^{-\alpha d_t}, \tag{2.37}$$

where $d_t = \sum_i d_i$ is the sum of the random object depths through which the signal travels. If there are many objects between the transmitter and receiver, then by the central limit theorem we can approximate d_t by a Gaussian random variable. Thus, $\ln s(d_t) = -\alpha d_t$ will have a Gaussian distribution with mean μ and standard deviation σ. The value of σ will depend on the environment.

Example 2.4: In Example 2.3 we found that the exponent for the single-slope path loss model that best fits the measurements in Table 2.2 was $\gamma = 3.71$. Assuming the single-slope path loss model with this exponent, $d_r = 1$m, and the same $K = -31.53$ dB, find $\sigma^2_{\psi_{\text{dB}}}$, the variance of log-normal shadowing about the mean path loss based on these empirical measurements.

Solution:

The sample variance relative to the simplified path loss model with $\gamma = 3.71$ is

$$\sigma^2_{\psi_{\text{dB}}} = \frac{1}{5}\sum_{i=1}^{5}[M_{\text{measured}}(d_i) - M_{\text{model}}(d_i)]^2,$$

where $M_{\text{measured}}(d_i)$ is the path loss measurement in Table 2.2 at distance d_i and $M_{\text{model}}(d_i) = K - 37.1\log_{10}(d_i)$since $d_r = 1$m. This yields

$$\begin{aligned}\sigma^2_{\psi_{\text{dB}}} &= \frac{1}{5}[(-70 + 31.53 + 37.1)^2 + (-75 + 31.53 + 48.27)^2 \\ &\quad + (-90 + 31.53 + 63.03)^2 + (-110 + 31.53 + 74.2)^2 \\ &\quad + (-125 + 31.53 + 91.90)^2] \\ &= 13.28.\end{aligned}$$

Thus, the standard deviation of shadow fading on this path is $\sigma_{\psi_{\text{dB}}} = 3.64$ dB. Note that the bracketed term in the displayed expression equals the MMSE formula (2.27) from Example 2.3 with $\gamma = 3.71$.

Extensive measurements have been taken to characterize the empirical autocorrelation function of the shadow fading process over distance for different environments at different frequencies (see e.g. [105, 107, 108, 109, 110]). The most common analytical model for this function, first proposed by Gudmundson [105] and based on empirical measurements, assumes that the shadowing $\psi(d)$ is a first-order autoregressive process where the covariance between shadow fading at two points separated by distance δ is characterized by

$$A(\delta) = \mathbf{E}[(\psi_{\text{dB}}(d) - \mu_{\psi_{\text{dB}}})(\psi_{\text{dB}}(d+\delta) - \mu_{\psi_{\text{dB}}})] = \sigma^2_{\psi_{\text{dB}}}\rho_D^{\delta/D}, \tag{2.38}$$

where ρ_D is the normalized covariance between two points separated by a fixed distance D. This covariance must be obtained empirically, and it varies with the propagation environment and carrier frequency. Measurements indicate that for suburban macrocells with $f_c = 900$ MHz, $\rho_D = 0.82$ for $D = 100$ m and for urban microcells

with $f_c = 2$ GHz, $\rho_D = 0.3$ for $D = 10$ m [105, 109]. This model can be simplified and its empirical dependence removed by setting $\rho_D = 1/e$ for distance $D = X_c$, which yields

$$A(\delta) = \sigma^2_{\psi_{\text{dB}}} e^{-\delta/X_c}. \tag{2.39}$$

The *decorrelation distance* X_c in this model is the distance at which the signal autocovariance equals $1/e$ of its maximum value and is on the order of the size of the blocking objects or clusters of these objects. For outdoor systems, X_c typically ranges from 50 m to 100 m [108, 109]. For users moving at velocity v, the shadowing decorrelation in time τ is obtained by substituting $v\tau = \delta$ in (2.38) or (2.39). For a linear array of antennas, discussed in Chapter 10, shadowing can vary with the angle between the array and the signal's LOS path. Autocorrelation relative to this angle has been investigated in [107, 109].

The first-order autoregressive correlation model (2.38) and its simplified form (2.39) are easy to analyze and to simulate. Specifically, one can simulate ψ_{dB} by first generating a white Gaussian noise process with power $\sigma^2_{\psi_{\text{dB}}}$ and then passing it through a first-order filter with response $\rho_D^{\delta/D}$ for a covariance characterized by (2.38) or response $e^{-\delta/X_c}$ for a covariance characterized by (2.39). The filter output will produce a shadowing random process with the desired correlation properties [105, 106].

2.7 Combined Path Loss and Shadowing

Models for path loss and shadowing can be combined to capture power falloff versus distance along with the random attenuation about this path loss from shadowing. In this combined model, average dB path loss ($\mu_{\psi_{\text{dB}}}$) is characterized by the path loss model while shadow fading, with a mean of 0 dB, creates variations about this path loss, as illustrated by the path loss and shadowing curve in Figure 2.1. Specifically, this curve plots the combination of the single-slope path loss model (2.23) and the log-normal shadowing random process defined by (2.34) and (2.39).

2.7.1 Single-Slope Path Loss with Log-Normal Shadowing

For the combined model of single-slope path loss (2.24) and log-normal shadowing (2.33), the ratio of received to transmitted power in dB is given by

$$\frac{P_r}{P_t}\,\text{dB} = 10\log_{10} K - 10\gamma \log_{10}(d/d_r) - \psi_{\text{dB}}, \tag{2.40}$$

where ψ_{dB} is a Gauss-distributed random variable with variance $\sigma^2_{\psi_{\text{dB}}}$. The mean of ψ_{dB} is assumed to be zero when the term $10\log_{10} K$ captures average shadowing. When this is not the case, for example when K is calculated from the free-space path gain formula at $d = d_r$, then the mean of ψ_{dB} is positive and equal to the average shadowing loss over all distances. In (2.40), and as shown in Figure 2.1, the path loss increases linearly relative to $\log_{10} d/d_r$ with a slope of -10γ dB/decade, where γ is the path loss exponent. The variations due to shadowing change more rapidly, on the order of the decorrelation distance X_c.

Examples 2.3 and 2.4 illustrated the combined model for single-slope path loss and log-normal shadowing based on the measurements in Table 2.2, where path loss obeys the single-slope path loss model with $K = -31.53\,\text{dB}$ and path loss exponent $\gamma = 3.71$ and where shadowing obeys the log-normal model with mean given by the path loss model and standard deviation $\sigma_{\psi_{\text{dB}}} = 3.65\,\text{dB}$.

2.7.2 Outage Probability

The combined effects of path loss and shadowing have important implications for wireless system design. In wireless systems there is typically a target minimum received power level P_{min} or, equivalently, a minimum SNR $\text{SNR}_{\text{min}} = P_{\text{min}}/(N_0 B)$ below which performance becomes unacceptable (e.g., the caller's voice on a cell phone is unintelligible). However, with shadowing the received power at any given distance from the transmitter is log-normally distributed with some probability of falling below P_{min}. We define *outage probability* $P_{\text{out}}(P_{\text{min}}, d)$ to be the probability that the received power at a given distance d, $P_r(d)$, falls below P_{min}: $P_{\text{out}}(P_{\text{min}}, d) = p(P_r(d) < P_{\text{min}})$. For the combined path loss and shadowing model of Section 2.8 this becomes

$$p(P_r(d) \leq P_{\text{min}}) = 1 - Q\left(\frac{P_{\text{min}} - (P_t + 10\log_{10} K - 10\gamma \log_{10}(d/d_r))}{\sigma_{\psi_{\text{dB}}}}\right), \tag{2.41}$$

where the Q-function is defined as the probability that a Gaussian random variable X with mean 0 and variance 1 is greater than z:

$$Q(z) \triangleq p(X > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \tag{2.42}$$

The conversion between the Q-function and complementary error function is

$$Q(z) = \frac{1}{2}\,\text{erfc}\left(\frac{z}{\sqrt{2}}\right). \tag{2.43}$$

We will omit the arguments P_{min} and d of P_{out} when the context is clear or in generic references to outage probability.

Example 2.5: Find the outage probability at 150 m for a channel based on the single-slope path loss and shadowing models of Examples 2.3 and 2.4, assuming a transmit power of $P_t = 10$ mW and minimum power requirement of $P_{\text{min}} = -110.5$ dBm.
Solution: We have $P_t = 10\ \text{mW} = 10\ \text{dBm}$. Hence,

$$\begin{aligned}
&P_{\text{out}}(-110.5\ \text{dBm},\ 150\,\text{m}) \\
&\quad = p(P_r(150\,\text{m}) < -110.5\,\text{dBm}) \\
&\quad = 1 - Q\left(\frac{P_{\text{min}} - (P_t + 10\log_{10} K - 10\gamma \log_{10}(d/d_r))}{\sigma_{\psi_{\text{dB}}}}\right) \\
&\quad = 1 - Q\left(\frac{-110.5 - (10 - 31.53 - 37.1\log_{10}(150))}{3.65}\right) \\
&\quad = 0.0120.
\end{aligned}$$

An outage probability of 1% is a typical target in wireless system designs.

2.7.3 Cell Coverage Area and Percentage

The *coverage area* of a given cell in a cellular system is defined as the area of locations within the cell where the received power is above a given minimum. Consider a base station inside a circular cell of a given radius R. All

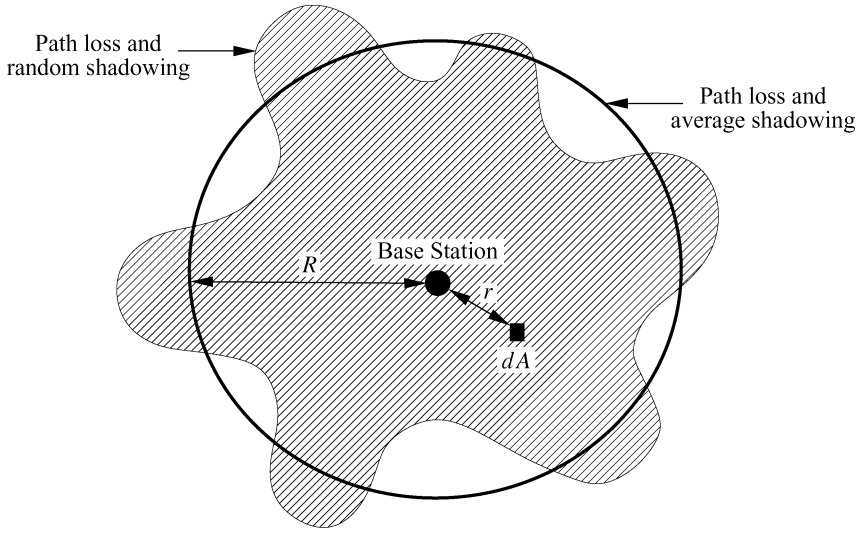

Figure 2.6: Contours of constant received power.

mobiles within the cell require some minimum received SNR for acceptable performance. Assuming a given model for noise, the SNR requirement translates to a minimum received power $P_{\min}$ throughout the cell. The transmit power at the base station is designed for an *average* received power at the cell boundary of $\overline{P}_R$, averaged over the shadowing variations. In the absence of shadowing, the coverage area of this system is πR^2 since all locations have received power above the required minimum. However, shadowing will cause some locations within the cell to have received power below $\overline{P}_R$, and others will have received power exceeding $\overline{P}_R$. This is illustrated in Figure 2.6, where we show contours of constant received power based on a fixed transmit power at the base station for path loss and average shadowing and for path loss and random shadowing. For path loss and average shadowing, constant power contours form a circle around the base station because combined path loss and average shadowing is the same at a uniform distance from the base station. For path loss and random shadowing, the contours form an amoeba-like shape due to the random shadowing variations about the average.

The constant power contours for combined path loss and random shadowing indicate the challenge that shadowing poses in cellular system design. Specifically, it is not possible for all users at the cell boundary to receive the same power level. Thus, either the base station must transmit extra power to ensure users affected by shadowing receive their minimum required power $P_{\min}$, which causes excessive interference to neighboring cells, or some users within the cell will find their minimum received power requirement unmet. In fact, since the Gaussian distribution has infinite tails, under this model *any* mobile in the cell has a nonzero probability of experiencing received power below its required minimum, even if the mobile is close to the base station. The model matches propagation scenarios in practice since a mobile may be in a tunnel or blocked by a large building, regardless of its proximity to the base station.

The *cell coverage percentage* is defined as the expected percentage of locations within a cell where received power exceeds $P_{\min}$. The cell coverage percentage under path loss and shadowing, also referred to as the fraction of useful service area, was derived by Reudink in [1, Chapter 2.5.3] as follows. The percentage of area within a cell where the received power exceeds the minimum required power $P_{\min}$ is obtained by taking an incremental area dA at radius r from the base station in the cell, as shown in Figure 2.6. Let $P_r(r)$ be the received power in dA from combined path loss and shadowing. Then the total area within the cell where the minimum power requirement is exceeded is obtained by integrating over all incremental areas where this minimum is exceeded:

$$C = \mathbf{E}\left[\frac{1}{\pi R^2}\int_{\text{cell area}} 1[P_r(r) > P_{\min} \text{ in } dA]\, dA\right]$$
$$= \frac{1}{\pi R^2}\int_{\text{cell area}} \mathbf{E}[1[P_r(r) > P_{\min} \text{ in } dA]]\, dA, \tag{2.44}$$

where $1[\cdot]$ denotes the indicator function. Define $P_A(r) = p(P_r(r) > P_{\min})$ in dA. Then $P_A(r) = \mathbf{E}[1[P_r(r) > P_{\min} \text{ in } dA]]$. Making this substitution in (2.44) and using polar coordinates for the integration yields

$$C = \frac{1}{\pi R^2}\int_{\text{cell area}} P_A(r)\, dA = \frac{1}{\pi R^2}\int_0^{2\pi}\int_0^R P_A(r) r\, dr\, d\theta. \tag{2.45}$$

The *outage probability within the cell* is defined as the percentage of area within the cell that does not meet its minimum power requirement $P_{\min}$; that is, $P_{\text{out}}^{\text{cell}} = 1 - C$. Given the log-normal distribution for the shadowing, we have

$$P_A = p(P_r(r) \geq P_{\min}) = Q\left(\frac{P_{\min} - (P_t + 10\log_{10} K - 10\gamma \log_{10}(r/d_0))}{\sigma_{\psi_{\text{dB}}}}\right)$$
$$= 1 - P_{\text{out}}(P_{\min}, r), \tag{2.46}$$

where P_{out} is the outage probability defined in (2.41) with $d = r$. Locations within the cell with received power below $P_{\min}$ are said to be *outage locations*.

Combining (2.45) and (2.46) yields[5]

$$C = \frac{2}{R^2}\int_0^R rQ\left(a + b\ln\frac{r}{R}\right) dr, \tag{2.47}$$

where

$$a = \frac{P_{\min} - \overline{P}_r(R)}{\sigma_{\psi_{\text{dB}}}}, \qquad b = \frac{10\gamma\log_{10}(e)}{\sigma_{\psi_{\text{dB}}}}, \tag{2.48}$$

and $\overline{P}_R = P_t + 10\log_{10} K - 10\gamma\log_{10}(R/d_0)$ is the received power at the cell boundary (distance R from the base station) due to path loss alone. Applying integration by parts to solve this integral yields a closed-form solution for C in terms of a and b:

$$C = Q(a) + \exp\left[\frac{2-2ab}{b^2}\right] Q\left(\frac{2-ab}{b}\right). \tag{2.49}$$

If the target minimum received power equals the average power at the cell boundary, $P_{\min} = \overline{P}_r(R)$, then $a = 0$ and the cell coverage percentage simplifies to

$$C = \frac{1}{2} + \exp\left[\frac{2}{b^2}\right] Q\left(\frac{2}{b}\right). \tag{2.50}$$

[5]Recall that (2.46) is generally valid only for $r \geq d_0$, yet to simplify the analysis we have applied the model for all r. This approximation will have little impact on cell coverage percentage, since d_0 is typically very small compared to R and the outage probability for $r < d_0$ is negligible.

Note that with this simplification C depends only on the ratio $\gamma/\sigma_{\psi_{\text{dB}}}$. Moreover, owing to the symmetry of the Gaussian distribution, under this assumption the outage probability at the cell boundary $P_{\text{out}}(\overline{P}_r(R), R) = 0.5$.

Example 2.6: Find the cell coverage percentage for a cell with the combined path loss and shadowing models of Examples 2.3 and 2.4, a cell radius of 600 m, a base station transmit power of $P_t = 100\,\text{mW} = 20\,\text{dBm}$, and a minimum received power requirement of $P_{\text{min}} = -110\,\text{dBm}$ and also one of $P_{\text{min}} = -120\,\text{dBm}$.

Solution: We first consider $P_{\text{min}} = -110$ and check if $a = 0$ to see whether we should use the full formula (2.49) or the simplified formula (2.50). We have $\overline{P}_r(R) = P_t + K - 10\gamma \log_{10}(600) = 20 - 31.53 - 37.1 \log_{10}(600) = -114.6\,\text{dBm} \neq -110\,\text{dBm}$, so we use (2.49). Evaluating a and b from (2.48) yields $a = (-110 + 114.6)/3.65 = 1.26$ and $b = (37.1 \cdot 0.434)/3.65 = 4.41$. Substituting these into (2.49) yields

$$C = Q(1.26) + \exp\left[\frac{2 - 2(1.26 \cdot 4.41)}{4.41^2}\right] Q\left(\frac{2 - (1.26)(4.41)}{4.41}\right) = .59,$$

which would be a very low cell coverage percentage value for an operational cellular system (lots of unhappy customers). Now considering the less stringent received power requirement $P_{\text{min}} = -120\,\text{dBm}$ yields $a = (-120 + 114.6)/3.65 = -1.479$ and the same $b = 4.41$. Substituting these values into (2.49) yields $C = .988$, a much more acceptable value for cell coverage percentage.

Example 2.7: Consider a cellular system designed so that $P_{\text{min}} = \overline{P}_r(R)$. That is, the received power due to path loss and average shadowing at the cell boundary equals the minimum received power required for acceptable performance. Find the cell coverage percentage for path loss values $\gamma = 2, 4, 6$ and $\sigma_{\psi_{\text{dB}}} = 4, 8, 12$, and explain how cell coverage percentage changes as γ and $\sigma_{\psi_{\text{dB}}}$ increase.

Solution: For $P_{\text{min}} = \overline{P}_r(R)$ we have $a = 0$, so cell coverage percentage is given by the formula (2.50). The cell coverage percentage thus depends only on the value for $b = 10\gamma \log_{10}(e)/\sigma_{\psi_{\text{dB}}}$, which in turn depends only on the ratio $\gamma/\sigma_{\psi_{\text{dB}}}$. Table 2.3 contains cell coverage percentage evaluated from (2.50) for the different γ and $\sigma_{\psi_{\text{dB}}}$ values.

Table 2.3: cell coverage percentage for different γ and $\sigma_{\psi_{\text{dB}}}$

γ	$\sigma_{\psi_{\text{dB}}}$=4 dB	$\sigma_{\psi_{\text{dB}}}$=8 dB	$\sigma_{\psi_{\text{dB}}}$=12 dB
2	.77	.67	.63
4	.85	.77	.71
6	.90	.83	.77

Not surprisingly, for fixed γ the cell coverage percentage increases as $\sigma_{\psi_{\text{dB}}}$ decreases; this is because a smaller $\sigma_{\psi_{\text{dB}}}$ means less variation about the mean path loss. Without shadowing we have 100% cell coverage percentage (since $P_{\text{min}} = \overline{P}_r(R)$) and so we expect that, as $\sigma_{\psi_{\text{dB}}}$ decreases to zero, cell coverage percentage increases to 100%. It is a bit more puzzling that for a fixed $\sigma_{\psi_{\text{dB}}}$ the cell coverage percentage increases as γ increases, since a larger γ implies that received signal power falls off more quickly. But recall that we have set $P_{\text{min}} = \overline{P}_r(R)$, so the faster power falloff is already taken into account (i.e., we need to transmit at much higher power with $\gamma = 6$ than with $\gamma = 2$ for this equality to hold). The reason cell coverage percentage increases with path loss exponent under this assumption is that, as γ increases, the transmit power must increase to satisfy $P_{\text{min}} = \overline{P}_r(R)$. This results in higher average power throughout the cell, yielding a higher percentage of locations in the cell that have the desired minimum power.

2.8 General Ray Tracing

In a typical urban or indoor environment, a radio signal transmitted from a fixed source will encounter multiple objects in the environment, as shown in Figure 2.7. These objects produce reflected, diffracted, or scattered copies of the transmitted signal, which are not captured by the free-space and two-ray models discussed earlier. These additional copies of the transmitted signal, known as multipath components, can be attenuated in power, delayed in time, and shifted in phase and/or frequency with respect to the LOS signal path at the receiver. The transmitted signal and its multipath components are summed together at the receiver, which can produce rapid fluctuations in received signal power due to constructive and destructive combining of the components. This multipath fading was also exhibited in the two-ray model. If the arriving signal components have delay differences that exceed a symbol time, this leads to distortion in the received signal relative to the transmitted signal.

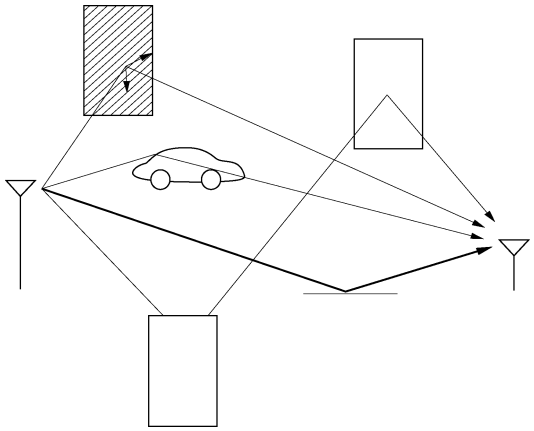

Figure 2.7: Reflected, diffracted, and scattered wave components.

In ray tracing we assume a finite number of reflectors with known location and dielectric properties. The details of the multipath propagation can then be solved using Maxwell's equations with appropriate boundary conditions. However, the computational complexity of this solution makes it impractical as a general modeling tool. Ray-tracing techniques approximate the propagation of electromagnetic waves by representing the wavefronts as simple particles. Thus, the effects of reflection, diffraction, and scattering on the wavefront are approximated using simple geometric equations instead of Maxwell's more complex wave equations. The error of the ray-tracing approximation is smallest when the receiver is many wavelengths from the nearest scatterer and when all the scatterers are large relative to a wavelength and fairly smooth. Comparison of the ray-tracing method with certain empirical data sets shows that it can accurately model received signal power in rural areas [24], along city streets when both the transmitter and receiver are close to the ground [24, 26, 27], and in indoor environments with appropriately adjusted diffraction coefficients [28].

If the transmitter, receiver, and reflectors are all immobile, then the characteristics of the multiple received signal paths are fixed. However, if the transmitter or receiver are moving then the characteristics of the multiple paths vary with time. These time variations are deterministic when the number, location, and characteristics of the reflectors are known over time. Otherwise, statistical models must be used. Similarly, if the number of reflectors is large, or if the reflector surfaces are not smooth such that each reflector generates many signal paths, then we must use statistical approximations to characterize the received signal. We will discuss statistical fading models

for propagation effects in Chapter 3. Hybrid models, which combine ray tracing and statistical fading, can also be found in the literature [29, 30]; however, we will not describe them here.

The most general ray tracing model includes all attenuated, diffracted, and scattered multipath components. This model uses all of the geometrical and dielectric properties of the objects surrounding the transmitter and receiver [37, 38, 39, 31]. Since this information is site-specific, general ray tracing methods are not used to obtain general theories about system performance and layout but rather for modeling propagation for a given transmit-receive configuration in a given environment. General ray tracing uses geometrical optics to trace the propagation of the LOS and reflected signal components as well as signal components from object diffraction and diffuse scattering. There is no limit to the number of multipath components at a given receiver location: the strength of each component is derived explicitly based on the locations and dielectric properties of the objects that generate them. In general, the LOS and reflected paths provide the dominant components of the received signal, since diffraction and scattering losses are high. However, in regions close to scattering or diffracting surfaces these other multipath components may dominate, especially if the LOS and reflected rays are blocked. Open source and commercial software for ray tracing has been developed for propagation modeling in both indoor and outdoor environments. In some of these programs, computer graphics are combined with aerial photographs (outdoor channels) or architectural drawings (indoor channels) to obtain a three-dimensional geometric picture of the environment.

The following section describes a ray tracing model capturing reflected rays only, including the commonly-used 6-ray and 10-ray models for a signal propagating along a straight street or hallway. We then present ray tracing models for signals that are reflected as well as diffracted and scattered. We also define the important parameter of local mean received power associated with these ray tracing models.

2.8.1 Multi-Ray Reflections

Most wireless channels have more than just the single reflected ray captured in the two-ray model described in Section 2.4. The multi-ray reflection model, developed in [25, 26], is a generalization of the two-ray model based on a propagation environment called a dielectric "canyon." This model captures all rays that experience one or more reflections on the path between the transmitter and receiver up to a certain maximum number of reflections.

In an outdoor setting, a dielectric canyon approximates propagation in a city with rectilinear streets[6] with buildings along both sides of the street as well as transmitter and receiver antenna heights that are close to street level. Theoretically, an infinite number of rays can be reflected off the building fronts to arrive at the receiver; in addition, rays may also be back-reflected from buildings behind the transmitter or receiver. However, since some of the signal energy is dissipated with each reflection, signal paths corresponding to more than two or three reflections can generally be ignored. When the street layout is relatively straight, back reflections are usually negligible also. Experimental data show that the dielectric canyon model of ten rays closely approximates signal propagation through cities with a rectilinear street layout [26]. The ten-ray model incorporates the LoS path as well as all nine paths with one, two, or three reflections: specifically, there is the ground-reflected, single-wall reflected, double-wall reflected, triple-wall reflected, wall-ground reflected, and ground-wall reflected paths. There are two of each type of wall-reflected path, one for each side of the street. An overhead view of the ten-ray model is shown in Figure 2.8. A simpler model, the six-ray model, includes a subset of these ten rays: the LOS, ground reflection, two single-wall reflections, and two double-wall reflections.

For the ten-ray model, the received signal is given by

$$r_{\text{10-ray}}(t) = \text{Re}\left\{\frac{\lambda}{4\pi}\left[\sum_{i=0}^{9}\frac{R_i\sqrt{G_i}u(t-\tau_i)e^{-j2\pi d_i/\lambda}}{d_i}\right]e^{j2\pi f_c t}\right\}, \tag{2.51}$$

[6] A rectilinear city is flat and has linear streets that intersect at 90° angles, as in midtown Manhattan.

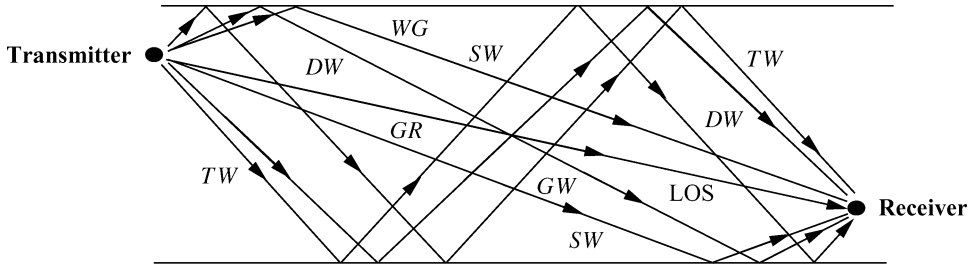

Figure 2.8: Overhead view of the ten-ray model. Includes the line-of-sight (LOS), ground-reflected (GR), two single-wall (SW) reflected, two double-wall (DW) reflected, two triple-wall (TW) reflected, wall-ground (WG) reflected, and ground-wall (GW) reflected rays

where d_i denotes the total distance traveled by the ith ray, $\tau_i = d_i/c$, and G_i is the product of the transmit and receive antenna power gains corresponding to the ith ray. The first two terms in the summation of (2.51) correspond to the LoS and ground reflected path in the two-ray model (2.12), hence $R_0 = 1$ since there is no reflection of the LoS path. For the ith reflection path, $i > 0$, the coefficient R_i is either a single reflection coefficient given by (2.16) or, if the path corresponds to multiple reflections, the product of the reflection coefficients corresponding to each reflection. The dielectric constants used in (2.51) are approximately the same as the ground dielectric, so $\varepsilon_r = 15$ is used for all the calculations of R_i. The delay spread of the 10-ray and 6-ray models equals the difference between the delay of the LOS ray and that of the reflected ray with the largest delay: $\max_i \tau_i - \tau_0$.

If we again assume a narrowband model such that $u(t - \tau_0) \approx u(t - \tau_i)$ for all i, then the received power corresponding to (2.51) is

$$P_r = P_t \left[\frac{\lambda}{4\pi}\right]^2 \left|\sum_{i=0}^{9} \frac{R_i \sqrt{G_i} e^{-j\Delta\phi_i}}{d_i}\right|^2, \tag{2.52}$$

where $\Delta\phi_i = 2\pi(d_i - d_0)/\lambda$.

Power falloff with distance in both the ten-ray model (2.52) and urban empirical data [15, 32, 33] for transmit antennas both above and below the building skyline is typically proportional to d^{-2}, even at relatively large distances. Moreover, the falloff exponent is relatively insensitive to the transmitter height. This falloff with distance squared is due to the dominance of the multipath rays, which decay as d^{-2}, over the combination of the LOS and ground-reflected rays (two-ray model), which decays as d^{-4}. Other empirical studies [34, 35, 36] have obtained power falloff with distance proportional to $d^{-\gamma}$, where γ lies anywhere between 2 and 6.

2.8.2 Diffraction

The propagation model for the LOS and reflected paths was outlined in the previous section. Diffraction allows a radio signal to "bend around" an object in its path to the receiver, as shown in Figure 2.9 for a wedge-shaped object. This bending phenomenon can be explained using Huygen's principle, which states that all points on the signal wavefront can be considered as point sources for a secondary wavefront called a *wavelet*. When a signal is diffracted, these wavelets combine together to produce a wavefront in the new direction of propagation caused by the bending [29, Chapter 4.3]. Diffraction results from many phenomena, including hilly or irregular terrain, building and rooftop edges, or obstructions blocking the LOS path between the transmitter and receiver [2, 11, 16]. Diffraction can be accurately characterized using the geometrical theory of diffraction (GTD) [40], but the complexity of this approach has precluded its use in wireless channel modeling. Wedge diffraction simplifies the GTD by assuming the diffracting object is a wedge rather than a more general shape. This model has been

used to characterize the mechanism by which signals are diffracted around street corners, which can result in signal attenuation exceeding 100 dB for some incident angles on the wedge [28, 39, 41, 42]. Although wedge diffraction simplifies the GTD, it still requires a numerical solution for the resulting attenuation [40, 43] and thus is not commonly used. Diffraction is most commonly modeled by the *Fresnel knife-edge diffraction model* because of its simplicity. The geometry of this model is shown in Figure 2.9, where the diffracting object is assumed to be asymptotically thin, which is not generally the case for hills, rough terrain, or wedge diffractors. In particular, this model does not consider diffractor parameters such as polarization, conductivity, and surface roughness, which can lead to inaccuracies [41].

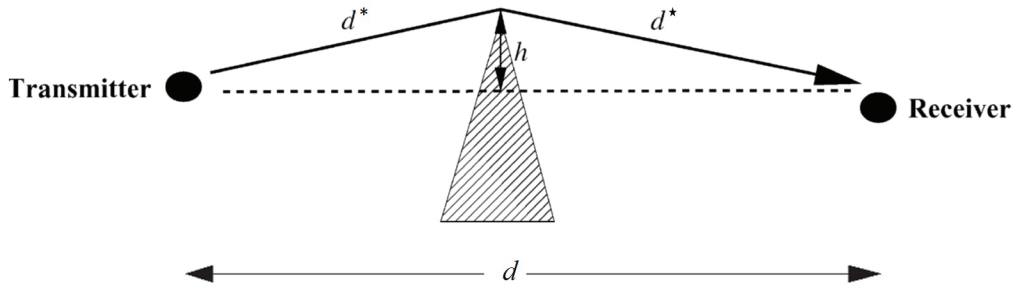

Figure 2.9: Knife-edge diffraction.

The diffracted signal of Figure 2.9 travels a distance $d^* + d^\star$, resulting in a phase shift of $\phi = 2\pi(d^* + d^\star)/\lambda$. For a LOS path of distance d, the geometry of Figure 2.9 indicates that, for h small relative to d^* and $d^\star$, the signal must travel an additional distance relative to the LOS path of approximately

$$\Delta d = d^* + d^\star - d \approx \frac{h^2(d^* + d^\star)}{2d^* d^\star};$$

the corresponding phase shift relative to the LOS path is approximately

$$\Delta\phi = \frac{2\pi\Delta d}{\lambda} \approx \frac{\pi}{2}v^2, \tag{2.53}$$

where

$$v = h\sqrt{\frac{2(d^* + d^\star)}{\lambda d^* d^\star}} \tag{2.54}$$

is called the *Fresnel–Kirchhoff diffraction parameter*. The path loss of a signal at distance $d^* + d^\star$ experiencing knife-edge diffraction at distance d^* is generally a function of v. However, computing this path loss is fairly complex, requiring the use of Huygens's principle, Fresnel zones, and the complex Fresnel integral [2]. Moreover, the resulting diffraction loss cannot generally be found in closed form. Approximations for the additional path loss due to knife-edge diffraction relative to free-space path loss (in dB) as a function of v in (2.54) are given in [16, 44] as

$$L(v)\text{dB} = \begin{cases} 0, & v < -1, \\ 20\log_{10}[.5 - .62v], & -0.8 \le v < 0, \\ 20\log_{10}[.5e^{-.95v}], & 0 \le v < 1, \\ 20\log_{10}[.4 - \sqrt{.1184 - (.38 - .1v)^2}], & 1 \le v \le 2.4, \\ 20\log_{10}[.225/v], & v > 2.4. \end{cases} \tag{2.55}$$

The knife-edge diffraction received signal model is obtained by adding this loss to the free-space path loss of (2.11), yielding the following formula for the received diffracted signal:

$$r(t) = \mathrm{Re}\left\{ \frac{\lambda}{4\pi} \left[\frac{L(v)\sqrt{G_d}u(t-\tau)e^{-j2\pi(d^*+d^\star)/\lambda}}{d^*+d^\star} \right] e^{j2\pi f_c t} \right\}, \tag{2.56}$$

where $\sqrt{G_d}$ is the antenna gain and $\tau = (d^* + d^\star)/c$ is the delay associated with the defracted ray.

In addition to diffracted rays, there may also be rays that are diffracted multiple times, or rays that are both reflected and diffracted. Models exist for including all possible permutations of reflection and diffraction [45]; however, the attenuation of the corresponding signal components is generally so large that these components are negligible relative to the noise. Diffraction models can also be specialized to a particular environment or frequency band. For example, a model for diffraction from rooftops and buildings in cellular systems was developed by Walfisch and Bertoni in [46]. This rooftop diffraction model was extended to multi-antenna base stations in [47], where it produced a "keyhole" phenomenon that reduced the effective spatial dimensions of the channel. Two-dimensional diffraction models for urban small cells, where signals propagate below rooftop levels, is developed in [48]. Application of knife-edge and more general diffraction models to millimeter wave propagation is discussed in [49, 50].

2.8.3 Scattering

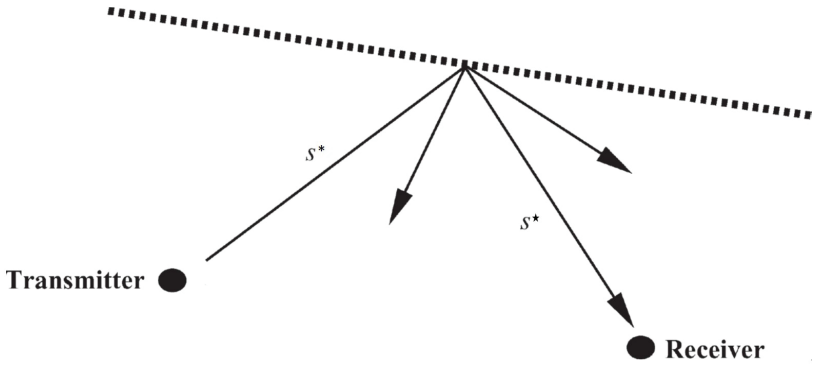

Figure 2.10: Scattering.

A scattered ray, shown in Figure 2.10 by the segments s_α and s_β, has a path loss proportional to the product of s_α and s_β. This multiplicative dependence is due to the additional spreading loss that the ray experiences after scattering. The received signal due to a scattered ray is given by the bistatic radar equation [51]:

$$r(t) = \mathrm{Re}\left\{ \frac{\lambda}{4\pi} \left[\frac{\sqrt{G_s}\sigma u(t-\tau)e^{-j2\pi(s^*+s^\star)/\lambda}}{\sqrt{4\pi}s^* s^\star} \right] e^{j2\pi f_c t} \right\}, \tag{2.57}$$

where $\tau = (s^* + s^\star)/c$ is the delay associated with the scattered ray; σ (in square meters) is the radar cross-section of the scattering object, which depends on the roughness, size, and shape of the scatterer; and G_s is the antenna power gain. The model assumes that the signal propagates from the transmitter to the scatterer based on free-space propagation and is then re-radiated by the scatterer with transmit power equal to σ times the received power at the scatterer. From (2.57), the path loss associated with scattering is

$$\begin{aligned} P_r \text{ dBm} = \quad & P_t \text{ dBm} + 10\log_{10}(G_s) + 20\log_{10}(\lambda) + 10\log_{10}(\sigma) \\ & -30\log(4\pi) - 20\log_{10}(s^*) - 20\log_{10}(s^\star). \end{aligned} \tag{2.58}$$

Empirical values of $10\log_{10}\sigma$ were determined in [52] for different buildings in several cities. Results from this study indicate that $10\log_{10}\sigma$ in dBm^2 ranges from -4.5 dBm^2 to 55.7 dBm^2, where dBm^2 denotes the dB value of the σ measurement with respect to one square meter.

2.8.4 Multipath Model with Reflection, Diffraction, and Scattering

The received signal is determined from the superposition of all the components due to the multiple rays. Thus, if we have a LOS ray, N_r reflected rays, N_d diffracted rays, and N_s scattered rays, the total received signal is

$$\begin{aligned} r_{\text{total}}(t) = \text{Re}\Bigg\{ \frac{\lambda}{4\pi} \Bigg[& \frac{\sqrt{G_0}u(t)e^{-j2\pi d_0/\lambda}}{d_0} + \sum_{i=1}^{N_r} \frac{R_i\sqrt{G_i}u(t-\tau_i)e^{-j2\pi d_i/\lambda}}{d_i} \\ & + \sum_{j=1}^{N_d} \frac{L_j(v)\sqrt{G_{d_j}}u(t-\tau_j)e^{-j2\pi(d_j^*+d_j^\star)/\lambda}}{d_j^* d_j^\star} \\ & + \sum_{k=1}^{N_s} \frac{\sqrt{G_{s_k}}\sigma_k u(t-\tau_k)e^{-j2\pi(s_k^*+s_k^\star)/\lambda}}{\sqrt{4\pi}s_k^* s_k^\star} \Bigg] e^{j2\pi f_c t} \Bigg\}, \end{aligned} \tag{2.59}$$

where τ_j and τ_k are, respectively, the time delays of the given diffracted and scattered rays, $(d_{\alpha_j}, d_{\beta_j})$ and $(s_{\alpha_k}, s_{\beta_k})$ are, respectively, the distance pair the signal travels before and after the object causing the diffraction or scattering, and the other parameters are as defined in the model (2.51) with the LoS and reflections only. The received power P_r of $r_{\text{total}}(t)$ and the corresponding path loss P_r/P_t are then obtained from (2.59). The delay spread of this model equals the difference between the delay of the LOS ray and that of the reflected, diffracted, or scattered ray that has the largest delay.

Any of these multipath components may have an additional attenuation factor if its propagation path is blocked by objects such as buildings outdoors or walls indoors. In this case, the attenuation factor of the obstructing object multiplies the component's path loss term in (2.59). This attenuation loss will vary widely, depending on the material and depth of the object [11, 53]. Models for random loss due to attenuation are described in Section 2.6.

2.8.5 Multi-Antenna and MIMO Systems

The ray tracing techniques described in Sections 2.8.1- 2.8.4 assume single antennas at the transmitter and receiver. The same ray tracing techniques can be used for systems with multiple antennas at the transmitter and/or receiver by superposition, whereby the single-antenna analysis is applied to each transmit-receive antenna pair and the results are summed together. When both the transmitter and receiver have multiple antennas, the system is called multiple-input multiple output (MIMO) system. While the application of ray tracing to MIMO systems is straightforward, the computational complexity of this approach grows linearly with the product of the number of transmit and receive antennas. As a result, ray tracing approximations to lower this complexity have been developed, including clustering of rays [54] and preprocessing environmental data [56]. A summary of ray tracing methods for MIMO systems can be found in [55].

2.8.6 Local Mean Received Power

The path loss computed from all ray-tracing models is associated with a fixed transmitter and receiver location. In addition, ray tracing can be used to compute the *local mean received power* $\overline{P}_r$ in the vicinity of a given receiver location by adding the squared magnitude of all the received rays. This has the effect of averaging out local spatial variations due to phase changes around the given location. Local mean received power is a good indicator of link quality and is often used in cellular system functions like power control and handoff [83].

2.9 Measurement-Based Propagation Models

Most mobile communication systems operate in complex propagation environments that cannot be accurately modeled by free-space path loss or ray tracing. A number of analytical path loss models have been developed over the years by researchers as well as standards bodies to model path loss in typical wireless environments including urban macrocells and small cells, suburban areas, rural areas, and inside buildings with different characteristics [11, 57, 59, 74, 17]. These models are generally based on large empirical measurement campaigns that can range over a variety of distances, frequency ranges, geographical regions for outdoor models, and building types for indoor models. These analytical models have the highest accuracy when they are applied to propagation conditions similar to those under which the empirical measurements that the models are based on were made. In this section we describe the most common analytical path loss models based on empirical measurements for both indoor and outdoor systems.

Analytical models characterize P_r/P_t as a function of distance, so path loss is well-defined. In contrast, empirical measurements of P_r/P_t as a function of distance include the effects of path loss, shadowing, multipath, and other site-specific factors that affect propagation. In order to remove multipath effects, empirical measurements for path loss typically average their received power measurements and the corresponding path loss at a given distance over several wavelengths. This average path loss is called the *local mean attenuation* (LMA) at distance d, and it generally decreases with d owing to free-space path loss and signal obstructions. The LMA in a given environment, such as a city, depends on the specific location of the transmitter and receiver corresponding to the LMA measurement. To characterize LMA more generally, measurements are typically taken throughout the environment and possibly in multiple environments with similar characteristics. Thus, the *empirical path loss* $P_L(d)$ for a given environment (a city, suburban area, or office building) is defined as the average of the LMA measurements at distance d averaged over all available measurements in the given environment. For example, empirical path loss for a generic downtown area with a rectangular street grid might be obtained by averaging LMA measurements in New York City, downtown San Francisco, and downtown Chicago. The empirical path loss models given in this section are all obtained from average LMA measurements. Empirical path loss models can also be developed or refined using measurement-driven learning applied to continuous LMA data collection [75].

2.9.1 Okumura Model

One of the most well-known models for signal prediction in large urban macrocells is the Okumura model [84]. This model is applicable over distances of 1–100 km and frequency ranges of 150–1500 MHz. Okumura used extensive measurements of base station-to-mobile signal attenuation throughout Tokyo to develop a set of curves giving median attenuation relative to free space of signal propagation in irregular terrain. The base station heights for these measurements were 30–100 m, a range whose upper end is higher than typical base stations today. The empirical path loss formula of Okumura at distance d parameterized by the carrier frequency f_c is given by

$$P_L(d)\ \text{dB} = L(f_c, d) + A_\mu(f_c, d) - G(h_t) - G(h_r) - G_{\text{AREA}}, \tag{2.60}$$

where $L(f_c, d)$ is free-space path loss at distance d and carrier frequency f_c, $A_\mu(f_c, d)$ is the median attenuation in addition to free-space path loss across all environments, $G(h_t)$ is the base station antenna height gain factor, $G(h_r)$ is the mobile antenna height gain factor, and G_{AREA} is the gain due to the type of environment. The values of $A_\mu(f_c, d)$ and G_{AREA} are obtained from Okumura's empirical plots [84, 11]. Okumura derived empirical formulas for $G(h_t)$ and $G(h_r)$ as follows:

$$G(h_t) = 20 \log_{10}(h_t/200), \quad 30 \text{ m} < h_t < 1000 \text{ m}; \tag{2.61}$$

$$G(h_r) = \begin{cases} 10 \log_{10}(h_r/3), & h_r \leq 3 \text{ m}, \\ 20 \log_{10}(h_r/3), & 3 \text{ m} < h_r < 10 \text{ m}. \end{cases} \tag{2.62}$$

Correction factors related to terrain are also developed in [84] that improve the model's accuracy. Okumura's model has a 10–14 dB empirical standard deviation between the path loss predicted by the model and the path loss associated with one of the measurements used to develop the model. The expected error in using Okumura's model for environments not based on these measurements is generally higher.

2.9.2 Hata Model

The Hata model [85] is an empirical formulation of the graphical path loss data provided by Okumura and is valid over roughly the same range of frequencies, 150–1500 MHz. This empirical model simplifies calculation of path loss because it is a closed-form formula and is not based on empirical curves for the different parameters. The standard formula for empirical path loss in urban areas under the Hata model is

$$\begin{aligned} P_{L,\text{urban}}(d) \text{ dB} = 69.55 + 26.16 \log_{10}(f_c) - 13.82 \log_{10}(h_t) - a(h_r) \\ + (44.9 - 6.55 \log_{10}(h_t)) \log_{10}(d). \end{aligned} \tag{2.63}$$

The parameters in this model are the same as under the Okumura model, and $a(h_r)$ is a correction factor for the mobile antenna height based on the size of the coverage area [85, 11]. For small to medium-sized cities, this factor is given by

$$a(h_r) = (1.1 \log_{10}(f_c) - .7)h_r - (1.56 \log_{10}(f_c) - .8) \text{ dB},$$

and for larger cities at frequencies $f_c > 300$ MHz by

$$a(h_r) = 3.2(\log_{10}(11.75h_r))^2 - 4.97 \text{ dB}.$$

Corrections to the urban model are made for suburban and rural propagation, so that these models are (respectively)

$$P_{L,\text{suburban}}(d) \text{ dB} = P_{L,\text{urban}}(d) \text{ dB} - 2[\log_{10}(f_c/28)]^2 - 5.4 \tag{2.64}$$

and

$$P_{L,\text{rural}}(d) \text{ dB} = P_{L,\text{urban}}(d) \text{ dB} - 4.78[\log_{10}(f_c)]^2 + 18.33 \log_{10}(f_c) - K, \tag{2.65}$$

where K ranges from 35.94 (countryside) to 40.94 (desert). Unlike the Okumura model, the Hata model does not provide for any path-specific correction factors. The Hata model well approximates the Okumura model for distances $d > 1$ km. Hence it is a good model for early-generation cellular systems operating in cities, but it

does not model propagation well in cellular systems with smaller cell sizes, multiple antennas, or higher frequencies. Indoor environments are also not captured by the Hata model. As cellular systems evolved beyond their early deployments, research into more complex propagation models was undertaken by multiple organizations, as described in the next section.

2.9.3 Cellular System Models

Following the success of early cellular systems, funding agencies and standards bodies worldwide launched projects to develop standardized cellular system channel models for the environments, carrier frequencies, cell sizes, and multiple antenna characteristics of these evolving systems. These standardized channel models can be used to evaluate different candidate cellular technologies in a uniform manner, and also provide guidelines for expected system performance in typical environments and operating conditions. Cellular system channel models span frequency ranges from .5 GHz to 100 GHz, indoor and outdoor environments with both LOS and non-LOS propagation, small and large cells sizes, and multiple antenna transmitters and receiver of different antenna heights, elevation angles, and polarizations. Models for all of these scenarios have been developed by 3GPP and the ITU [57, 58], whereas other standardized models are applicable to a subset of these settings. To illustrate some basic properties of these standardized models, in this section we provide the LOS path loss formulas for a few standardized models, which follow either the single-slope or dual-slope formulas ((2.24) and (2.28), respectively) with model-specific constants. Extensions of these models to more complex environments is also discussed. More details on the unique characteristics of millimeter wave propagation are provided in Section 2.9.5.

An early project to standardize channel models for cellular systems was undertaken by the European Cooperation in Science and Technology (COST). The COST 207 program, launched in 1984, brought together industry, government, and academia under a common umbrella to develop common propagation models for the emerging 2G cellular systems. The subsequent COST 231 project was focused on extending earlier channel models to carrier frequencies up to 2 GHz. The resulting model, referred to as the COST 231 extension to the Hata model [86], has the following path loss expression:

$$\begin{aligned} P_{L,\,\text{urban}}(d)\ \text{dB} = 46.3 + 33.9\log_{10}(f_c) - 13.82\log_{10}(h_t) - a(h_r) \\ + (44.9 - 6.55\log_{10}(h_t))\log_{10}(d) + C_M, \end{aligned} \tag{2.66}$$

where $a(h_r)$ is the same correction factor as in the Hata model and $C_M = 0$ dB for medium-sized cities and suburbs while $C_M = 3$ dB for metropolitan areas. The COST 231 extension to the Hata model is restricted to the following range of parameters: 1.5 GHz $< f_c <$ 2 GHz, $30\,\text{m} < h_t < 200\,\text{m}, 1\,\text{m} < h_r < 10\,\text{m}$, and 1 km $< d <$ 20 km. Given the Okumura measurements on which it is based, this model is most accurate for heterogeneous macrocell-only cellular system architectures with single-antenna base stations and terminals. The COST 231 model was further extended to incorporate new features and environments of the evolving cellular systems and standards, including diffraction and non-LOS propagation (COST-231-WI)[29], multiple base station antennas (COST 259) [88], and multiple antennas at both the base stations and the mobile terminals (COST 273, COST 2100) [89, 90].

The 3GPP standards body maintains a family of channel models characterizing signal propagation for the different environments and frequencies in which their cellular systems operate [57]. These include path loss models for both LOS and non-LOS propagation in outdoor macrocells and small cells as well as for indoor office buildings. The 3GPP outdoor propagation models follow the dual-slope path loss model (2.28); they have a distance-dependent pathloss exponent that takes on two different values, γ_1 and γ_2, for distances that are, respectively, smaller or greater than the breakpoint distance. The breakpoint distance is set to the critical distance $d_c = 4h_th_r/\lambda$ given by (2.22) in the two-ray model based on the effective transmit (h_t) and receive (h_r) antenna

heights[7]. In particular, the 3GPP path loss for LOS propagation in outdoor macrocells and small cells is given by

$$P_L(d)\,\mathrm{dB} = \begin{cases} 20\log_{10}(f_c) + \kappa + 10\gamma_1 \log_{10}(d), & d \le d_c, \\ 20\log_{10}(f_c) + \kappa + 10\gamma_2 \log_{10}(d) + \eta \log_{10}\left(d_c^2 + (h_t - h_r)^2\right), & d > d_c \end{cases} \tag{2.67}$$

for d in meters and f_c in GHz. The 3GPP indoor office model is governed by the first terms of (2.67) for all distances (i.e., there is no breakpoint distance in the indoor model due to the shorter distances involved). At distances below d_c this model has three separate terms: the constant κ that captures environment and antenna parameters independent of distance and frequency, the term $20\log_{10} f_c$ that captures frequency-dependent path loss, and the term $10\gamma_i \log_{10}(d)$ that captures distance-dependent path loss. The frequency-dependent path loss $20\log_{10} f_c$ is the same as in free space (2.10), and this term along with κ are captured by the constant K in the single-slope model (2.24). At distances above d_c the macrocell and small cell models have a fourth term that captures path loss dependence on antenna heights. The values of κ, γ_i, and η for the macrocell, small cell, and indoor office models are given in Table 2.4.

Table 2.4: Parameters for 3GPP LOS Path Loss Models

Parameter	Macrocell	Small Cell	Indoor Office
κ	28	32.4	32.4
γ	$\gamma_1 = 22, \gamma_2 = 40$	$\gamma_1 = 21, \gamma_2 = 40$	$\gamma_1 = 17.3$
η	-9	-9.5	NA

The 3GPP family of channel models also include non-LOS propagation models that take path loss to be the maximum of the LOS path loss model and a different non-LOS path loss model. The non-LOS path loss model is of the same form as (2.67) but with different parameters and a fourth term that depends only on the receiver height h_r. There is also a 3GPP rural macrocell model that includes additional parameters of the environment such as building heights and street widths. The 3GPP Spatial Channel Model (SCM) incorporates MIMO parameters. Details can be found in [57]. The ITU channel models are similar to the 3GPP channel family and cover the same frequency range of .5-100 GHz [58].

The European WINNER (Wireless World Initiative New Radio) I and II projects created a similar family of channel models as 3GPP for cellular system propagation under different conditions and environments [59]. The WINNER path loss model is similar to the single-slope model (2.24) with the constant K broken into 3 separate terms. Specifically, the path loss in the WINNER propagation model $10\log_{10}(P_t/P_r)$ is given by

$$P_L\mathrm{dBm} = 10\gamma \log_{10}(d) + \kappa + \beta \log_{10}(.2f_c) + X \tag{2.68}$$

for d in meters and f_c in GHz, where γ is the path loss exponent, κ is the path loss at the antenna near-field reference distance, typically a few signal wavelengths [12], which is independent of frequency, $C\log_{10}(.2f_c)$ is the frequency-dependence term, and X is an optional environment-specific term that may, for example, capture a fixed attenuation at all distances. The Winner II model includes MIMO channel parameters.

The COST, 3GPP, ITU, and WINNER families of channel models are widely used by researchers and practitioners for modeling path loss in cellular systems at frequencies below 6 GHz, with the 3GPP and ITU models covering frequencies up to 100 GHz. These models also include delay spread parameters ranging from tens of nanoseconds for indoor environments to hundreds of nanoseconds for outdoor environments. Alternative empirical channel models specifically for MIMO systems include the WIMAX 802.16 and Stanford University Interim

[7] In the 3GPP model the effective antennas heights are the actual heights reduced by 1m.

(SUI) channel models [76]. As cellular system architectures have evolved to include large and small cells, device-to-device communications, base stations mounted on UAVs and drones, large antenna arrays, millimeter wave frequencies, and very high speed mobile devices, there have been extensive measurement campaigns to characterize propagation under these new configurations and conditions [77, 78, 79]. In addition, learning algorithms that continually update channel models based on collected data have also been developed [80, 81, 82]. Clearly measurement-based models for cellular systems will continue to evolve as new system architectures, frequency bands, device requirements, and channel measurement tools are introduced.

2.9.4 Wi-Fi Channel Models

Prior to 802.11n, performance evaluation of 802.11 WLAN systems was done using channel models for 2.4 GHz (802.11g) and 5 GHz (802.11a) that had been developed by individual researchers or companies. There was no general consensus on which models to use for performance evaluation of these earlier systems. To address this shortcoming, the 802.11n task group (TG) was formed to develop a standard set of models, called the TGn models, to evaluate 802.11n systems. As per the specifications of the 802.11n standard, these models include both 20 MHz and 40 MHz channels with up to 4 antennas at the transmitter and receiver. Six different TGn channel models (models A-F) comprise the TGn family to characterize propagation in different size homes and offices as well as outdoors. All six follow the dual-slope path loss model (2.28) with a path loss exponent $\gamma_1 = 2$ prior to the breakpoint distance $d_{\rm BP}$ and $\gamma_2 = 3.5$ above this distance. The breakpoint distance $d_{\rm BP} = 5$ m in models A, B, and C, which correspond to LOS only, small home, and large home, respectively. This distance ranges from 10-30 m in models D-F (corresponding to small office, large office, and outdoors, respectively). The delay spread of the multipath, normalized with respect to the LOS path, becomes the difference between the delay of the multipath component with the largest delay and the LOS path. In these models this normalized delay spread ranges from 0 ns for the A model to 150 ns for the F model. Multipath reflections are assumed to arrive in clusters based on the indoor propagation model developed in [60]. Antenna correlation is modeled using the Kronecker model described in [61]. Shadow fading follows the log-normal model with standard deviation $\sigma_{\psi_{\rm dB}} = 3$ for distances $d < d_{\rm BP}$. For $d > d_{\rm BP}$, $\sigma_{\psi_{\rm dB}} = 4$ for Models A and B, $\sigma_{\psi_{\rm dB}} = 5$ for Models C and D, and $\sigma_{\psi_{\rm dB}} = 6$ for Models E and F. More details on the TGn models can be found in [62].

The TGac family of channels extends the TGn model to support performance evaluation of 802.11ac systems. This model increases the number of transmit and receive antennas using the same correlation model as TGn, increases channel bandwidth up to 160 MHz, and modifies the multipath clustering models to account for simultaneous transmission to multiple users (multiuser MIMO). The TGax family of channel models provides additional environmental scenarios for dense deployments both indoors and outdoors. Details of these extensions can be found in [63, 64].

2.9.5 Millimeter Wave Models

The millimeter (mmW) frequency band, spanning the 30-300 GHz range, was not used in the first few generations of Wi-Fi and cellular systems due to its high path loss as well as the high cost and poor performance of mmW hardware. However, as demand for increased data rates in wireless systems continued unabated with each new generation, utilizing the large amount of available spectrum in this band grew increasingly appealing. Moreover, in 2005 the FCC in the US created more flexible rules for spectrum allocation in the lower part of this band to increase its use [65]. Following this ruling, a few commercial mmW communication systems were developed for in-home use, but the technology was not yet competitive with Wi-Fi. Interest in utilizing mmW wave spectrum for cellular communications also started to grow, leading to outdoor channel measurement campaigns as well as cellular system designs to overcome the propagation challenges at these high frequencies. This section provides an overview of mmW propagation characteristics.

Since path loss is inversely proportional to frequency squared in the free space path loss formula (2.6), this loss is much higher at mmW frequencies than at lower frequencies. However, this large free-space path loss can be compensated by high gains associated with directional or multiple antenna technologies. In addition to free-space path loss, signal propagation in the mmW band experiences significant attenuation at specific carrier frequencies due to interactions with oxygen, water, and other molecular components in the air. Attenuation due to shadowing from objects and people is more severe at these frequencies and shadowing can also cause scattering of directed beams. As a result, there is typically only a small number of multipath components in mmWave channels whose received power is above the noise floor.

The resonant frequency of oxygen (O_2) molecules leads to heavy absorption, up to 15 dB/km, for signals transmitted through the air at carrier frequencies around 60 GHz. There is also significant absorption due to the resonant frequencies of O_2 or water vapor (H_2O) at 119, 183, and 325 GHz, with a small peak at 24 GHz. The combined effects of free-space path loss and the O_2 and H_2O absorption in non-rainy conditions is illustrated in Fig. 2.11 from [67]. It can be seen from this figure that atmospheric absorption is somewhat larger at sea level than at higher altitudes due to increased water vapor concentration.

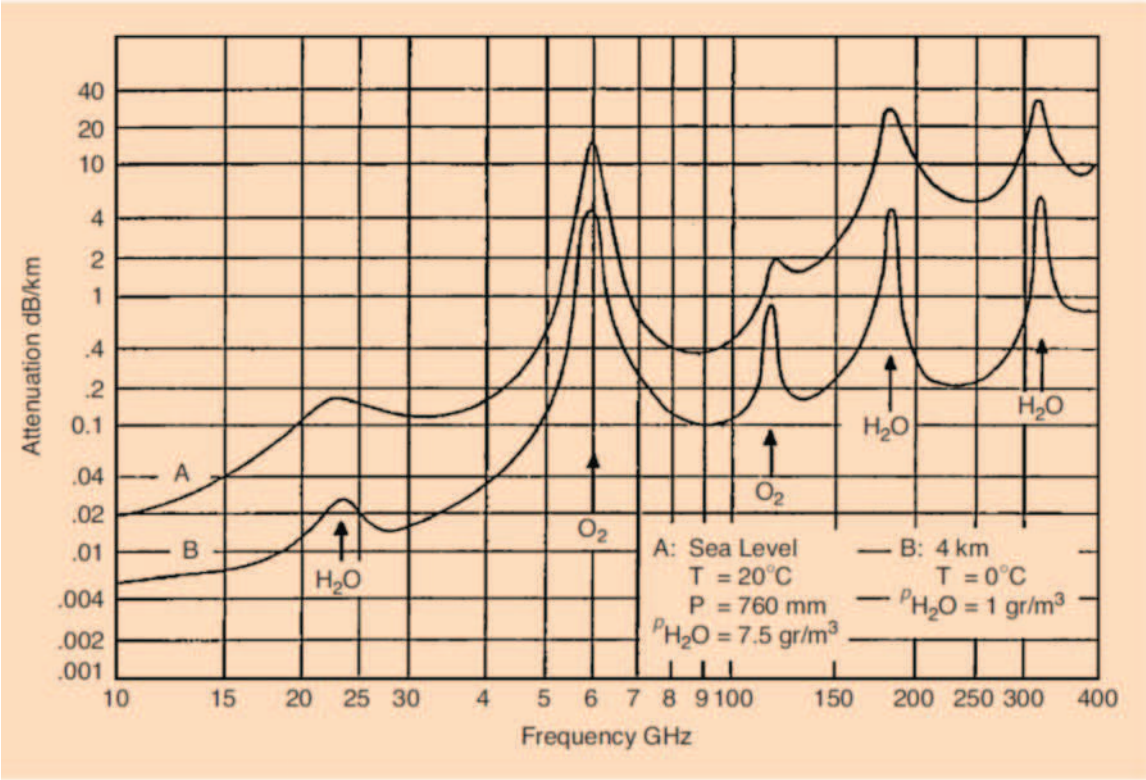

Figure 2.11: Path Loss, O2, and H2O Absorption in Millimeter Wave Propagation [67].

Millimeter wave propagation is also affected by rain absorption. In particular, rain causes scattering since the wavelength of a mmW signal is roughly the same size as a raindrop. The higher the rain rate, the more scattering occurs and hence the higher the absorption. A model for attenuation at different rain rates for mmW propagation was developed by the ITU [68], as illustrated in Fig. 2.12 from [66]. A simple approximation, developed in [69], gives the rain attenuation as $.95R^{.77}$ for R the rain rate in mm/hr.

While atmospheric absorption and rain attentuation entail very large path loss over large distances, most

mmWave systems are designed to operate over relatively short distances, less than 500 m. The losses illustrated in Figs. 2.11-2.12 for these short distances can be compensated through a combination of increased transmit power, antenna techniques, modulation choice, and channel coding.

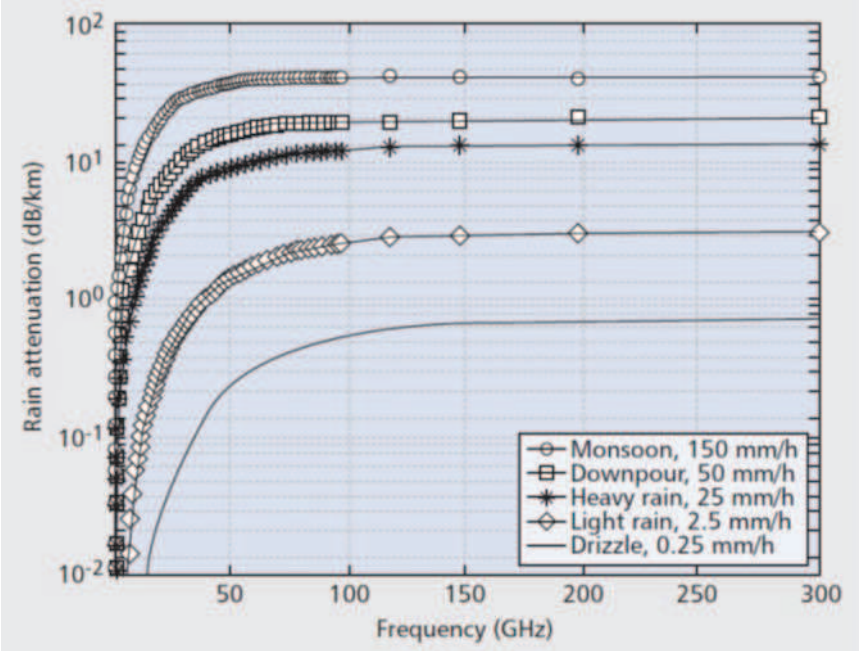

Figure 2.12: Rain Attenuation in Millimeter Wave Propagation [66].

As interest in mmW frequency utilization for cellular and Wi-Fi grew, extensive measurement campaigns were undertaken in different indoor and outdoor environments [70], with analytical channel models developed to approximate these empirical measurements. The most widely known of these models, summarized in [71], are based on work done by four different organizations: the 3GPP standards body, the 5G Channel Model (5GCM) ad hoc group, the Mobile and wireless communications Enablers for the Twentytwenty (2020) Information Society (METIS) research project, and the Millimeter-Wave Based Mobile Radio Access Network for 5G Integrated Communications (mmMAGIC) research project. The models for each group capture propagation for four different categories of environments: outdoor macrocell, outdoor small cell, home, and commercial building. In addition to models for path loss and shadowing, these mmW propagation models add a new component called *LOS probability*, which captures the probability that the LOS signal path between the transmitter and receiver is not blocked or heavily attenuated. While this probability is different for the four environment categories within each family of models, in all cases it depends on the transmit-receive distance and does not depend on the carrier frequency [72]. Millimeter wave path loss in all these models is assumed to follow the dual-slope model (2.28), where the breakpoint distance is different for each of the four environments. The shadowing standard deviation for each of the four environments is a fixed constant. As expected, shadowing in these models is larger for non-LOS compared to LOS environments, and ranges from about 1-8 dB. A summary of the differences between mmW and microwave channel models can be found in [72]. The sensitivity of performance analysis to the particular channel model adopted is demonstrated in [79]. The terahertz (THz) band (.3-3 THz) offers even more spectrum than the mmw band, along with more challenging propagation characteristics [73].

Table 2.5: Typical partition losses at $f_c \approx 1$ GHz

Partition type	Partition loss (dB)
Cloth partition	1.4
Double plasterboard wall	3.4
Foil insulation	3.9
Concrete wall	13
Aluminum siding	20.4
All metal	26

2.9.6 Indoor Attentuation Models

Indoor environments differ widely in the materials used for walls and floors, the layout of rooms, hallways, windows, and open areas, the location and material in obstructing objects, the size of each room, and the number of floors. All of these factors have a significant impact on path loss in an indoor environment. While the TGn, TGac, and TGax Wi-Fi models capture propagation in several generic indoor environments, the nature of these generic models make them inaccurate in characterizing propagation for any particular indoor environment. In this section we describe indoor models that can be tailored to such environments.

Indoor path loss models must accurately capture the effects of attenuation across floors due to partitions as well as between floors. Measurements across a wide range of building characteristics and signal frequencies indicate that the attenuation per floor is greatest for the first floor that is passed through and decreases with each subsequent floor. Specifically, measurements in [18, 94, 19, 20] indicate that, at 900 MHz, the attenuation when transmitter and receiver are separated by a single floor ranges from 10–20 dB, while subsequent attenuation is 6–10 dB per floor for the next three floors and then a few decibels per floor for more than four floors. At higher frequencies the attenuation loss per floor is typically larger [94, 95]. The attenuation per floor is thought to decrease as the number of attenuating floors increases because of the scattering up the side of the building and reflections from adjacent buildings. Partition materials and dielectric properties vary widely and thus so do partition losses. Measurements for the partition loss at different frequencies for different partition types can be found in [11, 18, 96, 97, 23], and Table 2.5 indicates a few examples of partition losses measured at around 1 GHz from this data. The partition loss obtained by different researchers for the same partition type at the same frequency often varies widely, so it is difficult to make generalizations about partition loss from a specific data set.

The experimental data for floor and partition loss can be added to an analytical or empirical dB path loss model $P_L(d)$ as

$$P_r\,\mathrm{dBm} = P_t\,\mathrm{dBm} - P_L(d) - \sum_{i=1}^{N_f} \mathrm{FAF}_i - \sum_{i=1}^{N_p} \mathrm{PAF}_i, \tag{2.69}$$

where FAF_i represents the floor attenuation factor for the ith floor traversed by the signal and PAF_i represents the partition attenuation factor associated with the ith partition traversed by the signal. The number of floors and partitions traversed by the signal are N_f and N_p, respectively.

Another important factor for indoor systems whose transmitter is located outside the building is outdoor-to-indoor penetration loss. Measurements indicate that this penetration loss is a function of frequency, height, and the exterior wall or window material. For carrier frequencies from .9-3 GHz penetration loss measurements range from 8 dB to 20 dB [2, 21, 100, 101]. Penetration loss for mmWave signals generally increases with frequency. Hence, the 3GPP, ITU, 5GCM, and mmMAGIC propagation models include a component for outdoor-to-indoor penetration loss as a function of both frequency and the exterior wall or window material [57, 58, 70]. The ITU penetration loss model as a function of frequency f_c and for different wall and window materials is given in Table 2.6. In general penetration loss decreases by about 1.4 dB per floor at floors above the ground floor. This

Table 2.6: ITU penetration loss model

Material	Penetration loss (dB)
Standard multi-pane glass	$2 + 0.2f_c$
Infrared reflective glass	$23 + .3f_c$
Concrete	$5 + 4f_c$
Wood	$4.85 + 0.12f_c$

decrease is due to reduced clutter at higher floors and the higher likelihood of a LOS path.

Chapter 2 Problems

1. Under the free-space path loss model, find the transmit power required to obtain a received power of 1 dBm for a wireless system with isotropic antennas ($G_t = G_r = 1$) and a carrier frequency $f_c = 5$ GHz, assuming a distance $d = 10\,\text{m}$. Repeat for $d = 100\,\text{m}$.

2. For the two-ray model with transmitter–receiver separation $d = 100\,\text{m}$, $h_t = 10\,\text{m}$, and $h_r = 2\,\text{m}$, find the delay spread between the two signals.

3. For the two-ray model, show how a Taylor series approximation applied to (2.14) results in the approximation

$$\Delta\phi = \frac{2\pi(d_1 - d_0)}{\lambda} \approx \frac{4\pi h_t h_r}{\lambda d}.$$

4. For the two-ray model, derive an approximate expression for the distance values below the critical distance d_c at which signal nulls occur.

5. Find the critical distance d_c under the two-ray model for a large macrocell in a suburban area with the base station mounted on a tower or building ($h_t = 20\,\text{m}$), the receivers at height $h_r = 3\,\text{m}$, and $f_c = 2$ GHz. Is this a good size for cell radius in a suburban macrocell? Why or why not?

6. Suppose that, instead of a ground reflection, a two-ray model consists of a LOS component and a signal reflected off a building to the left (or right) of the LOS path. Where must the building be located relative to the transmitter and receiver for this model to be the same as the two-ray model with a LOS component and ground reflection?

7. Consider a two-ray channel with impulse response $h(t) = \alpha_1\delta(t) + \alpha_2\delta(t - .022\,\mu\text{s})$, so for input signal $s(t)$ to the channel, the output is $r(t) = h(t) * s(t)$. Find the distance separating the transmitter and receiver, as well as α_1 and α_2, assuming free-space path loss on each path with a reflection coefficient of -1. Assume the transmitter and receiver are located 8 m above the ground and that the carrier frequency is 900 MHz.

8. Directional antennas are a powerful tool to reduce the effects of multipath as well as interference. In particular, directional antennas along the LOS path for the two-ray model can reduce the attenuation effect of ground wave cancellation, as will be illustrated in this problem. Assume the reference distance $d_r = 1$ m. Plot the dB power ($10\log_{10} P_r$) versus log distance ($\log_{10} d$) for the two-ray model with parameters $f_c = 900$ MHz, $R = -1$, $h_t = 50$ m, $h_r = 2$ m, $G_0 = 1$, and the following values for $G_1 : G_1 = 1, .316, .1,$ and .01 (i.e., $G_1 = 0,\ -5,\ -10,$ and -20 dB, respectively). Each of the four plots should range in distance from $d = 1$ m to $d = 100$ km. Also calculate and mark the critical distance $d_c = 4h_t h_r/\lambda$ on each plot, and normalize the plots to start at approximately 0 dB. Finally, show the piecewise linear model with flat power falloff up to distance h_t, falloff $10\log_{10}(d^{-2})$ for $h_t < d < d_c$, and falloff $10\log_{10}(d^{-4})$ for $d \geq d_c$. (On the power loss versus log distance plot, the piecewise linear curve becomes a set of three straight lines of slope $0, 2,$ and 4, respectively.) Note that at large distances it becomes increasingly difficult to have $G_1 \ll G_0$ because this requires extremely precise angular directivity in the antennas.

9. Under what conditions is the single-slope path loss model (2.23) the same as the free-space path loss model (2.7)?

10. Consider a receiver with noise power –160 dBm within the signal bandwidth of interest. Assume a single-slope path loss model with $d_0 = 1\ \text{m}, K$ obtained from the free-space path loss formula with isotropic antennas and $f_c = 1$ GHz, and $\gamma = 4$. For a transmit power of $P_t = 10\,\text{mW}$, find the maximum distance between the transmitter and receiver such that the received signal-to-noise power ratio is 20 dB.

11. For the set of empirical measurements of P_r/P_t given in Table 2.2, find the path loss exponent γ and parameter K that minimizes the MSE between the single-slope model (2.24) in dB and the empirical dB power measurements, assuming that $d_r = 1\,\mathrm{m}$. This problem differs from Example 2.5.1 in that now the two parameters γ and K are jointly optimized to minimize MSE, rather than fixing the value of K based on free-space path loss or the measured attenuation at the reference distance d_r. Find the received power at 150 m for this single-slope path loss model with transmit power of 1 mW (0 dBm) and compare with the result in Example 2.5.1. Does the better fitting of the data through this two-dimensional optimization lead to a larger or smaller value for this received power? Will this always be the case for any set of empirical measurements (why or why not)?

12. Find parameters for a multi-slope model with three segments to approximate the two-ray model path loss (2.13) over distances between 10 and 1000 meters, assuming $h_t = 10\,\mathrm{m}$, $h_r = 2\,\mathrm{m}$, and $G_0 = G_1 = 1$. Plot the path loss and the piecewise linear approximation using these parameters over this distance range.

13. This problem shows how different propagation models can lead to very different SNRs (and therefore different link performance) for a given system design. Consider a linear cellular system using frequency division, as might operate along a highway or rural road (see Figure 2.13). Each cell is allocated a certain band of frequencies, and these frequencies are reused in cells spaced a distance d away. Assume the system has square cells, 2 km per side, and that all mobiles transmit at the same power P. For the following propagation models, determine the minimum distance that the cells operating in the same frequency band must be spaced so that uplink SNR (the ratio of the minimum received signal-to-interference or S/I power from mobiles to the base station) is greater than 20 dB. You can ignore all interferers except those from the two nearest cells operating at the same frequency.

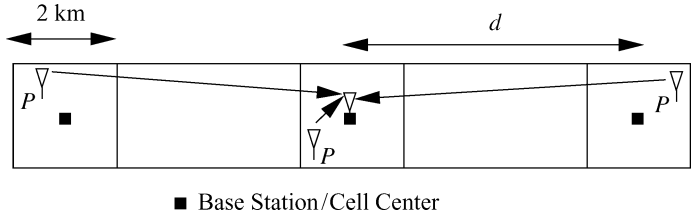

Figure 2.13: Linear cellular system for Problem 2-14.

    (a) Propagation for both signal and interference follow a free-space model.

    (b) Propagation for both signal and interference follow the single-slope path loss model (2.23) with $d_0 = 100\,\mathrm{m}$, $K = 1$, and $\gamma = 3$.

    (c) Propagation for the signal follows the single-slope path loss model with $d_0 = 100\,\mathrm{m}$, $K = 1$, and $\gamma = 2$, while propagation of the interfererence follows the same model but with $\gamma = 4$.

14. Show that the log-normal distribution for ψ given in (2.31) yields the Gaussian distribution for ψ_{dB} given in (2.34).

15. Table 2.7 lists a set of empirical path loss measurements. Assume a carrier frequency $f_c = 2$ GHz.

    (a) Find the parameters of a single-slope path loss model plus log-normal shadowing that best fit this data assuming K is calculated from free-space path loss at the reference distance $d_r = 1$ m.

    (b) Find the path loss at 2 km based on this model.

Table 2.7: path loss measurements for Problem 2-14

.

Distance from transmitter	P_r/P_t
5 m	–60 dB
25 m	–80 dB
65 m	–105 dB
110 m	–115 dB
400 m	–135 dB
1000 m	–150 dB

(c) Find the outage probability at a distance d assuming the received power at d due to path loss alone is 10 dB above the required power for non-outage.

16. Consider a cellular system operating at 900 MHz where propagation follows free-space path loss with variations about this path loss due to log-normal shadowing with standard deviation $\sigma = 6\,\mathrm{dB}$. Suppose that for acceptable voice quality a signal-to-noise power ratio of 15 dB is required at the mobile. Assume the base station transmits at 1 W and that its antenna has a 3 dB gain. There is no antenna gain at the mobile, and the receiver noise in the bandwidth of interest is $-60\ \mathrm{dBm}$. Find the maximum cell size such that a mobile on the cell boundary will have acceptable voice quality 90% of the time.

17. In this problem we will simulate the log-normal fading process over distance based on the autocovariance model (2.39). As described in the text, the simulation first generates a white noise process and then passes it through a first-order filter with a pole at $e^{-\delta/X_c}$. Assume $X_c = 20\,\mathrm{m}$ and plot the resulting log-normal fading process over a distance d ranging from $0\ \mathrm{m}$ to $200\ \mathrm{m}$, sampling the process every meter. You should normalize your plot about $0\ \mathrm{dB}$, since the mean of the log-normal shadowing is captured by path loss.

18. In this problem we will explore the impact of different log-normal shadowing parameters on outage probability. Consider a cellular system where the received signal power in dB has a Gaussian distribution with mean $P_r\ \mathrm{dBm}$ and standard deviation $\sigma\ \mathrm{dB}$. Assume the received signal power must be above 10 dBm for acceptable performance.

    (a) What is the outage probability when $P_r = 15\ \mathrm{dBm}$ and $\sigma = 8\ \mathrm{dB}$?

    (b) For $\sigma = 4\ \mathrm{dB}$, find the value of P_r required for the outage probability to be less than l% – a typical value for cellular systems.

    (c) Repeat part (b) for $\sigma = 12\ \mathrm{dB}$.

    (d) One proposed technique for reducing outage probability is to use *macrodiversity*, where a mobile unit's signal is received by multiple base stations and then combined. This can only be done if multiple base stations are able to receive a given mobile's signal. Explain why this might reduce outage probability.

19. Derive the formula for cell coverage percentage (2.49) by applying integration by parts to (2.47).

20. The cell coverage percentage is independent of the decorrelation distance X_c of the shadowing. Explain why. Now suppose it is known that a particular set of cell locations near the transmitter are in outage. Will the cell coverage percentage with this side information now depend on the decorrelation distance and, if so, how and why?

21. Find the coverage area for a microcellular system where path loss follows the single-slope model (with $\gamma = 3,\ d_0 = 1$ and $K = 0$ dB) and there is also log-normal shadowing with $\sigma = 4$ dB. Assume a cell radius of 100 m, a transmit power of 80 mW, and a minimum received power requirement of $P_{\min} = -100$ dBm.

22. Consider a cellular system where (a) path loss follows the single-slope model with $\gamma = 6$ and (b) there is also log-normal shadowing with $\sigma = 8$ dB. If the received power at the cell boundary due to path loss is 20 dB higher than the minimum required received power for non-outage, find the cell coverage area.

23. In microcells, path loss exponents usually range from 2 to 6 and shadowing standard deviation typically ranges from 4 to 12. Given a cellular system in which the received power due to path loss at the cell boundary equals the desired level for non-outage, find the path loss and shadowing parameters within these ranges that yield the best and worst coverage area. What is the coverage area when these parameters are in the middle of their typical ranges?

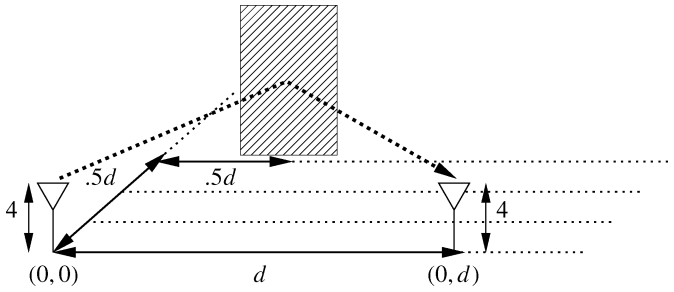

Figure 2.14: System with scattering for Problem 2-11.

24. What average power falloff with distance do you expect for the ten-ray model if $G_0 = G_1 >> G_i$ for $i = 2, ..., 9$? Repeat for $G_0 = G_1 = G_2 >> G_i$ for $i = 3, ..., 9$. Explain your answers in both cases.

25. For the ten-ray model, assume that the transmitter and receiver are at the same height in the middle of a street of width 20 m. The transmitter-receiver separation is 500 m. Find the delay spread for this model.

26. Consider a system with a transmitter, receiver, and scatterer as shown in Figure 2.14. Assume the transmitter and receiver are both at heights $h_t = h_r = 4$ m and are separated by distance d, with the scatterer at distance $.5d$ along both dimensions in a two-dimensional grid of the ground – that is, on such a grid the transmitter is located at $(0, 0)$, the receiver at $(0, d)$, and the scatterer at $(.5d,\ .5d)$. Assume a radar cross-section of 20 dBm^2, $G_s = 1$, and $f_c = 900$ MHz. Find the path loss of the scattered signal for $d = 1, 10, 100$, and 1000 meters. Compare with the path loss at these distances if the signal is only reflected, with reflection coefficient $R = -1$.

27. Find the median path loss under the Hata model assuming $f_c = 900$ MHz, $h_t = 20$ m, $h_r = 5$ m, and $d = 100$ m for a large urban city, a small urban city, a suburb, and a rural area. Explain qualitatively the path loss differences for these four environments.

28. Consider a wireless link operating outdoors over 200m in the mmW freqency band. Assume the path loss follows the single-slope model with a path loss exponent $\gamma = 2$ and reference distance for antenna far field $d_r = 1$ m. Without any additional attenuation caused by atmospheric and rain conditions, the model attenuation factor $K = 1$. Assume rain attenuation follows the simple approximation $K_{rain} = .95R^{.77}$ for R the rain rate in mm/hr. Assume the atmospheric absorption given in Fig. 2.11.

(a) What is the received signal power due to path loss and oxygen absorption at carrier frequencies of 60 GHz and of 80 GHz assuming a transmit power of 1W?

(b) Consider now only the 80 GHz link. Assume a day where it is dry at 8 am, there is a heavy drizzle of 2.5mm/Hr at 12 pm and a heavy downpour of 50mm/Hr at 5 pm. What is the required transmit power at 8 am, 12 pm and 5 pm if we desire a received signal power of -50dBm at each of these time instances?

29. Using the indoor attentuation model, determine the required transmit power for a desired received power of –110 dBm for a signal transmitted over 100 m that goes through three floors with attenuation 15 dB, 10 dB, and 6 dB (respectively) as well as two double plaster-board walls. Assume a reference distance $d_r = 1$, exponent $\gamma = 4$, and constant $K = 0$ dB.

Bibliography

[1] W. C. Jakes, Jr., *Microwave Mobile Communications*, Wiley, New York, 1974 [reprinted by IEEE Press].

[2] D. Parsons, *The Mobile Radio Propagation Channel*, 2nd Ed., Wiley, New York, 2000.

[3] M. Pätzold, *Mobile Fading Channels*, 2nd Ed., Wiley, New York, 2011.

[4] H. L. Bertoni, *Radio Propagation for Modern Wireless Systems*, Prentice Hall, 1999.

[5] J. Cavers, *Mobile Channel Characteristics*, Springer 2000.

[6] *IEEE J. Sel. Areas Commun.*, Special Issue on Channel and Propagation Modeling for Wireless Systems Design, April 2002 and August 2002.

[7] N. Costa and S. Haykin, *Multiple-input Multiple-output Channel Models: Theory and Practice*, John Wiley and Sons, 2010.

[8] *IEEE J. Sel. Areas Commun.*, Special Issue on Ultra-Wideband Radio in Multiaccess Wireless Communications, December 2002.

[9] T. Rappaport, R. W. Heath Jr, R. C. Daniels, and J. N. Murdock. *Millimeter Wave Wireless Communications*. Pearson Education, 2014.

[10] G. H. Spencer and M. V. R. K. Murty, "General Ray-Tracing Procedure". *Journal of the Optical Society of America*, Vol. 52, Issue 6, pp. 672-678, June 1962.

[11] T. S. Rappaport, *Wireless Communications-Principles and Practice*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2001.

[12] J. L. Volakis, *Antenna Engineering Handbook*, 4th Ed. McGraw-Hill, 2007.

[13] H. T. Friis, "A note on a simple transmission formula", *Proc. IRE*, vol. 34, no. 5, pp. 254-256, May 1946.

[14] A. S. Y. Poon and R. W. Brodersen, "The role of multiple-antenna systems in emerging open access environments," *EE Times Commun. Design Conf.*, October 2003.

[15] A. J. Rustako, Jr., N. Amitay, G. J. Owens, and R. S. Roman, "Radio propagation at microwave frequencies for line-of-sight microcellular mobile and personal communications," *IEEE Trans. Veh. Tech.*, pp. 203–10, February 1991.

[16] W. C. Y. Lee, *Mobile Communications Engineering*, McGraw-Hill, New York, 1982.

[17] V. Erceg, L. J. Greenstein, S. Y. Tjandra, S. R. Parkoff, A. Gupta, B. Kulic, A. A. Julius, and R. Bianchi, "An empirically based path loss model for wireless channels in suburban environments," *IEEE J. Sel. Areas Commun*., pp. 1205–11, July 1999.

[18] S. Y. Seidel and T. S. Rappaport, "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings," *IEEE Trans. Ant. Prop*., pp. 207–17, February 1992.

[19] A. F. Toledo and A. M. D. Turkmani, "Propagation into and within buildings at 900, 1800, and 2300 MHz," *Proc. IEEE Veh. Tech. Conf*., pp. 633–6, May 1992.

[20] F. C. Owen and C. D. Pudney, "Radio propagation for digital cordless telephones at 1700 MHz and 900 MHz," *Elec. Lett*., pp. 52–3, September 1988.

[21] A. F. Toledo, A. M. D. Turkmani, and J. D. Parsons, "Estimating coverage of radio transmission into and within buildings at 900, 1800, and 2300 MHz," *IEEE Pers. Commun. Mag*., pp. 40–7, Apri1 1998.

[22] D. M. J. Devasirvathan, R. R. Murray, and D. R. Woiter, "Time delay spread measurements in a wireless local loop test bed," *Proc. IEEE Veh. Tech. Conf*., pp. 241–5, May 1995.

[23] G. Durgin, T. S. Rappaport, and H. Xu, "Partition-based path loss analysis for in-home and residential areas at 5.85 GHz," *Proc. IEEE Globecom Conf*., pp. 904–9, November 1998.

[24] T. Kurner, D. J. Cichon, and W. Wiesbeck, "Concepts and results for 3D digital terrain-based wave propagation models: An overview," *IEEE J. Sel. Areas Commun*., pp. 1002–12, September 1993.

[25] A. J. Rustako, Jr., N. Amitay, G. J. Owens, and R.S. Roman, Radio propagation at microwave frequencies for line-of-sight microcellular mobile and personal communications, IEEE Trans. Veh. Technol., vol. 40, pp. 203–210, February 1991.

[26] N. Amitay, "Modeling and computer simulation of wave propagation in lineal line-of-sight microcells," *IEEE Trans. Veh. Tech*., pp. 337–42, November 1992.

[27] J. W. McKown and R. L. Hamilton, Jr., "Ray tracing as a design tool for radio networks," *IEEE Network*, pp. 27–30, November 1991.

[28] K. A. Remley, H. R. Anderson, and A. Weisshar, "Improving the accuracy of ray-tracing techniques for indoor propagation modeling," *IEEE Trans. Veh. Tech*., pp. 2350–8, November 2000.

[29] A. Domazetovic, L. J. Greenstein, N. Mandayan, and I. Seskar, "A new modeling approach for wireless channels with predictable path geometries," *Proc. IEEE Veh. Tech. Conf*., September 2002.

[30] J. H. Tarng, W.-S. Liu, Y.-F. Huang, and J.-M. Huang, "A novel and efficient hybrid model of radio multipath-fading channels in indoor environments," *IEEE Trans. Ant. Prop*., pp. 585–94, March 2003.

[31] Z. Yun and M. F. Iskander. "Ray tracing for radio propagation modeling: Principles and applications." *IEEE Access* Vol. 3, 2015, pp. 1089-1100.

[32] J.-F. Wagen, "Signal strength measurements at 881 MHz for urban microcells in downtown Tampa," *Proc. IEEE Globecom Conf*., pp. 1313–17, December 1991.

[33] R. J. C. Bultitude and G. K. Bedal, "Propagation characteristics on microcellular urban mobile radio channels at 910 MHz," *IEEE J. Sel. Areas Commun*., pp. 31–9, January 1989.

[34] J.-E. Berg, R. Bownds, and F. Lotse, "Path loss and fading models for microcells at 900 MHz," *Proc. IEEE Veh. Tech. Conf.*, pp. 666–71, May 1992.

[35] J. H. Whitteker, "Measurements of path loss at 910 MHz for proposed microcell urban mobile systems," *IEEE Trans. Veh. Tech.*, pp. 125–9, August 1988.

[36] H. Börjeson, C. Bergljung, and L. G. Olsson, "Outdoor microcell measurements at 1700 MHz," *Proc. IEEE Veh. Tech. Conf.*, pp. 927–31, May 1992.

[37] K. Schaubach, N. J. Davis IV, and T. S. Rappaport, "A ray tracing method for predicting path loss and delay spread in microcellular environments," *Proc. IEEE Veh. Tech. Conf.*, pp. 932–5, May 1992.

[38] F. Ikegami, S. Takeuchi, and S. Yoshida, "Theoretical prediction of mean field strength for urban mobile radio," *IEEE Trans. Ant. Prop.*, pp. 299–302, March 1991.

[39] M. C. Lawton and J. P. McGeehan, "The application of GTD and ray launching techniques to channel modeling for cordless radio systems," *Proc. IEEE Veh. Tech. Conf.*, pp. 125–30, May 1992.

[40] J. B. Keller, "Geometrical theory of diffraction," *J. Opt. Soc. Amer.*, 52, pp. 116–30, 1962.

[41] R. J. Luebbers, "Finite conductivity uniform GTD versus knife edge diffraction in prediction of propagation path loss," *IEEE Trans. Ant. Prop.*, pp. 70–6, January 1984.

[42] C. Bergljung and L. G. Olsson, "Rigorous diffraction theory applied to street microcell propagation," *Proc. IEEE Globecom Conf.*, pp. 1292–6, December 1991.

[43] R. G. Kouyoumjian and P. H. Pathak, "A uniform geomenical theory of diffraction for an edge in a perfectly conducting surface," *Proc. IEEE*, pp. 1448–61, November 1974.

[44] G. K. Chan, "Propagation and coverage prediction for cellular radio systems," *IEEE Trans. Veh. Tech.*, pp. 665–70, November 1991.

[45] K. C. Chamberlin and R. J. Luebbers, "An evaluation of Longley–Rice and GTD propagation models," *IEEE Trans. Ant. Prop.*, pp. 1093–8, November 1982.

[46] J. Walfisch and H. L. Bertoni, "A theoretical model of UHF propagation in urban environments," *IEEE Trans. Ant. Prop.*, pp. 1788–96, October 1988.

[47] D. Chizhik, G. J. Foschini, M. J. Gans and R. A. Valenzuela, "Keyholes, correlations, and capacities of multielement transmit and receive antennas," *IEEE Transactions on Wireless Communications*, vol. 1, no. 2, pp. 361-368, April 2002.

[48] W. Zhang, "Fast two-dimensional diffraction modeling for site-specific propagation prediction in urban microcellular environments," *IEEE Trans. Veh. Techn.*, vol. 49, no. 2, pp. 428436, 2000.

[49] S. Deng, G. R. MacCartney and T. S. Rappaport, "Indoor and Outdoor 5G Diffraction Measurements and Models at 10, 20, and 26 GHz," *2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1-7.

[50] M. Jacob, S. Priebe, R. Dickhoff, T. Kleine-Ostmann, T. Schrader and T. Kurner, "Diffraction in mm and Sub-mm Wave Indoor Propagation Channels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 3, pp. 833-844, March 2012.

[51] M. I. Skolnik, *Introduction to Radar Systems*, 2nd ed., McGraw-Hill, New York, 1980.

[52] S. Y. Seidel, T. S. Rappaport, S. Jain, M. L. Lord, and R. Singh, "Path loss, scattering, and multipath delay statistics in four European cities for digital cellular and microcellular radiotelephone," *IEEE Trans. Veh. Tech.*, pp. 721–30, November 1991.

[53] S. T. S. Chia, "1700 MHz urban microcells and their coverage into buildings," *Proc. IEEE Ant. Prop. Conf.*, pp. 504–11, York, U.K., Apri1 1991.

[54] T. Fugen, J. Maurer, W. Sorgel and W. Wiesbeck, "Characterization of multipath clusters with ray-tracing in urban MIMO propagation environments at 2 GHz," *in 2005 IEEE Antennas and Propagation Society International Symposium*, Washington, DC, 2005, pp. 410-413.

[55] M. A. Jensen and J. W. Wallace, "A review of antennas and propagation for MIMO wireless communications," *IEEE Transactions on Antennas and Propagation*, pp. 2810-2824, Nov. 2004.

[56] K. H. Ng, E. K. Tameh, A. Doufexi, M. Hunukumbure and A. R. Nix, "Efficient multielement ray Tracing with site-specific comparisons using measured MIMO channel data," *IEEE Transactions on Vehicular Technology*, pp. 1019-1032, May 2007.

[57] 3GPP, Study on channel model for frequencies from 0.5 to 100 GHz, 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 38.901, March 2017, Website: www.3gpp.org/DynaReport/38901.htm

[58] IMT, "Guidelines for evaluation of radio interface technologies for IMT-2020," Tech Rep. ITU-R M.2412-0, Oct. 2017, Website: www.itu.int/pub/R-REP-M.2412.

[59] P. Kyosti et al., WINNER II channel models, Eur. Commission, IST-4-027756-WINNER, Brussels, Belgium, Tech. Rep. D1.1.2, Sep. 2007

[60] A. A. M. Saleh and R. Valenzuela, "A Statistical Model for Indoor Multipath Propagation," in IEEE Journal on Selected Areas in Communications, vol. 5, no. 2, pp. 128-137, February 1987.

[61] L. Schumacher, K. I. Pedersen, and P.E. Mogensen, From antenna spacings to theoretical capacities guidelines for simulating MIMO systems, in Proc. PIMRC Conf., vol. 2, Sept. 2002, pp. 587-592.

[62] V. Erceg et al., TGn Channel Models. Doc. IEEE802.11-03/940r4.

[63] G. Breit et. al., "TGac Channel Model Addendum. IEEE 802.11-09/0308r12.

[64] J. Liu et. al., "IEEE 802.11ax Channel Model Document." IEEE 802.11-14/0882r4.

[65] FCC Memorandum and Opinion Order 05-45 *WTB Allocations and Service Rules for the 71-76 GHz, 81-86 GHz and 92-95 GHz Bands*. Feb. 2005.

[66] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems, *IEEE Commun. Mag.*, pp. 101107, June 2011.

[67] M. Marcus and B. Pattan, Millimeter wave propagation: Spectrum management implications, *IEEE Microwave Mag.*, pp. 5462, June 2005.

[68] ITU-R P.838-3, Specific Model for Rain for Use in Prediction Methods, 2005.

[69] J. Ostrometzky and H. Messer, "Accumulated rainfall estimation using maximum attenuation of microwave radio signal," *Proc. IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 193-196, June 2014.

[70] M. Xiao et al., "Millimeter Wave Communications for Future Mobile Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909-1935, Sept. 2017.

[71] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios and J. Zhang, "Overview of Millimeter Wave Communications for Fifth-Generation (5G) Wireless Networks With a Focus on Propagation Models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213-6230, Dec. 2017.

[72] M. Shafi, J. Zhang, H. Tataria, A. F. Molisch, S. Sun, T. S. Rappaport, F. Tufvesson, S. Wu, K. Kitao, "Microwave vs. Millimeter-Wave Propagation Channels: Key Differences and Impact on 5G Cellular Systems," *IEEE Communications Magazine*, Vol. 56, No. 12., pp. 1420, Dec. 2018.

[73] I. F. Akyildiza, J. M. Jornet, C. Han, "Terahertz band: Next frontier for wireless communications," *Elsevier J. Physical Commun.*, Volume 12, pp. 16-21, Sept. 2014.

[74] METIS. METIS Channel Model METIS2020, Deliverable D1.4 v3. [Online]. Available: https://www.metis2020.com/wpcontent/uploads/deliverables/METIS_D1.4_v1.0.pdf. July 2015.

[75] E. Ostlin, H.-J. Zepernick, and H. Suzuki, "Macrocell path-loss prediction using artificial neural networks," *IEEE Trans. Veh. Techn.*, vol. 59, no. 6, pp. 27352747, 2010.

[76] P. Almers et. al., Survey of Channel and Radio Propagation Models for Wireless MIMO Systems, EURASIP J. Wireless Commun. Net., vol. 2007.

[77] J. Medbo et al., "Channel modelling for the fifth generation mobile communications," The 8th European Conference on Antennas and Propagation (EuCAP 2014), The Hague, 2014, pp. 219-223.

[78] P. Ferrand, M. Amara, S. Valentin and M. Guillaud, "Trends and challenges in wireless channel modeling for evolving radio access," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 93-99, July 2016.

[79] S. Sun, T. Rappaport, M. Shafi, P. Tang, J. Zhang, and P.J. Smith, "Propagation Models and Performance Evaluation for 5G Millimeter-Wave Bands," *IEEE Transactions on Vehicular Technology*, Vol. 67 , No. 9 , Sept. 2018.

[80] M. Kasparick, R. L. G. Cavalcante, S. Valentin, S. Stanczak, and M. Yukawa, Kernel-based adaptive online reconstruction of coverage maps with side information, IEEE Trans. Veh. Technol., vol. 65, no. 7, pp. 54615473, 2016.

[81] D. Romero, S.-J. Kim, G. B. Giannakis, and R. Lopez-Valcarce, Learning power spectrum maps from quantized power measurements, IEEE Trans. Signal Process., vol. 65, no. 10, pp. 25472560, 2017.

[82] R. Nikbakht, A. Jonsson, and A. Lozano, Dual-kernel online reconstruction of power maps, IEEE Global Commun. Conf. (GLOBECOM18), 2018.

[83] D. Wong and D. C. Cox, "Estimating local mean signal power level in a Rayleigh fading environment," *IEEE Trans. Veh. Tech.*, pp. 956–9, May 1999.

[84] T. Okumura, E. Ohmori, and K. Fukuda, "Field strength and its variability in VHF and UHF land mobile service," *Rev. Elec. Commun. Lab.*, pp. 825–73, September/October 1968.

[85] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Veh. Tech.*, pp. 317-325, August 1980.

[86] European Cooperative in the Field of Science and Technical Research EURO-COST 231, "Urban transmission loss models for mobile radio in the 900 and 1800 MHz bands," rev. 2, The Hague, September 1991.

[87] *Wireless Communications* A. F. Molisch, John Wiley and Sons, 2nd ed. 2010.

[88] A. F. Molisch, H. Asplund, R. Heddergott, M. Steinbauer, and T. Zwick, The COST 259 directional channel model-part I: Overview and methodology, IEEE Trans. on Wireless Communications, vol. 5, no. 12, pp. 3421 3433, Dec. 2006.

[89] N. Czink and C. Oestges, "The COST 273 MIMO Channel Model: Three Kinds of Clusters," *IEEE 10th International Symposium on Spread Spectrum Techniques and Applications*, pp. 282-286, Aug. 2008.

[90] L. Liu et al., "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, pp. 92-99, Dec. 2012.

[91] E. McCune and K. Feher, "Closed-form propagation model combining one or more propagation constant segments," *Proc. IEEE Veh. Tech. Conf.*, pp. 1108–12, May 1997.

[92] M. Feuerstein, K. Blackard, T. Rappaport, S. Seidel, and H. Xia, "Path loss, delay spread, and outage models as functions of antenna height for microcellular system design," *IEEE Trans. Veh. Tech.*, pp. 487–98, August 1994.

[93] P. Harley, "Short distance attenuation measurements at 900 MHz and 1.8 GHz using low antenna heights for microcells," *IEEE J. Sel. Areas Commun.*, pp. 5–11, January 1989.

[94] A. J. Motley and J. M. P. Keenan, "Personal communication radio coverage in buildings at 900 MHz and 1700 MHz," *Elec. Lett.*, pp. 763–4, June 1988.

[95] S. Y. Seidel, T. S. Rappaport, M. J. Feuerstein, K. L. Blackard, and L. Grindstaff, "The impact of surrounding buildings on propagation for wireless in-building personal communications system design," *Proc. IEEE Veh. Tech. Conf.*, pp. 814–18, May 1992.

[96] C. R. Anderson, T. S. Rappaport, K. Bae, A. Verstak, N. Tamakrishnan, W. Trantor, C. Shaffer, and L. T. Waton, "In-building wideband multipath characteristics at 2.5 and 60 GHz," *Proc. IEEE Veh. Tech. Conf.*, pp. 24–8, September 2002.

[97] L.-S. Poon and H.-S. Wang, "Propagation characteristic measurement and frequency reuse planning in an office building," *Proc. IEEE Veh. Tech. Conf.*, pp. 1807–10, June 1994.

[98] A. Ghosh et al., Millimeter-wave enhanced local area systems: A high data-rate approach for future wireless networks, *IEEE J. Sel. Areas Commun.*, pp. 11521163, Jun. 2014.

[99] X. Zhang and J.G. Andrews, "Downlink cellular network analysis with multi-slope path loss models," *IEEE Trans. Commun.*, pp. 1881-1894, May 2015.

[100] R. Hoppe, G. Wölfle, and F. M. Landstorfer, "Measurement of building penetration loss and propagation models for radio transmission into buildings," *Proc. IEEE Veh. Tech. Conf.*, pp. 2298–2302, Apri1 1999.

[101] E. H. Walker, "Penetration of radio signals into buildings in cellular radio environments," *Bell Systems Tech. J.*, pp. 2719–34, September 1983.

[102] S. S. Ghassemzadeh, L. J. Greenstein, A. Kavcic, T. Sveinsson, and V. Tarokh, "Indoor path loss model for residential and commercial buildings," *Proc. IEEE Veh. Tech. Conf.*, pp. 3115–19, October 2003.

[103] A. J. Goldsmith, L. J. Greenstein, and G. J. Foschini, "Error statistics of real-time power measurements in cellular channels with multipath and shadowing," *IEEE Trans. Veh. Tech.*, pp. 439–46, August 1994.

[104] A. J. Goldsmith and L. J. Greenstein, "A measurement-based model for predicting coverage areas of urban microcells," *IEEE J. Sel. Areas Commun.*, pp. 1013–23, September 1993.

[105] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Elec. Lett.*, pp. 2145–6, November 7, 1991.

[106] G. L. Stüber, *Principles of Mobile Communications*, 2nd ed., Kluwer, Dordrecht, 2001.

[107] A. Algans, K. I. Pedersen, and P. E. Mogensen, "Experimental analysis of the joint statistical properties of azimuth spread, delay spread, and shadow fading," *IEEE J. Sel. Areas Commun.*, pp. 523–31, April 2002.

[108] M. Marsan and G. C. Hess, "Shadow variability in an urban land mobile radio environment," *Elec. Lett.*, pp. 646–8, May 1990.

[109] J. Weitzen and T. Lowe, "Measurement of angular and distance correlation properties of lognormal shadowing at 1900 MHz and its application to design of PCS systems," *IEEE Trans.Veh. Tech.*, pp. 265–73, March 2002.

[110] W. Turin, R. Jana, S. S. Ghassemzadeh, V.W. Rice, and V. Tarokh, "Autoregressive modeling of an indoor UWB channel," *Proc. IEEE Conf. UWB Syst. Technol.*, pp. 71–4, May 2002.

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. Although these multipath effects are captured in the ray-tracing models from Chapter 2 for deterministic channels, we rarely have sufficiently precise knowledge about wireless propagation environments to characterize them deterministically. Thus wireless channels are often characterized statistically. The statistical characterization of the random multipath channel is based on its time-varying impulse response. This chapter develops a statistical characterization of this channel model and describes its important properties.

If a single pulse is transmitted over a multipath channel then the received signal is a pulse train, with each pulse in the train corresponding to the LOS component or a multipath component associated with a distinct reflector. The *delay spread* of a multipath channel characterizes the difference in arrival times of the first received signal component (LOS or, if blocked, the first received multipath component) and the last received signal component of significant energy associated with a single transmitted pulse. For multi-ray reflection models such as the 10-ray and 6-ray models of Chapter 2.8.1, the delay spread equals the delay difference between the first and last ray. The delay spread for random channels is characterized statistically, as described in 3.3.2. 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, if the delay spread is large then there is significant time spreading of the received signal, which leads to substantial signal distortion.

Another characteristic of the multipath channel is its time-varying nature. This time variation arises because either the transmitter, receiver, or objects in the environment are moving. Hence the location of reflectors in the transmission path, which gives rise to multipath, will change over time. Thus, if we repeatedly transmit pulses from a transmitter given these changing reflector locations, we will observe changes in the amplitudes, delays, and 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 scatterer locations. We will first describe the time-varying channel impulse response that models these changing multipath components. We 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 featuring a fixed number of multipath components, each with fixed path loss and shadowing. For this 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 MIMO channel models are also discussed.

3.1 Time-Varying Channel Impulse Response

Let the transmitted signal be as in (2.1):

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

where $u(t)$ is the equivalent lowpass signal for $s(t)$ with bandwidth B_u and f_c is its carrier frequency. Neglecting noise, the corresponding received signal is the sum of the LOS path and all *resolvable* multipath components:

$$r(t) = \mathrm{Re}\left\{\sum_{i=0}^{N(t)-1} \alpha_i(t)u(t-\tau_i(t))e^{j(2\pi f_c(t-\tau_i(t))+\phi_{D_i}(t))}\right\}, \tag{3.2}$$

where the first term in the summation, corresponding to $i = 0$, is associated with the LOS path. The unknowns in this expression are the number of *resolvable* signal components $N(t)$ and, for each signal component (LOS and multipath), its path length $x_i(t)$ and corresponding delay $\tau_i(t) = x_i(t)/c$, Doppler phase shift $\phi_{D_i}(t)$, and amplitude $\alpha_i(t)$. The phase change associated with user movement and the resolvability of different signal components will be discussed in the next two paragraphs. If at time t the LOS path is blocked or significantly attenuated then this will be captured in the amplitude term $\alpha_0(t)$. Similarly, the amplitude $\alpha_i(t)$ of the ith multipath component is determined by the path loss and shadowing along its propagation path. In addition, the amplitudes $\alpha_i(t), i = 0, \ldots, N(t)-1$ may experience multipath fading due to constructive and destructive addition of nonresolvable multipath components.

When the transmitter or receiver is moving, the delay $\tau_i(t)$ in (3.2) changes over time and hence so does the associated phase change. This change is captured in (3.2) as $e^{-j2\pi f_c \tau_i(t)}$, however $\tau_i(t)$ is rarely precisely known in practice. Alternatively the phase change due to small changes in $\tau_i(t)$ can be captured by the signal's Doppler shift $f_{D_i}(t) = v(t)\cos\theta_i(t)/\lambda$ for $\theta_i(t)$ the component's angle of arrival (AoA) relative to the direction of motion at velocity $v(t)$. This Doppler frequency shift, which is generally easy to measure in practice, leads to a Doppler phase shift of $\phi_{D_i}(t) = \int_0^t 2\pi f_{D_i}(\nu)d\nu$. Note that the phase change of the ith multipath component due to user movement in (3.2) should either be captured in $e^{-j2\pi f_c\tau_i(t)}$ or in $\phi_{D_i}(t)$, but not in both terms as then the phase shift due to user movement is double-counted. It is generally assumed that $\tau_i(t) = x_i(t)/c$ does not capture changes in $x_i(t)$ of less than the signal wavelength λ, so the phase changes due to these small user movements are captured by the Doppler term in (3.2).

We now define resolvability of multipath components. We say that two multipath components with delay τ_i and τ_j are *resolvable* if their delay difference significantly exceeds the inverse signal bandwidth: $|\tau_i - \tau_j| \gg B_u^{-1}$. Multipath components that do not satisfy this resolvability criteria cannot be separated out at the receiver because $u(t-\tau_i) \approx u(t-\tau_j)$, and thus these components are *nonresolvable*. These nonresolvable multipath components are summed together to create a single resolvable multipath component in (3.2) corresponding to delay τ_i, with a resulting amplitude and phase corresponding to the summed signal. Any multipath components with delay τ_j for which $|\tau_j - \tau_0| < B_u^{-1}$ are combined with the LOS signal. The amplitudes of the LOS and resolvable multipath components in (3.2) that include nonresolvable multipath components will typically undergo rapid multipath fading due to the constructive and destructive combining of the different signal components. For wideband channels (B_u relatively large) each term in the summation of (3.2) typically corresponds to a single (smooth or non-smooth[1]) reflector, or to a cluster of closely-spaced reflectors, as shown in Figure 3.1. By contrast, narrowband channels (B_u relatively small) are likely to have one or more of the signal components in (3.2) correspond to a summation of nonresolvable signal components.

[1]A non-smooth reflector has a surface that is sufficiently rough such that it generates different multipath components with slightly different delays.

Since the parameters $\alpha_i(t), \tau_i(t)$, and $\phi_{D_i}(t)$ associated with each resolvable multipath component change over time, they are characterized as random processes that 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 (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 – owing to constructive and destructive addition of the different components.

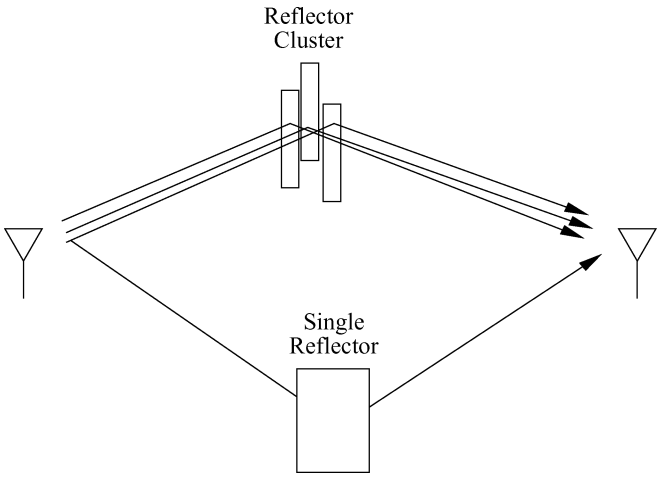

Figure 3.1: A single reflector and a cluster of closely-spaced reflectors. The reflector cluster generates multipath components that are typically resolvable in wideband channels and non-resolvable in narrowband channels.

We can simplify $r(t)$ by letting

$$\phi_i(t) = 2\pi f_c \tau_i(t) - \phi_{D_i}(t). \tag{3.3}$$

Then the received signal can be rewritten as

$$r(t) = \mathrm{Re}\left\{ \left[\sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} u(t-\tau_i(t)) \right] e^{j2\pi f_c t} \right\}. \tag{3.4}$$

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

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

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

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 $\tau_i(t) = \tau$, then $c(\tau, t) = 0$. Although the definition of the time-varying channel impulse response might at first seem counterintuitive, $c(\tau, t)$ must be defined in this way

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

to be consistent with the special case of time-invariant channels. Specifically, for time-invariant channels we have $c(\tau, t) = c(\tau, t+T)$ for any T; that is, 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 time zero.[3]

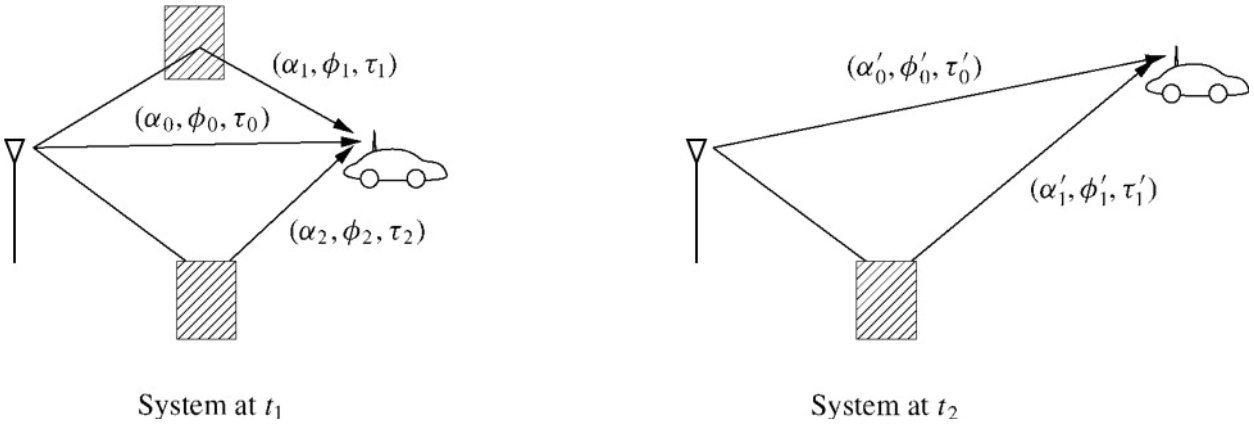

Figure 3.2: System multipath at two different measurement times.

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

$$c(\tau, t) = \sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} \delta(\tau - \tau_i(t)). \tag{3.6}$$

Substituting (3.6) back into (3.5) yields (3.4), thereby confirming that (3.6) is the channel's equivalent lowpass time-varying impulse response:

$$\begin{aligned} r(t) &= \mathrm{Re}\left\{ \left[\int_{-\infty}^{\infty} c(\tau, t) u(t-\tau) d\tau \right] e^{j2\pi f_c t} \right\} \\ &= \mathrm{Re}\left\{ \left[\int_{-\infty}^{\infty} \sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} \delta(\tau - \tau_i(t)) u(t-\tau) d\tau \right] e^{j2\pi f_c t} \right\} \\ &= \mathrm{Re}\left\{ \left[\sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} \left(\int_{-\infty}^{\infty} \delta(\tau - \tau_i(t)) u(t-\tau) d\tau \right) \right] e^{j2\pi f_c t} \right\} \\ &= \mathrm{Re}\left\{ \left[\sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} u(t - \tau_i(t)) \right] e^{j2\pi f_c t} \right\}, \end{aligned} \tag{3.7}$$

where the last equality follows from the sifting property of delta functions:

$$\int_{-\infty}^{\infty} \delta(\tau - \tau_i(t)) u(t-\tau) d\tau = \delta(t - \tau_i(t)) * u(t) = u(t - \tau_i(t)). \tag{3.8}$$

[3]By definition, $c(\tau, 0)$ is the response at time zero to an impulse at time $-\tau$, but since the channel is time invariant, this equals the response at time τ to an impulse at time zero.

Some channel models assume a continuum of multipath delays, in which case the sum in (3.6) becomes an integral that simplifies to a time-varying complex amplitude associated with each multipath delay τ:

$$c(\tau, t) = \int_{-\infty}^{\infty} \alpha(\xi, t) e^{-j\phi(\xi,t)} \delta(\tau - \xi) d\xi = \alpha(\tau, t) e^{-j\phi(\tau,t)}. \tag{3.9}$$

For 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, each with amplitude, phase, and delay triple $(\alpha_i, \phi_i, \tau_i), i = 0, 1, 2$. Thus, impulses that were launched into the channel at time $t_1 - \tau_i, i = 0, 1, 2$, 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

$$c(\tau, t_1) = \sum_{i=0}^{2} \alpha_i e^{-j\phi_i} \delta(\tau - \tau_i), \tag{3.10}$$

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 having amplitude, phase, and delay triple $(\alpha_i', \phi_i', \tau_i'), i = 0, 1$. Thus, impulses that were launched into the channel at time $t_2 - \tau_i', i = 0, 1$, 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

$$c(\tau, t_2) = \sum_{i=0}^{1} \alpha_i' e^{-j\phi_i'} \delta(\tau - \tau_i') \tag{3.11}$$

and is also shown in Figure 3.3.

If the channel is time invariant, then its response to an impulse at time t_1 is just a shifted version of its response to an impulse at time $t_2 \neq t_1$. Hence, the time-varying parameters in $c(\tau, t)$ become constant and $c(\tau, t) = c(\tau)$ is just a function of τ:

$$c(\tau) = \sum_{i=0}^{N} \alpha_i e^{-j\phi_i} \delta(\tau - \tau_i). \tag{3.12}$$

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 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 simply corresponds to the LOS path. For $t = nT_0$, the channel impulse response includes 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 ith multipath component will have $f_c \tau_i(t) \gg 1$. For example, with $f_c = 1$ GHz and $\tau_i = 50$ ns (a typical value for an indoor system), $f_c \tau_i = 50 \gg 1$. Outdoor wireless systems

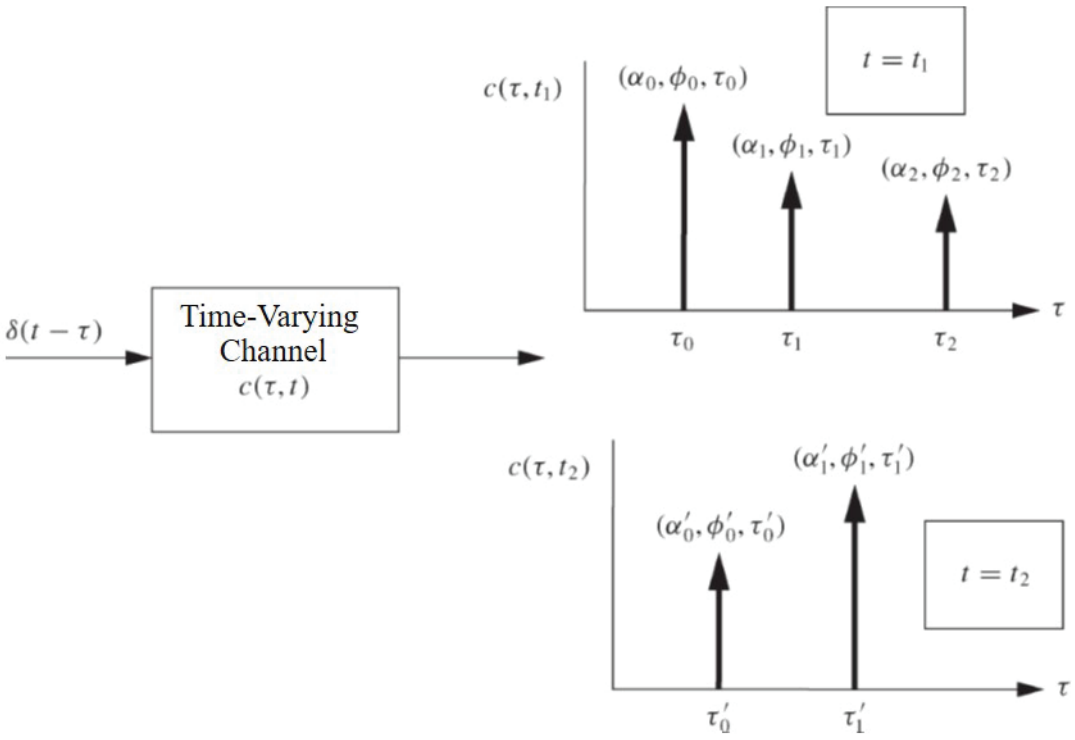

Figure 3.3: Response of time-varying channel.

have multipath delays much greater than 50 ns, so this property also holds for these systems. If $f_c\tau_i(t) \gg 1$ then a small change in the path delay $\tau_i(t)$ can lead to a large phase change in the ith multipath component with phase $\phi_i(t) = 2\pi f_c\tau_i(t) - \phi_{D_i}(t) - \phi_0$. Rapid phase changes in each multipath component give rise to constructive and destructive addition of the multipath components constituting the received signal, which in turn causes rapid variation in the received signal strength. This multipath fading phenomenon was exhibited in the deterministic channel models of Chapter 2 (see, e.g., Figure 2.4 illustrating the received signal power for the two-ray multipath model). This phenomenon for statistical multipath models 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 nonresolvable, leading to the narrowband fading model described in the next section. If the delay spread is large then some number of the LOS and multipath components are resolvable, leading to the wideband fading model of Section 3.3. 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 (3.12), 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 $T_m = \max_i[\tau_i - \tau_0]$. However, if the demodulator synchronizes to a multipath component with delay equal to the mean delay $\bar{\tau}$, then the delay spread is given by $T_m = \max_i |\tau_i - \bar{\tau}|$. 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 3.3.2. Specifically, two common

characterizations of channel delay spread – average delay spread and rms (root mean square) 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 [1, Chap. 5.4.1; 2, Chap. 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. We will adopt the rms delay spread for our characterization as it is the one most widely used in the literature. 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 urban microcells it ranges from 1–10 microseconds, and in rural macrocells is ranges from 10–100 microseconds [1, Chap. 5].

3.2 Narrowband Fading Model

Suppose the delay spread T_m of a channel is small relative to the inverse baseband signal bandwidth B_u of the transmitted signal; that is, suppose $T_m \ll B_u^{-1}$. We refer to the fading in this scenario as *narrowband fading* which is a function of both the signal bandwidth B_u and the channel delay spread T_m. As discussed previously, the delay spread T_m for time-varying channels is usually characterized by the rms delay spread, but it can also be characterized in other ways. Under most delay spread characterizations, $T_m \ll B_u^{-1}$ implies that the delay associated with the ith multipath component $\tau_i \leq T_m$ for all i, so $u(t-\tau_i) \approx u(t)$ for all i. Since in the narrowband fading model the delay associated with all multipath components results in minimal signal distortion in time, we say the channel under this model is *nondispersive*. We can rewrite (3.4) as

$$r(t) = \mathrm{Re}\left\{ u(t)e^{j2\pi f_c t} \left(\sum_i \alpha_i(t) e^{-j\phi_i(t)} \right) \right\}. \tag{3.13}$$

Equation (3.13) differs from the original transmitted signal by the complex scale factor in large parentheses. This scale factor is independent of the transmitted signal $s(t)$ and, in particular, of the equivalent lowpass signal $u(t)$ as long as the narrowband assumption $T_m \ll 1/B_u$ 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:

$$s(t) = \mathrm{Re}\{e^{j(2\pi f_c t + \phi_0)}\} = \cos(2\pi f_c t + \phi_0), \tag{3.14}$$

which is narrowband for any T_m.

With this assumption the received signal becomes

$$r(t) = \mathrm{Re}\left\{ \left[\sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} \right] e^{j2\pi f_c t} \right\} = r_I(t)\cos(2\pi f_c t) - r_Q(t)\sin(2\pi f_c t), \tag{3.15}$$

where the in-phase and quadrature components are given by

$$r_I(t) = \sum_{i=0}^{N(t)-1} \alpha_i(t)\cos\phi_i(t), \tag{3.16}$$

$$r_Q(t) = \sum_{i=0}^{N(t)-1} \alpha_i(t)\sin\phi_i(t) \tag{3.17}$$

and where the phase term

$$\phi_i(t) = 2\pi f_c \tau_i(t) - \phi_{D_i}(t) - \phi_0 \tag{3.18}$$

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

Suppose $N(t)$ is large and all signal (multipath as well as the LOS) components have random amplitudes with approximately the same distribution, and similarly for the random phases of each component. Then, given that $\alpha_i(t)$ and $\phi_i(t)$ are also independent for different components, we can invoke the Central Limit Theorem (CLT) for sums of independent identically-distributed (i.i.d.) random variables to approximate $r_I(t)$ and $r_Q(t)$ as jointly Gaussian random processes. The Gaussian property also holds for small $N(t)$ if the $\alpha_i(t)$ are Rayleigh distributed (given in (3.33) below) and the $\phi_i(t)$ are uniformly distributed on $[-\pi, \pi]$. This happens, again by the CLT, when the ith multipath component results from a reflection cluster with a large number of nonresolvable multipath components [3]. If the LoS or any of the multipath components has a much larger amplitude than the average multipath component amplitude, then this component dominates the received signal; hence the Gaussian approximation is no longer applicable as the terms in (3.15) are not i.i.d.

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 $r_Q(t)$, as well as the received signal $r(t)$. We focus on these correlations over a time period where the transmitter or receiver moves on the order of a few signal wavelengths. As we will see, multipath can cause the received signal to completely decorrelate over a signal wavelength due to changes in phase of the individual multipath components such that they combine either constructively or destructively. Our derivations of these correlations are based on some key assumptions that generally apply to propagation models without a dominant LOS component such that all nonzero terms in (3.15) are approximately i.i.d. We will also determine the power spectral density (PSD) of the received signal under certain assumptions that make its autocorrelation function time-invariant. In this case the PSD is just the Fourier transform of the autocorrelation function.

A transmitter or receiver moving at velocity v tranverses a distance equal to a signal wavelenth λ over time $\tau = \lambda/v$. We will assume in this section that the LoS and multipath amplitudes $\alpha_i(t)$, delays $\tau_i(t) = d_i(t)/c$, and Doppler frequencies $f_{D_i}(t) = v(t)\cos\theta_i(t)/\lambda$ are changing slowly enough to be considered constant over the time interval τ for which we compute the received signal autocorrelation and cross-correlation. This is generally true for $\tau \approx \lambda/v$ since over time τ or, equivalently, over distance λ the number of signal components $N(t) = N$ doesn't change. Moreover, since the shadowing associated with $\alpha_i(t)$ decorrelates over distance X_c which is generally much bigger than λ, $\alpha_i(t) \approx \alpha_i$ for some constant α_i. Similarly, the distance a given signal component travels, the component's AoA, and the velocity of the transmitter or receiver change little over the time interval τ, and hence we can assume $\tau_i(t) \approx \tau_i$ for some average delay τ_i and $f_{D_i}(t) \approx f_{D_i}$ for some average Doppler frequency f_{D_i}. With a constant Doppler frequency, the Doppler phase shift[4] is $\phi_{D_i}(t) = \int_0^t 2\pi f_{D_i} d\nu = 2\pi f_{D_i} t$ and the phase of the ith multipath component becomes $\phi_i(t) = 2\pi f_c \tau_i - 2\pi f_{D_i} t - \phi_0$.

We now make a key assumption; we assume that for the ith multipath component, the term $2\pi f_c \tau_i - 2\pi f_{D_i} t$ in $\phi_i(t)$ changes rapidly relative to all other phase terms in the expression. This is a reasonable assumption because f_c is large and hence the term $2\pi f_c \tau_i(t)$ can go through a 360° rotation for a small change in multipath delay $\tau_i(t)$. As discussed earlier, this change is typically captured in the Doppler term $2\pi f_{D_i} t$ for $\tau_i(t)$ assumed constant. Under this assumption, $\phi_i(t)$ is uniformly distributed on $[-\pi, \pi]$. Thus

$$\mathbf{E}[r_I(t)] = \mathbf{E}\left[\sum_i \alpha_i \cos\phi_i(t)\right] = \sum_i \mathbf{E}[\alpha_i]\mathbf{E}[\cos\phi_i(t)] = 0, \tag{3.19}$$

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

where the second equality follows from the independence of α_i and $\phi_i(t)$ and the last equality follows from the uniform distribution on $\phi_i(t)$. Similarly we can show that $\mathbf{E}[r_Q(t)] = 0$. Thus, the received signal also has $\mathbf{E}[r(t)] = 0$: it is a zero-mean Gaussian process. If there is a dominant LOS component in the channel then 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 $A_{r_I}(t, t+\tau) = \mathbf{E}[r_I(t)r_I(t+\tau)]$ of the in-phase component. Using the independence of α_i and $\phi_i(t)$, the independence of $\phi_i(t)$ and $\phi_j(t)$ $(i \neq j)$, and the uniform distribution of $\phi_i(t)$, we get that

$$\begin{aligned} A_{r_I}(t, t+\tau) &= \mathbf{E}\left[\sum_i \alpha_i \cos\phi_i(t) \sum_j \alpha_j \cos\phi_j(t+\tau)\right] \\ &= \sum_i \sum_j \mathbf{E}[\alpha_i\alpha_j]\mathbf{E}[\cos\phi_i(t)\cos\phi_j(t+\tau)] \\ &= \sum_i \mathbf{E}[\alpha_i^2]\mathbf{E}[\cos\phi_i(t)\cos\phi_i(t+\tau)]. \end{aligned} \tag{3.20}$$

Since $\phi_i(t) = 2\pi f_c\tau_i - 2\pi f_{D_i}t - \phi_0$ and $\phi_i(t+\tau) = 2\pi f_c\tau_i - 2\pi f_{D_i}(t+\tau) - \phi_0$ we have

$$\mathbf{E}[\cos\phi_i(t)\cos\phi_i(t+\tau)] = .5\mathbf{E}[\cos 2\pi f_{D_i}\tau] + .5\mathbf{E}[\cos(4\pi f_c\tau_i - 4\pi f_{D_i}t - 2\pi f_{D_i}\tau - 2\phi_0)]. \tag{3.21}$$

Because $2\pi f_c\tau_i$ changes rapidly relative to all other phase terms and is uniformly distributed modulo 2π, the second expectation term in (3.21) is zero and thus

$$A_{r_I}(t, t+\tau) = .5\sum_i \mathbf{E}[\alpha_i^2]\mathbf{E}[\cos(2\pi f_{D_i}\tau)] = .5\sum_i \mathbf{E}[\alpha_i^2]\mathbf{E}_{\theta_i}\left[\cos\left(\frac{2\pi v\tau}{\lambda}\cos\theta_i\right)\right], \tag{3.22}$$

since $f_{D_i} = v\cos(\theta_i)/\lambda$. The last expectation in (3.22) is with respect to the angle of arrival θ_i since that is the only random parameter in the expression over which the expectation is taken. Observe that $A_{r_I}(t, t+\tau)$ depends only on τ and hence $r_I(t)$ is a wide-sense stationary (WSS) random process. Thus we can write $A_{r_I}(t, t+\tau) = A_{r_I}(\tau)$ as a function of τ only. Moreover, using a similar derivation as (3.22) we can show that the quadrature component is also WSS with the same autocorrelation as $A_{r_I}(\tau)$, so $A_{r_I}(\tau) = A_{r_Q}(\tau)$.

We now consider the cross-correlation between the in-phase and quadrature components. Using the same trigonometric properties as in the derivation of (3.22) and the fact that $2\pi f_c\tau_i$ is uniformly distributed we get

$$A_{r_I,r_Q}(t, t+\tau) = \mathbf{E}[r_I(t)r_Q(t+\tau)] = -.5\sum_i \mathbf{E}[\alpha_i^2]\mathbf{E}_{\theta_i}\left[\sin\left(\frac{2\pi v\tau}{\lambda}\cos\theta_i\right)\right] \tag{3.23}$$

and

$$A_{r_Q,r_I}(t, t+\tau) = \mathbf{E}[r_Q(t)r_I(t+\tau)] = .5\sum_i \mathbf{E}[\alpha_i^2]\mathbf{E}_{\theta_i}\left[\sin\left(\frac{2\pi v\tau}{\lambda}\cos\theta_i\right)\right]. \tag{3.24}$$

We see from (3.23) and (3.24) that these cross-corrections only depend on τ so we can write these as $A_{r_I,r_Q}(t, t+\tau) = A_{r_I,r_Q}(\tau)$ and $A_{r_Q,r_I}(t, t+\tau) = A_{r_Q,r_I}(\tau)$. Moreover, we see from these equations that $A_{r_I,r_Q}(\tau) = -A_{r_Q,r_I}(\tau)$. Finally, setting $\tau = 0$ we see that $A_{r_I,r_Q}(0) = -A_{r_Q,r_I}(0) = 0$. Thus, $r_I(t)$ and $r_Q(t)$ are uncorrelated and, since they are jointly Gaussian processes, this means they are independent.

We now consider the autocorrelation of the received signal

$$r(t) = r_I(t)\cos(2\pi f_c t) - r_Q(t)\sin(2\pi f_c t).$$

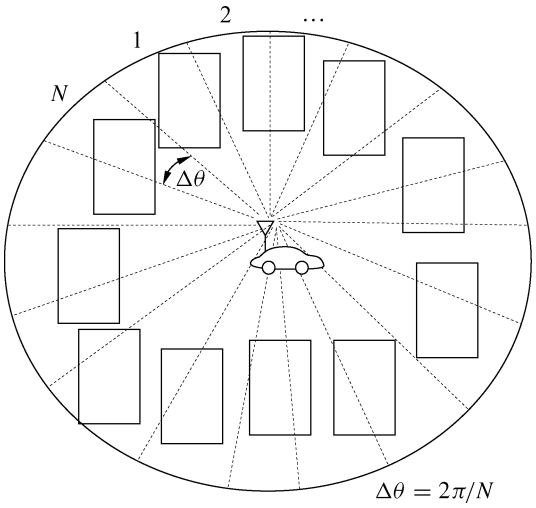

Figure 3.4: Dense scattering environment.

We have

$$\begin{aligned}
A_{r(t,t+\tau)} &= \mathbf{E}[r(t)r(t+\tau)] \\
&= \mathbf{E}[(r_I(t)\cos(2\pi f_c t) - r_Q(t)\sin(2\pi f_c t))(r_I(t+\tau)\cos(2\pi f_c(t+\tau)) - r_Q(t+\tau)\sin(2\pi f_c(t+\tau))] \\
&= \mathbf{E}[r_I(t)r_I(t+\tau)\cos(2\pi f_c t)\cos(2\pi f_c(t+\tau)) - r_Q(t)r_I(t+\tau)\sin(2\pi f_c t)\cos(2\pi f_c(t+\tau)) \\
&\quad - r_I(t)r_Q(t+\tau))\cos(2\pi f_c t)\sin(2\pi f_c(t+\tau)) + r_Q(t)r_Q(t+\tau)\sin(2\pi f_c t)\sin(2\pi f_c(t+\tau))] \\
&= A_{r_I}(\tau)\cos(2\pi f_c t)\cos(2\pi f_c(t+\tau)) + A_{r_Q}(\tau)\sin(2\pi f_c t)\sin(2\pi f_c(t+\tau)) \\
&\quad - A_{r_Q,r_I}(\tau)\sin(2\pi f_c t)\cos(2\pi f_c(t+\tau)) - A_{r_I,r_Q}(\tau)\cos(2\pi f_c t)\sin(2\pi f_c(t+\tau)) \\
&= A_{r_I}(\tau)\cos(2\pi f_c \tau) + A_{r_I,r_Q}(\tau)\sin(2\pi f_c \tau), \qquad (3.25)
\end{aligned}$$

where the last equality follows from trigonometric product formulas and that fact that $A_{r_I}(\tau) = A_{r_Q}(\tau)$ and $A_{r_I,r_Q}(\tau) = -A_{r_Q,r_I}(\tau)$. Since (3.25) is only a function of the time difference τ, the received signal $r(t)$ is WSS along with its in-phase and quadradure components.

In order to further simplify (3.22) and (3.23), we must make an additional assumption about the propagation environment with respect to the power angle spectrum (PAS), which characterizes the received signal power as a function of its angle of arrival. We will focus on the *uniform scattering environment* introduced by Clarke [4] and further developed by Jakes [5, Chap. 1]. In this model, the channel consists of many scatterers densely packed with respect to their angles of arrival, as shown in Figure 3.4. Note that this is a two-dimensional model, so that the elevation angle of arrival for the signal components is ignored. In the uniform scattering model we assume N (LoS and multipath) signal components with angles of arrival $\theta_i = i\triangle\theta$, $i = 0, \ldots, N-1$, where $\triangle\theta = 2\pi/N$. We also assume that each signal component has the same received power and so $\mathbf{E}[\alpha_i^2] = 2P_r/N$, where P_r is the total received power. Then (3.22) becomes

$$A_{r_I}(\tau) = \frac{P_r}{N}\sum_{i=0}^{N-1}\cos\left(\frac{2\pi v\tau}{\lambda}\cos\left(i\triangle\theta\right)\right). \qquad (3.26)$$

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

$$A_{r_I}(\tau) = \frac{P_r}{2\pi} \sum_{i=0}^{N-1} \cos\left(\frac{2\pi v\tau}{\lambda} \cos(i\triangle\theta)\right) \triangle\theta. \tag{3.27}$$

We now take the limit as the number of scatterers grows to infinity, which corresponds to uniform scattering from all directions. Then $N \to \infty$, $\triangle\theta \to 0$, and the summation in (3.27) becomes an integral:

$$A_{r_I}(\tau) = \frac{P_r}{2\pi} \int_0^{2\pi} \cos\left(\frac{2\pi v\tau}{\lambda} \cos\theta\right) d\theta = P_r J_0(2\pi f_D \tau), \tag{3.28}$$

where $f_D = 2\pi v/\lambda$ and

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

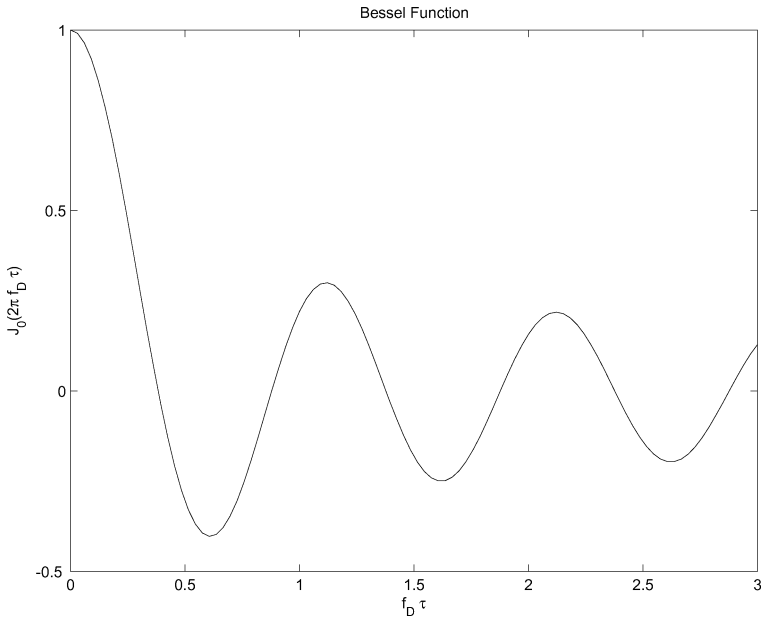

Figure 3.5: Bessel function versus $f_D\tau$.

is the Bessel function of zeroth order.[5] Similarly, for this uniform scattering environment,

$$A_{r_I,r_Q}(\tau) = \frac{P_r}{2\pi} \int \sin\left(\frac{2\pi v\tau}{\lambda} \cos(\theta)\right) d\theta = 0. \tag{3.29}$$

A plot of $J_0(2\pi f_D\tau)$ is shown in Figure 3.5. There are several interesting observations to make 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_i 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 independent fading paths obtained from multiple antennas can be combined to remove some of the negative effects of fading. For independent fading the antenna spacing must be $.4\lambda$ based on the analysis here.

[5]Equation (3.28) can also be derived by assuming that $2\pi v\tau \cos\theta_i/\lambda$ in (3.22) and (3.23) is random with θ_i uniformly distributed, and then taking expectations with respect to θ_i. However, based on the underlying physical model, θ_i can be uniformly distributed only in a dense scattering environment. So the derivations are equivalent.

However, combining paths that have a low correlation leads to almost the same gains as combining independent fading paths [6, Chap. 9.6.5]. 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 $.4\lambda$. Because of this recorrelation property, a Markov model is not completely accurate for a uniform scattering environment. However, since the recorrelation once the separation distance is greater than a half-wavelength is below .3, this is a reasonable approximation in analyses where such a low correlation has little impact on performance.

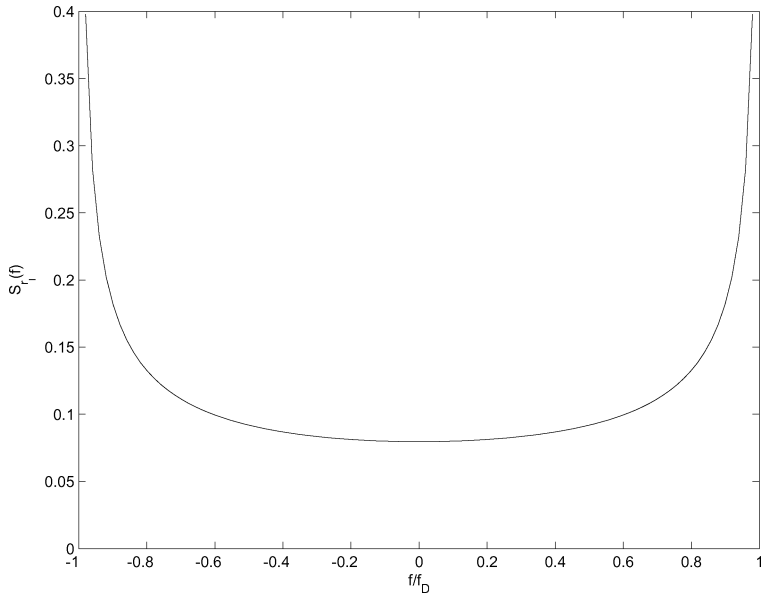

Figure 3.6: In-phase and quadrature PSD: $S_{r_I}(f) = S_{r_Q}(f)$.

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

$$S_{r_I}(f) = S_{r_Q}(f) = \mathcal{F}[A_{r_I}(\tau)] = \begin{cases} \frac{2P_r}{\pi f_D} \frac{1}{\sqrt{1-(f/f_D)^2}}, & |f| \le f_D, \\ 0, & \text{otherwise.} \end{cases} \tag{3.30}$$

This PSD is shown in Figure 3.6.

To obtain the PSD of the received signal $r(t)$ under uniform scattering we use (3.25) with $A_{r_I,r_Q}(\tau) = 0$, (3.30), and properties of baseband random processes modulated by a sinusoid (discussed in Appendix B.3) to obtain

$$\begin{aligned} S_r(f) &= \mathcal{F}[A_r(\tau)] = .25[S_{r_I}(f - f_c) + S_{r_I}(f + f_c)] \\ &= \begin{cases} \frac{P_r}{2\pi f_D} \frac{1}{\sqrt{1-(|f-f_c|/f_D)^2}}, & |f - f_c| \le f_D, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3.31}$$

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 probability density function (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 $f_D(\theta)$.

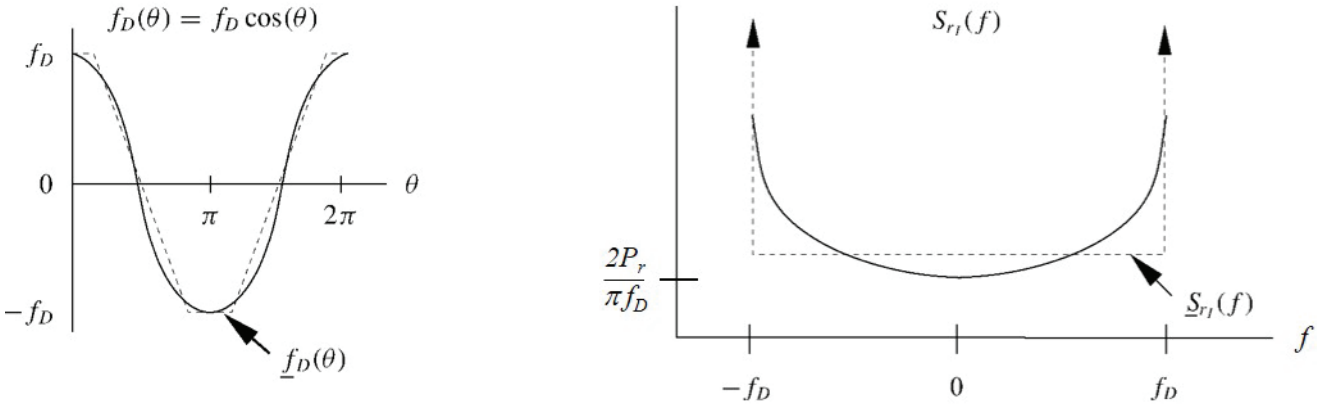

Figure 3.7: Cosine and PSD approximation by straight line segments.

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 pdf $p_{f_D(\theta)}(f)$ of the random Doppler frequency $f_D(\theta)$ can then be obtained from the pdf of θ. By definition, $p_{f_D(\theta)}(f)$ 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 $p_{f_D(\theta)}(f)$. For this characterization, in Figure 3.7 we plot $f_D(\theta) = f_D \cos(\theta) = \frac{v}{\lambda}\cos(\theta)$ along with a (dashed) 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 dashed 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 PSD 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. This explains the shape of $S_{r_I}(f)$ under uniform scattering. Formulas for the autocorrelation and PSD in nonuniform scattering – corresponding to more typical microcell and indoor environments – can be found in [5, Chap. 1; 7, Chap. 2]. Three-dimensional channel models that incorporate elevation angle are described in [8] and the references therein.

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 a frequency response $H(f)$ that satisfies

$$S_{r_I}(f) = S_{r_Q}(f) = \frac{N_0}{2}|H(f)|^2. \tag{3.32}$$

The filter outputs then correspond, respectively, 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. the Matlab communications toolbox) 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 [1, 7, 9, 10].

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 single-slope path loss and log-normal shadowing models developed in Chapter 2. In this figure we see the signal power due to path loss decreasing at a slope of -10γ relative to $\log_{10}(d/d_r)$ for γ the path-loss exponent and d_r the reference distance at which the path loss equals the constant K. The more rapid variations due to shadowing change on the order of the decorrelation distance X_c, and the very rapid variations due to multipath fading change on the order of half the signal wavelength. If we blow up a small segment of this figure over distances where path loss and shadowing are constant we obtain Figure 3.9, which plots the dB value of P_r/P_t versus linear distance $d = vt$ (not log distance). In this figure the average value of P_r/P_t is normalized to 0 dB. A mobile receiver traveling at fixed velocity v would experience stationary and ergodic received power variations over time as illustrated in this figure.

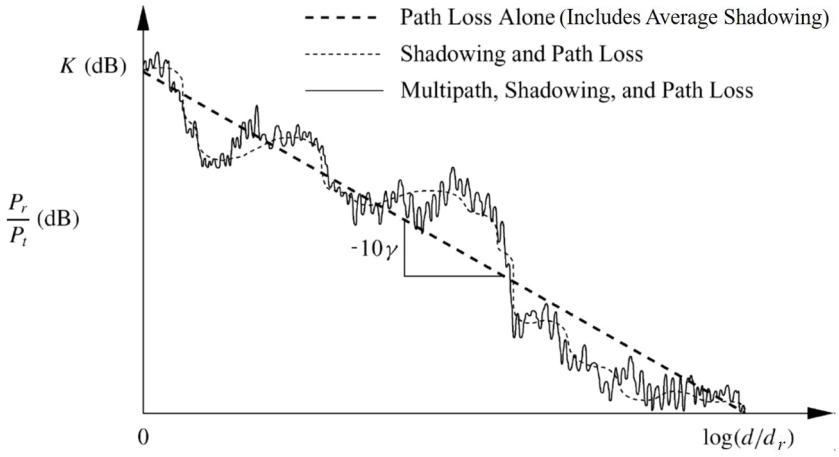

Figure 3.8: Combined single-slope path loss, log-normal shadowing, and narrowband fading.

3.2.2 Envelope and Power Distributions

We now consider the distribution of the envelope and power for the narrowband received signal $r(t) = r_I(t)\cos(2\pi f_c t) - r_Q(t)\sin(2\pi f_c t)$. It can be shown that, for any two Gaussian random variables X and Y, both with mean zero and equal variance σ^2, $Z = \sqrt{X^2 + Y^2}$ is Rayleigh distributed and Z^2 is exponentially distributed. We have seen that, for $\phi_i(t)$ uniformly distributed, r_I and r_Q are both zero-mean Gaussian random variables. If we assume a variance of $\sigma^2 = .5\sum_i \mathbf{E}[\alpha_i^2]$ for both in-phase and quadrature components, then the signal envelope $z(t) = |r(t)| = \sqrt{r_I^2(t) + r_Q^2(t)}$ is Rayleigh distributed with distribution

$$p_Z(z) = \frac{2z}{P_r}\exp\left[-\frac{z^2}{P_r}\right] = \frac{z}{\sigma^2}\exp\left[-\frac{z^2}{2\sigma^2}\right], \quad z \geq 0, \tag{3.33}$$

where $P_r = 2\sigma^2$ is the power of $z(t)$ which equals the power of $r(t)$.

We obtain the power distribution by making the change of variables $z^2(t) = |r(t)|^2$ in (3.33) to obtain

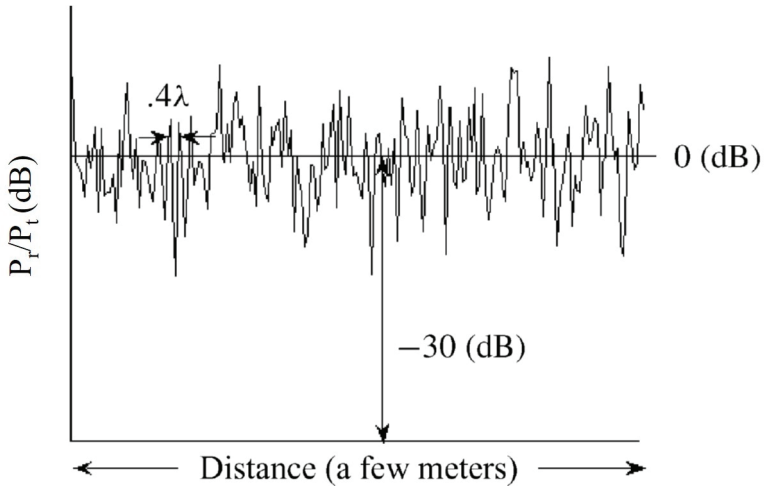

Figure 3.9: Narrowband fading.

$$p_{Z^2}(x) = \frac{1}{P_r} e^{-x/P_r} = \frac{1}{2\sigma^2} e^{-x/2\sigma^2}, \quad x \geq 0, \tag{3.34}$$

i.e. the power in $r(t)$ is exponentially distributed with mean $P_r = 2\sigma^2$. Thus, $r(t)$ has a Rayleigh-distributed amplitude and exponentially-distributed power with mean $P_r = 2\sigma^2$. The equivalent lowpass signal for $r(t)$ is given by $r_{LP}(t) = r_I(t) + j r_Q(t)$, which has the same power as $z(t)$ and 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 $|r_{LP}|$.

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 $P_r = 20\ \mathrm{dBm} = 100\ \mathrm{mW}$. We want to find the probability that $Z^2 < 10\ \mathrm{dBm} = 10\ \mathrm{mW}$. Thus

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

If the channel has a LOS component with a much larger signal power than the other multipath components, then $r_I(t)$ and $r_Q(t)$ are not zero-mean random processes. That is because the average signal power is dominated by the LOS component, with small fluctuations about this average due to constructive and destructive addition of the multipath components. In this scenario the received signal equals the superposition of a complex Gaussian component and a higher-power LOS component. The signal envelope in this case can be shown to have a Rician distribution [11] given by

$$p_Z(z) = \frac{z}{\sigma^2} \exp\left[\frac{-(z^2+s^2)}{2\sigma^2}\right] I_0\left(\frac{zs}{\sigma^2}\right), \quad z \geq 0, \tag{3.35}$$

where $2\sigma^2 = \sum_{i,i\neq 0} \mathbf{E}[\alpha_i^2]$ 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 zeroth order. The average received power in the Rician fading is given by

$$P_r = \int_0^\infty z^2 p_Z(z) dz = s^2 + 2\sigma^2. \tag{3.36}$$

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

$$K = \frac{s^2}{2\sigma^2}. \tag{3.37}$$

Thus, K is the ratio of the power in the LOS component to the power in the other (non-LOS) multipath components, which is typically a random variable. 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 relatively mild fading. Making the substitutions $s^2 = KP_r/(K+1)$ and $2\sigma^2 = P_r/(K+1)$, we can write the Rician distribution in terms of K and P_r as

$$p_Z(z) = \frac{2z(K+1)}{P_r} \exp\left[-K - \frac{(K+1)z^2}{P_r}\right] I_0\left(2z\sqrt{\frac{K(K+1)}{P_r}}\right), \quad z \geq 0. \tag{3.38}$$

Both the Rayleigh and Rician distributions for the received signal envelope can be obtained by using mathematics to capture the underlying physical properties of the channel models [3, 11]. However, some experimental data does not fit well into either of these distributions. Thus, a more general fading distribution for the received signal envelope 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

$$p_Z(z) = \frac{2m^m z^{2m-1}}{\Gamma(m) P_r^m} \exp\left[\frac{-mz^2}{P_r}\right], \quad m \geq .5, \tag{3.39}$$

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 (3.39) reduces to Rayleigh fading. For $m = (K+1)^2/(2K+1)$ the distribution in (3.39) is approximately Rician fading with parameter K. For $m = \infty$ there is no fading: $Z = \sqrt{P_r}$ is a constant. Thus, the Nakagami distribution can model both Rayleigh and Rician distributions as well as more general ones. Note that some empirical measurements support values of the m-parameter less than unity, 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

$$p_{Z^2}(x) = \left(\frac{m}{P_r}\right)^m \frac{x^{m-1}}{\Gamma(m)} \exp\left[\frac{-mx}{P_r}\right]. \tag{3.40}$$

While narrowband fading is often modeled using the Rayleigh, Rician, or Nakagami distribution, these distributions are not accurate for many environments that arise in practice. This has led to several important extensions of these models [6]. In particular, the Hoyt (or Nakagami-q) distribution models narrowband fading of the received signal envelope assuming instaneous values of the in-phase and quadrature components are zero-mean Gaussian random variables with unequal variances. The Beckman distribution generalizes the Hoyt model such that the instantaneous in-phase and quadrature components are Gaussian random variables with unequal means and unequal variances. The Weibull fading model generalizes both Rayleigh and exponential distributions, and with proper parameterization has been shown to fit with with experimental data. The pdf formulas for the Hoyt, Beckman, and Weibull distributions can be found in [6].

3.2.3 Level Crossing Rate and Average Fade Duration

In this section we derive the average fade duration of a narrowband fading signal, which equals the average length of time the signal is continuously below a given value. To compute the average fade duration, we must first compute the signal's level crossing rate. The level crossing rate is a characteristic of any random process. It is defined as the expected rate, in crossings per second, that the process crosses a particular value or level. The rate can be based on crossing the level in either the upward or downward direction. In order to compute the average fade duration, we consider the level crossing rate in the downward direction, since the signal is then below the desired value. We thus define the envelope level crossing rate L_Z of signal envelope $z(t) = |r(t)|$ 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 $p(z, \dot{z})$ of the signal envelope $z = |r|$ and its derivative with respect to time, $\dot{z}$. We now derive L_Z based on this joint distribution.

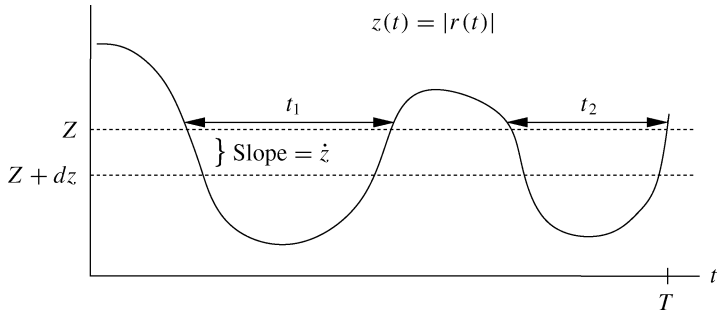

Figure 3.10: Level crossing rate and fade duration for fading process.

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

$$\int_0^T \dot{z}p(Z, \dot{z})d\dot{z}\, dt = \dot{z}p(Z, \dot{z})d\dot{z}T. \tag{3.41}$$

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

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

Finally, the expected number of crossings of the envelope level Z per second – that is, the level crossing rate – is

$$L_Z = \frac{N_Z}{T} = \int_{-\infty}^0 \dot{z}p(Z, \dot{z})d\dot{z}. \tag{3.43}$$

Note that this derivation of level crossing rate 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 [7, Chap. 2.1]. The level crossing rate for Rician fading is then obtained by using this pdf in (3.43), yielding

$$L_Z = \sqrt{2\pi(K + 1)} f_D \rho e^{-K-(K+1)\rho^2} I_0(2\rho\sqrt{K(K + 1)}), \tag{3.44}$$

where $\rho = Z/\sqrt{\bar{P}_r}$. For Rayleigh fading ($K = 0$) the level crossing rate simplifies to

$$L_Z = \sqrt{2\pi} f_D \rho e^{-\rho^2}. \tag{3.45}$$

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 such as bit error rate. If the signal amplitude or power falls below its target then we say the system is in outage. 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

$$p(z(t) < Z) = \frac{1}{T} \sum_i t_i. \tag{3.46}$$

Thus, for T sufficiently large, the average fade duration is

$$\bar{t}_Z = \frac{1}{TL_Z} \sum_{i=1}^{L_Z T} t_i \approx \frac{p(z(t) < Z)}{L_Z}. \tag{3.47}$$

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

$$\bar{t}_Z = \frac{e^{\rho^2} - 1}{\rho f_D \sqrt{2\pi}} \tag{3.48}$$

with $\rho = Z/\sqrt{\bar{P}_r}$. Note that (3.48) is the average fade duration for the signal envelope (amplitude) level with Z the target amplitude and $\sqrt{\bar{P}_r}$ the average envelope level. By a change of variables it is easily shown that (3.48) also yields the average fade duration for the signal power level with $\rho = \sqrt{P_0/\bar{P}_r}$, where P_0 is the target power level and $\bar{P}_r$ is the average power level. The average fade duration (3.48) decreases with the Doppler frequency f_D, 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 $\rho \gg 1$. That is because the signal is more likely to be below the target as the target level increases relative to the average. The average fade duration for Rician fading is more difficult to compute; it can be found in [7, Chap. 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 $z < Z$. In this case, if $T_b \approx \bar{t}_Z$ then 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 $T_b \ll \bar{t}_Z$ then many subsequent bits are received with $z < Z$, so large bursts of errors are likely. Finally, if $T_b \gg \bar{t}_Z$ then, since the fading is integrated over a bit time in the demodulator, the fading gets averaged out and so can be neglected. These issues will be explored in more detail in Chapter 8, where 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 $P_0 = .5\bar{P}_r$ and thus $\rho = \sqrt{.5}$. We require

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

and thus $f_D \geq (e^{.5} - 1)/(.060\sqrt{\pi}) = 6.1$ Hz.

3.2.4 Block-Fading and Finite-State Markov Fading

The complex mathematical characterization of the narrowband fading model described in the previous sections can be difficult to incorporate into wireless performance analysis. Therefore, simpler models that capture the main features of narrowband fading channels are often used for these analytical calculations. This section describes two such models: block-fading and finite-state Markov fading channels.

In a block-fading channel, the channel SNR γ is assumed constant over some period of time T called the *block time*. In this model we denote $\gamma(i)$ as the channel SNR over the ith block. It is typically assumed that $\gamma(i)$ is i.i.d. between blocks although a more complex model capturing correlation of the channel gain across blocks can also be used. Often T is set equal to an integer number of modulation symbol times T_s: $T = NT_s$ where NT_s is approximately equal to the channel coherence time defined in Section 3.3.4 below. A rapidly-fading channel might be approximated by setting $N = 1$ such that the fading is constant over each symbol time and changes to a different i.i.d. value between subsequent symbols. A variation on the block-fading model allows for continuous variation of the channel gain within a block, typically using the model of Section 3.2.1, with an abrupt transition to an independent channel gain at each block time. This model is often used for systems where blocks of symbols are sent intermittently and hence each block experiences a completely different propagation environment.

The finite-state Markov channel (FSMC) is a special case of block-fading where a finite set of channel SNR values is assumed with a Markov transition probability of the fading between blocks. Specifically in the FSMC the channel SNR is assumed to belong to a finite set of values, called channel states, and to follow a discrete-time Markov process with time discretized to the block time T. The channel varies over its possible 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 [12], indoor channels [13], Rayleigh fading channels [14, 15], Rician fading channels [16], and Nakagami-m fading channels [17]. They have also been used for system design and system performance analysis [15, 18]. First-order FSMC models are deficient when the fading recorrelation affects performance analysis, in which case higher-order models are used. The FSMC models for fading typically model amplitude variations only, although there has been some work on FSMC models for phase in fading [19] or phase-noisy channels [20].

A detailed FSMC model for Rayleigh fading was developed in [14]. In this model the time-varying SNR γ associated with the Rayleigh fading lies in the range $0 \leq \gamma < infty$. 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 [14] as

$$p_{j,j+1} = \frac{L_{j+1}T}{\pi_j}, \quad p_{j,j-1} = \frac{L_j T}{\pi_j}, \quad p_{j,j} = 1 - p_{j,j+1} - p_{j,j-1}, \tag{3.49}$$

where L_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 Model

Suppose the transmitted signal is not narrowband relative to the inverse delay spread. Since the signal bandwidth $B \approx 1/T$ for T the signal duration, the received signal will have duration $T + T_m >> T$ for T_m is the delay spread. Thus, the duration of the received signal will be significantly increased, and we refer to the channel in this case as a *dispersive* channel. This phenomenon is illustrated in Figure 3.11. In the 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.[6] If the multipath delay spread $T_m \ll T$ then the multipath components are received roughly on top of one another, as shown in the upper right of the figure. Hence the channel is nondispersive. 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 $T_m \gg T$, 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 (dashed pulses in the figure). This effect is called *intersymbol interference* (ISI).

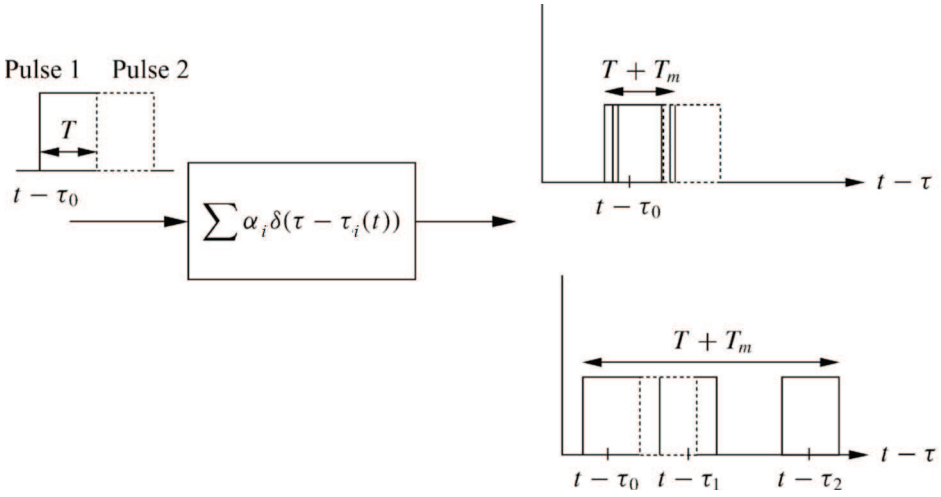

Figure 3.11: Multipath resolution.

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. Mitigating ISI is not necessary if $T \gg T_m$. However, most wireless systems require data rates $R \approx 1/T$ to be much greater than $1/T_m$ for typical delay spreads of both indoor and outdoor channels. 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 3.3.3.

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

Although the approximation in (3.13) 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 com-

[6] Linear modulation typically uses nonsquare pulse shapes for bandwidth efficiency, as discussed in Section 5.4.

ponent 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 are not resolvable if their delay difference is less than the inverse of the signal bandwidth, so the multipath components characterized in (3.6) combine at the receiver to yield the original transmitted signal with amplitude and phase characterized by random processes. These random processes are in turn characterized by their autocorrelation (or PSD) and 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.

3.3.1 Autocorrelation and Scattering Function

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[7] deterministic function of τ and t. Recall that $c(\tau,\ t)$ represents the response of the channel at time t to an impulse at time $t-\tau$. Hence, time variations of the impulse response are captured by the parameter t. We can take the Fourier transform of $c(\tau,\ t)$ with respect to t as

$$C(\tau,\ \rho) = \int_{-\infty}^{\infty} c(\tau,\ t)e^{-j2\pi\rho t}dt. \tag{3.50}$$

We call $C(\tau,\ \rho)$ the *deterministic scattering function* of the equivalent lowpass channel impulse response $c(\tau,\ t)$. Recall that time variations of the channel due to movement of the transmitter or receiver result in a Doppler frequency shift f_D. Since $C(\tau,\ \rho)$ is the Fourier transform of $c(\tau,\ t)$ with respect to the time variation parameter t, the deterministic scattering function $C(\tau,\ \rho)$ captures the Doppler characteristics of the channel via the frequency parameter ρ. The derivation of this relationship for random channels is given in Section 3.3.4

In general, the time-varying channel impulse response $c(\tau,\ t)$ given by (3.6) is random instead of deterministic because of 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 and the LOS signal component does not dominate the received signal, we can invoke the CLT to assume that $c(\tau,\ t)$ is a zero-mean complex Gaussian process and hence that 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

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

Most time-varying channels associated with deployed wireless systems are WSS, so that the joint statistics of a channel measured at two different times t and $t+\triangle t$ depends only on the time difference $\triangle t$. For WSS channels, the autocorrelation of the corresponding bandpass channel $h(\tau,\ t) = \mathrm{Re}\{c(\tau,\ t)e^{j2\pi f_C t}\}$ can be obtained from

[7]The wideband channel characterizations in this section can also be made for discrete-time channels (discrete with respect to τ) by changing integrals to sums and Fourier transforms to discrete Fourier transforms.

$A_c(\tau_1,\ \tau_2; t,\ t+\triangle t)$ as[8] $A_h(\tau_1,\ \tau_2; t,\ t+\triangle t) = .5\,\mathrm{Re}\{A_c(\tau_1,\ \tau_2; t,\ t+\triangle t)e^{j2\pi f_c \triangle t}\}$ [21]. We will assume that our channel model is WSS, in which case the autocorrelation becomes independent of t:

$$A_c(\tau_1,\ \tau_2; \triangle t) = \mathbf{E}[c^*(\tau_1; t)c(\tau_2; t+\triangle t)]. \tag{3.52}$$

Moreover, in real environments 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 denote channels that are WSS with US as WSSUS channels. The WSSUS channel model was first introduced by Bello in his landmark paper [21], where he also developed two-dimensional transform relationships associated with this autocorrelation. These relationships will be discussed in Section 3.3.5. Incorporating the US property into (3.52) yields

$$\mathbf{E}[c^*(\tau_1; t)c(\tau_2; t+\triangle t)] = A_c(\tau_1; \triangle t)\delta[\tau_1 - \tau_2]. \tag{3.53}$$

Now setting $\tau_1 = \tau_2 = \tau$ we get

$$\mathbf{E}[c^*(\tau; t)c(\tau; t+\triangle t)] = A_c(\tau; \triangle t), \tag{3.54}$$

where $A_c(\tau; \triangle t)$ gives the average output power associated with the channel as a function of the multipath delay $\tau = \tau_1 = \tau_2$ and the difference $\triangle t$ in observation time. The autocorrelation function (3.54) is based on the approximation that $\tau_1 = \tau_2$ when τ_1 and τ_2 satisfy $|\tau_1 - \tau_2| > B_u^{-1}$, since otherwise the receiver can't resolve the multipath components with these delays. In this case the two components are modeled as a single combined multipath component with delay $\tau \approx \tau_1 \approx \tau_2$ and autocorrelation $A_c(\tau; \triangle t)$.

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

$$S_c(\tau,\ \rho) = \int_{-\infty}^{\infty} A_c(\tau,\ \triangle t)e^{-j2\pi\rho\triangle t}d\triangle t. \tag{3.55}$$

The scattering function characterizes the average output power associated with the channel as a function of the multipath delay τ and Doppler ρ. By constrast, for the deterministic channel, its scattering function $C(\tau,\ \rho)$ characterizes the complex channel gain as a function of multipath delay τ and Doppler ρ. An example scattering function for an indoor channel with the power density and delay spread normalized by their mean values is shown in Figure 3.12.

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

3.3.2 Power Delay Profile

The *power delay profile* $A_c(\tau)$, also called the *multipath intensity profile*, is defined as the autocorrelation (3.54) with $\triangle t = 0 : A_c(\tau) \triangleq A_c(\tau, 0)$. The power delay profile represents the average power associated with a given multipath delay, and it is easily measured empirically. The average and rms delay spread are typically defined in terms of the power delay profile $A_c(\tau)$ as

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

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

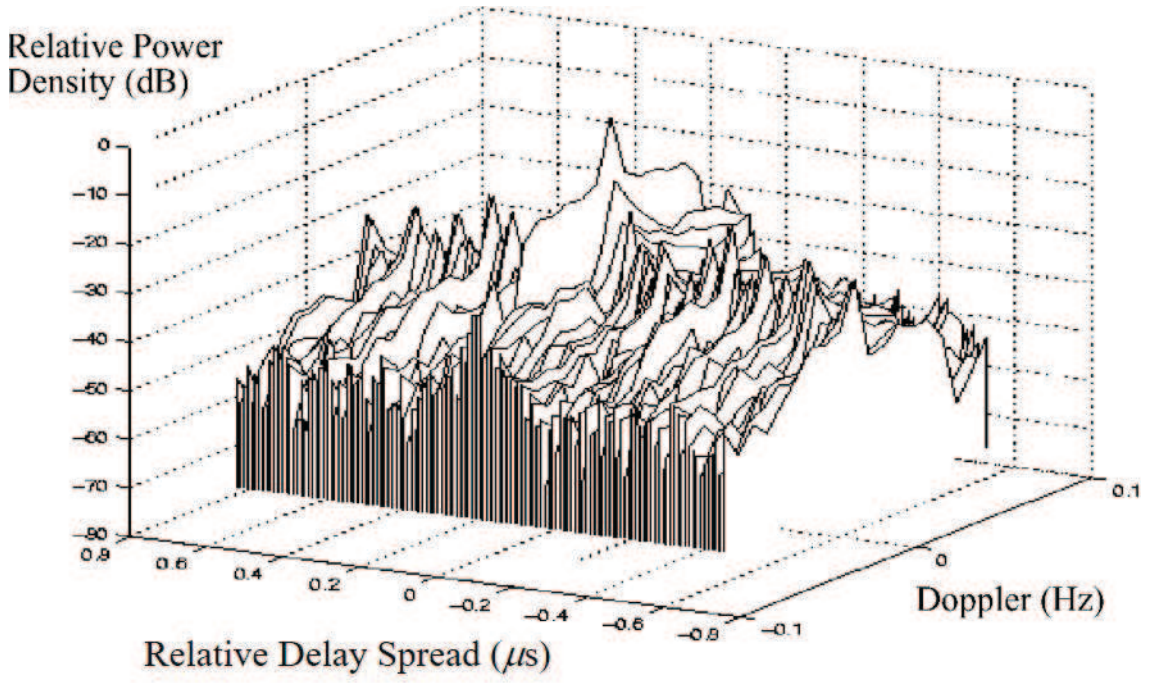

Figure 3.12: Scattering function with power spectral density and delay spread normalized by their mean values.

and

$$\sigma_{T_m} = \sqrt{\frac{\int_0^\infty (\tau - \mu_{T_m})^2 A_c(\tau) d\tau}{\int_0^\infty A_c(\tau) d\tau}}. \tag{3.57}$$

Note that $A_c(\tau) \geq 0$ for all τ, so if we define the distribution p_{T_m} of the random delay spread T_m as

$$p_{T_m}(\tau) = \frac{A_c(\tau)}{\int_0^\infty A_c(\tau) d\tau} \tag{3.58}$$

then μ_{T_m} and σ_{T_m} are (respectively) the mean and rms values of T_m, relative to this distribution. Defining the distribution of T_m by (3.58) – or, equivalently, defining the mean and rms delay spread by (3.56) and (3.57), respectively – weighs 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 affect 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. For example, we assume $A_c(\tau) \approx 0$ for $\tau > 3\sigma_{T_m}$. With this approximation, a linearly modulated signal with symbol period T_s experiences significant ISI if $T_s \ll \sigma_{T_m}$. Conversely, when $T_s \gg \sigma_{T_m}$ the system experiences negligible ISI. For calculations one can assume that $T_s \ll \sigma_{T_m}$ implies $T_s < \sigma_{T_m}/10$ and that $T_s \gg \sigma_{T_m}$ implies $T_s > 10\sigma_{T_m}$. If 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. In later chapters we will study the performance degradation due to ISI in linearly modulated systems as well as ISI mitigation methods.

Although $\mu_{T_m} \approx \sigma_{T_m}$ 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 $A_c(\tau)$. For a channel with no LOS component and a small number of

multipath components with approximately the same large delay, $\mu_{T_m} \gg \sigma_{T_m}$. 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 roughly 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 profile 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 (3.56) is $\mu_{T_m} = \overline{T}_m$ and find the rms delay spread (3.57).

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

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

and the rms delay spread is

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

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

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

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

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

The rms delay spread is

$$\sigma_{T_m} = \sqrt{\frac{\int_0^{20\cdot 10^{-6}} (\tau - \mu_{T_m})^2 e^{-\tau/.00001} d\tau}{\int_0^{20\cdot 10^{-6}} e^{-\tau/.00001} d\tau}} = 5.25\ \mu\text{s}.$$

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 $T_s = 52.5\ \mu$s 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), voice requires on the order of 32 kbps and data has orders of magnitude higher rates, with some systems operating at Gbps data rates.

3.3.3 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

$$C(f;t) = \int_{-\infty}^{\infty} c(\tau;t)e^{-j2\pi f\tau}d\tau. \tag{3.59}$$

Because $c(\tau;t)$ is a complex zero-mean Gaussian random variable in t, the Fourier transform in (3.59) represents the sum[9] of complex zero-mean Gaussian random processes; hence $C(f;t)$ is also a zero-mean Gaussian random process that is completely characterized by its autocorrelation. Since $c(\tau;t)$ is WSS, its integral $C(f;t)$ is also. Thus, the autocorrelation of (3.59) is given by

$$A_C(f_1,\ f_2;\triangle t) = \mathbf{E}[C^*(f_1;t)C(f_2;t+\triangle t)]. \tag{3.60}$$

We can simplify $A_C(f_1,\ f_2;\triangle t)$ as follows:

$$\begin{aligned} A_C(f_1,\ f_2;\triangle t) &= \mathbf{E}\left[\int_{-\infty}^{\infty} c^*(\tau_1;t)e^{j2\pi f_1\tau_1}d\tau_1 \int_{-\infty}^{\infty} c(\tau_2;t+\triangle t)e^{-j2\pi f_2\tau_2}d\tau_2\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{E}[c^*(\tau_1;t)c(\tau_2;t+\triangle t)]e^{j2\pi f_1\tau_1}e^{-j2\pi f_2\tau_2}d\tau_1 d\tau_2 \\ &= \int_{-\infty}^{\infty} A_c(\tau,\ \triangle t)e^{-j2\pi(f_2-f_1)\tau}d\tau \\ &= A_C(\triangle f;\triangle t), \end{aligned} \tag{3.61}$$

where $\triangle f = f_2 - f_1$ 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 $\triangle f$. The function $A_C(\triangle f;\triangle t)$ can be measured in practice by transmitting a pair of sinusoids through the channel that are separated in frequency by $\triangle f$ and then calculating their cross-correlation at the receiver for the time separation $\triangle t$.

If we define $A_C(\triangle f) \triangleq A_C(\triangle f;0)$ then, by (3.61),

$$A_C(\triangle f) = \int_{-\infty}^{\infty} A_c(\tau)e^{-j2\pi\triangle f\tau}d\tau. \tag{3.62}$$

Thus $A_C(\triangle f)$ is the Fourier transform of the power delay profile. Because $A_C(\triangle f) = \mathbf{E}[C^*(f;t)C(f+\triangle f;t)]$ is an autocorrelation, it follows that the channel response is approximately independent at frequency separations $\triangle f$ where $A_C(\triangle f) \approx 0$. The frequency B_c where $A_C(\triangle f) \approx 0$ for all $\triangle f > B_c$ is called the *coherence bandwidth* of the channel. By the Fourier transform relationship between $A_c(\tau)$ and $A_C(\triangle f)$, if $A_c(\tau) \approx 0$ for $\tau > T$ then $A_C(\triangle f) \approx 0$ for $\triangle f > 1/T$. Hence, 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 [22] has shown that $B_c \approx .02/\sigma_{T_m}$ approximates the

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

range of frequencies over which channel correlation exceeds 0.9 whereas $B_c \approx .2/\sigma_{T_m}$ approximates the range of frequencies over which this correlation exceeds 0.5.

In general, if we are transmitting a narrowband signal with bandwidth $B \ll B_c$, then fading across the entire signal bandwidth is highly correlated; that is, 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 $B \gg B_c$, 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 fading is called *frequency selective*. If $B \approx B_c$ 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 $T_s \approx 1/B \gg 1/B_c \approx \sigma_{T_m}$ – that is, the case where the channel experiences negligible ISI. Frequency-selective fading corresponds to $T_s \approx 1/B \ll 1/B_c = \sigma_{T_m}$, 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; this degrades performance, as will be discussed in Chapters 12 and 13.

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

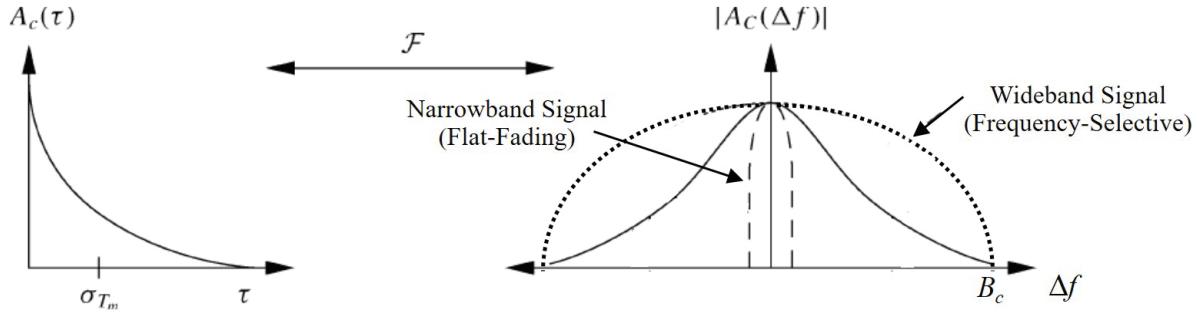

Figure 3.13: Power delay profile, rms delay spread, and coherence bandwidth.

Example 3.6: Consider an indoor channel with $\sigma_{T_m} = 30$ ns and a large macrocell channel with $\sigma_{T_m} = 300$ ns. Find the maximum symbol rate $R_s = 1/T_s$ for these environments such that a linearly modulated signal transmitted through them experiences negligible ISI.

Solution: We assume that negligible ISI requires $T_s \gg \sigma_{T_m}$ (i.e., $T_s \geq 10\sigma_{T_m}$). This translates into a symbol rate of $R_s = 1/T_s \leq .1/\sigma_{T_m}$. For $\sigma_{T_m} = 30$ ns this yields $R_s \leq 3.33$ Mbps and for $\sigma_{T_m} = 300$ ns this yields $R_s \leq 333$ Kbps. Note that high-performance wireless systems today can support Gbps data rates indoors and out. To maintain these data rates for a linearly modulated signal without severe performance degradation by ISI, some form of ISI mitigation is needed. Moreover, ISI is less severe in indoor than in outdoor systems owing to the former's lower delay spread values, which is why indoor systems tend to have higher data rates than outdoor systems.

3.3.4 Doppler Power Spectrum and Channel Coherence Time

The time variations of the channel that 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 $A_C(\triangle f; \triangle t)$ relative to $\triangle t$:

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

In order to characterize Doppler at a single frequency, we set $\triangle f$ to zero and then define $S_C(\rho) \triangleq S_C(0; \rho)$. It is easily seen that

$$S_C(\rho) = \int_{-\infty}^{\infty} A_C(\triangle t) e^{-j2\pi\rho\triangle t} d\triangle t, \tag{3.64}$$

where $A_C(\triangle t) \triangleq A_C(\triangle f = 0; \triangle t)$. Note that $A_C(\triangle t)$ is an autocorrelation function defining how the channel impulse response decorrelates over time. In particular, $A_C(\triangle t = T) = 0$ indicates that observations of the channel impulse response at times separated by T are uncorrelated and therefore independent, since the channel impulse response is a Gaussian random process. We define the *channel coherence time* T_c to be the range of $\triangle t$ values over which $A_C(\triangle t)$ 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. Since $S_C(\rho)$ is the Fourier transform of an autocorrelation, 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, denoted by B_D. By the Fourier transform relationship between $A_C(\triangle t)$ and $S_C(\rho)$, we have $B_D \approx 1/T_c$. 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 $\triangle t = .4/f_D$, 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 $A_C(\triangle_t)$ in Figure 3.14.

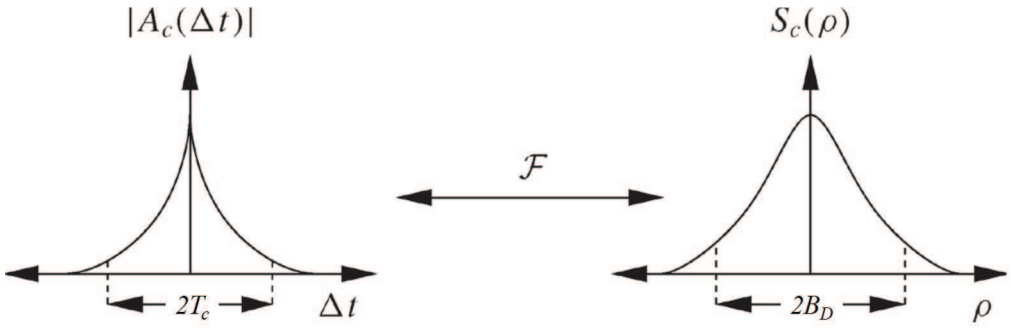

Figure 3.14: Doppler power spectrum, Doppler spread, and coherence time.

Example 3.7: For a channel with Doppler spread $B_D = 80$ Hz, find the time separation required in samples of the received signal in order for the samples to be 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. Thus, given the Gaussian properties of the underlying random process, these samples are approximately independent.

3.3.5 Transforms for Autocorrelation and Scattering Functions

From (3.63) we see that the scattering function $S_c(\tau;\rho)$ defined in (3.55) is the inverse Fourier transform of $S_C(\triangle f;\rho)$ in the $\triangle f$ variable. Furthermore $S_c(\tau;\rho)$ and $A_C(\triangle f;\triangle t)$ are related by the double Fourier transform

$$S_c(\tau;\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A_C(\triangle f;\triangle t)e^{-j2\pi\rho\triangle t}e^{j2\pi\tau\triangle f}d\triangle t\, d\triangle f. \tag{3.65}$$

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

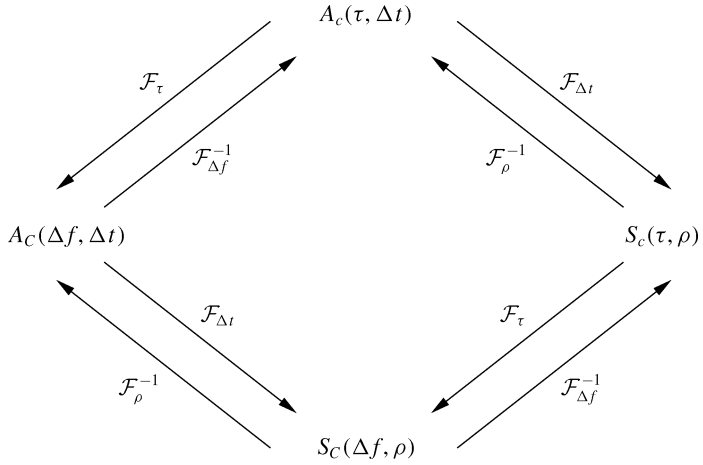

Figure 3.15: Fourier transform relationships.

Empirical measurements of the scattering function for a given channel are often used to approximate 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 $A_c(\tau,\ \triangle t) = \mathcal{F}_\rho^{-1}[S_c(\tau;\rho)]$ with $\triangle t = 0$ 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 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 [23], is especially useful in the study of spread-spectrum systems and RAKE receivers (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_i are combined, and incoming paths on

different subpath clusters with delays τ_i and τ_j, where $|\tau_i - \tau_j| > 1/B_u$, can be resolved. Here B_u denotes the baseband signal bandwidth.

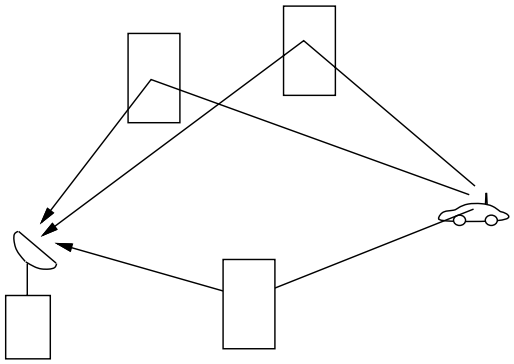

Figure 3.16: Point scatterer channel model.

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

$$c(\tau;t) = \sum_{i=0}^{N-1} \alpha_i(t) e^{-j\phi_i(t)} \delta(\tau - \tau_i(t)). \tag{3.66}$$

The statistics of the received signal for a given t are thus given by the statistics of $\{\tau_i\}_0^N$, $\{\alpha_i\}_0^N$, and $\{\phi_i\}_0^N$. Note that $c(\tau;t)$ in (3.66) has infinite bandwidth since it consists of a sum of delta functions. Since the model has infinite bandwidth, its discrete-time approximation associated with any finite sampling rate entails sampling distortion, as the channel's Nyquist rate is infinite.

With this sampling distortion in mind, the model of (3.66) can be approximated in discrete-time 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 being approximated. 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: $T \approx 1/B_u$. The statistics for the ith bin are that $r_i, 1 \leq i \leq M$, is a binary indicator of the existence of a multipath component in the ith bin: so $r_i = 1$ if there is a multipath component in the ith bin and 0 otherwise. If $r_i = 1$ then $(a_i,\ \theta_i)$, the amplitude and phase corresponding to this multipath component, follow an empirically determined distribution. This distribution is obtained by sample averages of $(a_i,\ \theta_i)$ for each i at different locations in the propagation environment. The empirical distribution of $(a_i,\ \theta_i)$ and $(a_j,\ \theta_j), i \neq j$, is generally different; it may correspond to the same family of fading but with different parameters (e.g., Rician fading with different K factors) or to different fading distributions altogether (e.g., Rayleigh fading for the ith bin, Nakagami fading for the jth bin).

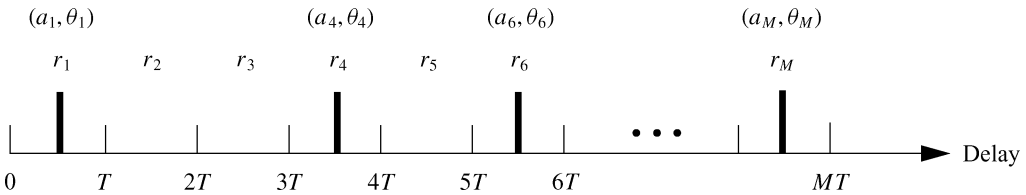

Figure 3.17: Discrete-time approximation.

This completes our statistical model of the discrete-time approximation for a single snapshot. A sequence of profiles will model the signal over time as the channel impulse response changes – for example, the impulse

response seen by a receiver moving at some nonzero velocity through a city. Thus, the model must include not only the first-order statistics of $(\tau_i, \alpha_i, \phi_i)$ for each profile (equivalently, each t) but also the temporal and spatial correlations (assumed to be Markov) between them. More details on the model and the empirically derived distributions for N and for $(\tau_i, \alpha_i, \phi_i)$ can be found in [23].

3.5 MIMO Channel Models

Multiple antennas at the transmitter and/or receiver are common in wireless systems because of their diversity and capacity benefits. These MIMO systems with multiple antennas require channel models that characterize both spatial and temporal characteristics of the channel. For a channel with M transmit antennas and N receive antennas, the MIMO channel can be written as a matrix

$$H(t,\tau) = \begin{pmatrix} h_{11}(t,\tau) & h_{12}(t,\tau) & \cdots & h_{1M}(t,\tau) \\ h_{21}(t,\tau) & h_{22}(t,\tau) & \cdots & h_{2M}(t,\tau) \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1}(t,\tau) & h_{N2}(t,\tau) & \cdots & h_{NM}(t,\tau) \end{pmatrix} \tag{3.67}$$

where $h_{ij}(t,\tau)$ is the time-varying impulse response between the jth transmit antenna and the ith receiver antenna as given by (3.6). When the channel is narrowband and time-invariant, each of these terms reduces to a complex channel gain h_{ij}. In this case the MIMO channel model consists of the distributions and cross correlations of these random channel gains. The class of MIMO channel models that capture only these gains is often referred to as the *analytical* MIMO channel model. The complex channel gains h_{ij} in MIMO systems are often assumed to be i.i.d. zero mean, unit variance, circularly symmetric Gaussian random variables, meaning that a fixed phase rotation has no impact on the distribution. This distribution for the MIMO channel gains is based on a physical environment where there is a sufficient number of scatterers to yield independent fading over each transmit-receive signal path. Environments with this characteristic are said to exhibit *rich scattering*. Models for more general environments with antenna gains that are correlated or dependent include the Rice, Kronecker, UIU, Weichselberger, and keyhole models [24]. For each transmit-receive antenna pair in the MIMO system, the single-antenna wideband fading model described in Section 3.3 with channel gain h_{ij} can be used to capture the effects of time variations and delay spread. However, this modeling approach does not capture dependencies across all three MIMO dimensions of space, time, and frequency.

To capture dependencies across all three of these dimensions, we extend the time-varying channel impulse response model (3.6) to take into account multiple transmit and receive antennas. As in the single-antenna case, the propagation environment assumes multiple scatterers that generate the multipath components [24, 25]. The locations of the scatterers at time t dictate the AoA $\theta_i(t)$ at the receiver array and the angle-of-departure (AoD) $\psi_i(t)$ at the transmitter array for the ith multipath component. Models can be either two-dimensional or three-dimensional. Since knowledge of the environment via the angles of arrival and departure is used in this model, it is sometimes referred to as the *parametric* MIMO channel model to contrast it with the analytical model.

The equivalent lowpass time-varying impulse response model (3.6) is extended to incorporate these angles of arrival and departure as follows:

$$c(\tau,t,\theta,\psi) = \sum_{i=0}^{N(t)-1} \alpha_i(t) e^{-j\phi_i(t)} \delta(\tau - \tau_i(t)) \delta(\theta - \theta_i) \delta(\psi - \psi_i) \tag{3.68}$$

where $N(t)$, $\alpha_i(t)$, $\phi_i(t)$ and $\tau_i(t)$ are defined as in (3.6). This time-varying impulse response is zero for all values of the AoA θ that do not correspond to one of the incoming multipath components at time t associated with the

physical environment, and similarly for all values of the AoD ψ. In three dimensions the angles of arrival and departure of each multipath component depends on the elevation and azimuth angles, the polarization of antenna elements, as well as the scatterer locations. Survey papers on such 3-D channel models include [26, 27].

Chapter 3 Problems

1. Consider a two-ray 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 for the length of the reflected path to be approximated by $d_{11} + d_{12} \approx d + 2h^2/d$. Assume a carrier frequency $f = 2$ GHz, unit gain antennas at the transmitter and receiver, and reflection coefficient $R = -1$.

2. Find a formula for the multipath delay spread T_m for a two-ray 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 ns, and another multipath component at delay 67 ns. 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 $f_c\tau_i$ 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, $Z = \sqrt{X^2 + Y^2}$ is Rayleigh-distributed and Z^2 is exponentially-distributed.

6. Assume a Rayleigh fading channel with 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. Suppose we have 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 cumulative distribution function (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. The em Generalized Two Ray (GTR) fading model, proposed in [28], is a generalization of the Rayleigh and Rician fading models. In this model there are two dominant signal components as well as many lower-power multipath components. More precisely, the received signal V_r can be modeled as

$$V_r = V_1 e^{j\phi_1} + V_2 e^{j\phi_2} + X + jY,$$

where V_1 and V_2 are non-negative constants, ϕ_i and ϕ_2 are i.i.d. and uniformly distributed over $[0, 2\pi]$, i.e. $\phi_i \sim \mathcal{U}[0, 2\pi], i = 1, 2$, and X and Y are i.i.d. Gauss-distributed random variables with mean zero and variance σ^2. The GTR model thus reduces to the Rician fading model if $V_1 > 0, V_2 = 0$ and to the Rayleigh

fading model if $V_1 = V_2 = 0$. The model is parameterized in terms of K and Δ where,

$$\begin{aligned} K &= \frac{V_1^2 + V_2^2}{2\sigma^2} \\ \Delta &= \frac{2V_1V_2}{V_1^2 + V_2^2} \end{aligned}$$

(a) Find the range of parameters K and Δ.

(b) Describe the channel model that corresponds to the following GTR parameters:

- $K \to \infty, \Delta \to 0$
- $K \to \infty, \Delta \to 1$
- $K \to 0$
- $K > 0, \Delta \to 0$

(c) Show that the received signal given that the phase difference between the 2 dominant components is constant,i.e., $\phi_1 - \phi_2 = \alpha$ for a constant α, consists of one dominant component and the multiple diffuse lower-power multipath components. This implies that the GTR model reduces to the Rician fading model when the phase difference between the LoS components in constant. Find the parameter $\bar{K}(\alpha, \Delta, K)$ of the equivalent Rician model as a function of α, Δ and K.

(d) Show that $\alpha \pmod{2\pi} \sim \mathcal{U}[0, 2\pi]$. Use this to show that

$$p_{\text{GTR}}(r \mid K, \Delta) = \frac{1}{2\pi} \int_0^{2\pi} p_{\text{Rice}}(r \mid \bar{K}(\alpha, \Delta, K)) \mathrm{d}\alpha,$$

where $p_{\text{GTR}}(r \mid \Sigma)$ is the pdf of the received signal amplitude of the GTR channel fading model with parameters Σ and $p_{\text{Rice}}(r \mid \Sigma)$ is the pdf of the received signal amplitude of the GTR channel fading model with parameters Σ.

11. 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 *macro-diversity*, 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. The term W is the received signal power while Z_1 and Z_2 are independent noise terms.

(a) 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.

(b) 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

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

(c) 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).

12. 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 simulation tool can be used.

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

13. For a Rayleigh fading channel with average power $\overline{P}_r = 30$ dB and Doppler frequency $f_D = 10$ Hz, compute the average fade duration for target fade values $P_0 = 0$ dB, $P_0 = 15$ dB, and $P_0 = 30$ dB.

14. Derive a formula for the average length of time a Rayleigh fading process with average power $\overline{P}_r$ stays **above** a given target fade value P_0. Evaluate this average length of time for $\overline{P}_r = 20$ dB, $P_0 = 25$ dB, and $f_D = 50$ Hz and compare with the average fade depth below P_0 associated with the same parameters.

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

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

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

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 eight 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?

17. In this problem we examine the narrowband approximation in the two-ray model. Consider the two-ray model discussed in Chapter 2.4 with reflection coefficient $R = -1$, free space path loss for each ray and carrier frequency 2.4 GHz. Both the transmitter and the receiver are 40 m high.

(a) For a separation distance of 20 m, find the impulse response of the equivalent baseband channel in the time domain. What is the delay spread T_m?

(b) Repeat the computation of the impulse response and delay spread for a separation distance of 2000 m.

(c) Plot the baseband channel output in both the time and frequency domain in each of the above two channels for channel input $\text{sinc}(ft) \triangleq \frac{\sin(\pi ft)}{\pi ft}$ and $f = 20$ MHz. Is there ISI in either channel? Find the coherence bandwidth of the channel in each case and whether this bandwidth implies flat or frequency selective fading.

18. 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 $\text{sinc}(x) = \sin(\pi x)/(\pi x)$.

(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.

19. 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 sinusoids $\cos(2\pi f_1 t)$ and $\cos(2\pi f_2 t)$ separated in frequency by $\Delta f = |f_1 - f_2|$. 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 $u_1(t) = \sin(2\pi ft)$ and $u_2(t) = \sin(2\pi f(t + \Delta t))$, find 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?

20. Consider a MIMO channel $\mathbf{H}$ with a time-invariant complex gain h_{ij} between the jth transmit antenna and the ith receive antenna. A MIMO channel is called a *keyhole* channel when the channel matrix $\mathbf{H}$ can be written as

$$\mathbf{H} = h_r h_t^*$$

where h_t is an $M \times 1$ vector of i.i.d. zero-mean unit variance complex Gaussian random variables, and h_r is an $N \times 1$ vector of i.i.d. zero-mean unit variance complex Gaussian random variables. Thus, h_{ij} is the product of two independent complex Gaussian random variables.

(a) Show that the power of h_{ij} has a distribution $p(x) = 2K_0(2\sqrt{x})$ where K_0 is the modified Bessel function of the second kind.

(b) Show that the matrix $\mathbf{H}$ has rank 1, i.e. show that all of its columns are linearly dependent, and the same for its rows.

The fact that the matrix $\mathbf{H}$ is rank 1 is what gives rise to the name keyhole channel, since the MIMO channel's multiple spatial dimensions reduce to a single dimension, as if the signal between the transmit and receive antenna arrays travels through a keyhole.

Bibliography

[1] T. S. Rappaport, *Wireless Communications – Principles and Practice*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2001.

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

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

[4] R. H. Clarke, "A statistical theory of mobile radio reception," *Bell System Tech. J.*, pp. 957–1000, July/August 1968.

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

[6] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*: *A Unified Approach to Performance Analysis*, 2nd Ed., Wiley, New York, 2004.

[7] G. L. Stuber, *Principles of Mobile Communications*, 2nd ed., Kluwer, Dordrecht, 2001.

[8] J. Zhang, C. Pan, F. Pei, G. Liu, and X. Cheng, Three-dimensional fading channel models: a survey of elevation angle research, IEEE Commun. Mag., vol. 52, no. 6, pp. 218226, 2014.

[9] M. Pätzold, *Mobile Fading Channels*, Wiley, New York, 2002.

[10] J. Cavers, *Mobile Channel Characteristics*, Springer 2000.

[11] S. O. Rice, "Mathematical analysis of random noise," *Bell System Tech. J.*, pp. 282–333, July 1944, and pp. 46–156, January 1945.

[12] F. Babich, G. Lombardi, and E. Valentinuzzi, "Variable order Markov modeling for LEO mobile satellite channels," *Elec. Lett.*, pp. 621–3, April 1999.

[13] A. M. Chen and R. R. Rao, "On tractable wireless channel models," *Proc. Intemat. Sympos. Pers., Indoor, Mobile Radio Commun.*, pp. 825–30, September 1998.

[14] H. S. Wang and N. Moayeri, "Finite-state Markov channel – A useful model for radio communication channels," *IEEE Trans. Veh. Tech.*, pp. 163–71, February 1995.

[15] C. C. Tan and N. C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Trans. Commun.*, pp. 2032–40, December 2000.

[16] C. Pimentel and I. F. Blake, "Modeling burst channels using partitioned Fritchman's Markov models," *IEEE Trans. Veh. Tech.*, pp. 885–99, August 1998.

[17] Y. L. Guan and L. F. Tumer, "Generalised FSMC model for radio channels with correlated fading," *IEE Proc. Commun*., pp. 133–7, April 1999.

[18] M. Chu and W. Stark, "Effect of mobile velocity on communications in fading channels," *IEEE Trans. Veh. Tech*., pp. 202–10, January 2000.

[19] C. Komninakis and R. D. Wesel, "Pilot-aided joint data and channel estimation in flat correlated fading," *Proc. IEEE Globecom Conf*., pp. 2534–9, November 1999.

[20] M. Peleg, S. Shamai (Shitz), and S. Galan, "Iterative decoding for coded noncoherent MPSK communications over phase-noisy AWGN channels," *IEE Proc. Commun*., pp. 87–95, April 2000.

[21] P. A. Bello, "Characterization of randomly time-variant linear channels," *IEEE Trans. Commun. Syst*., pp. 360–93, December 1963.

[22] W. C. Y. Lee, *Mobile Cellular Telecommunications Systems*, McGraw-Hill, New York, 1989.

[23] G. L. Turin, "Introduction to spread spectrum antimultipath techniques and their application to urban digital radio," *Proc. IEEE*, pp. 328–53, March 1980.

[24] R. W. Heath and A. Lozano, *Foundations of MIMO Communication,* Cambridge University Press, 2019.

[25] R. Ertel, P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Personal Communications Magazine*., pp. 10–22, February 1998.

[26] U. Karabulut, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, Spatial and Temporal Channel Characteristics of 5G 3D Channel Model with Beamforming for User Mobility Investigations *IEEE Communications Magazine*, Special Issue on Channel Models and Measurements for 5G, Dec. 2018.

[27] J. Zhang, C. Pan, F. Pei, G. Liu, and X. Cheng, Three-dimensional fading channel models: a survey of elevation angle research, *IEEE Communications Magazine*, June 2014.

[28] G. D. Durgin, T. S. Rappaport, D. A. de Wolf, "New analytical models and probability density functions for fading in wireless communications, *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 1005-1015, Jun. 2002.

Chapter 4

Capacity of Wireless Channels

The growing demand for wireless communication makes it important to determine the capacity limits of the underlying channels for these systems. 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. The mathematical theory of communication underlying channel capacity was pioneered by Claude Shannon in the late 1940s. This theory is based on the notion of mutual information between the input and output of a channel [1, 2, 3]. In particular, Shannon defined channel capacity as the channel's mutual information maximized over all possible input distributions. The significance of this mathematical construct was Shannon's coding theorem and its converse. The coding theorem proved that a code did exist that could achieve a data rate close to capacity with negligible probability of error. The converse proved 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: the high data rates he predicted for telephone channels, and his 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 and so, 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 transmitter and/or receiver have a single antenna. The capacity of single-user MIMO systems where the transmitter and receiver both have multiple antennas is treated in Chapter 10 and that of multiuser systems 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 additive white Gaussian noise (AWGN) channel and then consider capacity of time-varying flat fading channels. Unlike the AWGN case, here the capacity of a flat fading channel is not given by a single formula because 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 first consider flat fading channel capacity where only the fading distribution is known at the transmitter and receiver. Capacity under this assumption is typically 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 when 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 fade level 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 such a channel is known and the capacity-achieving power allocation 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 AWGN channel with channel input/out 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 received signal power P. The received signal-to-noise ratio (SNR) – the power in $x[i]$ divided by the power in $n[i]$ – is constant and given by $\gamma = P/(N_0 B)$, where $N_0/2$ is the power spectral density (PSD) 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 per 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.

Shannon capacity for AWGN channels provides interesting insights along the asymptotic regimes of bandwidth and power. In particular, it is easily shown that

$$\lim_{B\to\infty} B \log_2\left(1 + \frac{P}{N_0 B}\right) = \log_2(e)\frac{P}{N_0}. \tag{4.2}$$

This regime is called the *large bandwidth regime* for capacity. Interestingly, in this regime the capacity no longer depends on the bandwidth B. That is because there is not sufficient power to spread over the large amount of bandwidth available, so as the bandwidth grows large, there is diminishing returns in capacity growth. The *low power regime* captures a similar scenario, where there is not sufficient power to utilize the available bandwidth. In particular, using the approximation that $\log(1+x) \approx x$ for x small yields that, for P small,

$$B \log_2\left(1 + \frac{P}{N_0 B}\right) \approx \log_2(e)\frac{P}{N_0}. \tag{4.3}$$

Here we also see that capacity no longer depends on the channel bandwidth. Another interesting regime for capacity is the *high power regime*, where P grows large. We now consider how capacity scales with bandwidth in this regime. Using the high-power approximation $1 + P/(N_0 B) \approx P/(N_0 B)$, we get the following ratio of AWGN channel capacity $C(kB)$ given bandwidth kB compared with $C(B)$ for any constant $k > 1$ in the limit of asymptotically large power:

$$\lim_{P\to\infty} \frac{C(kB)}{C(B)} = \lim_{P\to\infty} \frac{kB\log_2\left(1+\frac{P}{N_0 kB}\right)}{B\log_2\left(1+\frac{P}{N_0 B}\right)} \approx \lim_{P\to\infty} \frac{kB\log_2\left(\frac{P}{N_0 kB}\right)}{B\log_2\left(\frac{P}{N_0 B}\right)} = \lim_{P\to\infty} \frac{k\left[\log_2\left(\frac{P}{N_0 B}\right) + \log_2(1/k)\right]}{\log_2\left(\frac{P}{N_0 B}\right)} = k. \tag{4.4}$$

Thus, in the high power regime, scaling the bandwidth up by a factor of k leads to the same increase in capacity. That is because there is sufficient power to fully utilize the capacity of the additional bandwidth.

Shannon's coding theorem and its converse are proved using the concept of mutual information between the channel input and output. For a discrete memoryless time-invariant channel with random input x and random output y, the channel's *mutual information* is defined as

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \, \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \tag{4.5}$$

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 discrete 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:

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \sum_{x,y} p(x,y) \, \log \left(\frac{p(x,y)}{p(x)p(y)} \right). \tag{4.6}$$

For the AWGN channel, the sum in (4.6) becomes an integral over continuous alphabets and 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, the coding theorem, and its converse can be found in [2, 5, 6].

The proofs of the coding theorem and its converse place no constraints on the complexity or delay of the communication system. Indeed, the capacity-achieving strategy requires a code with an asymptotically large blocklength to drive the error probability to zero. Such codes require asymptotically large transmission energy and asymptotically large delay and complexity to decode. Since practical decoders are constrained in delay and complexity, Shannon capacity is generally used as an upper bound on the data rates that can be achieved under real system constraints. The exact capacity penalty due to finite blocklength codes was derived in [7] as a function of blocklength, error probability, and the *channel dispersion*, which measures the stochastic variability of the channel relative to a deterministic channel with the same capacity.

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 useful bound for real systems. However, by the 1990s breakthroughs in hardware, modulation, and coding techniques brought commercial modems very close to the speeds predicted by Shannon in the 1940s. In fact, modems can exceed this 30-kbps 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 his initial calculation. Advances in coding techniques over the last few decades include turbo codes [8], LDPC codes (originally developed by Gallager in the 1960s and rediscovered in the 1990s) [9], lattice codes [10], and polar codes [11]. All of these codes come within a fraction of a decibel of the Shannon capacity limit using practical blocklengths. These codes and their performance will be covered in Chapter 8.

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 $d_0 = 10\,\text{m}$. Assume the channel has bandwidth $B = 30$ kHz and AWGN with noise PSD $N_0/2$,

where $N_0 = 10^{-9}$ 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 $\gamma = P_r(d)/(N_0B) = .1^3/(10^{-9} \cdot 30 \cdot 10^3) = 33 = 15\,\text{dB}$ for $d = 100\,\text{m}$ and $\gamma = .01^3/(10^{-9} \cdot 30 \cdot 10^3) = .033 = -15\,\text{dB}$ for $d = 1000\,\text{m}$. The corresponding capacities are $C = B\log_2(1+\gamma) = 30000\log_2(1+33) = 152.6$ kbps for $d = 100\,\text{m}$ and $C = 30000\log_2(1+.033) = 1.4$ kbps for $d = 1000\,\text{m}$. Note the significant decrease in capacity at greater distances due to the path-loss exponent of 3, which greatly reduces received power as distance increases.

Figure 4.1: Flat fading channel and system model.

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 $\sqrt{g[i]},\ 0 \le g[i]$, and AWGN $n[i]$, as shown in Figure 4.1. The channel power gain $g[i]$ follows a given distribution $p(g)$; for example, with 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 independent and identically distributed (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 $\bar{P}$ denote the average transmit signal power, $N_0/2$ the noise PSD of $n[i]$, and B the received signal bandwidth. The instantaneous received SNR is then $\gamma[i] = \bar{P}g[i]/(N_0B),\ 0 \le \gamma[i] < \infty$, and its expected value over all time is $\bar{\gamma} = \bar{P}\bar{g}/(N_0B)$. Since $\bar{P}/(N_0B)$ 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, which reconstructs an estimate $\hat{\mathbf{w}}$ of the transmitted message $\mathbf{w}$ from the received signal. 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 as follows.

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 to the receiver at time i, and both the transmitter and receiver know the distribution of $g[i]$.

3. *Transmitter and receiver CSI*: The value of $g[i]$ is known to 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, leading to the highest capacity of the three scenarios. Note that since the instantaneous SNR $\gamma[i]$ is just $g[i]$ multipled by the constant $\bar{P}/(N_0B)$, 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 [12, 13].

4.2.2 Channel Distribution Information 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 (4.6), but solving for the capacity-achieving input distribution (i.e., the distribution achieving the maximum in that equation) can be quite complicated depending on the nature of the fading distribution. Moreover, fading correlation introduces channel memory, in which case the capacity-achieving input distribution is found by optimizing over input blocks, and this 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 are known for two specific models of interest: i.i.d. Rayleigh fading channels and finite-state Markov channels. In i.i.d. Rayleigh fading, the channel power gain is exponentially distributed and changes independently with each channel use. The optimal input distribution for this channel was shown in [14] 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 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 [15], and these bounds are tight at high SNRs.

Approximating Rayleigh fading channels via FSMCs was discussed in Chapter 3.2.4. This model approximates the fading correlation as a Markov process. Although 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 [16, 17] and for general inputs in [18]. 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 final result and complexity of the capacity analysis are high for this relatively simple fading model. This shows the difficulty of obtaining the capacity and related design insights on 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 to the receiver at time i. Equivalently, $\gamma[i]$ is known to 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 because 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. This is defined as the maximum rate that can be transmitted over a channel with an 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 a slow deep fade. By allowing the system to lose some data in the event of such 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 $\bar{P}$ can be obtained from results in [19] as

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

Note that this formula is a probabilistic average: the capacity C is equal to Shannon capacity for an AWGN channel with SNR γ, given by $B\log_2(1+\gamma)$ and averaged over the distribution of γ. That is why Shannon capacity is also called ergodic capacity. However, care must be taken in interpreting (4.7) as an average. In particular, it is incorrect to interpret (4.7) to mean that this average capacity is achieved by maintaining a capacity $B\log_2(1+\gamma)$ when the instantaneous SNR is γ, for only the receiver knows the instantaneous SNR $\gamma[i]$ and so the data rate transmitted over the channel is constant, regardless of γ. Note also that the capacity-achieving code must be sufficiently long that a received codeword is affected by all possible fading states. This can result in significant delay.

By Jensen's inequality,

$$\begin{aligned}\mathbf{E}[B\log_2(1+\gamma)] &= \int B\log_2(1+\gamma)p(\gamma)d\gamma \\ &\le B\log_2(1+\mathbf{E}[\gamma]) = B\log_2(1+\bar{\gamma}),\end{aligned} \tag{4.8}$$

where $\bar{\gamma}$ 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 $\sqrt{g[i]}$, which can take on three possible values: $\sqrt{g_1} = .05$ with probability $p_1 = .1$, $\sqrt{g_2} = .5$ with probability $p_2 = .5$, and $\sqrt{g_3} = 1$ with probability $p_3 = .4$. The transmit power is 10 mW, the noise power spectral density $N_0/2$ has $N_0 = 10^{-9}$ 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 three possible received SNRs: $\gamma_1 = P_t g_1/(N_0B) = .01\cdot(.05^2)/(30000\cdot 10^{-9}) = .8333 = -.79\,\text{dB}$, $\gamma_2 = P_t g_2/(N_0B) = .01\cdot(.5^2)/(30000\cdot 10^{-9}) = 83.333 = 19.2\,\text{dB}$, and $\gamma_3 = P_t g_3/(N_0B) = .01/(30000\cdot 10^{-9}) = 333.33 = 25\,\text{dB}$. The probabilities associated with each of these SNR values are $p(\gamma_1) = .1$, $p(\gamma_2) = .5$, and $p(\gamma_3) = .4$. Thus, the Shannon capacity is given by

$$\begin{aligned}C &= \sum_i B\log_2(1+\gamma_i)p(\gamma_i) \\ &= 30000(.1\log_2(1.8333) + .5\log_2(84.333) + .4\log_2(334.33)) \\ &= 199.26\,\text{kbps}.\end{aligned}$$

The average SNR for this channel is $\bar{\gamma} = .1(.8333) + .5(83.33) + .4(333.33) = 175.08 = 22.43$ 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 $\gamma_{\min}$ 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 $\gamma_{\min}$ [21, 22]. If the received SNR is below $\gamma_{\min}$ then the bits received over that transmission burst cannot be decoded correctly with probability approaching 1, and the receiver declares an outage. The probability of outage is thus $P_{\text{out}} = p(\gamma < \gamma_{\min})$. The average rate correctly received over many transmission bursts is $C_{\text{out}} = (1-P_{\text{out}})B\log_2(1+\gamma_{\min})$ since data is only correctly received on $1-P_{\text{out}}$ transmissions. The value of $\gamma_{\min}$ 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 $P_{\text{out}} = p(\gamma < \gamma_{\min})$ for a Rayleigh fading channel (γ exponentially distributed) with $\bar{\gamma} = 20\,\text{dB}$. We see that capacity approaches zero for small outage probability, due to the requirement that bits transmitted under severe fading must be decoded correctly, 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 $\gamma_{\min}$ (or, equivalently, the P_{out}) that maximizes C_{out}.

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 $P_{\text{out}} < .1$, $P_{\text{out}} = .1$, and $P_{\text{out}} = .6$.

Solution: For time-varying channels with discrete SNR values, the capacity versus outage is a staircase function. Specifically, for $P_{\text{out}} < .1$ 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 $.1 \le P_{\text{out}} < .6$ 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 $.6 \le P_{\text{out}} < 1$ 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 $P_{\text{out}} < .1$, $C = 191.94$ kbps for $.1 \le P_{\text{out}} < .6$, and $C = 251.55$ kbps for $.6 \le P_{\text{out}} < 1$.

For $P_{\text{out}} < .1$, data transmitted at rates close to capacity $C = 26.23$ kbps are always correctly received because the channel can always support this data rate. For $P_{\text{out}} = .1$ we transmit at rates close to $C = 191.94$ kbps, but we can correctly decode these data only when the SNR is γ_2 or γ_3, so the rate correctly received is $(1-.1)191940 = 172.75$ kbps. For $P_{\text{out}} = .6$ we transmit at rates close to $C = 251.55$ kbps but we can correctly

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

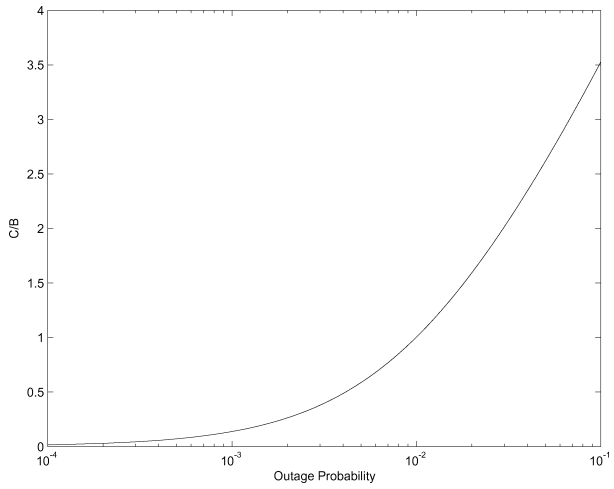

Figure 4.2: Normalized capacity (C/B) versus outage probability.

decode these data only when the SNR is γ_3, so the rate correctly received is $(1 - .6)251550 = 100.62$ 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 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 are discussed in Chapter 14.

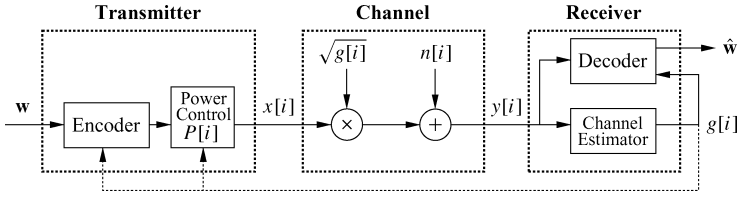

Figure 4.3: System model with transmitter and receiver CSI.

4.2.4 Channel Side Information at Transmitter and Receiver

If both the transmitter and receiver have CSI then 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; we also introduce alternate capacity definitions and their power and rate adaptation strategies.

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 S of discrete memoryless channels. Let C_s denote the capacity of a particular channel $s \in S$ and let $p(s)$ 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 [23]:

$$C = \sum_{s \in \mathcal{S}} C_s p(s). \tag{4.9}$$

We now apply this formula to the system model in Figure 4.3. We know that 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 distribution of the received SNR. By (4.9), the capacity of the fading channel with transmitter and receiver side information is thus[2]

$$C = \int_0^\infty C_\gamma p(\gamma) d\gamma = \int_0^\infty B \log_2(1+\gamma) p(\gamma) d\gamma. \tag{4.10}$$

We see that, without power adaptation, (4.7) and (4.10) are the same, so transmitter side information does not increase capacity unless power is also adapted.

Now let us allow the transmit power $P(\gamma)$ to vary with γ subject to an average power constraint $\bar{P}$:

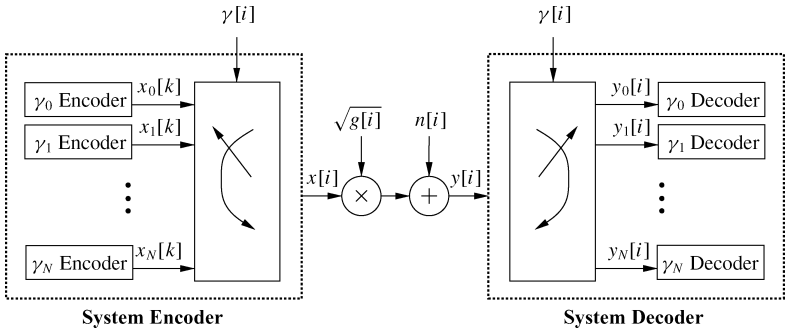

Figure 4.4: Multiplexed coding and decoding.

$$\int_0^\infty P(\gamma) p(\gamma) d\gamma \le \bar{P}. \tag{4.11}$$

With this additional constraint, we cannot apply (4.10) directly to obtain the capacity. However, we expect that the capacity with this average power constraint will be the average capacity given by (4.10) with the power optimally distributed over time. This motivates our definition of the fading channel capacity with average power constraint (4.11) as

$$C = \max_{P(\gamma): \int P(\gamma)p(\gamma)d\gamma = \bar{P}} \int_0^\infty B \log_2\left(1 + \frac{P(\gamma)\gamma}{\bar{P}}\right) p(\gamma) d\gamma. \tag{4.12}$$

It is proved in [24] that the capacity given in (4.12) can be achieved and that 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 $\{\gamma_j : 1 \le j \le N\}$. 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 $P(\gamma_j)\gamma_j/\bar{P}$. These encoder–decoder pairs

[2]Wolfowitz's result was for γ ranging over a finite set, but it can be extended to infinite sets [24].

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 then 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 operates only 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 (4.12).

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

$$J(P(\gamma)) = \int_0^\infty B \log_2 \left(1 + \frac{\gamma P(\gamma)}{\bar{P}}\right) p(\gamma) d\gamma - \lambda \int_0^\infty P(\gamma) p(\gamma) d\gamma. \tag{4.13}$$

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

$$\frac{\partial J(P(\gamma))}{\partial P(\gamma)} = \left[\left(\frac{B/\ln(2)}{1 + \gamma P(\gamma)/\bar{P}}\right)\frac{\gamma}{\bar{P}} - \lambda\right] p(\gamma) = 0. \tag{4.14}$$

Solving for $P(\gamma)$ with the constraint that $P(\gamma) > 0$ yields the optimal power adaptation that maximizes (4.12) as

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} 1/\gamma_0 - 1/\gamma & \gamma \geq \gamma_0, \\ 0 & \gamma < \gamma_0, \end{cases} \tag{4.15}$$

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 used at time i only if $\gamma_0 \leq \gamma[i] < \infty$. Substituting (4.15) into (4.12) then yields the capacity formula:

$$C = \int_{\gamma_0}^\infty B \log_2 \left(\frac{\gamma}{\gamma_0}\right) p(\gamma) d\gamma. \tag{4.16}$$

The multiplexing nature of the capacity-achieving coding strategy indicates that (4.16) is achieved with a time-varying data rate, where the rate corresponding to the 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 (4.16): it can also be achieved by adapting the transmit power and sending at a fixed rate [25]. 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 SNR – in contrast to the case of receiver CSI only, where fading always decreases capacity.

Note that the optimal power allocation policy (4.15) depends on the fading distribution $p(\gamma)$ only through the cutoff value γ_0. This cutoff value is found from the power constraint. Specifically, rearranging the power constraint (4.11) and replacing the inequality with equality (since using the maximum available power will always be optimal) yields the power constraint

$$\int_0^\infty \frac{P(\gamma)}{\bar{P}} p(\gamma) d\gamma = 1. \tag{4.17}$$

If we now substitute the optimal power adaptation (4.15) into this expression then the cutoff value γ_0 must satisfy

$$\int_{\gamma_0}^\infty \left(\frac{1}{\gamma_0} - \frac{1}{\gamma}\right) p(\gamma) d\gamma = 1. \tag{4.18}$$

Observe that this expression depends only on the distribution $p(\gamma)$. The value for γ_0 must be found numerically [26] because no closed-form solutions exist for typical continuous distributions $p(\gamma)$.

Since γ is time varying, the maximizing power adaptation policy of (4.15) 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

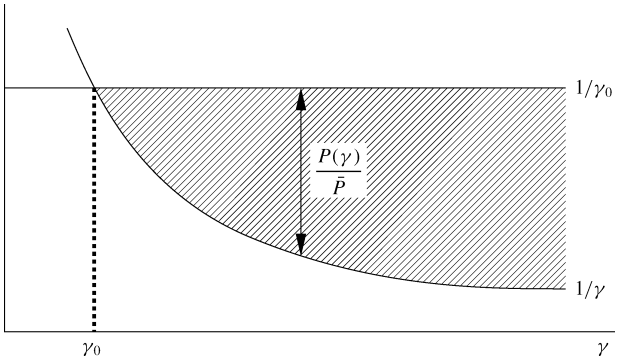

Figure 4.5: Optimal power allocation: water-filling.

channel conditions: when channel conditions are good (γ large), more power and a higher data rate are sent over the channel. As channel quality degrades (γ small), less power and rate are sent over the channel. If the instantaneous SNR falls below the cutoff value, the channel is not used. Adaptive modulation and coding techniques that follow this principle were developed in [27, 28] and are discussed in Chapter 9.

Note that the multiplexing argument sketching how capacity (4.12) is achieved applies to any power adaptation policy. That is, for any power adaptation policy $P(\gamma)$ with average power $\bar{P}$, the capacity

$$C = \int_0^\infty B \log_2 \left(1 + \frac{P(\gamma)\gamma}{\bar{P}}\right) p(\gamma) d\gamma \tag{4.19}$$

can be achieved with arbitrarily small error probability. Of course this capacity cannot exceed (4.12), 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 (4.12).

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 that 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 (4.18) given by

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

We first assume that all channel states are used to obtain γ_0 (i.e., we 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 the discrete power allocation formula to our channel model yields

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

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, this result is inconsistent because 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 the discrete power allocation becomes

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

Solving for γ_0 yields $\gamma_0 = .89$. Hence, by assuming that 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

$$\begin{aligned} C &= \sum_{i=2}^{3} B\log_2\left(\frac{\gamma_i}{\gamma_0}\right)p(\gamma_i) \\ &= 30000\left(.5\log_2\frac{83.33}{.89}+.4\log_2\frac{333.33}{.89}\right)=200.82\,\text{kbps}. \end{aligned}$$

Comparing with the results of Example 4.3 we see that this rate is only slightly higher than for the case of receiver CSI only, and it is still significantly below that of an AWGN channel with the same average SNR. This is because the average SNR for the channel in this example is relatively high: for low-SNR channels, capacity with flat fading can exceed that of the AWGN channel with the same average SNR if we take advantage of the rare times when the fading 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; that is, 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 $P(\gamma)/\bar{P} = \sigma/\gamma$, where σ equals the constant received SNR that can be maintained with the transmit power constraint (4.11). The constant σ thus satisfies $\int(\sigma/\gamma)p(\gamma)d\gamma = 1$, so $\sigma = 1/\mathbf{E}[1/\gamma]$.

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

$$C = B\log_2[1+\sigma] = B\log_2\left[1+\frac{1}{\mathbf{E}[1/\gamma]}\right]. \tag{4.20}$$

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 (4.20) 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. In Rayleigh fading, for example, $\mathbf{E}[1/\gamma]$ is infinite and thus the zero-outage capacity given by (4.20) is zero. Channel inversion is common in spread-spectrum systems with near–far interference imbalances [29]. It is also the simplest scheme to implement because 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

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

we have $C = 30000\log_2[1 + 1/.1272] = 94.43$ 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 that zero-outage capacity may be significantly smaller than Shannon capacity on a fading channel is the requirement of maintaining 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 non-outage channel states multiplied by the probability of non-outage. Outage capacity is achieved with a *truncated channel inversion* policy for power adaptation that compensates for fading only above a certain cutoff fade depth γ_0:

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} \sigma/\gamma & \gamma \geq \gamma_0, \\ 0 & \gamma < \gamma_0, \end{cases} \tag{4.21}$$

where γ_0 is based on the outage probability: $P_{\text{out}} = p(\gamma < \gamma_0)$. Since the channel is only used when $\gamma \geq \gamma_0$, the power constraint (4.11) yields $\sigma = 1/\mathbf{E}_{\gamma_0}[1/\gamma]$, where

$$\mathbf{E}_{\gamma_0}\left[\frac{1}{\gamma}\right] \triangleq \int_{\gamma_0}^{\infty} \frac{1}{\gamma} p(\gamma) d\gamma. \tag{4.22}$$

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

$$C(P_{\text{out}}) = B\log_2\left(1 + \frac{1}{\mathbf{E}_{\gamma_0}[1/\gamma]}\right) p(\gamma \geq \gamma_0). \tag{4.23}$$

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

$$C = \max_{\gamma_0} B\log_2\left(1 + \frac{1}{\mathbf{E}_{\gamma_0}[1/\gamma]}\right) p(\gamma \geq \gamma_0). \tag{4.24}$$

This maximum outage capacity will still be less than Shannon capacity (4.16) because 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 $C = B\log_2(1 + 1/\mathbf{E}_{\gamma_0}[1/\gamma])p(\gamma \geq \gamma_0) = 30000\log_2(1 + 138.88) \cdot .9 = 192.457$ kbps. 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 $C = B\log_2(1 + 1/\mathbf{E}_{\gamma_0}[1/\gamma])p(\gamma \geq \gamma_0) = 30000\log_2(1 + 833.33) \cdot .4 = 116.45$ 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 larger SNR occurs only 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 for improving 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 requires first 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 already given to obtain the capacity under diversity combining. The specific capacity formula used depends on the assumptions about channel side information; for example, in the case of perfect transmitter and receiver CSI the formula (4.16) would be used. For different CSI assumptions, capacity (under both maximal ratio and selection combining diversity) was computed in [26]. It was found that, as expected, the capacity with perfect transmitter and receiver CSI is greater than with receiver CSI only, which in turn is greater than with channel inversion. The performance gap of these different formulas decreases as the number of antenna branches increases. This trend is expected because a large number of antenna branches makes the channel look like an AWGN channel, 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 [30, 31, 32] indicating that the capacity of a fading channel with multiple inputs and outputs (a MIMO channel) is M times larger than 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 MIMO channels 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 (4.7), (4.16), (4.20), and (4.24) as a function of average received SNR for log-normal fading ($\sigma = 8\,\mathrm{dB}$ 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

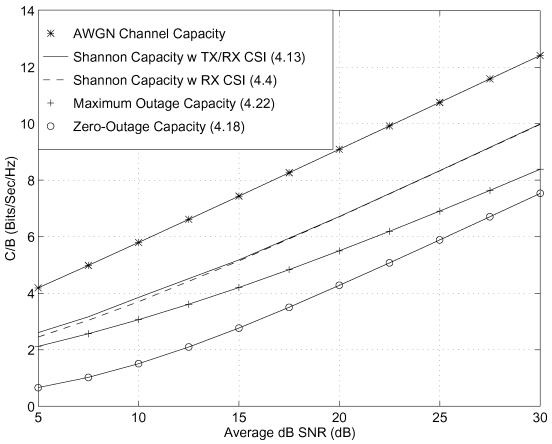

Figure 4.6: Capacity in log-normal fading.

average power is also shown for comparison. Note that the capacity in log-normal fading is plotted relative to average dB SNR ($\mu_{\rm dB}$), not average SNR in dB ($10\log_{10}\mu$): the relation between these values, as given by (2.45), is $10\log_{10}\mu = \mu_{\rm dB} + \sigma^2_{\rm dB}\ln(10)/20$.

Several observations in this comparison are worth noting. First, the figures show 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.

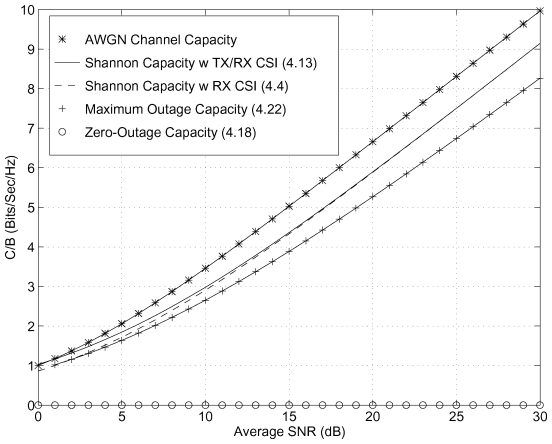

Figure 4.7: Capacity in Rayleigh fading.

That is because the AWGN channel always has the same low SNR, thereby limiting its capacity. A fading channel with this same low average SNR will occasionally have a high SNR, since the distribution has infinite range. Thus, if high power and rate are transmitted over the channel during these very infrequent large SNR values, the capacity will be greater 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 of the fading decreases (Rayleigh to Nakagami with $m = 2$), the capacity difference between the various adaptive

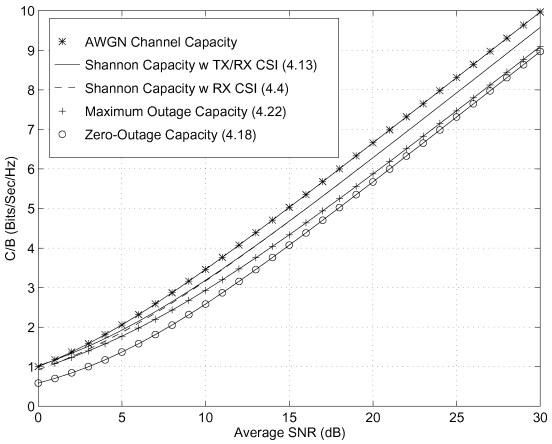

Figure 4.8: Capacity in Nakagami fading ($m = 2$).

policies also decreases, and their respective capacities approach that of the AWGN channel.

The difference between the capacity curves under transmitter and receiver CSI (4.16) and receiver CSI only (4.7) are negligible in all cases. Recalling that capacity under receiver CSI only (4.7) and under transmitter and receiver CSI without power adaptation (4.10) are the same, we conclude that, if the transmission rate is adapted relative to the channel, then 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 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 trade-off between capacity and complexity. The adaptive policy with transmitter and receiver side information requires more complexity in the transmitter (and 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 thus are 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 non-adaptive techniques for real systems. Achievable rates for adaptive trellis-coded MQAM have been investigated in [28], where a simple four-state trellis code combined with adaptive six-constellation MQAM modulation was shown to achieve rates within 7 dB of the Shannon capacity (4.12) 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 examine 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 like that of a flat fading channel but with the time axis replaced by the frequency axis. Then 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 to both the transmitter and receiver. The capacity of time-invariant channels under different assumptions about channel knowledge is discussed in [23, 25].

Let us first assume that $H(f)$ is *block fading*, so that frequency is divided into subchannels of bandwidth B with $H(f) = H_j$ constant over each subchannel, as shown in Figure 4.10.

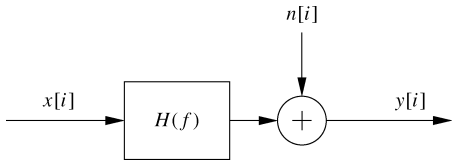

Figure 4.9: Time-invariant frequency-selective fading channel.

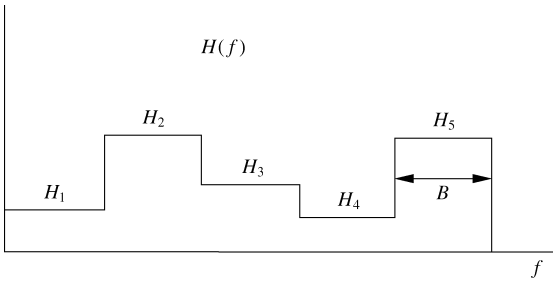

Figure 4.10: Block frequency-selective fading.

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 on each channel with power optimally allocated over all channels [5, 6]:

$$C = \sum_{\max P_j : \sum_j P_j \leq P} B \log_2 \left(1 + \frac{|H_j|^2 P_j}{N_0 B}\right). \tag{4.25}$$

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

$$\frac{P_j}{P} = \begin{cases} 1/\gamma_0 - 1/\gamma_j & \gamma_j \geq \gamma_0, \\ 0 & \gamma_j < \gamma_0, \end{cases} \tag{4.26}$$

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 P.[3] This optimal power allocation is illustrated in Figure 4.11. The cutoff value

[3] We define γ_j with respect to the total power P because of the normalization by P on the left hand side of (4.26).

is obtained by substituting the power adaptation formula into the power constraint, so γ_0 must satisfy

$$\sum_j \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_j} \right) = 1. \tag{4.27}$$

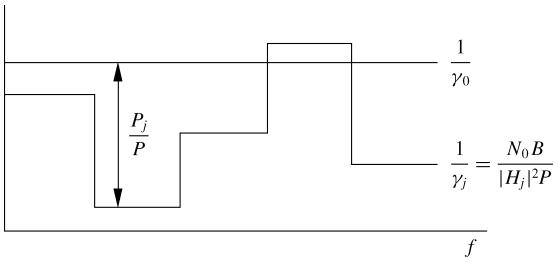

Figure 4.11: Water-filling in frequency-selective block fading.

The capacity then becomes

$$C = \sum_{j:\gamma_j \geq \gamma_0} B \log_2 \left(\frac{\gamma_j}{\gamma_0} \right). \tag{4.28}$$

This capacity is achieved by transmitting at different rates and powers over each subchannel. Multicarrier modulation uses the same technique in adaptive loading, as discussed in more detail in Section 12.3.

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

$$C = \max_{P(f): \int P(f) df \leq P} \int \log_2 \left(1 + \frac{|H(f)|^2 P(f)}{N_0} \right) df. \tag{4.29}$$

The expression inside the integral can be thought of as the incremental capacity associated with 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, Chap. 8.5]. An alternate proof [33] decomposes the channel into a parallel set using the discrete Fourier transform (DFT); the same premise is used in the discrete implementation of multicarrier modulation described in Section 12.4.

The power allocation over frequency, $P(f)$, that maximizes (4.29) is found via the Lagrangian technique. The resulting optimal power allocation is water-filling over frequency:

$$\frac{P(f)}{P} = \begin{cases} 1/\gamma_0 - 1/\gamma(f) & \gamma(f) \geq \gamma_0, \\ 0 & \gamma(f) < \gamma_0, \end{cases} \tag{4.30}$$

where $\gamma(f) = |H(f)|^2 P/N_0$. This results in channel capacity

$$C = \int_{f:\gamma(f) \geq \gamma_0} \log_2 \left(\frac{\gamma(f)}{\gamma_0} \right) df. \tag{4.31}$$

Example 4.7: Consider a time-invariant frequency-selective block fading channel that has three subchannels of bandwidth $B = 1$ MHz. The frequency responses associated with each subchannel are $H_1 = 1$, $H_2 = 2$, and $H_3 =$

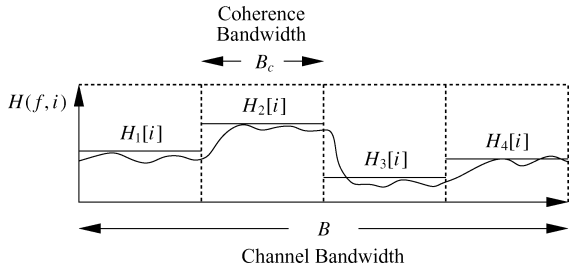

Figure 4.12: Channel division in frequency-selective fading.

3, respectively. The transmit power constraint is $P = 10$ mW and the noise PSD $N_0/2$ has $N_0 = 10^{-9}$ W/Hz. Find the Shannon capacity of this channel and the optimal power allocation that achieves this capacity. *Solution:* We first Find $\gamma_j = |H_j|^2 P/(N_0 B)$ for each subchannel, yielding $\gamma_1 = 10,\ \gamma_2 = 40$, and $\gamma_3 = 90$. The cutoff γ_0 must satisfy (4.27). Assuming that 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 $H(f) = H(f,\ i)$; that is, 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 – because of the effects of self-interference (ISI). In the case of transmitter and receiver side information, the optimal adaptation scheme must consider (a) the effect of the channel on the past sequence of transmitted bits and (b) how the ISI resulting from these bits will affect future transmissions [34]. The capacity of time-varying frequency-selective fading channels is in general unknown, but there do exist upper and lower bounds as well as limiting formulas [34, 35].

We can approximate channel capacity in time-varying frequency-selective fading by taking the channel bandwidth B of interest and then dividing 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 $\bar{P}_j$ that we allocate to each subchannel, subject to a total power constraint $\bar{P}$. 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 and averaged over both time and frequency:

$$C = \max_{\{\bar{P}_j\}:\sum_j \bar{P}_j \le \bar{P}} \sum_j C_j(\bar{P}_j), \tag{4.32}$$

where $C_j(\bar{P}_j)$ is the capacity of the flat fading subchannel with average power $\bar{P}_j$ and bandwidth B_c given by (4.16), (4.7), (4.20), or (4.24) for Shannon capacity under different side information and power allocation policies. We can also define $C_j(\bar{P}_j)$ 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 upperbounds capacity under any other side information assumptions or suboptimal power allocation strategies. We know that if we fix the average power per subchannel then the optimal power adaptation follows a water-filling formula. We expect that the optimal average power to be allocated to each subchannel will also follow a water-filling formula, 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 $\gamma_j[i] = |H_j[i]|^2\bar{P}/(N_0B)$ to be the instantaneous SNR on the jth subchannel at time i assuming the total power $\bar{P}$ 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):

$$C = \max_{P_j(\gamma_j):\sum_j \int_0^\infty P_j(\gamma_j)p(\gamma_j)d\gamma_j \le \bar{P}} \sum_j \int_0^\infty B_c \log_2\left(1 + \frac{P_j(\gamma_j)\gamma_j}{\bar{P}}\right) p(\gamma_j)d\gamma_j. \tag{4.33}$$

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

$$\begin{aligned} &J(P_j(\gamma_j)) \\ &\quad = \sum_j \int_0^\infty B_c \log_2\left(1 + \frac{P_j(\gamma_j)\gamma_j}{\bar{P}}\right) p(\gamma_j)d\gamma_j - \lambda \sum_j \int_0^\infty P_j(\gamma_j)p(\gamma_j)d\gamma_j. \end{aligned} \tag{4.34}$$

Note that (4.34) is similar to the Lagrangian (4.13) for the flat fading channel 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:

$$\frac{\partial J(P_j(\gamma_j))}{\partial P_j(\gamma_j)} = \left[\left(\frac{B_c/\ln(2)}{1 + \gamma_j P_j(\gamma_j)/\bar{P}}\right)\frac{\gamma_j}{\bar{P}} - \lambda\right] p(\gamma_j) = 0. \tag{4.35}$$

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

$$\frac{P_j(\gamma_j)}{\bar{P}} = \begin{cases} 1/\gamma_0 - 1/\gamma_j & \gamma_j \ge \gamma_0, \\ 0 & \gamma_j < \gamma_0, \end{cases} \tag{4.36}$$

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

$$\sum_j \int_0^\infty P_j(\gamma_j)p(\gamma_j)d\gamma_j = \bar{P}. \tag{4.37}$$

Thus, the optimal power allocation (4.36) is a two-dimensional water-filling with a common cutoff value γ_0. Dividing the constraint (4.37) by $\bar{P}$ and substituting into the optimal power allocation (4.36), we get that γ_0 must satisfy

$$\sum_j \int_{\gamma_0}^\infty \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_j}\right) p(\gamma_j)d\gamma_j = 1. \tag{4.38}$$

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 (4.37) into the capacity expression (4.33) yields

$$C = \sum_{j} \int_{\gamma_0}^{\infty} B_c \log_2 \left(\frac{\gamma_j}{\gamma_0} \right) p(\gamma_j) d\gamma_j. \tag{4.39}$$

Chapter 4 Problems

1. Capacity in AWGN is given by $C = B\log_2(1 + P/(N_0B))$, where P is the received signal power, B is the signal bandwidth, and $N_0/2$ is the noise PSD. Find capacity in the limit of infinite bandwidth $B \to \infty$ as a function of P.

2. Consider an AWGN channel with bandwidth 50 MHz, received signal power 10 mW, and noise PSD $N_0/2$ where $N_0 = 2 \cdot 10^{-9}$ 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. Now, 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 and where, for a fixed transmit power $\bar{P}$, the received SNR is one of six values: $\gamma_1 = 20\,\text{dB},\ \gamma_2 = 15\,\text{dB},\ \gamma_3 = 10\,\text{dB},\ \gamma_4 = 5\,\text{dB},\ \gamma_5 = 0\,\text{dB}$, and $\gamma_6 = -5\,\text{dB}$. The probabilities associated with each state are $p_1 = p_6 = .1,\ p_2 = p_4 = .15$, and $p_3 = p_5 = .25$. Assume that only the receiver has CSI.

   (a) Find the Shannon capacity of this channel.

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

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

   (a) Find the optimal power adaptation policy $P[i]/\bar{P}$ for this channel and its corresponding Shannon capacity per unit hertz (C/B).

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

   (c) Find the truncated channel inversion power adaptation policy for this channel and associated outage capacity per unit bandwidth for three different outage probabilities: $P_{\text{out}} = .1,\ P_{\text{out}} = .25$, 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 $d_0 = 100\,\mathrm{m}$ 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 $d = 1000\,\mathrm{m}$ from the transmitter, with an average transmit power constraint of $P_t = 100\,\mathrm{mW}$ and a receiver noise power of .1 mW. Assume that both transmitter and receiver have CSI.

   (a) Compute the distribution of the received SNR.

   (b) Derive the optimal power adaptation 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. In this problem we explore power allocation under a fixed energy constraint to maximize the number of bits transmitted. Consider a cellular system with a mobile receiver. The receiver stops at point A (100m from the cell phone tower) for 10 minutes and point B (500m from the cell phone tower) for 20 minutes. Assume the signal from the cell phone tower to the receiver only experiences path loss under the simplified path loss model with $\gamma = 3,\ d_0 = 1$m, and $K = 1$ along with AWGN, that the channel bandwidth is 20MHz, and the receiver noise PSD is $N_0 = 5 \times 10^{-14}$mW/Hz. Also, assume a fixed transmit power of 1mW.

   (a) What is the SNR received at points A and B?

   (b) What is total amount of energy (in Joules) that is transmitted during the two stops.

   (c) Assuming the transmitter sends data with a rate equal to the Shannon capacity of the channel at each of the stops, what is the total number of bits received?

   (d) Suppose now that the cell phone tower does not send at a fixed transmit power but fixes the total number of Joules it will transmit to the receiver at each of the two stops. Using the total amount of energy derived in part (b), determine the optimal allocation of this total energy at each of the stops to maximize the number of bits received. Assume that the transmitter sends data at a rate equal to the Shannon capacity with SNR based on your optimal energy allocation. How many bits are received with this optimal energy allocation and how does it compare to your answer in part (c)?

8. In this problem we explore why channel inversion leads to zero capacity in Rayleigh fading. Assume a Rayleigh fading channel with $\overline{\gamma} = 1$ so the receiver power has an exponential distribution $p(\gamma) = e^{-\gamma}$.

   (a) Show using a first-order Taylor series expansion of $e^{-\gamma}$ around $\gamma = 0$ that $p(\gamma) = e^{-\gamma} > 1 - \gamma$.

   (b) We will now show that $\mathbf{E}\left[\frac{1}{\gamma}\right]$ is unbounded by taking the expectation for $\gamma \in (0, c)$ for some small constant c. Fix this small constant c and a smaller constant $\epsilon < c$. Then compute the following lower bound for $\mathbf{E}\left[\frac{1}{\gamma}\right]$ by taking the expectation over the interval (ϵ, c) and using the lower bound on $p(\gamma)$ from part (a):
   $$\mathbf{E}\left[\frac{1}{\gamma}\right] > \int_\epsilon^c \frac{1}{\gamma}(1-\gamma)d\gamma.$$
   Show that as $\epsilon \to 0$, the above expression becomes unbounded. This suggests that the transmit power needed to maintain a constant received power by inverting Rayleigh fading is unbounded.

9. 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 $\bar{\gamma} = 10$ dB. Assume a channel bandwidth of 10 MHz.

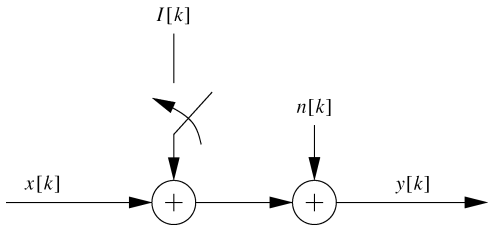

Figure 4.13: Interference channel for Problem 4.8.

(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 when only the receiver knows $\gamma[i]$.

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

(f) Repeat parts (b), (c), and (d) – that is, 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 $\bar{\gamma} = -5$ 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 behavior occurs.

10. This problem illustrates the capacity gains that can be obtained from interference estimation and also how a malicious jammer can wreak havoc on link performance. Consider the interference channel depicted in Figure 4.13. 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., switch up) with probability .75. The average transmit power is 10 mW, the noise PSD has $N_0 = 10^{-8}\,\mathrm{W/Hz}$, the channel bandwidth B is 10 kHz (receiver noise power is N_0B), 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 has 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?

11. Consider the malicious interferer of Problem 4.8. 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 because the transmitter sacrifices some power to cancel out the interference. Which strategy results in higher capacity? *Note*: There is a third strategy, in which 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, as described in Chapter 14.

12. 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 (4.26).

13. 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 $f > 0$ and $H(-f) = H(f)$. For a transmit power of 10 mW and a noise PSD of $.001\,\mu$W per Hertz, find the optimal power allocation and corresponding Shannon capacity of this channel.

14. 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 (4.30).

15. Consider a frequency-selective fading channel with total bandwidth 12 MHz and coherence bandwidth $B_c = 4$ MHz. Divide the total bandwidth into three 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 $\mathbf{E}[|H_1[i]|^2] = 1,\ \mathbf{E}[|H_2[i]|^2] = .5$, and $\mathbf{E}[|H_3[i]|^2] = .125$. Assume a total transmit power of 30 mW and a receiver noise PSD with $N_0 = .001\,\mu$W/Hz.

    (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 $H_j[i],\ j = 1, 2, 3$.

    (b) Compare the capacity derived in 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.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, pp. 379–423, 623–56, 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*, University of Illinois Press, Urbana, 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. 933–46, May 2000.

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

[6] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 2nd Ed, 2006.

[7] Y. Polyanskiy, H. V. Poor and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[8] C. Berrou, A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes", *IEEE Transactions on Commununications*, vol. 44, no. 10, pp. 1261-1271, Oct. 1996.

[9] R.G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA: MIT Press, 1963 (Sc.D. MIT, 1960).

[10] R. Urbanke, B. Rimoldi, "Lattice codes can achieve capacity on the AWGN channel", *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 273-278, Jan. 1998.

[11] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels", *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051-3073, July 2009.

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

[13] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, pp. 18–26, January 1991.

[14] 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.

[15] 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. 2426–67, October 2003.

[16] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inform. Theory*, pp. 868–86, May 1996.

[17] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channel," *IEEE Trans. Inform. Theory*, pp. 1277–90, November 1989.

[18] T. Holliday, A. Goldsmith, and P. Glynn, "Capacity of finite state Markov channels with general inputs," *Proc. IEEE Internat. Sympos. Inform. Theory*, p. 289, July 2003.

[19] R. J. McEliece and W. E. Stark, "Channels with block interference," *IEEE Trans. Inform. Theory*, pp. 44–53, January 1984.

[20] A. Lapidoth and S. Shamai, "Fading channels: How perfect need 'perfect side information' be?" *IEEE Trans. Inform. Theory*, pp. 1118–34, November 1997.

[21] G. J. Foschini, D. Chizhik, M. Gans, C. Papadias, and R. A. Valenzuela, "Analysis and performance of some basic space-time architectures," *IEEE J. Sel. Areas Commun.*, pp. 303–20, April 2003.

[22] W. L. Root and P. P. Varaiya, "Capacity of classes of Gaussian channels," *SIAM J. Appl. Math.*, pp. 1350–93, November 1968.

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

[24] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, pp. 1986–92, November 1997.

[25] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Trans. Inform. Theory*, pp. 2007–19, September 1999.

[26] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity combining techniques," *IEEE Trans. Veh. Tech.* pp. 1165–81, July 1999.

[27] S.-G. Chua and A. J. Goldsmith, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, pp. 1218–30, October 1997.

[28] S.-G. Chua and A. J. Goldsmith, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, pp. 595–602, May 1998.

[29] 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. 303–12, May 1991.

[30] G. J. Foschini, "Layered space-time architecture for wireless communication in fading environments when using multi-element antennas," *Bell System Tech. J.*, pp. 41–59, Autumn 1996.

[31] E. Teletar, "Capacity of multi-antenna Gaussian channels," AT&T Bell Labs Intemal Tech. Memo, June 1995.

[32] G. J. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, pp. 311–35, March 1998.

[33] W. Hirt and J. L. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inform. Theory*, pp. 380–8, May 1988.

[34] A. Goldsmith and M. Medard, "Capacity of time-varying channels with channel side information," *IEEE Trans. Inform. Theory* Vol. 53, No. 3, pp. 881-899, March 2007.

[35] S. Diggavi, 'Analysis of multicarrier transmission in time-varying channels," *Proc. IEEE Internat. Conf. Commun*., pp. 1191–5, June 1997.

Chapter 5

Digital Modulation and Detection

The advances over the last several decades in hardware and digital signal processing have made digital transceivers much cheaper, faster, and more power efficient than analog transceivers. More importantly, digital modulation offers a number of other advantages over analog modulation, including higher spectral efficiency, powerful error correction techniques, resistance to channel impairments, more efficient multiple access strategies, and better security and privacy. Specifically, high-level digital modulation techniques such as MQAM allow much more efficient use of spectrum than is possible with analog modulation. Advances in coding and coded modulation applied to digital signaling make the signal much less susceptible to noise and fading, and equalization or multicarrier techniques can be used to mitigate intersymbol interference (ISI). Spread-spectrum techniques applied to digital modulation can simultaneously remove or combine multipath, resist interference, and detect multiple users. Finally, digital modulation is much easier to encrypt, resulting in a higher level of security and privacy for digital systems. For all these reasons, systems currently being built or proposed for wireless applications are all digital systems.

Digital modulation and detection consist of transferring information in the form of bits over a communication channel. The bits are binary digits taking on the values of either 1 or 0. These information bits are derived from the information source, which may be a digital source or an analog source that has been passed through an A/D converter. Both digital and A/D-converted analog sources may be compressed to obtain the information bit sequence. Digital modulation consists of mapping the information bits into an analog signal for transmission over the channel. Detection consists of estimating the original bit sequence based on the signal received over the channel. The main considerations in choosing a particular digital modulation technique are high data rate; high spectral efficiency (minimum bandwidth occupancy); high power efficiency (minimum required transmit power); robustness to channel impairments (minimum probability of bit error); and low power/cost implementation. Often these are conflicting requirements, and the choice of modulation is based on finding the technique that achieves the best trade-off between them.

There are two main categories of digital modulation: amplitude/phase modulation and frequency modulation. Since frequency modulation typically has a constant signal envelope and is generated using nonlinear techniques, this modulation is also called *constant envelope modulation* or *nonlinear modulation*, and amplitude/phase modulation is also called *linear modulation*. Linear modulation generally has better spectral properties than nonlinear modulation, since nonlinear processing leads to spectral broadening. However, amplitude and phase modulation embeds the information bits into the amplitude or phase of the transmitted signal, which is more susceptible to variations from fading and interference. In addition, amplitude and phase modulation techniques typically require linear amplifiers, which are more expensive and less power efficient than the nonlinear amplifiers that can be used with nonlinear modulation. Thus, the trade-off between linear versus nonlinear modulation is one of better spectral efficiency for the former technique and better power efficiency and resistance to channel impairments for the latter. Once the modulation technique is determined, the constellation size must be chosen. Modulations with large

constellations have higher data rates for a given signal bandwidth, but they are more susceptible to noise, fading, and hardware imperfections. Finally, some demodulators require a coherent phase reference with respect to the transmitted signal. Obtaining this coherent reference may be difficult or significantly increase receiver complexity. Thus, modulation techniques that do not require a coherent phase reference in the receiver are desirable.

We begin this chapter with a general discussion of signal space concepts. These concepts greatly simplify the design and analysis of modulation and demodulation techniques by mapping infinite-dimensional signals to a finite-dimensional vector space. The general principles of signal space analysis will then be applied to the analysis of amplitude and phase modulation techniques, including pulse amplitude modulation (PAM), phase-shift keying (PSK), and quadrature amplitude modulation (QAM). We will also discuss constellation shaping and quadrature offset techniques for these modulations, as well as differential encoding to avoid the need for a coherent phase reference. We then describe frequency modulation techniques and their properties, including frequency-shift keying (FSK), minimum-shift keying (MSK), and continuous-phase FSK (CPFSK). Both coherent and noncoherent detection of these techniques will be discussed. Pulse-shaping techniques to improve the spectral properties of the modulated signals will also be covered, along with issues associated with carrier phase recovery and symbol synchronization.

5.1 Signal Space Analysis

Digital modulation encodes a bit stream of finite length into one of several possible transmitted signals. Intuitively, the receiver minimizes the probability of detection error by decoding the received signal as the signal in the set of possible transmitted signals that is "closest" to the one received. Determining the distance between the transmitted and received signals requires a metric for the distance between signals. By representing signals as projections onto a set of basis functions, we obtain a one-to-one correspondence between the set of transmitted signals and their vector representations. Thus, we can analyze signals in finite-dimensional vector space instead of infinite-dimensional function space, using classical notions of distance for vector spaces. In this section we show how digitally modulated signals can be represented as vectors in an appropriately defined vector space and how optimal demodulation methods can be obtained from this vector-space representation. This general analysis will then be applied to specific modulation techniques in later sections.

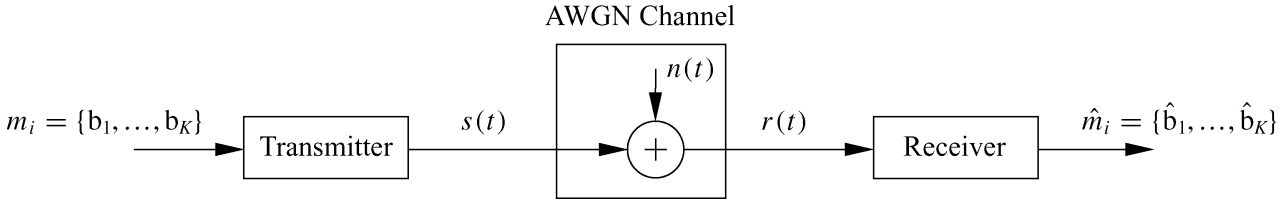

Figure 5.1: Communication system model.

5.1.1 Signal and System Model

Consider the communication system model shown in Figure 5.1. Every T seconds, the system sends $K = \log_2 M$ bits of information through the channel for a data rate of $R = K/T$ bits per second (bps). There are $M = 2^K$ possible sequences of K bits, and we say that each bit sequence of length K comprises a message $m_i = \{b_1, \ldots, b_K\} \in \mathcal{M}$, where $\mathcal{M} = \{m_1, \ldots, m_M\}$ is the set of all such messages. The messages have probability p_i of being selected for transmission, where $\sum_{i=1}^{M} p_i = 1$.

Suppose message m_i is to be transmitted over the channel during the time interval $[0, T)$. Because the channel is analog, the message must be embedded into an analog signal for channel transmission. Hence, each message

$m_i \in \mathcal{M}$ is mapped to a unique analog signal $s_i(t) \in S = \{s_1(t), \ldots, s_M(t)\}$, where $s_i(t)$ is defined on the time interval $[0, T)$ and has energy

$$E_{s_i} = \int_0^T s_i^2(t)dt, \quad i = 1, \ldots, M. \tag{5.1}$$

Since each message represents a bit sequence it follows that each signal $s_i(t) \in S$ also represents a bit sequence, and detection of the transmitted signal $s_i(t)$ at the receiver is equivalent to detection of the transmitted bit sequence. When messages are sent sequentially, the transmitted signal becomes a sequence of the corresponding analog signals over each time interval $[kT, (k+1)T) : s(t) = \sum_k s_i(t-kT)$, where $s_i(t)$ is a baseband or passband analog signal corresponding to the message m_i designated for the transmission interval $[kT, (k+1)T)$. This is illustrated in Figure 5.2, where we show the transmitted signal $s(t) = s_1(t) + s_2(t-T) + s_1(t-2T) + s_1(t-3T)$ corresponding to the string of messages m_1, m_2, m_1, m_1 with message m_i mapped to signal $s_i(t)$.

In the model of Figure 5.1, the transmitted signal is sent through an AWGN channel, where a white Gaussian noise process $n(t)$ of power spectral density $N_0/2$ is added to form the received signal $r(t) = s(t) + n(t)$. Given $r(t)$, the receiver must determine the best estimate of which $s_i(t) \in S$ was transmitted during each transmission interval $[kT, (k+1)T)$. This best estimate for $s_i(t)$ is mapped to a best estimate of the message $m_i \in \mathcal{M}$ and the receiver then outputs this best estimate $\hat{m} = \{\hat{\mathrm{b}}_1, \ldots, \hat{\mathrm{b}}_K\} \in \mathcal{M}$ of the transmitted bit sequence.

The goal of the receiver design in estimating the transmitted message is to minimize the probability of message error,

$$P_e = \sum_{i=1}^{M} p(\hat{m} \neq m_i | m_i \text{ sent}) p(m_i \text{ sent}), \tag{5.2}$$

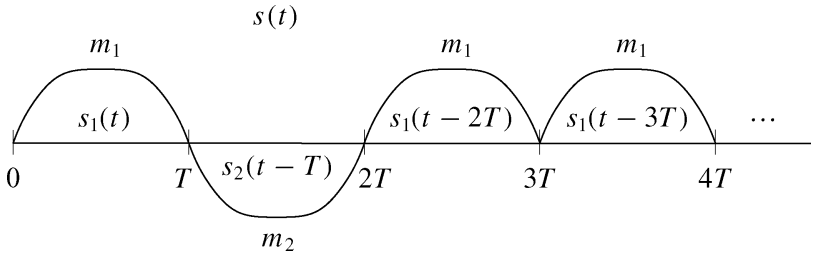

Figure 5.2: Transmitted signal for a sequence of messages.

over each time interval $[kT, (k+1)T)$. By representing the signals $\{s_i(t) : i = 1, \ldots, M\}$ geometrically, we can solve for the optimal receiver design in AWGN based on a minimum distance criterion. Note that, as we saw in previous chapters, wireless channels typically have a time-varying impulse response in addition to AWGN. We will consider the effect of an arbitrary channel impulse response on digital modulation performance in Chapter 6;methods to combat this performance degradation are discussed in Chapters 11–13.

5.1.2 Geometric Representation of Signals

The basic premise behind a geometrical representation of signals is the notion of a basis set. Specifically, using a Gram–Schmidt orthogonalization procedure [1, 2], it can be shown that any set of M real signals $S = \{s_1(t), \ldots, s_M(t)\}$ defined on $[0, T)$ with finite energy can be represented as a linear combination of $N \leq M$ real orthonormal basis functions $\{\phi_1(t), \ldots, \phi_N(t)\}$. We say that these basis functions *span* the set S. Thus, we can write each $s_i(t) \in S$ in terms of its *basis function representation* as

$$s_i(t) = \sum_{j=1}^{N} s_{ij}\phi_j(t), \quad 0 \le t < T, \tag{5.3}$$

where

$$s_{ij} = \int_0^T s_i(t)\phi_j(t)dt \tag{5.4}$$

is a real coefficient representing the projection of $s_i(t)$ onto the basis function $\phi_j(t)$ and

$$\int_0^T \phi_i(t)\phi_j(t)dt = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{5.5}$$

If the signals $\{s_i(t)\}$ are linearly independent then $N = M$, otherwise $N < M$. Moreover, the minimum number N of orthogonal basis functions needed to represent any signal $s_i(t)$ of duration T and bandwidth B is roughly $2BT$ [3, Chap. 5.3]. The signal $s_i(t)$ thus occupies $2BT$ orthogonal dimensions.

For linear passband modulation techniques, the basis set consists of the sine and cosine functions:

$$\phi_1(t) = \sqrt{\frac{2}{T}}\cos(2\pi f_c t) \tag{5.6}$$

and

$$\phi_2(t) = \sqrt{\frac{2}{T}}\sin(2\pi f_c t). \tag{5.7}$$

The $\sqrt{2/T}$ factor is needed for normalization so that $\int_0^T \phi_i^2(t)\,dt = 1, i = 1, 2$. In fact, with these basis functions we get only an approximation to (5.5), since

$$\int_0^T \phi_1^2(t)\,dt = \frac{2}{T}\int_0^T .5[1 + \cos(4\pi f_c t)]\,dt = 1 + \frac{\sin(4\pi f_c T)}{4\pi f_c T}. \tag{5.8}$$

The numerator in the second term of (5.8) is bounded by 1 and for $f_c T \gg 1$ the denominator of this term is very large. Hence this second term can be neglected. Similarly,

$$\int_0^T \phi_1(t)\phi_2(t)\,dt = \frac{2}{T}\int_0^T .5\sin(4\pi f_c t)\,dt = \frac{-\cos(4\pi f_c T)}{4\pi f_c T} \approx 0, \tag{5.9}$$

where the approximation is taken as an equality for $f_c T \gg 1$.

With the basis set $\phi_1(t) = \sqrt{2/T}\cos(2\pi f_c t)$ and $\phi_2(t) = \sqrt{2/T}\sin(2\pi f_c t)$, the basis function representation (5.3) corresponds to the equivalent lowpass representation of $s_i(t)$ in terms of its in-phase and quadrature components:

$$s_i(t) = s_{i1}\sqrt{\frac{2}{T}}\cos(2\pi f_c t) + s_{i2}\sqrt{\frac{2}{T}}\sin(2\pi f_c t). \tag{5.10}$$

Note that the carrier basis functions may have an initial phase offset ϕ_0. The basis set may also include a baseband pulse-shaping filter $g(t)$ to improve the spectral characteristics of the transmitted signal:

$$s_i(t) = s_{i1}g(t)\cos(2\pi f_c t) + s_{i2}g(t)\sin(2\pi f_c t). \tag{5.11}$$

In this case the pulse shape $g(t)$ must maintain the orthonormal properties (5.5) of basis functions; that is, we must have

$$\int_0^T g^2(t)\cos^2(2\pi f_c t)\,dt = 1 \tag{5.12}$$

and

$$\int_0^T g^2(t)\cos(2\pi f_c t)\sin(2\pi f_c t)\,dt = 0, \tag{5.13}$$

where the equalities may be approximations for $f_c T \gg 1$ as in (5.8) and (5.9). If the bandwidth of $g(t)$ satisfies $B \ll f_c$ then $g^2(t)$ is roughly constant over $T_c = 1/f_c$, so (5.13) is approximately true because the sine and cosine functions are orthogonal over one period T_c. The simplest pulse shape that satisfies (5.12) and (5.13) is the rectangular pulse shape $g(t) = \sqrt{2/T}, 0 \le t < T$.

Example 5.1: Binary phase-shift keying (BPSK) modulation transmits the signal $s_1(t) = \alpha\cos(2\pi f_c t)$, $0 \le t \le T$, to send a 1-bit and the signal $s_2(t) = -\alpha\cos(2\pi f_c t)$, $0 \le t \le T$, to send a 0-bit. Find the set of orthonormal basis functions and coefficients $\{s_{ij}\}$ for this modulation.
Solution: There is only one basis function for $s_1(t)$ and $s_2(t)$:

$$\phi(t) = \sqrt{2/T}\cos(2\pi f_c t),$$

where the $\sqrt{2/T}$ is needed for normalization. The coefficients are then given by $s_1 = \alpha\sqrt{T/2}$ and $s_2 = -\alpha\sqrt{T/2}$.

Let $\mathbf{s}_i = (s_{i1}, \ldots, s_{iN}) \in \mathbb{R}^N$ be the vector of coefficients $\{s_{ij}\}$ in the basis representation of $s_i(t)$. We call $\mathbf{s}_i$ the *signal constellation point* corresponding to the signal $s_i(t)$. The *signal constellation* consists of all constellation points $\{\mathbf{s}_1, \ldots, \mathbf{s}_M\}$. Given the basis functions $\{\phi_1(t), \ldots, \phi_N(t)\}$, there is a one-to-one correspondence between the transmitted signal $s_i(t)$ and its constellation point $\mathbf{s}_i$. Specifically, $s_i(t)$ can be obtained from $\mathbf{s}_i$ by (5.3) and $\mathbf{s}_i$ can be obtained from $s_i(t)$ by (5.4). Thus, it is equivalent to characterize the transmitted signal by $s_i(t)$ or $\mathbf{s}_i$. The representation of $s_i(t)$ in terms of its constellation point $\mathbf{s}_i \in \mathbb{R}^N$ is called its *signal space representation*, and the vector space containing the constellation is called the *signal space*. A two-dimensional signal space is illustrated in Figure 5.3, where we show $\mathbf{s}_i \in \mathbb{R}^2$ with the ith axis of $\mathbb{R}^2$ corresponding to the basis function $\phi_i(t)$, $i = 1, 2$. With this signal space representation we can analyze the infinite-dimensional functions $s_i(t)$ as vectors $\mathbf{s}_i$ in finite-dimensional vector space $\mathbb{R}^2$. This greatly simplifies the analysis of system performance as well as the derivation of optimal receiver design. Signal space representations for common modulation techniques like MPSK and MQAM are two-dimensional (corresponding to the in-phase and quadrature basis functions) and will be given later in the chapter.

In order to analyze signals via a signal space representation, we require a few definitions for vector characterization in the vector space $\mathbb{R}^N$. The length of a vector in $\mathbb{R}^N$ is defined as

$$\|\mathbf{s}_i\| \triangleq \sqrt{\sum_{j=1}^{N} s_{ij}^2}. \tag{5.14}$$

The distance between two signal constellation points $\mathbf{s}_i$ and $\mathbf{s}_k$ is thus

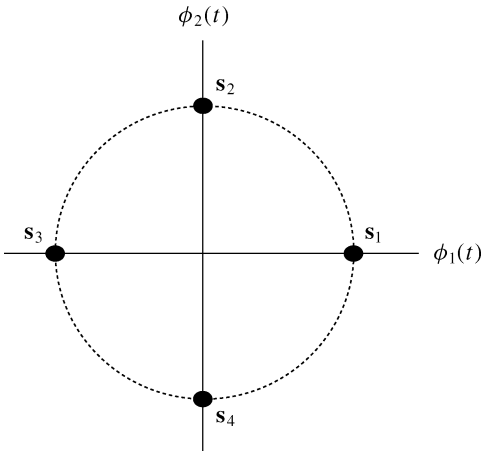

Figure 5.3: Signal space representation.

$$\|\mathbf{s}_i - \mathbf{s}_k\| = \sqrt{\sum_{j=1}^{N}(s_{ij} - s_{kj})^2} = \sqrt{\int_0^T (s_i(t) - s_k(t))^2 dt}, \tag{5.15}$$

where the second equality is obtained by writing $s_i(t)$ and $s_k(t)$ in their basis representation (5.3) and using the orthonormal properties of the basis functions. Finally, the inner product $\langle s_i(t),\, s_k(t)\rangle$ between two real signals $s_i(t)$ and $s_k(t)$ on the interval $[0, T]$ is

$$\langle s_i(t), s_k(t)\rangle = \int_0^T s_i(t) s_k(t)\, dt. \tag{5.16}$$

Similarly, the inner product $\langle \mathbf{s}_i, \mathbf{s}_k\rangle$ between two constellation points is

$$\langle \mathbf{s}_i, \mathbf{s}_k\rangle = \mathbf{s}_i \mathbf{s}_k^T = \int_0^T s_i(t) s_k(t)\, dt = \langle s_i(t), s_k(t)\rangle, \tag{5.17}$$

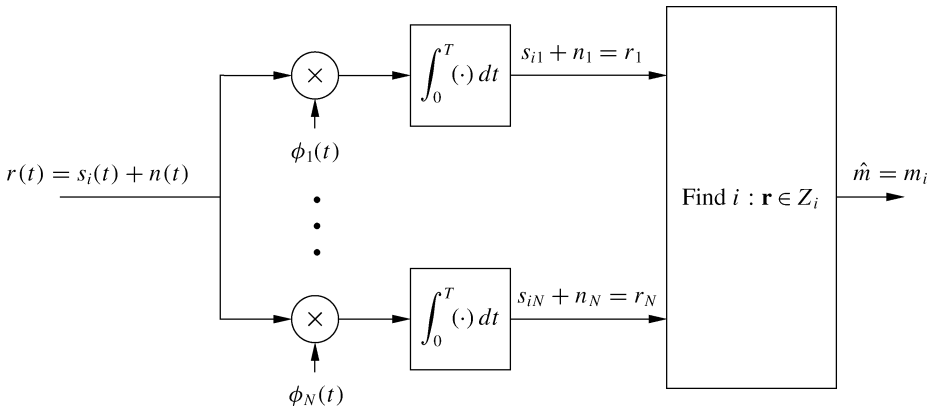

Figure 5.4: Receiver structure for signal detection in AWGN.

where the equality between the vector inner product and the corresponding signal inner product follows from the basis representation of the signals (5.3) and the orthonormal property of the basis functions (5.5). We say that two

signals are *orthogonal* if their inner product is zero. Thus, by (5.5), the basis functions are orthogonal functions.

5.1.3 Receiver Structure and Sufficient Statistics

Given the channel output $r(t) = s_i(t) + n(t)$, $0 \le t < T$, we now investigate the receiver structure to determine which constellation point $\mathbf{s}_i$ (or, equivalently, which message m_i) was sent over the time interval $[0, T)$. A similar procedure is done for each time interval $[kT, (k+1)T)$. We would like to convert the received signal $r(t)$ over each time interval into a vector, as this would allow us to work in finite-dimensional vector space when estimating the transmitted signal. However, this conversion should not be allowed to compromise the estimation accuracy. We now study a receiver that converts the received signal to a vector without compromising performance. Consider the receiver structure shown in Figure 5.4, where

$$s_{ij} = \int_0^T s_i(t)\phi_j(t)\,dt \tag{5.18}$$

and

$$n_j = \int_0^T n(t)\phi_j(t)\,dt. \tag{5.19}$$

We can rewrite $r(t)$ as

$$\sum_{j=1}^{N}(s_{ij} + n_j)\phi_j(t) + n_r(t) = \sum_{j=1}^{N} r_j\phi_j(t) + n_r(t); \tag{5.20}$$

here $r_j = s_{ij} + n_j$ and $n_r(t) = n(t) - \sum_{j=1}^{N} n_j\phi_j(t)$ denotes the "remainder" noise, which is the component of the noise orthogonal to the signal space. If we can show that the optimal detection of the transmitted signal constellation point $\mathbf{s}_i$ given received signal $r(t)$ does not make use of the remainder noise $n_r(t)$, then the receiver can make its estimate $\hat{m}$ of the transmitted message m_i as a function of $\mathbf{r} = (r_1, \ldots, r_N)$ alone. In other words, $\mathbf{r} = (r_1, \ldots, r_N)$ is a *sufficient statistic* for $r(t)$ in the optimal detection of the transmitted messages.

It is intuitively clear that the remainder noise $n_r(t)$ should not help in detecting the transmitted signal $s_i(t)$, since its projection onto the signal space is zero. This is illustrated in Figure 5.5, where we assume the transmitted signal lies in a space spanned by the basis set $(\phi_1(t), \phi_2(t))$ while the remainder noise lies in a space spanned by the basis function $\phi_{n_r}(t)$, which is orthogonal to $\phi_1(t)$ and $\phi_2(t)$. Specifically, the remainder noise in Figure 5.5 is represented by n_r, where $n_r(t) = n_r\phi_{n_r}(t)$. The received signal is represented by $\mathbf{r} + n_r$, where $\mathbf{r} = (r_1, r_2)$ with $r(t) - n_r(t) = r_1\phi_1(t) + r_2\phi_2(t)$. From the figure it appears that projecting $\mathbf{r} + n_r$ onto $\mathbf{r}$ will not compromise the detection of which constellation $\mathbf{s}_i$ was transmitted, since n_r lies in a space orthogonal to the space in which $\mathbf{s}_i$ lies. We now proceed to show mathematically why this intuition is correct.

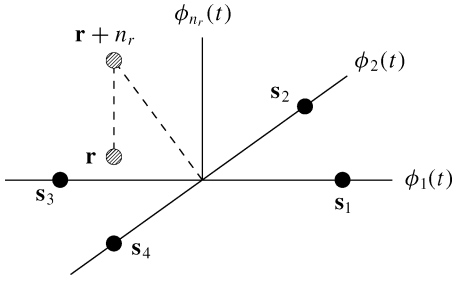

Figure 5.5: Projection of received signal onto received vector $\mathbf{r}$.

Let us first examine the distribution of $\mathbf{r}$. Since $n(t)$ is a Gaussian random process, if we condition on the transmitted signal $s_i(t)$ then the channel output $r(t) = s_i(t) + n(t)$ is also a Gaussian random process and $\mathbf{r} = (r_1, \ldots, r_N)$ is a Gaussian random vector. Recall that $r_j = s_{ij} + n_j$. Thus, conditioned on a transmitted constellation $\mathbf{s}_i$, we have that

$$\mu_{r_j|\mathbf{s}_i} = \mathbf{E}[r_j|\mathbf{s}_i] = \mathbf{E}[s_{ij} + n_j|s_{ij}] = s_{ij} \tag{5.21}$$

(since $n(t)$ has zero mean) and

$$\sigma_{r_j|\mathbf{s}_i} = \mathbf{E}[r_j - \mu_{r_j|\mathbf{s}_i}]^2 = \mathbf{E}[s_{ij} + n_j - s_{ij}|s_{ij}]^2 = \mathbf{E}[n_j^2]. \tag{5.22}$$

Moreover,

$$\begin{aligned}
\mathrm{Cov}[r_j r_k|\mathbf{s}_i] &= \mathbf{E}[(r_j - \mu_{r_j})(r_k - \mu_{r_k})|\mathbf{s}_i] \\
&= \mathbf{E}[n_j n_k] \\
&= \mathbf{E}\left[\int_0^T n(t)\phi_j(t)\,dt \int_0^T n(\tau)\phi_k(\tau)\,d\tau\right] \\
&= \int_0^T \int_0^T \mathbf{E}[n(t)n(\tau)]\phi_j(t)\phi_k(\tau)\,dt\,d\tau \\
&= \int_0^T \int_0^T \frac{N_0}{2}\delta(t-\tau)\phi_j(t)\phi_k(\tau)\,dt\,d\tau \\
&= \frac{N_0}{2}\int_0^T \phi_j(t)\phi_k(t)\,dt \\
&= \begin{cases} N_0/2 & j = k, \\ 0 & j \neq k, \end{cases}
\end{aligned} \tag{5.23}$$

where the last equality follows from the orthonormality of the basis functions. Thus, conditioned on the transmitted constellation $\mathbf{s}_i$, the r_j are uncorrelated and, since they are Gaussian, they are also independent. Furthermore, $\mathbf{E}[n_j^2] = N_0/2$.

We have shown that, conditioned on the transmitted constellation $\mathbf{s}_i$, r_j is a Gauss-distributed random variable that is independent of $r_k (k \neq j)$ with mean s_{ij} and variance $N_0/2$. Thus, the conditional distribution of $\mathbf{r}$ is given by

$$p(\mathbf{r} \mid \mathbf{s}_i \text{ sent}) = \prod_{j=1}^{N} p(r_j \mid m_i) = \frac{1}{(\pi N_0)^{N/2}} \exp\left[-\frac{1}{N_0}\sum_{j=1}^{N}(r_j - s_{ij})^2\right]. \tag{5.24}$$

It is also straightforward to show that $\mathbf{E}[r_j n_r(t) \mid \mathbf{s}_i] = 0$ for any t, $0 \leq t < T$. Thus, since r_j conditioned on $\mathbf{s}_i$ and $n_r(t)$ are Gaussian and uncorrelated, they are independent. Also, since the transmitted signal is independent of the noise, s_{ij} is independent of the process $n_r(t)$.

We now discuss the receiver design criterion and show that it is not affected by discarding $n_r(t)$. The goal of the receiver design is to minimize the probability of error in detecting the transmitted message m_i given received signal $r(t)$. In order to minimize $P_e = p(\hat{m} \neq m_i \mid r(t)) = 1 - p(\hat{m} = m_i \mid r(t))$, we maximize $p(\hat{m} = m_i \mid r(t))$. Therefore, the receiver output $\hat{m}$ given received signal $r(t)$ should correspond to the message m_i that maximizes $p(m_i \text{ sent} \mid r(t))$. Since there is a one-to-one mapping between messages and signal constellation

points, this is equivalent to maximizing $p(\mathbf{s}_i \text{ sent} \mid r(t))$. Recalling that $r(t)$ is completely described by $\mathbf{r} = (r_1, \ldots, r_N)$ and $n_r(t)$, we have

$$\begin{aligned} p(\mathbf{s}_i \text{ sent} \mid r(t)) &= p((s_{i1}, \ldots, s_{iN}) \text{ sent} \mid (r_1, \ldots, r_N),\, n_r(t)) \\ &= \frac{p((s_{i1}, \ldots, s_{iN}) \text{ sent},\ (r_1, \ldots, r_N),\, n_r(t))}{p((r_1, \ldots, r_N),\, n_r(t))} \\ &= \frac{p((s_{i1}, \ldots, s_{iN}) \text{ sent},\ (r_1, \ldots, r_N))p(n_r(t))}{p(r_1, \ldots, r_N)p(n_r(t))} \\ &= p((s_{i1}, \ldots, s_{iN}) \text{ sent} \mid (r_1, \ldots, r_N)), \end{aligned} \tag{5.25}$$

where the third equality follows because $n_r(t)$ is independent of both $(r_1, \ldots, r_N)$ and $(s_{i1}, \ldots, s_{iN})$. This analysis shows that $\mathbf{r} = (r_1, \ldots, r_N)$ is a sufficient statistic for $r(t)$ in detecting m_i – in the sense that the probability of error is minimized by using only this sufficient statistic to estimate the transmitted signal and discarding the remainder noise. Since $\mathbf{r}$ is a sufficient statistic for the received signal $r(t)$, we call $\mathbf{r}$ the *received vector* associated with $r(t)$.

5.1.4 Decision Regions and the Maximum Likelihood Decision Criterion

We saw in the previous section that the optimal receiver minimizes error probability by selecting the detector output $\hat{m}$ that maximizes $1 - P_e = p(\hat{m} \text{ sent} \mid \mathbf{r})$. In other words, given a received vector $\mathbf{r}$, the optimal receiver selects $\hat{m} = m_i$ corresponding to the constellation $\mathbf{s}_i$ that satisfies $p(\mathbf{s}_i \text{ sent} \mid \mathbf{r}) \geq p(\mathbf{s}_j \text{ sent} \mid \mathbf{r})$ for all $j \neq i$. Let us define a set of *decision regions* $\{Z_1, \ldots, Z_M\}$ that are subsets of the signal space $\mathbb{R}^N$ by

$$Z_i = \{\mathbf{r} \colon p(\mathbf{s}_i \text{ sent} \mid \mathbf{r}) > p(\mathbf{s}_j \text{ sent} \mid \mathbf{r})\ \forall j \neq i\}. \tag{5.26}$$

Clearly these regions do not overlap. Moreover, they partition the signal space if there is no $\mathbf{r} \in \mathbb{R}^N$ for which $p(\mathbf{s}_i \text{ sent} \mid \mathbf{r}) = p(\mathbf{s}_i \text{ sent} \mid \mathbf{r})$. If such points exist then the signal space is partitioned into decision regions by arbitrarily assigning such points to decision region Z_i or Z_j. Once the signal space has been partitioned by decision regions, then for a received vector $\mathbf{r} \in Z_i$ the optimal receiver outputs the message estimate $\hat{m} = m_i$. Thus, the receiver processing consists of computing the received vector $\mathbf{r}$ from $r(t)$, finding which decision region Z_i contains $\mathbf{r}$, and outputting the corresponding message m_i. This process is illustrated in Figure 5.6, where we show a two-dimensional signal space with four decision regions $Z_1, \ldots, Z_4$ corresponding to four constellations $\mathbf{s}_1, \ldots, \mathbf{s}_4$. The received vector $\mathbf{r}$ lies in region Z_1, so the receiver will output the message m_1 as the best message estimate given received vector $\mathbf{r}$.

We now examine the decision regions in more detail. We will abbreviate $p(\mathbf{s}_i \text{ sent} \mid \mathbf{r} \text{ received})$ as $p(\mathbf{s}_i \mid \mathbf{r})$ and $p(\mathbf{s}_i \text{ sent})$ as $p(\mathbf{s}_i)$. By Bayes' rule,

$$p(\mathbf{s}_i \mid \mathbf{r}) = \frac{p(\mathbf{r} \mid \mathbf{s}_i)p(\mathbf{s}_i)}{p(\mathbf{r})}. \tag{5.27}$$

To minimize error probability, the receiver output $\hat{m} = m_i$ corresponds to the constellation $\mathbf{s}_i$ that maximizes $p(\mathbf{s}_i \mid \mathbf{r})$; that is, $\mathbf{s}_i$ must satisfy

$$\arg\max_{\mathbf{s}_i} \frac{p(\mathbf{r} \mid \mathbf{s}_i)p(\mathbf{s}_i)}{p(\mathbf{r})} = \arg\max_{\mathbf{s}_i} p(\mathbf{r} \mid \mathbf{s}_i)p(\mathbf{s}_i), \quad i = 1, \ldots, M, \tag{5.28}$$

where the second equality follows from the fact that $p(\mathbf{r})$ is not a function of $\mathbf{s}_i$. Assuming equally likely messages $(p(\mathbf{s}_i) = 1/M)$, the receiver output $\hat{m} = m_i$ corresponds to the constellation $\mathbf{s}_i$ that satisfies

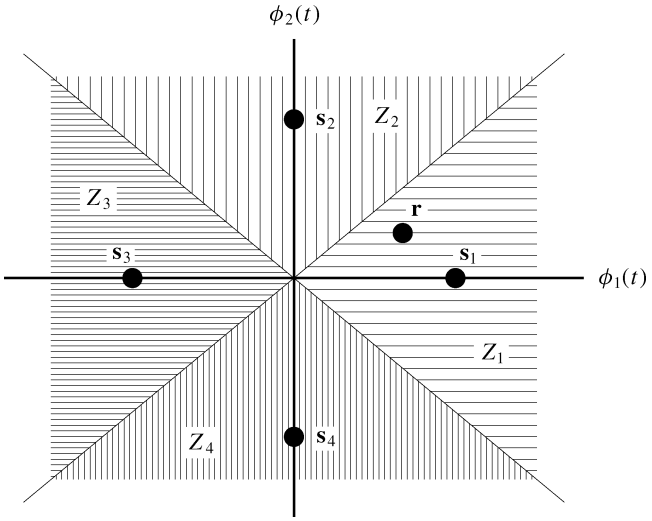

Figure 5.6: Decision regions.

$$\arg\max_{\mathbf{s}_i} p(\mathbf{r} \mid \mathbf{s}_i), \quad i = 1, \ldots, M. \tag{5.29}$$

Let us define the likelihood function associated with our receiver as

$$L(\mathbf{s}_i) = p(\mathbf{r} \mid \mathbf{s}_i). \tag{5.30}$$

Given a received vector $\mathbf{r}$, a *maximum likelihood receiver* outputs $\hat{m} = m_i$ corresponding to the constellation $\mathbf{s}_i$ that maximizes $L(\mathbf{s}_i)$. Since the log function is increasing in its argument, maximizing $L(\mathbf{s}_i)$ is equivalent to maximizing its log. Moreover, the constant factor $(\pi N_0)^{-N/2}$ in (5.24) does not affect the maximization of $L(\mathbf{s}_i)$ relative to $\mathbf{s}_i$. Thus, maximizing $L(\mathbf{s}_i)$ is equivalent to maximizing the *log likelihood function*, defined as $l(\mathbf{s}_i) = \ln[(\pi N_0)^{N/2} L(\mathbf{s}_i)]$. Using (5.24) for $L(\mathbf{s}_i) = p(\mathbf{r} \mid \mathbf{s}_i)$ then yields

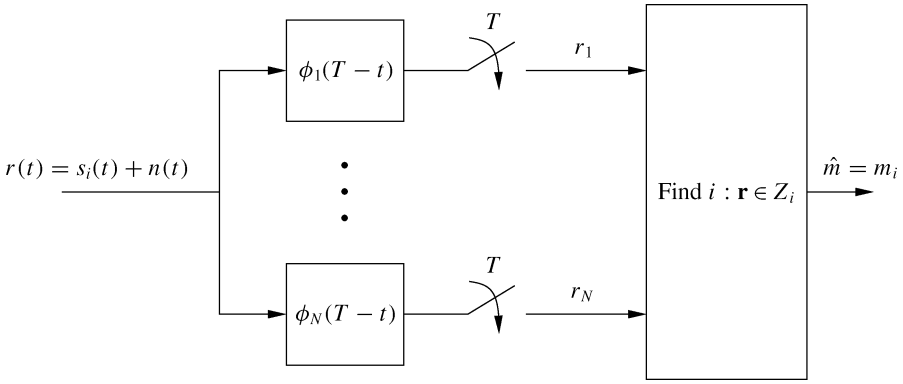

Figure 5.7: Matched filter receiver structure.

$$l(\mathbf{s}_i) = -\frac{1}{N_0} \sum_{j=1}^{N} (r_j - s_{ij})^2 = -\frac{1}{N_0} \|\mathbf{r} - \mathbf{s}_i\|^2. \tag{5.31}$$

Thus, the log likelihood function $l(\mathbf{s}_i)$ depends only on the distance between the received vector $\mathbf{r}$ and the constellation point $\mathbf{s}_i$. Moreover, from (5.31), $l(\mathbf{s}_i)$ is maximized by the constellation point $\mathbf{s}_i$ that is closest to the received vector $\mathbf{r}$.

The maximum likelihood receiver is implemented using the structure shown in Figure 5.4. First $\mathbf{r}$ is computed from $r(t)$, and then the signal constellation closest to $\mathbf{r}$ is determined as the constellation point $\mathbf{s}_i$ satisfying

$$\arg\min_{\mathbf{s}_i} \sum_{j=1}^{N} (r_j - s_{ij})^2 = \arg\min_{\mathbf{s}_i} \|\mathbf{r} - \mathbf{s}_i\|^2. \tag{5.32}$$

This $\mathbf{s}_i$ is determined from the decision region Z_i that contains $\mathbf{r}$, where Z_i is defined by

$$Z_i = \{\mathbf{r}\colon \|\mathbf{r} - \mathbf{s}_i\| < \|\mathbf{r} - \mathbf{s}_j\| \forall j = 1, \ldots, M, j \neq i\}, \quad i = 1, \ldots, M. \tag{5.33}$$

Finally, the estimated constellation $\mathbf{s}_i$ is mapped to the estimated message $\hat{m}$, which is output from the receiver. This result is intuitively satisfying, since the receiver decides that the transmitted constellation point is the one closest to the received vector. This maximum likelihood receiver structure is simple to implement because the decision criterion depends only on vector distances. This structure also minimizes the probability of message error at the receiver output when the transmitted messages are equally likely. However, if the messages and corresponding signal constellatations are not equally likely then the maximum likelihood receiver does not minimize error probability; in order to minimize error probability, the decision regions Z_i must be modified to take into account the message probabilities, as indicated in (5.27).

An alternate receiver structure is shown in Figure 5.7. This structure makes use of a bank of filters matched to each of the different basis functions. We call a filter with impulse response $\psi(t) = \phi(T - t), 0 \leq t \leq T$, the *matched filter* to the signal $\phi(t)$, so Figure 5.7 is also called a *matched filter receiver*. It can be shown that if a given input signal is passed through a filter matched to that signal then the output SNR is maximized. One can also show that the sampled matched filter outputs $(r_1, \ldots, r_n)$ in Figure 5.7 are the same as the $(r_1, \ldots, r_n)$ in Figure 5.4, so the receivers depicted in these two figures are equivalent.

Example 5.2: For BPSK modulation, find decision regions Z_1 and Z_2 corresponding to constellations $s_1 = A$ and $s_2 = -A$ for $A > 0$. *Solution:* The signal space is one-dimensional, so $\mathbf{r} = r \in \mathbb{R}$. By (5.33) the decision region $Z_1 \subset \mathbb{R}$ is defined by

$$Z_1 = \{r\colon \|r - A\| < \|r - (-A)\|\} = \{r\colon r > 0\}.$$

Thus, Z_1 contains all positive numbers on the real line. Similarly

$$Z_2 = \{r\colon \|r - (-A)\| < \|r - A\|\} = \{r\colon r < 0\}.$$

So Z_2 contains all negative numbers on the real line. For $r = 0$ the distance is the same to $s_1 = A$ and $s_2 = -A$, so we arbitrarily assign $r = 0$ to Z_2.

5.1.5 Error Probability and the Union Bound

We now analyze the error probability associated with the maximum likelihood receiver structure. For equally likely messages $p(m_i \text{ sent}) = 1/M$, we have

$$
\begin{aligned}
P_e &= \sum_{i=1}^{M} p(\mathbf{r} \notin Z_i \mid m_i \text{ sent}) p(m_i \text{ sent}) \\
&= \frac{1}{M} \sum_{i=1}^{M} p(\mathbf{r} \notin Z_i \mid m_i \text{ sent}) \\
&= 1 - \frac{1}{M} \sum_{i=1}^{M} p(\mathbf{r} \in Z_i \mid m_i \text{ sent}) \\
&= 1 - \frac{1}{M} \sum_{i=1}^{M} \int_{Z_i} p(\mathbf{r} \mid m_i)\, d\mathbf{r} \\
&= 1 - \frac{1}{M} \sum_{i=1}^{M} \int_{Z_i} p(\mathbf{r} = \mathbf{s}_i + \mathbf{n} \mid \mathbf{s}_i)\, d\mathbf{n} \\
&= 1 - \frac{1}{M} \sum_{i=1}^{M} \int_{Z_i - \mathbf{s}_i} p(\mathbf{n})\, d\mathbf{n}. \qquad (5.34)
\end{aligned}
$$

The integrals in (5.34) are over the N-dimensional subset $Z_i \subset \mathbb{R}^N$. We illustrate this error probability calculation in Figure 5.8, where the constellation points $\mathbf{s}_1, \ldots, \mathbf{s}_8$ are equally spaced around a circle with minimum separation $d_{\min}$. The probability of correct reception assuming the first symbol is sent, $p(\mathbf{r} \in Z_1 \mid m_1 \text{ sent})$, corresponds to the probability $p(\mathbf{r} = \mathbf{s}_1 + \mathbf{n} \in Z_1 \mid \mathbf{s}_1)$ that, when noise is added to the transmitted constellation $\mathbf{s}_1$, the resulting vector $\mathbf{r} = \mathbf{s}_1 + \mathbf{n}$ remains in the Z_1 region shown by the shaded area.

Figure 5.8 also indicates that the error probability is invariant to an angle rotation or axis shift of the signal constellation. The right side of the figure indicates a phase rotation of $\boldsymbol{\theta}$ and axis shift of $\mathbf{P}$ relative to the constellation on the left side. Thus, $\mathbf{s}_i' = \mathbf{s}_i e^{j\boldsymbol{\theta}} + \mathbf{P}$. The rotational invariance follows because the noise vector $\mathbf{n} = (n_1, \ldots, n_N)$ has components that are i.i.d. Gaussian random variables with zero mean; hence the polar representation $\mathbf{n} = |\mathbf{n}| e^{j\boldsymbol{\theta}}$ has $\boldsymbol{\theta}$ uniformly distributed, so the noise statistics are invariant to a phase rotation. The shift invariance follows from the fact that, if the constellation is shifted by some value $\mathbf{P} \in \mathbb{R}^N$, then the decision regions defined by (5.33) are also shifted by $\mathbf{P}$. Let $(\mathbf{s}_i, Z_i)$ denote a constellation point and corresponding decision region before the shift and $(\mathbf{s}_i', Z_i')$ the corresponding constellation point and decision region after the shift. It is then straightforward to show that $p(\mathbf{r} = \mathbf{s}_i + \mathbf{n} \in Z_i \mid \mathbf{s}_i) = p(\mathbf{r}' = \mathbf{s}_i' + \mathbf{n} \in Z_i' \mid \mathbf{s}_i')$. Thus, the error probability after an axis shift of the constellation points will remain unchanged.

Although (5.34) gives an exact solution to the probability of error, we cannot solve for this error probability in closed form. Therefore, we now investigate the union bound on error probability, which yields a closed-form expression that is a function of the distance between signal constellation points. Let A_{ik} denote the event that $\|\mathbf{r} - \mathbf{s}_k\| < \|\mathbf{r} - \mathbf{s}_i\|$ given that the constellation point $\mathbf{s}_i$ was sent. If the event A_{ik} occurs, then the constellation will be decoded in error because the transmitted constellation $\mathbf{s}_i$ is not the closest constellation point to the received vector $\mathbf{r}$. However, event A_{ik} does not necessarily imply that $\mathbf{s}_k$ will be decoded instead of $\mathbf{s}_i$, since there may be another constellation point $\mathbf{s}_l$ with $\|\mathbf{r} - \mathbf{s}_l\| < \|\mathbf{r} - \mathbf{s}_k\| < \|\mathbf{r} - \mathbf{s}_i\|$. The constellation is decoded correctly if $\|\mathbf{r} - \mathbf{s}_i\| < \|\mathbf{r} - \mathbf{s}_k\|$ for all $k \neq i$. Thus

$$
P_e(m_i \text{ sent}) = p\left(\bigcup_{\substack{k=1 \\ k \neq i}}^{M} A_{ik} \right) \leq \sum_{\substack{k=1 \\ k \neq i}}^{M} p(A_{ik}), \qquad (5.35)
$$

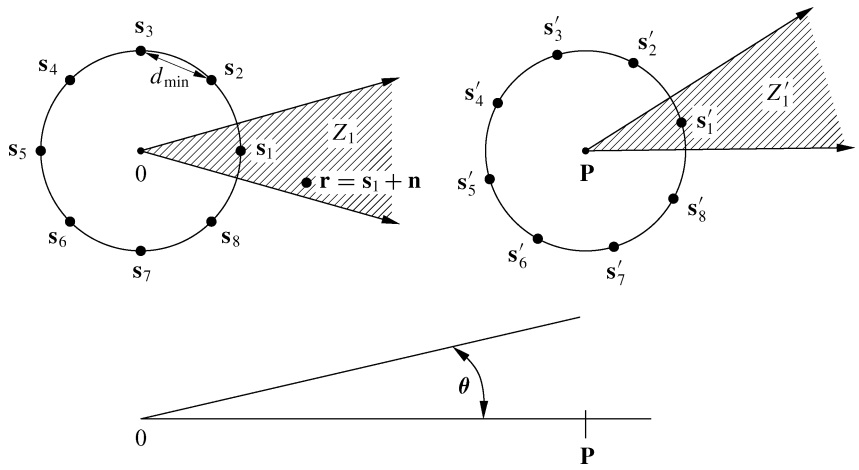

Figure 5.8: Error probability integral and its rotational/shift invariance.

where the inequality follows from the union bound on probability, defined below.

Let us now consider $p(A_{ik})$ more closely. We have

$$\begin{aligned} p(A_{ik}) &= p(\|\mathbf{s}_k - \mathbf{r}\| < \|\mathbf{s}_i - \mathbf{r}\| \mid \mathbf{s}_i \text{ sent}) \\ &= p(\|\mathbf{s}_k - (\mathbf{s}_i + \mathbf{n})\| < \|\mathbf{s}_i - (\mathbf{s}_i + \mathbf{n})\|) \\ &= p(\|\mathbf{n} + \mathbf{s}_i - \mathbf{s}_k\| < \|\mathbf{n}\|); \end{aligned} \tag{5.36}$$

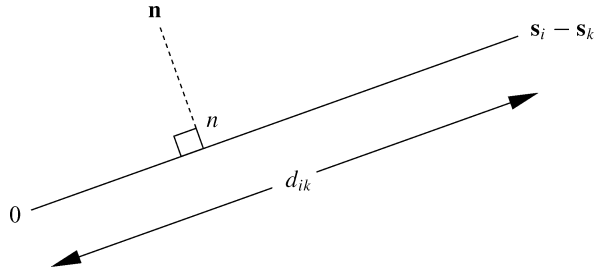

Figure 5.9: Noise projection.

that is, the probability of error equals the probability that the noise $\mathbf{n}$ is closer to the vector $\mathbf{s}_i - \mathbf{s}_k$ than to the origin. Recall that the noise has a mean of zero, so it is generally close to the origin. This probability does not depend on the entire noise component $\mathbf{n}$: it only depends on the projection of $\mathbf{n}$ onto the line connecting the origin and the point $\mathbf{s}_i - \mathbf{s}_k$, as shown in Figure 5.9. Given the properties of $\mathbf{n}$, the projection of $\mathbf{n}$ onto this one-dimensional line is a one-dimensional Gaussian random variable n with mean zero and variance $N_0/2$. The event A_{ik} occurs if n is closer to $\mathbf{s}_i - \mathbf{s}_k$ than to zero-that is, if $n > d_{ik}/2$, where $d_{ik} = \|\mathbf{s}_i - \mathbf{s}_k\|$ equals the distance between constellation points $\mathbf{s}_i$ and $\mathbf{s}_k$. Thus,

$$p(A_{ik}) = p\left(n > \frac{d_{ik}}{2}\right) = \int_{d_{ik}/2}^{\infty} \frac{1}{\sqrt{\pi N_0}} \exp\left[\frac{-v^2}{N_0}\right] dv = Q\left(\frac{d_{ik}}{\sqrt{2N_0}}\right). \tag{5.37}$$

Substituting (5.37) into (5.35) yields

$$P_e(m_i \text{ sent}) \leq \sum_{\substack{k=1 \\ k \neq i}}^{M} Q\left(\frac{d_{ik}}{\sqrt{2N_0}}\right), \tag{5.38}$$

where the Q-function, $Q(z)$, is defined as the probability that a Gaussian random variable X with mean 0 and variance 1 is greater than z:

$$Q(z) = p(X > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \tag{5.39}$$

Summing (5.38) over all possible messages yields the *union bound*

$$P_e = \sum_{i=1}^{M} p(m_i) P_e(m_i \text{ sent}) \leq \frac{1}{M} \sum_{i=1}^{M} \sum_{\substack{k=1 \\ k \neq i}}^{M} Q\left(\frac{d_{ik}}{\sqrt{2N_0}}\right). \tag{5.40}$$

Note that the Q-function cannot be solved for in closed form. It is related to the complementary error function as

$$Q(z) = \frac{1}{2} \text{ erfc}\left(\frac{z}{\sqrt{2}}\right). \tag{5.41}$$

We can also place an upper bound on $Q(z)$ with the closed-form expression

$$Q(z) \leq \frac{1}{z\sqrt{2\pi}} e^{-z^2/2}, \tag{5.42}$$

and this bound is tight for $z \gg 0$. A simpler upper bound can be obtained from the Chernoff bound [4] as

$$Q(z) \leq e^{-z^2/2} \tag{5.43}$$

Defining the *minimum distance* of the constellation as $d_{\min} = \min_{i,k} d_{ik}$, we can simplify (5.40) with the looser bound

$$P_e \leq (M-1) Q\left(\frac{d_{\min}}{\sqrt{2N_0}}\right). \tag{5.44}$$

Using (5.42) for the Q-function yields a closed-form bound

$$P_e \leq \frac{M-1}{d_{\min}\sqrt{\pi/N_0}} \exp\left[\frac{-d_{\min}^2}{4N_0}\right]. \tag{5.45}$$

Finally, P_e is sometimes approximated as the probability of error associated with constellations at the minimum distance $d_{\min}$ multiplied by the number $M_{d_{\min}}$ of neighbors at this distance:

$$P_e \approx M_{d_{\min}} Q\left(\frac{d_{\min}}{\sqrt{2N_0}}\right). \tag{5.46}$$

This approximation is called the *nearest neighbor approximation* to P_e. When different constellation points have a different number of nearest neighbors or different minimum distances, the bound can be averaged over the bound associated with each constellation point. Note that the nearest neighbor approximation will always be less than the loose bound (5.44) since $M \geq M_{d_{\min}}$. It will also be slightly less than the union bound (5.40), since the nearest neighbor approximation does not include the error associated with constellations farther apart than the minimum distance. However, the nearest neighbor approximation is quite close to the exact probability of symbol error at

high SNRs, since for x and y large with $x > y,\ Q(x) \ll Q(y)$ owing to the exponential falloff of the Gaussian distribution in (5.39). This indicates that the probability of mistaking a constellation point for another point that is not one of its nearest neighbors is negligible at high SNRs. A rigorous derivation for (5.46) is made in [5] and also referenced in [6]. Moreover, [5] indicates that (5.46) captures the performance degradation due to imperfect receiver conditions such as slow carrier drift with an appropriate adjustment of the constants. The appeal of the nearest neighbor approximation is that it depends only on the minimum distance in the signal constellation and the number of nearest neighbors for points in the constellation.

Example 5.3: Consider a signal constellation in $\mathbb{R}^2$ defined by $s_1 = (A, 0),\ s_2 = (0, A),\ s_3 = (-A, 0)$, and $s_4 = (0, -A)$. Assume $A/\sqrt{N_0} = 4$. Find the minimum distance and the union bound (5.40), looser bound (5.44), closed-form bound (5.45), and nearest neighbor approximation (5.46) on P_e for this constellation set. *Solution:* The constellation is as depicted in Figure 5.3 with the radius of the circle equal to A. By symmetry, we need only consider the error probability associated with one of the constellation points, since it will be the same for the others. We focus on the error associated with transmitting constellation point $\mathbf{s}_1$. The minimum distance to this constellation point is easily computed as $d_{\min} = d_{12} = d_{23} = d_{34} = d_{14} = \sqrt{A^2 + A^2} = \sqrt{2A^2}$. The distance to the other constellation points are $d_{13} = d_{24} = 2A$. By symmetry, $P_e(m_i \text{ sent}) = P_e(m_j \text{ sent})$ for $j \neq i$, so the union bound simplifies to

$$\begin{aligned} P_e &\leq \sum_{j=2}^{4} Q\left(\frac{d_{1j}}{\sqrt{2N_0}}\right) \\ &= 2Q\left(\frac{A}{\sqrt{N_0}}\right) + Q\left(\frac{\sqrt{2}A}{\sqrt{N_0}}\right) = 2Q(4) + Q(\sqrt{32}) = 3.1679 \cdot 10^{-5}. \end{aligned}$$

The looser bound yields

$$P_e \leq 3Q(4) = 9.5014 \cdot 10^{-5},$$

which is roughly a factor of 3 looser than the union bound. The closed-form bound yields

$$P_e \leq \frac{3}{\sqrt{2\pi A^2/N_0}} \exp\left[\frac{-.5A^2}{N_0}\right] = 1.004 \cdot 10^{-4},$$

which differs from the union bound by about an order of magnitude. Finally, the nearest neighbor approximation yields

$$P_e \approx 2Q(4) = 3.1671 \cdot 10^{-5},$$

which (as expected) is approximately equal to the union bound.

Note that, for binary modulation (where $M = 2$), there is only one way to make an error and $d_{\min}$ is the distance between the two signal constellation points, so the bound (5.44) is exact:

$$P_b = Q\left(\frac{d_{\min}}{\sqrt{2N_0}}\right). \tag{5.47}$$

The square of the minimum distance $d_{\min}$ in (5.45) and (5.47) is typically proportional to the SNR of the received signal, as discussed in Chapter 6.Thus, error probability is reduced by increasing the received signal power.

Recall that P_e is the probability of a symbol (message) error: $P_e = p(\hat{m} \neq m_i \mid m_i \text{ sent})$, where m_i corresponds to a message with $\log_2 M$ bits. However, system designers are typically more interested in the bit

error probability (also called the bit error rate, BER) than in the symbol error probability, because bit errors drive the performance of higher-layer networking protocols and end-to-end performance. Thus, we would like to design the mapping of the M possible bit sequences to messages m_i $(i = 1, \ldots, M)$ so that a decoding error associated with an adjacent decision region, which is the most likely way to make an error, corresponds to only one bit error. With such a mapping – and with the assumption that mistaking a signal constellation point for a point other than one of its nearest neighbors has a very low probability – we can make the approximation

$$P_b \approx \frac{P_e}{\log_2 M}. \tag{5.48}$$

The most common form of mapping in which mistaking a constellation point for one of its nearest neighbors results in a single bit error is called Gray coding. Mapping by Gray coding is discussed in more detail in Section 5.3. Signal space concepts are applicable to any modulation where bits are encoded as one of several possible analog signals, including the amplitude, phase, and frequency modulations discussed in what follows.

5.2 Passband Modulation Principles

The basic principle of passband digital modulation is to encode an information bit stream into a carrier signal, which is then transmitted over a communications channel. Demodulation is the process of extracting this information bit stream from the received signal. Corruption of the transmitted signal by the channel can lead to bit errors in the demodulation process. The goal of modulation is to send bits at a high data rate while minimizing the probability of data corruption.

In general, modulated carrier signals encode information in the amplitude $\alpha(t)$, frequency $f(t)$, or phase $\theta(t)$ of a carrier signal. Thus, the modulated signal can be represented as

$$s(t) = \alpha(t)\cos[2\pi(f_c + f(t))t + \theta(t) + \phi_0] = \alpha(t)\cos(2\pi f_c t + \phi(t) + \phi_0), \tag{5.49}$$

where $\phi(t) = 2\pi f(t)t + \theta(t)$ and ϕ_0 is the phase offset of the carrier. This representation combines frequency and phase modulation into angle modulation.

We can rewrite the right-hand side of (5.49) in terms of its in-phase and quadrature components as

$$\begin{aligned} s(t) &= \alpha(t)\cos(\phi(t) + \phi_0)\cos(2\pi f_c t) - \alpha(t)\sin(\phi(t) + \phi_0)\sin(2\pi f_c t) \\ &= s_I(t)\cos(2\pi f_c t) - s_Q(t)\sin(2\pi f_c t), \end{aligned} \tag{5.50}$$

where $s_I(t) = \alpha(t)(\cos\phi(t) + \phi_0)$ is the in-phase component of $s(t)$ and where $s_Q(t) = \alpha(t)(\sin\phi(t) + \phi_0)$ is its quadrature component. We can also write $s(t)$ in terms of its equivalent lowpass representation as

$$s(t) = \mathrm{Re}\{u(t)e^{j2\pi f_c t}\}, \tag{5.51}$$

where $u(t) = s_I(t) + js_Q(t)$. This representation, described in more detail in Appendix A, is useful because receivers typically process the in-phase and quadrature signal components separately.

5.3 Amplitude and Phase Modulation

In amplitude and phase modulation, the information bit stream is encoded in the amplitude and/or phase of the transmitted signal. Specifically: over a time interval of T_s, $K = \log_2 M$ bits are encoded into the amplitude and/or phase of the transmitted signal $s(t)$, $0 \le t < T_s$. The transmitted signal over this period, $s(t) = s_I(t)\cos(2\pi f_c t) -$

$s_Q(t)\sin(2\pi f_c t)$, can be written in terms of its signal space representation as $s(t) = s_{i1}\phi_1(t) + s_{i2}\phi_2(t)$ with basis functions $\phi_1(t) = g(t)\cos(2\pi f_c t + \phi_0)$ and $\phi_2(t) = -g(t)\sin(2\pi f_c t + \phi_0)$, where $g(t)$ is a shaping pulse. For $\phi_0 = 0$, to send the ith message over the time interval $[kT, (k+1)T)$ we set $s_I(t) = s_{i1}g(t)$ and $s_Q(t) = s_{i2}g(t)$. These in-phase and quadrature signal components are baseband signals with spectral characteristics determined by the pulse shape $g(t)$. In particular, their bandwidth B equals the bandwidth of $g(t)$, and the transmitted signal $s(t)$ is a passband signal with center frequency f_c and passband bandwidth $2B$. In practice we take $B = K_g/T_s$, where K_g depends on the pulse shape: for rectangular pulses $K_g = .5$ and for raised cosine pulses $.5 \leq K_g \leq 1$, as discussed in Section 5.5. Thus, for rectangular pulses the bandwidth of $g(t)$ is. $5/T_s$ and the bandwidth of $s(t)$ is $1/T_s$. The signal constellation for amplitude and phase modulation is defined based on the constellation points $\{(s_{i1}, s_{i2}) \in \mathbb{R}^2, i = 1, \ldots, M\}$. The equivalent lowpass representation of $s(t)$ is

$$s(t) = \mathrm{Re}\{x(t)e^{j\phi 0}e^{j2\pi f_c t}\}, \tag{5.52}$$

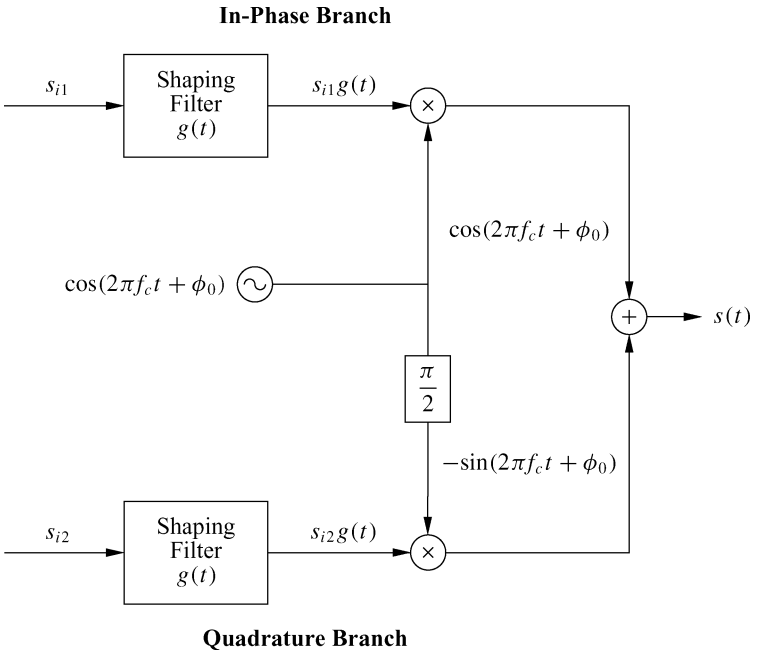

Figure 5.10: Amplitude/phase modulator.

where $x(t) = (s_{i1} + js_{i2})g(t)$. The constellation point $s_i = (s_{i1}, s_{i2})$ is called the *symbol* associated with the $\log_2 M$ bits, and T_s is called the *symbol time*. The bit rate for this modulation is K bits per symbol or $R = \log_2 M/T_s$ bits per second.

There are three main types of amplitude/phase modulation:

- pulse amplitude modulation (MPAM) – information encoded in amplitude only;
- phase-shift keying (MPSK) – information encoded in phase only;
- quadrature amplitude modulation (MQAM) – information encoded in both amplitude and phase.

The number of bits per symbol $K = \log_2 M$, the signal constellation $\{\mathbf{s}_i, i = 1, \ldots, M\}$, and the choice of pulse shape $g(t)$ determine the digital modulation design. The pulse shape $g(t)$ is chosen to improve spectral efficiency and combat ISI, as discussed in Section 5.5.

Amplitude and phase modulation over a given symbol period can be generated using the modulator structure shown in Figure 5.10. Note that the basis functions in this figure have an arbitrary phase ϕ_0 associated with the transmit oscillator. Demodulation over each symbol period is performed using the demodulation structure of Figure 5.11, which is equivalent to the structure of Figure 5.7 for $\phi_1(t) = g(t)\cos(2\pi f_c t + \phi)$ and $\phi_2(t) = -g(t)\sin(2\pi f_c t + \phi)$. Typically the receiver includes some additional circuitry for *carrier phase recovery* that matches the carrier phase ϕ at the receiver to the carrier phase ϕ_0 at the transmitter;[1] this is known as *coherent detection*. If $\phi - \phi_0 = \triangle\phi \neq 0$ then the in-phase branch will have an unwanted term associated with the quadrature branch and vice versa; that is, $r_1 = s_{i1}\cos(\triangle\phi) + s_{i2}\sin(\triangle\phi) + n_1$ and $r_2 = -s_{i1}\sin(\triangle\phi) + s_{i2}\cos(\triangle\phi) + n_2$, which can result in significant performance degradation. The receiver structure also assumes that the sampling function every T_s seconds is synchronized to the start of the symbol period, which is called *synchronization* or *timing recovery*. Receiver synchronization and carrier phase recovery are complex receiver operations that can be highly challenging in wireless environments. These operations are discussed in more detail in Section 5.6. We will assume perfect carrier recovery in our discussion of MPAM, MPSK, and MQAM and therefore set $\phi = \phi_0 = 0$ for their analysis.

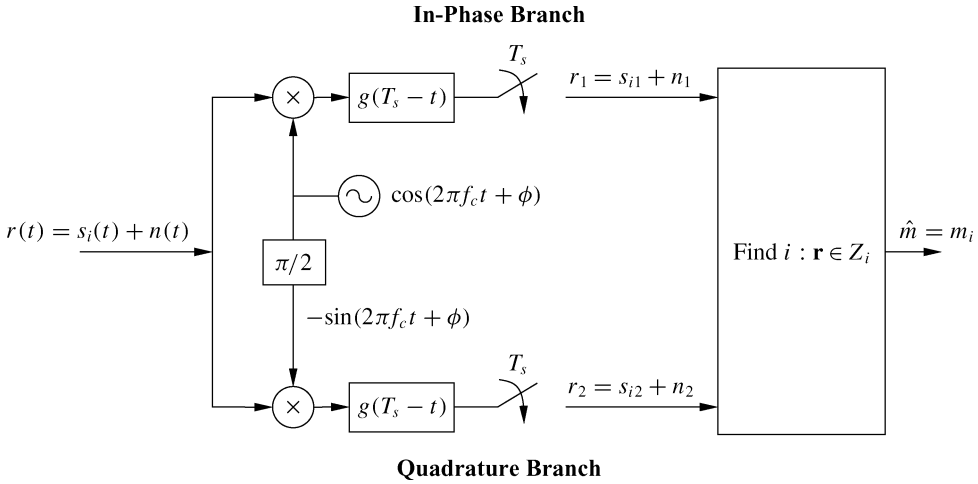

Figure 5.11: Amplitude/phase demodulator (coherent: $\phi = \phi_0$).

5.3.1 Pulse Amplitude Modulation (MPAM)

We will start by looking at the simplest form of linear modulation, one- dimensional MPAM, which has no quadrature component ($s_{i2} = 0$). For MPAM, all of the information is encoded into the signal amplitude A_i. The transmitted signal over one symbol time is given by

$$s_i(t) = \mathrm{Re}\{A_i g(t) e^{j2\pi f_c t}\} = A_i g(t)\cos(2\pi f_c t),\ 0 \leq t \leq T_s \gg 1/f_c, \tag{5.53}$$

where $A_i = (2i - 1 - M)d, i = 1, 2, \ldots, M$. The signal constellation is thus $\{A_i, i = 1, \ldots, M\}$, which is parameterized by the distance d. This distance is typically a function of the signal energy. The pulse shape $g(t)$ must satisfy (5.12) and (5.13). The minimum distance between constellation points is $d_{\min} = \min_{i,j}|A_i - A_j| = 2d$. The amplitude of the transmitted signal takes on M different values, which implies that each pulse conveys $\log_2 M = K$ bits per symbol time T_s.

[1] In fact, an additional phase term of $-2\pi f_c \tau$ will result from a propagation delay of τ in the channel. Thus, coherent detection requires the receiver phase $\phi = \phi_0 - 2\pi f_c \tau$, as discussed in more detail in Section 5.6.

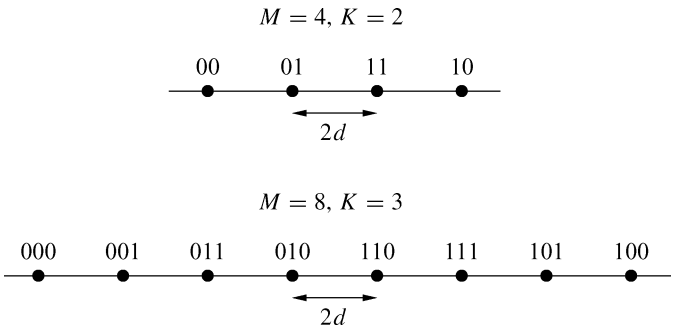

Figure 5.12: Gray encoding for MPAM.

Over each symbol period, the MPAM signal associated with the ith constellation has energy

$$E_{si} = \int_0^{T_s} s_i^2(t)\,dt = \int_0^{T_s} A_i^2 g^2(t)\cos^2(2\pi f_c t)\,dt = A_i^2, \tag{5.54}$$

since the pulse shape must satisfy (5.12).[2] Note that the energy is not the same for each signal $s_i(t), i = 1, \ldots, M$. Assuming equally likely symbols, the average energy is

$$\overline{E}_s = \frac{1}{M}\sum_{i=1}^{M} A_i^2. \tag{5.55}$$

The constellation mapping is usually done by Gray encoding, where the messages associated with signal amplitudes that are adjacent to each other differ by one bit value, as illustrated in Figure 5.12. With this encoding method, if noise causes the demodulation process to mistake one symbol for an adjacent one (the most likely type of error), the result is only a single bit error in the sequence of K bits. Gray codes can also be designed for MPSK and square MQAM constellations but not for rectangular MQAM.

Example 5.4: For $g(t) = \sqrt{2/T_s}(0 \le t < T_s)$ a rectangular pulse shape, find the average energy of 4-PAM modulation. *Solution:* For 4-PAM the A_i values are $A_i = \{-3d, -d, d, 3d\}$, so the average energy is

$$\overline{E}_s = \frac{d^2}{4}(9 + 1 + 1 + 9) = 5d^2.$$

The decision regions $Z_i, i = 1, \ldots, M$, associated with pulse amplitude $A_i = (2i - 1 - M)d$ for $M = 4$ and $M = 8$ are shown in Figure 5.13. Mathematically, for any M these decision regions are defined by

$$Z_i = \begin{cases} (-\infty, A_i + d) & i = 1, \\ [A_i - d, A_i + d) & 2 \le i \le M - 1, \\ [A_i - d, \infty) & i = M. \end{cases}$$

From (5.53) we see that MPAM has only a single basis function $\phi_1(t) = g(t)\cos(2\pi f_c t)$. Thus, the coherent demodulator of Figure 5.11 for MPAM reduces to the demodulator shown in Figure 5.14, where the multithreshold device maps r to a decision region Z_i and outputs the corresponding bit sequence $\hat{m} = m_i = \{\mathrm{b}_1, \ldots, \mathrm{b}_K\}$.

[2] Recall from (5.8) that (5.12) and hence (5.54) are not exact equalities but rather very good approximations for $f_c T_s \gg 1$.

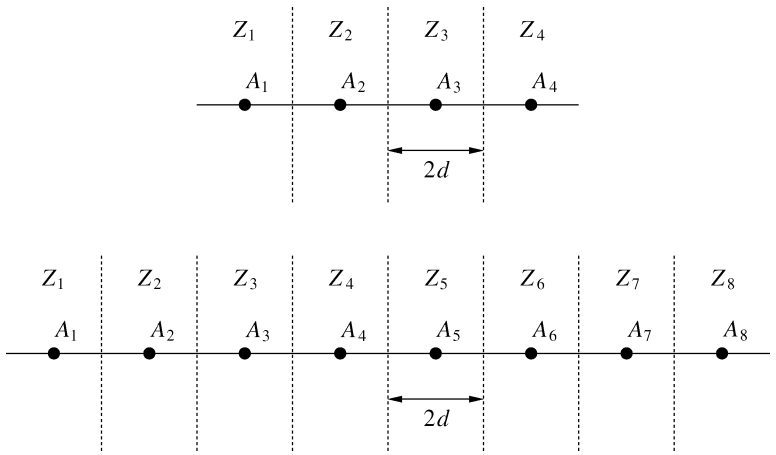

Figure 5.13: Decision regions for MPAM.

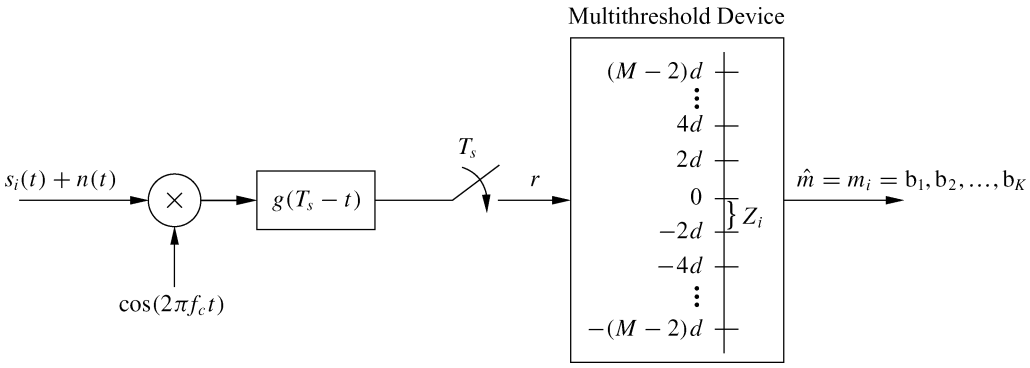

Figure 5.14: Coherent demodulator for MPAM.

5.3.2 Phase-Shift Keying (MPSK)

For MPSK, all of the information is encoded in the phase of the transmitted signal. Thus, the transmitted signal over one symbol time T_s is given by

$$\begin{aligned} s_i(t) &= \mathrm{Re}\{Ag(t)e^{j2\pi(i-1)/M}e^{j2\pi f_c t}\} \\ &= Ag(t)\cos\left[2\pi f_c t + \frac{2\pi(i-1)}{M}\right] \\ &= Ag(t)\cos\left[\frac{2\pi(i-1)}{M}\right]\cos 2\pi f_c t - Ag(t)\sin\left[\frac{2\pi(i-1)}{M}\right]\sin 2\pi f_c t \end{aligned} \tag{5.56}$$

for $0 \le t \le T_s$. Therefore, the constellation points or symbols (s_{i1}, s_{i2}) are given by $s_{i1} = A\cos[2\pi(i-1)/M]$ and $s_{i2} = A\sin[2\pi(i-1)/M]$ for $i = 1, \ldots, M$. The pulse shape $g(t)$ satisfies (5.12) and (5.13), and the $\theta_i = 2\pi(i-1)/M (i = 1, 2, \ldots, M = 2^K)$ are the different phases in the signal constellation points that convey the information bits. The minimum distance between constellation points is $d_{\min} = 2A\sin(\pi/M)$, where A is typically a function of the signal energy. Note that 2-PSK is often referred to as binary PSK or BPSK, while 4-PSK is often called quadrature phase shift keying (QPSK) and is the same as MQAM with $M = 4$ (defined in Section 5.3.3).

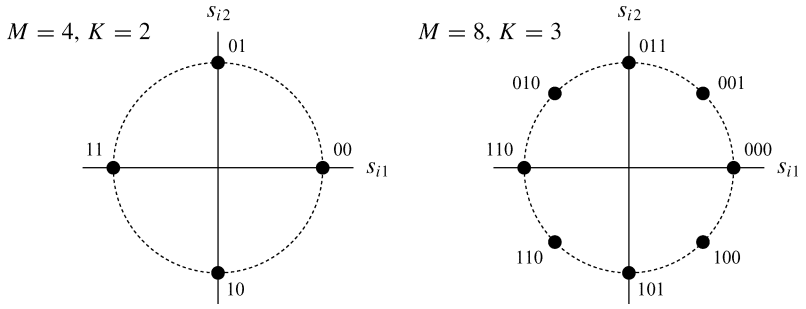

Figure 5.15: Gray encoding for MPSK.

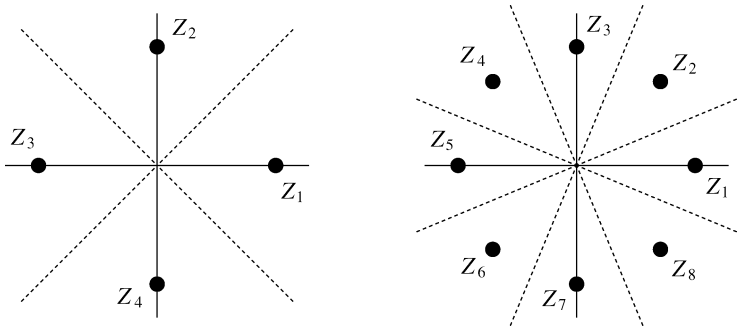

Figure 5.16: Decision regions for MPSK.

All possible transmitted signals $s_i(t)$ have equal energy:

$$E_{s_i} = \int_0^{T_s} s_i^2(t)dt = A^2. \tag{5.57}$$

Observe that for $g(t) = \sqrt{2/T_s}$ over a symbol time (i.e., a rectangular pulse) this signal has constant envelope, unlike the other amplitude modulation techniques MPAM and MQAM. However, rectangular pulses are spectrally inefficient, and more efficient pulse shapes give MPSK a nonconstant signal envelope. As for MPAM, constellation mapping is usually done by Gray encoding, where the messages associated with signal phases that are adjacent to each other differ by one bit value; see Figure 5.15. With this encoding method, mistaking a symbol for an adjacent one causes only a single bit error.

The decision regions $Z_i (i = 1, \ldots, M)$ associated with MPSK for $M = 4$ and $M = 8$ are shown in Figure 5.16. If we represent $\mathbf{r} = r_1 + jr_2 = re^{j\theta} \in \mathbb{R}^2$ in polar coordinates, then these decision regions for any M are defined by

$$Z_i = \{re^{j\theta} : 2\pi(i - 1.5)/M \leq \theta < 2\pi(i - .5)/M\}. \tag{5.58}$$

From (5.56) we see that MPSK has both in-phase and quadrature components, and thus the coherent demodulator is as shown in Figure 5.11. For the special case of BPSK, the decision regions as given in Example 5.2 simplify to $Z_1 = (r : r > 0)$ and $Z_2 = (r : r \leq 0)$. Moreover, BPSK has only a single basis function $\phi_1(t) = g(t)\cos(2\pi f_c t)$ and, since there is only a single bit transmitted per symbol time T_s, the bit time $T_b = T_s$. Thus, the coherent demodulator of Figure 5.11 for BPSK reduces to the demodulator shown in Figure 5.17, where the threshold device maps r to the positive or negative half of the real line and then outputs the corresponding bit value. We have assumed in this figure that the message corresponding to a bit value of $1, m_1 = 1$, is mapped to constellation

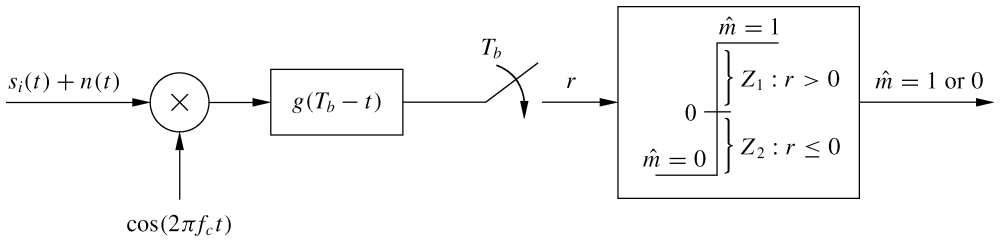

Figure 5.17: Coherent demodulator for BPSK.

point $s_1 = A$ and that the message corresponding to a bit value of $0, m_2 = 0$, is mapped to the constellation point $s_2 = -A$.

5.3.3 Quadrature Amplitude Modulation (MQAM)

For MQAM, the information bits are encoded in both the amplitude and phase of the transmitted signal. Thus, whereas both MPAM and MPSK have one degree of freedom in which to encode the information bits (amplitude or phase), MQAM has two degrees of freedom. As a result, MQAM is more spectrally efficient than MPAM and MPSK in that it can encode the most number of bits per symbol for a given average energy.

The transmitted signal is given by

$$\begin{aligned} s_i(t) &= \mathrm{Re}\{A_i e^{j\theta_i} g(t) e^{j2\pi f_c t}\} \\ &= A_i \cos(\theta_i) g(t) \cos(2\pi f_c t) - A_i \sin(\theta_i) g(t) \sin(2\pi f_c t), 0 \leq t \leq T_s, \end{aligned} \tag{5.59}$$

where the pulse shape $g(t)$ satisfies (5.12) and (5.13). The energy in $s_i(t)$ is

$$E_{s_i} = \int_0^{T_s} s_i^2(t) = A_i^2, \tag{5.60}$$

the same as for MPAM. The distance between any pair of symbols in the signal constellation is

$$d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\| = \sqrt{(s_{i1} - s_{j1})^2 + (s_{i2} - s_{j2})^2}. \tag{5.61}$$

For square signal constellations, where s_{i1} and s_{i2} take values on $(2i - 1 - L)d$ with $i = 1, 2, \ldots, L$, the minimum distance between signal points reduces to $d_{\min} = 2d$, the same as for MPAM. In fact, MQAM with square constellations of size L^2 is equivalent to MPAM modulation with constellations of size L on each of the in-phase and quadrature signal components. Common square constellations are 4-QAM and 16-QAM, which are shown in Figure 5.18. These square constellations have $M = L^2 = 2^{2l}$ constellation points, which are used to send $2l$ bits/symbol or/bits per dimension, where $l = .5 \log_2 M$. It can be shown that the average power of a square signal constellation with l bits per dimension, P_l, is proportional to $4^l/3$, and it follows that the average power for one more bit per dimension $P_{l+1} \approx 4P_l$. Thus, for square constellations it takes approximately 6 dB more power to send an additional 1 bit/dimension or 2 bits/symbol while maintaining the same minimum distance between constellation points.

Good constellation mappings can be hard to find for QAM signals, especially for irregular constellation shapes. In particular, it is hard to find a Gray code mapping where all adjacent symbols differ by a single bit. The decision regions $Z_i (i = 1, \ldots, M)$ associated with MQAM for $M = 16$ are shown in Figure 5.19. From

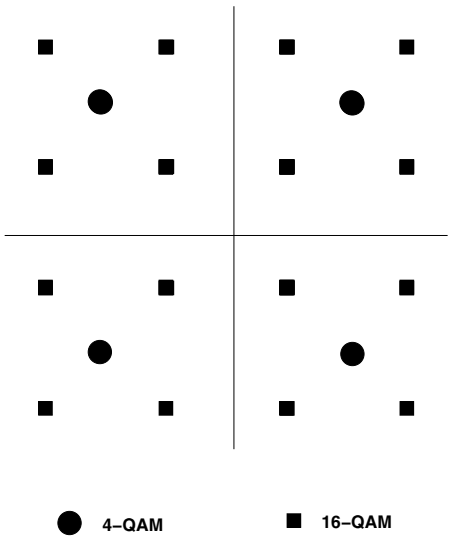

Figure 5.18: 4-QAM and 16-QAM constellations.

(5.59) we see that MQAM has both in-phase and quadrature components, and thus the coherent demodulator is as shown in Figure 5.11.

5.3.4 Differential Modulation

The information in MPSK and MQAM signals is carried in the signal phase. These modulation techniques therefore require coherent demodulation; that is, the phase of the transmitted signal carrier ϕ_0 must be matched to the phase of the receiver carrier ϕ. Techniques for phase recovery typically require more complexity and cost in the receiver, and they are also susceptible to phase drift of the carrier. Moreover, obtaining a coherent phase reference in a rapidly fading channel can be difficult. Issues associated with carrier phase recovery are discussed in more detail in Section 5.6. The difficulties as well as the cost and complexity associated with carrier phase recovery motivate the use of differential modulation techniques, which do not require a coherent phase reference at the receiver.

Differential modulation falls in the more general class of modulation with memory, where the symbol transmitted over time $[kT_s, (k+1)T_s)$ depends on the bits associated with the current message to be transmitted *and* on the bits transmitted over prior symbol times. The basic principle of differential modulation is to use the previous symbol as a phase reference for the current symbol, thus avoiding the need for a coherent phase reference at the receiver. Specifically, the information bits are encoded as the differential phase between the current symbol and the previous symbol. For example, in differential BPSK (referred to as DPSK), if the symbol over time $[(k-1)T_s, kT_s)$ has phase $\theta(k-1) = e^{j\theta i}$ for $\theta_i = 0, \pi$, then to encode a 0-bit over $[kT_s, (k+1)T_s)$ the symbol would have phase $\theta(k) = e^{j\theta i}$ and to encode a 1-bit the symbol would have phase $\theta(k) = e^{j(\theta i+\pi)}$. In other words: a 0-bit is encoded by no change in phase, whereas a 1-bit is encoded as a phase change of π. Similarly, in 4-PSK modulation with differential encoding, the symbol phase over symbol interval $[kT_s, (k+1)T_s)$ depends on the current information bits over this time interval and the symbol phase over the previous symbol interval. The phase transitions for DQPSK modulation are summarized in Table 5.1.

Specifically, suppose the symbol over time $[(k-1)T_s, kT_s)$ has phase $\theta(k-1) = e^{j\theta_i}$. Then, over symbol time $[kT_s, (k+1)T_s)$, if the information bits are 00 then the corresponding symbol would have phase $\theta(k) = e^{j\theta_i}$; that is, to encode the bits 00, the symbol from symbol interval $[(k-1)T_s, kT_s)$ is repeated over the next interval $[kT_s, (k+1)T_s)$. If the two information bits to be sent at time interval $[kT_s, (k+1)T_s)$ are 01, then the corresponding symbol has phase $\theta(k) = e^{j(\theta_i+\pi/2)}$. For information bits 10 the symbol phase is $\theta(k) = e^{j(\theta_i-\pi/2)}$,

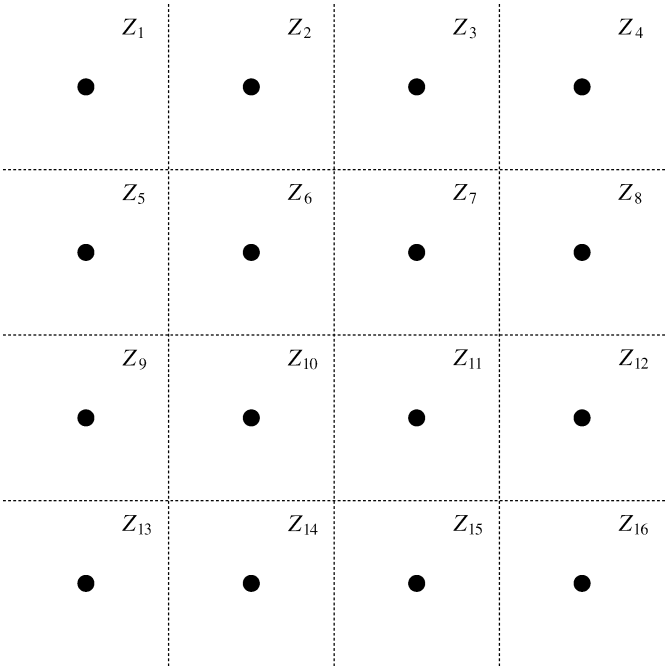

Figure 5.19: Decision regions for MQAM with $M = 16$.

Table 5.1: Mapping for DQPSK with Gray encoding

Bit sequence	**Phase transition**
00	0
01	$\pi/2$
10	$-\pi/2$
11	π

and for information bits 11 the symbol phase is $\theta(n) = e^{j(\theta_k+\pi)}$. We see that the symbol phase over symbol interval $[kT_s, (k+1)T_s)$ depends on the current information bits over this time interval and on the symbol phase θ_i over the previous symbol interval. Note that this mapping of bit sequences to phase transitions ensures that the most likely detection error – that of mistaking a received symbol for one of its nearest neighbors – results in a single bit error. For example, if the bit sequence 00 is encoded in the kth symbol then the kth symbol has the same phase as the $(k-1)$th symbol. Assume this phase is θ_i. The most likely detection error of the kth symbol is to decode it as one of its nearest neighbor symbols, which have phase $\theta_i \pm \pi/2$. But decoding the received symbol with phase $\theta_i \pm \pi/2$ would result in a decoded information sequence of either 01 or 10 – that is, it would differ by a single bit from the original sequence 00. More generally, we can use Gray encoding for the phase transitions in differential MPSK for any M, so that a message of all 0-bits results in no phase change, a message with a single 1-bit and the rest 0-bits results in the minimum phase change of $2\pi/M$, a message with two 1-bits and the rest 0-bits results in a phase change of $4\pi/M$, and so forth. Differential encoding is most common for MPSK signals, since the differential mapping is relatively simple. Differential encoding can also be done for MQAM with a more complex differential mapping. Differential encoding of MPSK is denoted by DMPSK, and for BPSK and QPSK by DPSK and DQPSK, respectively.

Example 5.5: Find the sequence of symbols transmitted using DPSK for the bit sequence 101110 starting at the kth symbol time, assuming the transmitted symbol at the $(k-1)$th symbol time was $\mathbf{s}(k-1) = Ae^{j\pi}$.

Solution: The first bit, a 1, results in a phase transition of π, so $\mathbf{s}(k) = A$. The next bit, a 0, results in no transition, so $\mathbf{s}(k+1) = A$. The next bit, a 1, results in another transition of π, so $\mathbf{s}(k+1) = Ae^{j\pi}$, and so on. The full symbol sequence corresponding to 101110 is A, A, $Ae^{j\pi}$, A, $Ae^{j\pi}$, $Ae^{j\pi}$.

The demodulator for differential modulation is shown in Figure 5.20. Assume the transmitted constellation at time k is $\mathbf{s}(k) = Ae^{j(\theta(k)+(\phi_0)}$. Then the received vector associated with the sampler outputs is

$$\mathbf{r}(k) = r_1(k) + jr_2(k) = Ae^{j(\theta(k)+(\phi_0)} + n(k), \tag{5.62}$$

where $n(k)$ is complex white Gaussian noise. The received vector at the previous time sample $k-1$ is thus

$$\mathbf{r}(k-1) = r_1(k-1) + jr_2(k-1) = Ae^{j(\theta(k-1)+(\phi_0)} + n(k-1). \tag{5.63}$$

The phase difference between $\mathbf{r}(k)$ and $\mathbf{r}(k-1)$ determines which symbol was transmitted. Consider

$$\begin{aligned}\mathbf{r}(k)\mathbf{r}^*(k-1) =& A^2e^{j(\theta(k)-\theta(k-1))} + Ae^{j(\theta(k)+\phi_0)}n^*(k-1) \\ &+ Ae^{-j(\theta(k-1)+\phi_0)}n(k) + n(k)n^*(k-1).\end{aligned} \tag{5.64}$$

In the absence of noise $(n(k) = n(k-1) = 0)$ only the first term in (5.64) is nonzero, and this term yields the desired phase difference. The phase comparator in Figure 5.20 extracts this phase difference and outputs the corresponding symbol.

Differential modulation is less sensitive to a random drift in the carrier phase. However, if the channel has a nonzero Doppler frequency then the signal phase can decorrelate between symbol times, making the previous symbol a noisy phase reference. This decorrelation gives rise to an irreducible error floor for differential modulation over wireless channels with Doppler, as we shall discuss in Chapter 6.

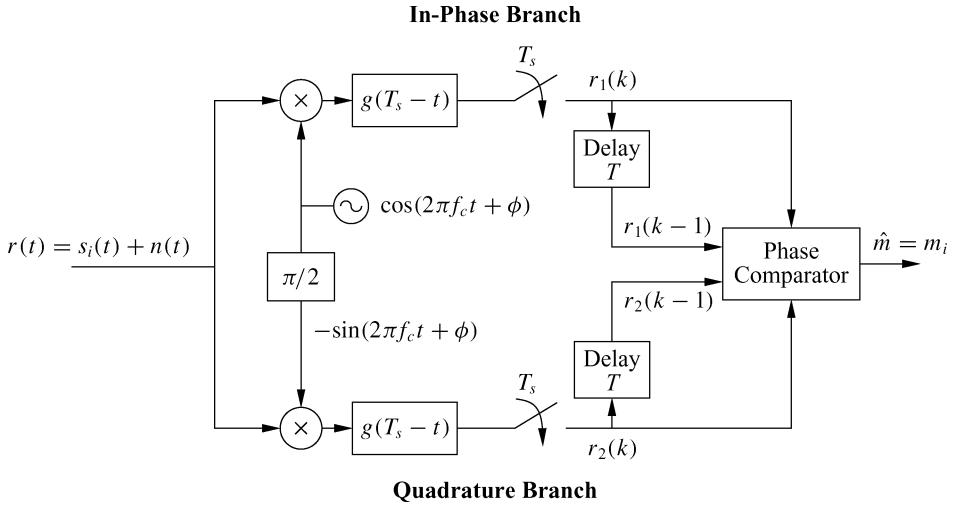

Figure 5.20: Differential PSK demodulator.

5.3.5 Constellation Shaping

Rectangular and hexagonal constellations have a better power efficiency than the square or circular constellations associated with MQAM and MPSK, respectively. These irregular constellations can save up to 1.3 dB of power at the xpense of increased complexity in the constellation map [7]. The optimal constellation shape is a sphere in N-dimensional space, which must be mapped to a sequence of constellations in two-dimensional space in order to be generated by the modulator shown in Figure 5.10. The general conclusion in [7] is that, for uncoded modulation, the increased complexity of spherical constellations is not worth their energy gains, since coding can provide much better performance at less complexity cost. However, if a complex channel code is already being used and little further improvement can be obtained by a more complex code, constellation shaping may obtain around 1 dB of additional gain. An in-depth discussion of constellation shaping (and of constellations that allow a noninteger number of bits per symbol) can be found in [7].

5.3.6 Quadrature Offset

A linearly modulated signal with symbol $\mathbf{s}_i = (s_{i1}, s_{i2})$ will lie in one of the four quadrants of the signal space. At each symbol time kT_s the transition to a new symbol value in a different quadrant can cause a phase transition of up to 180°, which may cause the signal amplitude to transition through the zero point; these abrupt phase transitions and large amplitude variations can be distorted by nonlinear amplifiers and filters. The abrupt transitions are avoided by offsetting the quadrature branch pulse $g(t)$ by half a symbol period, as shown in Figure 5.21. This *quadrature offset* makes the signal less sensitive to distortion during symbol transitions.

Phase modulation with quadrature offset is usually abbreviated as OMPSK, where the O indicates the offset. For example, QPSK modulation with quadrature offset is referred to as OQPSK. Offset QPSK has the same spectral properties as QPSK for linear amplification, but it has higher spectral efficiency under nonlinear amplification because the maximum phase transition of the signal is 90°, corresponding to the maximum phase transition in either the in-phase or quadrature branch but not both simultaneously. Another technique to mitigate the amplitude fluctuations of a 180° phase shift used in the IS-136 standard for digital cellular is $\pi/4$-QPSK [8, 9]. This technique allows for a maximum phase transition of 135° degrees, versus 90° for offset QPSK and 180° for QPSK. Thus, $\pi/4$-QPSK has worse spectral properties than OQPSK under nonlinear amplification. However, $\pi/4$-QPSK can be differentially encoded to eliminate the need for a coherent phase reference, which is a significant advantage. Using differential encoding with $\pi/4$-QPSK is called $\pi/4$-DQPSK. The $\pi/4$-DQPSK modulation works as follows: the information bits are first differentially encoded as in DQPSK, which yields one of the four QPSK constellation points. Then, every other symbol transmission is shifted in phase by $\pi/4$. This periodic phase shift has a similar effect as the time offset in OQPSK: it reduces the amplitude fluctuations at symbol transitions, which makes the signal more robust against noise and fading.

5.4 Frequency Modulation

Frequency modulation encodes information bits into the frequency of the transmitted signal. Specifically: at each symbol time, $K = \log_2 M$ bits are encoded into the frequency of the transmitted signal $s(t)$, $0 \leq t < T_s$, resulting in a transmitted signal $s_i(t) = A\cos(2\pi f_i t + \phi_i)$, where i is the index of the ith message corresponding to the $\log_2 M$ bits and ϕ_i is the phase associated with the ith carrier. The signal space representation is $s_i(t) = \sum_j s_{ij}\phi_j(t)$, where $s_{ij} = A\delta(i-j)$ and $\phi_j(t) = \cos(2\pi f_j t + \phi_j)$, so the basis functions correspond to carriers at different frequencies and only one such basis function is transmitted in each symbol period. The basis functions are orthogonal for a minimum carrier frequency separation of $\triangle f = \min_{ij} |f_j - f_i| = .5/T_s$ for $\phi_i = \phi_j$ and of $\triangle f = 1/T_s$ for $\phi_i \neq \phi_j$.

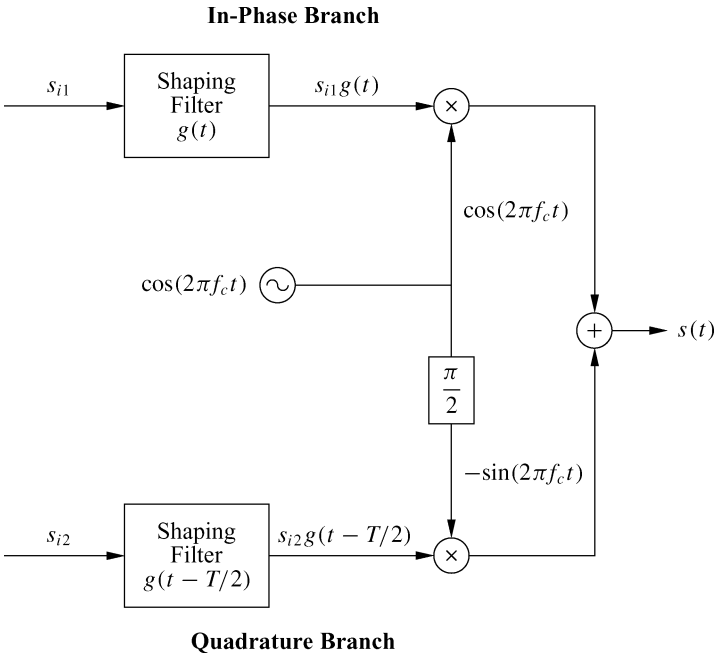

Figure 5.21: Modulator with quadrature offset.

Because frequency modulation encodes information in the signal frequency, the transmitted signal $s(t)$ has a constant envelope A. Since the signal is constant envelope, nonlinear amplifiers can be used with high power efficiency and hence the modulated signal is less sensitive to amplitude distortion introduced by the channel or the hardware. The price exacted for this robustness is a lower spectral efficiency: because the modulation technique is nonlinear, it tends to have a higher bandwidth occupancy than the amplitude and phase modulation techniques described in Section 5.3.

In its simplest form, frequency modulation over a given symbol period can be generated using the modulator structure shown in Figure 5.22. Demodulation over each symbol period is performed using the demodulation structure of Figure 5.23. Note that the demodulator of Figure 5.23 requires the jth carrier signal to be matched in phase to the jth carrier signal at the transmitter; this is similar to the coherent phase reference requirement in amplitude and phase modulation. An alternate receiver structure that does not require this coherent phase reference will be discussed in Section 5.4.3. Another issue in frequency modulation is that the different carriers shown in Figure 5.22 have different phases, $\phi_i \neq \phi_j$ for $i \neq j$, so at each symbol time T_s there will be a phase discontinuity in the transmitted signal. Such discontinuities can significantly increase signal bandwidth. Thus, in practice an alternate modulator is used that generates a frequency-modulated signal with continuous phase, as will be discussed in Section 5.4.2.

5.4.1 Frequency-Shift Keying (FSK) and Minimum-Shift Keying (MSK)

In MFSK the modulated signal is given by

$$s_i(t) = A\cos[2\pi f_c t + 2\pi\alpha_i \triangle f_c t + \phi_i], 0 \leq t < T_s, \tag{5.65}$$

where $\alpha_i = (2i - 1 - M)$ for $i = 1, 2, \ldots, M = 2^K$. The minimum frequency separation between FSK carriers is thus $2\triangle f_c$. MFSK consists of M basis functions $\phi_i(t) = \sqrt{2/T_s}\cos[2\pi f_c t + 2\pi\alpha_i \triangle f_c t + \phi_i], i = 1, \ldots, M$,

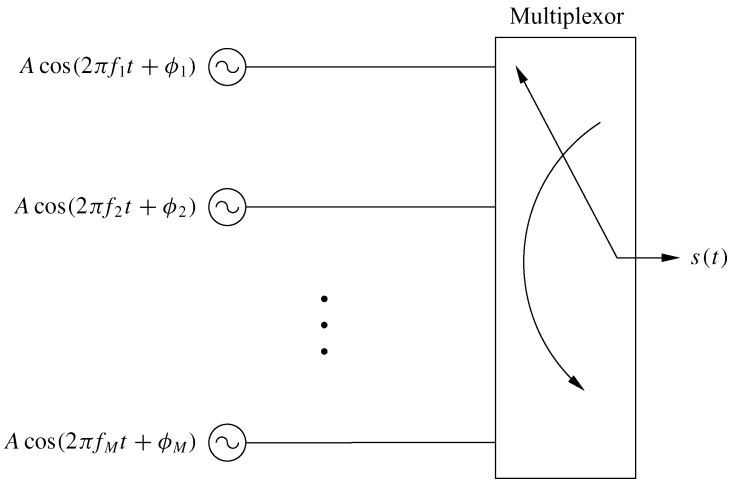

Figure 5.22: Frequency modulator.

where the $\sqrt{2/T_s}$ is a normalization factor to ensure that $\int_0^{T_s} \phi_i^2(t) = 1$. Over any given symbol time, only one basis function is transmitted through the channel.

A simple way to generate the MFSK signal is as shown in Figure 5.22, where M oscillators are operating at the different frequencies $f_i = f_c + \alpha_i \triangle f_c$ and the modulator switches between these different oscillators each symbol time T_s. However, this implementation entails a discontinuous phase transition at the switching times due to phase offsets between the oscillators, and this discontinuous phase leads to undesirable spectral broadening. (An FSK modulator that maintains continuous phase is discussed in the next section.) Coherent detection of MFSK uses the standard structure of Figure 5.23. For binary signaling the structure can be simplified to that shown in Figure 5.24, where the decision device outputs a 1-bit if its input is greater than zero and a 0-bit if its input is less than zero.

MSK is a special case of binary FSK where $\phi_1 = \phi_2$ and the frequency separation is $2\triangle f_c = .5/T_s$. Note that this is the minimum frequency separation that ensure $\langle s_i(t), s_j(t)\rangle = 0$ over a symbol time for $i \neq j$. Since signal orthogonality is required for demodulation, it follows that $2\triangle f_c = .5/T_s$ is the minimum possible frequency separation in FSK and so MSK is the minimum bandwidth FSK modulation.

5.4.2 Continuous-Phase FSK (CPFSK)

A better way to generate MFSK – one that eliminates the phase discontinuity – is to frequency modulate a single carrier with a modulating waveform, as in analog FM. In this case the modulated signal will be given by

$$s_i(t) = A\cos\left[2\pi f_c t + 2\pi\beta \int_{-\infty}^{t} u(\tau)\, d\tau\right] = A\cos[2\pi f_c t + \theta(t)], \tag{5.66}$$

where $u(t) = \sum_k a_k g(t - kT_s)$ is an MPAM signal modulated with the information bit stream, as described in Section 5.3.1. Clearly the phase $\theta(t)$ is continuous with this implementation. This form of MFSK is therefore called continuous-phase FSK, or CPFSK.

By Carson's rule [10], for β small the transmission bandwidth of $s(t)$ is approximately

$$B_s \approx 2M\triangle f_c + 2B_g, \tag{5.67}$$

where B_g is the bandwidth of the pulse shape $g(t)$ used in the MPAM modulating signal $u(t)$. By comparison, the bandwidth of a linearly modulated waveform with pulse shape $g(t)$ is roughly $B_s \approx 2B_g$. Thus, the spectral

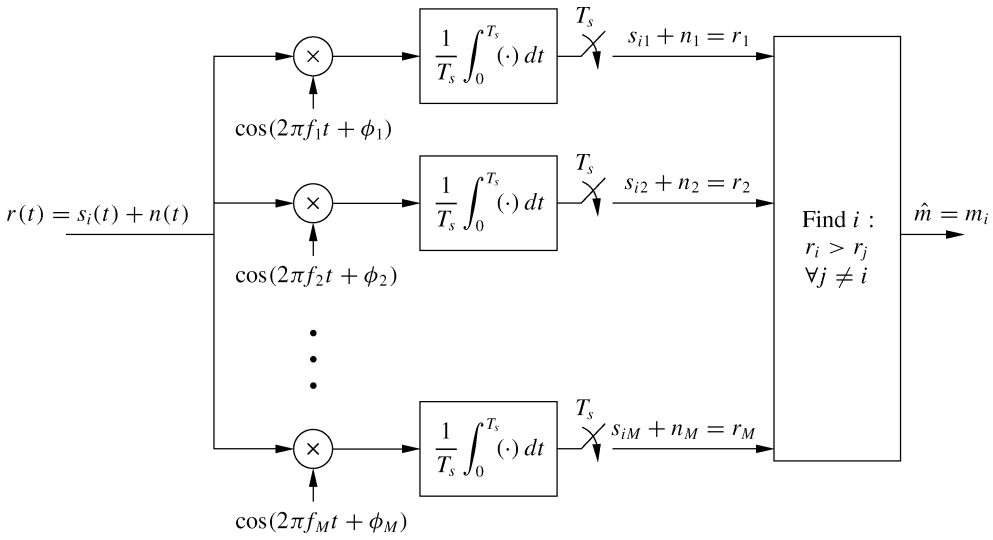

Figure 5.23: Frequency demodulator (coherent).

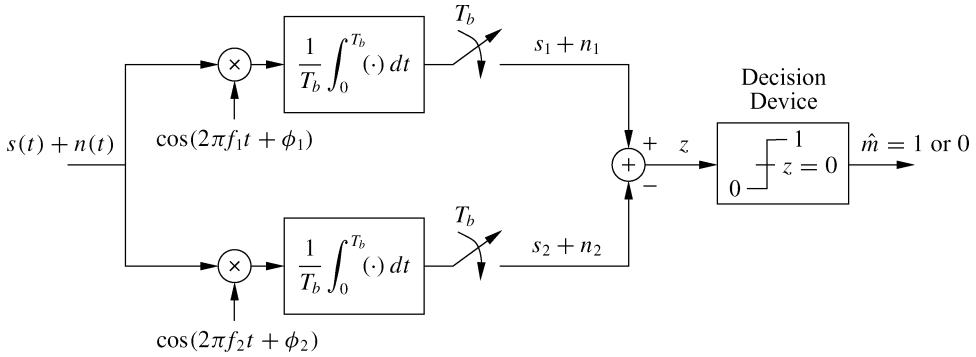

Figure 5.24: Demodulator for binary FSK.

occupancy of a CPFSK-modulated signal is larger than that of a linearly modulated signal by $M\triangle f_c \geq .5M/T_s$. The spectral efficiency penalty of CPFSK relative to linear modulation increases with data rate, in particular with the number of bits per symbol $K = \log_2 M$ and with the symbol rate $R_s = 1/T_s$.

Coherent detection of CPFSK can be done symbol-by-symbol or over a sequence of symbols. The sequence estimator is the optimal detector, since a given symbol depends on previously transmitted symbols and so it is optimal to detect (or estimate) all symbols simultaneously. However, sequence estimation can be impractical owing to the memory and computational requirements associated with making decisions based on sequences of symbols. Details on detectors for coherent demodulation of CPFSK can be found in [11, Chap. 4.9].

5.4.3 Noncoherent Detection of FSK

The receiver requirement for a coherent phase reference associated with each FSK carrier can be difficult and expensive to meet. The need for a coherent phase reference can be eliminated if the receiver first detects the energy of the signal at each frequency and, if the ith branch has the highest energy of all branches, then outputs message m_i. The modified receiver is shown in Figure 5.25.

Suppose the transmitted signal corresponds to frequency f_i:

$$
\begin{aligned}
s(t) &= A\cos(2\pi f_i t + \phi_i) \\
&= A\cos(\phi_i)\cos(2\pi f_i t) - A\sin(\phi_i)\sin(2\pi f_i t),\ 0 \le t < T_s. \qquad (5.68)
\end{aligned}
$$

Let the phase ϕ_i represent the phase offset between the transmitter and receiver oscillators at frequency f_i. A coherent receiver with carrier signal $\cos(2\pi f_i t)$ detects only the first term $A\cos(\phi_i)\cos(2\pi f_i t)$ associated with the received signal, which can be close to zero for a phase offset $\phi_i \approx \pm\pi/2$. To get around this problem, in Figure 5.25 the receiver splits the received signal into M branches corresponding to each frequency $f_j, j = 1, \ldots, M$. For each such carrier frequency f_j, the received signal is multiplied by a noncoherent in-phase and quadrature carrier at that frequency, integrated over a symbol time, sampled, and then squared. For the jth branch the squarer output associated with the in-phase component is denoted as $A_{jI} + n_{jI}$ and the corresponding output associated with the quadrature component is denoted as $A_{jQ} + n_{jQ}$, where n_{jI} and n_{jQ} are due to the noise $n(t)$ at the receiver input. Then, if $i = j$, we have $A_{jI} = A^2\cos^2(\phi_i)$ and $A_{jQ} = A^2\sin^2(\phi_i)$; if $i \neq j$ then $A_{jI} = A_{jQ} = 0$. In the absence of noise, the input to the decision device of the ith branch will be $A^2\cos^2(\phi_i) + A^2\sin^2(\phi_i) = A^2$, independent of ϕ_i, and all other branches will have an input of zero. Thus, over each symbol period, the decision device outputs the bit sequence corresponding to frequency f_j if the jth branch has the largest input to the decision device. Note that the noncoherent receiver of Figure 5.25 still requires accurate synchronization for sampling. Synchronization issues are discussed in Section 5.6.

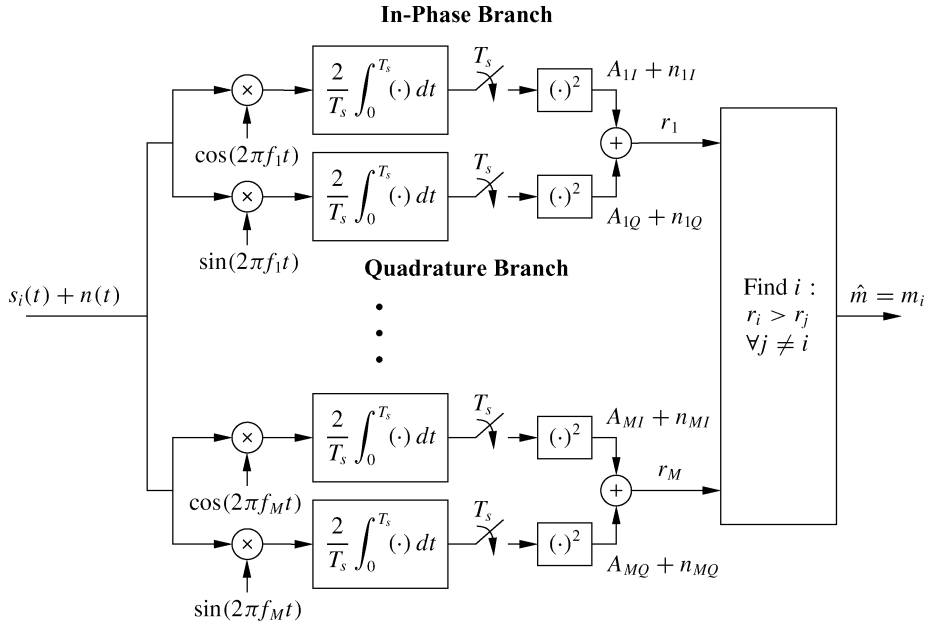

Figure 5.25: Noncoherent FSK demodulator.

5.5 Pulse Shaping

For amplitude and phase modulation, the bandwidth of the baseband and passband modulated signal is a function of the bandwidth of the pulse shape $g(t)$. If $g(t)$ is a rectangular pulse of width T_s, then the envelope of the signal is constant. However, a rectangular pulse has high spectral sidelobes, which can cause adjacent channel interference. Pulse shaping is a method for reducing sidelobe energy relative to a rectangular pulse; however, the shaping must

be done in such a way that intersymbol interference between pulses in the received signal is not introduced. Note that – prior to sampling the received signal – the transmitted pulse $g(t)$ is convolved with the channel impulse response $c(t)$ and the matched filter $g^*(-t)$; hence, in order to eliminate ISI prior to sampling, we must ensure that the effective received pulse $p(t) = g(t) * c(t) * g^*(-t)$ has no ISI. Since the channel model is AWGN, we assume $c(t) = \delta(t)$ so $p(t) = g(t) * g^*(-t)$ (in Chapter 11we will analyze ISI for more general channel impulse responses $c(t)$). To avoid ISI between samples of the received pulses, the effective pulse shape $p(t)$ must satisfy the *Nyquist criterion*, which requires the pulse to equal zero at the ideal sampling point associated with past or future symbols:

$$p(kT_s) = \begin{cases} p_0 = p(0) & k = 0, \\ 0 & k \neq 0. \end{cases}$$

In the frequency domain this translates to

$$\sum_{l=-\infty}^{\infty} P\left(f + \frac{l}{T_s}\right) = p_0 T_s. \tag{5.69}$$

The following pulse shapes all satisfy the Nyquist criterion.

1. *Rectangular pulses:* $g(t) = \sqrt{2/T_s}(0 \leq t \leq T_s)$, which yields the triangular effective pulse shape

$$p(t) = \begin{cases} 2 + 2t/T_s & -T_s \leq t < 0, \\ 2 - 2t/T_s & 0 \leq t < T_s, \\ 0 & \text{else.} \end{cases}$$

This pulse shape leads to constant envelope signals in MPSK but has poor spectral properties as a result of its high sidelobes.

2. *Cosine pulses:* $p(t) = \sin \pi t/T_s, 0 \leq t \leq T_s$. Cosine pulses are mostly used in OMPSK modulation, where the quadrature branch of the modulation is shifted in time by $T_s/2$. This leads to a constant amplitude modulation with sidelobe energy that is 10 dB lower than that of rectangular pulses.

3. *Raised cosine pulses:* These pulses are designed in the frequency domain according to the desired spectral properties. Thus, the pulse $p(t)$ is first specified relative to its Fourier transform:

$$P(f) = \begin{cases} T_s & 0 \leq |f| \leq \frac{1-\beta}{2T_s}, \\ \frac{T_s}{2}\left[1 - \sin\frac{\pi T_s}{\beta}\left(f - \frac{1}{2T_s}\right)\right] & \frac{1-\beta}{2T_s} \leq |f| \leq \frac{1+\beta}{2T_s}; \end{cases}$$

here β is defined as the rolloff factor, which determines the rate of spectral rolloff (see Figure 5.26). Setting $\beta = 0$ yields a rectangular pulse. The pulse $p(t)$ in the time domain corresponding to $P(f)$ is

$$p(t) = \frac{\sin \pi t/T_s}{\pi t/T_s} \frac{\cos \beta \pi t/T_s}{1 - 4\beta^2 t^2/T_s^2}.$$

The frequency- and time-domain properties of the raised cosine pulse are shown in Figures 5.26 and 5.27, respectively. The tails of this pulse in the time domain decay as $1/t^3$ (faster than for the previous pulse shapes), so a mistiming error in sampling leads to a series of inter- symbol interference components that converge. A variation of the raised cosine pulse is the root cosine pulse, derived by taking the square root of the frequency response for the raised cosine pulse. The root cosine pulse has better spectral properties than the raised cosine pulse but decays

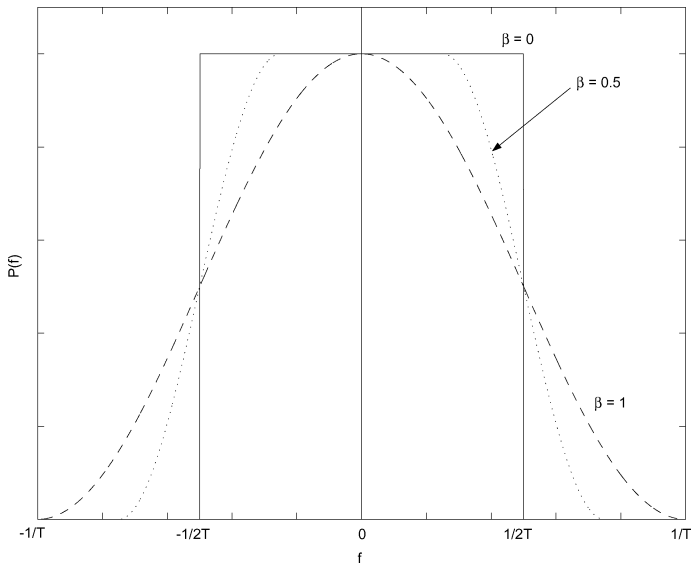

Figure 5.26: Frequency-domain (spectral) properties of the raised cosine pulse ($T = T_s$).

less rapidly in the time domain, which makes performance degradation due to synchronization errors more severe. Specifically, a mistiming error in sampling leads to a series of ISI components that may diverge.

Pulse shaping is also used with CPFSK to improve spectral efficiency, specifically in the MPAM signal that is frequency modulated to form the FSK signal. The most common pulse shape used in CPFSK is the Gaussian pulse shape, defined as

$$g(t) = \frac{\sqrt{\pi}}{\alpha} e^{-\pi^2 t^2/\alpha^2}, \tag{5.70}$$

where α is a parameter that dictates spectral efficiency. The spectrum of $g(t)$, which dictates the spectrum of the CPFSK signal, is given by

$$G(f) = e^{-\alpha^2 f^2}. \tag{5.71}$$

The parameter α is related to the 3-dB bandwidth of $g(t)$, B_g, by

$$\alpha = \frac{.5887}{B_g}. \tag{5.72}$$

Clearly, increasing α results in a higher spectral efficiency.

When the Gaussian pulse shape is applied to MSK modulation, it is abbreviated as GMSK. In general, GMSK signals have a high power efficiency (since they have a constant amplitude) and a high spectal efficiency (since the Gaussian pulse shape has good spectral properties for large α). For these reasons, GMSK is used in the GSM standard for digital cellular systems. Although this is a good choice for voice modulation, it is not the best choice for data. The Gaussian pulse shape does not satisfy the Nyquist criterion and so the pulse shape introduces ISI, which increases as α increases. Thus, improving spectral efficiency by increasing α leads to a higher ISI level, thereby creating an irreducible error floor from this self-interference. Since the required BER for voice is a relatively high $P_b \approx 10^{-3}$, the ISI can be fairly high and still maintain this target BER. In fact, it is generally used as a rule of thumb that $B_g T_s = .5$ is a tolerable amount of ISI for voice transmission with GMSK. However, a much lower BER is required for data, which will put more stringent constraints on the maximum α and corresponding

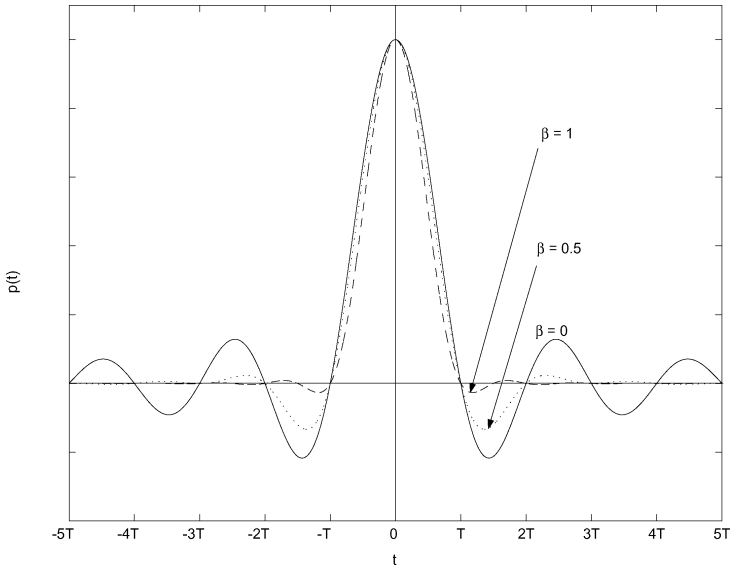

Figure 5.27: Time-domain properties of the raised cosine pulse ($T = T_s$).

minimum B_g, thereby decreasing the spectral efficiency of GMSK for data transmission. Techniques such as equalization can be used to mitigate the ISI in this case so that a tolerable BER is possible without significantly compromising spectral efficiency. However, it is more common to use linear modulation for spectrally efficient data transmission. Indeed, the data enhancements to GSM use linear modulation.

5.6 Symbol Synchronization and Carrier Phase Recovery

One of the most challenging tasks of a digital demodulator is to acquire accurate symbol timing and carrier phase information. Timing information, obtained via synchronization, is needed to delineate the received signal associated with a given symbol. In particular, timing information is used to drive the sampling devices associated with the demodulators for amplitude, phase, and frequency demodulation shown in Figures 5.11 and 5.23. Carrier phase information is needed in all coherent demodulators for both amplitude/phase and frequency modulation, as discussed in Sections 5.3 and 5.4.

This section gives a brief overview of standard techniques for synchronization and carrier phase recovery in AWGN channels. In this context the estimation of symbol timing and carrier phase falls under the broader category of signal parameter estimation in noise. Estimation theory provides the theoretical framework for studying this problem and for developing the maximum likelihood estimator of the carrier phase and symbol timing. However, most wireless channels suffer from time-varying multipath in addition to AWGN. Synchronization and carrier phase recovery is particularly challenging in such channels because multipath and time variations can make it extremely difficult to estimate signal parameters prior to demodulation. Moreover, there is little theory addressing good methods for estimation of carrier phase and symbol timing when these parameters are corrupted by time-varying multipath in addition to noise. In most performance analysis of wireless communication systems it is assumed that the receiver synchronizes to the multipath component with delay equal to the average delay spread;[3] then the channel is treated as AWGN for recovery of timing information and carrier phase. In practice, however, the receiver will sychronize to either the strongest multipath component or the first multipath component that exceeds a given power threshold. The other multipath components will then compromise the receiver's ability to acquire

[3]That is why delay spread is typically characterized by its rms value about its mean, as discussed in more detail in Chapter 2.

timing and carrier phase, especially in wideband systems like UWB. Multicarrier and spread-spectrum systems have additional considerations related to synchronization and carrier recovery, which will be discussed in Chapters 12 and 13.

The importance of synchronization and carrier phase estimation cannot be overstated: without them, wireless systems could not function. Moreover, as data rates increase and channels become more complex by adding additional degrees of freedom (e.g., multiple antennas), the tasks of receiver synchronizaton and phase recovery become even more complex and challenging. Techniques for synchronization and carrier recovery have been developed and analyzed extensively for many years, and they are continually evolving to meet the challenges associated with higher data rates, new system requirements, and more challenging channel characteristics. We give only a brief introduction to synchronizaton and carrier phase recovery techniques in this section. Comprehensive coverage of this topic and performance analysis of these techniques can be found in [12, 13]; more condensed treatments can be found in [bib5-10, Chap. 5;bib5-13].

5.6.1 Receiver Structure with Phase and Timing Recovery

The carrier phase and timing recovery circuitry for the amplitude and phase demodulator is shown in Figure 5.28. For BPSK only the in-phase branch of this demodulator is needed. For the coherent frequency demodulator of Figure 5.23, a carrier phase recovery circuit is needed for *each* of the distinct M carriers; the resulting circuit complexity motivates the need for noncoherent demodulators as described in Section 5.4.3. We see in Figure 5.28 that the carrier phase and timing recovery circuits operate directly on the received signal prior to demodulation.

Assuming an AWGN channel, the received signal $r(t)$ is a delayed version of the transmitted signal $s(t)$ plus AWGN $n(t)$: $r(t) = s(t-\tau) + n(t)$, where τ is the random propagation delay. Using the equivalent lowpass form we have $s(t) = \mathrm{Re}\{x(t)e^{j\phi_0}e^{j2\pi f_c t}\}$ and thus

$$r(t) = \mathrm{Re}\left\{x(t-\tau)e^{j\phi}e^{j2\pi f_c t}\right\} + n(t), \tag{5.73}$$

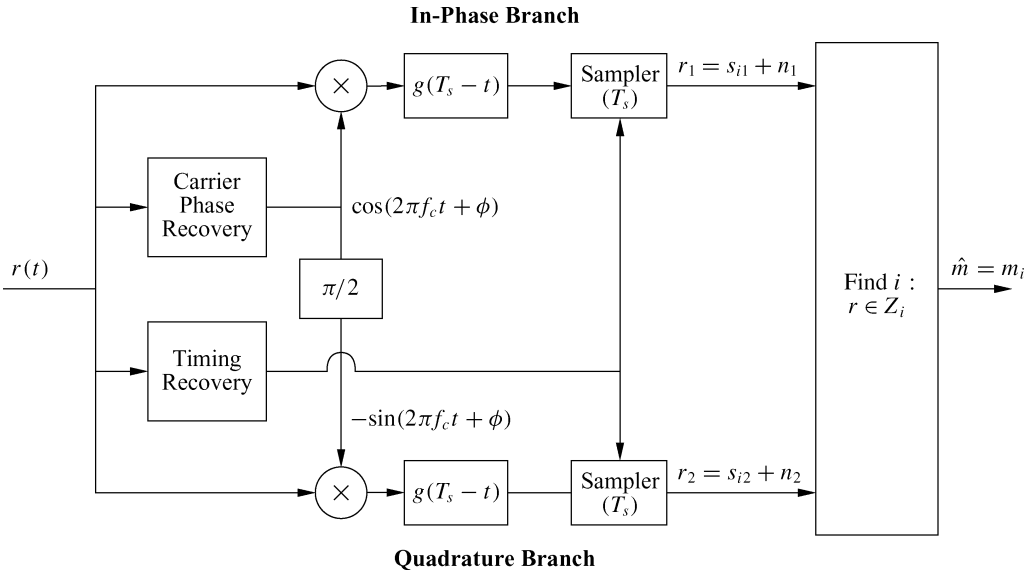

Figure 5.28: Receiver structure with carrier and timing recovery.

where $\phi = \phi_0 - 2\pi f_c \tau$ results from the transmit carrier phase and the propagation delay. Estimation of τ is needed for symbol timing, and estimation of ϕ is needed for carrier phase recovery. Let us express these two unknown

parameters as a vector $\psi = (\phi, \tau)$. Then we can express the received signal in terms of ψ as

$$r(t) = s(t; \psi) + n(t). \tag{5.74}$$

Parameter estimation must take place over some finite time interval $T_0 \geq T_s$. We call T_0 the *observation interval*. In practice, however, parameter estimation is initially done over this interval and is thereafter performed continually by updating the initial estimate using tracking loops. Our development here focuses on the initial parameter estimation over T_0. Discussion of parameter tracking can be found in [12, 13].

There are two common estimation methods for signal parameters in noise: the maximum likelihood (ML) criterion discussed in Section 5.1.4 in the context of receiver design, and the maximum a posteriori (MAP) criterion. The ML criterion chooses the estimate $\hat{\psi}$ that maximizes $p(r(t)|\psi)$ over the observation interval T_0, whereas the MAP criterion assumes some probability distribution $p(\psi)$ on ψ and then chooses the estimate $\hat{\psi}$ that maximizes

$$p(\psi|r(t)) = \frac{p(r(t)|\psi)p(\psi)}{p(r(t))}$$

over T_0. We assume that there is no prior knowledge of $\hat{\psi}$, so that $p(\psi)$ becomes uniform and hence the MAP and ML criteria are equivalent.

To characterize the distribution $p(r(t)|\psi)$, $0 \leq t < T_0$, let us expand $r(t)$ over the observation interval along a set of orthonormal basis functions $\{\phi_k(t)\}$ as

$$r(t) = \sum_{k=1}^{K} r_k \phi_k(t), \quad 0 \leq t < T_0.$$

Because $n(t)$ is white with zero mean and power spectral density $N_0/2$, the distribution of the vector $\mathbf{r} = (r_1, \ldots, r_K)$ conditioned on the unknown parameter ψ is given by

$$p(\mathbf{r}|\psi) = \left(\frac{1}{\sqrt{\pi N_0}}\right)^K \exp\left[-\sum_{k=1}^{K} \frac{(r_k - s_k(\psi))^2}{N_0}\right], \tag{5.75}$$

where (by the basis expansion)

$$r_k = \int_{T_0} r(t)\phi_k(t)\, dt$$

and

$$s_k(\psi) = \int_{T_0} s(t; \psi)\phi_k(t)\, dt.$$

From these basis expansions we can show that

$$\sum_{k=1}^{K} [r_k - s_k(\psi)]^2 = \int_{T_0} [r(t) - s(t; \psi)]^2\, dt. \tag{5.76}$$

Using this in (5.75) yields that maximizing $p(\mathbf{r}|\psi)$ is equivalent to maximizing the *likelihood function*

$$\Lambda(\psi) = \exp\left[-\frac{1}{N_0}\int_{T_0} [r(t) - s(t; \psi)]^2\right] dt. \tag{5.77}$$

Maximizing the likelihood function (5.77) results in the joint ML estimate of the carrier phase and symbol timing. Maximum likelihood estimation of the carrier phase and symbol timing can also be done separately, and in subsequent sections we will discuss this separate estimation in more detail. Techniques joint estimation are more complex; details of such techniques can be found in [10, Chap. 6.4; 11, Chaps. 8–9].

5.6.2 Maximum Likelihood Phase Estimation

In this section we derive the maximum likelihood phase estimate assuming the timing is known. The likelihood function (5.77) with timing known reduces to

$$\begin{aligned}\Lambda(\phi) &= \exp\left[-\frac{1}{N_0}\int_{T_0}[r(t)-s(t;\phi)]^2\,dt\right] \\ &= \exp\left[-\frac{1}{N_0}\int_{T_0}x^2(t)\,dt+\frac{2}{N_0}\int_{T_0}r(t)s(t;\phi)\,dt-\frac{1}{N_0}\int_{T_0}s^2(t;\phi)\,dt\right].\end{aligned} \tag{5.78}$$

We estimate the carrier phase as the value $\hat{\phi}$ that maximizes this function. Note that the first term in (5.78) is independent of ϕ. Moreover, we assume that the third integral, which measures the energy in $s(t;\phi)$ over the observation interval, is relatively constant in ϕ. Given these assumptions, we see that the $\hat{\phi}$ that maximizes (5.78) also maximizes

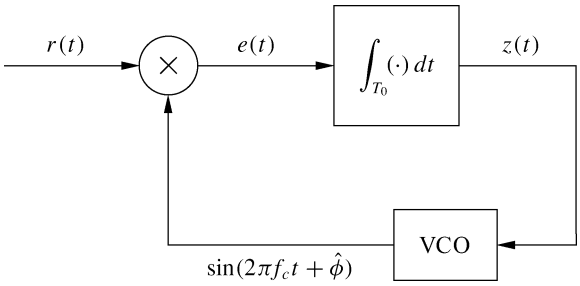

Figure 5.29: Phase-locked loop for carrier phase recovery (unmodulated carrier).

$$\Lambda'(\phi) = \int_{T_0} r(t)s(t;\phi)\,dt. \tag{5.79}$$

We can solve directly for the maximizing $\hat{\phi}$ in the simple case where the received signal is just an unmodulated carrier plus noise: $r(t) = A\cos(2\pi f_c t+\phi)+n(t)$. Then $\hat{\phi}$ must maximize

$$\Lambda'(\phi) = \int_{T_0} r(t)\cos(2\pi f_c t+\phi)\,dt. \tag{5.80}$$

Differentiating $\Lambda'(\phi)$ relative to ϕ and then setting it to zero yields that $\hat{\phi}$ satisfies

$$\int_{T_0} r(t)\sin(2\pi f_c t+\hat{\phi})\,dt = 0. \tag{5.81}$$

Solving (5.81) for $\hat{\phi}$ yields

$$\hat{\phi} = -\tan^{-1}\left[\frac{\int_{T_0} r(t)\sin(2\pi f_c t)\,dt}{\int_{T_0} r(t)\cos(2\pi f_c t)\,dt}\right]. \tag{5.82}$$

We could build a circuit to compute (5.82) from the received signal $r(t)$; in practice, however, carrier phase recovery is accomplished by using a phase lock loop to satisfy (5.81), as shown in Figure 5.29. In this figure, the integrator input in the absence of noise is given by $e(t) = r(t)\sin(2\pi f_c t + \hat{\phi})$ and the integrator output is

$$z(t) = \int_{T_0} r(t)\sin(2\pi f_c t + \hat{\phi})dt,$$

which is precisely the left-hand side of (5.81). Thus, if $z(t) = 0$ then the estimate $\hat{\phi}$ is the maximum likelihood estimate for ϕ. If $z(t) \neq 0$ then the voltage-controlled oscillator (VCO) adjusts its phase estimate $\hat{\phi}$ up or down depending on the polarity of $z(t)$: for $z(t) > 0$ it decreases $\hat{\phi}$ to reduce $z(t)$, and for $z(t) < 0$ it increases $\hat{\phi}$ to increase $z(t)$. In practice the integrator in Figure 5.29 is replaced with a *loop filter* whose output $.5A\sin(\hat{\phi} - \phi) \approx .5A(\hat{\phi} - \phi)$ is a function of the low-frequency component of its input $e(t) = A\cos(2\pi f_c t + \phi)\sin(2\pi f_c t + \hat{\phi}) = .5A\sin(\hat{\phi} - \phi) + .5A\sin(2\pi f_c t + \phi + \hat{\phi})$. This discussion of phase-locked loop (PLL) operation assumes that $\hat{\phi} \approx \phi$ because otherwise the polarity of $z(t)$ may not indicate the correct phase adjustment; that is, we would not necessarily have $\sin(\hat{\phi} - \phi) \approx \hat{\phi} - \phi$. The PLL typically exhibits some transient behavior in its initial estimation of the carrier phase. The advantage of a PLL is that it continually adjusts its estimate $\hat{\phi}$ to maintain $z(t) \approx 0$, which corrects for slow phase variations due to oscillator drift at the transmitter or changes in the propagation delay. In fact, the PLL is an example of a feedback control loop. More details on the PLL and its performance can be found in [11, 12].

The PLL derivation is for an unmodulated carrier, yet amplitude and phase modulation embed the message bits into the amplitude and phase of the carrier. For such signals there are two common carrier phase recovery approaches to deal with the effect of the data sequence on the received signal: the data sequence is either (a) assumed known or (b) treated as random with the phase estimate averaged over the data statistics. The first scenario is referred to as *decision-directed* parameter estimation, and this scenario typically results from sending a known training sequence. The second scenario is referred to as *non-decision directed* parameter estimation. With this technique the likelihood function (5.78) is maximized by averaging over the statistics of the data. One decision-directed technique uses data decisions to remove the modulation of the received signal: the resulting unmodulated carrier is then passed through a PLL. This basic structure is called a *decision-feedback PLL* because data decisions are fed back into the PLL for processing. The structure of a non–decision-directed carrier phase recovery loop depends on the underlying distribution of the data. For large constellations, most distributions lead to highly nonlinear functions of the parameter to be estimated. In this case the symbol distribution is often assumed to be Gaussian along each signal dimension, which greatly simplifies the recovery loop structure. An alternate non–decision-directed structure takes the Mth power of the signal ($M = 2$ for PAM and M for MPSK modulation), passes it through a bandpass filter at frequency Mf_c, and then uses a PLL. The nonlinear operation removes the effect of the amplitude or phase modulation so that the PLL can operate on an unmodulated carrier at frequency Mf_c. Many other structures for both decision-directed and non–decision-directed carrier recovery can be used, with different trade-offs in performance and complexity. A more comprehensive discussion of design and performance of carrier phase recovery can be found in [10, Chaps. 6.2.4–6.2.5; 11].

5.6.3 Maximum Likelihood Timing Estimation

In this section we derive the maximum likelihood estimate of delay τ assuming the carrier phase is known. Since we assume that the phase ϕ is known, the timing recovery will not affect the carrier phase recovery loop and associated downconversion shown in Figure 5.28. Thus, it suffices to consider timing estimation for the in-phase or quadrature equivalent lowpass signals of $r(t)$ and $s(t;\tau)$. We denote the in-phase and quadrature components for $r(t)$ as $r_I(t)$ and $r_Q(t)$ and for $s(t;\tau)$ as $s_I(t;\tau)$ and $s_Q(t;\tau)$. We focus on the in-phase branch since the timing recovered from this branch can be used for the quadrature branch. The equivalent lowpass in-phase signal is given by

$$s_I(t;\tau) = \sum_k s_I(k) g(t - kT_s - \tau), \tag{5.83}$$

where $g(t)$ is the pulse shape and $s_I(k)$ denotes the amplitude associated with the in-phase component of the message transmitted over the kth symbol period. The in-phase equivalent lowpass received signal is $r_I(t) = s_I(t;\tau) + n_I(t)$. As in the case of phase synchronization, there are two categories of timing estimators: those for which the information symbols output from the demodulator are assumed known (decision-directed estimators), and those for which this sequence is not assumed known (non–decision-directed estimators).

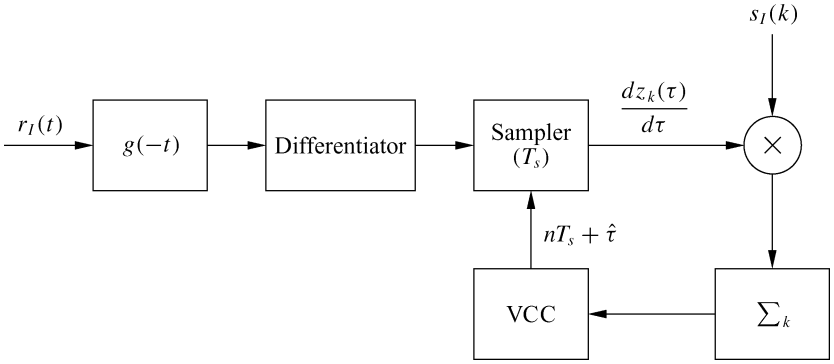

Figure 5.30: Decision-directed timing estimation.

The likelihood function (5.77) with known phase ϕ has a form similar to (5.78), the case of known delay:

$$\begin{aligned}\Lambda(\tau) &= \exp\left[-\frac{1}{N_0}\int_{T_0}[r_I(t) - s_I(t;\tau)]^2\,dt\right] \\ &= \exp\left[-\frac{1}{N_0}\int_{T_0} r_I^2(t)\,dt + \frac{2}{N_0}\int_{T_0} r_I(t)s_I(t;\tau)\,dt - \frac{1}{N_0}\int_{T_0} s_I^2(t;\tau)\,dt\right].\end{aligned} \tag{5.84}$$

Since the first and third terms in (5.84) do not change significantly with τ, the delay estimate $\hat{\tau}$ that maximizes (5.84) also maximizes

$$\begin{aligned}\Lambda'(\tau) &= \int_{T_0} r_I(t)s_I(t;\tau)\,dt \\ &= \sum_k s_I(k)\int_{T_0} r(t)g(t - kT_s - \tau)\,dt = \sum_k s_I(k)z_k(\tau),\end{aligned} \tag{5.85}$$

where

$$z_k(\tau) = \int_{T_0} r(t)g(t - kT_s - \tau)\,dt. \tag{5.86}$$

Differentiating (5.85) relative to τ and then setting it to zero yields that the timing estimate $\hat{\tau}$ must satisfy

$$\sum_k s_I(k)\frac{d}{d\tau}z_k(\tau) = 0. \tag{5.87}$$

For decision-directed estimation, (5.87) gives rise to the estimator shown in Figure 5.30. The input to the voltage-controlled clock (VCC) is (5.87). If this input is zero, then the timing estimate $\hat{\tau} = \tau$. If not, the clock (i.e., the timing estimate $\hat{\tau}$) is adjusted to drive the VCC input to zero. This timing estimation loop is also an example of a feedback control loop.

One structure for non–decision-directed timing estimation is the *early–late gate synchronizer* shown in Figure 5.31. This structure exploits two properties of the autocorrelation of $g(t)$, $R_g(\tau) = \int_0^{T_S} g(t)g(t-\tau)\,dt$ – namely, its symmetry ($R_g(\tau) = R_g(-\tau)$) and the fact that its maximum value is at $\tau = 0$. The input to the sampler in the upper branch of Figure 5.31 is proportional to the autocorrelation $R_g(\hat{\tau}-\tau+\delta) = \int_0^{T_S} g(t-\tau)g(t-\hat{\tau}-\delta)\,dt$, and the input to the sampler in the lower branch is proportional to the autocorrelation $R_g(\hat{\tau}-\tau-\delta) = \int_0^{T_S} g(t-\tau)g(t-\hat{\tau}+\delta)dt$. If $\hat{\tau} = \tau$ then, since $R_g(\delta) = R_g(-\delta)$, the input to the loop filter will be zero and the voltage-controlled clock will maintain its correct timing estimate. If $\hat{\tau} > \tau$ then $R_g(\hat{\tau} - \tau + \delta) < R_g(\hat{\tau} - \tau - \delta)$, and this negative input to the VCC will cause it to decrease its estimate of $\hat{\tau}$. Conversely, if $\hat{\tau} < \tau$ then $R_g(\hat{\tau} - \tau + \delta) > R_g(\hat{\tau} - \tau - \delta)$, and this positive input to the VCC will cause it to increase its estimate of $\hat{\tau}$.

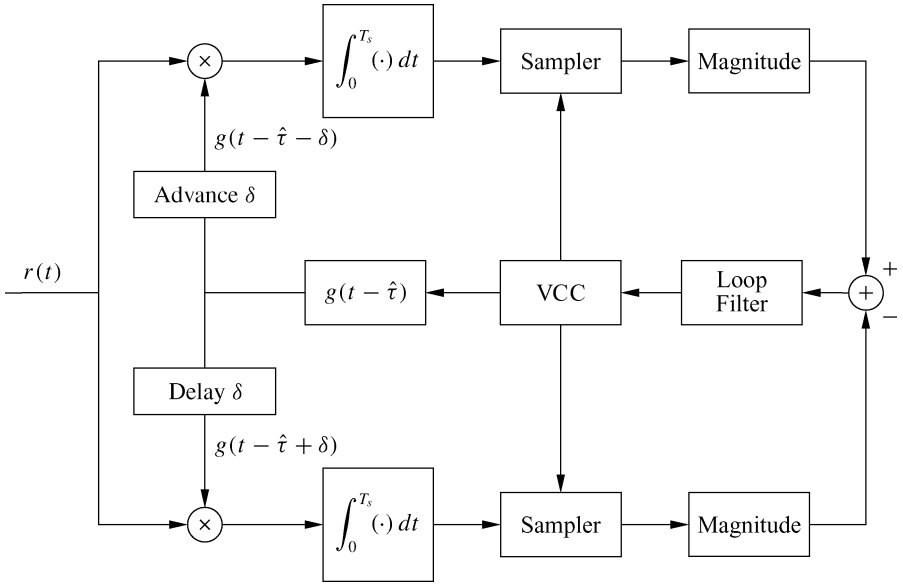

Figure 5.31: Early-late gate synchronizer.

More details on these and other structures for decision-directed and non–decision-directed timing estimation– as well as their performance trade-offs–can be found in [10, Chaps. 6.2.4–6.2.5; 11].

Chapter 5 Problems

1. Using properties of orthonormal basis functions, show that if $s_i(t)$ and $s_j(t)$ have constellation points $\mathbf{s}_i$ and $\mathbf{s}_j$ (respectively) then

$$\|\mathbf{s}_i - \mathbf{s}_j\|^2 = \int_0^T (s_i(t) - s_j(t))^2 dt.$$

2. Find an altemate set of orthonormal basis functions for the space spanned by $\cos(2\pi t/T)$ and $\sin(2\pi t/T)$.

3. Consider a set of M orthogonal signal waveforms $s_m(t)$, for $1 \leq m \leq M$ and $0 \leq t \leq T$, where each waveform has the same energy $\mathcal{E}$. Define a new set of M waveforms as

$$s'_m(t) = s_m(t) - \frac{1}{M}\sum_{i=1}^{M} s_i(t), \quad 1 \leq m \leq M,\ 0 \leq t \leq T.$$

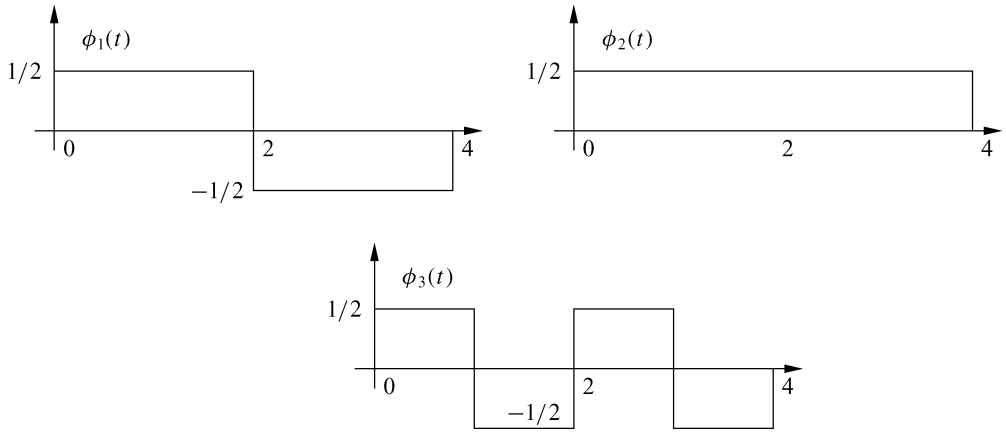

Figure 5.32: Signal waveforms for Problem 5-4.

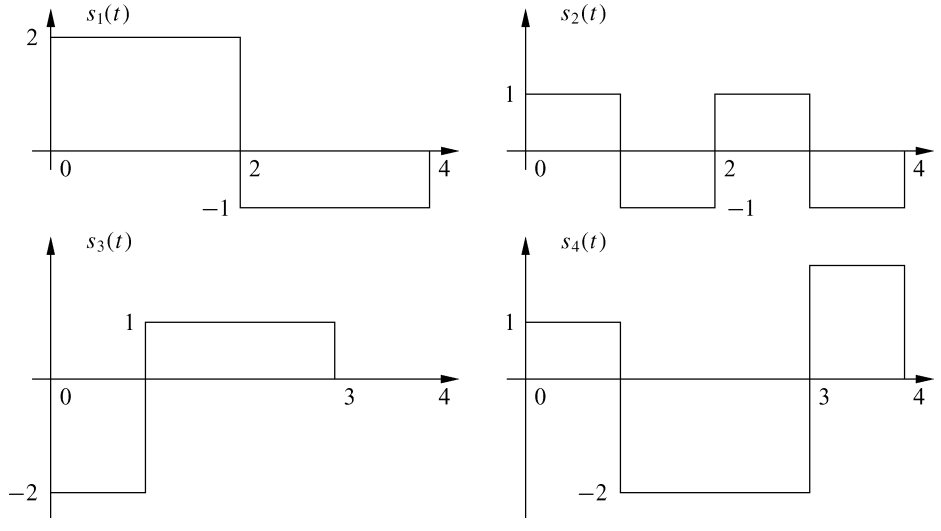

Figure 5.33: Signal waveforms for Problem 5-5.

Show that the M signal waveforms $\{s'_m(t)\}$ have equal energy, given by

$$\mathcal{E}' = (M-1)\mathcal{E}/M.$$

What is the inner product between any two waveforms?

4. Consider the three signal waveforms $\{\phi_1(t), \phi_2(t), \phi_3(t)\}$ shown in Figure 5.32.
   (a) Show that these waveforms are orthonormal.
   (b) Express the waveform $x(t)$ as a linear combination of $\{\phi_i(t)\}$ and find the coefficients, where $x(t)$ is given as

$$x(t) = \begin{cases} 2 & 0 \le t < 2, \\ 4 & 2 \le t \le 4. \end{cases}$$

5. Consider the four signal waveforms as shown in Figure 5.33.
   (a) Determine the dimensionality of the waveforms and a set of basis functions.
   (b) Use the basis functions to represent the four waveforms by vectors.
   (c) Determine the minimum distance between all the vector pairs.

6. Derive a mathematical expression for decision regions Z_i that minimize error probability assuming that messages are not equally likely – that is, assuming $p(m_i) = p_i (i = 1, \ldots, M)$, where p_i is not necessarily equal to $1/M$. Solve for these regions in the case of QPSK modulation with $s_1 = (A_c, 0)$, $s_2 = (0, A_c)$, $s_3 = (-A_c, 0)$, and $s_4 = (0, A_c)$, assuming $p(s_1) = p(s_3) = .2$ and $p(s_1) = p(s_3) = .3$.

7. Show that the remainder noise term $n_r(t)$ is independent of the correlator outputs r_i for all i. In other words, show that $\mathbf{E}[n_r(t)r_i] = 0$ for all i. Thus, since r_j (conditioned on $\mathbf{s}_i$) and $n_r(t)$ are Gaussian and uncorrelated, they are independent.

8. Show that output SNR is maximized when a given input signal is passed through a filter that is matched to that signal.

9. Find the matched filters $g(T-t), 0 \le t \le T$, and Find $\int_0^T g(t)g(T-t)\,dt$ for the following waveforms.
   (a) Rectangular pulse: $g(t) = \sqrt{2/T}$.
   (b) Sinc pulse: $g(t) = \mathrm{sinc}(t)$.
   (c) Gaussian pulse: $g(t) = (\sqrt{\pi}/\alpha)e^{-\pi^2 t^2/\alpha^2}$.

10. Show that the ML receiver of Figure 5.4 is equivalent to the matched filter receiver of Figure 5.7.

11. Use the Chernoff Bound to derive the upper bound (5.43).

12. Compute the three bounds (5.40), (5.44), (5.45) as well as the approximation (5.46) for an asymmetric signal constellation $s_1 = (A_c, 0)$, $s_2 = (0, 2A_c)$, $s_3 = (-2A_c,\ 0)$, and $s_4 = (0, -A_c)$, assuming that $A_c/\sqrt{N_0} = 4$.

13. Find the input to each branch of the decision device in Figure 5.11 if the transmit carrier phase ϕ_0 differs from the receiver carrier phase ϕ by $\triangle\phi$.

14. Consider a 4-PSK constellation with $d_{\min} = \sqrt{2}$. What is the additional energy required to send one extra bit (8-PSK) while keeping the same minimum distance (and thus with the same bit error probability)?

15. Show that the average power of a square signal constellation with l bits per dimension, P_l, is proportional to $4^l/3$ and that the average power for one more bit per dimension, keeping the same minimum distance, is $P_{l+1} \approx 4P_l$. Find P_l for $l = 2$ and compute the average energy of MPSK and MPAM constellations with the same number of bits per symbol.

16. For MPSK with differential modulation, let $\triangle\phi$ denote the phase drift of the channel over a symbol time T_s. In the absence of noise, how large must $\triangle\phi$ be in order for a detection error to occur?

17. Find the Gray encoding of bit sequences to phase transitions in differential 8-PSK. Then find the sequence of symbols transmitted using differential 8-PSK modulation with this Gray encoding for the bit sequence 101110100101110 starting at the kth symbol time, assuming the transmitted symbol at the $(k-1)$th symbol time is $\mathbf{s}(k-1) = Ae^{j\pi/4}$.

18. Consider the octal signal point constellation shown in Figure 5.34.

    (a) The nearest neighbor signal points in the 8-QAM signal constellation are separated by a distance of A. Determine the radii a and b of the inner and outer circles.

    (b) The adjacent signal points in the 8-PSK are separated by a distance of A. Determine the radius r of the circle.

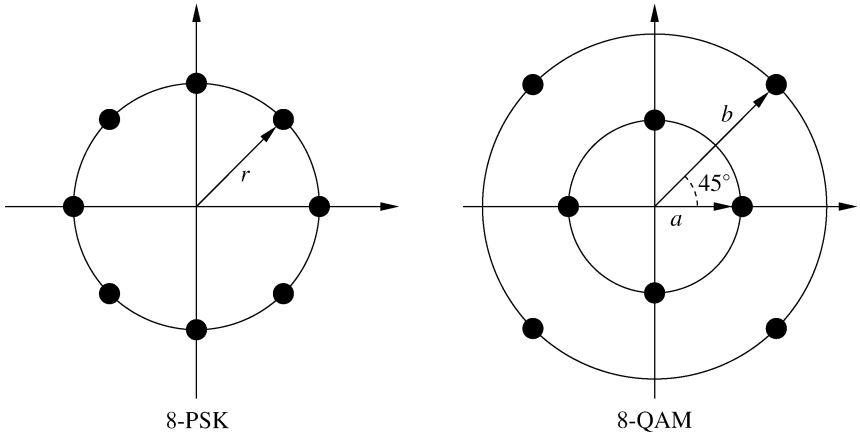

Figure 5.34: Octal signal point constellation for Problem 5-17.

    (c) Determine the average transmitter powers for the two signal constellations and compare the two powers. What is the relative power advantage of one constellation over the other? (Assume that all signal points are equally probable.)

    (d) Is it possible to assign three data bits to each point of the signal constellation such that nearest neighbor (adjacent) points differ in only one bit position?

    (e) Determine the symbol rate if the desired bit rate is 90 Mbps.

19. The $\pi/4$-QPSK modulation may be considered as two QPSK systems offset by $\pi/4$ radians.

    (a) Sketch the signal space diagram for a $\pi/4$-QPSK signal.

(b) Using Gray encoding, label the signal points with the corresponding data bits.

(c) Determine the sequence of symbols transmitted via $\pi/4$-QPSK for the bit sequence 0100100111100101.

(d) Repeat part (c) for $\pi/4$-DQPSK, assuming that the last symbol transmitted on the in-phase branch had a phase of π and that the last symbol transmitted on the quadrature branch had a phase of $-3\pi/4$.

20. Show that the minimum frequency separation for FSK such that the $\cos(2\pi f_j t)$ and $\cos(2\pi f_i t)$ are orthogonal is $\triangle f = \min_{ij} |f_j - f_i| = .5/T_s$.

21. Show that the Nyquist criterion for zero ISI pulses given by $p(kT_s) = p_0\delta(k)$ is equivalent to the frequency domain condition (5.69).

22. Show that the Gaussian pulse shape does not satisfy the Nyquist criterion.

Bibliography

[1] S. Haykin and M. Moher, *Communication Systems*, 5th ed., Wiley, New York, 2009.

[2] J. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed., Pearson, 2001.

[3] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.

[4] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics,* Vol. 23, pp. 493507, 1952.

[5] M. Fitz, "Further results in the unified analysis of digital communication systems," *IEEE Trans. Commun*., pp. 521–32, March 1992.

[6] R. Ziemer, "An overview of modulation and coding for wireless communications," *Proc. IEEE Veh. Tech. Conf*., pp. 26–30, April 1996.

[7] G. D. Forney, Jr., and L.-F. Wei, "Multidimensional constellations-Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Sel. Areas Commun*., pp. 877–92, August 1989.

[8] T. S. Rappaport, *Wireless Communications – Principles and Practice*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2001.

[9] G. L. Stuber, *Principles of Mobile Communications*, 4th ed., Springer, 2017.

[10] S. Haykin and M. Moher, *An Introduction to Analog and Digital Communications*, 2nd ed., Wiley, New York, 2006.

[11] J. G. Proakis and M. Salehi, *Digital Communications*, 5th ed., McGraw-Hill, New York, 2007.

[12] U. Mengali and A. N. D'Andrea, *Synchronization Techniques for Digital Receivers*, Plenum, New York, 1997.

[13] H. Meyr, M. Moeneclaey, and S. A. Fechtel, *Digital Communication Receivers*, vol. 2, *Synchronization, Channel Estimation, and Signal Processing*, Wiley, New York, 1997.

[14] L. E. Franks, "Carrier and bit synchronization in data communication – A tutorial review," *IEEE Trans. Commun*., pp. 1107–21, August 1980.

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 probability 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 (small at typical 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 Section 5.1.

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

In an AWGN channel the modulated signal $s(t) = \mathrm{Re}\{u(t)e^{j2\pi f_c t}\}$ 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 (PSD) $N_0/2$. The received signal is thus $r(t) = s(t) + n(t)$.

We define the received 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 and 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 PSD $N_0/2$, the total noise power within the bandwidth $2B$ is $N = N_0/2 \cdot 2B = N_0 B$. Hence the received SNR is given by

$$\text{SNR} = \frac{P_r}{N_0 B}.$$

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. This is a reasonable approximation if the interference statistics approximate those of Gaussian noise. The received SINR is given by

$$\text{SINR} = \frac{P_r}{N_0 B + P_I},$$

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

$$\text{SNR} = \frac{P_r}{N_0 B} = \frac{E_s}{N_0 B T_s} = \frac{E_b}{N_0 B T_b}, \tag{6.1}$$

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 pulse shaping with $T_s = 1/B$ (e.g., raised consine pluses 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, with M-ary 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 Section 5.1 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-ary 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 send a 0-bit and $s_2(t) = -Ag(t)\cos(2\pi f_c t)$ to send a 1-bit for $A > 0$. From (5.47) we have that the probability of error is

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

From Section 5.3.2, $d_{\min} = \|s_1 - s_0\| = \|A - (-A)\| = 2A$. Let us now relate A to the energy per bit. We have

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

by (5.57). 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

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

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 5.1.5, the union bound (5.40) on P_s for QPSK is

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

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

$$P_s \le 2Q(\sqrt{\gamma_s}) + Q(\sqrt{2\gamma_s}) \le 3Q(\sqrt{\gamma_s}). \tag{6.10}$$

The closed-form bound (5.45) becomes

$$P_s \le \frac{3}{\sqrt{2\pi\gamma_s}} \exp[-.5\gamma_s]. \tag{6.11}$$

Using the fact that the minimum distance between constellation points is $d_{\min} = \sqrt{2A^2}$ in (5.46), we obtain the nearest neighbor approximation

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

Note that with Gray encoding we can approximate P_b from P_s by $P_b \approx P_s/2$, since QPSK has two 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 $P_b \approx P_s/2$ based on the assumption of Gray coding. Finally, compute P_s based on the nearest neighbor bound using $\gamma_s = 2\gamma_b$ and then compare with the exact P_s.

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

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

The exact symbol error probability P_s is

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

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

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

which matches well with the exact P_s.

6.1.3 Error Probability for MPSK

The signal constellation for MPSK has $s_{i1} = A\cos[2\pi(i-1)/M]$ and $s_{i2} = A\sin[2\pi(i-1)/M]$ for $A > 0$ and $i = 1, \ldots, M$. The symbol energy is $E_s = A^2$, so $\gamma_s = A^2/N_0$. From (5.58) it follows that, for the received vector $\mathbf{r} = re^{j\theta}$ 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 [1, Chap. 5.2.7], which yields

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

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

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

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

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

A closed-form solution to this integral does not exist for $M > 4$ and so 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 (5.46) to P_s is given by

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

This nearest neighbor approximation can differ significantly from the exact value of P_s. However, it is much simpler to compute than the numerical integration of (6.15) that is required to obtain the exact P_s. This formula can also 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 in the first line of (6.15) yields (6.16).

Example 6.2: Compare the probability of bit error for 8-PSK and 16-PSK assuming $\gamma_b = 15$ dB and using the P_s approximation given in (6.16) along with the approximations (6.3) and (6.2).

Solution: From (6.2) we have that, for 8-PSK, $\gamma_s = (\log_2 8) \cdot 10^{15/10} = 94.87$. Substituting this into (6.16) yields

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

Now, using (6.3), we get $P_b = P_s/3 = 4.52 \cdot 10^{-8}$. For 16-PSK we have $\gamma_s = (\log_2 16) \cdot 10^{15/10} = 126.49$. Substituting this into (6.16) yields

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

and by using (6.3) we get $P_b = P_s/4 = 4.79 \cdot 10^{-4}$. Note that P_b is much larger for 16-PSK than for 8-PSK given the same γ_b. This result is expected because 16-PSK packs more bits per symbol into a given constellation and 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 to 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, Apx. C; 2, Chap. 4.3.2; 3; 4]. These works indicate that, as expected, significant phase offset leads to an irreducible bit error probability. Moreover, nonbinary signaling is more sensitive than BPSK to phase offset because of the resulting cross-coupling between 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 owing to constructive and destructive multipath interference. Even with differential modulation, phase changes over and between symbol times can produce irreducible errors [5]. Timing errors can also degrade performance; analysis of timing errors in MPSK performance can be found in [2, Chap. 4.3.3; 6].

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

$$\begin{aligned} P_s &= \frac{1}{M}\sum_{i=1}^{M} P_s(\mathbf{s}_i) \\ &= \frac{M-2}{M} 2Q\left(\sqrt{\frac{2d^2}{N_0}}\right) + \frac{2}{M} Q\left(\sqrt{\frac{2d^2}{N_0}}\right) \\ &= \frac{2(M-1)}{M} Q\left(\sqrt{\frac{2d^2}{N_0}}\right). \end{aligned} \tag{6.18}$$

From (5.55), the average energy per symbol for MPAM is

$$\overline{E}_s = \frac{1}{M}\sum_{i=1}^{M} A_i^2 = \frac{1}{M}\sum_{i=1}^{M}(2i - 1 - M)^2 d^2 = \frac{1}{3}(M^2 - 1)d^2. \tag{6.19}$$

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

$$P_s = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{6\overline{\gamma}_s}{M^2-1}}\right). \tag{6.20}$$

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 (6.20) with M replaced by $L = \sqrt{M}$ and $\overline{\gamma}_s$ equal to the average energy per symbol in the MQAM constellation:

$$P_{s,\,\text{branch}} = \frac{2(\sqrt{M}-1)}{\sqrt{M}} Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right). \tag{6.21}$$

Note that $\overline{\gamma}_s$ is multiplied by a factor of 3 in (6.21) instead of the factor of 6 in (6.20), 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

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

The nearest neighbor approximation to probability of symbol error depends on whether the constellation point is an inner or outer point. Inner points have four nearest neighbors, while outer points have either two or three nearest neighbors; in both cases the distance between nearest neighbors is $2d$. If we take a conservative approach and set the number of nearest neighbors to be four, we obtain the nearest neighbor approximation

$$P_s \approx 4Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right). \tag{6.23}$$

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

$$P_s \leq 1 - \left[1 - 2Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right)\right]^2 \leq 4Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right), \tag{6.24}$$

which is the same as (6.23) for square constellations. The nearest neighbor approximation for nonrectangular constellations is

$$P_s \approx M_{d_{\min}} Q\left(\frac{d_{\min}}{\sqrt{2N_0}}\right), \tag{6.25}$$

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 16-QAM with $\gamma_b = 15$ dB ($\gamma_s = \log_2 M \cdot \gamma_b$), compare the exact probability of symbol error (6.22) with (a) the nearest neighbor approximation (6.23) and (b) the symbol error probability for 16-PSK with the same γ_b (which was obtained in Example 6.2).

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

$$P_s = 1 - \left(1 - \frac{2(4-1)}{4} Q\left(\sqrt{\frac{3.126.49}{15}}\right)\right)^2 = 7.37 \cdot 10^{-7}.$$

The nearest neighbor approximation is given by

$$P_s \approx 4Q\left(\sqrt{\frac{3 \cdot 126.49}{15}}\right) = 9.82 \cdot 10^{-7},$$

which is slightly larger than the exact value owing to the conservative approximation that every constellation point has four nearest neighbors. The symbol error probability for 16-PSK from Example 6.2 is $P_s \approx 1.916 \cdot 10^{-3}$, which is roughly four orders of magnitude larger than the exact P_s for 16-QAM. 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 therefore 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 symbol are not skewed in amplitude or phase. The analysis of performance degradation due to phase estimation error is similar to the case of MPSK discussed previously. The channel amplitude is used to scale the decision regions so that they 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 [7, Chap. 10.3; 8].

6.1.5 Error Probability for FSK and CPFSK

Let us first consider the error probability of binary FSK with the coherent demodulator of Figure 5.24. Since demodulation is coherent, we can neglect any phase offset in the carrier signals. The transmitted signal is defined by

$$s_i(t) = A\sqrt{2T_b}\cos(2\pi f_i t), \quad i = 1, 2. \tag{6.26}$$

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

$$z = s_1 + n_1 - s_2 - n_2. \tag{6.27}$$

The device outputs a 1-bit if $z > 0$ or a 0-bit if $z \leq 0$. Let us assume that $s_1(t)$ is transmitted; then

$$z|1 = A + n_1 - n_2. \tag{6.28}$$

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

$$z|0 = n_1 - A - n_2, \tag{6.29}$$

and an error occurs if $z = n_1 - A - n_2 > 0$. 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,

$$P_b = .5p(A + n_1 - n_2 \leq 0) + .5p(n_1 - A - n_2 > 0) = Q(A/\sqrt{N_0}) = Q(\sqrt{\gamma_b}). \tag{6.30}$$

The derivation of P_s for coherent MFSK with $M > 2$ is more complex and does not lead to a closed-form solution [2, eq. (4.92)]. The probability of symbol error for noncoherent MFSK is derived in [9, Chap. 8.1] as

$$P_s = \sum_{m=1}^{M-1} (-1)^{m+1} \binom{M-1}{m} \frac{1}{m+1} \exp\left[\frac{-m\gamma_s}{m+1}\right]. \tag{6.31}$$

The error probability of CPFSK depends on whether the detector is coherent or noncoherent and also on whether it uses symbol-by-symbol detection or sequence estimation. Analysis of error probability for CPFSK is complex because the memory in the modulation requires error probability analysis over multiple symbols. The formulas for error probability can also become quite complicated. Detailed derivations of error probability for these different CPFSK structures can be found in [1; Chap. 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, Chap. 5.2; 10; 11].

Table 6.1: Approximate symbol and bit error probabilities for coherent modulations

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

6.1.6 Error Probability Approximation for Coherent Modulations

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

$$P_s(\gamma_s) \approx \alpha_M Q(\sqrt{\beta_M \gamma_s}), \tag{6.32}$$

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 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 (6.32) yields a simple formula for P_b as a function of γ_b:

$$P_b(\gamma_b) = \hat{\alpha}_M Q(\sqrt{\hat{\beta}_M \gamma_b}), \tag{6.33}$$

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 (6.32). This conversion is used in what follows to obtain P_b versus γ_b from the general form of P_s versus γ_s in (6.32).

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 5.20. Specifically, the phase comparator extracts the phase of

$$\begin{aligned}\mathbf{r}(k)\mathbf{r}^*(k-1) =& A^2 e^{j(\theta(k)-\theta(k-1))} + Ae^{j(\theta(k)+\phi_0)} n^*(k-1) \\ &+ Ae^{-j(\theta(k-1)+\phi_0)} n(k) + n(k)n^*(k-1)\end{aligned} \tag{6.34}$$

in order to determine the transmitted symbol. By symmetry we can assume a given phase difference when computing the error probability. Assuming then a phase difference of zero, $\theta(k) - \theta(k-1) = 0$, yields

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

Next we define new random variables

$$\tilde{n}(k) = n(k)e^{-j(\theta(k-1)+\phi_0)} \quad \text{and} \quad \tilde{n}(k-1) = n(k-1)e^{-j(\theta(k)+\phi_0)},$$

which have the same statistics as $n(k)$ and $n(k-1)$. Then

$$\mathbf{r}(k)\mathbf{r}^*(k-1) = A^2 + A(\tilde{n}^*(k-1) + \tilde{n}(k)) + \tilde{n}(k)\tilde{n}^*(k-1). \tag{6.36}$$

There are three terms in (6.36): 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

$$\tilde{z} = A + \mathrm{Re}\{\tilde{n}^*(k-1) + \tilde{n}(k)\} + j\ \mathrm{Im}\{\tilde{n}^*(k-1) + \tilde{n}(k)\}. \tag{6.37}$$

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

$$\theta_{\tilde{z}} = \tan^{-1} y/x. \tag{6.38}$$

Given that the phase difference was zero, an 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, by (6.37), we have two noise terms instead of one and so 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 5.20 when making a decision, so we set $x = \mathrm{Re}\{\tilde{z}\}$ in our analysis. In particular, assuming a zero is transmitted, if $x = A + \mathrm{Re}\{\tilde{n}^*(k-1) + \tilde{n}(k)\} < 0$ 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 quite general and can be applied to a wide variety of different modulation and detection types in both AWGN and fading [9, Chap. 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, Apx. B; 12, Apx. B], and the resulting bit error probability is given by

$$P_b = \frac{1}{2}e^{-\gamma b}. \tag{6.39}$$

The error probability of DQPSK is harder to obtain than that of DPSK, since the phase difference between two adjacent symbols depends on two information bits instead of one. An elegant analysis of the resulting error probability is obtained in [1, Apx. B] as a special case of the probability analysis for quadratic functions of complex Gaussian random variables. This yields, under Gray coding, the bit error probability [1, Eqn. 4.5-66]

$$P_b = Q_1(a,b) - \frac{1}{2}\exp\left[\frac{-(a^2+b^2)}{2}\right] I_0(ab), \tag{6.40}$$

where $Q_1(a,b)$ is the *Marcum Q function* defined as

$$Q_1(a,b) \triangleq \int_b^\infty z\exp\left[\frac{-(a^2+z^2)}{2}\right] I_0(az)dz, \tag{6.41}$$

for $I_0(x)$ the modified Bessel function of the first kind and zeroth order, $a = \sqrt{2\gamma_b(1-\sqrt{.5})}$, and $b = \sqrt{2\gamma_b(1+\sqrt{.5})}$.

6.2 Alternate Q-Function Representation

In (6.32) 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 0 and variance 1 exceeds the value z:

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

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 [13]. The alternate form is given by

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

This representation can also be deduced from the work of Weinstein [14] or Pawula et al. [5]. 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

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

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

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

(see [13]), where $\beta_M = \sin^2(\pi/M)$. Note that this formula does not correspond to the general form $\alpha_M Q(\sqrt{\beta_M \gamma_s})$: the general form is an approximation, whereas (6.45) is exact. Note also that (6.45) is obtained via a finite-range integral of simple trigonometric functions that is easily computed using 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 as a result of shadowing and/or multipath fading. Thus, in fading, γ_s is a random variable with distribution $p_{\gamma_s}(\gamma)$ and so $P_s(\gamma_s)$ 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 γ_0 corresponding to the maximum allowable P_s;
- the average error probability, $\bar{P}_s$, 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 fading coherence time is on the order of a symbol time ($T_s \approx T_c$), so that the signal fade level is roughly constant over a symbol period. Since many error correction coding techniques can recover from a few bit errors and since end-to-end performance is typically not seriously degraded by a few simultaneous bit errors (since the erroneous bits can be dropped or retransmitted), the average error probability is a reasonably good figure of merit for the channel quality under these conditions.

However, if the signal fading is changing slowly ($T_s \ll T_c$) then a deep fade will affect many simultaneous symbols. Hence 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 if $T_c \ll T_s$ then 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

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

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 because it generally can't be detected by the human ear. Thus, for a BPSK signal in Rayleigh fading, $\gamma_b < 7\,\text{dB}$ would be declared an outage; hence we set $\gamma_0 = 7\,\text{dB}$.

In Rayleigh fading the outage probability becomes

$$P_{\text{out}} = \int_0^{\gamma 0} \frac{1}{\bar{\gamma}_s} e^{-\gamma_s/\bar{\gamma}_s} d\gamma_s = 1 - e^{-\gamma_0/\bar{\gamma}_s}. \tag{6.47}$$

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

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

In decibels this means that $10 \log \bar{\gamma}_s$ must exceed the target $10 \log \gamma_0$ by

$$F_d = -10 \log[-\ln(1 - P_{\text{out}})]$$

in order to maintain acceptable performance more than $100 \cdot (1 - P_{\text{out}})$ percent of the time. The quantity F_d is typically called the *dB fade margin*.

Example 6.4: Determine the required $\bar{\gamma}_b$ for BPSK modulation in slow Rayleigh fading such that, for 95% of the time (or in 95% of the locations), $P_b(\gamma_b) < 10^{-4}$.

Solution: For BPSK modulation in AWGN the target BER is obtained at $\gamma_b = 8.5\,\text{dB}$. That is, for $P_b(\gamma_b) = Q(\sqrt{2\gamma_b})$ we have $P_b(10^{.85}) = 10^{-4}$. Thus, $\gamma_0 = 8.5\,\text{dB}$. We want $P_{\text{out}} = p(\gamma_b < \gamma_0) = .05$, so

$$\bar{\gamma}_b = \frac{\gamma_0}{-\ln(1 - P_{\text{out}})} = \frac{10^{.85}}{-\ln(1 - .05)} = 21.4\,\text{dB}.$$

6.3.2 Average Probability of Error

The average probability of error is used as a performance metric when $T_s \approx T_c$. We can therefore assume that γ_s is roughly constant over a symbol time. Then the average probability of error is computed by integrating the error probability in AWGN over the fading distribution:

$$\bar{P}_s = \int_0^\infty P_s(\gamma) p_{\gamma_s}(\gamma)\, d\gamma, \tag{6.49}$$

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 (e.g., Rayleigh, Rician, log-normal), we compute $p_{\gamma_s}(\gamma)$ by making the change of variable

$$p_{\gamma_s}(\gamma) d\gamma = p(r)\, dr. \tag{6.50}$$

For example, in Rayleigh fading the received signal amplitude r has the Rayleigh distribution

$$p(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \quad r \geq 0, \tag{6.51}$$

and the signal power is exponentially distributed with mean $2\sigma^2$. The SNR per symbol for a given amplitude r is

$$\gamma = \frac{r^2 T_s}{2\sigma_n^2}, \tag{6.52}$$

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

$$d\gamma = \frac{rT_s}{\sigma_n^2}\, dr. \tag{6.53}$$

Substituting (6.52) and (6.53) into (6.51) and then (6.50) yields

$$p_{\gamma_s}(\gamma) = \frac{\sigma_n^2}{\sigma^2 T_s} e^{-\gamma\sigma_n^2/\sigma^2 T_s}. \tag{6.54}$$

Since the average SNR per symbol $\bar{\gamma}_s$ is just $\sigma^2 T_s/\sigma_n^2$, we can rewrite (6.54) as

$$p_{\gamma}(\gamma) = \frac{1}{\bar{\gamma}_s} e^{-\gamma/\bar{\gamma}_s}, \tag{6.55}$$

which is the exponential distribution. For binary signaling this reduces to

$$p_{\gamma b}(\gamma) = \frac{1}{\bar{\gamma}_b} e^{-\gamma/\bar{\gamma} b}. \tag{6.56}$$

Integrating (6.6) over the distribution (6.56) yields the following average probability of error for BPSK in Rayleigh fading:

$$\bar{P}_b = \frac{1}{2}\left[1 - \sqrt{\frac{\bar{\gamma}_b}{1+\bar{\gamma}_b}}\right] \approx \frac{1}{4\bar{\gamma}_b} \quad \text{(BPSK)}, \tag{6.57}$$

where the approximation holds for large $\bar{\gamma}_b$. A similar integration of (6.30) over (6.56) yields the average probability of error for binary FSK in Rayleigh fading as

$$\bar{P}_b = \frac{1}{2}\left[1 - \sqrt{\frac{\bar{\gamma}_b}{2+\bar{\gamma}_b}}\right] \approx \frac{1}{2\bar{\gamma}_b} \quad \text{(binary FSK)}. \tag{6.58}$$

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

$$\bar{P}_b = \frac{1}{2(1+\bar{\gamma}_b)} \approx \frac{1}{2\bar{\gamma}_b} \text{ (DPSK)}, \tag{6.59}$$

where again the approximation holds for large $\bar{\gamma}_b$. Thus, at high SNRs, the performance of DPSK and binary FSK converge, and this performance is approximately 3 dB worse than that of BPSK. This was also observed in AWGN and represents 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

$$\bar{P}_s \approx \int_0^\infty \alpha_M Q(\sqrt{\beta_M \gamma}) \cdot \frac{1}{\bar{\gamma}_s} e^{-\gamma/\bar{\gamma}_s}\, d\gamma_s = \frac{\alpha_M}{2}\left[1 - \sqrt{\frac{.5\beta_M \bar{\gamma}_s}{1+.5\beta_M \bar{\gamma}_s}}\right] \approx \frac{\alpha_M}{2\beta_M \bar{\gamma}_s}, \tag{6.60}$$

where the last approximation is in the limit of high SNR.

It is interesting to compare the bit error probabilities of the different modulation schemes in AWGN and in 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 $\bar{\gamma}_b$. 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, whereas it takes 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 (6.23) and (6.60), is shown in Figure 6.2. From these figures it is clear that minimizing transmit 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.

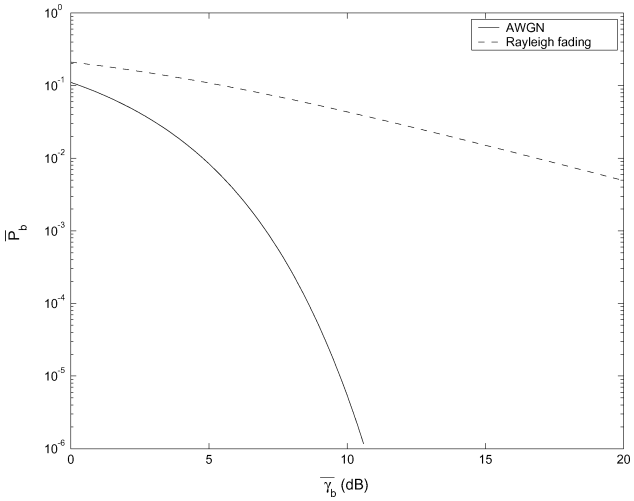

Figure 6.1: Average P_b for BPSK in Rayleigh fading and AWGN.

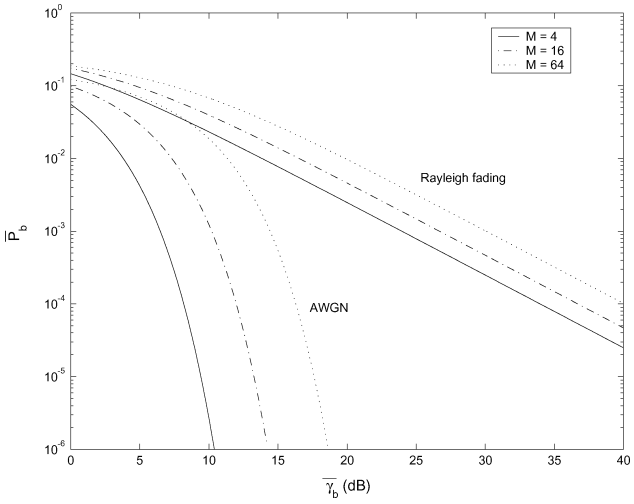

Figure 6.2: Average P_b for MQAM in Rayleigh fading and AWGN.

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.

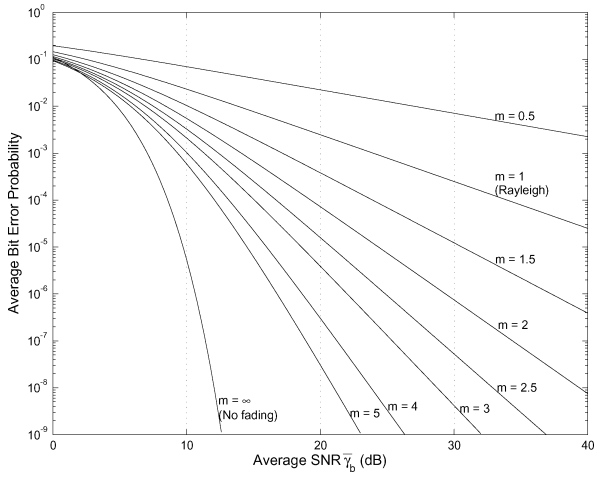

Figure 6.3: Average P_b for BPSK in Nakagami fading.

6.3.3 Moment Generating Function Approach to Average Error Probability

The *moment generating function* (MGF) is a useful tool for 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 MGF for a nonnegative random variable γ with distribution $p_\gamma(\gamma), \gamma \geq 0$, is defined as

$$\mathcal{M}_\gamma(s) = \int_0^\infty p_\gamma(\gamma) e^{s\gamma}\, d\gamma. \tag{6.61}$$

Note that this function is just the Laplace transform of the distribution $p_\gamma(\gamma)$ 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 [9, Chap. 5.1].

- Rayleigh:

$$\mathcal{M}_{\gamma_s}(s) = (1 - s\bar{\gamma}_s)^{-1}. \tag{6.62}$$

- Rician with factor K:

$$\mathcal{M}_{\gamma_s}(s) = \frac{1+K}{1+K-s\bar{\gamma}_s} \exp\left[\frac{Ks\bar{\gamma}_s}{1+K-s\bar{\gamma}_s}\right]. \tag{6.63}$$

- Nakagami-m:

$$\mathcal{M}_{\gamma_s}(s) = \left(1 - \frac{s\bar{\gamma}_s}{m}\right)^{-m}. \tag{6.64}$$

As indicated by its name, the moments $\mathbf{E}[\gamma^n]$ of γ can be obtained from $\mathcal{M}_\gamma(s)$ as

$$\mathbf{E}[\gamma^n] = \frac{\partial^n}{\partial s^n}[\mathcal{M}_{\gamma_s}(s)]|_{s=0}. \tag{6.65}$$

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,

$$P_s = a\exp[-b\gamma_s] \tag{6.66}$$

for constants a and b, or as a finite-range integral of such an exponential function:

$$P_s = \int_0^B a\exp[-b(x)\gamma_s]\,dx, \tag{6.67}$$

where the constant $b(x)$ may depend on the integrand but the SNR γ_s does not (and is not in the limits of integration). These forms allow the average probability of error to be expressed in terms of the MGF for the fading distribution. Specifically, if $P_s = a\exp[-b\gamma_s]$, then

$$\bar{P}_s = \int_0^\infty a\exp[-b\gamma]p_{\gamma_s}(\gamma)d\gamma = a\mathcal{M}_{\gamma_s}(-b). \tag{6.68}$$

Since DPSK is in this form with $a = 1/2$ and $b = 1$, we see that the average probability of bit error for DPSK in any type of fading is

$$\bar{P}_b = \tfrac{1}{2}\mathcal{M}_{\gamma_s}(-1), \tag{6.69}$$

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 (6.62) with $s = -1$ yields $\bar{P}_b = [2(1+\bar{\gamma}_b)]^{-1}$, which is the same as we obtained in (6.59). If P_s is in the integral form of (6.67) then

$$\begin{aligned}\bar{P}_s &= \int_0^\infty \int_0^B a\exp[-b(x)\gamma]dx\, p_{\gamma_s}(\gamma)\,d\gamma \\ &= a\int_0^B \left[\int_0^\infty \exp[-b(x)\gamma]p_{\gamma s}(\gamma)d\gamma\right]dx \\ &= a\int_0^B \mathcal{M}_{\gamma_s}(-b(x))\,dx.\end{aligned} \tag{6.70}$$

In this latter case, the average probability of symbol error is a single finite-range integral of the MGF of the fading distribution, which typically can be found in closed form or easily evaluated numerically.

Let us now apply the MGF approach to specific modulations and fading distributions. Consider any modulation for which its symbol error probability P_s in AWGN is given or approximated by an expression of the following form:

$$P_s(\gamma_s) = \alpha Q(\sqrt{\beta\gamma_s}). \tag{6.71}$$

Note that (6.32) provided an approximation to P_s in AWGN for coherent MPSK, MQAM, and MPAM in this form, with constants α and β that depend on M.

Using the alternate Q-function representation (6.43), we get that

$$P_s = \frac{\alpha}{\pi}\int_0^{\pi/2} \exp\left[\frac{-.5\beta\gamma}{\sin^2\phi}\right]d\phi, \tag{6.72}$$

which is in the desired form (6.67). Thus, the average error probability in fading for modulations with $P_s = \alpha Q(\sqrt{\beta\gamma_s})$ in AWGN is given by

$$\begin{aligned}\bar{P}_s &= \frac{\alpha}{\pi}\int_0^{\infty}\int_0^{\pi/2}\exp\left[\frac{-.5\beta\gamma}{\sin^2\phi}\right]d\phi\, p_{\gamma_s}(\gamma)\, d\gamma d\phi \\ &= \frac{\alpha}{\pi}\int_0^{\pi/2}\left[\int_0^{\infty}\exp\left[\frac{-.5\beta\gamma}{\sin^2\phi}\right]p_{\gamma_s}(\gamma)d\gamma\right]d\phi \\ &= \frac{\alpha}{\pi}\int_0^{\pi/2}\mathcal{M}_{\gamma_s}\left(\frac{-.5\beta}{\sin^2\phi}\right)d\phi, \end{aligned} \tag{6.73}$$

where $\mathcal{M}_{\gamma_s}(s)$ is the MGF associated with the distribution $p_{\gamma_s}(\gamma)$ as defined by (6.61). Recall that Table 6.1 approximates the error probability in AWGN for many modulations of interest as $P_s \approx \alpha Q(\sqrt{\beta\gamma_s})$, so (6.73) 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 (6.73) by noting that Craig's formula for P_s of MPSK in AWGN given by (6.45) is in the desired form (6.67). Thus, the exact average probability of error for MPSK becomes

$$\begin{aligned}\bar{P}_s &= \int_0^{\infty}\frac{1}{\pi}\int_0^{(M-1)\pi/M}\exp\left[\frac{-\beta_M\gamma_s}{\sin^2\phi}\right]d\phi\, p_{\gamma_s}(\gamma)\, d\gamma \\ &= \frac{1}{\pi}\int_0^{(M-1)\pi/M}\left[\int_0^{\infty}\exp\left[\frac{-\beta_M\gamma_s}{\sin^2\phi}\right]p_{\gamma_s}(\gamma)\, d\gamma\right]d\phi \\ &= \frac{1}{\pi}\int_0^{(M-1)\pi/M}\mathcal{M}_{\gamma s}\left(-\frac{\beta_M}{\sin^2\phi}\right)d\phi, \end{aligned} \tag{6.74}$$

where $\beta_M = \sin^2(\pi/M)$ depends on the size of the MPSK constellation. The MGF $\mathcal{M}_{\gamma_s}(s)$ for Rayleigh, Rician, and Nakagami-m distributions were given by (6.62), (6.63), and (6.64), respectively. Substituting $s = -\beta_M/\sin^2\phi$ in these expressions yields the following equations.

- Rayleigh:

$$\mathcal{M}_{\gamma_s}\left(-\frac{\beta_M}{\sin^2\phi}\right) = \left(1+\frac{\beta_M\bar{\gamma}_s}{\sin^2\phi}\right)^{-1}. \tag{6.75}$$

- Rician with factor K:

$$\mathcal{M}_{\gamma_s}\left(-\frac{\beta_M}{\sin^2\phi}\right) = \frac{(1+K)\sin^2\phi}{(1+K)\sin^2\phi+\beta_M\bar{\gamma}_s}\exp\left[-\frac{K\beta_M\bar{\gamma}_s}{(1+K)\sin^2\phi+\beta_M\bar{\gamma}_s}\right]. \tag{6.76}$$

- Nakagami-m:

$$\mathcal{M}_{\gamma_s}\left(-\frac{\beta_M}{\sin^2\phi}\right) = \left(1+\frac{\beta_M\bar{\gamma}_s}{m\sin^2\phi}\right)^{-m}. \tag{6.77}$$

All of these functions are simple trigonometrics and are therefore easy to integrate over the finite range in (6.73) or (6.74).

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 BPSK for an AWGN channel has $P_b = Q(\sqrt{2\gamma_b})$, so $\alpha = 1$ and $\beta = 2$ in (6.71). The moment generating function for Nakagami-*m* fading is given by (6.77), and substituting this into (6.73) with $\alpha = g = 1$ yields

$$\bar{P}_b = \frac{1}{\pi}\int_0^{\pi/2}\left(1 + \frac{\bar{\gamma}_b}{m\sin^2\phi}\right)^{-m} d\phi.$$

From (6.22) 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 $Q^2(z)$ allows us to apply the same techniques used here for MPSK to MQAM modulation. Specifically, an alternate representation of $Q^2(z)$ is derived in [15] as

$$Q^2(z) = \frac{1}{\pi}\int_0^{\pi/4}\exp\left[\frac{-z^2}{2\sin^2\phi}\right] d\phi. \tag{6.78}$$

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 (6.22) in terms of the alternate representations for $Q(z)$ and $Q^2(z)$ as

$$\begin{aligned} P_s(\gamma_s) = {} & \frac{4}{\pi}\left(1 - \frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\exp\left(-\frac{\beta_M\gamma_s}{\sin^2\phi}\right) d\phi \\ & - \frac{4}{\pi}\left(1 - \frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\exp\left[-\frac{\beta_M\gamma_s}{\sin^2\phi}\right] d\phi, \end{aligned} \tag{6.79}$$

where $\beta_M = 1.5/(M-1)$ is a function of the MQAM constellation size. Then the average probability of symbol error in fading becomes

$$\begin{aligned} \bar{P}_s &= \int_0^\infty P_s(\gamma)p_{\gamma_s}(\gamma)d\gamma \\ &= \frac{4}{\pi}\left(1 - \frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\int_0^\infty \exp\left[-\frac{\beta_M\gamma}{\sin^2\phi}\right] p_{\gamma_s}(\gamma)\, d\gamma\, d\phi \\ &\quad - \frac{4}{\pi}\left(1 - \frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\int_0^\infty \exp\left[-\frac{\beta_M\gamma}{\sin^2\phi}\right] p_{\gamma_s}(\gamma)\, d\gamma\, d\phi \\ &= \frac{4}{\pi}\left(1 - \frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\mathcal{M}_{\gamma_s}\left(-\frac{\beta_M}{\sin^2\phi}\right) d\phi \\ &\quad - \frac{4}{\pi}\left(1 - \frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\mathcal{M}_{\gamma_s}\left(-\frac{\beta_M}{\sin^2\phi}\right) d\phi. \end{aligned} \tag{6.80}$$

Thus, the exact average probability of symbol error is obtained via two finite-range integrals of the MGF over the fading distribution, which typically can 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 MFSK, with P_s in AWGN given by (6.31), which is a finite sum of the desired form (6.66). Thus, in fading, the average symbol error probability of noncoherent MFSK is given by

$$\begin{aligned}\bar{P}_s &= \int_0^\infty \sum_{m=1}^{M} (-1)^{m+1} \binom{M-1}{m} \frac{1}{m+1} \exp\left[\frac{-m\gamma}{m+1}\right] p_{\gamma_s}(\gamma)\, d\gamma \\ &= \sum_{m=1}^{M} (-1)^{m+1} \binom{M-1}{m} \frac{1}{m+1} \left[\int_0^\infty \exp\left[\frac{-m\gamma}{m+1}\right] p_{\gamma_s}(\gamma)\, d\gamma\right] \\ &= \sum_{m=1}^{M} (-1)^{m+1} \binom{M-1}{m} \frac{1}{m+1} \mathcal{M}_{\gamma_s}\left(-\frac{m}{m+1}\right). \end{aligned} \tag{6.81}$$

Finally, for differential MPSK it can be shown [16] that the average probability of symbol error is given by

$$P_s = \frac{\sqrt{\beta_M}}{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\exp[-\gamma_s(1-\sqrt{1-\beta_M}\cos\theta)]}{1-\sqrt{1-\beta_M}\cos\theta}\, d\theta \tag{6.82}$$

for $\beta_M = \sin^2(\pi/M)$, which is in the desired form (6.67). Thus we can express the average probability of symbol error in terms of the MGF of the fading distribution as

$$\bar{P}_s = \frac{\sqrt{\beta_M}}{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\mathcal{M}_{\gamma s}(-(1-\sqrt{1-\beta_M}\cos\theta))}{1-\sqrt{1-\beta_M}\cos\theta}\, d\theta. \tag{6.83}$$

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 [9, Chap. 8.2].

6.3.4 Combined Outage and Average Error Probability

When the fading environment is a superposition of both fast and slow fading (e.g., 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 non-outage is obtained by averaging over the fast fading. We use the following notation.

- $\bar{\bar{\gamma}}_s$ denotes the average SNR per symbol for a fixed path loss with averaging over fast fading and shadowing.
- $\bar{\gamma}_s$ denotes the (random) SNR per symbol for a fixed path loss and random shadowing but averaged over fast fading. Its average value, averaged over the shadowing, is $\bar{\bar{\gamma}}_s$.
- γ_s denotes the random SNR due to fixed path loss, shadowing, and multipath. Its average value, averaged over multipath only, is $\bar{\gamma}_s$. Its average value, averaged over both multipath and shadowing, is $\bar{\bar{\gamma}}_s$.

With this notation we can specify an average error probability $\bar{P}_s$ with some probability $1 - P_{\text{out}}$. An outage is declared when the received SNR per symbol due to shadowing and path loss alone, $\bar{\gamma}_s$, falls below a given target value $\bar{\gamma}_{s_0}$. When not in outage ($\bar{\gamma}_s \geq \bar{\gamma}_{s_0}$), the average probability of error is obtained by averaging over the distribution of the fast fading conditioned on the mean SNR:

$$\bar{P}_s = \int_0^\infty P_s(\gamma_s) p(\gamma_s \mid \bar{\gamma}_s)\, d\gamma_s. \tag{6.84}$$

The criterion used to determine the outage target $\bar{\gamma}_{s_0}$ is typically based on a given maximum acceptable average probability of error $\bar{P}_{s_0}$. The target $\bar{\gamma}_{s_0}$ must then satisfy

$$\bar{P}_{s_0} = \int_0^\infty P_s(\gamma_s) p(\gamma_s \mid \bar{\gamma}_{s_0})\, d\gamma_s. \tag{6.85}$$

It is clear that, whenever $\bar{\gamma}_s > \bar{\gamma}_{s_0}$, the average error probability $\bar{P}_s$ will be below the target maximum value $\bar{P}_{s_0}$.

Example 6.6: Consider BPSK modulation in a channel with both log-normal shadowing ($\sigma_{\psi_{\rm dB}} = 8\,{\rm dB}$) and Rayleigh fading. The desired maximum average error probability is $\bar{P}_{b_0} = 10^{-4}$, which requires $\bar{\gamma}_{b_0} = 34\,{\rm dB}$. Determine the value of $\bar{\bar{\gamma}}_b$ that will ensure $\bar{P}_b \le 10^{-4}$ with probability $1 - P_{\rm out} = .95$.

Solution: We must find $\bar{\bar{\gamma}}_b$, the average of γ_b in both the fast and slow fading, such that $p(\bar{\gamma}_b > \bar{\gamma}_{b_0}) = 1 - P_{\rm out}$. For log-normal shadowing we compute this as

$$p(\bar{\gamma}_b > 34) = p\left(\frac{\bar{\gamma}_b - \bar{\bar{\gamma}}_b}{\sigma_{\psi_{\rm dB}}} \ge \frac{34 - \bar{\bar{\gamma}}_b}{\sigma_{\psi_{\rm dB}}}\right) = Q\left(\frac{34 - \bar{\bar{\gamma}}_b}{\sigma_{\psi_{\rm dB}}}\right) = 1 - P_{\rm out}, \tag{6.86}$$

since, assuming units in dB, $(\bar{\gamma}_b - \bar{\bar{\gamma}}_b)/\sigma_{\psi_{\rm dB}}$ is a Gauss-distributed random variable with mean 0 and standard deviation 1. Thus, the value of $\bar{\bar{\gamma}}_b$ is obtained by substituting the values of $P_{\rm out}$ and $\sigma_{\psi_{\rm dB}}$ in (6.86) and using a table of Q-functions or an inversion program, which yields $(34 - \bar{\bar{\gamma}}_b)/8 = -1.6$ or $\bar{\bar{\gamma}}_b = 46.8\,{\rm dB}$.

6.4 Doppler Spread

One consequence of Doppler spread is 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 consequent 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 they 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 [1, Apx. B; 12, Apx. B]. A different approach utilizing alternate forms of the Marcum Q-function can also be used [9, Chap. 8.2.5]. The resulting average bit error probability for DPSK is

$$\bar{P}_b = \frac{1}{2}\left(\frac{1 + K + \bar{\gamma}_b(1 - \rho_C)}{1 + K + \bar{\gamma}_b}\right) \exp\left[-\frac{K\bar{\gamma}_b}{1 + K + \bar{\gamma}_b}\right], \tag{6.87}$$

where ρ_C is the channel correlation coefficient after a bit time T_b, K is the fading parameter of the Rician distribution, and $\bar{\gamma}_b$ is the average SNR per bit. For Rayleigh fading ($K = 0$) this simplifies to

$$\bar{P}_b = \frac{1}{2}\left(\frac{1 + \bar{\gamma}_b(1 - \rho_C)}{1 + \bar{\gamma}_b}\right). \tag{6.88}$$

Letting $\bar{\gamma}_b \to \infty$ in (6.87) yields the irreducible error floor:

$$\bar{P}_{\rm floor} = \frac{(1 - \rho_C)e^{-K}}{2}\ ({\rm DPSK}). \tag{6.89}$$

A similar approach is used in [18] to bound the bit error probability of DQPSK in fast Rician fading as

$$\bar{P}_b \leq \frac{1}{2}\left(1 - \sqrt{\frac{(\rho_C\bar{\gamma}_s/\sqrt{2})^2}{(\bar{\gamma}_s+1)^2 - (\rho_C\bar{\gamma}_s/\sqrt{2})^2}}\right)\exp\left[-\frac{(2-\sqrt{2})K\bar{\gamma}_s/2}{(\bar{\gamma}_s+1)-(\rho_C\bar{\gamma}_s/\sqrt{2})}\right], \tag{6.90}$$

where K is as before, ρ_C is the channel correlation coefficient after a symbol time T_s, and $\bar{\gamma}_s$ is the average SNR per symbol. Letting $\bar{\gamma}_s \to \infty$ yields the irreducible error floor:

$$\bar{P}_{\text{floor}} = \frac{1}{2}\left(1 - \sqrt{\frac{(\rho_C/\sqrt{2})^2}{1-(\rho_C/\sqrt{2})^2}}\right)\exp\left[-\frac{(2-\sqrt{2})(K/2)}{1-\rho_C/\sqrt{2}}\right] \text{(DQPSK)}. \tag{6.91}$$

As discussed in Section 3.2.1, the channel correlation $A_C(\tau)$ over time τ 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 at $T = T_b$ for DPSK. Table 6.2, from [19], 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 (6.89) yields an irreducible error for DPSK of

Table 6.2: Correlation coefficients for different Doppler power spectra models

Type	**Doppler power spectrum $S_C(f)$**	$\rho_C = A_C(T)/A_C(0)$
Rectangular	$P_0/2B_D, \lvert f\rvert < B_D$	$\text{sinc}(2B_DT)$
Gaussian	$(P_0/\sqrt{\pi}B_D)e^{-f^2/B_D^2}$	$e^{-(\pi B_D T)^2}$
Uniform scattering	$P_{0/\pi}\sqrt{B_D^2 - f^2}, \lvert f\rvert < B_D$	$J_0(2\pi B_D T)$
1st-order Butterworth	$P_0 B_D/\pi(f^2 + B_D^2)$	$e^{-2\pi B_D T}$

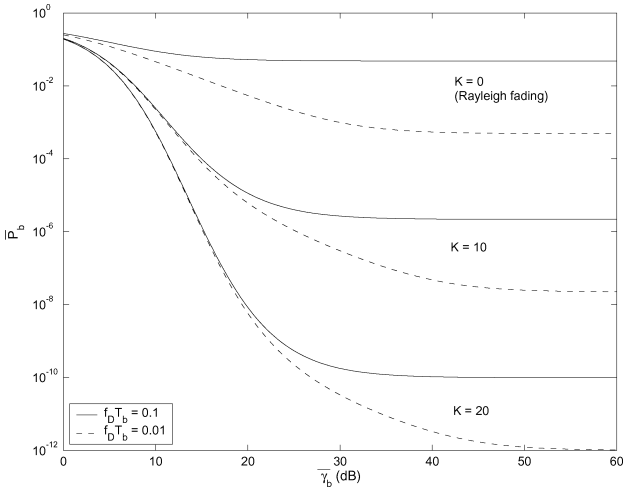

Figure 6.4: Average P_b for DPSK in fast Rician fading with uniform scattering.

$$\overline{P}_{\text{floor}} = \frac{1 - J_0(2\pi f_D T_b)}{2} \approx .5(\pi f_D T_b)^2, \tag{6.92}$$

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 decorrelate between transmitted symbols. This phenomenon is one of the few instances in digital communications where performance improves as data rate increases.

A plot of (6.87), 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 $f_D = 80$ Hz. For what approximate range of data rates will the irreducible error floor of DPSK be below 10^{-4}?
Solution: We have $\overline{P}_{\text{floor}} \approx .5(\pi f_D T_b)^2 < 10^{-4}$. 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, simulations 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-136 TDMA standard) in [20, 21]. 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 typically decorrelates more over a given symbol time.

6.5 Intersymbol Interference

Frequency-selective fading gives rise to intersymbol interference, 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 because it depends on the modulation format and the ISI characteristics, which in turn depend on the characteristics of the channel and the sequence of transmitted symbols.

The first extensive analysis of the degradation in symbol error probability due to ISI was done by Bello and Nelin [22]. 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 complex and must be approximated for evaluation. A more accurate analytical analysis can be found in [23, Chap. 8.2], where the irreducible error floor is evaluated based on both the worst-case sequence of transmitted symbols and by averaging over all possible symbol sequences. These expressions are also complex to evaluate owing to their dependence on the channel and symbol sequence characteristics. A simple approximation to symbol error probability with ISI can be obtained by treating the ISI as uncorrelated white Gaussian noise. Then the SNR becomes

$$\hat{\gamma}_s = \frac{P_r}{N_0 B + I}, \tag{6.93}$$

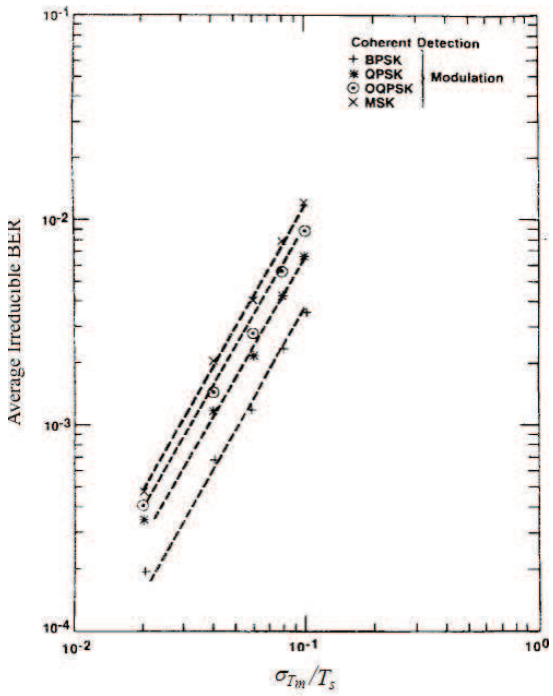

Figure 6.5: lrreducible error versus normalized rms delay spread for Gaussian power delay profile. (Reprinted by permission from [24, Fig. 9], © 1987 IEEE.)

where P_r is the received power associated with the line-of-sight signal component, and I is the received 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 LOS signal component and the ISI experience flat fading, then $\hat{\gamma}_s$ will be a random variable with distribution $p(\hat{\gamma}_s)$. The outage probability is then $P_{\text{out}} = p(\hat{\gamma}_s < \gamma_0)$ and the average symbol error probability is $\overline{P}_s = \int P_s(\hat{\gamma}_s)p(\hat{\gamma}_s)\, d\gamma_s$. Note that $\hat{\gamma}_s$ is the ratio of two random variables – the LOS received power P_r and the ISI received power I – and thus the resulting distribution $p(\hat{\gamma}_s)$ may be hard to obtain and typically 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 [24, 20, 21, 25, 26, 27]. The earliest of these works is [24]. The extensive simulations in this work determined the irreducible error floor due to ISI for BPSK, DPSK, QPSK, OQPSK and MSK modulations with 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 [24] 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: for the raised cosine pulses discussed in Section 5.5, increasing β from 0 to 1 can reduce the error floor by over an order of magnitude.

The simulation results of [24] are 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 the figure that, for all modulations, we can approximately bound the irreducible error floor as

$$P_{\text{floor}} \leq K \left(\frac{\sigma_{T_m}}{T_s} \right)^2 \tag{6.94}$$

for K a constant that depends on the modulation type, the transmit and receive filtering, and the power delay profile. For the modulation types in Figure 6.5 we see that K ranges from about 0.4 for BPSK to 1 for MSK. Other simulation results [27, 25, 26] support this bound as well. This bound can also be derived analytically based on group delay. Details can be found in [29, Chapter 12.3.2].

This bound imposes severe constraints on data rate even when symbol error probabilities on the order of 10^{-2} are acceptable, and usually this is a very high error probability even with strong error correction codes. For example, the rms delay spread in a typical urban environment is approximately $\sigma_{T_m} = 2.5$ μs. 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, $\sigma_{T_m} \approx 25$ μs, which reduces the maximum data rate to 4 kbaud.

Example 6.8: Using the approximation $\overline{P}_{\text{floor}} \leq K(\sigma_{T_m}/T_s)^2$ of (6.94) with $K = 0.4$ for BPSK and $K = 0.7$ for QPSK, find the maximum data rate that can be transmitted through a channel with delay spread $\sigma_{T_m} = 3$ μs, 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 $\overline{P}_{\text{floor}} = .4(\sigma_{T_m}/T_b)^2$ and so require $T_b \geq \sigma_{T_m}/\sqrt{\overline{P}_{\text{floor}}/.4} = 94.87$ μs, which leads to a data rate of $R = 1/T_b = 15$ kbps. For QPSK, we have $T_s \geq \sigma_{T_m}/\sqrt{\overline{P}_{\text{floor}}/.7} = 113.39$ μs. Since there are two bits per symbol, this leads to a data rate of $R = 2/T_s = 17.64$ kbps. This indicates that, for a given data rate, QPSK is more robust to ISI than BPSK because its symbol time is slower, even though the K value in the error floor bound (6.94) is higher for the larger constellation size of QPSK versus BPSK. In general K increases in the bound of (6.94) as the constellation size increases, but large constellations are still more robust to ISI than smaller ones due to their longer symbol times.

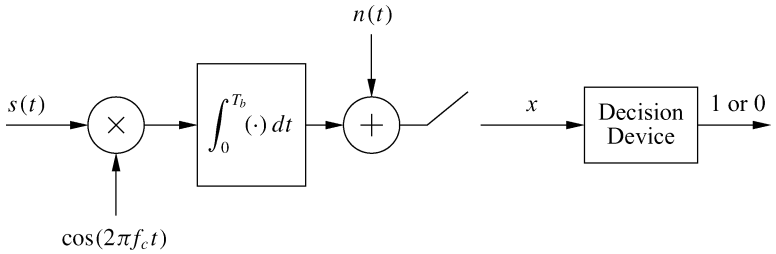

Figure 6.6: BPSK demodulator for Problem 6-4.

Chapter 6 Problems

1. Consider a system in which data is transferred at a rate of 100 bits per second over the channel.

   (a) Find the symbol duration if we use a sinc pulse for signaling and the channel bandwidth is 10 kHz.

   (b) Suppose 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 a priori 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 $s_2(t) = -A\cos(2\pi f_c t)$ for $A > 0$, assuming the receiver structure is as shown in Figure 5.17.

   (b) Suppose you can change the threshold value in the receiver of Figure 5.17. Find the threshold value that yields equal error probability regardless of which bit is transmitted – that is, 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 $A > 0$ and $B > 0$ so that the receiver of Figure 5.17 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 and $N_0 = .1$. For which system is P_b minimized?

3. Consider a BPSK receiver whose 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 5.17 is $\cos(2\pi f_c t + \phi)$. Determine the threshold level in the threshold device of Figure 5.17 that minimizes probability of bit error, and find this minimum error probability.

4. Assume a BPSK demodulator in which the receiver noise is added after the integrator, as shown in Figure 6.6. The decision device outputs a 1 if its input x has $\mathrm{Re}\{x\} \geq 0$, 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 $A_c = \sqrt{2/T_b} = 1$ 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 signal constellations shown in Figure 6.7.

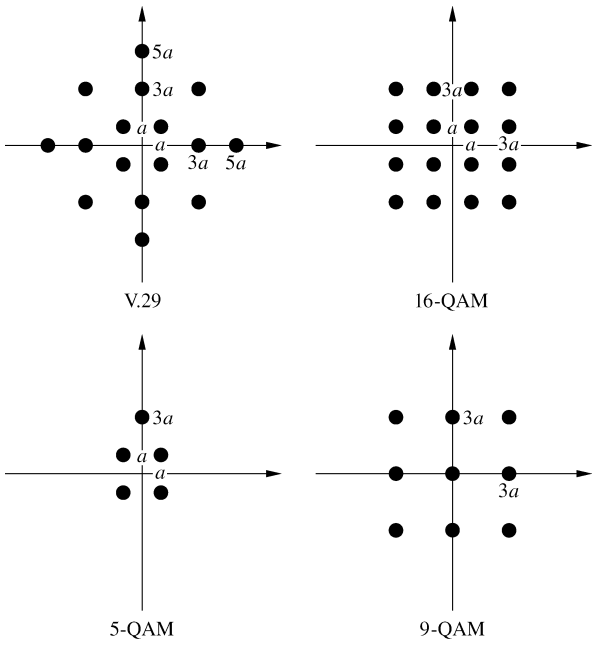

Figure 6.7: Signal constellations for Problem 6-5.

6. Plot the exact symbol error probability and the approximation from Table 6.1 of 16-QAM with $0 \leq \gamma_s \leq 30$ dB. Does the error in the approximation increase or decrease with γ_s? 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? 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 of this, corresponding 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

$$I_x(a) \triangleq \int_0^\infty \frac{e^{-at^2}}{x^2 + t^2} dt, \tag{6.97}$$

and 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

$$I_x(a) \triangleq \int_0^\infty \frac{e^{-at^2}}{x^2+t^2}\,dt = \frac{\pi}{2x}e^{ax^2}\mathrm{erfc}(x\sqrt{a}). \tag{6.99}$$

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 left-hand side of (6.99), derive the alternate representation of the erfc function:

$$\mathrm{erfc}(x) = \frac{2}{\pi}\int_0^{\pi/2} e^{-x^2/\sin^2\theta}d\theta.$$

(d) Convert this alternate representation of the erfc function to the alternate representation of the Q-function.

9. Consider a communication system that uses BPSK signaling, with average signal power of 100 W and noise power at the receiver of 4 W. Based on its error probability, can this system be used for transmission of data? Can it be used for voice? Now consider the presence of fading with an average SNR $\overline{\gamma}_b = 20$ dB. How do your answers to the previous questions 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 l mph, 10 mph, and 100 mph.

11. Derive the expression for the moment generating function for SNR in Rayleigh fading.

12. This problem illustrates why satellite systems that must compensate for shadow fading are going bankrupt. Consider an LEO satellite system orbiting 500 km above the earth. Assume that 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 power spectral density of 10^{-16} mW/Hz (total noise power is N_0B). 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, find the transmit power needed for all users in the cell to meet the 10^{-6} BER target.

(b) Repeat part (a) assuming that the channel also experiences log-normal shadowing with $\sigma_{\psi_{\mathrm{dB}}} = 8$ 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 from BPSK to the higher-level signal modulation of 16-PSK.

(a) Find the minimum distance between constellation points in 16-PSK modulation as a function of signal energy E_s.

(b) Find α_M and β_M such that the symbol error probability of 16-PSK in AWGN is approximately

$$P_s \approx \alpha_M Q(\sqrt{\beta_M \gamma_s}).$$

(c) Using your expression in part (b), find an approximation for the average symbol error probability of 16-PSK in Rayleigh fading in terms of $\overline{\gamma}_s$.

(d) Convert the expressions for average symbol error probability of 16-PSK in Rayleigh fading to an expression 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 16-PSK. 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 dB to 20 dB 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 – that is, 1 in 100 customers can be unhappy at any given time. In nonoutage, the voice BER requirement is $\overline{P}_b = 10^{-3}$. Assume a noise PSD $N_0/2$ with $N_0 = 10^{-16}$ mW/Hz, a signal bandwidth of 30 kHz, a carrier frequency of 900 MHz, free-space path-loss propagation with nondirectional antennas, and a shadowing standard deviation of $\sigma_{\psi_{\rm dB}} = 6$ 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 of radius 100 meters. Assume that propagation follows the simplified path-loss model with $K = 1, d_0 = 1$ m, and $\gamma = 3$. Assume the signal experiences (in addition to path loss) log-normal shadowing with $\sigma_{\psi_{\rm dB}} = 4$ 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 $N_0/2$ has $N_0 = 10^{-14}$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 of less than 10^{-4}.

    (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 $\overline{P}_b < 10^{-4}$ under BPSK modulation.

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 0-bit or a 1-bit. Let us assume that over time kT_b a 0-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 $\mathbf{r}(k-1) = g_{k-1}\mathbf{s}(k-1) + n(k-1)$ and $\mathbf{r}(k) = g_k\mathbf{s}(k-1) + n(k)$, 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 that is input to the phase comparator of Figure 5.20 in order to extract the phase difference is $\mathbf{r}(k)\mathbf{r}^*(k-1) = g_k g_{k-1}^* + g_k \mathbf{s}(k-1)^* n_{k-1} + g_{k-1}^* s_{k-1}^* n_k + n_k n_{k-1}^*$.

Assuming a reasonable SNR, the last term $n_k n_{k-1}^*$ of this expression can be neglected. So neglecting this term and then 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 0-bit was transmitted over time kT_b, an error is made if $x = \mathrm{Re}\{\tilde{z}\} < 0$, so we must determine the distribution of x. The characteristic function for x is the two-sided Laplace transform of the distribution of x:

$$\Phi_X(s) = \int_{-\infty}^{\infty} p_X(s) e^{-sx}\, dx = \mathbf{E}[e^{-sx}].$$

This function will have a left-plane pole p_1 and a right-plane pole p_2, so it 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 distribution $p_X(x)$ for $x \geq 0$, and the rightplane pole corresponds to the distribution $p_X(x)$ 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 $x = \mathrm{Re}\{\tilde{z}\} < 0$, so here we need only consider the distribution $p_X(x)$ for $x < 0$ corresponding to the second term of $\Phi_X(s)$.

(c) Show that the inverse Laplace transform of the second term of $\Phi_X(s)$ from part (b) is

$$p_X(x) = \frac{p_1 p_2}{p_2 - p_1} e^{p_2 X},\ x < 0.$$

(d) Use part (c) to show that $P_b = -p_1/(p_2 - p_1)$.

In $x = \mathrm{Re}\{\tilde{z}\} = \mathrm{Re}\{g_k g_{k-1}^* + g_k \tilde{n}_{k-1}^* + g_{k-1}^* \tilde{n}_k\}$, the channel gains g_k, g_{k-1} and noises $\tilde{n}_k, \tilde{n}_{k-1}$ are complex Gaussian random variables. Thus, the poles p_1, p_2 in $p_X(x)$ are derived using the general quadratic form of complex Gaussian random variables [1, Apx. B; 12, Apx. B] as

$$p_1 = \frac{-1}{N_0(\overline{\gamma}_b[1+\rho_c]+1)} \text{ and } p_2 = \frac{1}{N_0(\overline{\gamma}_b[1-\rho_c]+1)}$$

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_b expression from part (d).

(f) Show that this reduces to the average probability of error $\overline{P}_b = 1/2(1+\overline{\gamma}_b)$ 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 dB 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 due to Doppler for DQPSK modulation with a data rate of 40 kbps, 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 ns and a Doppler spread of 80 Hz. Given the error floors due to Doppler and ISI – and assuming DQPSK modulation in Rayleigh fading and uniform scattering – approximately what range of data rates can be transmitted over this channel with a BER of 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 $\sigma_{T_m} = 3$ μs (using BPSK, QPSK, or MSK modulation) such that the probability of bit error P_b is less than 10^{-3}.

Bibliography

[1] J. G. Proakis and M. Salehi, *Digital Communications*, 5th ed., McGraw-Hill, New York, 2008.

[2] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[3] S. Rhodes, "Effect of noisy phase reference on coherent detection of offset-QPSK signals," *IEEE Trans. Commun.*, pp. 1046–55, August 1974.

[4] N. R. Sollenberger and J. C.-I. Chuang, "Low-overhead symbol timing and carrier recovery for portable TDMA radio systems," *IEEE Trans. Commun.*, pp. 1886–92, October 1990.

[5] R. Pawula, S. Rice, and J. Roberts, "Distribution of the phase angle between two vectors perturbed by Gaussian noise," *IEEE Trans. Commun.*, pp. 1828–41, August 1982.

[6] W. Cowley and L. Sabel, "The performance of two symbol timing recovery algorithms for PSK demodulators," *IEEE Trans. Commun.*, pp. 2345–55, June 1994.

[7] W. T. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation*, IEEE/Pentech Press, London, 1994.

[8] X. Tang, M.-S. Alouini, and A. Goldsmith, "Effect of channel estimation error on M-QAM BER performance in Rayleigh fading," *IEEE Trans. Commun.*, pp. 1856–64, December 1999.

[9] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*, 2nd Ed., Wiley, New York, 2004.

[10] 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.*, pp. 922–33, February–April 1995.

[11] E. Grayver and B. Daneshrad, "A low-power all-digital FSK receiver for deep space applications," *IEEE Trans. Commun.*, pp. 911–21, May 2001.

[12] M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, New York, 1966 [reprinted 1995 by Wiley/IEEE Press].

[13] J. Craig, "New, simple and exact result for calculating the probability of error for two-dimensional signal constellations," *Proc. Military Commun. Conf.*, pp. 25.5.1–25.5.5, November 1991.

[14] F. S. Weinstein, "Simplified relationships for the probability distribution of the phase of a sine wave in narrowband normal noise," *IEEE Trans. Inform. Theory*, pp. 658–61, September 1974.

[15] M. K. Simon and D. Divsalar, "Some new twists to problems involving the Gaussian probability integral," *IEEE Trans. Commun.*, pp. 200–10, Febmary 1998.

[16] R. F. Pawula, "A new formula for MDPSK symbol error probability," *IEEE Commun. Lett*., pp. 271–2, October 1998.

[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*., pp. 160–8, June 1962.

[18] P. Y. Kam, "Tight bounds on the bit-error probabilities of 2DPSK and 4DPSK in nonselective Rician fading," *IEEE Trans. Commun*., pp. 860–2, July 1998.

[19] P. Y. Kam, "Bit error probabilities of MDPSK over the nonselective Rayleigh fading channel with diversity reception," *IEEE Trans. Commun*., pp. 220–4, February 1991.

[20] 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. Sel. Areas Commun*., pp. 393–405, April 1993.

[21] S. Chennakeshu and G. J. Saulnier, "Differential detection of $\pi/4$-shifted-DQPSK for digital cellular radio," *IEEE Trans. Veh. Tech*., pp. 46–57, Febmary 1993.

[22] 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*., pp. 170–86, June 1963.

[23] M. B. Pursley, *Introduction to Digital Communications*, Prentice-Hall, Englewood Cliffs, NJ, 2005.

[24] J. C.-I. Chuang, "The effects of time delay spread on portable radio communications channels with digital modulation," *IEEE J. Sel. Areas Commun*., pp. 879–89, June 1987.

[25] N. R. Sollenberger, J. C. I. Chuang, L. F. Chang, S. Ariyavisitakul and H. Arnold, "Architecture and implementation of an efficient and robust TDMA frame structure for digital portable communications," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 1, pp. 250-260, Feb. 1991.

[26] J. C. I. Chuang, "The effects of delay spread on 2-PSK, 4-PSK, 8-PSK and 16-QAM in a portable radio environment," *IEEE Transactions on Vehicular Technology*, vol. 38, no. 2, pp. 43-45, May 1989.

[27] C. Liu and K. Feher, "Bit error rate performance of $\pi/4$ DQPSK in a frequency selective fast Rayleigh fading channel," *IEEE Trans. Veh. Tech*., pp. 558–68, August 1991.

[28] S. Gumnathan and K. Feher, "Multipath simulation models for mobile radio channels," *Proc. IEEE Veh. Tech. Conf*., pp. 131–4, May 1992.

[29] *Wireless Communications*, 2nd Ed, A. F. Molisch, John Wiley and Sons, 2010.

Chapter 7

Diversity

In Chapter 6 we saw that both Rayleigh fading and log-normal shadowing exact a large power penalty on the performance of modulation over wireless channels. One of the best techniques to mitigate the effects of fading is diversity combining of independently fading signal paths. Diversity combining exploits the fact that independent signal paths have a low probability of experiencing deep fades simultaneously. Thus, the idea behind diversity is to send the same data over independent fading paths. These independent paths are combined in such a way that the fading of the resultant signal is reduced. For example, consider a system with two antennas at either the transmitter or receiver that experience independent fading. If the antennas are spaced sufficiently far apart, it is unlikely that they both experience deep fades at the same time. By selecting the antenna with the strongest signal, a technique known as *selection combining*, we obtain a much better signal than if we had just one antenna. This chapter focuses on common methods used at the transmitter and receiver to achieve diversity. A more detailed treatment of diversity combining methods and their performance, including the effects of correlated fading on the different paths as well as channel estimation errors, can be found in [1, Chap. 9].

Diversity techniques that mitigate the effect of multipath fading are called *microdiversity*, and that is the focus of this chapter. Diversity to mitigate the effects of shadowing from buildings and objects is called *macrodiversity*. Macrodiversity is generally implemented by combining signals received by several base stations or access points, which requires coordination among these different stations or points. Such coordination is implemented as part of the networking protocols in infrastructure-based wireless networks. We will therefore defer discussion of macrodiversity until Chapter 15, where we examine the design of such networks.

7.1 Realization of Independent Fading Paths

There are many ways of achieving independent fading paths in a wireless system. One method is to use multiple transmit or receive antennas, also called an antenna array, where the elements of the array are separated in distance. This type of diversity is referred to as *space diversity*. Note that with receiver space diversity, independent fading paths are realized without an increase in transmit signal power or bandwidth. Moreover, coherent combining of the diversity signals increases the signal-to-noise power ratio at the receiver over the SNR that would be obtained with just a single receive antenna. This SNR increase, called array gain, can also be obtained with transmitter space diversity by appropriately weighting the antenna transmit powers relative to the channel gains. In addition to array gain, space diversity also provides diversity gain, defined as the change in slope of the error probability resulting from the diversity combining. We will describe both the array gain and diversity gain for specific diversity-combining techniques in subsequent sections.

The maximum diversity gain for either transmitter or receiver space diversity typically requires that the separation between antennas be such that the fading amplitudes corresponding to each antenna are approximately

independent. For example, from equation (3.28) we know that, in a uniform scattering environment with omni-directional transmit and receive antennas, the minimum antenna separation required for independent fading on each antenna is approximately one half-wavelength ($.38\lambda$, to be exact). If the transmit or receive antennas are directional, the scattering is not uniform, or a dominant LOS component exists, then a larger antenna separation is typically required to obtain independent fading samples [2, Chap. 13].

A second method of achieving diversity is by using either two transmit antennas or two receive antennas with different polarization (e.g., vertically and horizontally polarized waves). The two transmitted waves follow the same path. However, since the multiple random reflections distribute the power nearly equally relative to both polarizations, the average receive power corresponding to either polarized antenna is approximately the same. Since the scattering angle relative to each polarization is random, it is highly improbable that signals received on the two differently polarized antennas would be simultaneously in a deep fade. There are two disadvantages of polarization diversity. First, you can have at most two diversity branches, corresponding to the two types of polarization. The second disadvantage is that polarization diversity loses effectively half the power (3 dB) because the transmit or receive power is divided between the two differently polarized antennas.

Directional antennas provide angle (or directional) diversity by restricting the receive antenna beamwidth to a given angle. In the extreme, if the angle is very small then at most one of the multipath rays will fall within the receive beamwidth, so there is no multipath fading from multiple rays. However, this diversity technique requires either a sufficient number of directional antennas to span all possible directions of arrival or a single antenna whose directivity can be steered to the arrival angle of one of the multipath components (preferably the strongest one). Note also that with this technique the SNR may decrease owing to the loss of multipath components that fall outside the receive antenna beamwidth – unless the directional gain of the antenna is sufficiently large to compensate for this lost power. *Smart antennas* are antenna arrays with adjustable phase at each antenna element: such arrays form directional antennas that can be steered to the incoming angle of the strongest multipath component [3].

Frequency diversity is achieved by transmitting the same narrowband signal at different carrier frequencies, where the carriers are separated by the coherence bandwidth of the channel. This technique requires additional transmit power to send the signal over multiple frequency bands. Spread-spectrum techniques, discussed in Chapter 13, are sometimes described as providing frequency diversity in that the channel gain varies across the bandwidth of the transmitted signal. However, this is not equivalent to sending the same information signal over independently fading paths. As discussed in Section 13.2.4, spread spectrum with RAKE reception does provide independently fading paths of the information signal and thus is a form of path diversity. Time diversity is achieved by transmitting the same signal at different times, where the time difference is greater than the channel coherence time (the inverse of the channel Doppler spread). Time diversity does not require increased transmit power but it does lower the data rate, since data is repeated in the diversity time slots rather than sending new data in those time slots. Time diversity can also be achieved through coding and interleaving, as will be discussed in Chapter 8. Clearly time diversity cannot be used for stationary applications, since the channel coherence time is infinite and thus fading is highly correlated over time.

In this chapter we focus on space diversity as a reference when describing diversity systems and the different combining techniques, although the techniques can be applied to any type of diversity. Thus, the combining techniques will be defined as operations on an antenna array. Receiver and transmitter diversity are treated separately, since their respective system models and combining techniques have important differences.

7.2 Receiver Diversity

7.2.1 System Model

In receiver diversity, the independent fading paths associated with multiple receive antennas are combined to obtain a signal that is then passed through a standard demodulator. The combining can be done in several ways, which vary in complexity and overall performance. Most combining techniques are linear: the output of the combiner is just a weighted sum of the different fading paths or *branches*, as shown in Figure 7.1 for M-branch diversity. In this model the complex fading associated with the ith branch is $r_i e^{j\theta_i}$. The signal on the ith branch is multiplied by complex gain α_i and then summed with the other weighted branches to form the combiner output. When all but one of the complex α_i are zero, only one path is passed to the combiner output. If more than one of the α_i are nonzero then the combiner adds together multiple paths, though each path may be weighted by a different value. Combining more than one branch signal requires *co-phasing*, where the phase θ_i of the ith branch is removed through multiplication by $\alpha_i = a_i e^{-j\theta_i}$ for some real-valued a_i. Without co-phasing, the branch signals would not add up coherently in the combiner, so the resulting output could still exhibit significant fading due to constructive and destructive addition of the signals in all the branches.

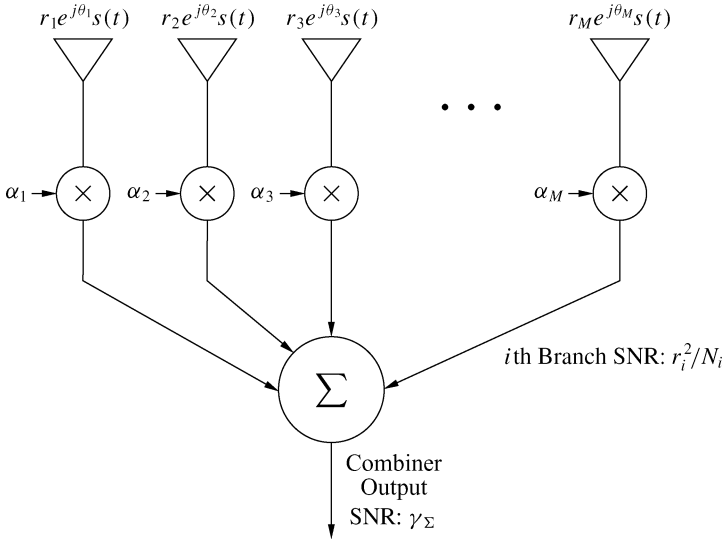

Figure 7.1: Linear combiner.

Since co-phasing requires knowledge of the branch phase prior to multiplication by α_i, coherent detection to determine this phase value is required in each branch, as shown in Figure 7.2. This coherent detection can also determine the value of the branch signal envelope r_i or SNR $\gamma_i = r_i^2/N_i$ for N_i the noise power on the ith branch,[1] which α_i depends on in some combining techniques. A dedicated coherent detector for each branch of a diversity system significantly increases the hardware complexity and power consumption relative to a single-branch system, particularly for a large number of branches. When the combiner utilizes just one branch at any given time, i.e. $\alpha_i = 0$ for all except one of the M branches, then the coherent detector can be placed after the combiner in Figure 7.1, as co-phasing multiple signals is no longer needed. We will assume throughout this chapter that the noise power in each branch $N_i = N = N_0 B$ is the same. When that is not the case, the results of this chapter can be applied by scaling the amplitude term r_i by $\sqrt{N/N_i}$ for average noise power $N = N_0 B$ across all branches.

The main purpose of diversity is to coherently combine the independent fading paths so that the effects of

[1] In practice $r_i^2 + N_i$ is easier to measure than SNR, since the former involves finding only the total power in the received signal.

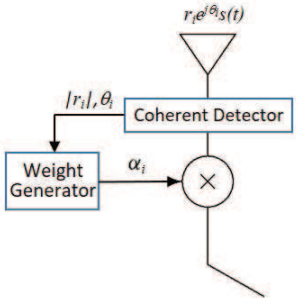

Figure 7.2: Branch Coherent Detection

fading are mitigated. The signal output from the combiner equals the original transmitted signal $s(t)$ multiplied by a random real-valued amplitude term $\alpha_\Sigma = \sum_i a_i r_i$. This amplitude term results in a random SNR γ_Σ at the combiner output, where the distribution of γ_Σ is a function of the number of diversity paths, the fading distribution on each path, and the combining technique, as we shall describe in more detail.

The array gain in receiver space diversity results from coherent combining of multiple receive signals. Even in the absence of fading, this can lead to an increase in average received SNR. For example, suppose there is no fading and that each branch has the same SNR $\gamma_i = r_i^2/(N_0 B) = \gamma$. Since both the noise and the signal are multiplied by a_i, the SNR of the combined output is

$$\gamma_\Sigma = \frac{\left(\sum_{i=1}^M a_i r_i\right)^2}{\sum_{i=1}^M a_i^2 N_0 B}. \tag{7.1}$$

Now set $a_i = r_i$ in (7.1). We will see later that setting $a_i = cr_i$ for any constant c is optimal for maximal-ratio combining in fading. Then,

$$\gamma_\Sigma = \frac{\left(\sum_{i=1}^M r_i^2\right)^2}{N_0 B \sum_{i=1}^M r_i^2} = \frac{\sum_{i=1}^M r_i^2}{N_0 B} = M\gamma. \tag{7.2}$$

Thus, in the absence of fading, with appropriate weighting there is an M-fold increase in SNR due to the coherent combining of the M signals received from the different antennas. This SNR increase in the absence of fading is referred to as the *array gain*. More precisely, array gain A_g is defined as the increase in the average combined SNR $\bar{\gamma}_\Sigma$ over the average branch SNR $\bar{\gamma}$:

$$A_g = \frac{\bar{\gamma}_\Sigma}{\bar{\gamma}}.$$

Array gain occurs for all diversity combining techniques but is most pronounced in maximal-ratio combining. The array gain allows a system with multiple transmit or receive antennas in a fading channel to achieve better performance than a system without diversity in an AWGN channel with the same average SNR. We will see this effect in performance curves for both maximal-ratio and equal-gain combining with a large number of antennas.

With fading, the combining of multiple independent fading paths leads to a more favorable distribution for γ_Σ than would be the case with just a single path. In particular, the performance of a diversity system (whether it uses space diversity or another form of diversity) in terms of $\bar{P}_s$ and P_{out} is as defined in Sections 6.3.1 and 6.3.2:

$$\bar{P}_s = \int_0^\infty P_s(\gamma) p_{\gamma_\Sigma}(\gamma) d\gamma, \tag{7.3}$$

where $P_s(\gamma)$ is the probability of symbol error for demodulation of $s(t)$ in AWGN with SNR γ; and

$$P_{\text{out}} = p(\gamma_\Sigma \leq \gamma_0) = \int_0^{\gamma_0} p_{\gamma_\Sigma}(\gamma) d\gamma \tag{7.4}$$

for some target SNR value γ_0. The more favorable distribution for γ_Σ resulting from diversity combining leads to a decrease in $\bar{P}_s$ and P_{out}, and the resulting performance advantage is called the *diversity gain*.

When $\bar{P}_s$ as a function of $p_{\gamma_\Sigma}(\gamma)$ has a particular form, the performance benefit of diversity combining is defined in terms of the *diversity order* of the system. In particular, if the average probability of error with diversity combining can be expressed in the form $\bar{P}_s = c\bar{\gamma}^{-M}$, where c is a constant that depends on the specific modulation and coding and $\bar{\gamma}$ is the average received SNR per branch, then M is defined as the *diversity order* of the system. The diversity order indicates how the *slope* of the average probability of error as a function of average SNR changes with diversity (Figures 7.4 and 7.7 below show these slope changes as a function of M for different combining techniques). Recall from (6.60) that a general approximation for average error probability in Rayleigh fading with no diversity is $\bar{P}_s \approx \alpha_M/(2\beta_M\bar{\gamma})$. This expression has a diversity order of 1, which is consistent with a single fading path, e.g. from a single receive antenna. The maximum diversity order of a system with M fading paths, e.g. from M antennas, is M. When the diversity order equals the number of independent fading paths that are combined via diversity, the system is said to achieve *full diversity order*.

In the following sections we will describe the different combining techniques and their performance in more detail. These techniques entail various trade-offs between performance and complexity.

7.2.2 Selection Combining

In selection combining (SC), the combiner outputs the signal on the branch with the highest SNR. Because only one branch is used at a time, SC requires just one receiver that is switched into the active antenna branch. However, a dedicated receiver on each antenna branch may be needed for systems that transmit continuously in order to monitor SNR on each branch simultaneously. With SC, the path output from the combiner has an SNR equal to the maximum SNR of all the branches. Moreover, since only one branch output is used, co-phasing of multiple branches is not required; hence this technique can be used with either coherent or differential modulation.

For M-branch diversity, the cumulative distribution function (cdf) of γ_Σ is given by

$$P_{\gamma_\Sigma}(\gamma) = p(\gamma_\Sigma < \gamma) = p(\max[\gamma_1, \gamma_2, \ldots, \gamma_M] < \gamma) = \prod_{i=1}^{M} p(\gamma_i < \gamma). \tag{7.5}$$

We obtain the distribution of γ_Σ by differentiating $P_{\gamma_\Sigma}(\gamma)$ relative to γ, and we obtain the outage probability by evaluating $P_{\gamma_\Sigma}(\gamma)$ at $\gamma = \gamma_0$. Assume that we have M branches with uncorrelated Rayleigh fading amplitudes r_i and corresponding instantaneous SNR $\gamma_i = r_i^2/N_0B$. Defining the average SNR on the ith branch as $\bar{\gamma}_i = \mathbf{E}[\gamma_i]$, the SNR distribution will be exponential:

$$p(\gamma_i) = \frac{1}{\bar{\gamma}_i} e^{-\gamma_i/\bar{\gamma}_i}. \tag{7.6}$$

From (6.47), the outage probability for a target γ_0 on the ith branch in Rayleigh fading is

$$P_{\text{out}}(\gamma_0) = 1 - e^{-\gamma_0/\bar{\gamma}_i}. \tag{7.7}$$

The outage probability of the selection combiner for the target γ_0 is then

$$P_{\text{out}}(\gamma_0) = \prod_{i=1}^{M} p(\gamma_i < \gamma_0) = \prod_{i=1}^{M} [1 - e^{-\gamma_0/\bar{\gamma}_i}]. \tag{7.8}$$

If the average SNR for all of the branches are the same ($\bar{\gamma}_i = \bar{\gamma}$ for all i), then this reduces to

$$P_{\text{out}}(\gamma_0) = p(\gamma_\Sigma < \gamma_0) = [1 - e^{-\gamma_0/\bar{\gamma}}]^M. \tag{7.9}$$

Differentiating (7.9) relative to γ_0 yields the distribution for γ_Σ:

$$p_{\gamma_\Sigma}(\gamma) = \frac{M}{\bar{\gamma}}[1 - e^{-\gamma/\bar{\gamma}}]^{M-1} e^{-\gamma/\bar{\gamma}}. \tag{7.10}$$

From (7.10) we see that the average SNR of the combiner output in independent and identically distributed Rayleigh fading is

$$\begin{aligned}
\bar{\gamma}_\Sigma &= \int_0^\infty \gamma p_{\gamma_\Sigma}(\gamma) d\gamma \\
&= \int_0^\infty \frac{\gamma M}{\bar{\gamma}}[1 - e^{-\gamma/\bar{\gamma}}]^{M-1} e^{-\gamma/\bar{\gamma}} d\gamma \\
&= \bar{\gamma} \sum_{i=1}^{M} \frac{1}{i},
\end{aligned} \tag{7.11}$$

where (7.11) follows from the binomial expansion of $[1 - e^{-\gamma/\bar{\gamma}}]^{M-1}$ and induction to obtain the final result. Thus, the average SNR gain and corresponding array gain increase with M, but not linearly. The biggest gain is obtained by going from no diversity to two-branch diversity. Increasing the number of diversity branches from two to three will give much less gain than going from one to two, and in general increasing M yields diminishing returns in terms of the array gain. This trend is also illustrated in Figure 7.3, which shows P_{out} versus $\bar{\gamma}/\gamma_0$ for different M in i.i.d. Rayleigh fading. We see that there is dramatic improvement even with two-branch selection combining: going from $M = 1$ to $M = 2$ at 1% outage probability yields an approximate 12-dB reduction in required SNR, and at .01% outage probability there is an approximate 20-dB reduction in required SNR. However, at .01% outage, going from two-branch to three-branch diversity results in an additional reduction of about 7 dB, and going from three-branch to four-branch results in an additional reduction of about 4 dB. Clearly the power savings is most substantial when going from no diversity to two-branch diversity, with diminishing returns as the number of branches is increased. It should be noted also that, even with Rayleigh fading on all branches, the distribution of the combiner output SNR is no longer exponential.

Example 7.1: Find the outage probability of BPSK modulation at $P_b = 10^{-3}$ for a Rayleigh fading channel with SC diversity for $M = 1$ (no diversity), $M = 2$, and $M = 3$. Assume equal branch SNRs of $\bar{\gamma} = 15\,\text{dB}$.
Solution: A BPSK modulated signal with $\gamma_b = 7\,\text{dB}$ has $P_b = 10^{-3}$. Thus, we have $\gamma_0 = 7\,\text{dB}$. Substituting $\gamma_0 = 10^{.7}$ and $\bar{\gamma} = 10^{1.5}$ into (7.9) yields $P_{\text{out}} = .1466$ for $M = 1$, $P_{\text{out}} = .0215$ for $M = 2$, and $P_{\text{out}} = .0031$ for $M = 3$. We see that each additional branch reduces outage probability by almost an order of magnitude.

The average probability of symbol error is obtained from (7.3) with $P_s(\gamma)$ the probability of symbol error in AWGN for the signal modulation and $p_{\gamma_\Sigma}(\gamma)$ the distribution of the combiner SNR. For most fading distributions and coherent modulations, this result cannot be obtained in closed form and must be evaluated numerically or by approximation. In Figure 7.4 we plot $\bar{P}_b$ versus $\bar{\gamma}_b$ in i.i.d. Rayleigh fading, obtained by a numerical evaluation of $\int Q(\sqrt{2\gamma}) p_{\gamma_\Sigma}(\gamma) d\gamma$ for $p_{\gamma_\Sigma}(\gamma)$ given by (7.10). The figure shows that the diversity system for $M \geq 8$ has a lower error probability than an AWGN channel with the same SNR, owing to the array gain of the combiner. The same will be true for the performance of maximal-ratio and equal-gain combining. Closed-form results do exist

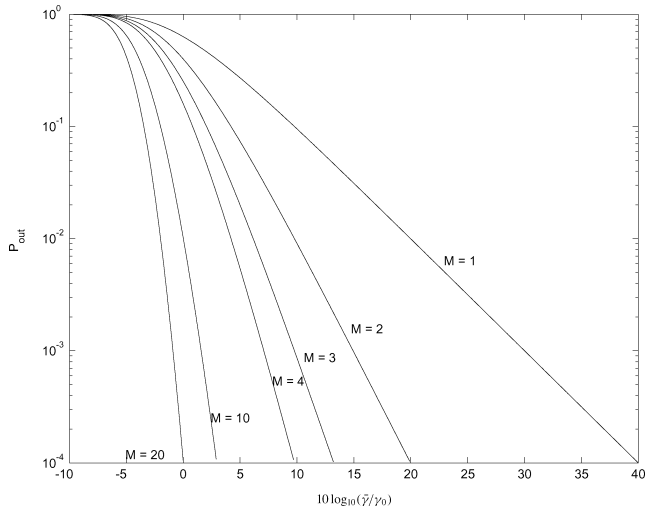

Figure 7.3: Outage probability of selection combining in Rayleigh fading.

for differential modulation under i.i.d. Rayleigh fading on each branch [1, Chap. 9.7; 4, Chap. 6.1]. For example, it can be shown for DPSK with $p_{\gamma_\Sigma}(\gamma)$ given by (7.10) that the average probability of symbol error is given by

$$\bar{P}_b = \int_0^\infty \frac{1}{2} e^{-\gamma} p_{\gamma_\Sigma}(\gamma) d\gamma = \frac{M}{2} \sum_{m=0}^{M-1} (-1)^m \frac{\binom{M-1}{m}}{1+m+\bar{\gamma}}. \tag{7.12}$$

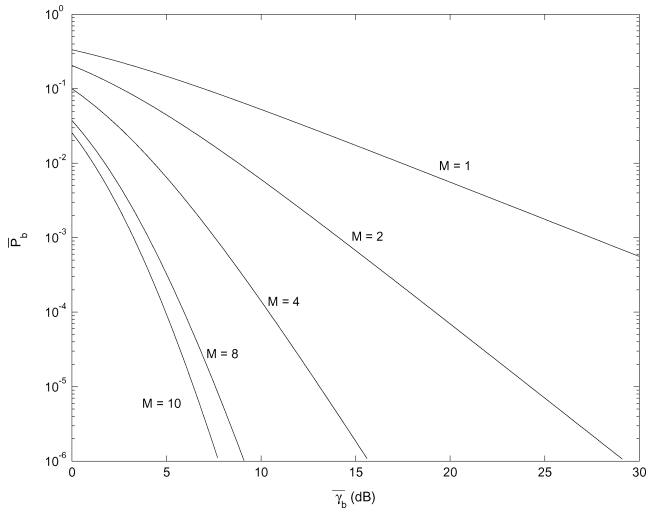

Figure 7.4: Average P_b of BPSK under selection combining with i.i.d. Rayleigh fading.

In the foregoing derivations we assume that there is no correlation between the branch amplitudes. If the correlation is nonzero then there is a slight degradation in performance, which is almost negligible for correlations below .5. Derivation of the exact performance degradation due to branch correlation can be found in [1, Chap. 9.7; 2].

7.2.3 Threshold Combining

Selection combining for systems that transmit continuously may require a dedicated receiver on each branch to continuously monitor branch SNR. A simpler type of combining, called threshold combining, avoids the need for a dedicated receiver on each branch by scanning each of the branches in sequential order and outputting the first signal whose SNR is above a given threshold γ_T. As in SC, co-phasing is not required because only one branch output is used at a time. Hence this technique can be used with either coherent or differential modulation.

Once a branch is chosen, the combiner outputs that signal as long as the SNR on that branch remains above the desired threshold. If the SNR on the selected branch falls below the threshold, the combiner switches to another branch. There are several criteria the combiner can use for determining which branch to switch to [5]. The simplest criterion is to switch randomly to another branch. With only two-branch diversity this is equivalent to switching to the other branch when the SNR on the active branch falls below γ_T. This method is called *switch-and-stay combining* (SSC). The switching process and SNR associated with SSC is illustrated in Figure 7.5. Since the SSC does not select the branch with the highest SNR, as illustrated in Figure 7.5 on the third switch, its performance is between that of no diversity and ideal SC.

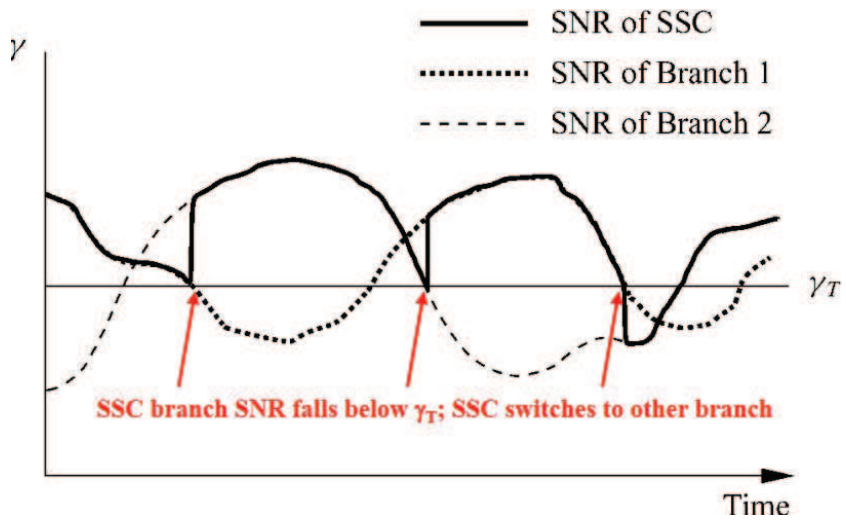

Figure 7.5: SNR of the SSC technique. Switching when the SNR falls below a threshold does not always select the branch with the highest SNR.

Let us denote the SNR on the ith branch by γ_i and the SNR of the combiner output by γ_Σ. The cdf of γ_Σ will depend on the threshold level γ_T and the cdf of γ_i. For two-branch diversity with i.i.d. branch statistics, the cdf of the combiner output $P_{\gamma_\Sigma}(\gamma) = p(\gamma_\Sigma \leq \gamma)$ can be expressed in terms of the cdf $P_{\gamma_i}(\gamma) = p(\gamma_i \leq \gamma)$ and the distribution $p_{\gamma_i}(\gamma)$ of the individual branch SNRs as

$$P_{\gamma_\Sigma}(\gamma) = \begin{cases} P_{\gamma_1}(\gamma_T)P_{\gamma_2}(\gamma) & \gamma < \gamma_T, \\ p(\gamma_T \leq \gamma_2 \leq \gamma) + P_{\gamma_1}(\gamma_T)P_{\gamma_2}(\gamma) & \gamma \geq \gamma_T. \end{cases} \tag{7.13}$$

For Rayleigh fading in each branch with $\bar{\gamma}_i = \bar{\gamma}$ $(i = 1, 2)$, this yields

$$P_{\gamma_\Sigma}(\gamma) = \begin{cases} 1 - e^{-\gamma_T/\bar{\gamma}} - e^{-\gamma/\bar{\gamma}} + e^{-(\gamma_T+\gamma)/\bar{\gamma}} & \gamma < \gamma_T, \\ 1 - 2e^{-\gamma/\bar{\gamma}} + e^{-(\gamma_T+\gamma)/\bar{\gamma}} & \gamma \geq \gamma_T. \end{cases} \tag{7.14}$$

The outage probability P_{out} associated with a given γ_0 is obtained by evaluating $P_{\gamma_\Sigma}(\gamma)$ at $\gamma = \gamma_0$:

$$P_{\text{out}}(\gamma_0) = P_{\gamma_\Sigma}(\gamma_0) = \begin{cases} 1 - e^{-\gamma_T/\bar{\gamma}} - e^{-\gamma_0/\bar{\gamma}} + e^{-(\gamma_T+\gamma_0)/\bar{\gamma}} & \gamma_0 < \gamma_T, \\ 1 - 2e^{-\gamma_0/\bar{\gamma}} + e^{-(\gamma_T+\gamma_0)/\bar{\gamma}} & \gamma_0 \geq \gamma_T. \end{cases} \tag{7.15}$$

The performance of SSC under other types of fading, as well as the effects of fading correlation, is studied in [1, Chap. 9.8; 6; 7]. In particular, it is shown in [1, Chap. 9.8] that, for any fading distribution, SSC with an optimized threshold of $\gamma_T = \gamma_0$ has the same outage probability as SC.

Example 7.2: Find the outage probability of BPSK modulation at $P_b = 10^{-3}$ for two-branch SSC diversity with i.i.d. Rayleigh fading on each branch for threshold values of $\gamma_T = 5\,\text{dB}$, 7 dB, and 10 dB. Assume the average branch SNR is $\bar{\gamma} = 15\,\text{dB}$. Discuss how the outage probability changes with γ_T. Also compare outage probability under SSC with that of SC and no diversity from Example 7.1.

Solution: As in Example 7.1, $\gamma_0 = 7\,\text{dB}$. For $\gamma_T = 5\,\text{dB}$ we have $\gamma_0 \geq \gamma_T$, so we use the second line of (7.15) to get

$$P_{\text{out}} = 1 - 2e^{-10^{.7}/10^{1.5}} + e^{-(10^{.5}+10^{.7})/10^{1.5}} = .0654.$$

For $\gamma_T = 7\,\text{dB}$ we have $\gamma_0 = \gamma_T$, so we again use the second line of (7.15) and obtain

$$P_{\text{out}} = 1 - 2e^{-10^{.7}/10^{1.5}} + e^{-(10^{.7}+10^{.7})/10^{1.5}} = .0215.$$

For $\gamma_T = 10\,\text{dB}$ we have $\gamma_0 < \gamma_T$, so we use the first line of (7.15) to get

$$P_{\text{out}} = 1 - e^{-10/10^{1.5}} - e^{-10^{.7}/10^{1.5}} + e^{-(10+10^{.7})/10^{1.5}} = .0397.$$

We see that the outage probability is smaller for $\gamma_T = 7\,\text{dB}$ than for the other two values. At $\gamma_T = 5\,\text{dB}$ the threshold is too low, so the active branch can be below the target γ_0 for a long time before a switch is made; this contributes to a large outage probability. At $\gamma_T = 10\,\text{dB}$ the threshold is too high: the active branch will often fall below this threshold value, which will cause the combiner to switch to the other antenna even though that other antenna may have a lower SNR than the active one. This example shows that the threshold γ_T minimizing P_{out} equals γ_0.

From Example 7.1, SC has $P_{\text{out}} = .0215$. Thus, $\gamma_T = 7\,\text{dB}$ is the optimal threshold where SSC performs the same as SC. We also see that performance with an unoptimized threshold can be much worse than SC. However, the performance of SSC under all three thresholds is better than the performance without diversity, derived as $P_{\text{out}} = .1466$ in Example 7.1.

We obtain the distribution of γ_Σ by differentiating (7.13) relative to γ. Then the average probability of error is obtained from (7.3) with $P_s(\gamma)$ the probability of symbol error in AWGN and $p_{\gamma_\Sigma}(\gamma)$ the distribution of the SSC output SNR. For most fading distributions and coherent modulations, this result cannot be obtained in closed form and so must be evaluated numerically or by approximation. However, for i.i.d. Rayleigh fading we can differentiate (7.14) to obtain

$$p_{\gamma_\Sigma}(\gamma) = \begin{cases} (1 - e^{-\gamma_T/\bar{\gamma}})(1/\bar{\gamma})e^{-\gamma/\bar{\gamma}} & \gamma < \gamma_T, \\ (2 - e^{-\gamma_T/\bar{\gamma}})(1/\bar{\gamma})e^{-\gamma/\bar{\gamma}} & \gamma \geq \gamma_T. \end{cases} \tag{7.16}$$

As with SC, for most fading distributions and coherent modulations the resulting average probability of error is not in closed form and must be evaluated numerically. However, closed form results do exist for differential

modulation under i.i.d. Rayleigh fading on each branch. In particular, the average probability of symbol error for DPSK is given by

$$\bar{P}_b = \int_0^\infty \frac{1}{2} e^{-\gamma} p_{\gamma_\Sigma}(\gamma)\, d\gamma = \frac{1}{2(1+\bar{\gamma})}(1 - e^{-\gamma_T/\bar{\gamma}} + e^{-\gamma_T} e^{-\gamma_T/\bar{\gamma}}). \tag{7.17}$$

Example 7.3: Find the average probability of error for DPSK modulation under two-branch SSC diversity with i.i.d. Rayleigh fading on each branch for threshold values of $\gamma_T = 3\,\text{dB}$, 7 dB, and 10 dB. Assume the average branch SNR is $\bar{\gamma} = 15\,\text{dB}$. Discuss how the average proability of error changes with γ_T. Also compare average error probability under SSC with that of SC and with no diversity.

Solution: Evaluating (7.17) with $\bar{\gamma} = 15\,\text{dB}$ and $\gamma_T = 3, 7$, and 10 dB yields (respectively) $\bar{P}_b = .0029$, $P_b = .0023$, and $\bar{P}_b = .0042$. As in the previous example, there is an optimal threshold that minimizes average probability of error. Setting the threshold too high or too low degrades performance. From (7.12) we have that SC yields $\bar{P}_b = .5(1+10^{1.5})^{-1} - .5(2+10^{1.5})^{-1} = 4.56 \cdot 10^{-4}$, which is roughly an order of magnitude less than with SSC and an optimized threshold. With no diversity we have $\bar{P}_b = .5(1+10^{1.5})^{-1} = .0153$, which is roughly an order of magnitude worse than with two-branch SSC.

7.2.4 Equal-Gain Combining

Equal-gain combining (EGC) is a more complex technique than those discussed in prior sections, as it co-phases the signals on each branch and then combines them with equal weighting, $\alpha_i = e^{-\theta_i}$. The co-phasing requires a coherent phase detector on each branch, but the instantaneous SNR of the branch is not required, in contrast to maximal-ratio combining. The SNR of the combiner output, assuming equal noise power $N_0 B$ in each branch, is then given by

$$\gamma_\Sigma = \frac{1}{M N_0 B} \left(\sum_{i=1}^{M} r_i \right)^2. \tag{7.18}$$

The distribution and cdf of γ_Σ do not exist in closed form. For i.i.d. Rayleigh fading with two-branch diversity and average branch SNR $\bar{\gamma}$, an expression for the cdf in terms of the Q-function can be derived [4, Chap. 6.4; 8, Chap. 5.6] as

$$P_{\gamma_\Sigma}(\gamma) = 1 - e^{-2\gamma/\bar{\gamma}} - \sqrt{\pi\gamma/\bar{\gamma}} e^{-\gamma/\bar{\gamma}} (1 - 2Q(\sqrt{2\gamma/\bar{\gamma}})). \tag{7.19}$$

The resulting outage probability is given by

$$P_{\text{out}}(\gamma_0) = 1 - e^{-2\gamma_R} - \sqrt{\pi\gamma_R} e^{-\gamma_R} (1 - 2Q(\sqrt{2\gamma_R})), \tag{7.20}$$

where $\gamma_R = \gamma_0/\bar{\gamma}$. Differentiating (7.19) with respect to γ yields the distribution

$$p_{\gamma_\Sigma}(\gamma) = \frac{1}{\bar{\gamma}} e^{-2\gamma/\bar{\gamma}} + \sqrt{\pi} e^{-\gamma/\bar{\gamma}} \left(\frac{1}{\sqrt{4\gamma\bar{\gamma}}} - \frac{1}{\bar{\gamma}} \sqrt{\frac{\gamma}{\bar{\gamma}}} \right) \left(1 - 2Q\left(\sqrt{\frac{2\gamma}{\bar{\gamma}}} \right) \right). \tag{7.21}$$

Substituting this into (7.3) for BPSK yields the average probability of bit error:

$$\bar{P}_b = \int_0^\infty Q(\sqrt{2\gamma}) p_{\gamma_\Sigma}(\gamma) d\gamma = .5 \left(1 - \sqrt{1 - \left(\frac{1}{1+\bar{\gamma}} \right)^2} \right). \tag{7.22}$$

7.2.5 Maximal-Ratio Combining

In maximal-ratio combining (MRC) the output is a weighted sum of all branches, so the α_i in Figure 7.1 are all nonzero, and the weights are determined to maximize the combiner output's SNR. The signals are co-phased and so $\alpha_i = a_i e^{-j\theta_i}$, where θ_i is the phase of the incoming signal on the ith branch. Thus, the envelope of the combiner output will be $r = \sum_{i=1}^{M} a_i r_i$. Assuming the same noise power $N_0 B$ in each branch yields a total noise power at the combiner output of $\sum_{i=1}^{M} a_i^2 N_0 B$. Thus, the output SNR of the combiner is

$$\gamma_\Sigma = \frac{1}{N_0 B} \frac{\left(\sum_{i=1}^{M} a_i r_i\right)^2}{\sum_{i=1}^{M} a_i^2}. \tag{7.23}$$

The goal of MRC is to choose the a_i to maximize γ_Σ. Intuitively, branches with a high SNR should be weighted more than branches with a low SNR. We find the a_is that maximize γ_Σ by taking partial derivatives of (7.23) or finding the weights that achieve equality in the Cauchy-Schwartz inequality[2]. Solving for the optimal weights yields $a_i^2 = c\gamma_i = c r_i^2/(N_0 B)$ for any constant c where the constant c cancels out in the SNR calculation. Setting $a_i = r_i$ yields the combiner SNR

$$\gamma_\Sigma = \sum_{i=1}^{M} \frac{r_i^2}{N_0 B} = \sum_{i=1}^{M} \gamma_i. \tag{7.24}$$

Thus, the SNR of the combiner output is the sum of SNRs on each branch. Hence, the average combiner SNR and corresponding array gain increase linearly with the number of diversity branches M, in contrast to the diminishing returns associated with the average combiner SNR in SC given by (7.11). As with SC, the distribution of the combiner output SNR does not remain exponential even when there is Rayleigh fading on all branches.

To obtain the distribution of γ_Σ, we take the product of the exponential moment generating functions or characteristic functions. Assuming i.i.d. Rayleigh fading on each branch with equal average branch SNR $\bar{\gamma}$, the distribution of γ_Σ is χ^2 with $2M$ degrees of freedom, expected value $\bar{\gamma}_\Sigma = M\bar{\gamma}$, and variance $2M\bar{\gamma}$:

$$p_{\gamma_\Sigma}(\gamma) = \frac{\gamma^{M-1} e^{-\gamma/\bar{\gamma}}}{\bar{\gamma}^M (M-1)!}, \quad \gamma \geq 0. \tag{7.25}$$

The corresponding outage probability for a given threshold γ_0 is given by

$$P_{\text{out}} = p(\gamma_\Sigma < \gamma_0) = \int_0^{\gamma_0} p_{\gamma_\Sigma}(\gamma)\, d\gamma = 1 - e^{-\gamma_0/\bar{\gamma}} \sum_{k=1}^{M} \frac{(\gamma_0/\bar{\gamma})^{k-1}}{(k-1)!}. \tag{7.26}$$

Figure 7.6 is a plot of P_{out} for maximal-ratio combining indexed by the number of diversity branches.

The average probability of symbol error is obtained from (7.3) with $P_s(\gamma)$ the probability of symbol error in AWGN for the signal modulation and $p_{\gamma_\Sigma}(\gamma)$ the distribution of γ_Σ. For BPSK modulation with i.i.d. Rayleigh fading, where $p_{\gamma_\Sigma}(\gamma)$ is given by (7.25), it can be shown [4, Chap. 6.3] that

$$\bar{P}_b = \int_0^{\infty} Q(\sqrt{2\gamma}) p_{\gamma_\Sigma}(\gamma)\, d\gamma = \left(\frac{1-\Gamma}{2}\right)^M \sum_{m=0}^{M-1} \binom{M-1+m}{m} \left(\frac{1+\Gamma}{2}\right)^m, \tag{7.27}$$

where $\Gamma = \sqrt{\bar{\gamma}/(1+\bar{\gamma})}$. This equation is plotted in Figure 7.7. Comparing the outage probability for MRC in Figure 7.6 with that of SC in Figure 7.3, or comparing the average probability of error for MRC in Figure 7.7 with

[2]The Cauchy-Schwartz inequality for two vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ is $\sum_i x_i y_i \leq \sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}$

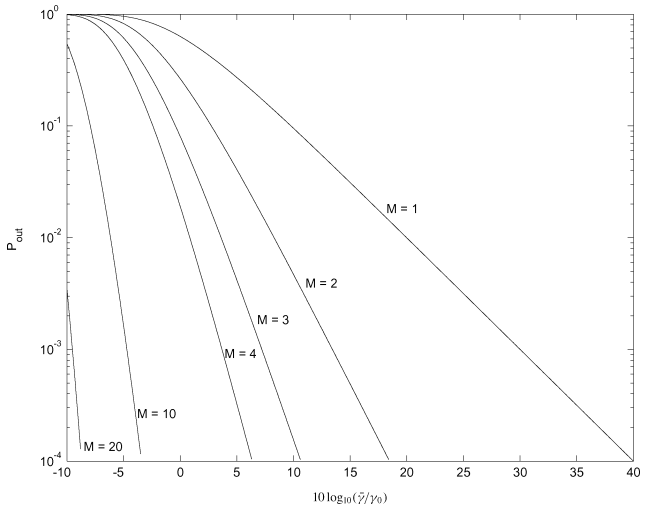

Figure 7.6: P_{out} for maximal-ratio combining with i.i.d. Rayleigh fading.

that of SC in Figure 7.4, indicates that MRC has significantly better performance than SC. In Section 7.4 we will use a different analysis based on moment generating functions to compute average error probability under MRC, which can be applied to any modulation type, any number of diversity branches, and any fading distribution on the different branches.

For modulations whose probability of error in AWGN can be approximated as $P_s = \alpha Q(\sqrt{\beta\gamma})$ for SNR γ, e.g. BPSK, MPSK, and MQAM, we can obtain a simple upper bound on the average probability of error by applying the Chernoff bound $Q(x) \leq e^{-x^2/2}$ from (5.43) to the Q-function. For these modulations, we have

$$P_s = \alpha Q(\sqrt{\beta\gamma_\Sigma}) \leq \alpha e^{-\beta\gamma_\Sigma/2} = \alpha e^{-\beta(\gamma_1+\cdots+\gamma_M)/2}. \tag{7.28}$$

Integrating over the χ^2 distribution for γ_Σ yields

$$\bar{P}_s \leq \alpha \prod_{i=1}^{M} \frac{1}{1+\beta\bar{\gamma}_i/2}. \tag{7.29}$$

In the limit of high SNR and assuming that the γ_i are identically distributed with $\bar{\gamma}_i = \bar{\gamma}$, we have

$$\bar{P}_s \approx \alpha \left(\frac{\beta\bar{\gamma}}{2}\right)^{-M}. \tag{7.30}$$

Thus, at high SNR, the diversity order of MRC is M (the number of antennas) and so MRC achieves full diversity order.

MRC is the only diversity-combining technique that achieves full diversity order. However, it is shown in [8, Chap. 5.7] that performance of EGC is quite close to that of MRC, typically exhibiting less than 1 dB of power penalty. This is the price paid for the reduced complexity of using equal gains. A more extensive performance comparison between SC, MRC, and EGC can be found in [1, Chap. 9].

Example 7.4: Compare the average probability of bit error with BPSK under MRC and EGC two-branch diversity, assuming i.i.d. Rayleigh fading with an average SNR of 10 dB on each branch.

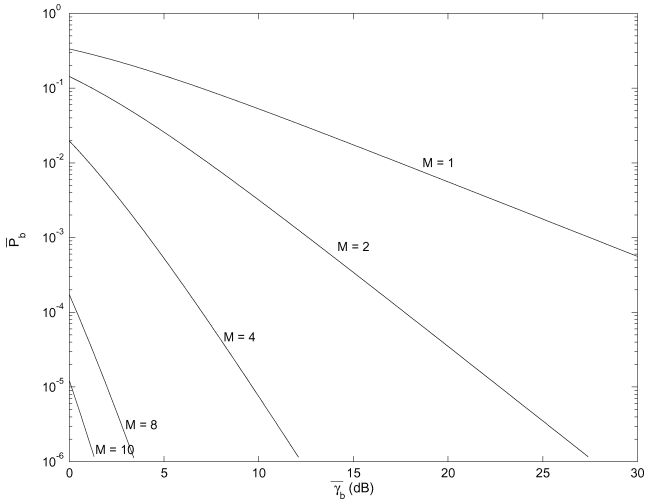

Figure 7.7: Average P_b for maximal-ratio combining with i.i.d. Rayleigh fading.

Solution: By (7.27), under MRC we have

$$\bar{P}_b = \left(\frac{1-\sqrt{10/11}}{2}\right)^2 (2+\sqrt{10/11}) = 1.60 \cdot 10^{-3}.$$

By (7.22), under EGC we have

$$\bar{P}_b = .5\left(1-\sqrt{1-\left(\frac{1}{11}\right)^2}\right) = 2.07 \cdot 10^{-3}.$$

Thus we see that the performance of MRC and EGC are almost the same.

7.2.6 Generalized Combining

The three most common diversity-combining schemes used in practice are MRC, EGC, and SC. MRC has the best performance and achieves full diversity order. However, MRC also has the highest complexity, as it requires coherent detection in each branch that determines the branch phase and amplitude. SC is the simplest of these three techniques as only one coherent detector is needed. The price paid for this simplicity is in performance, with diminishing diversity gain with each branch that is added. EGC has complexity and performance between the extremes of MRC and SC. Generalized Combining (GC), also referred to as Generalized Selection Combining or Hybrid Combining, utilizes SC to pick a subset of the branch signals which are then combined using MRC or EGC. In particular, a generalized combiner has $M > 2$ diversity branches of which it selects the $L > 1$ branches, $L < M$, with the highest SNR; these L branches are then combined using MRC or EGC. When MRC is used, the performance is better than an L-branch MRC system and worse than an M-branch MRC system, and similarly for EGC, due to the selection of the $L < M$ highest-SNR branches among the M available branches. Performance analysis of generalized combining under both MRC and EGC in different fading environments has been studied extensively (see [1, Chapter 9.11] and the references therein). This performance is best analyzed via the Moment Generating Function analysis for diversity systems, as will be explored in Section 7.4.

7.3 Transmitter Diversity

In transmit diversity there are multiple transmit antennas, and the transmit power is divided among these antennas. Transmit diversity is desirable in systems where more space, power, and processing capability is available on the transmit side than on the receive side, as best exemplified by cellular systems. Transmit diversity design depends on whether or not the complex channel gain is known to the transmitter. When this gain is known, the system is quite similar to receiver diversity. However, without this channel knowledge, transmit diversity gain requires a combination of space and time diversity via a novel technique called the *Alamouti scheme* and its extensions. We now discuss transmit diversity under different assumptions about channel knowledge at the transmitter, assuming the channel gains are known at the receiver.

7.3.1 Channel Known at Transmitter

Consider a transmit diversity system with M transmit antennas and one receive antenna. We assume that the path gain $r_i e^{j\theta_i}$ associated with the ith antenna is known at the transmitter; this is referred to as having channel side information (CSI) at the transmitter, or CSIT. Let $s(t)$ denote the transmitted signal with average power P_t. The signal is multiplied by a complex gain $\alpha_i = a_i e^{-j\theta_i} (0 \leq a_i \leq 1)$ and then sent through the ith antenna. This complex multiplication performs both co-phasing and weighting relative to the channel gains. Because of the average power constraint P_t, the weights must satisfy $\sum_{i=1}^{M} a_i^2 = 1$. The weighted signals transmitted over all antennas are added "in the air", which leads to a received signal given by

$$r(t) = \sum_{i=1}^{M} a_i r_i s(t). \tag{7.31}$$

Let $N_0 B$ denote the noise power in the receiver and assume Nyquist pulses such that $BT_s = 1$.

Suppose we wish to set the branch weights to maximize received SNR. Using a similar analysis as in receiver MRC diversity, we see that the weights a_i that achieve the maximum SNR are given by

$$a_i = \frac{r_i}{\sqrt{\sum_{i=1}^{M} r_i^2}}, \tag{7.32}$$

and the resulting SNR is

$$\gamma_\Sigma = \frac{P_t}{N_0 B} \sum_{i=1}^{M} r_i^2 = \sum_{i=1}^{M} \gamma_i \tag{7.33}$$

for $\gamma_i = r_i^2 P_t / (N_0 B)$ equal to the branch SNR between the ith transmit antenna and the receive antenna. Thus we see that transmit diversity when the channel gains are known to the transmitter is very similar to receiver diversity with MRC: the received SNR is the sum of SNRs on each of the individual branches. In particular, if all antennas have the same gain $r_i = r$ then $\gamma_\Sigma = M r^2 P_t / (N_0 B)$, so there is an array gain of M corresponding to an M-fold increase in SNR over a single antenna transmitting with full power.

We again consider modulations whose error probability in AWGN can be approximated as $P_s = \alpha Q(\sqrt{\beta\gamma})$ for SNR γ, such as BPSK, MPSK, and MQAM. Using the Chernoff bound as in Section 7.2.5 we see that, for such modulations and static gains on each diversity branch,

$$P_s = \alpha Q(\sqrt{\beta\gamma_\Sigma}) \leq \alpha e^{-\beta\gamma_\Sigma/2} = \alpha e^{-\beta(\gamma_1 + \cdots + \gamma_M)/2}. \tag{7.34}$$

Integrating over the χ^2 distribution for γ_Σ yields the same bound as (7.29):

$$\bar{P}_s \leq \alpha \prod_{i=1}^{M} \frac{1}{1 + \beta\bar{\gamma}_i/2}. \tag{7.35}$$

In the limit of high SNR and assuming that the γ_i are identically distributed with $\bar{\gamma}_i = \bar{\gamma}$, we have

$$\bar{P}_s \approx \alpha \left(\frac{\beta \bar{\gamma}}{2} \right)^{-M}. \tag{7.36}$$

Thus, at high SNR, the diversity order of transmit diversity with MRC is M, so both transmit and receive MRC achieve full diversity order. The analysis for EGC and SC assuming transmitter channel knowledge is the same as under receiver diversity.

The complication of transmit diversity is to obtain the channel phase and, for SC and MRC, the channel gain, at the transmitter. These channel values can be measured at the receiver using a pilot technique and then fed back to the transmitter. Alternatively, in cellular systems with time division, the base station can measure the channel gain and phase on transmissions from the mobile to the base and then use these measurements when transmitting back to the mobile, since under time division the forward and reverse links are reciprocal.

7.3.2 Channel Unknown at Transmitter – The Alamouti Scheme

We now consider the same model as in Section 7.3.1 but assume here that the transmitter no longer knows the channel gains $r_i e^{j\theta_i}$, so there is no CSIT. In this case it is not obvious how to obtain diversity gain. Consider, for example, a naive strategy whereby for a two-antenna system we divide the transmit energy equally between the two antennas. Thus, the transmit signal on antenna i will be $s_i(t) = \sqrt{.5}s(t)$ for $s(t)$ the transmit signal with energy per symbol E_s. Assume that each antenna has a complex Gaussian channel gain $h_i = r_i e^{j\theta_i} (i = 1, 2)$ with mean 0 and variance 1. The received signal is then

$$r(t) = \sqrt{.5}(h_1 + h_2)s(t). \tag{7.37}$$

Note that $h_1 + h_2$ is the sum of two complex Gaussian random variables and thus is itself a complex Gaussian, with mean equal to the sum of means (0) and variance equal to the sum of variances (2). Hence $\sqrt{.5}(h_1 + h_2)$ is a complex Gaussian random variable with mean 0 and variance 1, so the received signal has the same distribution as if we had just used one antenna with the full energy per symbol. In other words, we have obtained no performance advantage from the two antennas, since we could neither divide our energy intelligently between them nor obtain coherent combining via co-phasing.

Transmit diversity gain can be obtained even in the absence of channel information, given an appropriate scheme to exploit the antennas. A particularly simple and prevalent scheme for this diversity that combines both space and time diversity was developed by Alamouti in [11]. Alamouti's scheme is designed for a digital communication system with two-antenna transmit diversity. The scheme works over two symbol periods and it is assumed that the channel gain is constant over this time. Over the first symbol period, two different symbols s_1 and s_2 (each with energy $E_s/2$) are transmitted simultaneously from antennas 1 and 2, respectively. Over the next symbol period, symbol $-s_2^*$ is transmitted from antenna 1 and symbol s_1^* is transmitted from antenna 2, each again with symbol energy $E_s/2$.

Assume complex channel gains $h_i = r_i e^{j\theta_i} (i = 1, 2)$ between the ith transmit antenna and the receive antenna. The received symbol over the first symbol period is $y_1 = h_1 s_1 + h_2 s_2 + n_1$ and the received symbol over the second symbol period is $y_2 = -h_1 s_2^* + h_2 s_1^* + n_2$, where $n_i (i = 1, 2)$ is the AWGN sample at the receiver associated with the ith symbol transmission. We assume that the noise sample has mean 0 and power N.

The receiver uses these sequentially received symbols to form the vector $\mathbf{y} = [y_1 \;\; y_2^*]^T$ given by

$$\mathbf{y} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2^* \end{bmatrix} = \mathbf{H}_A \mathbf{s} + \mathbf{n},$$

where $\mathbf{s} = [s_1 \;\; s_2]^T$, $\mathbf{n} = [n_1 \;\; n_2]^T$, and

$$\mathbf{H}_A = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix}.$$

Let us define the new vector $\mathbf{z} = \mathbf{H}_A^H \mathbf{y}$. The structure of $\mathbf{H}_A$ implies that

$$\mathbf{H}_A^H \mathbf{H}_A = (|h_1^2| + |h_2^2|)\mathbf{I}_2 \tag{7.38}$$

is diagonal and thus

$$\mathbf{z} = [z_1 \;\; z_2]^T = (|h_1^2| + |h_2^2|)\mathbf{I}_2 \mathbf{s} + \tilde{\mathbf{n}}, \tag{7.39}$$

where $\tilde{\mathbf{n}} = \mathbf{H}_A^H \mathbf{n}$ is a complex Gaussian noise vector with mean 0 and covariance matrix $E[\tilde{\mathbf{n}}\tilde{\mathbf{n}}^*] = (|h_1^2| + |h_2^2|)N_0\mathbf{I}_2$. The diagonal nature of $\mathbf{z}$ effectively decouples the two symbol transmissions, so that each component of $\mathbf{z}$ corresponds to one of the transmitted symbols:

$$z_i = (|h_1^2| + |h_2^2|)s_i + \tilde{n}_i, \quad i = 1, 2. \tag{7.40}$$

The received SNR thus corresponds to the SNR for z_i given by

$$\gamma_i = \frac{(|h_1^2| + |h_2^2|)P_t}{2N_0 B}, \tag{7.41}$$

where the factor of 2 comes from the fact that s_i is transmitted using half the total power P_t. The received SNR is thus equal to the sum of SNRs on each branch divided by 2. By (7.40) the Alamouti scheme achieves a diversity order of 2 – the maximum possible for a two-antenna transmit system – despite the fact that channel knowledge is not available at the transmitter. However, by (7.41) it only achieves an array gain of 1, whereas MRC can achieve an array gain and a diversity gain of 2. The Alamouti scheme can be generalized for $M > 2$; this generalization falls into the category of orthogonal space-time block code design [10, Chap. 6.3.3; 11, Chap. 7.4].

7.4 Moment Generating Functions in Diversity Analysis

In this section we use the MGFs introduced in Section 6.3.3 to greatly simplify the analysis of average error probability under diversity. The use of MGFs in diversity analysis arises from the difficulty in computing the distribution $p_{\gamma_\Sigma}(\gamma)$ of the combiner SNR γ_Σ. Specifically, although the average probability of error and outage probability associated with diversity combining are given by the simple formulas (7.3) and (7.4), these formulas require integration over the distribution $p_{\gamma_\Sigma}(\gamma)$. This distribution is often not in closed form for an arbitrary number of diversity branches with different fading distributions on each branch, regardless of the combining technique that is used. In particular, the distribution for $p_{\gamma_\Sigma}(\gamma)$ is often in the form of an infinite-range integral, in which case the expressions for (7.3) and (7.4) become double integrals that can be difficult to evaluate numerically. Even when $p_{\gamma_\Sigma}(\gamma)$ is in closed form, the corresponding integrals (7.3) and (7.4) may not lead to closed-form solutions and may also be difficult to evaluate numerically. A large body of work over many decades has addressed approximations and numerical techniques for computing the integrals associated with average probability of symbol error for different modulations, fading distributions, and combining techniques (see [14] and the references therein). Expressing the average error probability in terms of the MGF for γ_Σ instead of its distribution often eliminates these integration difficulties. Specifically, when the diversity fading paths are independent but not necessarily identically distributed, the average error probability based on the MGF of γ_Σ is typically in closed form or consists of a single finite-range integral that can be easily computed numerically.

The simplest application of MGFs in diversity analysis is for coherent modulation with MRC, so this is treated first. We then discuss the use of MGFs in the analysis of average error probability under EGC, SC, and GC.

7.4.1 Diversity Analysis for MRC

The simplicity of using MGFs in the analysis of maximal-ratio combining stems from the fact that, as derived in Section 7.2.5, the combiner SNR γ_Σ is the sum of the γ_i, the branch SNRs:

$$\gamma_\Sigma = \sum_{i=1}^{M} \gamma_i. \tag{7.42}$$

As in the analysis of average error probability without diversity (Section 6.3.3), let us again assume that the probability of error in AWGN for the modulation of interest can be expressed either as an exponential function of γ_s, as in (6.66), or as a finite range integral of such a function, as in (6.67).

We first consider the case where $P_s = a\exp[-b\gamma]$ for SNR γ as given in (6.66), e.g. for DPSK with $a = .5$ and $b = 1$. Then the average probability of symbol error under MRC is

$$\bar{P}_s = \int_0^\infty a\exp[-b\gamma]p_{\gamma_\Sigma}(\gamma)\, d\gamma. \tag{7.43}$$

We assume that the branch SNRs are independent, so that their joint distribution becomes a product of the individual distributions: $p_{\gamma_1,\ldots,\gamma_M}(\gamma_1,\ldots,\gamma_M) = p_{\gamma_1}(\gamma_1)\ldots p_{\gamma_M}(\gamma_M)$. Using this factorization and substituting $\gamma = \gamma_1 + \cdots + \gamma_M$ in (7.43) yields

$$\bar{P}_s = a\underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{M\text{-fold}} \exp[-b(\gamma_1 + \cdots + \gamma_M)]p_{\gamma_1}(\gamma_1)\ldots p_{\gamma_M}(\gamma_M)d\gamma_1\ldots d\gamma_M. \tag{7.44}$$

Now using the product forms $\exp[b(\gamma_1 + \cdots + \gamma_M)] = \prod_{i=1}^M \exp[b\gamma_i]$ and $p_{\gamma_1}(\gamma_1)\ldots p_{\gamma_M}(\gamma_M) = \prod_{i=1}^M p_{\gamma_i}(\gamma_i)$ in (7.44) yields

$$\bar{P}_s = a\underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{M\text{-fold}} \prod_{i=1}^{M} \exp[-b\gamma_i]p_{\gamma_i}(\gamma_i)d\gamma_i. \tag{7.45}$$

Finally, switching the order of integration and multiplication in (7.45) yields our desired final form:

$$\bar{P}_s = a\prod_{i=1}^{M}\int_0^\infty \exp[-b\gamma_i]p_{\gamma_i}(\gamma_i)d\gamma_i = a\prod_{i=1}^{M}\mathcal{M}_{\gamma_i}(-b), \tag{7.46}$$

where $\mathcal{M}_{\gamma_i}(s)$ is the MGF of the fading distribution for the ith diversity branch as given by (6.62), (6.63), and (6.64) for (respectively) Rayleigh, Rician, and Nakagami fading. Thus, the average probability of symbol error is just the product of MGFs associated with the SNR on each branch.

We now consider modulations where $P_s = \alpha Q(\sqrt{\beta\gamma})$ in AWGN for γ the SNR. Using the alternate Q-function representation we get from (6.72) that

$$\begin{aligned}
\bar{P}_s &= \int_0^\infty \int_0^{\pi/2} a\exp[-b(\phi)\gamma]d\phi\, p_{\gamma_\Sigma}(\gamma)d\gamma \\
&= \underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{M\text{-fold}} \int_0^{\pi/2} a\prod_{i=1}^{M}\exp[-b(\phi)\gamma_i]p_{\gamma_i}(\gamma_i)\, d\gamma_i d\phi.
\end{aligned} \tag{7.47}$$

with $a = \alpha/\pi$ and $b(\phi) = .5\beta/\sin^2\phi$; Again, switching the order of integration and multiplication yields our desired final form:

$$\bar{P}_s = \frac{\alpha}{\pi}\int_0^{\pi/2}\prod_{i=1}^{M}\int_0^{\infty}\exp\left[\frac{-.5\beta}{\sin^2\phi}\gamma_i\right]p_{\gamma_i}(\gamma_i)d\gamma_i dx = \frac{\alpha}{\pi}\int_0^{\pi/2}\prod_{i=1}^{M}\mathcal{M}_{\gamma_i}\left(\frac{-.5\beta}{\sin^2\phi}\right)\,d\phi. \tag{7.48}$$

Thus, the average probability of symbol error is just a single finite-range integral of the product of MGFs associated with the SNR on each branch. The simplicity of (7.46) and (7.48) are quite remarkable, given that these expressions apply for any number of diversity branches and any type of fading distribution on each branch (as long as the branch SNRs are independent).

We now apply these general results to specific modulations and fading distributions. Let us first consider DPSK, where $P_b(\gamma_b) = .5e^{-\gamma b}$ in AWGN is in the form of (6.66) with $a = 1/2$ and $b = 1$. Then, by (7.46), the average probability of bit error in DPSK under M-fold MRC diversity is

$$\bar{P}_b = \frac{1}{2}\prod_{i=1}^{M}\mathcal{M}_{\gamma_i}(-1). \tag{7.49}$$

Note that this reduces to the probability of average bit error without diversity given by (6.59) for $M = 1$.

Example 7.5: Compute the average probability of bit error for DPSK modulation under three-branch MRC, assuming i.i.d. Rayleigh fading in each branch with $\bar{\gamma}_1 = 15$ dB and $\bar{\gamma}_2 = \bar{\gamma}_3 = 5$ dB. Compare with the case of no diversity for $\bar{\gamma} = 15$ dB.

Solution: From (6.62), $\mathcal{M}_{\gamma_i}(s) = (1 - s\bar{\gamma}_i)^{-1}$. Using this MGF in (7.49) with $s = -1$ yields

$$\bar{P}_b = \frac{1}{2}\frac{1}{1+10^{1.5}}\left(\frac{1}{1+10^{.5}}\right)^2 = 8.85\cdot 10^{-4}.$$

With no diversity we have

$$\bar{P}_b = \frac{1}{2(1+10^{1.5})} = 1.53\cdot 10^{-2}.$$

This indicates that additional diversity branches can significantly reduce average BER, even when the SNR on these branches is somewhat low.

Example 7.6: Compute the average probability of bit error for DPSK modulation under three-branch MRC, assuming: Nakagami fading in the first branch with $m = 2$ and $\bar{\gamma}_1 = 15$ dB; Rician fading in the second branch with $K = 3$ and $\bar{\gamma}_2 = 5$ dB; and Nakagami fading in the third branch with $m = 4$ and $\bar{\gamma}_3 = 5$ dB. Compare with the results of the prior example.

Solution: From (6.63) and (6.64), for Nakagami fading $\mathcal{M}_{\gamma_i}(s) = (1 - s\bar{\gamma}_i/m)^{-m}$ and for Rician fading

$$\mathcal{M}_{\gamma_s}(s) = \frac{1+K}{1+K-s\bar{\gamma}_s}\exp\left[\frac{Ks\bar{\gamma}_s}{1+K-s\bar{\gamma}_s}\right].$$

Using these MGFs in (7.49) with $s = -1$ yields

$$\bar{P}_b = \frac{1}{2}\left(\frac{1}{1+10^{1.5}/2}\right)^2\frac{4}{4+10^{.5}}\exp\left[\frac{-3\cdot 10^{.5}}{4+10^{.5}}\right]\left(\frac{1}{1+10^{.5}/4}\right)^4 = 6.9\cdot 10^{-5},$$

which is more than an order of magnitude lower than the average error probability under i.i.d. Rayleigh fading with the same branch SNRs derived in the previous problem. This indicates that Nakagami and Rician fading are much more benign distributions than Rayleigh, especially when multiple branches are combined under MRC. This example also illustrates the power of the MGF approach: computing average probability of error when the branch SNRs follow different distributions simply consists of multiplying together different functions in closed form; the result is then also in closed form. In contrast, computing the distribution of the sum of random variables from different families involves the convolution of their distributions, which rarely leads to a closed-form distribution.

For BPSK we see from (6.44) that P_b has the same form as (6.67) with $a = 1/\pi$, and $b(\phi) = 1/\sin^2\phi$. Thus we obtain the average bit error probability for BPSK with M-fold diversity as

$$\bar{P}_b = \frac{1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{M}\mathcal{M}_{\gamma_i}\left(-\frac{1}{\sin^2\phi}\right)d\phi. \tag{7.50}$$

Similarly, if $P_s = \alpha Q(\sqrt{\beta\gamma_s})$ then the average probability of error in fading with M-branch diversity is given in (7.48) as

$$\bar{P}_s = \frac{\alpha}{\pi}\int_0^{\pi/2}\prod_{i=1}^{M}\mathcal{M}_{\gamma_i}\left(-\frac{.5\beta}{\sin^2\phi}\right)d\phi. \tag{7.51}$$

If the branch SNRs are i.i.d. then this expression simplifies to

$$\bar{P}_s = \frac{\alpha}{\pi}\int_0^{\pi/2}\left(\mathcal{M}_{\gamma}\left(-\frac{.5\beta}{\sin^2\phi}\right)\right)^{M}d\phi, \tag{7.52}$$

where $\mathcal{M}_\gamma(s)$ is the common MGF for the branch SNRs.

We can also use the MGF approach to determine the exact probability of symbol error for MPSK in fading. For this analysis, i.e. in equations (7.55)-(7.56), we use N to denote the number of branches and M for the signal constellation size. This analysis extends the derivation of (6.74) to multiple branches. As in that derivation, we note that Craig's formula for P_s of MPSK in AWGN is in the desired form (6.67) for MGF analysis. Using this form, we get the average symbol error probability in N-branch diversity as

$$\begin{aligned}
\bar{P}_s &= \int_0^{\infty}\int_0^{(M-1)\pi/M} a\exp[-b(\phi)\gamma]d\phi\, p_{\gamma_\Sigma}(\gamma)d\gamma &(7.53)\\
&= \underbrace{\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}}_{N\text{-fold}}\int_0^{(M-1)\pi/M} a\prod_{i=1}^{N}\exp[-b(\phi)\gamma_i]p_{\gamma_i}(\gamma_i)\,d\gamma_i d\phi &(7.54)\\
&= \bar{P}_s = \frac{1}{\pi}\int_0^{(M-1)\pi/M}\prod_{i=1}^{N}\mathcal{M}_{\gamma_i}\left(-\frac{\beta_M}{\sin^2\phi}\right)d\phi. &(7.55)
\end{aligned}$$

for $\beta_M = \sin^2(\pi/M)$. For i.i.d. fading this simplifies to

$$\bar{P}_s = \frac{1}{\pi}\int_0^{(M-1)\pi/M}\left(\mathcal{M}_{\gamma}\left(-\frac{\beta_M}{\sin^2\phi}\right)\right)^{N}d\phi. \tag{7.56}$$

Example 7.7: Find an expression for the average symbol error probability for 8-PSK modulation for two-branch

MRC combining, where each branch is Rayleigh fading with average SNR of 20 dB.

Solution: The MGF for Rayleigh fading is $\mathcal{M}_{\gamma_i}(s) = (1 - s\bar{\gamma}_i)^{-1}$. Using this MGF in (7.56) with $s = -(\sin^2 \pi/8)/\sin^2 \phi$ and $\bar{\gamma} = 100$ yields

$$\bar{P}_s = \frac{1}{\pi} \int_0^{7\pi/8} \left(\frac{1}{1 + (100 \sin^2 \pi/8)/\sin^2 \phi} \right)^2 d\phi.$$

This expression does not lead to a closed-form solution and so must be evaluated numerically, which results in $\bar{P}_s = 1.56 \cdot 10^{-3}$.

We can use similar techniques to extend the derivation of the exact error probability for MQAM in fading, given by (6.80), to include MRC. As with MPSK, we let N denote the number of branches and M the constellation size in the expression (7.57) for $\bar{P}_s$. For this analysis we first integrate the expression for P_s in AWGN, expressed in (6.79) using the alternate representation of Q and Q^2, over the distribution of γ_Σ. Since $\gamma_\Sigma = \sum_i \gamma_i$ and the SNRs are independent, the exponential function and distribution in the resulting expression can be written in product form. Then we use the same reordering of integration and multiplication as in the MPSK derivation. The resulting average probability of symbol error for MQAM modulation with MRC combining of N branches is given by

$$\begin{aligned} \bar{P}_s =& \frac{4}{\pi} \left(1 - \frac{1}{\sqrt{M}} \right) \int_0^{\pi/2} \prod_{i=1}^{N} \mathcal{M}_{\gamma_i} \left(-\frac{.5\beta}{\sin^2 \phi} \right) d\phi \\ & - \frac{4}{\pi} \left(1 - \frac{1}{\sqrt{M}} \right)^2 \int_0^{\pi/4} \prod_{i=1}^{N} \mathcal{M}_{\gamma_i} \left(-\frac{.5\beta}{\sin^2 \phi} \right) d\phi. \end{aligned} \tag{7.57}$$

More details on the use of MGFs to obtain average probability of error under M-fold MRC diversity for a broad class of modulations can be found in [1, Chap. 9.2].

7.4.2 Diversity Analysis for EGC, SC, and GC

Moment generating functions are less useful in the analysis of EGC and SC than in MRC. The reason is that, with MRC, $\gamma_\Sigma = \sum_i \gamma_i$ and so $\exp[-b\gamma_\Sigma] = \prod_i \exp[-b\gamma_i]$. This factorization leads directly to the simple formulas whereby probability of symbol error is based on a product of MGFs associated with each of the branch SNRs. Unfortunately, neither EGC nor SC leads to this type of factorization. However, working with the MGF of γ_Σ can sometimes lead to simpler results than working directly with its distribution. This is illustrated in [1, Chap. 9.3.3]: the exact probability of symbol error for MPSK with EGC is obtained based on the characteristic function associated with each branch SNR, where the characteristic function is just the MGF evaluated at $s = j2\pi f$ (i.e., it is the Fourier transform of the distribution). The resulting average error probability, given by [1, eq. (9.78)], is a finite-range integral over a sum of closed-form expressions and thus is easily evaluated numerically.

In the case of GC where the $L < M$ fading branches with the L largest SNRs are combined using MRC, the MGF approach combined with order statistics can be used for performance analysis. Let $\gamma_i, i = 1, \ldots, M$, denote the SNR on the ith branch. We can order these random variables from smallest to largest as $\gamma_{(1)}, \ldots, \gamma_{(M)}$, $\gamma_{(i)} \leq \gamma_{(i+1)}$, with $\gamma_{(1)} = \min[\gamma_1, \gamma_2, \ldots, \gamma_M]$ and $\gamma_{(M)} = \max[\gamma_1, \gamma_2, \ldots, \gamma_M]$. Note that these ordered random variables are not i.i.d. The largest L of these values are $\gamma_{(M)}, \gamma_{(M-1)}, \ldots, \gamma_{(M-L+1)}$ where, by definition, $\gamma_{(M)} \geq \gamma_{(M-1)} \geq, \ldots, \geq \gamma_{(M-L+1)}$. For GC using MRC to combine diversity branches with these L largest SNRs, its SNR is

$$\gamma_\Sigma = \sum_{i=M-L+1}^{M} \gamma_{(i)}. \tag{7.58}$$

The distribution of γ_Σ is difficult to obtain directly for general distributions on the branch SNRs, as is the case for MRC of all branches (i.e. when $M = L$). However, as with MRC, we can obtain the distribution of γ_Σ through the MGF approach as follows.

Since $\gamma_i, i = 1, \ldots, M$ are i.i.d., the joint distribution of $\gamma_{(M)}, \ldots, \gamma_{(M-L+1)}$, i.e.

$$p_{\gamma_{(M)},\ldots,\gamma_{(M-L+1)}}(x_{(M)}, \ldots, x_{(M-L+1)}) = p(\gamma_{(M)} = x_{(M)}, \ldots, \gamma_{(M-L+1)} = x_{(M-L+1)}) \tag{7.59}$$

can be shown to equal [1, Eqn. (9.419)]

$$p_{\gamma_{(M)},\ldots,\gamma_{(M-L+1)}}(x_{(M)}, \ldots, x_{(M-L+1)}) = \frac{M!}{(M-L)!}\left[P_\gamma(x_{(M-L+1)})\right]^{M-L}\prod_{i=1}^{L} p_\gamma(x_{(M-L+i)}), \tag{7.60}$$

where $p_\gamma(x) = p(\gamma = x)$ is the pdf of the branch SNR and $P_\gamma(x) = p(\gamma < x)$ is the cdf of the branch SNR. For example, in i.i.d. Rayleigh fading, $p_\gamma(\gamma)$ is exponential with mean $\overline{\gamma}$ and $P_\gamma(\gamma) = 1 - e^{-\gamma/\overline{\gamma}}$.

The MGF of γ_Σ given by (7.58) is

$$M_{\gamma_\Sigma}(s) = E_{\gamma_\Sigma}\left[e^{s\gamma_\Sigma}\right], \tag{7.61}$$

where $E_{\gamma_\Sigma}[\cdot]$ is the expectation with respect to the distribution of γ_Σ as given by (7.60). A closed-form expression for $M_{\gamma_\Sigma}(s)$ can then be obtained by writing γ_Σ in terms of the $L-1$ differences or *spaces* x_i between the L largest branch SNRs, given by $x_i = \gamma_{(M-i)} - \gamma_{(M-i-1)}, i = 0, \ldots, L-1$. For i.i.d. Rayleigh fading on each of the M total branches with branch SNR $\overline{\gamma}$ these spaces have been shown by Sukhatme [1, Theorem 9.1] to follow the exponential distribution

$$p_{x_i}(x) = \frac{1}{\overline{\gamma}}e^{-ix/\overline{\gamma}}. \tag{7.62}$$

Expressing the MGF (7.58) in terms of these x_is yields, after some manipulation, a closed-form result for the MGF of GC with MRC of the L highest-SNR branches given by

$$M_{\gamma_\Sigma}(s) = (1 - s\overline{\gamma})^{-L+1}\prod_{i=L}^{M}\left(1 - \frac{s\overline{\gamma}L}{i}\right)^{-1}. \tag{7.63}$$

Differentiating this expression yields the pdf of γ_Σ which can then be used to obtain P_{out} and $\overline{P}_s$. More details can be found in [1, Chapter 9.11.2.1].

7.4.3 Diversity Analysis for Noncoherent and Differentially Coherent Modulation

A similar MGF approach to determining the average symbol error probability of noncoherent and differentially coherent modulations with diversity combining is presented in [1, 15]. This approach differs from that of the coherent modulation case in that it relies on an alternate form of the Marcum Q-function instead of the Gaussian Q-function, since the BER of noncoherent and differentially coherent modulations in AWGN are given in terms of the Marcum Q-function. Otherwise the approach is essentially the same as in the coherent case, and it leads to BER expressions involving a single finite-range integral that can be readily evaluated numerically. More details on this approach are found in [1] and [15].

Chapter 7 Problems

1. Find the outage probability of QPSK modulation at $P_s = 10^{-3}$ for a Rayleigh fading channel with SC diversity for $M = 1$ (no diversity), $M = 2$, and $M = 3$. Assume branch SNRs of $\bar{\gamma}_1 = 10\,\text{dB}$, $\bar{\gamma}_2 = 15\,\text{dB}$, and $\bar{\gamma}_3 = 20\,\text{dB}$.

2. This problem explores diversity gain under both outage probability and average probability of error metrics, and shows they are the same for extreme SNR values. Consider an M-branch diversity system with independent Rayleigh fading on each path. The average SNR at each branch is given by $\overline{\gamma}$. Assume DPSK transmission.

   (a) Defining outage probability for $M > 1$ as the probability of the instantaneous SNR γ at all the receiver antennas being less than a constant c, compute $-\frac{\log(\texttt{outage probability})}{\log(\overline{\gamma})}$ as $\overline{\gamma} \to \infty$. Does it depend on c ?

   (b) Now consider an alternative definition of diversity gain based on average error probability as

   $$-\frac{\log(\texttt{average probability of error with MRC})}{\log(\overline{\gamma})}.$$

   For $M > 1$, evaluate this as $\overline{\gamma} \to \infty$.

   (c) Now evaluate the diversity gain according to both the definitions as $\overline{\gamma} \to 0$. Are these definitions of diversity gain more suitable at high or low average SNR to describe the gains from having $M > 1$ independent fading paths?

3. This problem explores the derivation of (7.11). Show that the average SNR of the combiner output under SC diversity in independent and identically distributed Rayleigh fading is

   $$\bar{\gamma}_\Sigma = \int_0^\infty \frac{\gamma M}{\bar{\gamma}}[1 - e^{-\gamma/\bar{\gamma}}]^{M-1} e^{-\gamma/\bar{\gamma}} d\gamma = \bar{\gamma} \sum_{i=1}^{M} \frac{1}{i}$$

   by expressing $[1 - e^{-\gamma/\bar{\gamma}}]^{M-1}$ using its binomial expansion, performing the integration in the resulting expression, and applying induction to this simplified expression to get the final result.

4. This problem uses order statistics as an alternate method to derive the result of (7.11). For a set of M identically and independently distributed random variables, $X_1, X_2, \ldots, X_M$, the order statistics of these variables are given by $X_{(1)}, X_{(2)}, \ldots, X_{(M)}$ where

   $$X_{(1)} = \min\left[X_1, X_2, \ldots, X_M\right] \qquad X_{(M)} = \max\left[X_1, X_2, \ldots, X_M\right].$$

   In general, $X_{(i)}$ denotes the ith largest value of $X_1, X_2, \ldots, X_n$, with $X_{(1)} \leq X_{(2))} \leq \ldots \leq X_{(M)}$. A well-known result in order statistics states that if $X_1, \ldots X_M$ are i.i.d random variables drawn from an exponential distribution with mean $\overline{X}$, then the order statistics of this collection of variables are given by

   $$X_{(j)} = \overline{X}\left(\sum_{j=1}^{i} \frac{Z_j}{n - j + 1}\right), \tag{7.64}$$

   where $Z_1, \ldots, Z_M$ are independent exponentially distributed random variables with mean 1.

   (a) For selection-combining (SC) of M i.i.d. branches under Rayleigh fading where each branch has the same average received power $\bar{\gamma}$, use the result of (7.64) to derive (7.11).

(b) Suppose that $\bar{\gamma} = 20$ dB. Use order statistics of i.i.d. exponentially-distributed random variables to find P_{out} for the cases listed below assuming BPSK modulation with $P_b = 10^{-4}$.

i. $M = 1$ (no diversity)

ii. $M = 3$. Also provide an expression for P_{out} in terms of the cdf of $X_{(3)}$.

(c) Suppose that instead of an SC diversity scheme, we have a single transmitter sending to three users, each with a single antenna. Assume BPSK modulation is sent over each user's single-antenna channel and these channels experience i.i.d. Rayleigh fading with SNR $\bar{\gamma} = 20$ dB. A user is in outage if $P_b > 10^{-4}$. Compute the probability associated with each of the following events and express these probabilities in terms of the cdfs of the variables $X_{(1)}$, $X_{(2)}$, and $X_{(3)}$:

i. No users are in outage.

ii. One or fewer users are in outage.

iii. Two or fewer users are in outage.

iv. All users are in outage.

5. Plot the distribution $p_{\gamma_\Sigma}(\gamma)$ given by (7.10) for the selection combiner SNR in Rayleigh fading with M-branch diversity assuming $M = 1, 2, 4, 8$, and 10. Assume that each branch has an average SNR of 10 dB. Your plot should be linear on both axes and should focus on the range of linear γ values $0 \leq \gamma \leq 60$. Discuss how the distribution changes with increasing M and why this leads to lower probability of error.

6. Derive the average probability of bit error for DPSK under SC with i.i.d. Rayleigh fading on each branch as given by (7.12).

7. Derive a general expression for the cdf of the two-branch SSC output SNR for branch statistics that are not i.i.d., and show that it reduces to (7.13) for i.i.d. branch statistics. Evaluate your expression assuming Rayleigh fading in each branch with different average SNRs $\bar{\gamma}_1$ and $\bar{\gamma}_2$.

8. Derive the average probability of bit error for DPSK under SSC with i.i.d. Rayleigh fading on each branch as given by (7.17).

9. Plot the average probability of bit error for DPSK under MRC with $M = 2, 3$, and 4, assuming i.i.d. Rayleigh fading on each branch and an average branch SNR ranging from 0 dB to 20 dB.

10. Show that the weights a_i maximizing γ_Σ under receiver diversity with MRC are $a_i^2 = r_i^2/N_0$ for $N_0/2$ the common noise PSD on each branch. Also show that, with these weights, $\gamma_\Sigma = \sum_i \gamma_i$.

11. This problem illustrates that, because of array gain, you can get performance gains from diversity combining even without fading. Consider an AWGN channel with N-branch diversity combining and $\gamma_i = 10\,\text{dB}$ per branch. Assume MQAM modulation with $M = 4$ and use the approximation $P_b = .2e^{-1.5\gamma/(M-1)}$ for bit error probability, where γ is the received SNR.

(a) Find P_b for $N = 1$.

(b) Find N so that, under MRC, $P_b < 10^{-6}$.

12. Derive the average probability of bit error for BPSK under MRC with i.i.d. Rayleigh fading on each branch as given by (7.27).

13. Derive the average probability of bit error for BPSK under EGC with i.i.d. Rayleigh fading on each branch as given by (7.22).

14. Compare the average probability of bit error for BPSK modulation under no diversity, two-branch SC, two-branch SSC with $\gamma_T = \gamma_0$, two-branch EGC, and two-branch MRC. Assume i.i.d. Rayleigh fading on each branch with equal branch SNRs of 10 dB and of 20 dB. How does the relative performance change as the branch SNRs increase?

15. Plot the average probability of bit error for BPSK under both MRC and EGC assuming two-branch diversity with i.i.d. Rayleigh fading on each branch and average branch SNR ranging from 0 dB to 20 dB. What is the maximum dB penalty of EGC as compared to MRC?

16. Compare the outage probability of BPSK modulation at $P_b = 10^{-3}$ under MRC versus EGC, assuming two-branch diversity with i.i.d. Rayleigh fading on each branch and average branch SNR $\bar{\gamma} = 10\,\mathrm{dB}$.

17. Compare the average probability of bit error for BPSK under MRC versus EGC, assuming two-branch diversity with i.i.d. Rayleigh fading on each branch and average branch SNR $\bar{\gamma} = 10\,\mathrm{dB}$.

18. Consider a GC diversity system with M receiver branches and L receivers, whereby the combiner chooses the L highest SNR signals from the M branches and then combines them using MRC. Assume the SNR on the ith branch is γ_i which is i.i.d. distributed uniformly between 0 and 1 in linear units. Denote the SNR for the combiner output as γ_Σ; the outage probability for threshold γ_0 is $P_{out} = p(\gamma_\Sigma < \gamma_0)$. Assume that there are $M = 3$ branches and the threshold is $\gamma_0 = 1$.

    (a) Suppose $L = 1$, state the form of diversity combining for this special case and compute P_{out}.

    (b) Repeat part (a) if $L = 3$.

    (c) Suppose $L = 2$. Show that $p(\gamma_{\max} < \gamma_0/2) \leq P_{out} = p(\gamma_\Sigma < \gamma_0) \leq p(\gamma_{\min} < \gamma_0/2)$ for $\gamma_{\min} = \min_i \gamma_i, \gamma_{\max} = \max_i \gamma_i$. Compute upper and lower bounds for P_{out}.

19. Compute the average BER of a channel with two-branch transmit diversity under the Alamouti scheme, assuming the average branch SNR is 10 dB.

20. Consider a fading distribution $p(\gamma)$ where $\int_0^\infty p(\gamma)e^{-x\gamma}d\gamma = .01\bar{\gamma}/\sqrt{x}$. Find the average P_b for a BPSK modulated signal where (a) the receiver has two-branch diversity with MRC combining and (b) each branch has an average SNR of 10 dB and experiences independent fading with distribution $p(\gamma)$.

21. Consider a fading channel that is BPSK modulated and has three-branch diversity with MRC, where each branch experiences independent fading with an average received SNR of 15 dB. Compute the average BER of this channel for Rayleigh fading and for Nakagami fading with $m = 2$. *Hint*: Using the alternate Q-function representation greatly simplifies this computation, at least for Nakagami fading.

22. Plot the average probability of error as a function of branch SNR for a two-branch MRC system with BPSK modulation, where the first branch has Rayleigh fading and the second branch has Nakagami-m fading with $m = 2$. Assume the two branches have the same average SNR; your plots should have this average branch SNR ranging from 5 dB to 20 dB.

23. Plot the average probability of error as a function of branch SNR for an M-branch MRC system with 8-PSK modulation for $M = 1, 2, 4, 8$. Assume that each branch has Rayleigh fading with the same average SNR. Your plots should have this SNR ranging from 5 dB to 20 dB.

24. Derive the average probability of symbol error for MQAM modulation under MRC diversity given by (7.57) from the probability of error in AWGN (6.79) by utilizing the alternate representation of Q and Q^2.

25. Compare the average probability of symbol error for 16-PSK and 16-QAM modulation, assuming three-branch MRC diversity with Rayleigh fading on the first branch and Rician fading on the second and third branches with $K = 2$. Assume equal average branch SNRs of 10 dB.

26. Plot the average probability of error as a function of branch SNR for an M-branch MRC system with 16-QAM modulation for $M = 1, 2, 4, 8$. Assume that each branch has Rayleigh fading with the same average SNR. Your plots should have an SNR ranging from 5 dB to 20 dB.

Bibliography

[1] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*, Wiley, New York, 2nd Ed., 2004.

[2] *Wireless Communications* A. F. Molisch, John Wiley and Sons, 2nd ed. 2010.

[3] J. Winters, "Signal acquisition and tracking with adaptive arrays in the digital mobile radio system IS-54 with flat fading," *IEEE Trans. Veh. Tech*., pp. 1740–51, November 1993.

[4] G. L. Stuber, *Principles of Mobile Communications*, 2nd ed., Kluwer, Dordrecht, 2001.

[5] M. Blanco and K. Zdunek, "Performance and optimization of switched diversity systems for the detection of signals with Rayleigh fading," *IEEE Trans. Commun*., pp. 1887–95, December 1979.

[6] A. Abu-Dayya and N. Beaulieu, "Switched diversity on microcellular Ricean channels," *IEEE Trans. Veh. Tech*., pp. 970–6, November 1994.

[7] A. Abu-Dayya and N. Beaulieu, "Analysis of switched diversity systems on generalized-fading channels," *IEEE Trans. Commun*., pp. 2959–66, November 1994.

[8] M. Yacoub, *Principles of Mobile Radio Engineering*, CRC Press, Boca Raton, FL, 1993.

[9] T. Eng, N. Kong and L. B. Milstein, "Comparison of diversity combining techniques for Rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 44, no. 9, pp. 1117-1129, Sept. 1996.

[10] GC2 M.-S. Alouini and M. K. Simon, '1'An MGF-based performance analysis of generalized selection combining over Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 401-415, March 2000.

[11] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun*., pp. 1451–8, October 1998.

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

[13] E. G. Larsson and P. Stoica, *Space-Time Block Coding for Wireless Communications*, Cambridge University Press, 2003.

[14] M. K. Simon and M.-S. Alouini, "A unified approach to the performance analysis of digital communications over generalized fading channels," *Proc. IEEE*, pp. 1860–77, September 1998.

[15] M. K. Simon and M.-S. Alouini, "A unified approach for the probability of error for noncoherent and differentially coherent modulations over generalized fading channels," *IEEE Trans. Commun*., pp. 1625–38, December 1998.

Chapter 8

Coding for Wireless Channels

This chapter has not yet been updated for the 2nd edition

Coding allows bit errors introduced by transmission of a modulated signal through a wireless channel to be either detected or corrected by a decoder in the receiver. Coding can be considered as the embedding of signal constellation points in a higher-dimensional signaling space than is needed for communications. By going to a higher-dimensional space, the distance between points can be increased, which provides for better error correction and detection.

In this chapter we describe codes designed for additive white Gaussian noise channels and for fading channels. Codes designed for AWGN channels typically do not work well on fading channels because they cannot correct for long error bursts that occur in deep fading. Codes for fading channels are mainly based on using an AWGN channel code combined with interleaving, but the criterion for the code design changes to provide fading diversity. Other coding techniques to combat performance degradation due to fading include unequal error protection codes and joint source and channel coding.

We first provide an overview of code design in both fading and AWGN, along with basic design parameters such as minimum distance, coding gain, bandwidth expansion, and diversity order. Sections 8.2 and 8.3 provide a basic overview of block and convolutional code designs for AWGN channels. Although these designs are not directly applicable to fading channels, codes for fading channels and other codes used in wireless systems (e.g., spreading codes in CDMA) require background in these fundamental techniques. Concatenated codes and their evolution to turbo codes – as well as low-density parity-check (LDPC) codes – for AWGN channels are also described. These extremely powerful codes exhibit near-capacity performance with reasonable complexity levels. Coded modulation was invented in the late 1970s as a technique to obtain error correction through a joint design of the modulation and coding. We will discuss the basic design principles behind trellis and more general lattice coded modulation along with their performance in AWGN.

Code designs for fading channels are covered in Section 8.8. These designs combine block or convolutional codes designed for AWGN channels with interleaving and then modify the AWGN code design metric to incorporate maximum fading diversity. Diversity gains can also be obtained by combining coded modulation with symbol or bit interleaving, although bit interleaving generally provides much higher diversity gain. Thus, coding combined with interleaving provides diversity gain in the same manner as other forms of diversity, with the diversity order built into the code design. Unequal error protection is an alternative to diversity in fading mitigation. In these codes bits are prioritized, and high-priority bits are encoded with stronger error protection against deep fades. Since bit priorities are part of the source code design, unequal error protection is a special case of joint source and channel coding, which we also describe.

Coding is a very broad and deep subject, with many excellent books devoted solely to this topic. This chapter assumes no background in coding, and thus it provides an in-depth discussion of code designs for AWGN channels

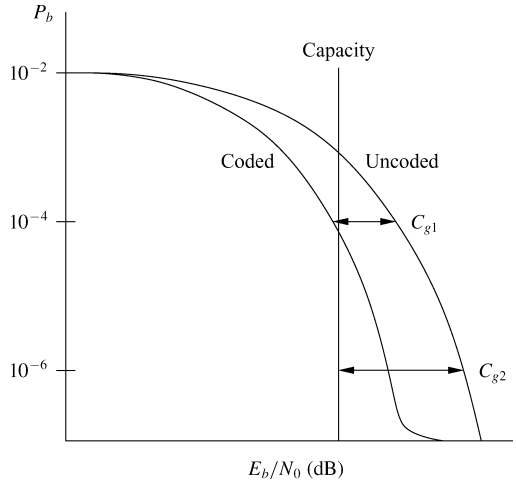

Figure 8.1: Coding gain in AWGN channels.

before designs for wireless systems can be treated. This in-depth discussion can be omitted for a more cursory treatment of coding for wireless channels by focusing on Sections 8.1 and 8.8.

8.1 Overview of Code Design

The main reason to apply error correction coding in a wireless system is to reduce the probability of bit or block error. The bit error probability P_b for a coded system is the probability that a bit is decoded in error. The block error probability P_{bl}, also called the packet error rate, is the probability that one or more bits in a block of coded bits is decoded in error. Block error probability is useful for packet data systems where bits are encoded and transmitted in blocks. The amount of error reduction provided by a given code is typically characterized by its coding gain in AWGN and its diversity gain in fading.

Coding gain in AWGN is defined as the amount that the bit energy or signal-to-noise power ratio can be reduced under the coding technique for a given P_b or P_{bl}. We illustrate coding gain for P_b in Figure 8.1. We see in this figure that the gain C_{g1} at $P_b = 10^{-4}$ is less than the gain C_{g2} at $P_b = 10^{-6}$, and there is negligible coding gain at $P_b = 10^{-2}$. In fact, codes designed for high-SNR channels can have negative coding gain at low SNRs, since the extra redundancy of the code does not provide sufficient performance gain in P_b or P_{bl} at low SNRs to compensate for spreading the bit energy over multiple coded bits. Thus, unexpected fluctuations in channel SNR can significantly degrade code performance. The coding gain in AWGN is generally a function of the minimum Euclidean distance of the code, which equals the minimum distance in signal space between codewords or error events. Thus, codes designed for AWGN channels maximize their Euclidean distance for good performance.

Error probability with or without coding tends to fall off with SNR as a waterfall shape at low to moderate SNRs. Whereas this waterfall shape holds at all SNRs for uncoded systems and many coded systems, some codes (such as turbo codes) exhibit error floors as SNR grows. The error floor, also shown in Figure 8.1, kicks in at a threshold SNR that depends on the code design. For SNRs above this threshold, the slope of the error probability curve decreases because minimum distance error events dominate code performance in this SNR regime.

Code performance is also commonly measured against channel capacity. The capacity curve is associated with the SNR (E_b/N_0) where Shannon capacity $B\log_2(1 + \text{SNR})$ equals the data rate of the system. At rates up to capacity, the capacity-achieving code has a probability of error that goes to zero, as indicated by the straight line in Figure 8.1. The capacity curve thus indicates the best possible performance that any practical code can achieve.

For many codes, the error correction capability of a code does not come for free. This performance enhancement is paid for by increased complexity and – for block codes, convolutional codes, turbo codes, and LDPC codes – by either a decrease in data rate or an increase in signal bandwidth. Consider a code with n coded bits for every k uncoded bits. This code effectively embeds a k-dimensional subspace into a larger n-dimensional space to provide larger distances between coded symbols. However, if the data rate through the channel is fixed at R_b, then the information rate for a code that uses n coded bits for every k uncoded bits is $(k/n)R_b$; that is, coding decreases the data rate by the fraction k/n. We can keep the information rate constant and introduce coding gain by decreasing the bit time by k/n. This typically results in an expanded bandwidth of the transmittted signal by n/k. Coded modulation uses ajoint design of the code and modulation to obtain coding gain without this bandwidth expansion, as discussed in more detail in Section 8.7.

Codes designed for AWGN channels do not generally work well in fading owing to bursts of errors that cannot be corrected for. However, good performance in fading can be obtained by combining AWGN channel codes with interleaving and by designing the code to optimize its inherent diversity. The interleaver spreads out bursts of errors over time, so it provides a form of time diversity. This diversity is exploited by the inherent diversity in the code. In fact, codes designed in this manner exhibit performance similar to MRC diversity, with diversity order equal to the minimum Hamming distance of the code. Hamming distance is the number of coded symbols that differ between different codewords or error events. Thus, coding and interleaving designed for fading channels maximize their Hamming distance for good performance.

8.2 Linear Block Codes

Linear block codes are conceptually simple codes that are basically an extension of single-bit parity-check codes for error detection. A single-bit parity-check code is one of the most common forms of detecting transmission errors. This code uses one extra bit in a block of n data bits to indicate whether the number of l-bits in a block is odd or even. Thus, if a single error occurs, either the parity bit is corrupted or the number of detected 1-bits in the information bit sequence will be different from the number used to compute the parity bit; in either case the parity bit will not correspond to the number of detected 1-bits in the information bit sequence, so the single error is detected. Linear block codes extend this notion by using a larger number of parity bits to either detect more than one error or correct for one or more errors. Unfortunately, linear block codes – along with convolutional codes – trade their error detection or correction capability for either bandwidth expansion or a lower data rate, as we shall discuss in more detail. We will restrict our attention to binary codes, where both the original information and the corresponding code consist of bits taking a value of either 0 or 1.

8.2.1 Binary Linear Block Codes

A binary block code generates a block of n coded bits from k information bits. We call this an (n, k) binary block code. The coded bits are also called *codeword symbols*. The n codeword symbols can take on 2^n possible values corresponding to all possible combinations of the n binary bits. We select 2^k codewords from these 2^n possibilities to form the code, where each k bit information block is uniquely mapped to one of these 2^k codewords. The rate of the code is $R_c = k/n$ information bits per codeword symbol. If we assume that codeword symbols are transmitted across the channel at a rate of R_s symbols per second, then the information rate associated with an (n, k) block code is $R_b = R_c R_s = (k/n)R_s$ bits per second. Thus we see that block coding reduces the data rate compared to what we obtain with uncoded modulation by the code rate R_c.

A block code is called a *linear* code when the mapping of the k information bits to the n codeword symbols is a linear mapping. In order to describe this mapping and the corresponding encoding and decoding functions in more detail, we must first discuss properties of the vector space of binary n-tuples and its corresponding subspaces.

The set of all binary n-tuples B_n is a vector space over the binary field, which consists of the two elements 0 and 1. All fields have two operations, addition and multiplication: for the binary field these operations correspond to binary addition (modulo 2 addition) and standard multiplication. A subset S of B_n is called a *subspace* if it satisfies the following conditions.

1. The all-zero vector is in S.
2. The set S is closed under addition; that is, if $S_i \in S$ and $S_j \in S$ then $S_i + S_j \in S$.

An (n, k) block code is linear if the 2^k length-n codewords of the code form a subspace of B_n. Thus, if $\mathbf{C}_i$ and $\mathbf{C}_j$ are two codewords in an (n, k) linear block code, then $\mathbf{C}_i + \mathbf{C}_j$ must form another codeword of the code.

Example 8.1:

The vector space B_3 consists of all binary tuples of length 3:

$$B_3 = \{[000], [001], [010], [011], [100], [101], [110], [111]\}.$$

Note that B_3 is a subspace of itself, since it contains the all-zero vector and is closed under addition. Determine which of the following subsets of B_3 form a subspace:

- $A_1 = \{[000], [001], [100], [101]\}$;
- $A_2 = \{[000], [100], [110], [111]\}$;
- $A_3 = \{[001], [100], [101]\}$.

Solution: It is easily verified that A_1 is a subspace, since it contains the all-zero vector and the sum of any two tuples in A_1 is also in A_1. A_2 is not a subspace because it is not closed under addition, as $110 + 111 = 001 \notin A_2$. A_3 is not a subspace because it is not closed under addition $(001 + 001 = 000 \notin A_3)$ and it does not contain the all-zero vector.

Intuitively, the greater the distance between codewords in a given code, the less chance that errors introduced by the channel will cause a transmitted codeword to be decoded as a different codeword. We define the *Hamming distance* between two codewords $\mathbf{C}_i$ and $\mathbf{C}_j$, denoted as $d(\mathbf{C}_i, \mathbf{C}_j)$ or d_{ij}, as the number of elements in which they differ:

$$d_{ij} = \sum_{l=1}^{n} (\mathbf{C}_i(l) + \mathbf{C}_j(l)), \tag{8.1}$$

where $\mathbf{C}_m(l)$ denotes the lth bit in $\mathbf{C}_m$. For example, if $\mathbf{C}_i = [00101]$ and $\mathbf{C}_j = [10011]$ then $d_{ij} = 3$. We define the *weight* of a given codeword $\mathbf{C}_i$ as the number of 1-bits in the codeword, so $\mathbf{C}_i = [00101]$ has weight 2. The weight of a given codeword $\mathbf{C}_i$ is just its Hamming distance d_{0i} from the all-zero codeword $\mathbf{C}_0 = [00\ldots0]$ or, equivalently, the sum of its elements:

$$w(\mathbf{C}_i) = \sum_{l=1}^{n} \mathbf{C}_i(l). \tag{8.2}$$

Since $0+0=1+1=0$, the Hamming distance between $\mathbf{C}_i$ and $\mathbf{C}_j$ is equal to the weight of $\mathbf{C}_i+\mathbf{C}_j$. For example, with $\mathbf{C}_i=[00101]$ and $\mathbf{C}_j=[10011]$ as given before, $w(\mathbf{C}_i)=2$, $w(\mathbf{C}_j)=3$, and $d_{ij}=w(\mathbf{C}_i+\mathbf{C}_j)=w([10110])=3$. Since the Hamming distance between any two codewords equals the weight of their sum and since this sum is also a codeword, we can determine the minimum distance between all codewords in a code by just looking at the minimum distance between all nonzero codewords and the all-zero codeword $\mathbf{C}_0$. Thus, we define the minimum distance of a code as

$$d_{\min}=\min_{i,i\neq 0} d_{0i}. \tag{8.3}$$

We will see in Section 8.2.6 that the minimum distance of a linear block code is a critical parameter in determining its probability of error.

8.2.2 Generator Matrix

The generator matrix is a compact description of how codewords are generated from information bits in alinear block code. The design goal in linear block codes is to find generator matrices such that their corresponding codes are easy to encode and decode yet have powerful error correction/detection capabilities. Consider an (n,k) code with k information bits, denoted as

$$\mathbf{U}_i=[u_{i1},\ldots,u_{ik}],$$

that are encoded into the codeword

$$\mathbf{C}_i=[c_{i1},\ldots,c_{in}].$$

We represent the encoding operation as a set of n equations defined by

$$c_{ij}=u_{i1}g_{1j}+u_{i2}g_{2j}+\cdots+u_{ik}g_{kj},\quad j=1,\ldots,n, \tag{8.4}$$

where g_{ij} is binary (0 or 1) and where binary (standard) multiplication is used. We can write these n equations in matrix form as

$$\mathbf{C}_i=\mathbf{U}_i\mathbf{G}, \tag{8.5}$$

where the $k\times n$ *generator matrix* $\mathbf{G}$ for the code is defined as

$$\mathbf{G}=\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ g_{k1} & g_{k2} & \cdots & g_{kn} \end{bmatrix}. \tag{8.6}$$

If we denote the lth row of $\mathbf{G}$ as $\mathbf{g}_l=[g_{l1},\ldots,g_{ln}]$, then we can write any codeword $\mathbf{C}_i$ as linear combinations of these row vectors as follows:

$$\mathbf{C}_i=u_{i1}\mathbf{g}_1+u_{i2}\mathbf{g}_2+\cdots+u_{ik}\mathbf{g}_k. \tag{8.7}$$

Since a linear (n,k) block code is a subspace of dimension k in the larger n-dimensional space, it follows that the k row vectors $\{\mathbf{g}_l\}_{l=1}^{k}$ of $\mathbf{G}$ must be linearly independent so that they span the k-dimensional subspace associated

with the 2^k codewords. Hence, $\mathbf{G}$ has rank k. Since the set of basis vectors for this subspace is not unique, the generator matrix is also not unique.

A *systematic* linear block code is described by a generator matrix of the form

$$\mathbf{G} = [\mathbf{I}_k|\mathbf{P}] = \left[\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & p_{11} & p_{12} & \cdots & p_{1(n-k)} \\ 0 & 1 & \cdots & 0 & p_{21} & p_{22} & \cdots & p_{2(n-k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & p_{k1} & p_{k2} & \cdots & p_{k(n-k)} \end{array}\right], \tag{8.8}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{P}$ is a $k \times (n-k)$ matrix that determines the redundant, or parity, bits to be used for error correction or detection. The codeword output from a systematic encoder is of the form

$$\mathbf{C}_i = \mathbf{U}_i\mathbf{G} = \mathbf{U}_i[\mathbf{I}_k|\mathbf{P}] = [u_{i1}, \ldots, u_{ik}, p_1, \ldots, p_{(n-k)}], \tag{8.9}$$

where the first k bits of the codeword are the original information bits and the last $(n-k)$ bits of the codeword are the parity bits obtained from the information bits as

$$p_j = u_{i1}p_{1j} + \cdots + u_{ik}p_{kj}, \quad j = 1, \ldots, n-k. \tag{8.10}$$

Note that any generator matrix for an (n, k) linear block code can be reduced by row operations and column permutations to a generator matrix in systematic form.

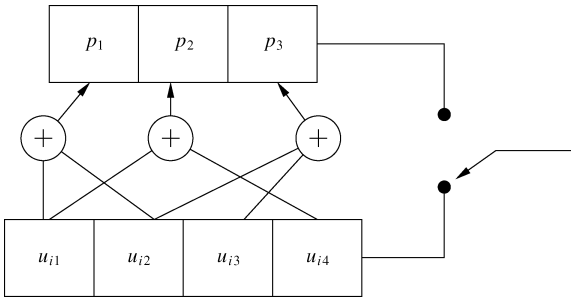

Figure 8.2: Implementation of (7,4) binary code.

Example 8.2: Systematic linear block codes are typically implemented with $n-k$ modulo-2 adders tied to the appropriate stages of a shift register. The resulting parity bits are appended to the end of the information bits to form the codeword. Find the corresponding implementation for generating a (7,4) binary code with the generator matrix

$$\mathbf{G} = \left[\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right]. \tag{8.11}$$

Solution: The matrix $\mathbf{G}$ is already in systematic form with

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{8.12}$$

Let P_{lj} denote the ljth element of $\mathbf{P}$. By (8.10) we see that the first parity bit in the codeword is $p_1 = u_{i1}P_{11} + u_{i2}P_{21} + u_{i3}P_{31} + u_{i4}P_{41} = u_{i1} + u_{i2}$. Similarly, the second parity bit is $p_2 = u_{i1}P_{12} + u_{i2}P_{22} + u_{i3}P_{32} + u_{i4}P_{42} = u_{i1} + u_{i4}$ and the third parity bit is $p_3 = u_{i1}P_{13} + u_{i2}P_{23} + u_{i3}P_{33} + u_{i4}P_{43} = u_{i2} + u_{i3}$. The shift register implementation to generate these parity bits is shown in Figure 8.2. The codeword output is $[u_{i1}u_{i2}u_{i3}u_{i4}p_1p_2p_3]$, where the switch is in the down position to output the systematic bits $u_{ij}(j = 1, \ldots, 4)$ of the code or in the up position to output the parity bits $p_j(j = 1, 2, 3)$ of the code.

8.2.3 Parity-Check Matrix and Syndrome Testing

The parity-check matrix is used to decode linear block codes with generator matrix $\mathbf{G}$. The parity-check matrix $\mathbf{H}$ corresponding to a generator matrix $\mathbf{G} = [\mathbf{I}_k|\mathbf{P}]$ is defined as

$$\mathbf{H} = [\mathbf{P}^T|\mathbf{I}_{n-k}]. \tag{8.13}$$

It is easily verified that $\mathbf{G}\mathbf{H}^T = \mathbf{0}_{k,n-k}$, where $\mathbf{0}_{k,n-k}$ denotes an all-zero $k \times (n-k)$ matrix. Recall that a given codeword $\mathbf{C}_i$ in the code is obtained by multiplication of the information bit sequence $\mathbf{U}_i$ by the generator matrix $\mathbf{G}$: $\mathbf{C}_i = \mathbf{U}_i\mathbf{G}$. Thus,

$$\mathbf{C}_i\mathbf{H}^T = \mathbf{U}_i\mathbf{G}\mathbf{H}^T = \mathbf{0}_{n-k} \tag{8.14}$$

for any input sequence $\mathbf{U}_i$, where $\mathbf{0}_{n-k}$ denotes the all-zero row vector of length $n-k$. Thus, multiplication of any valid codeword with the parity-check matrix results in an all-zero vector. This property is used to determine whether the received vector is a valid codeword or has been corrupted, based on the notion of *syndrome testing*, which we now define.

Let $\mathbf{R}$ be the received codeword resulting from transmission of codeword $\mathbf{C}$. In the absence of channel errors, $\mathbf{R} = \mathbf{C}$. However, if the transmission is corrupted, then one or more of the codeword symbols in $\mathbf{R}$ will differ from those in $\mathbf{C}$. We therefore write the received codeword as

$$\mathbf{R} = \mathbf{C} + \mathbf{e}, \tag{8.15}$$

where $\mathbf{e} = [e_1, e_2, \ldots, e_n]$ is the *error pattern* indicating which codeword symbols were corrupted by the channel. We define the *syndrome* of $\mathbf{R}$ as

$$\mathbf{S} = \mathbf{R}\mathbf{H}^T. \tag{8.16}$$

If $\mathbf{R}$ is a valid codeword (i.e., $\mathbf{R} = \mathbf{C}_i$ for some i) then $\mathbf{S} = \mathbf{C}_i\mathbf{H}^T = \mathbf{0}_{n-k}$ by (8.14). Thus, the syndrome equals the all-zero vector if the transmitted codeword is not corrupted – or is corrupted in a manner such that the received codeword is a valid codeword in the code but is different from the transmitted codeword. If the received codeword $\mathbf{R}$ contains detectable errors, then $\mathbf{S} \neq \mathbf{0}_{n-k}$. If the received codeword contains correctable errors, the syndrome identifies the error pattern corrupting the transmitted codeword, and these errors can then be corrected. Note that the syndrome is a function only of the error pattern $\mathbf{e}$ and not the transmitted codeword $\mathbf{C}$, since

$$\mathbf{S} = \mathbf{R}\mathbf{H}^T = (\mathbf{C} + \mathbf{e})\mathbf{H}^T = \mathbf{C}\mathbf{H}^T + \mathbf{e}\mathbf{H}^T = \mathbf{0}_{n-k} + \mathbf{e}\mathbf{H}^T. \tag{8.17}$$

Because $\mathbf{S} = \mathbf{e}\mathbf{H}^T$ corresponds to $n - k$ equations in n unknowns, there are 2^k possible error patterns that can produce a given syndrome $\mathbf{S}$. However, since the probability of bit error is typically small and independent for each bit, the most likely error pattern is the one with minimal weight, corresponding to the least number of errors introduced in the channel. Thus, if an error pattern $\hat{\mathbf{e}}$ is the most likely error associated with a given syndrome $\mathbf{S}$, the transmitted codeword is typically decoded as

$$\hat{\mathbf{C}} = \mathbf{R} + \hat{\mathbf{e}} = \mathbf{C} + \mathbf{e} + \hat{\mathbf{e}}. \tag{8.18}$$

If the most likely error pattern does occur, then $\hat{\mathbf{e}} = \mathbf{e}$ and $\hat{\mathbf{C}} = \mathbf{C}$ – that is, the corrupted codeword is correctly decoded. The decoding process and associated error probability will be covered in Section 8.2.6

Let $\mathbf{C}_w$ denote a codeword in a given (n, k) code with minimum weight (excluding the all-zero codeword). Then $\mathbf{C}_w\mathbf{H}^T = \mathbf{0}_{n-k}$ is just the sum of $d_{\min}$ columns of $\mathbf{H}^T$, since $d_{\min}$ equals the number of l-bits (the weight) in the minimum weight codeword of the code. Since the rank of $\mathbf{H}^T$ is at most $n - k$, this implies that the minimum distance of an (n, k) block code is upper bounded by

$$d_{\min} \leq n - k + 1, \tag{8.19}$$

which is referred to as the *Singelton bound*.

8.2.4 Cyclic Codes

Cyclic codes are a subclass of linear block codes in which all codewords in a given code are cyclic shifts of one another. Specifically, if the codeword $\mathbf{C} = [c_0, c_1, \ldots, c_{n-1}]$ is a codeword in a given code, then a cyclic shift by 1, denoted as $\mathbf{C}^{(1)} = [c_{n-1}, c_0, \ldots, c_{n-2}]$, is also a codeword. More generally, any cyclic shift $C^{(i)} = [c_{n-i}, c_{n-i+1}, \ldots, c_{n-i-1}]$ is also a codeword. The cyclic nature of cyclic codes creates a nice structure that allows their encoding and decoding functions to be of much lower complexity than the matrix multiplications associated with encoding and decoding for general linear block codes. Thus, most linear block codes used in practice are cyclic codes.

Cyclic codes are generated via a *generator polynomial* instead of a generator matrix. The generator polynomial $g(X)$ for an (n, k) cyclic code has degree $n - k$ and is of the form

$$g(X) = g_0 + g_1 X + \cdots + g_{n-k} X^{n-k}, \tag{8.20}$$

where g_i is binary (0 or 1) and $g_0 = g_{n-k} = 1$. The k-bit information sequence $[u_0, \ldots, u_{k-1}]$ is also written in polynomial form as the *message polynomial*

$$u(X) = u_0 + u_1 X + \cdots + u_{k-1} X^{k-1}. \tag{8.21}$$

The codeword associated with a given k-bit information sequence is obtained from the polynomial coefficients of the generator polynomial multiplied by the message polynomial; thus, the codeword $C = [c_0, \ldots, c_{n-1}]$ is obtained from

$$c(X) = u(X)g(X) = c_0 + c_1 X + \cdots + c_{n-1} X^{n-1}. \tag{8.22}$$

A codeword described by a polynomial $c(X)$ is a valid codeword for a cyclic code with generator polynomial $g(X)$ if and only if $g(X)$ divides $c(X)$ with no remainder (no remainder polynomial terms) – that is, if and only if

$$\frac{c(X)}{g(X)} = q(X) \tag{8.23}$$

for a polynomial $q(X)$ of degree less than k.

Example 8.3: Consider a (7, 4) cyclic code with generator polynomial $g(X) = 1 + X^2 + X^3$. Determine if the codewords described by polynomials $c_1(X) = 1 + X^2 + X^5 + X^6$ and $c_2(X) = 1 + X^2 + X^3 + X^5 + X^6$ are valid codewords for this generator polynomial.

Solution: Division of binary polynomials is similar to division of standard polynomials except that, under binary addition, subtraction is the same as addition. Dividing $c_1(X) = 1 + X^2 + X^5 + X^6$ by $g(X) = 1 + X^2 + X^3$, we have:

$$\begin{array}{r} X^3 + 1 \\ X^3 + X^2 + 1 \;\overline{)\,X^6 + X^5 + \;X^2 + 1} \\ \underline{X^6 + X^5 + \;X^3} \\ X^3 + X^2 + 1 \\ \underline{X^3 + X^2 + 1} \\ 0. \end{array}$$

Since $g(X)$ divides $c(X)$ with no remainder, it is a valid codeword. In fact, we have $c_1(X) = (1 + X^3)g(X) = u(X)g(X)$ and so the information bit sequence corresponding to $c_1(X)$ is $\mathbf{U} = [1001]$, corresponding to the coefficients of the message polynomial $u(X) = 1 + X^3$.

Dividing $c_2(X) = 1 + X^2 + X^3 + X^5 + X^6$ by $g(X) = 1 + X^2 + X^3$ yields

$$\begin{array}{r} X^3 + 1 \\ X^3 + X^2 + 1 \;\overline{)\,X^6 + X^5 + X^3 + X^2 + 1} \\ \underline{X^6 + X^5 + X^3} \\ X^2 + 1, \end{array}$$

where we note that there is a remainder of $X^2 + 1$ in the division. Thus, $c_2(X)$ is not a valid codeword for the code corresponding to this generator polynomial.

Recall that systematic linear block codes have the first k codeword symbols equal to the information bits and the remaining codeword symbols equal to the parity bits. A cyclic code can be put in systematic form by first multiplying the message polynomial $u(X)$ by X^{n-k}, yielding

$$X^{n-k}u(X) = u_0X^{n-k} + u_1X^{n-k+1} + \cdots + u_{k-1}X^{n-1}. \tag{8.24}$$

This shifts the message bits to the k rightmost digits of the codeword polynomial. If we next divide (8.24) by $g(X)$, we obtain

$$\frac{X^{n-k}u(X)}{g(X)} = q(X) + \frac{p(X)}{g(X)}, \tag{8.25}$$

where $q(X)$ is a polynomial of degree at most $k-1$ and $p(X)$ is a remainder polynomial of degree at most $n-k-1$. Multiplying (8.25) through by $g(X)$, we now have

$$X^{n-k}u(X) = q(X)g(X) + p(X). \tag{8.26}$$

Adding $p(X)$ to both sides yields

$$p(X) + X^{n-k}u(X) = q(X)g(X). \tag{8.27}$$

This implies that $p(X) + X^{n-k}u(X)$ is a valid codeword, since it is divisible by $g(X)$ with no remainder. The codeword is described by the n coefficients of the codeword polynomial $p(X) + X^{n-k}u(X)$. Note that we can express $p(X)$ (of degree $n-k-1$) as

$$p(X) = p_0 + p_1X + \cdots + p_{n-k-1}X^{n-k-1}. \tag{8.28}$$

Combining (8.24) and (8.28), we get

$$\begin{aligned} &p(X) + X^{n-k}u(X) \\ &\quad = p_0 + p_1X + \cdots + p_{n-k-1}X^{n-k-1} + u_0X^{n-k} + u_1X^{n-k+1} + \cdots + u_{k-1}X^{n-1}. \end{aligned} \tag{8.29}$$

Thus, the codeword corresponding to this polynomial has the first k bits consisting of the message bits $[u_0, \ldots, u_k]$ and the last $n-k$ bits consisting of the parity bits $[p_0, \ldots, p_{n-k-1}]$, as is required for the systematic form.

Note that the systematic codeword polynomial is generated in three steps: first multiplying the message polynomial $u(X)$ by X^{n-k}; then dividing $X^{n-k}u(X)$ by $g(X)$ to obtain the remainder polynomial $p(X)$ (along with the quotient polynomial $q(X)$, which is not used); and finally adding $p(X)$ to $X^{n-k}u(X)$ to get (8.29). The polynomial multiplications are straightforward to implement, and the polynomial division is easily implemented with a feedback shift register [1, Chap. 8.1; 2, Chap. 6.7]. Thus, codeword generation for systematic cyclic codes has very low cost and low complexity.

Let us now consider how to characterize channel errors for cyclic codes. The codeword polynomial corresponding to a transmitted codeword is of the form

$$c(X) = u(X)g(X). \tag{8.30}$$

The received codeword can also be written in polynomial form as

$$r(X) = c(X) + e(X) = u(X)g(X) + e(X), \tag{8.31}$$

where $e(X)$ is the error polynomial with coefficients equal to 1 where errors occur. For example, if the transmitted codeword is $\mathbf{C} = [1011001]$ and the received codeword is $\mathbf{R} = [1111000]$, then $e(X) = X + X^{n-1}$. The *syndrome polynomial* $s(X)$ for the received codeword is defined as the remainder when $r(X)$ is divided by $g(X)$, so $s(X)$ has degree $n-k-1$. But by (8.31), the syndrome polynomial $s(X)$ is equivalent to the error polynomial $e(X)$ modulo $g(X)$. Moreover, we obtain the syndrome through a division circuit similar to the one used for generating the code. As stated previously, this division circuit is typically implemented using a feedback shift register, resulting in a low-cost implementation of low complexity.

8.2.5 Hard Decision Decoding (HDD)

The probability of error for linear block codes depends on whether the decoder uses soft decisions or hard decisions. In hard decision decoding (HDD), each coded bit is demodulated as a 0 or 1 – that is, the demodulator detects each coded bit (symbol) individually. For example, in BPSK, the received symbol is decoded as a 1 if it is closer to $\sqrt{E_b}$ and as a 0 if it is closer to $-\sqrt{E_b}$. This form of demodulation removes information that can be used by the channel decoder. In particular, for the BPSK example the distance of the received bit from $\sqrt{E_b}$ and $-\sqrt{E_b}$ can be used in the channel decoder to make better decisions about the transmitted codeword. In soft decision decoding, these distances are used in the decoding process. Soft decision decoding of linear block codes, which is more common in wireless systems than hard decision decoding, is treated in Section 8.2.7.

Hard decision decoding uses *minimum distance* decoding based on Hamming distance. In minimum distance decoding the n bits corresponding to a codeword are first demodulated to a 0 or 1, and the demodulator output is then passed to the decoder. The decoder compares this received codeword to the 2^k possible codewords that constitute the code and decides in favor of the codeword that is closest in Hamming distance to (differs in the least number of bits from) the received codeword. Mathematically, for a received codeword $\mathbf{R}$, the decoder uses the formula

$$\text{pick } \mathbf{C}_j \text{ s.t. } d(\mathbf{C}_j, \mathbf{R}) \leq d(\mathbf{C}_i, \mathbf{R}) \forall i \neq j. \tag{8.32}$$

If there is more than one codeword with the same minimum distance to $\mathbf{R}$, one of these is chosen at random by the decoder.

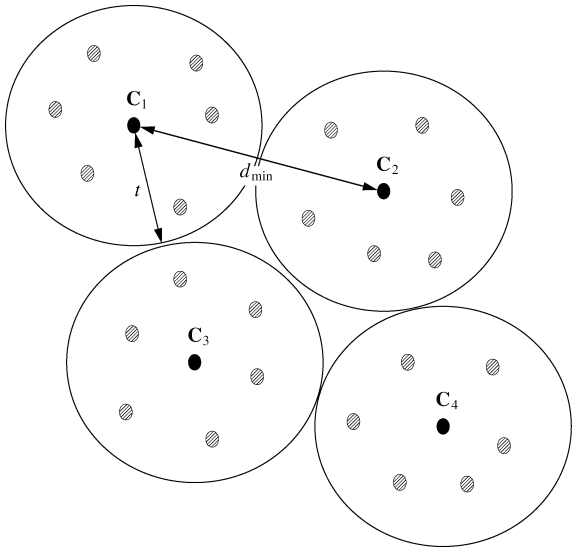

Figure 8.3: Maximum likelihood decoding in code space.

Maximum likelihood decoding picks the transmitted codeword that has the highest probability of having produced the received codeword. In other words, given the received codeword $\mathbf{R}$, the maximum likelihood decoder choses the codeword $\mathbf{C}_j$ as

$$\mathbf{C}_j = \arg\max_i p(\mathbf{R}|\mathbf{C}_i), \;\; i = 0, \ldots, 2^k - 1. \tag{8.33}$$

Since the most probable error event in an AWGN channel is the event with the minimum number of errors needed to produce the received codeword, the minimum distance criterion (8.32) and the maximum likelihood criterion

(8.33) are equivalent. Once the maximum likelihood codeword $\mathbf{C}_i$ is determined, it is decoded to the k bits that produce codeword $\mathbf{C}_i$.

Because maximum likelihood detection of codewords is based on a distance decoding metric, we can best illustrate this process in code space, as shown in Figure 8.3. The minimum Hamming distance between codewords, which are illustrated by the black dots in this figure, is $d_{\min}$. Each codeword is centered inside a sphere of radius $t = \lfloor .5d_{\min} \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. The shaded dots represent received sequences where one or more bits differ from those of the transmitted codeword. The figure indicates the Hamming distance between $\mathbf{C}_1$ and $\mathbf{C}_2$.

Minimum distance decoding can be used either to detect or to correct errors. Detected errors in a data block cause either the data to be dropped or a retransmission of the data. Error correction allows the corruption in the data to be reversed. For error correction the minimum distance decoding process ensures that a received codeword lying within a Hamming distance t from the transmitted codeword will be decoded correctly. Thus, the decoder can correct up to t errors, as can be seen from Figure 8.3: since received codewords corresponding to t or fewer errors will lie within the sphere centered around the correct codeword, it will be decoded as that codeword using minimum distance decoding. We see from Figure 8.3 that the decoder can detect all error patterns of $d_{\min} - 1$ errors. In fact, a decoder for an (n, k) code can detect $2^n - 2^k$ possible error patterns. The reason is that there are $2^k - 1$ nondetectable errors, corresponding to the case where a corrupted codeword is exactly equal to a codeword in the set of possible codewords (of size 2^k) that is not equal to the transmitted codeword. Since there are $2^n - 1$ total possible error patterns, this yields $2^n - 2^k$ detectable error patterns. Note that this is not hard decision decoding because we are not correcting errors, just detecting them.

Example 8.4: A $(5, 2)$ code has codewords $\mathbf{C}_0 = [00000]$, $\mathbf{C}_1 = [01011]$, $\mathbf{C}_2 = [10101]$, and $\mathbf{C}_3 = [11110]$. Suppose the all-zero codeword $\mathbf{C}_0$ is transmitted. Find the set of error patterns corresponding to nondetectable errors for this codeword transmission.

Solution:

The nondetectable error patterns correspond to the three nonzero code-words. That is, $\mathbf{e}_1 = [01011]$, $\mathbf{e}_2 = [10101]$, and $\mathbf{e}_3 = [11110]$ are nondetectable error patterns, since adding any of these to $\mathbf{C}_0$ results in a valid codeword.

8.2.6 Probability of Error for HDD in AWGN

The probability of codeword error, P_e, is defined as the probability that a transmitted codeword is decoded in error. Under hard decision decoding a received codeword *may* be decoded in error if it contains more than t errors (it will not be decoded in error if there is not an alternative codeword closer to the received codeword than the transmitted codeword). The error probability is thus bounded above by the probability that more than t errors occur. Since the bit errors in a codeword occur independently on an AWGN channel, this probability is given by

$$P_e \leq \sum_{j=t+1}^{n} \binom{n}{j} p^j (1-p)^{n-j}, \tag{8.34}$$

where p is the probability of error associated with transmission of the bits in the codeword. Thus, p corresponds to the error probability associated with uncoded modulation for the given energy per codeword symbol, as treated in Chapter 6 for AWGN channels. For example, if the codeword symbols are sent via coherent BPSK modulation then $p = Q(\sqrt{2E_c/N_0})$, where E_c is the energy per codeword symbol and $N_0/2$ is the noise power spectral density. Since there are k/n information bits per codeword symbol, the relationship between the energy per bit and the energy per symbol is $E_c = kE_b/n$. Thus, powerful block codes with a large number of parity bits (k/n

small) reduce the channel energy per symbol and therefore increase the error probability in demodulating the codeword symbols. However, the error correction capability of these codes typically more than compensates for this reduction, especially at high SNRs. At low SNRs this may not happen, in which case the code exhibits *negative coding gain* – it has a higher error probability than uncoded modulation. The bound (8.34) holds with equality when the decoder corrects exactly t or fewer errors in a codeword and cannot correct for more than t errors in a codeword. A code with this property is called a *perfect* code.

At high SNRs, the most likely way to make a codeword error is to mistake a codeword for one of its nearest neighbors. Nearest neighbor errors yield a pair of upper and lower bounds on error probability. The lower bound is the probability of mistaking a codeword for a given nearest neighbor at distance $d_{\min}$:

$$P_e \geq \sum_{j=t+1}^{d_{\min}} \binom{d_{\min}}{j} p^j (1-p)^{d_{\min}-j}. \tag{8.35}$$

The upper bound, a union bound, assumes that all of the other $2^k - 1$ codewords are at distance $d_{\min}$ from the transmitted codeword. Thus, the union bound is just $2^k - 1$ times (8.35), the probability of mistaking a given codeword for a nearest neighbor at distance $d_{\min}$:

$$P_e \leq (2^k - 1) \sum_{j=t+1}^{d_{\min}} \binom{d_{\min}}{j} p^j (1-p)^{d_{\min}-j}. \tag{8.36}$$

When the number of codewords is large or the SNR is low, both of these bounds are quite loose.

A tighter upper bound can be obtained by applying the Chemoff bound, $P(X \geq x) \leq e^{-x^2/2}$ for X a zero-mean, unit-variance, Gaussian random variable, to compute codeword error probability. Using this bound, it can be shown [3, Chap. 5.2] that the probability of decoding the all-zero codeword as the jth codeword with weight w_j is upper bounded by

$$P(w_j) \leq [4p(1-p)]^{w_j/2}. \tag{8.37}$$

Since the probability of decoding error is upper bounded by the probability of mistaking the all-zero codeword for any of the other codewords, we obtain the upper bound

$$P_e \leq \sum_{j=1}^{2^k-1} [4p(1-p)]^{w_j/2}. \tag{8.38}$$

This bound requires the weight distribution $\{w_j\}_{j=1}^{2^k-1}$ for all codewords (other than the all-zero codeword corresponding to $j = 0$) in the code. A simpler, slightly looser upper bound is obtained from (8.38) by using $d_{\min}$ instead of the individual codeword weights. This simplification yields the bound

$$P_e \leq (2^k - 1)[4p(1-p)]^{d_{\min}/2}. \tag{8.39}$$

Note that the probability of codeword error P_e depends on p, which is a function of the Euclidean distance between modulation points associated with the transmitted codeword symbols. In fact, the best codes for AWGN channels should not be based on Hamming distance: they should be based on maximizing the Euclidean distance between the codewords after modulation. However, this requires that the channel code be designed jointly with the modulation. This is the basic concept of coded modulation, which will be discussed in Section 8.7. However, Hamming distance is a better measure of code performance in fading when codes are combined with interleaving, as discussed in Section 8.8.

The probability of bit error after decoding the received codeword depends in general on the specific code and decoder and in particular on how bits are mapped to codewords, as in the bit mapping procedure associated with nonbinary modulation. This bit error probability is often approximated as

$$P_b \approx \frac{1}{n} \sum_{j=t+1}^{n} j \binom{n}{j} p^j (1-p)^{n-j} \tag{8.40}$$

[2, Chap. 6.5]; for $t = 1$, this can be simplified [2] to $P_b \approx p - p(1-p)^{n-1}$.

Example 8.5: Consider a $(24, 12)$ linear block code with a minimum distance $d_{\min} = 8$ (an extended Golay code, discussed in Section 8.2.8, is one such code). Find P_e based on the loose bound (8.39), assuming the codeword symbols are transmitted over the channel using BPSK modulation with $E_b/N_0 = 10\,\text{dB}$. Also find P_b for this code using the approximation $P_b = P_e/k$ and compare with the bit error probability for uncoded modulation.
Solution: For $E_b/N_0 = 10\,\text{dB}$ we have $E_c/N_0 = \frac{12}{24}10 = 5$. Thus, $p = Q(\sqrt{10}) = 7.82 \cdot 10^{-4}$. Using this value in (8.39) with $k = 12$ and $d_{\min} = 8$ yields $P_e \leq 3.92 \cdot 10^{-7}$. Using the P_b approximation we get $P_b \approx (1/k)P_e = 3.27 \cdot 10^{-8}$. For uncoded modulation we have $P_b = Q(\sqrt{2E_b/N_0}) = Q(\sqrt{20}) = 3.87 \cdot 10^{-6}$. Thus we obtain over two orders of magnitude performance gain with this code. Note that the loose bound can be orders of magnitude away from the true error probability, so this calculation may significantly underestimate the performance gain of the code.

8.2.7 Probability of Error for SDD in AWGN

The HDD described in the previous section discards information that can reduce probability of codeword error. For example, in BPSK, the transmitted signal constellation is $\pm\sqrt{E_b}$ and the received symbol after matched filtering is decoded as a 1 if it is closer to $\sqrt{E_b}$ and as a 0 if it is closer to $-\sqrt{E_b}$. Thus, the distance of the received symbol from $\sqrt{E_b}$ and $-\sqrt{E_b}$ is not used in decoding, yet this information can be used to make better decisions about the transmitted codeword. When these distances are used in the channel decoder it is called *soft decision decoding* (SDD), since the demodulator does not make a hard decision about whether a 0 or 1 bit was transmitted but rather makes a soft decision corresponding to the distance between the received symbol and the symbol corresponding to a 0-bit or a 1-bit transmission. We now describe the basic premise of SDD for BPSK modulation; these ideas are easily extended to higher-level modulations.

Consider a codeword transmitted over a channel using BPSK. As in the case of HDD, the energy per codeword symbol is $E_c = (k/n)E_b$. If the jth codeword symbol is a 1 it will be received as $r_j = \sqrt{E_c} + n_j$ and if it is a 0 it will be received as $r_j = -\sqrt{E_c} + n_j$, where n_j is the AWGN sample of mean zero and variance $N_0/2$ associated with the receiver. In SDD, given a received codeword $\mathbf{R} = [r_1, \ldots, r_n]$, the decoder forms a *correlation metric* $C(\mathbf{R},\ \mathbf{C}_i)$ for each codeword $\mathbf{C}_i (i = 0, \ldots, 2^k - 1)$ in the code and then the decoder chooses the codeword $\mathbf{C}_i$ with the highest correlation metric. The correlation metric is defined as

$$C(\mathbf{R},\ \mathbf{C}_i) = \sum_{j=1}^{n} (2c_{ij} - 1) r_j, \tag{8.41}$$

where c_{ij} denotes the jth coded bit in the codeword $\mathbf{C}_i$. If $c_{ij} = 1$ then $2c_{ij} - 1 = 1$ and if $c_{ij} = 0$ then $2c_{ij} - 1 = -1$. Thus the received codeword symbol is weighted by the polarity associated with the corresponding symbol in the codeword for which the correlation metric is being computed. Hence $C(\mathbf{R}, \mathbf{C}_i)$ is large when most of the received symbols have a large magnitude and the same polarity as the corresponding symbols in $\mathbf{C}_i$, is smaller when most of the received symbols have a small magnitude and the same polarity as the corresponding

symbols in $\mathbf{C}_i$, and is typically negative when most of the received symbols have a different polarity than the corresponding symbols in $\mathbf{C}_i$. In particular, at very high SNRs, if $\mathbf{C}_i$ is transmitted then $C(\mathbf{R}, \mathbf{C}_i) \approx n\sqrt{E_c}$ while $C(\mathbf{R}, \mathbf{C}_j) < n\sqrt{E_c}$ for $j \neq i$.

For an AWGN channel, the probability of codeword error is the same for any codeword of a linear code. Error analysis is typically easiest when assuming transmission of the all-zero codeword. Let us therefore assume that the all-zero codeword $\mathbf{C}_0$ is transmitted and the corresponding received codeword is $\mathbf{R}$. To correctly decode $\mathbf{R}$, we must have that $C(\mathbf{R},\ \mathrm{C}_0) > C(\mathbf{R},\ \mathbf{C}_i), i = 1, \ldots, 2^k - 1$. Let w_i denote the Hamming weight of the ith codeword $\mathbf{C}_i$, which equals the number of 1-bits in $\mathbf{C}_i$. Then, conditioned on the transmitted codeword $\mathbf{C}_i$, it follows that $C(\mathbf{R},\ \mathbf{C}_i)$ is Gauss distributed with mean $\sqrt{E_c}n(1-2w_i/n)$ and variance $nN_0/2$. Note that the correlation metrics are not independent, since they are all functions of $\mathbf{R}$. The probability $P_e(\mathbf{C}_i) = p(C(\mathbf{R},\ \mathbf{C}_0) < C(\mathbf{R},\ \mathbf{C}_i)$ can be shown to equal the probability that a Gauss-distributed random variable with variance $2w_iN_0$ is less than $-2w_i\sqrt{E_c}$; that is,

$$P_e(\mathbf{C}_i) = Q\left(\frac{2w_i\sqrt{E_c}}{\sqrt{2w_iN_0}}\right) = Q(\sqrt{2w_i\gamma_bR_c}). \tag{8.42}$$

Then, by the union bound, the probability of error is upper bounded by the sum of pairwise error probabilities relative to each $\mathbf{C}_i$:

$$P_e \leq \sum_{i=1}^{2^k-1} P_e(\mathbf{C}_i) = \sum_{i=1}^{2^k-1} Q(\sqrt{2w_i\gamma_bR_c}). \tag{8.43}$$

Computing (8.43) requires the weight distribution w_i $(i = 1, \ldots, 2^k - 1)$ of the code. This bound can be simplified by noting that $w_i \geq d_{\min}$, so

$$P_e \leq (2^k - 1)Q(\sqrt{2\gamma_bR_cd_{\min}}). \tag{8.44}$$

The Chemoff bound on the Q-function is $Q(\sqrt{2x}) < e^{-x}$. Applying this bound to (8.43) yields

$$P_e \leq (2^k - 1)e^{-\gamma_bR_cd_{\min}} < 2^ke^{-\gamma_bR_cd_{\min}} = e^{-\gamma_bR_cd_{\min}+k\ln 2}. \tag{8.45}$$

Comparing this bound with that of uncoded BPSK modulation,

$$P_b = Q(\sqrt{2\gamma_b}) < e^{-\gamma_b}, \tag{8.46}$$

we get a dB coding gain of approximately

$$G_c = 10\log_{10}[(\gamma_bR_cd_{\min} - k\ln 2)/\gamma_b] = 10\log_{10}[R_cd_{\min} - (k\ln 2)/\gamma_b]. \tag{8.47}$$

Note that the coding gain depends on the code rate, the number of information bits per code-word, the minimum distance of the code, and the channel SNR. In particular, the coding gain decreases as γ_b decreases, and it becomes negative at sufficiently low SNRs. In general the performance of SDD is about 2–3 dB better than HDD [1, Chap. 8.1].

Example 8.6: Find the approximate coding gain of SDD over uncoded modulation for the (24, 12) code with $\mathrm{d}_{\min} = 8$ considered in Example 8.5, with $\gamma_b = 10$ dB.
Solution: Setting $\gamma_b = 10$, $R_c = 12/24$, $d_{\min} = 8$, and $k = 12$ in (8.47) yields $G_c = 5$ dB. This significant coding gain is a direct result of the large minimum distance of the code.

8.2.8 Common Linear Block Codes

We now describe some common linear block codes. More details can be found in [1, 2, 3, 4]. The most common type of block code is a Hamming code, which is parameterized by an integer $m \geq 2$. For an $(n,\ k)$ Hamming code, $n = 2^m - 1$ and $k = 2^m - m - 1$, so $n - k = m$ redundant bits are introduced by the code. The minimum distance of all Hamming codes is $d_{\min} = 3$, so $t = 1$ error in the $n = 2^m - 1$ codeword symbols can be corrected. Although Hamming codes are not very powerful, they are perfect codes and thus have probability of error given exactly by the right side of (8.34).

Golay and extended Golay codes are another class of channel codes with good performance. The Golay code is a linear (23, 12) code with $d_{\min} = 7$ and $t = 3$. The extended Golay code is obtained by adding a single parity bit to the Golay code, resulting in a $(24, 12)$ block code with $d_{\min} = 8$ and $t = 3$. The extra parity bit does not change the error correction capability (since t remains the same), but it greatly simplifies implementation because the information bit rate is exactly half the coded bit rate. Thus, both uncoded and coded bit streams can be generated by the same clock, using every other clock sample to generate the uncoded bits. These codes have higher $d_{\min}$ and thus better error correction capabilities than Hamming codes, but at a cost of more complex decoding and a lower code rate $R_c = k/n$. The lower code rate implies that the code either has a lower data rate or requires additional bandwidth.

Another powerful class of block codes is the Bose–Chadhuri–Hocquenghem (BCH) codes. These are cyclic codes, and at high rates they typically outperform all other block codes with the same n and k at moderate to high SNRs. This code class provides a large selection of blocklengths, code rates, and error correction capabilities. In particular, the most common BCH codes have $n = 2^m - 1$ for any integer $m \geq 3$.

The P_b for a number of BCH codes under hard decision decoding and coherent BPSK modulation is shown in Figure 8.4. The plot is based on the approximation (8.40), where for coherent BPSK we have

$$p = Q(\sqrt{\frac{2E_c}{N_0}}) = Q(\sqrt{2R_c\gamma_b}). \tag{8.48}$$

In the figure, the BCH (127, 36) code actually has a negative coding gain at low SNRs. This is not uncommon for powerful channel codes owing to their reduced energy per symbol, as discussed in Section 8.2.6.

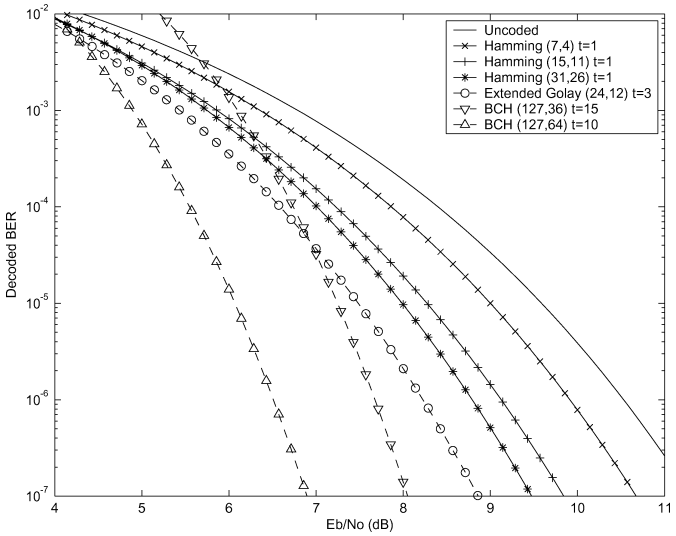

Figure 8.4: P_b for different BCH codes.

8.2.9 Nonbinary Block Codes: The Reed Solomon Code

A nonbinary block code has similar properties as the binary code: it has K information symbols mapped into codewords of length N. However, the N codeword symbols of each codeword are chosen from a nonbinary alphabet of size $q > 2$. Thus the codeword symbols can take any value in $\{0, 1, \ldots, q-1\}$. Usually $q = 2^k$, so that k bits can be mapped into one symbol.

The most common nonbinary block code is the Reed Solomon (RS) code, used in a range of applications from magnetic recording to cellular digital packet data (CDPD). Reed Solomon codes have $N = q - 1 = 2^k - 1$ and $K = 1, 2, \ldots, N-1$. The value of K dictates the error correction capability of the code. Specifically, an RS code can correct up to $t = .5\lfloor N - K \rfloor$ codeword symbol errors. In nonbinary codes the minimum distance between codewords is defined as the number of codeword symbols in which the codewords differ. Reed Solomon codes achieve a minimum distance of $d_{\min} = N - K + 1$, which is the largest possible minimum distance between codewords for any linear code with the same encoder input and output block lengths. Reed Solomon codes are often *shortened* to meet the requirements of a given system [4, Chap. 5.10].

Because nonbinary codes – and RS codes in particular – generate symbols corresponding to 2^k bits, they are sometimes used with M-ary modulation where $M = 2^k$. In particular, with 2^k-ary modulation each codeword symbol is transmitted over the channel as one of 2^k possible constellation points. If the error probability associated with the modulation (the probability of mistaking the received constellation point for a constellation point other than the transmitted point) is P_M, then the probability of codeword error associated with the non-binary code is upper bounded by

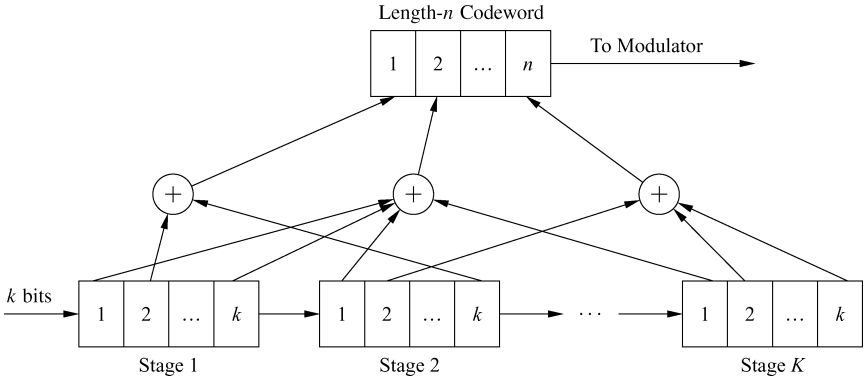

Figure 8.5: Convolutional encoder.

$$P_e \leq \sum_{j=t+1}^{N} \binom{N}{j} P_M^j (1 - P_M)^{N-j}, \tag{8.49}$$

which is similar to the form of (8.34) for the binary code. We can then approximate the probability of information symbol error as

$$P_s \approx \frac{1}{N} \sum_{j=t+1}^{N} j \binom{N}{j} P_M^j (1 - P_M)^{N-j}. \tag{8.50}$$

8.3 Convolutional Codes

A convolutional code generates coded symbols by passing the information bits through a linear finite-state shift register, as shown in Figure 8.5. The shift register consists of K stages with k bits per stage. There are n binary addition operators with inputs taken from all K stages: these operators produce a codeword of length n for each k-bit input sequence. Specifically, the binary input data is shifted into each stage of the shift register k bits at a time, and each of these shifts produces a coded sequence of length n. The rate of the code is $R_c = k/n$. The maximum span of output symbols that can be influenced by a given input bit in a convolutional code is called the *constraint length* of the code. It is clear from Figure 8.5 that a length-n codeword depends on kK input bits–in contrast to a block code, which only depends on k input bits. Thus, the constraint length of the encoder is kK bits or, equivalently, K k-bit bytes. Convolutional codes are said to have memory since the current codeword depends on more input bits (kK) than the number input to the encoder to generate it (k). Note that in general a convolutional encoder may not have the same number of bits per stage.

8.3.1 Code Characterization: Trellis Diagrams

When a length-n codeword is generated by the convolutional encoder of Figure 8.5, this code-word depends both on the k bits input to the first stage of the shift register as well as the *state* of the encoder, defined as the contents in the other $K-1$ stages of the shift register. In order to characterize this convolutional code, we must characterize how the codeword generation depends both on the k input bits and the encoder state, which has $2^{k(K-1)}$ possible values. There are multiple ways to characterize convolutional codes, including a tree diagram, state diagram, and trellis diagram [1, Chap. 8.2]. The tree diagram represents the encoder in the form of a tree, where each branch represents a different encoder state and the corresponding encoder output. A state diagram is a graph showing the different states of the encoder and the possible state transitions and corresponding encoder outputs. A trellis diagram uses the fact that the tree representation repeats itself once the number of stages in the tree exceeds the constraint length of the code. The trellis diagram simplifies the tree representation by merging nodes in the tree corresponding to the same encoder state. In this section we focus on the trellis representation of a convolutional code, since this is the most common characterization. The details of trellis diagram representation are best described by an example.

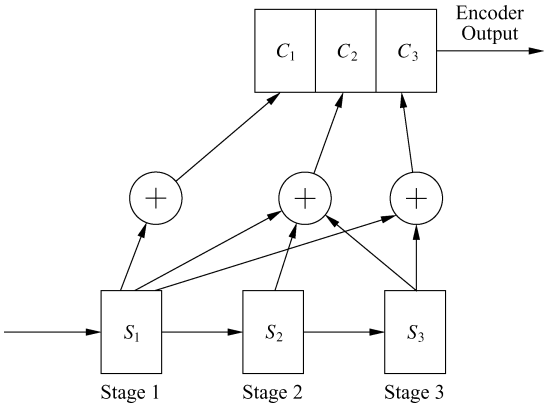

Figure 8.6: Convolutional encoder example $(n = 3,\ k = 1,\ k = 3)$.

Consider the convolutional encoder shown in Figure 8.6 with $n = 3, k = 1$, and $K = 3$. In this encoder, one bit at a time is shifted into Stage 1 of the three-stage shift register. At a given time t we denote the bit in Stage i of the shift register as S_i. The three stages of the shift register are used to generate a codeword of length 3, $C_1C_2C_3$; from the figure we see that $C_1 = S_1, C_2 = S_1 + S_2 + S_3$, and $C_3 = S_1 + S_3$. A bit sequence

$\mathbf{U}$ shifted into the encoder generates a sequence of coded symbols, which we denote by $\mathbf{C}$. Note that the coded symbols corresponding to C_1 are just the original information bits. As with block codes, when one of the coded symbols in a convolutional code corresponds to the original information bits we say that the code is *systematic*. We define the encoder state as $S = S_2S_3$ (i.e., the contents of the last two stages of the encoder), and there are $2^2 = 4$ possible values for this encoder state. To characterize the encoder, we must show for each input bit and each possible encoder state what the encoder output will be, and we must also show how the new input bit changes the encoder state for the next input bit.

The trellis diagram for this code is shown in Figure 8.7. The solid lines in Figure 8.7 indicate the encoder state transition when a 0-bit is input to Stage 1 of the encoder, and the dashed lines indicate the state transition corresponding to a 1-bit input. For example, starting at state $S = 00$, if a 0-bit is input to Stage 1 then, when the shift register transitions, the new state will remain as $S = 00$ (since the 0 in Stage 1 transitions to Stage 2, and the 0 in Stage 2 transitions to Stage 3, resulting in the new state $S = S_2S_3 = 00$). On the other hand, if a 1-bit is input to Stage 1 then, when the shift register transitions, the new state will become $S = 10$ (since the 1 in Stage 1 transitions to Stage 2, and the 0 in Stage 2 transitions to Stage 3, resulting in the new state $S = S_2S_3 = 10$). The encoder output corresponding to a particular encoder state S and input S_1 is written next to the transition lines in the figure. This output is the encoder output that results from the encoder addition operations on the bits S_1, S_2, and S_3 in each stage of the encoder. For example, if $S = 00$ and $S_1 = 1$ then the encoder output $C_1C_2C_3$ has $C_1 = S_1 = 1, C_2 = S_1 + S_2 + S_3 = 1$, and $C_3 = S_1 + S_3 = 1$. This output 111 is drawn next to the dashed line transitioning from state $S = 00$ to state $S = 10$ in Figure 8.7. Note that the encoder output for $S_1 = 0$ and $S = 00$ is always the all-zero codeword regardless of the addition operations that form the codeword $C_1C_2C_3$, since summing together any number of 0s always yields 0. The portion of the trellis between time t_i and t_{i+1} is called the ith *branch* of the trellis. Figure 8.7 indicates that the initial state at time t_0 is the all-zero state. The trellis achieves *steady state*, defined as the point where all states can be entered from either of two preceding states, at time t_3. After this steady state is reached, the trellis repeats itself in each time interval. Note also that, in steady state, each state transitions to one of two possible new states. In general, trellis structures starting from the all-zero state at time t_0 achieve steady state at time t_K.

For general values of k and K, the trellis diagram will have 2^{K-1} states, where each state has 2^k paths entering each node and 2^k paths leaving each node. Thus, the number of paths through the trellis grows exponentially with k, K, and the length of the trellis path.

Example 8.7: Consider the convolutional code represented by the trellis in Figure 8.7. For an initial state $S = S_2S_3 = 01$, find the state sequence S and the encoder output C for input bit sequence $\mathrm{U} = 011$.
Solution: The first occurrence of $S = 01$ in the trellis is at time t_2. We see at t_2 that if the information bit $S_1 = 0$ then we follow the solid line in the trellis from $S = 01$ at t_2 to $S = 00$ at t_3, and the output corresponding to this path through the trellis is $C = 011$. Now at t_3, starting at $S = 00$, for the information bit $S_1 = 1$ we follow the dashed line in the trellis to $S = 10$ at t_4, and the output corresponding to this path through the trellis is $C = 111$. Finally, at t_4, starting at $S = 10$, for the information bit $S_1 = 1$ we follow the dashed line in the trellis to $S = 11$ at t_5, and the output corresponding to this path through the trellis is $C = 101$.

8.3.2 Maximum Likelihood Decoding

The convolutional code generated by the finite state shift register is basically a finite-state machine. Thus, unlike an $(n,\ k)$ block code – where maximum likelihood detection entails finding the length-n codeword that is closest to the received length-n codeword – maximum likelihood detection of a convolutional code entails finding the most likely sequence of coded symbols $\mathbf{C}$ given the received sequence of coded symbols, which we denote by $\mathbf{R}$. In

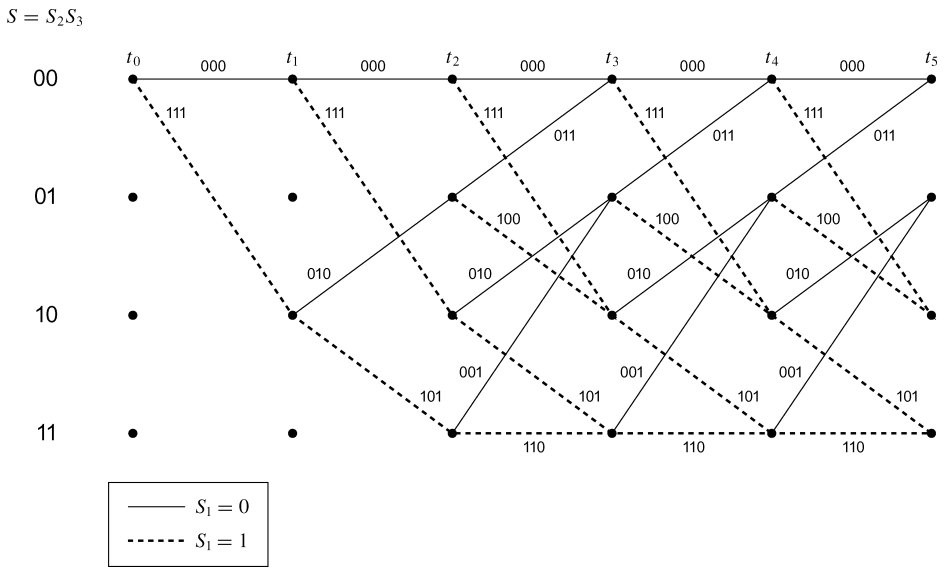

Figure 8.7: Trellis diagram.

particular, for a received sequence $\mathbf{R}$, the decoder decides that coded symbol sequence $\mathbf{C}^*$ was transmitted if

$$p(\mathbf{R}|\mathbf{C}^*) \geq p(\mathbf{R}|\mathbf{C}) \quad \forall \mathbf{C}. \tag{8.51}$$

Since each possible sequence $\mathbf{C}$ corresponds to one path through the trellis diagram of the code, maximum likelihood decoding corresponds to finding the maximum likelihood path through the trellis diagram. For an AWGN channel, noise affects each coded symbol independently. Thus, for a convolutional code of rate $1/n$, we can express the likelihood (8.51) for a path of length L through the trellis as

$$p(\mathbf{R}|\mathbf{C}) = \prod_{i=0}^{L-1} p(R_i|C_i) = \prod_{i=0}^{L-1} \prod_{j=1}^{n} p(R_{ij}|C_{ij}), \tag{8.52}$$

where C_i is the portion of the code sequence $\mathbf{C}$ corresponding to the ith branch of the trellis, R_i is the portion of the received code sequence $\mathbf{R}$ corresponding to the ith branch of the trellis, C_{ij} is the jth coded symbol corresponding to C_i, and R_{ij} is the jth received coded symbol corresponding to R_i. The $\log$ likelihood function is defined as the $\log$ of $p(\mathbf{R}|\mathbf{C})$, given as

$$\log p(\mathbf{R}|\mathbf{C}) = \sum_{i=0}^{L-1} \log p(R_i|C_i) = \sum_{i=0}^{L-1} \sum_{j=1}^{n} \log p(R_{ij}|C_{ij}). \tag{8.53}$$

The expression

$$B_i = \sum_{j=1}^{n} \log p(R_{ij}|C_{ij}) \tag{8.54}$$

is called the *branch metric* because it indicates the component of (8.53) associated with the ith branch of the trellis. The sequence or path that maximizes the likelihood function also maximizes the $\log$ likelihood function, since the $\log$ is monotonically increasing. However, it is computationally more convenient for the decoder to use the $\log$ likelihood function because it involves a summation rather than a product. The $\log$ likelihood function associated

with a given path through the trellis is also called the *path metric*, which by (8.53) is equal to the sum of branch metrics along each branch of the path. The path through the trellis with the maximum path metric corresponds to the maximum likelihood path.

The decoder can use either hard or soft decisions for the expressions $\log p(R_{ij}|C_{ij})$ in the log likelihood metric. For hard decision decoding, the R_{ij} is decoded as a 1 or a 0. The probability of hard decision decoding error depends on the modulation and is denoted as p. If $\mathbf{R}$ and $\mathbf{C}$ are N symbols long and differ in d places (i.e., their Hamming distance is d), then

$$p(\mathbf{R}|\mathbf{C}) = p^d(1-p)^{N-d}$$

and

$$\log p(\mathbf{R}|\mathbf{C}) = -d\log\frac{1-p}{p} + N\log(1-p). \tag{8.55}$$

Since $p < .5$, (8.55) is maximized when d is minimized. So the coded sequence $\mathbf{C}$ with minimum Hamming distance to the received sequence $\mathbf{R}$ corresponds to the maximum likelihood sequence.

In soft decision decoding, the value of the received coded symbols (R_{ij}) are used directly in the decoder, rather than quantizing them to 1 or 0. For example, if the C_{ij} are sent via BPSK over an AWGN channel with a 1 mapped to $\sqrt{E_c}$ and a 0 mapped to $-\sqrt{E_c}$, then

$$R_{ij} = \sqrt{E_c}(2C_{ij}-1) + n_{ij}, \tag{8.56}$$

where $E_c = kE_b/n$ is the energy per coded symbol and n_{ij} denotes Gaussian noise of mean zero and variance $\sigma^2 = .5N_0$. Thus,

$$p(R_{ij}|C_{ij}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(R_{ij}-\sqrt{E_c}(2C_{ij}-1))^2}{2\sigma^2}\right]. \tag{8.57}$$

Maximizing this likelihood function is equivalent to choosing the C_{ij} that is closest in Euclidean distance to R_{ij}. In determining which sequence $\mathbf{C}$ maximizes the log likelihood function (8.53), any terms that are common to two different sequences $\mathbf{C}_1$ and $\mathbf{C}_2$ can be neglected, since they contribute the same amount to the summation. Similarly, we can scale all terms in (8.53) without changing the maximizing sequence. Thus, by neglecting scaling factors and terms in (8.57) that are common to any C_{ij}, we can replace $\sum_{j=1}^{n}\log p(R_{ij}|C_{ij})$ in (8.53) with the *equivalent branch metric*

$$\mu_i = \sum_{j=1}^{n} R_{ij}(2C_{ij}-1) \tag{8.58}$$

and obtain the same maximum likelihood output.

We now illustrate the path metric computation under both hard and soft decisions for the convolutional code of Figure 8.6 with the trellis diagram in Figure 8.7. For simplicity, we will consider only two possible paths through the trellis and compute their corresponding likelihoods for a given received sequence $\mathbf{R}$. Assume we start at time t_0 in the all-zero state. The first path we consider is the all-zero path, corresponding to the all-zero input sequence. The second path we consider starts in state $S = 00$ at time t_0 and transitions to state $S = 10$ at time t_1, then to state $S = 01$ at time t_2, and finally to state $S = 00$ at time t_3, at which point this path merges with the all-zero path. Since the paths and therefore their branch metrics at times $t < t_0$ and $t \geq t_3$ are the same, the maximum likelihood path corresponds to the path whose sum of branch metrics over the branches in which the two paths differ is smaller. From Figure 8.7 we see that the all-zero path through the trellis generates the coded sequence $\mathbf{C}_0 = 000000000$

over the first three branches in the trellis. The second path generates the coded sequence $\mathbf{C}_1 = 111010011$ over the first three branches in the trellis.

Let us first consider hard decision decoding with error probability p. Suppose the received sequence over these three branches is $\mathbf{R} = 100110111$. Note that the Hamming distance between $\mathbf{R}$ and $\mathbf{C}_0$ is 6 while the Hamming distance between $\mathbf{R}$ and $\mathbf{C}_1$ is 4. As discussed previously, the most likely path therefore corresponds to $\mathbf{C}_1$, since it has minimum Hamming distance to $\mathbf{R}$. The path metric for the all-zero path is

$$M_0 = \sum_{i=0}^{2} \sum_{j=1}^{3} \log P(R_{ij}|C_{ij}) = 6 \log p + 3 \log(1-p), \tag{8.59}$$

while the path metric for the other path is

$$M_1 = \sum_{i=0}^{2} \sum_{j=1}^{3} \log P(R_{ij}|C_{ij}) = 4 \log p + 5 \log(1-p). \tag{8.60}$$

Assuming $p \ll 1$, which is generally the case, this yields $M_0 \approx 6 \log p$ and $M_1 \approx 4 \log p$. Since $\log p < 1$, this confirms that the second path has a larger path metric than the first.

Let us now consider soft decision decoding over time t_0 to t_3. Suppose the received sequence (before demodulation) over these three branches, for $E_c = 1$, is $\mathbf{R} = (.8, -.35, -.15, 1.35, 1.22, -.62, .87, 1.08, .91)$. The path metric for the all-zero path is

$$M_0 = \sum_{i=0}^{2} \mu_i = \sum_{i=0}^{2} \sum_{j=1}^{3} R_{ij}(2C_{ij} - 1) = \sum_{i=0}^{2} \sum_{j=1}^{3} -R_{ij} = -5.11,$$

and the path metric for the second path is

$$M_1 = \sum_{i=0}^{2} \sum_{j=1}^{3} R_{ij}(2C_{ij} - 1) = 1.91.$$

Thus, the second path has a higher path metric than the first. In order to determine if the second path is the maximum likelihood path, we must compare its path metric to that of all other paths through the trellis.

The difficulty with maximum likelihood decoding is that the complexity of computing the $\log$ likelihood function (8.53) grows exponentially with the memory of the code, and this computation must be done for every possible path through the trellis. The Viterbi algorithm, discussed in the next section, reduces the complexity of maximum likelihood decoding by taking advantage of the structure of the path metric computation.

8.3.3 The Viterbi Algorithm

The Viterbi algorithm, introduced by Viterbi in 1967 [5], reduces the complexity of maximum likelihood decoding by systematically removing paths from consideration that cannot achieve the highest path metric. The basic premise is to look at the partial path metrics associated with all paths *entering* a given node (node N) in the trellis. Since the possible paths through the trellis *leaving* node N are the same for each *entering* path, the complete trellis path with the highest path metric that goes through node N must coincide with the path that has the highest partial path metric up to node N. This is illustrated in Figure 8.8, where path 1, path 2, and path 3 enter node N (at trellis depth n) with partial path metrics $P^l = \sum_{k=0}^{n-1} B_k^l (l = 1, 2, 3)$ up to this node. Assume P^1 is the largest of these partial path metrics. The complete path with the highest metric has branch metrics $\{B_k\}$ after node N. The maximum likelihood path starting from node N (i.e., the path starting from node N with the largest path metric) has partial path metric $\sum_{k=n}^{\infty} B_k$. The complete path metric for path $l (l = 1, 2, 3)$ up to node N and the maximum likelihood

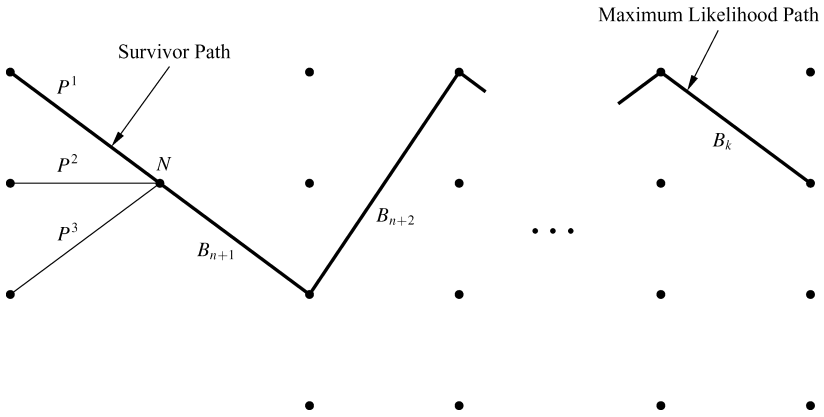

Figure 8.8: Partial path metrics on maximum likelihood path.

path after node N is $P^l + \sum_{k=n}^{\infty} B_k (l = 1, 2, 3)$, and thus the path with the maximum partial path metric P^l up to node N (path 1 in this example) must correspond to the path with the largest path metric that goes through node N.

The Viterbi algorithm takes advantage of this structure by discarding all paths entering a given node except the path with the largest partial path metric up to that node. The path that is not discarded is called the *survivor path*. Thus, for the example of Figure 8.8, path 1 is the survivor at node N and paths 2 and 3 are discarded from further consideration. Hence, at every stage in the trellis there are 2^{K-1} surviving paths, one for each possible encoder state. A branch for a given stage of the trellis cannot be decoded until all surviving paths at a subsequent trellis stage overlap with that branch; see Figure 8.9, which shows the surviving paths at time t_{k+3}. We see in the figure that all of these surviving paths can be traced back to a *common stem* from time t_k to t_{k+1}. At this point the decoder can output the codeword symbol C_i associated with this branch of the trellis. Note that there is not a fixed decoding delay associated with how far back in the trellis a common stem occurs for a given set of surviving paths, since this delay depends on k, K, and the specific code properties. To avoid a random decoding delay, the Viterbi algorithm is typically modified so that, at a given stage in the trellis, the most likely branch n stages back is decided upon based on the partial path metrics up to that point. Although this modification does not yield exact maximum likelihood decoding, for n sufficiently large (typically $n \geq 5K$) it is a good approximation.

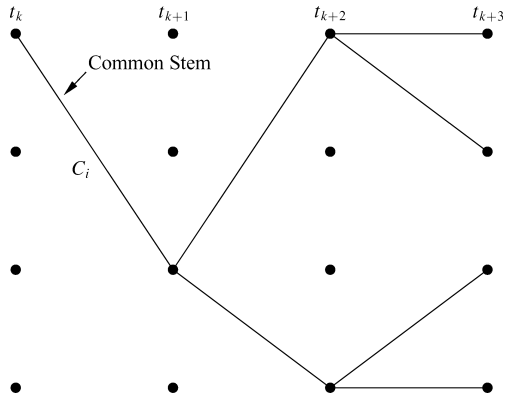

Figure 8.9: Common stem for all survivor paths in the trellis.

The Viterbi algorithm must keep track of $2^{k(K-1)}$ surviving paths and their corresponding metrics. At each stage, in order to determine the surviving path, 2^k metrics must be computed for each node corresponding to the

2^k paths entering each node. Thus, the number of computations in decoding and the memory requirements for the algorithm increase exponentially with k and K. This implies that practical implementations of convolutional codes are restricted to relatively small values of k and K.

8.3.4 Distance Properties

As with block codes, the error correction capability of convolutional codes depends on the distance between codeword sequences. Since convolutional codes are linear, the minimum distance between all codeword sequences can be found by determining the minimum distance from any sequence or, equivalently, determining any trellis path to the all-zero sequence/trellis path. Clearly the trellis path with minimum distance to the all-zero path will diverge and remerge with the all-zero path, so that the two paths coincide except over some number of trellis branches. To find this minimum distance path we must consider all paths that diverge from the all-zero state and then remerge with this state. As an example, in Figure 8.10 we draw all paths in Figure 8.7 between times t_0 and t_5 that diverge and remerge with the all-zero state. Note that path 2 is identical to path 1 – just shifted in time – and thus is not considered as a separate path. Note also that we could look over a longer time interval, but any paths that diverge and remerge over this longer interval would traverse the same branches (shifted in time) as one of these paths plus some additional branches and would therefore have larger path metrics. In particular, we see that path 4 traverses the trellis branches 00-10-01-10-01-00, whereas path 1 traverses the branches 00-10-01-00. Since path 4 traverses the same branches as path 1 on its first, second, and last transition – and since it has additional transitions – its path metric will be no smaller than the meteric of path 1. Thus we need not consider a longer time interval to find the minimum distance path. For each path in Figure 8.10 we label the Hamming distance of the codeword on each branch to the all-zero codeword in the corresponding branch of the all-zero path. By summing up the Hamming distances on all branches of each path, we see that path 1 has a Hamming distance of 6 and that paths 3 and 4 have Hamming distances of 8. Recalling that dashed lines indicate 1-bit inputs while solid lines indicate 0-bit inputs, we see that path 1 corresponds to an input bit sequence from t_0 to t_5 of 10000, path 3 corresponds to an input bit sequence of 11000, and path 4 corresponds to an input bit sequence of 10100. Thus, path 1 results in one bit error relative to the all-zero squence, and paths 3 and 4 result in two bit errors.

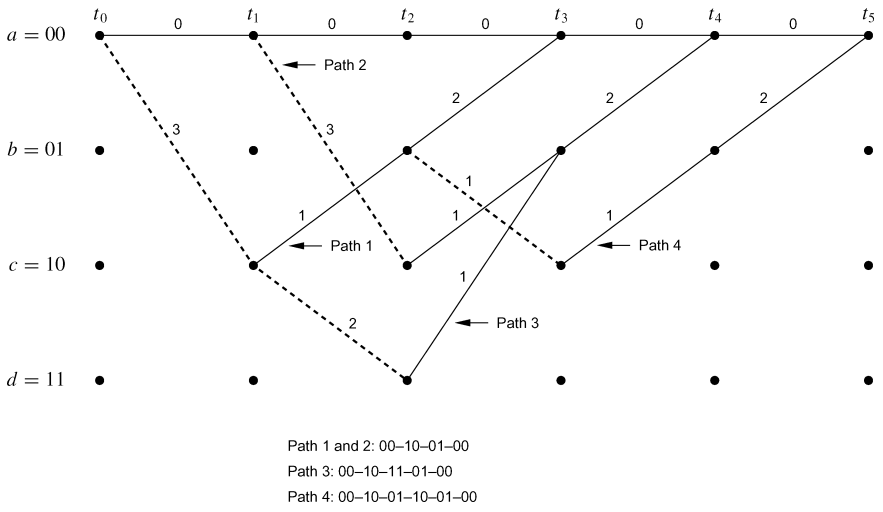

Figure 8.10: Path distances to the all-zero path.

We define the *minimum free distance* d_f of a convolutional code, also called simply the free distance, to be the minimum Hamming distance of all paths through the trellis to the all-zero path, which for this example is 6.

The error correction capability of the code is obtained in the same manner as for block codes, with $d_{\min}$ replaced by d_f, so that the code can correct t channels errors with $t = \lfloor .5d_f \rfloor$.

8.3.5 State Diagrams and Transfer Functions

The transfer function of a convolutional code is used to characterize paths that diverge and remerge from the all-zero path, and it is also used to obtain probability of error bounds. The transfer function is obtained from the code's state diagram representing possible transitions from the all-zero state to the all-zero state. The state diagram for the code illustrated in Figure 8.7 is shown in Figure 8.11, with the all-zero state $a = 00$ split into a second node e to facilitate representing paths that begin and end in this state. Transitions between states due to a 0 input bit are represented by solid lines, while transitions due to a 1 input bit are represented by dashed lines. The branches of the state diagram are labeled as either $D^0 = 1, D^1, D^2$, or D^3, where the exponent of D corresponds to the Hamming distance between the codeword (which is shown for each branch transition) and the all-zero codeword in the all-zero path. The self-loop in node a can be ignored because it does not contribute to the distance properties of the code.

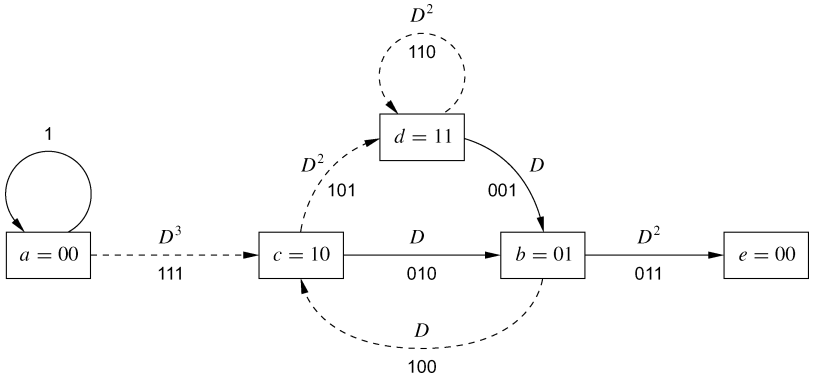

Figure 8.11: State diagram.

The state diagram can be represented by state equations for each state. For the example of Figure 8.11, we obtain state equations corresponding to the four states

$$X_c = D^3 X_a + D X_b, \quad X_b = D X_c + D X_d, \quad X_d = D^2 X_c + D^2 X_d, \quad X_e = D^2 X_b, \tag{8.61}$$

where $X_a, \ldots, X_e$ are dummy variables characterizing the partial paths. The transfer function of the code, describing the paths from state a to state e, is defined as $T(D) = X_e / X_a$. By solving the state equations for the code, which can be done using standard techniques such as Mason's formula, we obtain a transfer function of the form

$$T(D) = \sum_{d=d_f}^{\infty} a_d D^d, \tag{8.62}$$

where a_d is the number of paths with Hamming distance d from the all-zero path. As stated before, the minimum Hamming distance to the all-zero path is d_f, and the transfer function $T(D)$ indicates that there are a_{d_f} paths with this minimum distance. For the example of Figure 8.11, we can solve the state equations given in (8.61) to get the transfer function

$$T(D) = \frac{D^6}{1 - 2D^2} = D^6 + 2D^8 + 4D^{10} + \cdots . \tag{8.63}$$

We see from the transfer function that there is one path with minimum distance $d_f = 6$ and two paths with Hamming distance 8, which is consistent with Figure 8.10. The transfer function is a convenient shorthand for enumerating the number and corresponding Hamming distance of all paths in a particular code that diverge and later remerge with the all-zero path.

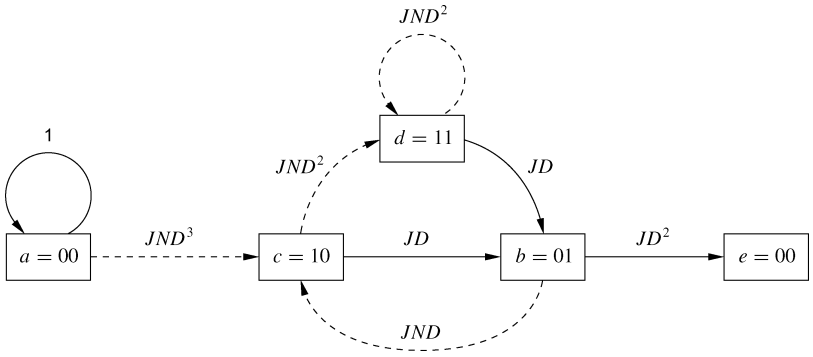

Figure 8.12: Extended state diagram.

Although the transfer function is sufficient to capture the number and Hamming distance of paths in the trellis to the all-zero path, we need a more detailed characterization to compute the bit error probability of the convolutional code. We therefore introduce two additional parameters into the transfer function, N and J, for this additional characterization. The factor N is introduced on all branch transitions associated with a 1 input bit (dashed lines in Figure 8.11). The factor J is introduced to every branch in the state diagram such that the exponent of J in the transfer function equals the number of branches in any given path from node a to node e. The extended state diagram corresponding to the trellis of Figure 8.7 is shown in Figure 8.12.

The extended state diagram can also be represented by state equations. For the example of Figure 8.12, these are given by:

$$\begin{aligned} X_c &= JND^3X_a + JNDX_b, \quad X_b = JDX_c + JDX_d, \\ X_d &= JND^2X_c + JND^2X_d, \quad X_e = JD^2X_b. \end{aligned} \tag{8.64}$$

Similarly to the previous transfer function definition, the transfer function associated with this extended state is defined as $T(D, N, J) = X_e/X_a$, which for this example yields

$$\begin{aligned} T(D, N, J) &= \frac{J^3ND^6}{1 - JND^2(1 + J)} \\ &= J^3ND^6 + J^4N^2D^8 + J^5N^2D^8 + J^5N^3D^{10} + \cdots. \end{aligned} \tag{8.65}$$

The factor J is most important when we are interested in transmitting finite-length sequences; for infinite-length sequences we typically set $J = 1$ to obtain the transfer function for the extended state:

$$T(D, N) = T(D, N, J = 1). \tag{8.66}$$

The transfer function for the extended state tells us more information about the diverging and remerging paths than just their Hamming distance; namely, the minimum distance path with Hamming distance 6 is of length 3 and results in a single bit error (exponent of N is unity), one path of Hamming distance 8 is of length 4 and results in two bit errors, and the other path of Hamming distance 8 is of length 5 and results in two bit errors, consistent

with Figure 8.10. The extended transfer function is a convenient shorthand for representing the Hamming distance, length, and number of bit errors that correspond to each diverging and remerging path of a code from the all-zero path. In the next section we show that this representation is useful in characterizing the probability of error for convolutional codes.

8.3.6 Error Probability for Convolutional Codes

Since convolutional codes are linear codes, the probability of error can be obtained by first assuming that the all-zero sequence is transmitted and then determining the probability that the decoder decides in favor of a different sequence. We will consider error probability for both hard decision and soft decision decoding; soft decisions are much more common in wireless systems owing to their superior performance.

First consider soft decision decoding. We are interested in the probability that the all-zero sequence is sent but a different sequence is decoded. If the coded symbols output from the convolutional encoder are sent over an AWGN channel using coherent BPSK modulation with energy $E_c = R_c E_b$, then it can be shown [1] that, if the all-zero sequence is transmitted, the probability of mistaking this sequence with a sequence Hamming distance d away is

$$P_2(d) = Q\left(\sqrt{\frac{2E_c}{N_0}d}\right) = Q(\sqrt{2\gamma_b R_c d}). \tag{8.67}$$

We call this probability the *pairwise error probability*, since it is the error probability associated with a pairwise comparison of two paths that differ in d bits. The transfer function enumerates all paths that diverge and remerge with the all zero path, so by the union bound we can upper bound the probability of mistaking the all-zero path for another path through the trellis as

$$P_e \leq \sum_{d=d_f}^{\infty} a_d Q(\sqrt{2\gamma_b R_c d}), \tag{8.68}$$

where a_d denotes the number of paths of distance d from the all-zero path. This bound can be expressed in terms of the transfer function itself if we use the Chemoff upper bound for the Q-function, which yields

$$Q(\sqrt{2\gamma_b R_c d}) \leq e^{-\gamma_b R_c d}.$$

Using this in (8.68) we obtain the upper bound

$$P_e < T(D)|_{D=e^{-\gamma_b R_c}}. \tag{8.69}$$

This upper bound tells us the probability of mistaking one sequence for another, but it does not yield the more fundamental probability of bit error. We know that the exponent in the factor N of $T(D, N)$ indicates the number of information bit errors associated with selecting an incorrect path through the trellis. Specifically, we can express $T(D, N)$ as

$$T(D, N) = \sum_{d=d_f}^{\infty} a_d D^d N^{f(d)}, \tag{8.70}$$

where $f(d)$ denotes the number of bit errors associated with a path of distance d from the all-zero path. Then we can upper bound the bit error probability for $k = 1$ as

$$P_b \leq \sum_{d=d_f}^{\infty} a_d f(d) Q(\sqrt{2\gamma_b R_c d}) \tag{8.71}$$

[1, Chap. 8.2], which differs from (8.68) only in the weighting factor $f(d)$ corresponding to the number of bit errors in each incorrect path. If the Q-function is upper bounded using the Chemoff bound as before, we get the upper bound

$$P_b < \left. \frac{dT(D,N)}{dN} \right|_{N=1, D=e^{-\gamma_b R_c}}. \tag{8.72}$$

If $k > 1$ then we divide (8.71) or (8.72) by k to obtain P_b.

All of these bounds assume coherent BPSK transmission (or coherent QPSK, which is equivalent to two independent BPSK transmissions). For other modulations, the pairwise error probability $P_2(d)$ must be recomputed based on the probability of error associated with the given modulation.

Let us now consider hard decision decoding. The probability of selecting an incorrect path at distance d from the all-zero path, for d odd, is given by

$$P_2(d) = \sum_{k=.5(d+1)}^{d} \binom{d}{k} p^k (1-p)^{(d-k)}, \tag{8.73}$$

where p is the probability of error on the channel. This follows because the incorrect path will be selected only if the decoded path is closer to the incorrect path than to the all-zero path – that is, the decoder makes at least $.5(d+1)$ errors. If d is even then the incorrect path is selected when the decoder makes more than $.5d$ errors, and the decoder makes a choice at random if the number of errors is exactly $.5d$. We can bound the pairwise error probability as

$$P_2(d) < [4p(1-p)]^{d/2}. \tag{8.74}$$

Following the same approach as in soft decision decoding, we then obtain the error probability bound as

$$P_e < \sum_{d=d_f}^{\infty} a_d [4p(1-p)]^{d/2} < T(D)|_{D=\sqrt{4p(1-p)}}, \tag{8.75}$$

and

$$P_b < \sum_{d=d_f}^{\infty} a_d f(d) P_2(d) = \left. \frac{dT(D,N)}{dN} \right|_{N=1,\ D=\sqrt{4p(1-p)}}. \tag{8.76}$$

8.4 Concatenated Codes

A concatenated code uses two levels of coding: an inner code and an outer code, as shown in Figure 8.13. The inner code is typically designed to remove most of the errors introduced by the channel, and the outer code is typically a less powerful code that further reduces error probability when the received coded bits have a relatively low probability of error (since most errors are corrected by the inner code). Concatenated codes can be particularly effective at correcting bursts of errors, which are common in wireless channels as a result of deep fades. In addition, at low SNRs Viterbi decoding of a convolutional code tends to have errors that occur in bursts. To compensate for these error bursts, an inner convolutional code is often concatenated with an outer Reed Solomon code, since

RS codes have good burst error correcting properties. In addition, concatenated codes frequently have the inner and outer codes separated by an interleaver to break up bursts of errors. Interleaver design for different coding techniques is described in Section 8.8.

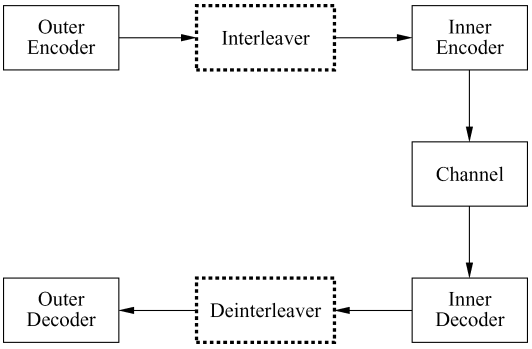

Figure 8.13: Concatenated coding.

Concatenated codes typically achieve very low error probability with less complexity than a single code with the same error probability performance. The decoding of concatenated codes is usually done in two stages, as indicated in the figure: first the inner code is decoded, and then the outer code is decoded separately. This is a suboptimal technique, since in fact both codes are working in tandem to reduce error probability. However, the maximum likelihood decoder for a concatenated code, which performs joint decoding, is highly complex. It was discovered in the mid-1990s that a near-optimal decoder for concatenated codes can be obtained based on iterative decoding. This is the basic premise behind turbo codes, described in the next section.

8.5 Turbo Codes

Turbo codes, introduced in 1993 in a landmark paper by Berrou, Glavieux, and Thitimajshima ([6]; see also [7]), are powerful codes that can come within a fraction of a decibel of the Shannon capacity limit on AWGN channels. Turbo codes and the more general family of codes on graphs with iterative decoding algorithms [8, 9] have been studied extensively, yet some of their characteristics are still not well understood. The main ideas behind codes on graphs were introduced by Gallager in the early sixties [10]; at the time, however, these cod- ing techniques were thought impractical and were generally not pursued by researchers in the field. The landmark 1993 paper on turbo codes [6] provided more than enough motivation to revisit the work of Gallager and others on iterative, graph-based decoding techniques.

As first described by Berrou et al., turbo codes consist of two key components: parallel concatenated encoding and iterative, "turbo" decoding [6, 11]. A typical parallel concatenated encoder is shown in Figure 8.14. It consists of two parallel convolutional encoders separated by an interleaver, with the input to the channel being the data bits m along with the parity bits X_1 and X_2 output from each of the encoders in response to input m. Since the m information bits are transmitted as part of the codeword, we consider this a systematic turbo code. The key to parallel concatenated encoding lies in the recursive nature of the encoders and the impact of the interleaver on the information stream. Interleavers also play a significant role in the reduction of error floors [11], which are commonly exhibited in turbo codes.

Iterative or "turbo" decoding exploits the component-code substructure of the turbo encoder by associating a component decoder with each of the component encoders. More specifically, each decoder performs soft input–soft output decoding, as shown in Figure 8.15 for the example encoder of Figure 8.14. In this figure decoder 1 generates a soft decision in the form of a probability measure $p(m_1)$ on the transmitted information bits based on

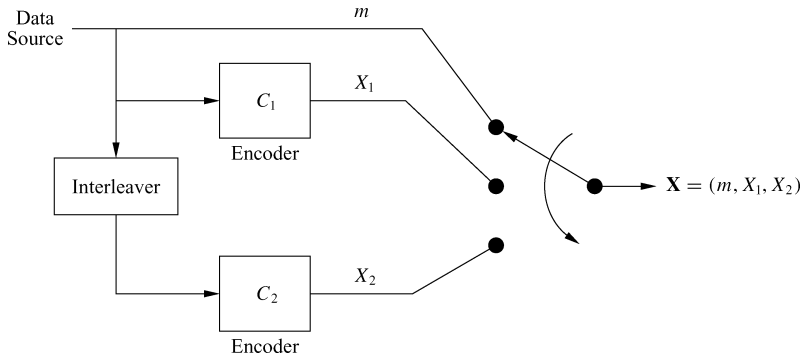

Figure 8.14: Parallel concatenated (turbo) encoder.

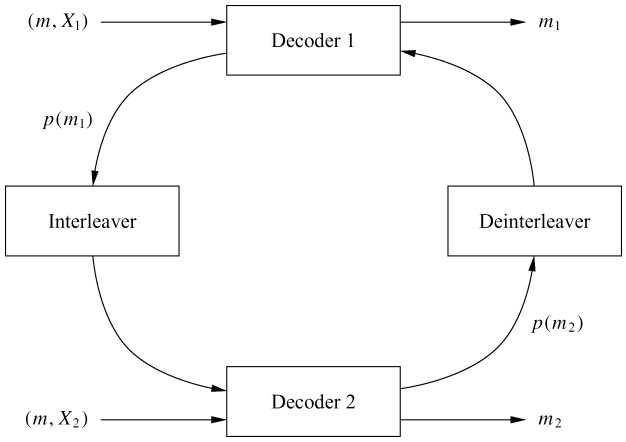

Figure 8.15: Turbo decoder.

the received codeword (m, X_1). The probability measure is generated based on either a maximum a posteriori (MAP) probability or a soft output Viterbi algorithm (SOVA), which attaches a reliability indicator to the VA hard decision outputs [4, Chap. 12.5]. This probability information is passed to decoder 2, which generates its own probability measure $p(m_2)$ from its received codeword (m, X_2) and the probability measure $p(m_1)$. This reliability information is input to decoder 1, which revises its measure $p(m_1)$ based on this information and the original received codeword. Decoder 1 sends the new reliability information to decoder 2, which revises its measure using this new information. Turbo decoding proceeds in an iterative manner, with the two component decoders alternately updating their probability measures. Ideally the decoders will eventually agree on probability measures that reduce to hard decisions $m = m_1 = m_2$. However, the stopping condition for turbo decoding is not well-defined, in part because there are many cases in which the turbo decoding algorithm does not converge: the decoders cannot agree on the value of m. Several methods have been proposed for detecting convergence (if it occurs), including bit estimate variance [7] and neural net–based techniques [12].

The simulated performance of turbo codes over multiple iterations of the decoder is shown in Figure 8.16 for a code composed of two convolutional codes of rate $1/2$ with constraint length $K = 5$ separated by an interleaver of depth $d = 2^{16} = 65536$. The decoder converges after approximately eighteen iterations. This curve indicates several important aspects of turbo codes. First, note their exceptional performance: bit error probability of 10^{-6} at an E_b/N_0 of less than 1 dB. In fact, the original turbo code proposed in [6] performed within .5 dB of the Shannon capacity limit at $P_b = 10^{-5}$. The intuitive explanation for the amazing performance of turbo codes is that the code complexity introduced by the encoding structure is similar to the codes that achieve Shannon capacity. The

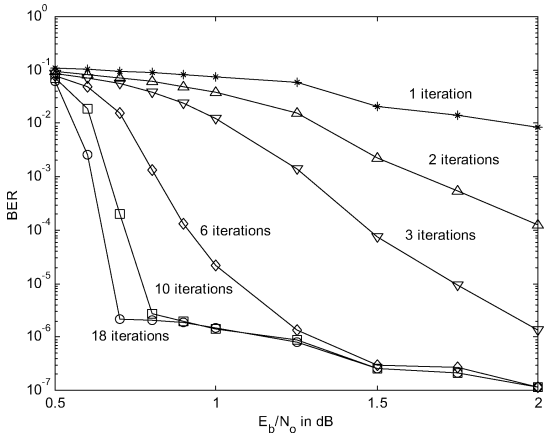

Figure 8.16: Turbo code performance (rate $1/2$, $K = 5$ component codes with interleaver depth 2^{16}).

iterative procedure of the turbo decoder allows these codes to be decoded without excessive complexity. However, note that the turbo code exhibits an error floor: in Figure 8.16 this floor occurs at 10^{-6}. This floor is problematic for systems that require extremely low bit error rates. Several mechanisms have been investigated to lower the error floor, including bit interleaving and increasing the constraint length of the component codes.

An alternative to parallel concatenated coding is serial concatenated coding [13]. In this coding technique, one component code serves as an outer code, and then the output of this first encoder is interleaved and passed to a second encoder. The output of the second encoder comprises the coded bits. Iterative decoding between the inner and outer codes is used for decoding. There has been much work comparing serial and parallel concatenated code performance (see e.g. [13, 14, 15]). Whereas both codes perform very well under similar delay and complexity conditions, in some cases serial concatenated coding performs better at low bit error rates and also can exhibit a lower error floor.

8.6 Low-Density Parity-Check Codes

Low-density parity-check (LDPC) codes were originally invented by Gallager [10]. However, these codes were largely ignored until the introduction of turbo codes, which rekindled some of the same ideas. Subsequent to the landmark paper [6] on turbo codes in 1993, LDPC codes were announced by Mackay and Neal [16] and Wiberg [17]. Shortly thereafter it was recognized that these new code designs were actually reinventions of Gallager's original ideas, and subsequently much work has been devoted to finding the capacity limits, encoder and decoder designs, and practical implementation of LDPC codes for different channels.

Low-density parity-check codes are linear block codes with a particular structure for the parity check matrix $\mathbf{H}$, which was defined in Section 8.2.3. Specifically, a (d_v, d_c) regular binary LDPC has a parity-check matrix $\mathbf{H}$ with d_v 1s in each column and d_c ls in each row, where d_v and d_c are chosen as part of the codeword design and are small relative to the codeword length. Since the fraction of nonzero entries in $\mathbf{H}$ is small, the parity-check matrix for the code has a low density – hence the name low-density parity-check codes.

Provided that the codeword length is long, LDPC codes achieve performance close to the Shannon limit and in some cases surpass the performance of parallel or serially concatenated codes [18]. The fundamental practical difference between turbo codes and LDPC codes is that turbo codes tend to have low encoding complexity (linear in blocklength) but high decoding complexity (due to their iterative nature and message passing). In contrast, LDPC codes tend to have relatively high encoding complexity but low decoding complexity. In particular, like turbo codes, LDPC decoding uses iterative techniques that are related to Pearl's belief propagation, which is commonly used by

the artificial intelligence community (see [19]). However, the belief propagation corresponding to LDPC decoding may be simpler than for turbo decoding [19, 20]. In addition, this belief propagation decoding is parallelizable and can be closely approximated with decoders of very low complexity [21], although this may also be possible for turbo decoding. Finally, the decoding algorithm for LDPC codes can determine when a correct codeword has been detected, which is not necessarily the case for turbo codes. The trade-offs between turbo and LDPC codes is an active research area, and many open questions remain regarding their relative performance.

Additional work in the area of LDPC codes includes finding capacity limits for these codes [21], determining effective code designs [22] and efficient encoding and decoding algorithms [21, 23], and expanding the code designs to include nonregular [18] and nonbinary LDPC codes [24] as well as coded modulation [25].

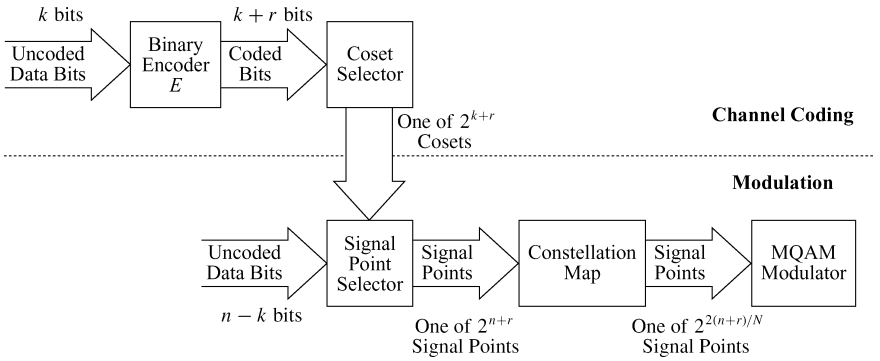

Figure 8.17: General coding scheme.

8.7 Coded Modulation

Although Shannon proved the capacity theorem for AWGN channels in the late 1940s, it wasn't until the 1990s that rates approaching the Shannon limit were attained – primarily for AWGN channels with binary modulation using turbo codes. Shannon's theorem predicted the possibility of reducing both energy and bandwidth simultaneously through coding. However, as described in Section 8.1, traditional error correction coding schemes (e.g., block, convolutional, and turbo codes) provide coding gain at the expense of increased bandwidth or reduced data rate.

The spectrally efficient coding breakthrough came when Ungerboeck [26] introduced a coded modulation technique to jointly optimize both channel coding and modulation. This joint optimization results in significant coding gains without bandwidth expansion. Ungerboeck's trellis coded modulation, which uses multilevel/phase signal modulation and simple convolutional coding with mapping by set partitioning, compares favorably to later developments in coded modulation (coset and lattice codes) as well as to more complex trellis codes [27]. We now outline the general principles of this coding technique. Comprehensive treatments of trellis, lattice, and coset codes can be found in [26, 27, 28].

The basic scheme for trellis and lattice coding – or, more generally, for any type of coset coding – is depicted in Figure 8.17. There are five elements required to generate the coded modulation:

1. a binary encoder E, block or convolutional, that operates on k uncoded data bits to produce $k+r$ coded bits;

2. a subset (coset) selector, which uses the coded bits to choose one of 2^{k+r} subsets from a partition of the N-dimensional signal constellation;

3. a point selector, which uses $n-k$ additional uncoded bits to choose one of the 2^{n-k} signal points in the selected subset;

4. a constellation map, which maps the selected point from N-dimensional space to a sequence of $N/2$ points in two-dimensional space; and

5. an MQAM modulator (or other M-ary modulator).

The first two elements constitute the channel coding, and the remaining elements are the modulation. The receiver essentially reverses the modulation and coding steps. After MQAM demodulation and an inverse $2/N$ constellation mapping, decoding is done in essentially two stages: first, the points within each subset that are closest to the received signal point are determined; then, the maximum likelihood subset sequence is calculated. When the encoder E is a convolutional encoder, this scheme is referred to as *trellis coded modulation*; for E a block encoder, it is called *lattice coded modulation*. These schemes are also referred to as trellis or lattice codes.

Steps 1–5 essentially decouple the channel coding gain from the gain associated with signal shaping in the modulation. Specifically, the code distance properties – and thus the channel coding gain – are determined by the encoder (E) properties and the subset partitioning, which are essentially decoupled from signal shaping. We will discuss the channel coding gain in more detail below. Optimal shaping of the signal constellation provides up to an additional 1.53 dB of shaping gain (for asymptotically large N), independent of the channel coding scheme.[1] However, the performance improvement from shaping gain is offset by the corresponding complexity of the constellation map, which grows exponentially with N. The size of the transmit constellation is determined by the average power constraint and does not affect the shaping or coding gain.

The channel coding gain results from a selection of sequences among all possible sequences of signal points. If we consider a sequence of N input bits as a point in N-dimensional space (the *sequence space*), then this selection is used to guarantee some minimum distance $d_{\min}$ in the sequence space between possible input sequences. Errors generally occur when a sequence is mistaken for its closest neighbor, and in AWGN channels this error probability is a decreasing function of $d^2_{\min}$. We can thus decrease the BER by increasing the separation between each point in the sequence space by a fixed amount ("stretching" the space). However, this will result in a proportional power increase, so no net coding gain is realized. The effective power gain of the channel code is therefore the minimum squared distance between selected sequence points (the sequence points obtained through coding) multiplied by the density of the selected sequence points. Specifically, if the minimum distance and density of all points in the sequence space are denoted by d_0 and $\triangle_0$, respectively, and if the minimum distance and density of points in the sequence space selected through coding are denoted by $d_{\min}$ and $\triangle$, respectively, then maximum likelihood sequence detection yields a channel coding gain of

$$G_c = \left(\frac{d^2_{\min}}{d^2_0}\right)\left(\frac{\triangle}{\triangle_0}\right). \tag{8.77}$$

The second term on the right side of this expression is also referred to as the *constellation expansion factor* and equals 2^{-r} (per N dimensions) for a redundancy of r bits in the encoder E [27].

Some of the nominal coding gain in (8.77) is lost owing to selected sequences having more than one nearest neighbor in the sequence space, which increases the possibility of incorrect sequence detection. This loss in coding gain is characterized by the *error coefficient*, which is tabulated for most common lattice and trellis coded modulations in [27]. In general, the error coefficient is larger for lattice codes than for trellis codes with comparable values of G_c.

Channel coding is done using set partitioning of lattices. A *lattice* is a discrete set of vectors in real Euclidean N-dimensional space that forms a group under ordinary vector addition, so the sum or difference of any two vectors in the lattice is also in the lattice. A *sublattice* is a subset of a lattice that is itself a lattice. The sequence space for

[1] A square constellation has 0 dB of shaping gain; a circular constellation, which is the geometrical figure with the least average energy for a given area, achieves the maximum shape gain for a given N [27].

uncoded MQAM modulation is just the N-cube,[2] so the minimum distance between points is no different than in the two-dimensional case. By restricting input sequences to lie on a lattice in N-dimensional space that is denser than the N-cube, we can increase $d_{\min}$ while maintaining the same density (or, equivalently, the same average power) in the transmit signal constellation; hence, there is no constellation expansion. The N-cube is a lattice, but for every $N > 1$ there are denser lattices in N-dimensional space. Finding the densest lattice in N dimensions is a well-known mathematical problem, and it has been solved for all N for which the decoder complexity is manageable.[3] Once the densest lattice is known, we can form partitioning subsets, or *cosets*, of the lattice via translation of any sublattice. The choice of the partitioning sublattice will determine the size of the partition – that is, the number of subsets that the subset selector in Figure 8.17 has to choose from. Data bits are then conveyed in two ways: through the sequence of cosets from which constellation points are selected, and through the points selected within each coset. The density of the lattice determines the distance between points within a coset, while the distance between subset sequences is essentially determined by the binary code properties of the encoder E and its redundancy r. If we let d_p denote the minimum distance between points within a coset and d_s the minimum distance between the coset sequences, then the minimum distance code is $d_{\min} = \min(d_p, d_s)$. The effective coding gain is given by

$$G_c = 2^{-2r/N} d_{\min}^2, \tag{8.78}$$

where $2^{-2r/N}$ is the constellation expansion factor (in two dimensions) from the r extra bits introduced by the binary channel encoder.

Returning to Figure 8.17, suppose we want to send $m = n + r$ bits per dimension, so that an N sequence conveys mN bits. If we use the densest lattice in N-dimensional space that lies within an N-dimensional sphere, where the radius of the sphere is just large enough to enclose 2^{mN} points, then we achieve a total coding gain that combines the coding gain (resulting from the lattice density and the encoder properties) with the shaping gain of the N-dimensional sphere over the N-dimensional rectangle. Clearly, the coding gain is decoupled from the shaping gain. An increase in signal power would allow us to use a larger N-dimensional sphere and hence transmit more uncoded bits. It is possible to generate maximum-density N-dimensional lattices for $N = 4, 8, 16$, and 24 using a simple partition of the two-dimensional rectangular lattice combined with either conventional block or convolutional coding. Details of this type of code construction, and the corresponding decoding algorithms, can be found in [28] for both lattice and trellis codes. For these constructions, an effective coding gain of approximately 1.5, 3.0, 4.5, and 6.0 dB is obtained with lattice codes for $N = 4, 8, 16$, and 24, respectively. Trellis codes exhibit higher coding gains with comparable complexity.

We conclude this section with an example of coded modulation: the $N = 8$, 3-dB–gain lattice code proposed in [28]. First, the two-dimensional signal constellation is partitioned into four subsets as shown in Figure 8.18, where the subsets are represented by the points A_0, A_1, B_0, and B_1. For example, a 16-QAM constellation would have four subsets, each consisting of four constellation points. Note that the distance between points in each subset is twice the distance between points in the (uncoded) constellation. From this subset partition, we form an 8-dimensional lattice by taking all sequences of four points in which (i) all points are either A-points or B-points and (ii) within a four-point sequence, the point subscripts satisfy the parity check $i_1 + i_2 + i_3 + i_4 = 0$ (so the sequence subscripts must be codewords in the (4, 3) parity-check code, which has a minimum Hamming distance of 2). Thus, three data bits and one parity check bit are used to determine the lattice subset. The square of the minimum distance resulting from this subset partition is four times that of the uncoded signal constellation, yielding a 6-dB gain. However, the extra parity check bit expands the constellation by 1/2 bit per dimension, which by Section 5.3.3 costs an additional power factor of $4^{.5} = 2$, or 3 dB. Thus, the net coding gain is $6 - 3 = 3$ dB. The remaining data bits are used to choose a point within the selected subset and so, for a data rate of m bits per symbol, the four

[2]The Cartesian product of two-dimensional rectangular lattices with points at odd integers.

[3]The complexity of the maximum likelihood decoder implemented with the Viterbi algorithm is roughly proportional to N.

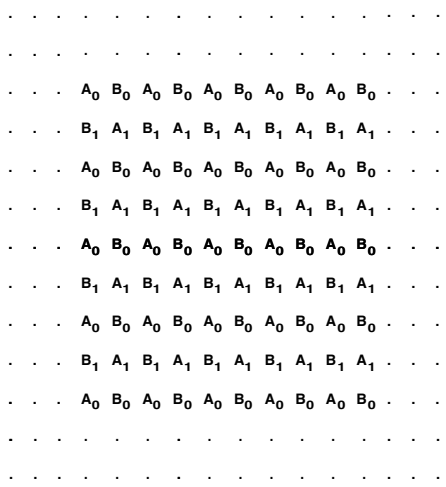

Figure 8.18: Subset partition for an 8-dimensional lattice.

lattice subsets must each have 2^{m-1} points.[4] For example, with a 16-QAM constellation the four subsets each have $2^{m-1} = 4$ points, so the data rate is 3 data bits per 16-QAM symbol.

Coded modulation using turbo codes has also been investigated [29, 30, 31]. This work shows that turbo trellis coded modulation can come very close to the Shannon limit for non-binary signaling.

8.8 Coding with Interleaving for Fading Channels

Block, convolutional, and coded modulation are designed for good performance in AWGN channels. In fading channels, errors associated with the demodulator tend to occur in bursts, corresponding to the times when the channel is in a deep fade. Most codes designed for AWGN channels cannot correct for the long bursts of errors exhibited in fading channels. Hence, codes designed for AWGN channels can exhibit worse performance in fading than an uncoded system.

To improve performance of coding in fading channels, coding is typically combined with *interleaving* to mitigate the effect of error bursts. The basic premise of coding and interleaving is to spread error bursts due to deep fades over many codewords so that each received codeword exhibits at most a few simultaneous symbol errors, which can be corrected for. The spreading out of burst errors is accomplished by the interleaver and the error correction is accomplished by the code. The size of the interleaver must be large enough that fading is independent across a received codeword. Slowly fading channels require large interleavers, which in turn can lead to large delays.

Coding and interleaving is a form of diversity, and performance of coding and interleaving is often characterized by the diversity order associated with the resulting probability of error. This diversity order is typically a function of the minimum Hamming distance of the code. Thus, designs for coding and interleaving on fading channels must focus on maximizing the diversity order of the code rather than on metrics like Euclidean distance, which are used as a performance criterion in AWGN channels. In the following sections we discuss coding and interleaving for block, convolutional, and coded modulation in more detail. We will assume that the receiver has knowledge of the channel fading, which greatly simplifies both the analysis and the decoder. Estimates of channel fading are commonly obtained through pilot symbol transmissions [32, 33]. Maximum likelihood detection of coded signals in fading without this channel knowledge is computationally intractable [34] and so usually requires

[4] This yields $m - 1$ bits/symbol, with the additional bit/symbol conveyed by the channel code.

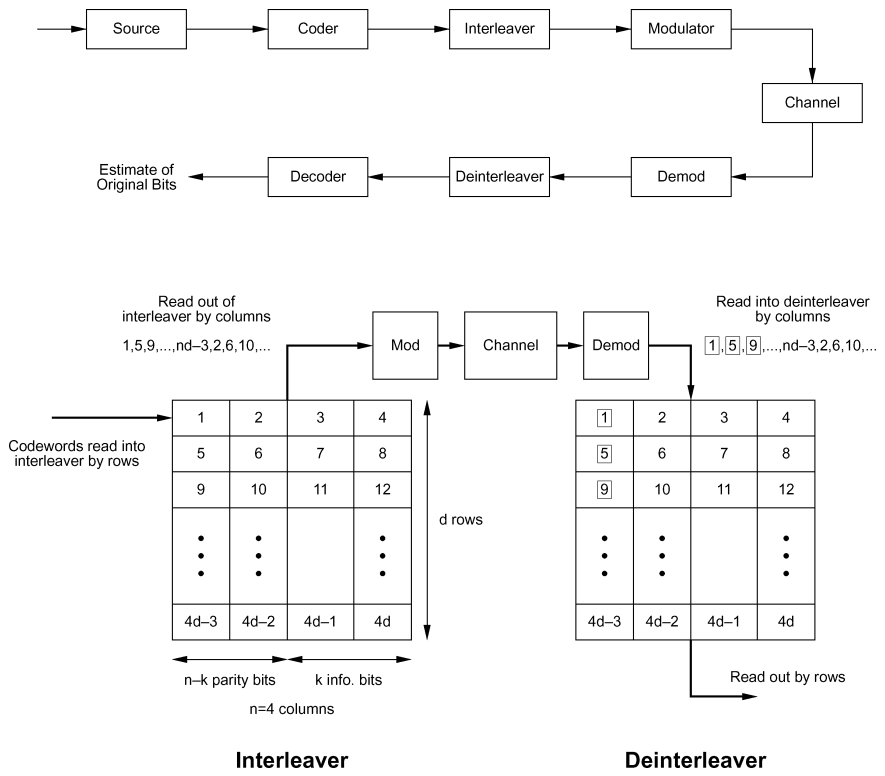

Figure 8.19: The interleaver/deinterleaver operation.

approximations to either the ML decoding metric or the channel [34, 35, 36, 37, 38, 39]. Note that turbo codes designed for AWGN channels, described in Section 8.5, have an interleaver inherent to the code design. However, the interleaver design considerations for AWGN channels are different than for fading channels. A discussion of interleaver design and performance analysis for turbo codes in fading channels can be found in [40, Chap. 8; 41; 42]. Low-density parity-check codes can also be designed for inherent diversity against fading, with performance similar to that of turbo codes optimized for fading diversity [43].

8.8.1 Block Coding with Interleaving

Block codes are typically combined with block interleaving to spread out burst errors from fading. A block interleaver is an array with d rows and n columns, as shown in Figure 8.19. For block interleavers designed for an (n, k) block code, codewords are read into the interleaver by rows so that each row contains an (n, k) codeword. The interleaver contents are read out by columns into the modulator for subsequent transmission over the channel. During transmission, codeword symbols in the same codeword are separated by $d - 1$ other symbols, so symbols in the same codeword experience approximately independent fading if their separation in time is greater than the channel coherence time – that is, if $dT_s > T_c \approx 1/B_D$, where T_s is the codeword symbol duration, T_c is the channel coherence time, and B_D is the channel Doppler spread. An interleaver is called a *deep interleaver* if the condition $dT_s > T_c$ is satisfied. The deinterleaver is an array identical to the interleaver. Bits are read into the deinterleaver from the demodulator by column so that each row of the deinterleaver contains a codeword (whose bits may have been corrupted by the channel). The deinterleaver output is read into the decoder by rows, one codeword at a time.

Figure 8.19 illustrates the ability of coding and interleaving to correct for bursts of errors. Suppose our coding

scheme is an (n, k) binary block code with error correction capability $t = 2$. If this codeword is transmitted through a channel with an error burst of three symbols, then three out of four of the codeword symbols will be received in error. Since the code can only correct two or fewer errors, the codeword will be decoded in error. However, if the codeword is put through an interleaver then, as shown in Figure 8.19, the error burst of three symbols will be spread out over three separate codewords. Since a single symbol error can be easily corrected by an (n, k) code with $t = 2$, the original information bits can be decoded without error. Convolutional interleavers are similar in concept to block interleavers and are better suited to convolutional codes, as will be discussed in Section 8.8.2.

Performance analysis of coding and interleaving requires pairwise error probability analysis or the application of Chernoff or union bounds. Details of this analysis can be found in [1, Chap. 14.6]. The union bound provides a simple approximation to performance. Assume a Rayleigh fading channel with deep interleaving such that each coded symbol fades independently. Then the union bound for an (n, k) block code with soft decision decoding under noncoherent FSK modulation yields a codeword error given as

$$P_e < (2^k - 1)[4p(1-p)]^{d_{\min}}, \tag{8.79}$$

where $d_{\min}$ is the minimum Hamming distance of the code and

$$p = \frac{1}{2 + R_c \bar{\gamma}_b}. \tag{8.80}$$

Similarly, for slowly fading channels in which a coherent phase reference can be obtained, the union bound on the codeword error probability of an (n, k) block code with soft decision decoding and BPSK modulation yields

$$P_e < 2^k \binom{2d_{\min} - 1}{d_{\min}} \left(\frac{1}{4R_c \bar{\gamma} b}\right)^{d_{\min}}. \tag{8.81}$$

Note that both (8.79) and (8.81) are similar to the formula for error probability under MRC diversity combining given by (7.30), with $d_{\min}$ providing the diversity order. Similar formulas apply for hard decoding, with diversity order reduced by a factor of 2 relative to soft decision decoding. Thus, designs for block coding and interleaving over fading channels optimize their performance by maximizing the Hamming distance of the code.

Coding and interleaving is a suboptimal coding technique, since the correlation of the fading that affects subsequent bits contains information about the channel that could be used in a true maximum likelihood decoding scheme. By essentially throwing away this information, the inherent capacity of the channel is decreased [44]. Despite this capacity loss, coding with interleaving using codes designed for AWGN channels is a common coding technique for fading channels, since the complexity required for maximum likelihood decoding on correlated coded symbols is prohibitive.

Example 8.8: Consider a Rayleigh fading channel with a Doppler of $B_D = 80\,\text{Hz}$. The system uses a (5,2) Hamming code with interleaving to compensate for the fading. If the codeword symbols are sent through the channel at 30 kbps, find the required interleaver depth needed to obtain independent fading on each symbol. What is the longest burst of codeword symbol errors that can be corrected and the total interleaver delay for this depth?
Solution: The (5, 2) Hamming code has a minimum distance of 3, so it can correct $t = \lfloor .5 \cdot 3 \rfloor = 1$ codeword symbol error. The codeword symbols are sent through the channel at a rate $R_s = 30$ kbps, so the symbol time is $T_s = 1/R_s = 3.3 \cdot 10^{-5}$. Assume a coherence time for the channel of $T_c = 1/B_D = .0125$ s. The bits in the interleaver are separated by dT_s, so we require $dT_s \geq T_c$ for independent fading on each codeword symbol. Solving for d yields $d \geq T_c/T_s = 375$. Since the interleaver spreads a burst of errors over the depth d of the interleaver, a burst of d symbol errors in the interleaved codewords will result in just one symbol error per codeword after

deinterleaving, which can be corrected. The system can therefore tolerate an error burst of 375 symbols. However, all rows of the interleaver must be filled before it can read out by columns, so the total delay of the interleaver is $ndT_s = 5 \cdot 375 \cdot 3.3 \cdot 10^{-5} = 62.5$ ms. This delay can degrade quality in a voice system. We thus see that the price paid for correcting long error bursts through coding and interleaving is significant delay.

8.8.2 Convolutional Coding with Interleaving

As with block codes, convolutional codes suffer performance degradation in fading channels because the code is not designed to correct for bursts of errors. Thus, it is common to use an interleaver to spread out error bursts. In block coding the interleaver spreads errors across different codewords. Since there is no similar notion of a codeword in convolutional codes, a slightly different interleaver design is needed to mitigate the effect of burst errors. The interleaver commonly used with convolutional codes, called a *convolutional interleaver*, is designed both to spread out burst errors and to work well with the incremental nature of convolutional code generation [45, 46].

A block diagram for a convolutional interleaver is shown in Figure 8.20. The encoder output is multiplexed into buffers of increasing size, from no buffering to a buffer of size $N-1$. The channel input is similarly multiplexed from these buffers into the channel. The reverse operation is performed at the decoder. Thus, the convolutional interleaver delays the transmission through the channel of the encoder output by progressively larger amounts, and this delay schedule is reversed at the receiver. This interleaver takes sequential outputs of the encoder and separates them by $N-1$ other symbols in the channel transmission, thereby breaking up burst errors in the channel. Note that a convolutional encoder can also be used with a block code, but it is most commonly used with a convolutional code. The total memory associated with the convolutional interleaver is $.5N(N-1)$ and the delay is $N(N-1)T_s$ [2, Chap. 8.2.2], where T_s is the symbol time for transmitting the coded symbols over the channel.

The probability of error analysis for convolutional coding and interleaving is given in [1, Chap. 14.6] under assumptions similar to those used in our block fading analysis. The Chernoff bound again yields a probability of error under soft decision decoding with a diversity order based on the minimum free distance of the code. Hard decision decoding reduces this diversity by a factor of 2.

Example 8.9: Consider a channel with coherence time $T_c = 12.5$ ms and a coded bit rate of $R_s = 100$ kilosymbols per second. Find the average delay of a convolutional interleaver that achieves independent fading between subsequent coded bits.

Solution: For the convolutional interleaver, each subsequent coded bit is separated by NT_s and we require $NT_s \geq T_c$ for independent fading, where $T_s = 1/R_s$. Thus we have $N \geq T_c/T_s = .0125/.00001 = 1250$. Note that

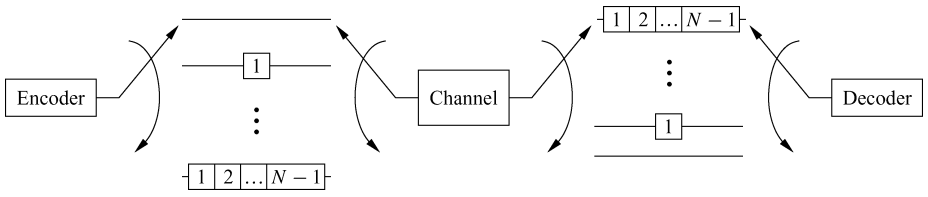

Figure 8.20: Convolutional coding and interleaving.

this is the same as the required depth for a block interleaver to get independent fading on each coded bit. The total delay is $N(N-1)T_s = 15$ s. This is quite a high delay for either voice or data.

8.8.3 Coded Modulation with Symbol/Bit Interleaving

As with block and convolutional codes, coded modulation designed for an AWGN channel performs poorly in fading. This leads to the notion of coded modulation with interleaving for fading channels. However, unlike block and convolutional codes, there are two options for interleaving in coded modulation. One option is to interleave the bits and then map them to modulated symbols. This is called bit-interleaved coded modulation (BICM). Alternatively, the modulation and coding can be done jointly as in coded modulation for AWGN channels and the resulting symbols interleaved prior to transmission. This technique is called symbol-interleaved coded modulation (SICM).

Symbol-interleaved coded modulation seems at first like the best approach because it preserves joint coding and modulation, the main design premise behind coded modulation. However, the coded modulation design criterion must be changed in fading, since performance in fading depends on the code diversity as characterized by its Hamming distance rather than its Euclidean distance. Initial work on coded modulation for fading channels focused on techniques to maximize diversity in SICM. However, good design criteria were hard to obtain, and the performance of these codes was somewhat disappointing [47, 48, 49].

A major breakthrough in the design of coded modulation for fading channels was the discovery of bit-interleaved coded modulation [50, 51]. In BICM the code diversity is equal to the smallest number of distinct bits (rather than channel symbols) along any error event. This is achieved by bitwise interleaving at the encoder output prior to symbol mapping, with an appropriate soft decision bit metric as an input to the Viterbi decoder. Although this breaks the coded modulation paradigm of joint modulation and coding, it provides much better performance than SICM. Moreover, analytical tools for evaluating the performance of BICM as well as design guidelines for good performance are known [50]. BICM is now the dominant technique for coded modulation in fading channels.

8.9 Unequal Error Protection Codes

When not all bits transmitted over the channel have the same priority or bit error probability requirement, multiresolution or unequal error protection (UEP) codes can be used. This scenario arises, for example, in voice and data systems where voice is typically more tolerant to bit errors than data: data received in error must be retransmitted, so $P_b < 10^{-6}$ is typically required, whereas good quality voice requires only on the order of $P_b < 10^{-3}$. This scenario also arises for certain types of compression. For example, in image compression, bits corresponding to the low-resolution reproduction of the image are required, whereas high-resolution bits simply refine the image. With multiresolution channel coding, all bits are received correctly with a high probability under benign channel conditions. However, if the channel is in a deep fade, only the high-priority bits or those requiring low P_b will be received correctly with high probability.

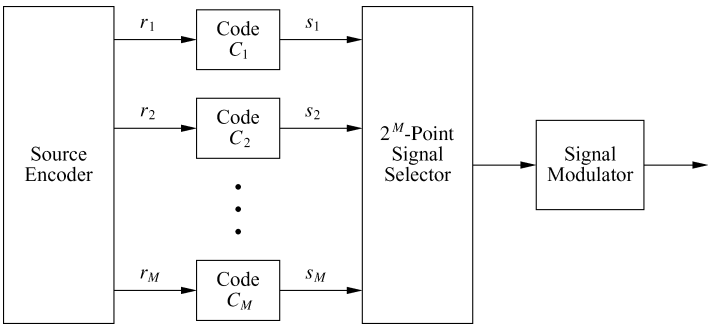

Figure 8.21: Multilevel encoder.

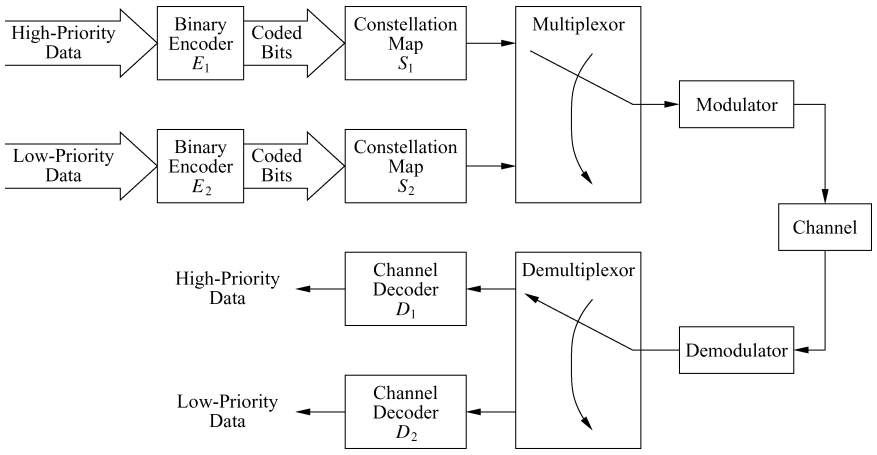

Figure 8.22: Transceiver for time-multiplexed coded modulation.

Practical implementation of a UEP code was first studied by Imai and Hirakawa [52]. Binary UEP codes were later considered for combined speech and channel coding [53] as well as for combined image and channel coding [54]. These implementations use traditional (block or convolutional) error correction codes, so coding gain is directly proportional to bandwidth expansion. Subsequently, two bandwidth-efficient implementations for UEP codes were proposed: time multiplexing of bandwidth-efficient coded modulation [55], and coded modulation techniques applied to both uniform and nonuniform signal constellations [56, 57]. All of these multilevel codes can be designed for either AWGN or fading channels. We now briefly summarize these UEP coding techniques; specifically, we describe the principles behind multilevel coding and multistate decoding as well as the more complex bandwidth-efficient implementations.

A block diagram of a general multilevel encoder is shown in Figure 8.21. The source encoder first divides the information sequence into M parallel bit streams of decreasing priority. The channel encoder consists of M different binary error correcting codes $C_1, \ldots, C_M$ with decreasing codeword distances. The ith-priority bit stream enters the ith encoder, which generates the coded bits s_i. If the 2^M points in the signal constellation are numbered from 0 to $2^M - 1$, then the point selector chooses the constellation point s corresponding to

$$s = \sum_{i=1}^{M} s_i \cdot 2^{i-1}. \tag{8.82}$$

For example, if $M = 3$ and the signal constellation is 8-PSK, then the chosen signal point will have phase $2\pi s/8$.

Optimal decoding of the multilevel code uses a maximum likelihood decoder, which determines the input sequence that maximizes the received sequence probability. The ML decoder must therefore jointly decode the code sequences $\{s_1\}, \ldots, \{s_m\}$. This can entail significant complexity even if the individual codes in the multilevel code have low complexity. For example, if the component codes are convolutional codes with 2^{μ_i} states, $i = 1, \ldots, M$, then the number of states in the optimal decoder is $2^{\mu 1 + \cdots + \mu_M}$. Because of the high complexity of optimal decoding, the suboptimal technique of multistage decoding, introduced in [52], is used for most implementations. Multistage decoding is accomplished by decoding the component codes sequentially. First the most powerful code, C_1, is decoded, then C_2, and so forth. Once the code sequence corresponding to encoder C_i is estimated, it is assumed to be correct for code decisions on the weaker code sequences.

The binary encoders of this multilevel code require extra code bits to achieve their coding gain, so they are not bandwidth efficient. An alternative approach proposed in [56] uses time multiplexing of the coded modulations described in Section 8.7. In this approach, different lattice or trellis coded modulations with different coding gains

are used for each priority class of input data. The transmit signal constellations corresponding to each encoder may differ in size (number of signal points), but the average power of each constellation is the same. The signal points output by each of the individual encoders are then time-multiplexed together for transmission over the channel, as shown in Figure 8.22 for two bit streams of different priority. Let R_i denote the bit rate of encoder C_i in this figure for $i = 1, 2$. If T_1 equals the fraction of time that the high-priority C_1 code is transmitted and if T_2 equals the fraction of time that the C_2 code is transmitted, then the total bit rate is $(R_1T_1 + R_2T_2)/(T_1 + T_2)$, with the high-priority bits constituting $R_1T_1/(R_1T_1 + R_2T_2)$ percent of this total.

The time-multiplexed coding method yields a higher gain if the constellation maps S_1 and S_2 of Figure 8.22 are designed jointly. This revised scheme is shown in Figure 8.23 for two encoders, where the extension to M encoders is straightforward. In trellis and lattice coded modulation, recall that bits are encoded to select the lattice subset and that uncoded bits choose the constellation point within the subset. The binary encoder properties reduce the P_b for the encoded bits only; the P_b for the uncoded bits is determined by the separation of the constellation signal points. We can easily modify this scheme to yield two levels of coding gain, where the high-priority bits are heavily encoded and are used to choose the subset of the partitioned constellation, while the low-priority bits are uncoded or lightly coded and are used to select the constellation signal point.

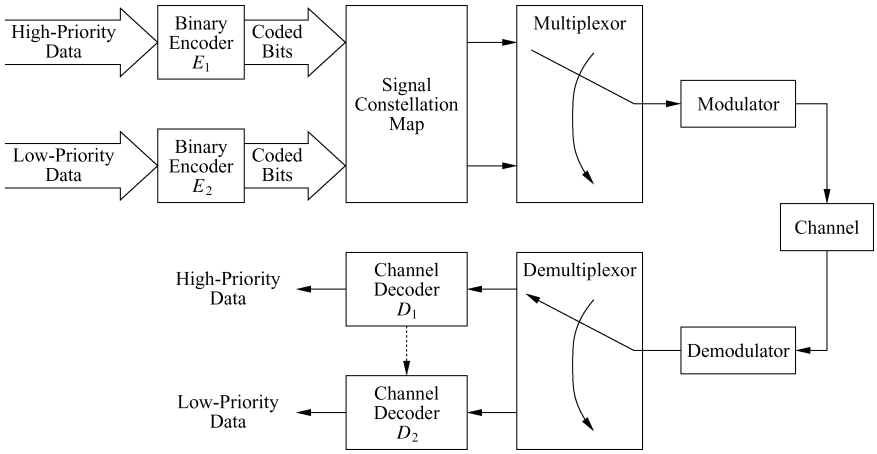

Figure 8.23: Joint optimization of signal constellation.

8.10 Joint Source and Channel Coding

The underlying premise of UEP codes is that the bit error probabilities of the channel code should be matched to the priority or P_b requirements associated with the bits to be transmitted. These bits are often taken from the output of a compression algorithm acting on the original data source. Hence, UEP coding can be considered as a joint design between compression (also called *source coding*) and channel coding. Although Shannon determined that the source and channel codes can be designed separately on an AWGN channel with no loss in optimality [58], this result holds only in the limit of infinite source code dimension, infinite channel code blocklength, and infinite complexity and delay. Thus, there has been much work on investigating the benefits of joint source and channel coding under more realistic system assumptions.

Work in the area of joint source and channel coding falls into several broad categories: source-optimized channel coding; channel-optimized source coding; and iterative algorithms, which combine these two code designs. In source-optimized channel coding, the source code is designed for a noiseless channel. A channel code is then designed for this source code to minimize end-to-end distortion over the given channel based on the distortion

associated with corruption of the different transmitted bits. Unequal error protection channel coding – where the P_b of the different component channel codes is matched to the bit priorities associated with the source code – is an example of this technique. Source-optimized channel coding has been applied to image compression with convolution channel coding and with rate-compatible punctured convolutional (RCPC) channel codes in [54, 59, 60]. A comprehensive treatment of matching RCPC channel codes or MQAM to subband and linear predictive speech coding – in both AWGN and Rayleigh fading channels – can be found in [61]. In source-optimized modulation, the source code is designed for a noiseless channel and then the modulation is optimized to minimize end-to-end distortion. An example of this approach is given in [62], where compression by a vector quantizer (VQ) is followed by multicarrier modulation, and the modulation provides unequal error protection to the different source bits by assigning different powers to each subcarrier.

Channel-optimized source coding is another approach to joint source and channel coding. In this technique the source code is optimized based on the error probability associated with the channel code, where the channel code is designed independently of the source. Examples of work taking this approach include the channel-optimized vector quantizer (COVQ) and its scalar variation [63, 64]. Source-optimized channel coding and modulation can be combined with channel-optimized source coding via an iterative design. This approach is used for the joint design of a COVQ and multicarrier modulation in [65] and for the joint design of a COVQ and RCPC channel code in [66]. Combined trellis coded modulation and *trellis coded quantization*, a source coding strategy that borrows from the basic premise of trellis coded modulation, is investigated in [67, 68]. All of this work on joint source and channel code design indicates that significant performance advantages are possible when the source and channel codes are jointly designed. Moreover, many sophisticated channel code designs, such as turbo and LDPC codes, have not yet been combined with source codes in a joint optimization. Thus, much more work is needed in the broad area of joint source and channel coding to optimize performance for different applications.

Chapter 8 Problems

1. Consider a $(3, 1)$ linear block code where each codeword consists of three data bits and one parity bit.

   (a) Find all codewords in this code.

   (b) Find the minimum distance of the code.

2. Consider a $(7, 4)$ code with generator matrix

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

   (a) Find all the codewords of the code.

   (b) What is the minimum distance of the code?

   (c) Find the parity check matrix of the code.

   (d) Find the syndrome for the received vector $\mathbf{R} = [1101011]$.

   (e) Assuming an information bit sequence of all 0s, find all minimum weight error patterns $\mathbf{e}$ that result in a valid codeword that is not the all-zero codeword.

   (f) Use row and column operations to reduce $\mathbf{G}$ to systematic form and find its corresponding parity-check matrix. Sketch a shift register implementation of this systematic code.

3. All Hamming codes have a minimum distance of 3. What are the error correction and error detection capabilities of a Hamming code?

4. The $(15, 11)$ Hamming code has generator polynomial $g(X) = 1 + X + X^4$. Determine if the codewords described by polynomials $c_1(X) = 1 + X + X^3 + X^7$ and $c_2(X) = 1 + X^3 + X^5 + X^6$ are valid codewords for this generator polynomial. Also find the systematic form of this polynomial $p(X) + X^{n-k}u(X)$ that generates the codewords in systematic form.

5. The $(7, 4)$ cyclic Hamming code has a generator polynomial $g(X) = 1 + X^2 + X^3$.

   (a) Find the generator matrix for this code in systematic form.

   (b) Find the parity-check matrix for the code.

   (c) Suppose the codeword $\mathbf{C} = [1011010]$ is transmitted through a channel and the corresponding received codeword is $\mathbf{C} = [1010011]$. Find the syndrome polynomial associated with this received codeword.

   (d) Find all possible received codewords such that, for the transmitted codeword $\mathbf{C} = [1011010]$, the received codeword has a syndrome polynomial of zero.

6. The weight distribution of a Hamming code of blocklength n is given by

$$N(x) = \sum_{i=0}^{n} N_i x^i = \frac{1}{n+1}[(1+x)^n + n(1+x)^{.5(n-1)}(1-x)^{.5(n+1)}],$$

   where N_i denotes the number of codewords of weight i.

(a) Use this formula to determine the weight distribution of a Hamming $(7, 4)$ code.

(b) Use the weight distribution from part (a) to find the union upper bound based on weight distribution (8.38) for a Hamming $(7, 4)$ code, assuming BPSK modulation of the coded bits with an SNR of 10 dB. Compare with the probability of error from the looser bound (8.39) for the same modulation.

7. Find the union upper bound on probability of codeword error for a Hamming code with $m = 7$. Assume the coded bits are transmitted over an AWGN channel using 8-PSK modulation with an SNR of 10 dB. Compute the probability of bit error for the code assuming a codeword error corresponds to one bit error, and compare with the bit error probability for uncoded modulation.

8. Plot P_b versus γ_b for a $(24, 12)$ linear block code with $d_{\min} = 8$ and $0 \leq E_b/N_0 \leq 20$ dB using the union bound for probability of codeword error. Assume that the coded bits are transmitted over the channel using QPSK modulation. Over what range of E_b/N_0 does the code exhibit negative coding gain?

9. Use (8.47) to find the approximate coding gain of a $(7, 4)$ Hamming code with SDD over uncoded modulation, assuming $\gamma_b = 15$ dB.

10. Plot the probability of codeword error for a $(24, 12)$ code with $d_{\min} = 8$ for $0 \leq \gamma_b \leq 10$ dB under both hard and soft decoding, using the union bound (8.36) for hard decoding and the approximation (8.44) for soft decoding. What is the difference in coding gain at high SNR for the two decoding techniques?

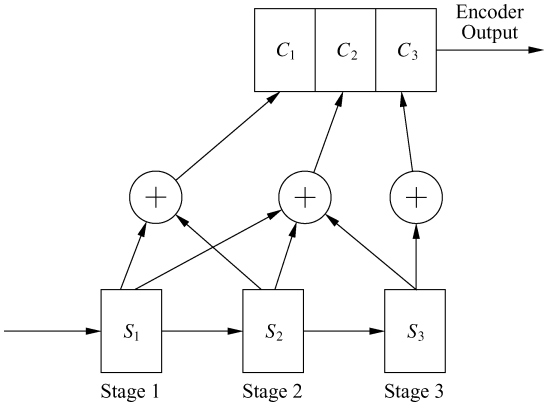

Figure 8.24: Convolutional encoder for Problems 8-14 and 8-15.

11. Evalute the upper and lower bounds on codeword error probability, (8.35) and (8.36) respectively, for an extended Golay code with HDD, assuming an AWGN channel with BPSK modulation and an SNR of 10 dB.

12. Consider a Reed Solomon code with $k = 3$ and $K = 4$, mapping to 8-PSK modulation. Find the number of errors that can be corrected with this code and its minimum distance. Also find its probability of bit error assuming the coded symbols transmitted over the channel via 8-PSK have a symbol error probability of 10^{-3}.

13. For the trellis of Figure 8.7, determine the state sequence and encoder output assuming an initial state $S = 00$ and information bit sequence $\mathbf{U} = [0110101101]$.

14. Consider the convolutional code generated by the encoder shown in Figure 8.24.

(a) Sketch the trellis diagram of the code.

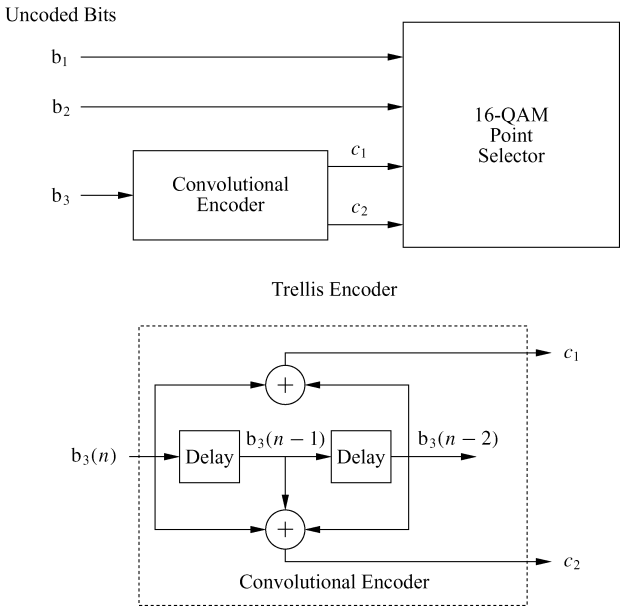

Figure 8.25: 16-QAM trellis encoder for Problem 8-16.

(b) For a received sequence $\mathbf{R} = [001010001]$, find the path metric for the all-zero path assuming probability of symbol error $p = 10^{-3}$.

(c) Find one path at a minimum Hamming distance from the all-zero path and compute its path metric for the same $\mathbf{R}$ and p as in part (b).

15. This problem is based on the convolutional encoder of Figure 8.24.

(a) Draw the state diagram for this convolutional encoder.

(b) Determine its transfer function $T(D,\ N,\ J)$.

(c) Determine the minimum distance of paths through the trellis to the all-zero path.

(d) Compute the upper bound (8.72) on probability of bit error for this code assuming SDD and BPSK modulation with $\gamma_b = 10$ dB.

(e) Compute the upper bound (8.76) on probability of bit error for this code assuming HDD and BPSK modulation with $\gamma_b = 10$ dB. How much coding gain is achieved with soft versus hard decoding?

16. Suppose you have a 16-QAM signal constellation that is trellis encoded via the scheme of Figure 8.25. Assume the set partitioning for 16-QAM shown in Figure 8.18.

(a) Assuming that parallel transitions–that is, transitions within a subset–dominate the error probability, find the coding gain of this trellis code relative to uncoded 8-PSK, given that d_0 for the 16-QAM is .632 and for the 8-PSK is .765.

(b) Draw the trellis for the convolutional encoder and assign subsets to the trellis transitions according to the following heuristic rules of Ungerboeck: (i) if k bits are to be encoded per modulation interval then the trellis must allow for $2k$ possible transitions from each state to the next state; (ii) more than one transition may occur between pairs of states; (iii) all waveforms should occur with equal frequency

and with a fair amount of regularity and symmetry; (iv) transitions originating from the same state are assigned waveforms either from the A or B subsets but not from both; (v) transitions entering into the same state are assigned waveforms either from the A or B subsets but not from both; and (vi) parallel transitions are assigned from only one subset.

(c) What is the minimum distance error event through the trellis relative to the path generated by the all-zero bit stream?

(d) Assuming that your answer to part (c) is the minimum distance error event for the trellis, what is $d_{\min}$ of the code?

17. Consider a channel with coherence time $T_c = 10$ ms and a coded bit rate of $R_s = 50$ kilosymbols per second. Find the average delay of a convolutional interleaver that achieves independent fading between subsequent coded bits. Also find the memory requirements of this system.

18. In a Rayleigh fading channel, determine an upper bound for the bit error probability P_b of a Golay $(23, 12)$ code with deep interleaving $(dT_s \gg T_c)$, BPSK modulation, soft decision decoding, and an average coded E_c/N_0 of 15 dB. Compare with the uncoded P_b in Rayleigh fading.

19. Consider a Rayleigh fading channel with BPSK modulation, average SNR of 10 dB, and a Doppler of 80 Hz. The data rate over the channel is 30 kbps. Assume that bit errors occur on this channel whenever $P_b(\gamma) \geq 10^{-2}$. Design an interleaver and associated $(n,\ k)$ block code that corrects essentially all of the bit errors, where the interleaver delay is constrained to be less than 50 ms. Your design should include the dimensions of the interleaver as well as the block code type and the values of n and k.

20. Assume a multilevel encoder as in Figure 8.21, where the information bits have three different error protection levels $(M = 3)$ and the three encoder outputs are modulated using 8-PSK modulation with an SNR of 10 dB. Assume the code C_i associated with the ith bit stream b_i is a Hamming code with parameter m_i, where $m_1 = 2, m_2 = 3$, and $m_3 = 4$.

(a) Find the probability of error for each Hamming code C_i, assuming it is decoded individually using HDD.

(b) If the symbol time of the 8-PSK modulation is $T_s = 10\ \mu$s, what is the data rate for each of the three bit streams?

(c) For what size code must the maximum likelihood decoder of this UEP code be designed?

21. Design a two-level UEP code using either Hamming or Golay codes and such that, for a channel with an SNR of 10 dB, the UEP code has $P_b = 10^{-3}$ for the low-priority bits and $P_b = 10^{-6}$ for the high priority bits.

Bibliography

[1] J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.

[2] B. Sklar, *Digital Communications–Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] D. G. Wilson, *Digital Modulation and Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[4] S. Lin and J. D. J. Costello, *Error Control Coding*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2004.

[5] A. J. Vlterbi, "Error bounds for convolutional codes and asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, pp. 260–9, April 1967.

[6] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," *Proc. IEEE Intemat. Conf. Commun*., pp. 54–8, May 1993.

[7] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. Commun*., pp. 1261–71, October 1996.

[8] *IEEE Trans. Inform. Theory*, Special Issue on Codes and Graphs and Iterative Algorithms, Febmary 2001.

[9] S. B. Wicker and S. Kim, *Codes, Graphs, and Iterative Decoding*, Kluwer, Boston, 2002.

[10] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inform. Theory*, pp. 21–8, January 1962.

[11] C. Heegard and S. B. Wicker, *Turbo Coding*, Kluwer, Boston, 1999.

[12] M. E. Buckley and S. B. Wicker, "The design and performance of a neural network for predicting decoder error in turbo-coded ARQ protocols," *IEEE Trans. Commun*., pp. 566–76, April 2000.

[13] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial concatenation ofinterleaved codes: Performance analysis, design and iterative decoding," *IEEE Trans. Inform. Theory*, pp. 909–26, May 1998.

[14] H. Jin and R. J. McEliece, "Coding theorems for turbo code ensembles," *IEEE Trans. Inform. Theory*, pp. 1451–61, June 2002.

[15] I. Sasan and S. Shamai, "Improved upper bounds on the ML decoding error probability of parallel and serial concatenated turbo codes via their ensemble distance spectrum," *IEEE Trans. Inform. Theory*, pp. 24–47, January 2000.

[16] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Elec. Lett*., p. 1645, August 1996.

[17] N. Wiberg, N.-A. Loeliger, and R. Kotter, "Codes and iterative decoding on general graphs," *Euro. Trans. Telecommun*., pp. 513–25, June 1995.

[18] T. Richardson, A. Shokrollahi, and R. Urbanke, "Design of capacity-approaching irregular lowdensity parity-check codes," *IEEE Trans. Inform. Theory*, pp. 619–37, Febmary 2001.

[19] R. McEliece, D. J. C. MacKay, and J.-F. Cheng, " Turbo decoding as an instance of Pearl's' belief propagation' algorithm," *IEEE J. Sel. Areas Commun*., pp. 140–52, Febmary 1998.

[20] F. R. Kschischang and D. Frey, "Iterative decoding of compound codes by probability propagation in graphical models," *IEEE J. Sel. Areas Commun*., pp. 219–30, Febmary 1998.

[21] T. Richardson and R. Urbanke, " The capacity of low-density parity-check codes under message passing decoding," *IEEE Trans. Inform. Theory*, pp. 599–618, February 2001.

[22] S.-Y. Chung, G. D. Fomey, T. Richardson, and R. Urbanke, "On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit," *IEEE Commun. Lett*., pp. 58–60, Febmary 2001.

[23] M. Fossorier, "Iterative reliability-based decoding of low-density parity check codes," *IEEE J. Sel. Areas Commun*., pp. 908–17, May 2001.

[24] M. C. Davey and D. MacKay, "Low density parity-check codes over GF(q)," *IEEE Commun. Lett*., pp. 165–7, June 1998.

[25] J. Hou, P. Siegel, L. Milstein, and H. D. Pfister, "Capacity-approaching bandwidth efficient coded modulation schemes based on low-density parity-check codes," *IEEE Trans. Inform. Theory*, pp. 2141–55, September 2003.

[26] G. Ungerboeck, "Channel coding with multi-level/phase signals," *IEEE Trans. Inform. Theory*, pp. 55–67, January 1982.

[27] G. D. Fomey, "Coset codes, I: Introduction and geomenical classification, and II: Binary lattices and related codes," *IEEE Trans. Inform. Theory*, pp. 1123–87, September 1988.

[28] G. D. Fomey, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff, and S. U. Quereshi, "Efficient modulation for band-limited channels," *IEEE J. Sel. Areas Commun*., pp. 632–47, September 1984.

[29] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Parallel concatenated trellis coded modulation," *Proc. IEEE Intemat. Conf. Commun*., pp. 974–8, June 1996.

[30] C. Fragouli and R. D. Wesel, "Turbo-encoder design for symbol-interleaved parallel concatenated trellis-coded modulation," *IEEE Trans. Commun*., pp. 425–35, March 2001.

[31] P. Robertson and T. Worz, "Bandwidth-efficient turbo trellis-coded modulation using punctured component codes," *IEEE J. Sel. Areas Commun*., pp. 206–18, February 1998.

[32] G. T. Irvine and P. J. Mclane, "Symbol-aided plus decision-directed reception for PSK TCM modulation on shadowed mobile satellite fading channels," *IEEE J. Sel. Areas Commun*., pp. 1289–99, October 1992.

[33] D. Subasinghe-Dias and K. Feher, "A coded 16-QAM scheme for fast fading mobile radio channels," *IEEE Trans. Commun*., pp. 1906–16, February–April 1995.

[34] P. Y. Kam and H. M. Ching, "Sequence estimation over the slow nonselective Rayleigh fading channel with diversity reception and its application to Vlterbi decoding," *IEEE J. Sel. Areas Commun*., pp. 562–70, April 1992.

[35] D. Makrakis, P. T. Mathiopoulos, and D. P. Bouras, "Optimal decoding of coded PSK and QAM signals in correlated fast fading channels and AWGN–A combined envelope, multiple differential and coherent detection approach," *IEEE Trans. Commun*., pp. 63–75, January 1994.

[36] M. J. Gertsman and J. H. Lodge, "Symbol-by-symbol MAP demodulation of CPM and PSK signals on Rayleigh flat-fading channels," *IEEE Trans. Commun*., pp. 788–99, July 1997.

[37] H. Kong and E. Shwedyk, "Sequence detection and channel state estimation over finite state Markov channels," *IEEE Trans. Veh. Tech*., pp. 833–9, May 1999.

[38] G. M. Vitetta and D. P. Taylor, "Maximum-likelihood decoding of uncoded and coded PSK signal sequences transmitted over Rayleigh flat-fading channels," *IEEE Trans. Commun*., pp. 2750–8, November 1995.

[39] L. Li and A. J. Goldsmith, "Low-complexity maximum-likelihood detection of coded signals sent over finite-state Markov channels," *IEEE Trans. Commun*., pp. 524–31, April 2002.

[40] B. Vucetic and J. Yuan, *Turbo Codes*: *Principles and Applications*, Kluwer, Dordrecht, 2000.

[41] E. K. Hall and S. G. Wilson, "Design and analysis of turbo codes on Rayleigh fading channels," *IEEE J. Sel. Areas Commun*., pp. 160–74, February 1998.

[42] C. Komninakis and R. D. Wesel, "Joint iterative channel estimation and decoding in flat correlated Rayleigh fading," *IEEE J. Sel. Areas Commun*. pp. 1706–17, September 2001.

[43] J. Hou, P. H. Siegel, and L. B. Milstein, "Performance analysis and code optimization of lowdensity parity-check codes on Rayleigh fading channels," *IEEE J. Sel. Areas Commun*., pp. 924–34, May 2001.

[44] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inform. Theory*, pp. 868–86, May 1996.

[45] G. D. Fomey, "Burst error correcting codes for the classic bursty channel," *IEEE Trans. Commun. Tech*., pp. 772–81, October 1971.

[46] J. L. Ramsey, "Realization of optimum interleavers," *IEEE Trans. Inform. Theory*, pp. 338–45, 1970.

[47] C.-E. W. Sundberg and N. Seshadri, "Coded modulation for fading channels-An overview," *Euro. Trans. Telecommun*., pp. 309–24, May/ June 1993.

[48] L.-F. Wei, "Coded M-DPSK with built-in time diversity for fading channels," *IEEE Trans. Inform. Theory*, pp. 1820–39, November 1993.

[49] S. H. Jamali and T. Le-Ngoc, *Coded-Modulation Techniques for Fading Channels*, Kluwer, New York, 1994.

[50] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, pp. 927–46, May 1998.

[51] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel," *IEEE Trans. Commun*., pp. 873–84, May 1992.

[52] H. Imai and S. Hirakawa, "A new multilevel coding method using error correcting codes," *IEEE Trans. Inform. Theory*, pp. 371–7, May 1977.

[53] R. V. Cox, J. Hagenauer, N. Seshadri, and C.-E. W. Sundberg, "Variable rate sub-band speech coding and matched convolutional channel coding for mobile radio channels," *IEEE Trans. Signal Proc.*, pp. 1717–31, August 1991.

[54] J. W. Modestino and D. G. Daut, "Combined source-channel coding of images," *IEEE Trans. Commun.*, pp. 1644–59, November 1979.

[55] A. R. Calderbank and N. Seshadri, "Multilevel codes for unequal error protection," *IEEE Trans. Inform. Theory*, pp. 1234–48, July 1993.

[56] L.-F. Wei, "Coded modulation with unequal error protection," *IEEE Trans. Commun.*, pp. 1439–49, October 1993.

[57] N. Seshadri and C.-E. W. Sundberg, "Multilevel trellis coded modulations for the Rayleigh fading channel," *IEEE Trans. Commun.*, pp. 1300–10, September 1993.

[58] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, part 4, pp. 142–63, 1959.

[59] N. Tanabe and N. Farvardin, "Subband image coding using entropy-coded quantization over noisy channels," *IEEE J. Sel. Areas Commun.*, pp. 926–43, June 1992.

[60] H. Jafarkhani, P. Ligdas, and N. Farvardin, "Adaptive rate allocation in ajoint source/channel coding framework for wireless channels," *Proc. IEEE Veh. Tech. Conf.*, pp. 492–6, April 1996.

[61] W. C. Wong, R. Steele, and C.-E. W. Sundberg, *Source-Matched Mobile Communications*, Pentech and IEEE Press, London and New York, 1995.

[62] K.-P. Ho and J. M. Kahn, "Transmission of analog signals using multicarrier modulation: A combined source-channel coding approach," *IEEE Trans. Commun.*, pp. 1432–43, November 1996.

[63] N. Farvardin and V. Vaishampayan, "On the performance and complexity of channel-optimized vector quantizers," *IEEE Trans. Inform. Theory*, pp. 155–60, January 1991.

[64] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Inform. Theory*, pp. 827–38, November 1987.

[65] K.-P. Ho and J. M. Kahn, "Combined source-channel coding using channel-optimized quantizer and multicanier modulation," *Proc. IEEE Intemat. Conf. Commun.*, pp. 1323–7, June 1996.

[66] A. J. Goldsmith and M. Effros, "Joint design of fixed-rate source codes and multiresolution channel codes," *IEEE Trans. Commun.*, pp. 1301–12, October 1998.

[67] E. Ayanoglu and R. M. Gray, "The design ofjoint source and channel trellis waveform coders," *IEEE Trans. Inform. Theory*, pp. 855–65, November 1987.

[68] T. R. Fischer and M. W. Marcellin, "Joint trellis coded quantization/modulation," *IEEE Trans. Commun.*, pp. 172–6, Febmary 1991.

Chapter 9

Adaptive Modulation and Coding

Adaptive modulation and coding enable robust and spectrally efficient transmission over time-varying channels. The basic premise is to estimate the channel at the receiver and feed this estimate back to the transmitter, so that the transmission scheme can be adapted relative to the channel characteristics. Modulation and coding techniques that do not adapt to fading conditions require a fixed link margin to maintain acceptable performance when the channel quality is poor. Thus, these systems are effectively designed for worst-case channel conditions. Since Rayleigh fading can cause a signal power loss of up to 30 dB, designing for the worst-case channel conditions can result in very inefficient utilization of the channel. Adapting to the channel fading can increase average throughput, reduce required transmit power, or reduce average probability of bit error by taking advantage of favorable channel conditions to send at higher data rates or lower power – and by reducing the data rate or increasing power as the channel degrades. In Section 4.2.4 we derived the optimal adaptive transmission scheme that achieves the Shannon capacity of a flat fading channel. In this chapter we describe more practical adaptive modulation and coding techniques to maximize average spectral efficiency while maintaining a given average or instantaneous bit error probability. The same basic premise can be applied to MIMO channels, frequency-selective fading channels with equalization, OFDM or CDMA, and cellular systems. The application of adaptive techniques to these systems will be described in subsequent chapters.

Adaptive transmission was first investigated in the late sixties and early seventies [1, 2]. Interest in these techniques was short-lived, perhaps due to hardware constraints, lack of good channel estimation techniques, and or systems focusing on point-to-point radio links without transmitter feedback. As technology evolved these issues became less constraining, resulting in a revived interest in adaptive modulation methods for 3G wireless systems [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. As a result, many wireless systems – including both GSM and CDMA cellular systems as well as wireless LANs – use adaptive transmission techniques [13, 14, 15, 16].

There are several practical constraints that determine when adaptive modulation should be used. Adaptive modulation requires a feedback path between the transmitter and receiver, which may not be feasible for some systems. Moreover, if the channel is changing faster than it can be reliably estimated and fed back to the transmitter, adaptive techniques will perform poorly. Many wireless channels exhibit variations on different time scales; examples include multipath fading, which can change very quickly, and shadowing, which changes more slowly. Often only the slow variations can be tracked and adapted to, in which case flat fading mitigation is needed to address the effects of multipath. Hardware constraints may dictate how often the transmitter can change its rate and/or power, and this may limit the performance gains possible with adaptive modulation. Finally, adaptive modulation typically varies the rate of data transmission relative to channel conditions. We will see that average spectral efficiency of adaptive modulation under an average power constraint is maximized by setting the data rate to be small or zero in poor channel conditions. However, with this scheme the quality of fixed-rate applications (such as voice or video) with hard delay constraints may be significantly compromised. Thus, in delay-constrained applications

the adaptive modulation should be optimized to minimize outage probability for a fixed data rate [17].

9.1 Adaptive Transmission System

In this section we describe the system associated with adaptive transmission. The model is the same as the model of Section 4.2.1 used to determine the capacity of flat fading channels. We assume linear modulation, where the adaptation takes place at a multiple of the symbol rate $R_s = 1/T_s$. We also assume the modulation uses ideal Nyquist data pulses (sinc$[t/T_s]$), so the signal bandwidth $B = 1/T_s$. We model the flat fading channel as a discrete-time channel in which each channel use corresponds to one symbol time T_s. The channel has stationary and ergodic time-varying gain $\sqrt{g[i]}$ that follows a given distribution $p(g)$ and AWGN $n[i]$, with power spectral density $N_0/2$. Let $\bar{P}$ denote the average transmit signal power, $B = 1/T_s$ the received signal bandwidth, and $\bar{g}$ the average channel gain. The instantaneous received signal-to-noise power ratio (SNR) is then $\gamma[i] = \bar{P}g[i]/N_0B$, $0 \leq \gamma[i] < \infty$, and its expected value over all time is $\bar{\gamma} = \bar{P}\bar{g}/N_0B$. Since $g[i]$ is stationary, the distribution of $\gamma[i]$ is independent of i, and we denote this distribution by $p(\gamma)$.

In adaptive transmission we estimate the power gain or the received SNR at time i and then adapt the modulation and coding parameters accordingly. The most common parameters to adapt are the data rate $R[i]$, transmit power $P[i]$, and coding parameters $C[i]$. For M-ary modulation the data rate $R[i] = \log_2 M[i]/T_s = B\log_2 M[i]$ bps. The *spectral efficiency* of the M-ary modulation is $R[i]/B = \log_2 M[i]$ bps/Hz. We denote the SNR estimate as $\hat{\gamma}[i] = \bar{P}\hat{g}[i]/N_0B$, which is based on the power gain estimate $\hat{g}[i]$. Suppose the transmit power is adapted relative to $\hat{\gamma}[i]$. We denote this adaptive transmit power at time i by $P(\hat{\gamma}[i]) = P[i]$, and the received power at time i is then $\gamma[i]P(\hat{\gamma}[i])/\bar{P}$. Similarly, we can adapt the data rate of the modulation $R(\hat{\gamma}[i]) = R[i]$ and/or the coding parameters $C(\hat{\gamma}[i]) = C[i]$ relative to the estimate $\hat{\gamma}[i]$. When the context is clear, we will omit the time reference i relative to γ, $P(\gamma)$, $R(\gamma)$, and $C(\gamma)$.

The system model is illustrated in Figure 9.1. We assume that an estimate $\hat{g}[i]$ of the channel power gain $g[i]$ at time i is available to the receiver after an estimation time delay of i_e and that this same estimate is available to the transmitter after a combined estimation and feedback path delay of $i_d = i_e + i_f$. The availability of this channel information at the transmitter allows it to adapt its transmission scheme relative to the channel variation. The adaptive strategy may take into account the estimation error and delay in $\hat{g}[i - i_d]$ or it may treat $\hat{g}[i - i_d]$ as the true gain: this issue will be discussed in more detail in Section 9.3.7. We assume that the feedback path does not introduce any errors, which is a reasonable assumption if strong error correction and detection codes are used on the feedback path and if packets associated with detected errors are retransmitted.

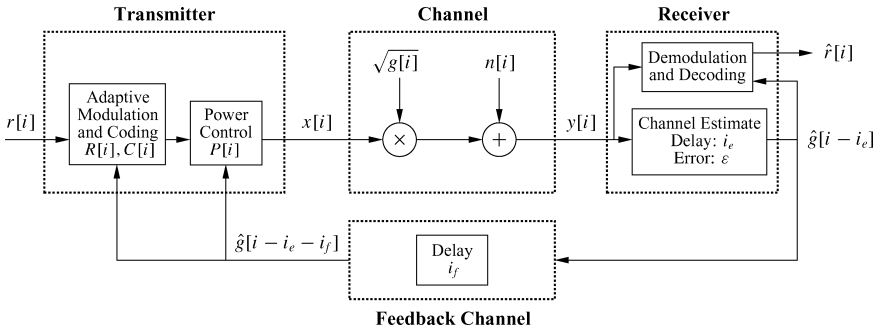

Figure 9.1: System model.

The rate of channel variation will dictate how often the transmitter must adapt its transmission parameters, and it will also affect the estimation error of $g[i]$. When the channel gain consists of both fast and slow fading components, the adaptive transmission may adapt to both if $g[i]$ changes sufficiently slowly, or it may adapt to

just the slow fading. In particular, if $g[i]$ corresponds to shadowing and multipath fading, then at low speeds the shadowing is essentially constant, and the multipath fading is sufficiently slow that it can be estimated and fed back to the transmitter with estimation error and delay that do not significantly degrade performance. At high speeds the system can no longer effectively estimate and feed back the multipath fading in order to adapt to it. In this case, the adaptive transmission responds to the shadowing variations only, and the error probability of the modulation must be averaged over the fast fading distribution. Adaptive techniques for combined fast and slow fading are discussed in Section 9.5.

9.2 Adaptive Techniques

There are many parameters that can be varied at the transmitter relative to the channel gain γ. In this section we discuss adaptive techniques associated with variation of the most common parameters: data rate, power, coding, error probability, and combinations of these adaptive techniques.

9.2.1 Variable-Rate Techniques

In variable-rate modulation the data rate $R[\gamma]$ is varied relative to the channel gain γ. This can be done by fixing the symbol rate $R_s = 1/T_s$ of the modulation and using multiple modulation schemes or constellation sizes, or by fixing the modulation (e.g. BPSK) and changing the symbol rate. Symbol rate variation is difficult to implement in practice because a varying signal bandwidth is impractical and complicates bandwidth sharing. In contrast, changing the constellation size or modulation type with a fixed symbol rate is fairly easy, and these techniques are used in current systems. Specifically, the GSM and IS-136 EDGE system as well as 802.11a wireless LANs vary their modulation and coding relative to channel quality [15]. In general the modulation parameters that dictate the transmission rate are fixed over a block or frame of symbols, where the frame size is a parameter of the design. Frames may also include pilot symbols for channel estimation and other control information.

When a discrete set of modulation types or constellation sizes are used, each value of γ must be mapped to one of the possible modulation schemes. This is often done to maintain the bit error probability of each scheme below a given value. These ideas are illustrated in the following example as well as in subsequent sections on specific adaptive modulation techniques.

Example 9.1: Consider an adaptive modulation system that uses QPSK and 8-PSK for a target P_b of approximately 10^{-3}. If the target P_b cannot be met with either scheme, then no data is transmitted. Find the range of γ-values associated with the three possible transmission schemes (no transmission, QPSK, and 8-PSK) as well as the average spectral efficiency of the system, assuming Rayleigh fading with $\bar{\gamma} = 20$ dB.

Solution: First note that the SNR $\gamma = \gamma_s$ for both QPSK and 8-PSK. From Section 6.1 we have $P_b \approx Q(\sqrt{\gamma})$ for QPSK and $P_b \approx .666Q(\sqrt{2\gamma}\sin(\pi/8))$ for 8-PSK. Since $\gamma > 14.79$ dB yields $P_b < 10^{-3}$ for 8-PSK, the adaptive modulation uses 8-PSK modulation for $\gamma > 14.79$ dB. Since $\gamma > 10.35$ dB yields $P_b < 10^{-3}$ for QPSK, the adaptive modulation uses QPSK modulation for 14.79 dB $\geq \gamma > 10.35$ dB. The channel is not used for $\gamma \leq 10.35$ dB.

We determine the average rate by analyzing how often each of the different transmission schemes is used. Since 8-PSK is used when $\gamma > 14.79$ dB = 30.1, in Rayleigh fading with $\bar{\gamma} = 20$ dB the spectral efficiency $R[\gamma]/B = \log_2 8 = 3$ bps/Hz is transmitted a fraction of time equal to $P_8 = \int_{30.1}^{\infty} \frac{1}{100} e^{-\gamma/100}\, d\gamma = .74$. QPSK is used when $10.35 < \gamma \leq 14.79$ dB, where 10.35 dB = 10.85 in linear units. So $R[\gamma] = \log_2 4 = 2$ bps/Hz is transmitted a fraction of time equal to $P_4 = \int_{10.85}^{30.1} \frac{1}{100} e^{-\gamma/100}\, d\gamma = .157$. During the remaining .103 portion of time there is no data transmission. So the average spectral efficiency is $.74 \cdot 3 + .157 \cdot 2 + .103 \cdot 0 = 2.534$ bps/Hz.

Note that if $\gamma \leq 10.35$ dB then – rather than suspending transmission, which leads to an outage probability of roughly .1 – either just one signaling dimension could be used (i.e., BPSK could be transmitted) or error correction coding could be added to the QPSK to meet the P_b target. If block or convolutional codes were used then the spectral efficiency for $\gamma \leq 10.35$ dB would be less than 2 bps/Hz but larger than a spectral efficiency of zero corresponding to no transmission. These variable-coding techniques are described in Section 9.2.4.

9.2.2 Variable-Power Techniques

Adapting the transmit power alone is generally used to compensate for SNR variation due to fading. The goal is to maintain a fixed bit error probability or, equivalently, a constant received SNR. The power adaptation thus inverts the channel fading so that the channel appears as an AWGN channel to the modulator and demodulator.[1] The power adaptation for channel inversion is given by

$$\frac{P(\gamma)}{\bar{P}} = \frac{\sigma}{\gamma}, \tag{9.1}$$

where σ equals the constant received SNR. The average power constraint $\bar{P}$ implies that

$$\int \frac{P(\gamma)}{\bar{P}} p(\gamma)\, d\gamma = \int \frac{\sigma}{\gamma} p(\gamma)\, d\gamma = 1. \tag{9.2}$$

Solving (9.2) for σ yields that $\sigma = 1/\mathbf{E}[1/\gamma]$, so σ is determined by $p(\gamma)$, which in turn depends on the average transmit power $\bar{P}$ through $\bar{\gamma}$. Thus, for a given average power $\bar{P}$, if the value for σ required to meet the target bit error rate is greater than $1/\mathbf{E}[1/\gamma]$ then this target cannot be met. Note that for Rayleigh fading, where γ is exponentially distributed, $\mathbf{E}[1/\gamma] = \infty$ and so no target P_b can be met using channel inversion.

The fading can also be inverted above a given cutoff γ_0, which leads to a truncated channel inversion for power adaptation. In this case the power adaptation is given by

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} \sigma/\gamma & \gamma \geq \gamma_0, \\ 0 & \gamma < \gamma_0. \end{cases} \tag{9.3}$$

The cutoff value γ_0 can be based on a desired outage probability $P_{\text{out}} = p(\gamma < \gamma_0)$ or on a desired target BER above a cutoff that is determined by the target BER and $p(\gamma)$. Since the channel is only used when $\gamma \geq \gamma_0$, given an average power $\bar{P}$ we have $\sigma = 1/\mathbf{E}_{\gamma_0}[1/\gamma]$, where

$$\mathbf{E}_{\gamma_0}\left[\frac{1}{\gamma}\right] \triangleq \int_{\gamma_0}^{\infty} \frac{1}{\gamma} p(\gamma)\, d\gamma. \tag{9.4}$$

Example 9.2: Find the power adaptation for BPSK modulation that maintains a fixed $P_b = 10^{-3}$ in non-outage for a Rayleigh fading channel with $\bar{\gamma} = 10$ dB. Also find the resulting outage probability.
Solution: The power adaptation is truncated channel inversion, so we need only find σ and γ_0. For BPSK modulation, with a constant SNR of $\sigma = 4.77$ we get $P_b = Q(\sqrt{2\sigma}) = 10^{-3}$. Setting $\sigma = 1/\mathbf{E}_{\gamma_0}[1/\gamma]$ and solving for γ_0, which must be done numerically, yields $\gamma_0 = .7423$. So $P_{\text{out}} = p(\gamma < \gamma_0) = 1 - e^{-\gamma_0/10} = .379$. Hence there is a high outage probability, which results from requiring $P_b = 10^{-3}$ in this relatively weak channel.

[1] Channel inversion and truncated channel inversion were discussed in Section 4.2.4 in the context of fading channel capacity.

9.2.3 Variable Error Probability

We can also adapt the instantaneous BER subject to an average BER constraint $\bar{P}_b$. In Section 6.3.2 we saw that in fading channels the instantaneous error probability varies as the received SNR γ varies, resulting in an average BER of $\bar{P}_b = \int P_b(\gamma)p(\gamma)\,d\gamma$. This is not considered an adaptive technique, since the transmitter does not adapt to γ. Thus, in adaptive modulation, error probability is typically adapted along with some other form of adaptation such as constellation size or modulation type. Adaptation based on varying both data rate and error probability to reduce transmit energy was first proposed by Hayes in [1], where a 4-dB power savings was obtained at a target average bit error probability of 10^{-4}.

9.2.4 Variable-Coding Techniques

In adaptive coding, different channel codes are used to provide different amounts of coding gain to the transmitted bits. For example, a stronger error correction code may be used when γ is small, with a weaker code or no coding used when γ is large. Adaptive coding can be implemented by multiplexing together codes with different error correction capabilities. However, this approach requires that the channel remain roughly constant over the block length or constraint length of the code [7]. On such slowly varying channels, adaptive coding is particularly useful when the modulation must remain fixed, as may be the case owing to complexity or peak-to-average power ratio constraints.

An alternative technique to code multiplexing is rate-compatible punctured convolutional (RCPC) codes [18]. This is a family of convolutional codes at different code rates $R_c = k/n$. The basic premise of RCPC codes is to have a single encoder and decoder whose error correction capability can be modified by not transmitting certain coded bits (*puncturing* the code). Moreover, RCPC codes have a rate compatibility constraint so that the coded bits associated with a high-rate (weaker) code are also used by all lower-rate (stronger) codes. Thus, to increase the error correction capability of the code, the coded bits of the weakest code are transmitted along with additional coded bits to achieve the desired level of error correction. The rate compatibility makes it easy to adapt the error protection of the code, since the same encoder and decoder are used for all codes in the RCPC family, with puncturing at the transmitter to achieve the desired error correction. Decoding is performed by a Viterbi algorithm operating on the trellis associated with the lowest rate code, with the puncturing incorporated into the branch metrics. Puncturing is a very effective and powerful adaptive coding technique, and it forms the basis of adaptive coding in the GSM and IS-136 EDGE protocol for data transmission [13].

Adaptive coding through either multiplexing or puncturing can be done for fixed modulation or combined with adaptive modulation as a hybrid technique. When the modulation is fixed, adaptive coding is often the only practical mechanism to address the channel variations [6, 7]. The focus of this chapter is on systems where adaptive modulation is possible, so adaptive coding on its own will not be further discussed.

9.2.5 Hybrid Techniques

Hybrid techniques can adapt multiple parameters of the transmission scheme, including rate, power, coding, and instantaneous error probability. In this case joint optimization of the different techniques is used to meet a given performance requirement. Rate adaptation is often combined with power adaptation to maximize spectral efficiency, and we apply this joint optimization to different modulations in subsequent sections. Adaptive modulation and coding has been widely investigated in the literature and is currently used in both cellular systems and wireless LANs [13, 15].

9.3 Variable-Rate Variable-Power MQAM

In the previous section we discussed general approaches to adaptive modulation and coding. In this section we describe a specific form of adaptive modulation where the rate and power of MQAM are varied to maximize spectral efficiency while meeting a given instantaneous P_b target. We study this specific form of adaptive modulation because it provides insight into the benefits of adaptive modulation and since, moreover, the same scheme for power and rate adaptation that achieves capacity also optimizes this adaptive MQAM design. We will show that there is a constant power gap between the spectral efficiency of this adaptive MQAM technique and capacity in flat fading, and this gap can be partially closed by coded modulation using a trellis or lattice code superimposed on the adaptive modulation.

Consider a family of MQAM signal constellations with a fixed symbol time T_s, where M denotes the number of points in each signal constellation. We assume $T_s = 1/B$ based on ideal Nyquist pulse shaping. Let $\bar{P}$, N_0, $\gamma = \bar{P}g/N_0B$, and $\bar{\gamma} = \bar{P}/N_0B$ be as given in our system model. Then the average E_s/N_0 equals the average SNR:

$$\frac{\bar{E}_s}{N_0} = \frac{\bar{P}T_s}{N_0} = \bar{\gamma}. \tag{9.5}$$

The spectral efficiency for fixed M is $R/B = \log_2 M$, the number of bits per symbol. This efficiency is typically parameterized by the average transmit power $\bar{P}$ and the BER of the modulation technique.

9.3.1 Error Probability Bounds

In [19] the BER for an AWGN channel with MQAM modulation, ideal coherent phase detection, and SNR γ is bounded by

$$P_b \leq 2e^{-1.5\gamma/(M-1)}. \tag{9.6}$$

A tighter bound, good to within 1 dB for $M \geq 4$ and $0 \leq \gamma \leq 30$ dB, is

$$P_b \leq .2e^{-1.5\gamma/(M-1)}. \tag{9.7}$$

Note that these expressions are only bounds, so they don't match the error probability expressions from Table 6.1. We use these bounds because they are easy to invert: so we can obtain M as a function of the target P_b and the power adaptation policy, as we will see shortly. Adaptive modulation designs can also be based on BER expressions that are not invertible or on BER simulation results, with numerical inversion used to obtain the constellation size and SNR associated with a given BER target.

In a fading channel with nonadaptive transmission (constant transmit power and rate), the average BER is obtained by integrating the BER in AWGN over the fading distribution $p(\gamma)$. Thus, we use the average BER expression to find the maximum data rate that can achieve a given average BER for a given average SNR. Similarly, if the data rate and average BER are fixed, we can find the required average SNR to achieve this target, as illustrated in the next example.

Example 9.3: Find the average SNR required to achieve an average BER of $\bar{P}_b = 10^{-3}$ for nonadaptive BPSK modulation in Rayleigh fading. What is the spectral efficiency of this scheme?
Solution: From Section 6.3.2, BPSK in Rayleigh fading has $\bar{P}_b \approx 1/4\bar{\gamma}$. Thus, without transmitter adaptation, for a target average BER of $\bar{P}_b = 10^{-3}$ we require $\bar{\gamma} = 1/4\bar{P}_b = 250 = 24$ dB. The spectral efficiency is $R/B = \log_2 2 = 1$ bps/Hz. We will see that adaptive modulation provides a much higher spectral efficiency at this same SNR and target BER.

9.3.2 Adaptive Rate and Power Schemes

We now consider adapting the transmit power $P(\gamma)$ relative to γ, subject to the average power constraint $\bar{P}$ and an instantaneous BER constraint $P_b(\gamma) = P_b$. The received SNR is then $\gamma P(\gamma)/\bar{P}$, and the P_b bound for each value of γ, using the tight bound (9.7), becomes

$$P_b(\gamma) \leq .2 \exp\left[\frac{-1.5\gamma}{M-1}\frac{P(\gamma)}{\bar{P}}\right]. \tag{9.8}$$

We adjust M and $P(\gamma)$ to maintain the target P_b. Rearranging (9.8) yields the following maximum constellation size for a given P_b:

$$M(\gamma) = 1 + \frac{1.5\gamma}{-\ln(5P_b)}\frac{P(\gamma)}{\bar{P}} = 1 + K\gamma\frac{P(\gamma)}{\bar{P}}, \tag{9.9}$$

where

$$K = \frac{-1.5}{\ln(5P_b)} < 1. \tag{9.10}$$

We maximize spectral efficiency by maximizing

$$\mathbf{E}[\log_2 M(\gamma)] = \int_0^\infty \log_2\left(1 + \frac{K\gamma P(\gamma)}{\bar{P}}\right) p(\gamma)\, d\gamma \tag{9.11}$$

subject to the power constraint

$$\int_0^\infty P(\gamma)p(\gamma)\, d\gamma = \bar{P}. \tag{9.12}$$

The power adaptation policy that maximizes (9.11) has the same form as the optimal power adaptation policy (4.15) that achieves capacity:

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} 1/\gamma_0 - 1/(\gamma K) & \gamma \geq \gamma_0/K, \\ 0 & \gamma < \gamma_0/K, \end{cases} \tag{9.13}$$

where γ_0/K is the optimized cutoff fade depth below which the channel is not used, for K given by (9.10). If we define $\gamma_K = \gamma_0/K$ and multiply both sides of (9.13) by K, we obtain

$$\frac{KP(\gamma)}{\bar{P}} = \begin{cases} 1/\gamma_K - 1/\gamma & \gamma \geq \gamma_K, \\ 0 & \gamma < \gamma_K, \end{cases} \tag{9.14}$$

where γ_K is a cutoff fade depth below which the channel is not used. This cutoff must satisfy the power constraint (9.12):

$$\int_{\gamma_K}^\infty \left(\frac{1}{\gamma_K} - \frac{1}{\gamma}\right) p(\gamma)\, d\gamma = K. \tag{9.15}$$

Substituting (9.13) or (9.14) into (9.9), we get that the instantaneous rate is given by

$$\log_2 M(\gamma) = \log_2\left(\frac{\gamma}{\gamma_K}\right), \tag{9.16}$$

and the corresponding average spectral efficiency (9.11) is given by

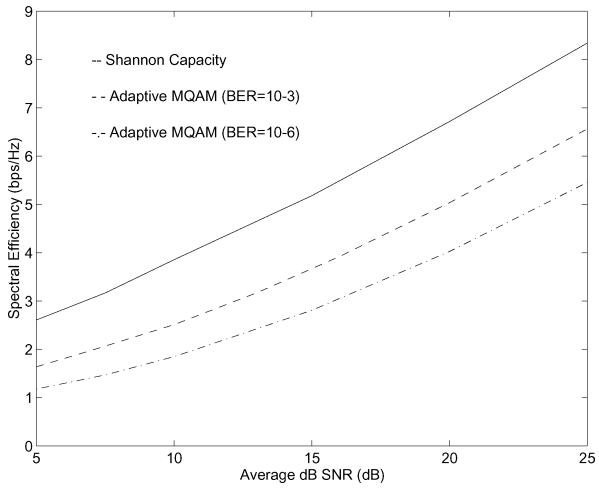

Figure 9.2: Average spectral efficiency in log-normal shadowing ($\sigma_{\psi_{\rm dB}} = 8$ dB).

$$\frac{R}{B} = \int_{\gamma_K}^{\infty} \log_2 \left(\frac{\gamma}{\gamma_K} \right) p(\gamma)\, d\gamma. \tag{9.17}$$

Comparing the power adaptation and average spectral efficiency – (4.15) and (4.16) – associated with the Shannon capacity of a fading channel with (9.13) and (9.17), the optimal power adaptation and average spectral efficiency of adaptive MQAM, we see that the power and rate adaptation are the same and lead to the same average spectral efficiency, with an effective power loss of K for adaptive MQAM as compared to the capacity-achieving scheme. Moreover, this power loss is independent of the fading distribution. Thus, if the capacity of a fading channel is R bps/Hz at SNR $\bar{\gamma}$, then uncoded adaptive MQAM requires a received SNR of $\bar{\gamma}/K$ to achieve the same rate. Equivalently, K is the maximum possible coding gain for this variable rate and power MQAM method. We discuss superimposing a trellis or lattice code on the adaptive MQAM to obtain some of this coding gain in Section 9.3.8.

We plot the average spectral efficiency (9.17) of adaptive MQAM at a target P_b of 10^{-3} and 10^{-6} for both log-normal shadowing and Rayleigh fading in Figures 9.2 and 9.3, respectively. We also plot the capacity in these figures for comparison. Note that the gap between the spectral efficiency of variable-rate variable-power MQAM and capacity is the constant K, which from (9.10) is a simple function of the BER.

9.3.3 Channel Inversion with Fixed Rate

We can also apply channel inversion power adaptation to maintain a fixed received SNR. We then transmit a single fixed-rate MQAM modulation that achieves the target P_b. The constellation size M that meets this target P_b is obtained by substituting the channel inversion power adaptation $P(\gamma)/\bar{P} = \sigma/\gamma$ of (9.2) into (9.9) with $\sigma = 1/\mathbf{E}[1/\gamma]$. Since the resulting spectral efficiency $R/B = M$, this yields the spectral efficiency of the channel inversion power adaptation as

$$\frac{R}{B} = \log_2 \left(1 + \frac{-1.5}{\ln(5P_b)\mathbf{E}[1/\gamma]} \right). \tag{9.18}$$

This spectral efficiency is based on the tight bound (9.7); if the resulting $M = R/B < 4$ then the loose bound (9.6) must be used, in which case $\ln(5P_b)$ is replaced by $\ln(.5P_b)$ in (9.18).

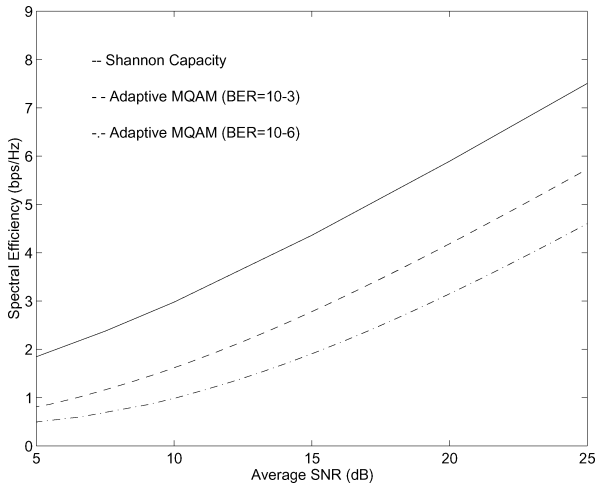

Figure 9.3: Average spectral efficiency in Rayleigh fading.

With truncated channel inversion the channel is only used when $\gamma > \gamma_0$. Thus, the spectral efficiency with truncated channel inversion is obtained by substituting $P(\gamma)/\bar{P} = \sigma/\gamma(\gamma > \gamma_0)$ into (9.9) and multiplying by the probability that $\gamma > \gamma_0$. The maximum value is obtained by optimizing relative to the cutoff level γ_0:

$$\frac{R}{B} = \max_{\gamma_0} \log_2 \left(1 + \frac{-1.5}{\ln(5P_b)\mathbf{E}_{\gamma_0}[1/\gamma]}\right) p(\gamma > \gamma_0). \tag{9.19}$$

The spectral efficiency of adaptive MQAM (with the optimal water-filling and truncated channel inversion power adaptation) in a Rayleigh fading channel with a target BER of 10^{-3} is shown in Figure 9.4, along with the capacity under the same two power adaptation policies. We see, surprisingly, that truncated channel inversion with fixed-rate transmission has almost the same spectral efficiency as optimal variable-rate variable-power MQAM. This suggests that truncated channel inversion is more desirable in practice, since it achieves almost the same spectral efficiency as variable-rate variable-power transmission but does not require that we vary the rate. However, this assumes there is no restriction on constellation size. Specifically, the spectral efficiencies (9.17), (9.18), and (9.19) assume that M can be any real number and that the power and rate can vary continuously with γ. Though MQAM modulation for noninteger values of M is possible, the complexity is quite high [20]. Moreover, it is difficult in practice to continually adapt the transmit power and constellation size to the channel fading, particularly in fast fading environments. Thus, we now consider restricting the constellation size to just a handful of values. This will clearly affect the spectral efficiency, though (as shown in the next section) not by very much.

9.3.4 Discrete-Rate Adaptation

We now assume the same model as in the previous section, but we restrict the adaptive MQAM to a limited set of constellations. Specifically, we assume a set of square constellations of size $M_0 = 0, M_1 = 2$, and $M_j = 2^{2(j-1)}$, $j = 2, \ldots, N-1$ for some N. We assume square constellations for $M > 2$ since they are easier to implement than rectangular ones [21]. We first analyze the impact of this restriction on the spectral efficiency of the optimal adaptation policy. We then determine the effect on the channel inversion policies.

Consider a variable-rate variable-power MQAM transmission scheme subject to these constellation restrictions. Thus, at each symbol time we transmit a symbol from a constellation in the set $\{M_j \colon j = 0, 1, \ldots, N-1\}$; the choice of constellation depends on the fade level γ over that symbol time. Choosing the M_0 constellation corresponds to no data transmission. For each value of γ, we must decide which constellation to transmit and what

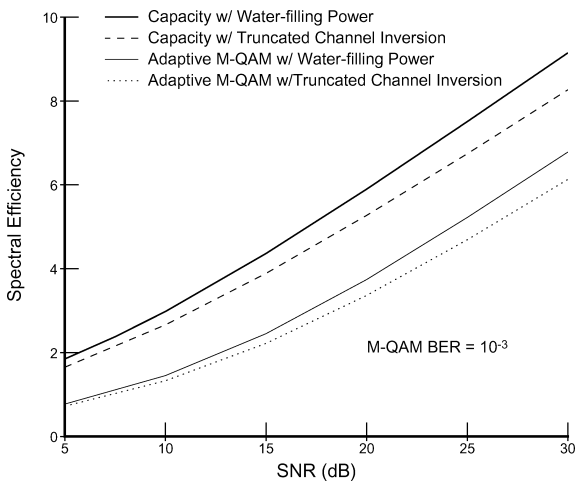

Figure 9.4: Spectral efficiency with different power adaptation policies (Rayleigh fading).

the associated transmit power should be. The rate at which the transmitter must change its constellation and power is analyzed below. Since the power adaptation is continuous while the constellation size is discrete, we call this a continuous-power discrete-rate adaptation scheme.

We determine the constellation size associated with each γ by discretizing the range of channel fade levels. Specifically, we divide the range of γ into *N fading regions* $R_j = [\gamma_{j-1}, \gamma_j)$, $j = 0, \ldots, N-1$, where $\gamma_{-1} = 0$ and $\gamma_{N-1} = \infty$. We transmit constellation M_j when $\gamma \in R_j$. The spectral efficiency for $\gamma \in R_j$ is thus $\log_2 M_j$ bps/Hz for $j > 0$.

The adaptive MQAM design requires that the boundaries of the R_j regions be determined. Although these boundaries can be optimized to maximize spectral efficiency, as derived in Section 9.4.2, the optimal boundaries cannot be found in closed form and require an exhaustive search to obtain. Thus, we will use a suboptimal technique to determine boundaries. These suboptimal boundaries are much easier to find than the optimal ones and have almost the same performance. Define

$$M(\gamma) = \frac{\gamma}{\gamma_K^*}, \tag{9.20}$$

where $\gamma_K^* > 0$ is a parameter that will later be optimized to maximize spectral efficiency. Note that substituting (9.13) into (9.9) yields (9.20) with $\gamma_K^* = \gamma_K$. Therefore, the appropriate choice of γ_K^* in (9.20) defines the optimal constellation size for each γ when there is no constellation restriction.

Assume now that γ_K^* is fixed and define $M_N = \infty$. To obtain the constellation size $M_j (j = 0, \ldots, N-1)$ for a given SNR γ, we first compute $M(\gamma)$ from (9.20). We then find j such that $M_j \leq M(\gamma) < M_{j+1}$ and assign constellation M_j to this γ-value. Thus, for a fixed γ, we transmit the largest constellation in our set $\{M_j \colon j = 0, \ldots, N-1\}$ that is smaller than $M(\gamma)$. For example, if the fade level γ satisfies $2 \leq \gamma/\gamma_K^* < 4$ then we transmit BPSK. The region boundaries other than $\gamma_{-1} = 0$ and $\gamma_{N-1} = \infty$ are located at $\gamma_j = \gamma_K^* M_{j+1}$, $j = 0, \ldots, N-2$. This mapping of γ values to fading regions R_j and associated constellations M_j is shown in Figure 9.5. Clearly, increasing the number of discrete signal constellations N yields a better approximation to the continuous adaptation (9.9), resulting in a higher spectral efficiency.

Once the regions and associated constellations are fixed, we must find a power adaptation policy that satisfies the BER requirement and the power constraint. By (9.9) we can maintain a fixed BER for the constellation $M_j > 0$ using the power adaptation policy

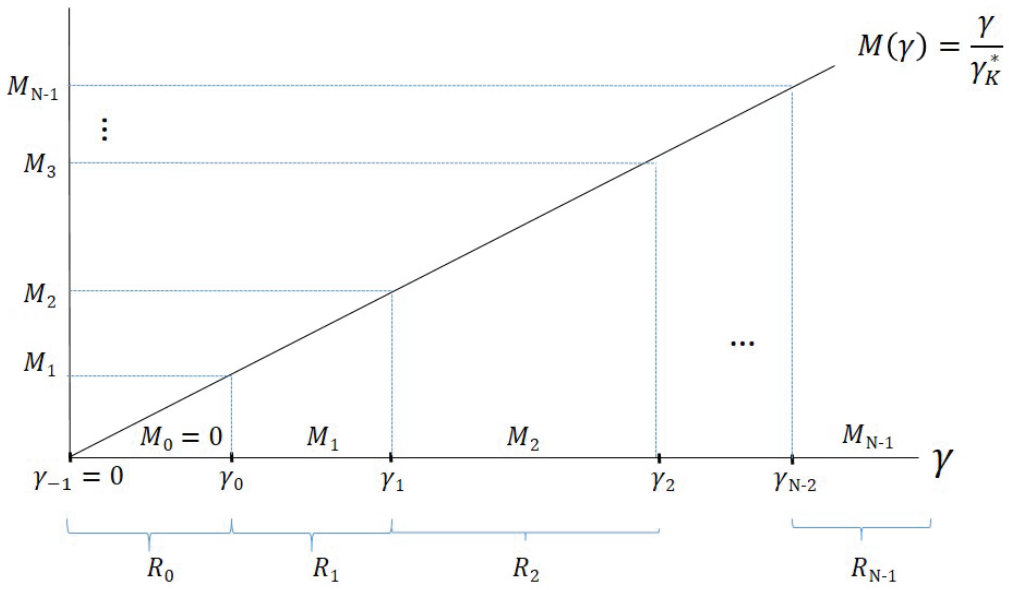

Figure 9.5: Mapping γ to Fading Regions R_j and Associated Constellations M_j in Discrete-Rate Adaptation

$$\frac{P_j(\gamma)}{\bar{P}} = \begin{cases} (M_j - 1)/(\gamma K) & M_j < \gamma/\gamma_K^* \leq M_{j+1}, \\ 0 & M_j = 0 \end{cases} \tag{9.21}$$

for $\gamma \in R_j$, since this power adaptation policy leads to a fixed received E_s/N_0 for the constellation M_j of

$$\frac{E_s(j)}{N_0} = \frac{\gamma P_j(\gamma)}{\bar{P}} = \frac{M_j - 1}{K}. \tag{9.22}$$

By definition of K, MQAM modulation with constellation size M_j and E_s/N_0 given by (9.22) results in the desired target P_b. In Table 9.1 we tabulate the constellation size and power adaptation as a function of γ and γ_K^* for five fading regions.

Table 9.1: Rate and power adaptation for five regions

Region (j)	**γ range**	M_j	$P_j(\gamma)/\bar{P}$
0	$0 \leq \gamma/\gamma_K^\star < 2$	0	0
1	$2 \leq \gamma/\gamma_K^\star < 4$	2	$1/(\gamma K)$
2	$4 \leq \gamma/\gamma_K^\star < 16$	4	$3/(\gamma K)$
3	$16 \leq \gamma/\gamma_K^\star < 64$	16	$15/(\gamma K)$
4	$64 \leq \gamma/\gamma_K^\star < \infty$	64	$63/(\gamma K)$

The spectral efficiency for this discrete-rate policy is just the sum of the data rates associated with each of the regions multiplied by the probability that γ falls in that region:

$$\frac{R}{B} = \sum_{j=1}^{N-1} \log_2(M_j) p\left(M_j \leq \frac{\gamma}{\gamma_K^*} < M_{j+1}\right). \tag{9.23}$$

Since M_j is a function of γ_K^*, we can maximize (9.23) relative to γ_K^*, subject to the power constraint

$$\sum_{j=1}^{N-1} \int_{\gamma_K^* M_j}^{\gamma_K^* M_{j+1}} \frac{P_j(\gamma)}{\bar{P}} p(\gamma)\, d\gamma = 1, \tag{9.24}$$

where $P_j(\gamma)/\bar{P}$ is defined in (9.21). There is no closed-form solution for the optimal γ_K^*: in the calculations that follow it was found using numerical search techniques.

In Figures 9.6 and 9.7 we show the maximum of (9.23) versus the number of fading regions N for log-normal shadowing and Rayleigh fading, respectively. We assume a BER of 10^{-3} for both plots. From Figure 9.6 we see that restricting our adaptive policy to just six fading regions $(M_j = 0, 2, 4, 16, 64, 256)$ results in a spectral efficiency that is within 1 dB of the efficiency obtained with continuous-rate adaptation (9.17) under log-normal shadowing. A similar result holds for Rayleigh fading using five regions $(M_j = 0, 2, 4, 16, 64)$. Essentially this indicates that with 5 regions, $M_j \approx M(\gamma) = \gamma/\gamma_K, j = 0, \ldots, N-1$ for $M(\gamma)$ the optimal rate adaptation with no constraints on the constellation size. Both these figures also show that a smaller number of regions leads to a lower average rate. That is because the approximation $M_j \approx M(\gamma) = \gamma/\gamma_K$ no longer holds, both because the smaller number of regions implies that M_j maintains a constant value for a larger range of γ values and because the maximum constellation $M_{N-1} << M(\gamma)$ at high SNRs. This latter effect is evident in both figures as, for high SNRs, the average rate levels off to that of the largest constellation in the available set of constellations (e.g. for 4 regions, the maximum constellation is $M_3 = 16$QAM whose spectral efficiency is 4 bps/Hz).

We can simplify our discrete-rate policy even further by using a constant transmit power for each constellation M_j. Thus, each fading region is associated with one signal constellation and one transmit power. This policy is called discrete-power discrete-rate adaptive MQAM. Since the transmit power and constellation size are fixed in each region, the BER will vary with γ in each region. Thus, the region boundaries and transmit power must be set to achieve a given target average BER.

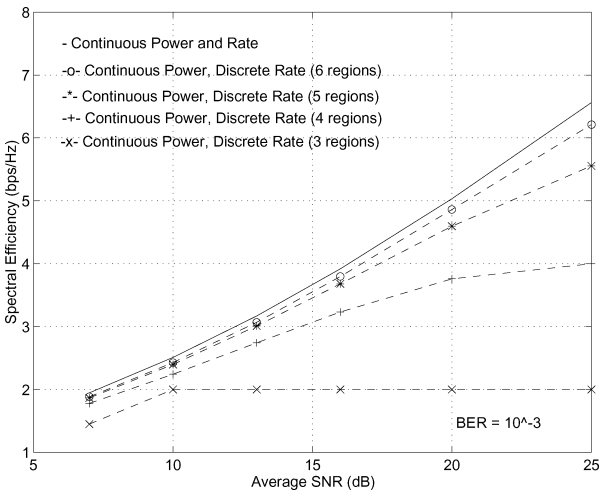

Figure 9.6: Discrete-rate efficiency in log-normal shadowing ($\sigma_{\psi_{\rm dB}} = 8$ dB).

A restriction on allowable signal constellations will also affect the total channel inversion and truncated channel inversion policies. Specifically, suppose that, with the channel inversion policies, the constellation must be chosen from a fixed set of possible constellations $\mathcal{M} = \{M_0 = 0,\ \ldots,\ M_{N-1}\}$. For total channel inversion the spectral efficiency with this restriction is thus

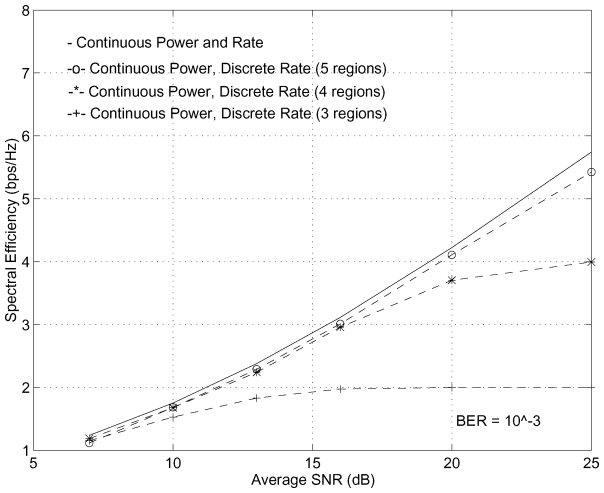

Figure 9.7: Discrete-rate efficiency in Rayleigh fading.

$$\frac{R}{B} = \log_2 \left\lfloor \left(1 + \frac{-1.5}{\ln(5P_b)\mathbf{E}[1/\gamma]}\right)\right\rfloor_{\mathcal{M}}, \tag{9.25}$$

where $\lfloor x \rfloor_{\mathcal{M}}$ denotes the largest number in the set $\mathcal{M}$ less than or equal to x. The spectral efficiency with this policy will be restricted to values of $\log_2 M (M \in \mathcal{M})$, with discrete jumps at the $\bar{\gamma}$-values where the spectral efficiency without constellation restriction (9.18) equals $\log_2 M$. For truncated channel inversion the spectral efficiency is given by

$$\frac{R}{B} = \max_{\gamma_0} \log_2 \left\lfloor \left(1 + \frac{-1.5}{\ln(5P_b)\mathbf{E}_{\gamma_0}[1/\gamma]}\right)\right\rfloor_{\mathcal{M}} p(\gamma > \gamma_0). \tag{9.26}$$

In Figures 9.8 and 9.9 we show the impact of constellation restriction on adaptive MQAM for the different power adaptation policies. When the constellation is restricted we assume six fading regions, so $\mathcal{M} = \{M_0 = 0, 2, 4, \ldots, 256\}$. The power associated with each fading region for the discrete-power discrete-rate policy was chosen to have an average BER equal to the instantaneous BER of the discrete-rate continuous-power adaptative policy. We see from these figures that, for variable-rate MQAM with a small set of constellations, restricting the power to a single value for each constellation degrades spectral efficiency by about $1-2$ dB relative to continuous-power adaptation. For comparison, we also plot the maximum efficiency (9.17) for continuous-power and -rate adaptation. All discrete-rate policies have performance that is within 3 dB of this theoretical maximum.

These figures also show the spectral efficiency of fixed-rate transmission with truncated channel inversion (9.26). The efficiency of this scheme is quite close to that of the discrete-power discrete-rate policy. However, to achieve this high efficiency, the optimal γ_0 is quite large, with a corresponding outage probability $P_{\text{out}} = p(\gamma \leq \gamma_0)$ ranging from .1 to .6. Thus, this policy is similar to packet radio, with bursts of high-speed data when the channel conditions are favorable. The efficiency of total channel inversion (9.25) is also shown for log-normal shadowing; this efficiency equals zero in Rayleigh fading. We also plot the spectral efficiency of nonadaptive transmission, where both the transmission rate and power are constant. As discussed in Section 9.3.1, the average BER in this case is obtained by averaging the probability of error (9.31) over the fade distribution $p(\gamma)$. The spectral efficiency is obtained by determining the value of M that yields a 10^{-3} average BER for the given value of $\bar{\gamma}$, as illustrated in Example 9.3. Nonadaptive transmission clearly suffers a large spectral efficiency loss in exchange for its simplicity. However, if the channel varies rapidly and cannot be accurately estimated, nonadaptive

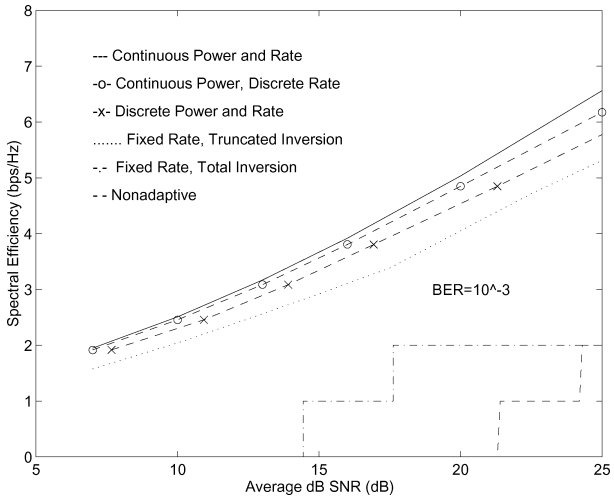

Figure 9.8: Efficiency in log-normal shadowing ($\sigma_{\psi_{\mathrm{dB}}} = 8$ dB).

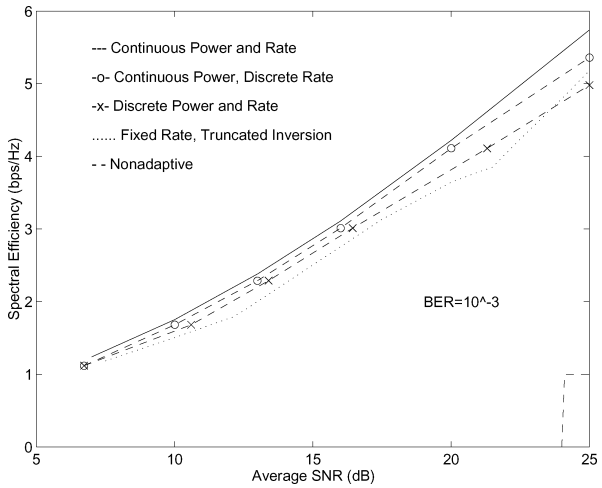

Figure 9.9: Efficiency in Rayleigh fading.

transmission may be the best alternative. Similar curves can be obtained for a target BER of 10^{-6}, with roughly the same spectral efficiency loss relative to a 10^{-3} BER as was exhibited in Figures 9.2 and 9.3.

9.3.5 Average Fade Region Duration

The choice of the number of regions to use in the adaptive policy will depend on how fast the channel is changing as well as on the hardware constraints, which dictate how many constellations are available to the transmitter and at what rate the transmitter can change its constellation and power. Channel estimation and feedback considerations along with hardware constraints may dictate that the constellation remains constant over tens or even hundreds of symbols. In addition, power-amplifier linearity requirements and out-of-band emission constraints may restrict the rate at which power can be adapted. An in-depth discussion of hardware implementation issues can be found in [22]. However, determining how long the SNR γ remains within a particular fading region R_j is of interest, since it determines the trade-off between the number of regions and the rate of power and constellation adaptation. We now investigate the time duration over which the SNR remains within a given fading region.

Table 9.2: Average fade region duration $\bar{\tau}_j$ for $f_D = 100$ Hz

$\mathbf{Region}(j)$	$\bar{\gamma} = \mathbf{10}\ dB$	$\bar{\gamma} = \mathbf{20}\ dB$
0	2.23 ms	0.737 ms
1	0.83 ms	0.301 ms
2	3.00 ms	1.06 ms
3	2.83 ms	2.28 ms
4	1.43 ms	3.84 ms

Let $\bar{\tau}_j$ denote the average time duration that γ stays within the jth fading region. Let $A_j = \gamma_K^* M_j$ for γ_K^* and M_j as previously defined. The jth fading region is then defined as $\{\gamma : A_j \leq \gamma < A_{j+1}\}$. We call $\bar{\tau}_j$ the jth average fade region duration (AFRD). This definition is similar to the average fade duration (AFD; see Section 3.2.3), except that the AFD measures the average time that γ stays below a single level, whereas we are interested in the average time that γ stays between two levels. For the worst-case region $(j = 0)$, these two definitions coincide.

Determining the exact value of $\bar{\tau}_j$ requires a complex derivation based on the joint density $p(\gamma, \dot{\gamma})$, and it remains an open problem. However, a good approximation can be obtained using the finite-state Markov model described in Section 3.2.4. In this model, fading is approximated as a discrete-time Markov process with time discretized to a given interval T, typically the symbol time. It is assumed (i) that the fade value γ remains within one region over a symbol period and (ii) that from a given region the process can only transition to the same region or to adjacent regions. Note that this approximation can lead to longer deep fade durations than more accurate models [23]. The transition probabilities between regions under this assumption are given as

$$p_{j,j+1} = \frac{L_{j+1}T}{\pi_j}, \quad p_{j,j-1} = \frac{L_j T}{\pi_j}, \quad p_{j,j} = 1 - p_{j,j+1} - p_{j,j-1}, \tag{9.27}$$

where L_j is the level-crossing rate at A_j and π_j is the steady-state distribution corresponding to the jth region: $\pi_j = p(A_j \leq \gamma < A_{j+1})$. Since the time over which the Markov process stays in a given state is geometrically distributed [24, Chap. 2.3], $\bar{\tau}_j$ is given by

$$\bar{\tau}_j = \frac{T}{p_{j,j+1} + p_{j,j-1}} = \frac{\pi_j}{L_{j+1} + L_j}. \tag{9.28}$$

The value of $\bar{\tau}_j$ is thus a simple function of the level crossing rate and the fading distribution. Whereas the level crossing rate is known for Rayleigh fading [25, Chap. 1.3.4], it cannot be obtained for log-normal shadowing because the joint distribution $p(\gamma, \dot{\gamma})$ for this fading type is unknown.

In Rayleigh fading, the level crossing rate is given by (3.45) as

$$L_j = \sqrt{\frac{2\pi A_j}{\bar{\gamma}}} f_D e^{-A_j/\bar{\gamma}}, \tag{9.29}$$

where $f_D = v/\lambda$ is the Doppler frequency. Substituting (9.29) into (9.28), we easily see that $\bar{\tau}_j$ is inversely proportional to the Doppler frequency. Moreover, π_j and A_j do not depend on f_D and so, if we compute $\bar{\tau}_j$ for a given Doppler frequency f_D, then we can compute $\hat{\bar{\tau}}_j$ corresponding to another Doppler frequency $\hat{f}_D$ as

$$\hat{\bar{\tau}}_j = \frac{f_D}{\hat{f}_D} \bar{\tau}_j. \tag{9.30}$$

Table 9.2 shows the $\bar{\tau}_j$ values corresponding to five regions ($M_j = 0, 2, 4, 16, 64$) in Rayleigh fading[2] for $f_D = 100$ Hz and two average power levels: $\bar{\gamma} = 10$ dB($\gamma_K^* = 1.22$) and $\bar{\gamma} = 20$ dB($\gamma_K^* = 1.685$). The AFRD for other Doppler frequencies is easily obtained using the table values and (9.30). This table indicates that, even at high velocities, for rates of 100 kilosymbols/second the discrete-rate discrete-power policy will maintain the same constellation and transmit power over tens to hundreds of symbols.

Example 9.4: Find the AFRDs for a Rayleigh fading channel with $\bar{\gamma} = 10$ dB, $M_j = 0, 2, 4, 16, 64, 64$, and $f_D = 50$ Hz.
Solution: We first note that all parameters are the same as used in the calculation of Table 9.2 except that the Doppler $\hat{f}_D = 50$ Hz is half the Doppler of $f_D = 100$ Hz used to compute the table values. Thus, from (9.30), we obtain the AFRDs with this new Doppler by multiplying each value in the table by $f_D/\hat{f}_D = 2$.

In shadow fading we can obtain a coarse approximation of $\bar{\tau}_j$ based on the shadowing autocorrelation function (2.39): $A(\delta) = \sigma^2_{\psi_{\rm dB}} e^{-\delta/X_c}$, where $\delta = v\tau$ for v the mobile's velocity. Specifically, we can approximate the AFRD for all regions as $\bar{\tau}_j \approx .1X_c/v$, since then the correlation between fade levels separated in time by $\bar{\tau}_j$ is .9. Thus, for a small number of regions it is likely that γ will remain within the same region over this time period.

9.3.6 Exact versus Approximate Bit Error Probability

The adaptive policies described in prior sections are based on the BER upper bounds of Section 9.3.1. Since these are upper bounds, they will lead to a lower BER than the target. We would like to see how the BER achieved with these policies differs from the target BER. A more accurate value for the BER achieved with these policies can be obtained by simulation or by using a better approximation for BER than the upper bounds. From Table 6.1, the BER of MQAM with Gray coding at high SNRs is well approximated by

$$P_b \approx \frac{4}{\log_2 M} Q\left(\sqrt{\frac{3\gamma}{M-1}}\right). \tag{9.31}$$

Moreover, for the continuous-power discrete-rate policy, $\gamma = E_s/N_0$ for the jth signal constellation is

$$\frac{E_s(j)}{N_0} = \frac{M_j - 1}{K}. \tag{9.32}$$

Thus, we can obtain a more accurate analytical expression for the average BER associated with our adaptive policies by averaging over the BER (9.31) for each signal constellation as

$$\bar{P}_b = \frac{\sum_{j=1}^{N-1} 4Q(\sqrt{3/K}) \int_{\gamma_K^* M_j}^{\gamma_K^* M_{j+1}} p(\gamma)\, d\gamma}{\sum_{j=1}^{N-1} \log_2 M_j \int_{\gamma_K^* M_j}^{\gamma_K^* M_{j+1}} p(\gamma)\, d\gamma} \tag{9.33}$$

with $M_N = \infty$.

We plot the analytical expression (9.33) along with the simulated BER for the variable rate and power MQAM with a target BER of 10^{-3} in Figures 9.10 and 9.11 for log-normal shadowing and Rayleigh fading, respectively. The simulated BER is slightly better than the analytical calculation of (9.33) because this equation is based on the nearest neighbor bound for the maximum number of nearest neighbors. Both the simulated and analytical BER are smaller than the target BER of 10^{-3} for $\bar{\gamma} > 10$ dB. The BER bound of 10^{-3} breaks down at low SNRs, since (9.7) is not applicable to BPSK, so we must use the looser bound (9.6). Since the adaptive policy often uses the

[2]The validity of the finite-state Markov model for Rayleigh fading channels has been investigated in [26].

BPSK constellation at low SNRs, the P_b will be larger than that predicted from the tight bound (9.7). That the simulated BER is less than the target at high SNRs implies that the analytical calculations in Figures 9.6 and 9.7 are pessimistic; a slightly higher efficiency could be achieved while still maintaining the target P_b of 10^{-3}.

9.3.7 Channel Estimation Error and Delay

In this section we examine the effects of estimation error and delay, where the estimation error $\varepsilon = \hat{\gamma}/\gamma \neq 1$ and the delay $i_d = i_f + i_e \neq 0$. We first consider the estimation error. Suppose the transmitter adapts its power and rate relative to a target BER P_b^t based on the channel estimate $\hat{\gamma}$ instead of the true value γ. By (9.8), the BER is then bounded by

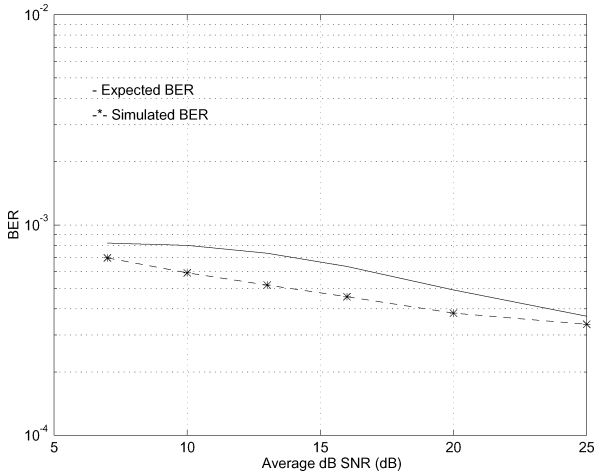

Figure 9.10: BER for log-normal shadowing (six regions).

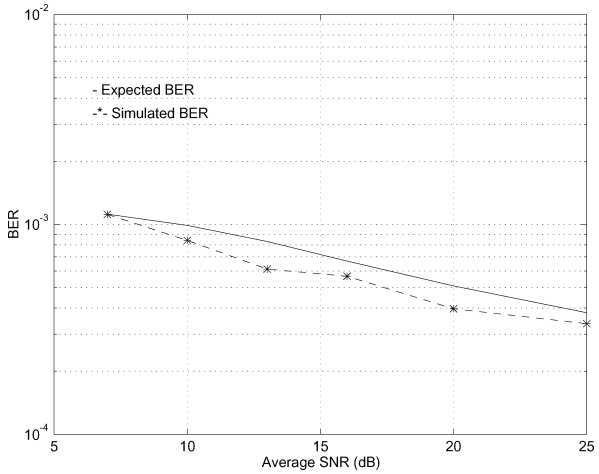

Figure 9.11: BER for Rayleigh fading (five regions).

$$P_b(\gamma, \hat{\gamma}) \leq .2 \exp\left[\frac{-1.5\gamma}{M(\hat{\gamma}) - 1} \frac{P(\hat{\gamma})}{\bar{P}}\right] = .2[5P_b^t]^{1/\epsilon}, \tag{9.34}$$

where the right-hand equality is obtained by substituting the optimal rate (9.9) and power (9.13) policies. For $\varepsilon = 1$, (9.34) reduces to the target BER P_b^t. For $\varepsilon \neq 1 : \varepsilon > 1$ yields an increase in BER above the target, and $\varepsilon < 1$ yields a decrease in BER.

The average probability of bit error is given by

$$\bar{P}_b = \frac{\mathbf{E}[\text{number of bits in error per transmission}]}{\mathbf{E}[\text{number of bits per transmission}]}. \tag{9.35}$$

Given the BER bound (9.34) for constellation $M(\hat{\gamma})$ with rate $\log_2 M(\hat{\gamma})$, we can bound the average probability of error (9.35) of adaptive modulation given channel estimation error and delay as

$$\bar{P}_b = \frac{\int_0^\infty \int_{\gamma_K}^\infty P_b(\gamma,\hat{\gamma}) M(\hat{\gamma}) p(\gamma,\hat{\gamma})\, d\hat{\gamma} d\gamma}{\int_0^\infty M(\hat{\gamma}) p(\hat{\gamma})\, d, \hat{\gamma}} \leq \frac{\int_0^\infty \int_{\gamma_K}^\infty .2[.5P_b^t]^{\gamma/\hat{\gamma}} M(\hat{\gamma}) p(\gamma,\hat{\gamma})\, d\hat{\gamma} d\gamma}{\int_0^\infty M(\hat{\gamma}) p(\hat{\gamma})\, d, \hat{\gamma}}. \tag{9.36}$$

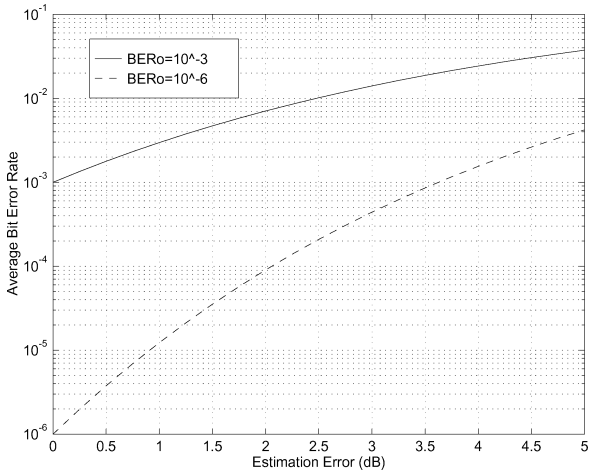

Figure 9.12: Effect of estimation error on BER.

The joint distribution $p(\gamma,\hat{\gamma})$ depends on the channel estimation technique. It has been shown [27] that, when the channel is estimated using pilot symbols, the joint distribution of the signal envelope and its estimate is bivariate Rayleigh. This joint distribution was used in [27] to obtain the probability of error for nonadaptive modulation with channel estimation errors. This analysis can be extended to adaptive modulation using a similar methodology.

If the estimation error stays within some finite range then we can bound its effect using (9.34). We plot the BER increase as a function of a constant ε in Figure 9.12. This figure shows that for a target BER of 10^{-3} the estimation error should be less than 1 dB, and for a target BER of 10^{-6} it should be less than .5 dB. These values are pessimistic, since they assume a constant value of estimation error. Even so, the estimation error can be kept within this range using the pilot-symbol assisted estimation technique described in [28] with appropriate choice of parameters. When the channel is underestimated $(\varepsilon < 1)$, the BER decreases but there will also be some loss in spectral efficiency, since the mean of the channel estimate $\bar{\hat{\gamma}}$ will differ from the true mean $\bar{\gamma}$. The effect of this average power estimation error is characterized in [29].

Suppose now that the channel is estimated perfectly $(\varepsilon = 1)$ but the delay i_d of the estimation and feedback path is nonzero. Thus, at time i the transmitter will use the delayed version of the channel estimate $\hat{\gamma}[i] = \gamma[i - i_d]$ to adjust its power and rate. It was shown in [30] that, conditioned on the outdated channel estimates, the received signal follows a Rician distribution, and the probability of error can then be computed by averaging over the distribution of the estimates. Moreover, [30] develops adaptive coding designs to mitigate the effect of estimation

delay on the performance of adaptive modulation. Alternatively, channel prediction can be used to mitigate these effects [31].

The increase in BER from estimation delay can also be examined in the same manner as in (9.34). Given a target BER P_b^t, the exact channel SNR $\gamma[i]$, and its delayed value $\gamma[i - i_d]$, we have

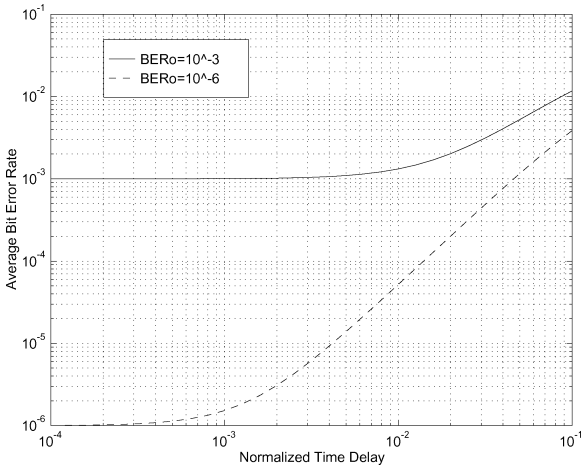

Figure 9.13: Effect of normalized delay $(i_d f_D)$ on BER.

$$P_b(\gamma[i],\, \gamma[i - i_d]) \leq .2 \exp\left[\frac{-1.5\gamma[i]}{M(\gamma[i - i_d]) - 1} \frac{P(\gamma[i - i_d])}{\bar{P}}\right]$$
$$= .2[5P_b^t]^{\gamma[i]/\gamma[i-i_d]}. \tag{9.37}$$

Define $\xi[i,\, i_d] = \gamma[i]/\gamma[i - i_d]$. Since $\gamma[i]$ is stationary and ergodic, the distribution of $\xi[i,\, i_d]$ conditioned on $\gamma[i]$ depends only on i_d and the value of $\gamma = \gamma[i]$. We denote this distribution by $p_{i_d}(\xi|\gamma)$. The instantaneous data rate $\log_2 M(\gamma[i - i_d])$ is based on the value of $\gamma[i - i_d]$ but is transmitted over the channel at time i. So the time delay i_d does not impact the average data rate. However, this delay impacts the instantaneous BER since there is a mismatch between the constellation to be transmitted at time i, $M(\gamma[i])$, and associated power control to achieve the target BER P_b^t, and the constellation that is transmitted at that time, $M(\gamma[i - i_d])$. As shown in [32], the average BER is obtained by integrating over the distribution of this instantaneous BER which depends only on ξ and γ:

$$P_b[i_d] = \int_{\gamma_K}^{\infty} \left[\int_0^{\infty} .2[5P_b^t]^{\xi} p_{i_d}(\xi|\gamma) d\xi\right] p(\gamma)\, d\gamma, \tag{9.38}$$

where γ_K is the cutoff level of the optimal policy and $p(\gamma)$ is the fading distribution. The outer integral is taken from γ_K rather than zero since we only transmit over the channel when $\gamma \geq \gamma_K$. The distribution $p_{i_d}(\xi|\gamma)$ will depend on the autocorrelation of the fading process. A closed-form expression for $p_{i_d}(\xi|\gamma)$ in Nakagami fading (of which Rayleigh fading is a special case) is derived in [32]. Using this distribution in (9.38), we obtain the average BER in Rayleigh fading as a function of the delay parameter i_d. A plot of (9.38) versus the normalized time delay $i_d f_D$ is shown in Figure 9.13. From this figure we see that the total estimation and feedback path delay must be kept to within. $001/f_D$ in order to keep the BER near its desired target.

9.3.8 Adaptive Coded Modulation

Additional coding gain can be achieved with adaptive modulation by superimposing trellis or lattice codes on the adaptive modulation. Specifically, by using the subset partitioning inherent to coded modulation, trellis (or lattice) codes designed for AWGN channels can be superimposed directly onto the adaptive modulation with the same approximate coding gain. The basic idea of adaptive coded modulation is to exploit the separability of code and constellation design that is characteristic of coded modulation, as described in Section 8.7.

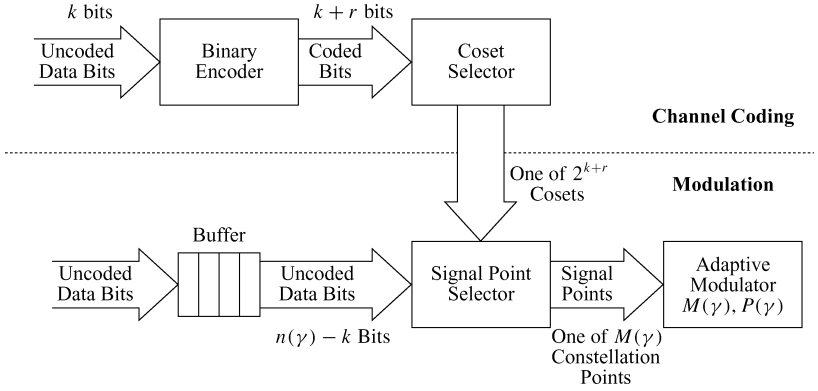

Figure 9.14: Adaptive coded modulation scheme.

Coded modulation is a natural coding scheme to use with variable-rate variable-power MQAM, since the channel coding gain is essentially independent of the modulation. We can therefore adjust the power and rate (number of levels or signal points) in the transmit constellation relative to the instantaneous SNR *without* affecting the channel coding gain, as we now describe in more detail.

The coded modulation scheme is shown in Figure 9.14. The coset code design is the same as it would be for an AWGN channel; that is, the lattice structure and conventional encoder follow the trellis or lattice coded modulation designs outlined in Section 8.7. Let G_c denote the coding gain of the coded modulation, as given by (8.78). The modulation works as follows. The signal constellation is a square lattice with an adjustable number of constellation points M. The size of the MQAM signal constellation from which the signal point is selected is determined by the transmit power, which is adjusted relative to the instantaneous SNR and the desired BER, as in the uncoded case described in Section 9.3.2.

Specifically, if the BER approximation (9.7) is adjusted for the coding gain, then for a particular SNR $= \gamma$ we have

$$P_b \approx .2e^{-1.5(\gamma G_c/M-1)}, \tag{9.39}$$

where M is the size of the transmit signal constellation. As in the uncoded case, using the tight bound (9.7) allows us to adjust the number of constellation points M and signal power relative to the instantaneous SNR in order to maintain a fixed BER:

$$M(\gamma) = 1 + \frac{1.5\gamma G_c}{-\ln(5P_b)}\frac{P(\gamma)}{\bar{P}}. \tag{9.40}$$

The number of uncoded bits required to select the coset point is

$$n(\gamma) - 2k/N = \log_2 M(\gamma) - 2(k+r)/N.$$

Since this value varies with time, these uncoded bits must be queued until needed, as shown in Figure 9.14.

The bit rate per transmission is $\log_2 M(\gamma)$, and the data rate is $\log_2 M(\gamma) - 2r/N$. Therefore, we maximize the data rate by maximizing $\mathbf{E}[\log_2 M]$ relative to the average power constraint. From this maximization, we obtain the optimal power adaptation policy for this modulation scheme:

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} 1/\gamma_0 - 1/\gamma K_c & \gamma \geq \gamma_0/K_c, \\ 0 & \gamma < \gamma_0/K_c, \end{cases} \tag{9.41}$$

where $\gamma_{K_c} = \gamma_0/K_c$ is the cutoff fade depth for $K_c = KG_c$ with K given by (9.10). This is the same as the optimal policy for the uncoded case (9.13) with K replaced by K_c. Thus, the coded modulation increases the effective transmit power by G_c relative to the uncoded variable-rate variable-power MQAM performance. The adaptive data rate is obtained by substituting (9.41) into (9.40) to get

$$M(\gamma) = \left(\frac{\gamma}{\gamma_{K_c}}\right). \tag{9.42}$$

The resulting spectral efficiency is

$$\frac{R}{B} = \int_{\gamma K_C}^{\infty} \log_2 \left(\frac{\gamma}{\gamma_{K_c}}\right) p(\gamma)\, d\gamma. \tag{9.43}$$

If the constellation expansion factor is not included in the coding gain G_c, then we must subtract $2r/N$ from (9.43) to get the data rate. More details on this adaptive coded modulation scheme can be found in [33], along with plots of the spectral efficiency for adaptive trellis coded modulation of varying complexity. These results indicate that adaptive trellis coded modulation can achieve within 5 dB of Shannon capacity at reasonable complexity and that the coding gains of superimposing a given trellis code onto uncoded adaptive modulation are roughly equal to the coding gains of the trellis code in an AWGN channel.

9.4 General M-ary Modulations

The variable rate and power techniques already described for MQAM can be applied to other M-ary modulations. For any modulation, the basic premise is the same: the transmit power and constellation size are adapted to maintain a given fixed instantaneous BER for each symbol while maximizing average data rate. In this section we will consider optimal rate and power adaptation for both continuous-rate and discrete-rate variation of general M-ary modulations.

9.4.1 Continuous-Rate Adaptation

We first consider the case where both rate and power can be adapted continuously. We want to find the optimal power $P(\gamma)$ and rate $k(\gamma) = \log_2 M(\gamma)$ adaptation for general M-ary modulation that maximizes the average data rate $\mathbf{E}[k(\gamma)]$ with average power $\bar{P}$ while meeting a given BER target. This optimization is simplified when the exact or approximate probability of bit error for the modulation can be written in the following form:

$$P_b(\gamma) \approx c_1 \exp\left[\frac{-c_2\gamma(P(\gamma)/\bar{P})}{2^{c_3 k(\gamma)} - c_4}\right], \tag{9.44}$$

where c_1, c_2, and c_3 are positive fixed constants and c_4 is a real constant. For example, in the BER bounds for MQAM given by (9.6) and (9.7), $c_1 = 2$ or .2, $c_2 = 1.5, c_3 = 1$, and $c_4 = 1$. The probability of bit error for most M-ary modulations can be approximated in this form with appropriate curve fitting.

The advantage of (9.44) is that, when $P_b(\gamma)$ is in this form, we can invert it to express the rate $k(\gamma)$ as a function of the power adaptation $P(\gamma)$ and the BER target P_b as follows:

$$k(\gamma) = \log_2 M(\gamma) = \begin{cases} \frac{1}{c_3}\log_2\left[c_4 - \frac{c_2\gamma}{\ln(P_b/c_1)}\frac{P(\gamma)}{P}\right] & P(\gamma) \geq 0,\ k(\gamma) \geq 0, \text{ (9.45)} \\ 0 & \text{else.} \end{cases} \tag{9.45}$$

To find the power and rate adaptation that maximize spectral efficiency $\mathbf{E}[k(\gamma)]$, we create the Lagrangian

$$J(P(\gamma)) = \int_0^\infty k(\gamma)p(\gamma)d\gamma + \lambda\left[\int_0^\infty P(\gamma)p(\gamma)d\gamma - \bar{P}\right]. \tag{9.46}$$

The optimal adaptation policy maximizes this Lagrangian with nonnegative rate and power, so it satisfies

$$\frac{\partial J}{\partial P(\gamma)} = 0,\ P(\gamma) \geq 0,\ k(\gamma) \geq 0. \tag{9.47}$$

Solving (9.47) for $P(\gamma)$ with (9.45) for $k(\gamma)$ yields the optimal power adaptation

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} -1/c_3(\ln 2)\lambda\bar{P} - 1/(\gamma K) & P(\gamma) \geq 0,\ k(\gamma) \geq 0, \\ 0 & \text{else,} \end{cases} \tag{9.48}$$

where

$$K = -\frac{c_2}{c_4\ln(P_b/c_1)}. \tag{9.49}$$

The power adaptation (9.48) can be written in the more simplified form

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} \mu - 1/(\gamma K) & P(\gamma) \geq 0,\ k(\gamma) \geq 0, \\ 0 & \text{else.} \end{cases} \tag{9.50}$$

The constant μ in (9.50) is determined from the average power constraint (9.12).

Although the analytical expression for the optimal power adaptation (9.50) looks simple, its behavior is highly dependent on the c_4 values in the P_b approximation (9.44). For (9.44) given by the MQAM approximations (9.6) or (9.7), the power adaptation is the water-filling formula given by (9.13). However, water-filling is not optimal in all cases, as we now show.

Based on (6.16), with Gray coding the BER for MPSK is tightly approximated as

$$P_b \approx \frac{2}{\log_2 M}Q\left(\sqrt{2\gamma}\sin\left(\frac{\pi}{M}\right)\right). \tag{9.51}$$

However, (9.51) is not in the desired form (9.44). In particular, the Q-function is not easily inverted to obtain the optimal rate and power adaptation for a given target BER. Let us therefore consider the following three P_b bounds for MPSK, which are valid for $k(\gamma) \geq 2$.

$$\text{Bound 1: } P_b(\gamma) \approx .05\exp\left[\frac{-6\gamma(P(\gamma)/\bar{P})}{2^{1.9k(\gamma)} - 1}\right]. \tag{9.52}$$

$$\text{Bound 2: } P_b(\gamma) \approx .2\exp\left[\frac{-7\gamma(P(\gamma)/\bar{P})}{2^{1.9k(\gamma)} + 1}\right]. \tag{9.53}$$

$$\text{Bound 3: } P_b(\gamma) \approx .25\exp\left[\frac{-8\gamma(P(\gamma)/\bar{P})}{2^{1.94k(\gamma)}}\right]. \tag{9.54}$$

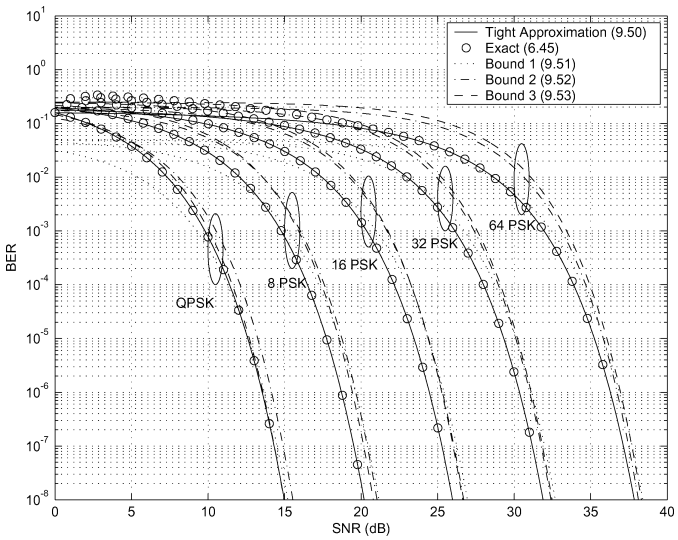

Figure 9.15: BER bounds for MPSK.

The bounds are plotted in Figure 9.15 along with the tight approximation (9.51). We see that all bounds well approximate the exact BER given by (6.45), especially at high SNRs.

In the first bound (9.52), $c_1 = .05, c_2 = 6, c_3 = 1.9$, and $c_4 = 1$. Thus, in (9.50), $K = -c_2/(c_4 \ln(P_b/c_1))$ is positive as long as the target P_b is less than .05, which we assume. Therefore μ must be positive for the power adaptation $P(\gamma)/\bar{P} = \mu - 1/(\gamma K)$ to be positive about a cutoff SNR γ_0. Moreover, for K positive, $k(\gamma) \geq 0$ for any $P(\gamma) \geq 0$. Thus, with μ and $k(\gamma)$ positive, (9.50) can be expressed as

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} 1/(\gamma_0 K) - 1/(\gamma K) & P(\gamma) \geq 0, \\ 0 & \text{else}, \end{cases} \tag{9.55}$$

where $\gamma_0 \geq 0$ is a cutoff fade depth below which no signal is transmitted. Like μ, this cutoff value is determined by the average power constraint (9.12). The power adaptation (9.55) is the same water-filling as in adaptive MQAM given by (9.13), which results from the similarity of the MQAM P_b bounds (9.7) and (9.6) to the MPSK bound (9.52). The corresponding optimal rate adaptation, obtained by substituting (9.55) into (9.45), is

$$k(\gamma) = \begin{cases} (1/c_3)\log_2(\gamma/\gamma_0) & \gamma \geq \gamma_0, \\ 0 & \text{else}, \end{cases} \tag{9.56}$$

which is also in the same form as the adaptive MQAM rate adaptation (9.16).

Let us now consider the second bound (9.53). Here $c_1 = .2, c_2 = 7, c_3 = 1.9$, and $c_4 = -1$. Thus, $K = -c_2/(c_4 \ln(P_b/c_1))$ is negative for a target $P_b < .2$, which we assume. From (9.45), with K negative we must have $\mu \geq 0$ in (9.50) to make $k(\gamma) \geq 0$. Then the optimal power adaptation such that $P(\gamma) \geq 0$ and $k(\gamma) \geq 0$ becomes

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} \mu - 1/(\gamma K) & k(\gamma) \geq 0, \\ 0 & \text{else}. \end{cases} \tag{9.57}$$

From (9.45), the optimal rate adaptation then becomes

$$k(\gamma) = \begin{cases} (1/c_3)\log_2(\gamma/\gamma_0) & \gamma \geq \gamma_0, \\ 0 & \text{else}, \end{cases} \tag{9.58}$$

where $\gamma_0 = -1/(K\mu)$ is a cutoff fade depth below which the channel is not used. Note that for the first bound (9.52) the positivity constraint on power ($P(\gamma) \geq 0$) dictates the cutoff fade depth, whereas for this bound (9.53) the positivity constraint on rate ($k(\gamma) \geq 0$) determines the cutoff. We can rewrite (9.57) in terms of γ_0 as

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} -1/(\gamma_0 K) - 1/(\gamma K) & \gamma \geq \gamma_0, \\ 0 & \text{else} \end{cases} \tag{9.59}$$

This power adaptation is an *inverse water-filling*: since K is negative, less power is used as the channel SNR increases above the optimized cutoff fade depth γ_0. As usual, the value of γ_0 is obtained based on the average power constraint (9.12).

Finally, for the third bound (9.54), $c_1 = .25$, $c_2 = 8$, $c_3 = 1.94$, and $c_4 = 0$. Thus, $K = -c_2/(c_4 \ln(P_b/c_1)) = \infty$ for a target $P_b < .25$, which we assume. From (9.50), the optimal power adaptation becomes

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} \mu & k(\gamma) \geq 0,\ P(\gamma) \geq 0, \\ 0 & \text{else}. \end{cases} \tag{9.60}$$

This is *on-off* power transmission: power is either zero or a constant nonzero value. By (9.45), the optimal rate adaptation $k(\gamma)$ with this power adaptation is

$$k(\gamma) = \begin{cases} (1/c_3)\log_2(\gamma/\gamma_0) & \gamma \geq \gamma_0, \\ 0 & \text{else}, \end{cases} \tag{9.61}$$

where $\gamma_0 = -\ln(P_b/c_1)/c_2\mu$ is a cutoff fade depth below which the channel is not used. As for the previous bound, it is the rate positivity constraint that determines the cutoff fade depth γ_0. The optimal power adaptation as a function of γ_0 is

$$\frac{P(\gamma)}{\bar{P}} = \begin{cases} K_0/\gamma_0 & \gamma \geq \gamma_0, \\ 0 & \text{else}, \end{cases} \tag{9.62}$$

where $K_0 = -\ln(P_b/c_1)/c_2$. The value of γ_0 is determined from the average power constraint to satisfy

$$\frac{K_0}{\gamma_0}\int_{\gamma_0}^{\infty} p(\gamma)\, d\gamma = 1. \tag{9.63}$$

Thus, for all three P_b approximations in MPSK, the optimal adaptive rate schemes (9.56), (9.58), and (9.61) have the same form whereas the optimal adaptive power schemes (9.55), (9.59), and (9.62) have different forms. The optimal power adaptations (9.55), (9.59), and (9.62) are plotted in Figure 9.16 for Rayleigh fading with a target BER of 10^{-3} and $\bar{\gamma} = 30$ dB. This figure clearly shows the water-filling, inverse water-filling, and on–off behavior of the different schemes. Note that the cutoff γ_0 for all these schemes is roughly the same. We also see from this figure that even though the power adaptation schemes are different at low SNRs, they are almost the same at high SNRs. Specifically, we see that for $\gamma < 10\,\text{dB}$, the optimal transmit power adaptations are dramatically different, whereas for $\gamma \geq 10\,\text{dB}$ they rapidly converge to the same constant value. From the cumulative distribution function of γ, also shown in Figure 9.16, the probability that γ is less than 10 is 0.01. Thus, although the optimal power adaptation corresponding to low SNRs is very different for the different techniques, this behavior has little impact on spectral efficiency because the probability of being at those low SNRs is quite small.

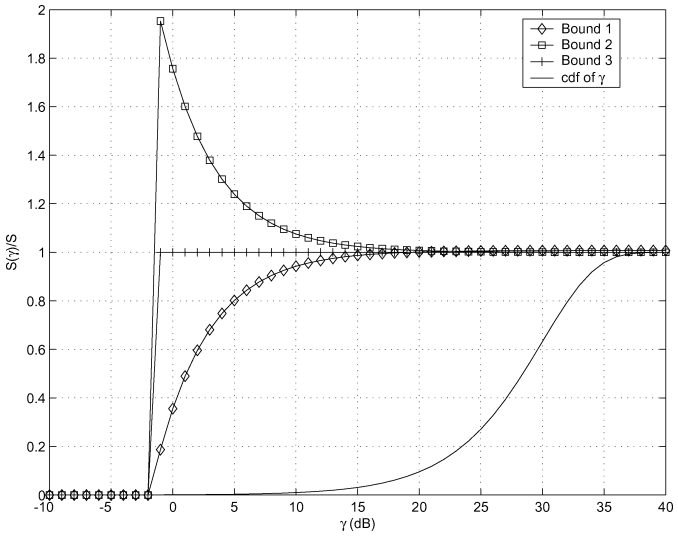

Figure 9.16: Power adaptation for MPSK BER bounds (Rayleigh fading, $P_{\rm b} = 10^{-3}, \bar{\gamma} = 30\,{\rm dB}$).

9.4.2 Discrete-Rate Adaptation

We now assume a given discrete set of constellations $\mathcal{M} = \{M_0 = 0,\ \ldots,\ M_{N-1}\}$, where M_0 corresponds to no data transmission. The rate corresponding to each of these constellations is $k_j = \log_2 M_j (j = 0, \ldots, N-1)$, where $k_0 = 0$. Each rate $k_j (j > 0)$ is assigned to a fading region of γ-values $R_j = [\gamma_{j-1}, \gamma_j), j = 0, \ldots, N-1$, for $\gamma - 1 = 0$ and $\gamma_{N-1} = \infty$. The boundaries $\gamma_j (j = 0, \ldots, N-2)$ are optimized as part of the adaptive policy. The channel is not used for $\gamma < \gamma_0$. We again assume that P_b is approximated using the general formula (9.44). Then the power adaptation that maintains the target BER above the cutoff γ_0 is

$$\frac{P(\gamma)}{\bar{P}} = \frac{h(k_j)}{\gamma},\ \gamma_{j-1} \le \gamma \le \gamma_j, \tag{9.64}$$

where

$$h(k_j) = -\frac{\ln(P_b/c_1)}{c_2}(2^{c_3 k_j} - c_4). \tag{9.65}$$

The region boundaries $\gamma_0, \ldots, \gamma_{N-2}$ that maximize spectral efficiency are found using the Lagrange equation

$$J(\gamma_0,\ \gamma_1, \ldots, \gamma_{N-2}) = \sum_{j=1}^{N-1} k_j \int_{\gamma_{j-1}}^{\gamma_j} p(\gamma)\, d\gamma + \lambda \left[\sum_{j=1}^{N-1} \int_{\gamma_{j-1}}^{\gamma_j} \frac{h(k_j)}{\gamma} p(\gamma) d\gamma - 1 \right]. \tag{9.66}$$

The optimal rate region boundaries are obtained by solving the following equation for γ_j:

$$\frac{\partial J}{\partial \gamma_j} = 0,\ 0 \le j \le N-2. \tag{9.67}$$

This yields

$$\gamma_0 = \frac{h(k_1)}{k_1}\rho \tag{9.68}$$

and

$$\gamma_j = \frac{h(k_{j+1}) - h(k_j)}{k_{j+1} - k_j}\rho,\ 1 \le i \le N-2, \tag{9.69}$$

where ρ is determined by the average power constraint

$$\sum_{j=1}^{N-1} \int_{\gamma_{j-1}}^{\gamma j} \frac{h(k_j)}{\gamma} p(\gamma)\, d\gamma = 1. \tag{9.70}$$

9.4.3 Average BER Target

Suppose now that we relax our assumption that the P_b target must be met on every symbol transmission, requiring instead that just the average P_b be below some target average $\bar{P}_b$. In this case, in addition to adapting rate and power, we can also adapt the instantaneous $P_b(\gamma)$ subject to the average constraint $\bar{P}_b$. This gives an additional degree of freedom in adaptation that may lead to higher spectral efficiencies. The average probability of error for adaptive modulation is given in (9.35).
When the bit rate $k(\gamma)$ is continuously adapted this becomes

$$\bar{P}_b = \frac{\int_0^\infty P_b(\gamma)k(\gamma)p(\gamma)\, d\gamma}{\int_0^\infty k(\gamma)p(\gamma)\, d\gamma}, \tag{9.71}$$

and when $k(\gamma)$ takes values in a discrete set this becomes

$$\bar{P}_b = \frac{\sum_{j=1}^{N-1} k_j \int_{\gamma_{j-1}}^{\gamma_j} P_b(\gamma)p(\gamma)\, d\gamma}{\sum_{j=1}^{N-1} k_j \int_{\gamma_{j-1}}^{\gamma_j} p(\gamma)\, d\gamma}. \tag{9.72}$$

We now derive the optimal continuous rate, power, and BER adaptation to maximize spectral efficiency $\mathbf{E}[k(\gamma)]$ subject to an average power constraint $\bar{P}$ and the average BER constraint (9.71). As with the instantaneous BER constraint, this is a standard constrained optimization problem, which we solve using the Lagrange method. We now require two Lagrangians for the two constraints: average power and average BER. Specifically, the Lagrange equation is

$$\begin{aligned} J(k(\gamma),\ P(\gamma)) = &\int_0^\infty k(\gamma)p(\gamma)\, d\gamma \\ &+ \lambda_1 \left[\int_0^\infty P_b(\gamma)k(\gamma)p(\gamma)\, d\gamma - \bar{P}_b \int_0^\infty k(\gamma)p(\gamma)\, d\gamma \right] \\ &+ \lambda_2 \left[\int_0^\infty P(\gamma)p(\gamma)\, d\gamma - \bar{P} \right]. \end{aligned} \tag{9.73}$$

The optimal rate and power adaptation must satisfy

$$\frac{\partial J}{\partial k(\gamma)} = 0 \text{ and } \frac{\partial J}{\partial P(\gamma)} = 0, \tag{9.74}$$

with the additional constraint that $k(\gamma)$ and $P(\gamma)$ be nonnegative for all γ.

Assume that P_b is approximated using the general formula (9.44). Define

$$f(k(\gamma)) = 2^{c_3 k(\gamma)} - c_4. \tag{9.75}$$

Then, using (9.44) in (9.73) and solving (9.74)), we obtain that the power and BER adaptation that maximize spectral efficiency satisfy

$$\frac{P(\gamma)}{\bar{P}} = \max\left[\frac{f(k(\gamma))}{\frac{\partial f(k(\gamma))}{\partial k(\gamma)}}\lambda_2\bar{P}(\lambda_1\bar{P}_b - 1) - \frac{f(k(\gamma))^2}{c_2\gamma\frac{\partial f(k(\gamma))}{\partial k(\gamma)}k(\gamma)}, 0\right] \tag{9.76}$$

for nonnegative $k(\gamma)$ and

$$P_b(\gamma) = \frac{\lambda_2\bar{P}f(k(\gamma))}{\lambda_1 c_2\gamma k(\gamma)}. \tag{9.77}$$

Moreover, from (9.44), (9.76), and (9.77) we get that the optimal rate adaptation $k(\gamma)$ is either zero or the nonnegative solution of

$$\frac{\lambda_1\bar{P}_b - 1}{\frac{\partial f(k(\gamma))}{\partial k(\gamma)}\lambda_2\bar{P}} - \frac{f(k(\gamma))}{c_2\gamma\frac{\partial f(k(\gamma))}{\partial k(\gamma)}k(\gamma)} = \frac{1}{\gamma c_2}\ln\left[\frac{\lambda_1 c_1 c_2\gamma k(\gamma)}{\lambda_2\bar{P}f(k(\gamma))}\right]. \tag{9.78}$$

The values of $k(\gamma)$ and the Lagrangians λ_1 and λ_2 must be found through a numerical search whereby the average power constraint $\bar{P}$ and average BER constraint (9.71) are satisfied.

In the discrete-rate case, the rate is varied within a fixed set $k_0, \ldots, k_{N-1}$, where k_0 corresponds to no data transmission. We must determine region boundaries $\gamma_0, \ldots, \gamma_{N-2}$ such that we assign rate k_j to the rate region $[\gamma_{j-1},\ \gamma_j)$, where $\gamma_{-1} = 0$ and $\gamma_{N-1} = \infty$. Under this rate assignment we wish to maximize spectral efficiency through optimal rate, power, and BER adaptation subject to an average power and BER constraint. Since the set of possible rates and their corresponding rate region assignments are fixed, the optimal rate adaptation corresponds to finding the optimal rate region boundaries $\gamma_j, j = 0, \ldots, N-2$. The Lagrangian for this constrained optimization problem is

$$\begin{aligned} J(\gamma_0, \gamma_1, \ldots, \gamma_{N-2},\ P(\gamma)) = &\sum_{j=1}^{N-1} k_j \int_{\gamma_{j-1}}^{\gamma_j} p(\gamma)\, d\gamma \\ &+ \lambda_1\left[\sum_{j=1}^{N-1} k_j \int_{\gamma_{j-1}}^{\gamma_j} (P_b(\gamma) - \bar{P}_b)p(\gamma)\, d\gamma\right] \\ &+ \lambda_2\left[\int_{\gamma_0}^{\infty} P(\gamma)p(\gamma)\, d\gamma - \bar{P}\right]. \end{aligned} \tag{9.79}$$

The optimal power adaptation is obtained by solving the following equation for $P(\gamma)$:

$$\frac{\partial J}{\partial P(\gamma)} = 0. \tag{9.80}$$

Similarly, the optimal rate region boundaries are obtained by solving the following set of equations for γ_j:

$$\frac{\partial J}{\partial \gamma_j} = 0,\ 0 \le j \le N-2. \tag{9.81}$$

From (9.80) we see that the optimal power and BER adaptation must satisfy

$$\frac{\partial P_b(\gamma)}{\partial P(\gamma)} = \frac{-\lambda_2}{k_j\lambda_1}, \gamma_{j-1} \le \gamma \le \gamma_j. \tag{9.82}$$

Substituting (9.44) into (9.82), we get that

$$P_b(\gamma) = \lambda \frac{f(k_j)}{\gamma k_j}, \gamma_{j-1} \leq \gamma \leq \gamma_j, \tag{9.83}$$

where $\lambda = \bar{P}\lambda_2/c_2\lambda_1$. This form of BER adaptation is similar to the water-filling power adaptation: the instantaneous BER decreases as the channel quality improves. Now, setting the BER in (9.44) equal to (9.83) and solving for $P(\gamma)$ yields

$$P(\gamma) = P_j(\gamma), \gamma_{j-1} \leq \gamma \leq \gamma_j, \tag{9.84}$$

where

$$\frac{P_j(\gamma)}{\bar{P}} = \ln\left[\frac{\lambda f(k_j)}{c_1 \gamma k_j}\right] \frac{f(k_j)}{-\gamma c_2},\ 1 \leq j \leq N-1, \tag{9.85}$$

and $P(\gamma) = 0$ for $\gamma < \gamma_0$. We see by (9.85) that $P(\gamma)$ is discontinuous at the γ_j boundaries.

Let us now consider the optimal region boundaries $\gamma_0, \ldots, \gamma_{N-2}$. Solving (9.81) for $P_b(\gamma_j)$ yields

$$P_b(\gamma_j) = \bar{P}_b - \frac{1}{\lambda_1} - \frac{\lambda_2}{\lambda_1} \frac{P_{j+1}(\gamma_j) - P_j(\gamma_j)}{k_{j+1} - k_j},\ 0 \leq j \leq N-2, \tag{9.86}$$

where $k_0 = 0$ and $P_0(\gamma) = 0$. Unfortunately, this set of equations can be difficult to solve for the optimal boundary points $\{\gamma_j\}$. However, if we assume that $P(\gamma)$ is continuous at each boundary, then (9.86) becomes

$$P_b(\gamma_j) = \bar{P}_b - \frac{1}{\lambda_1},\ 0 \leq j \leq N-2, \tag{9.87}$$

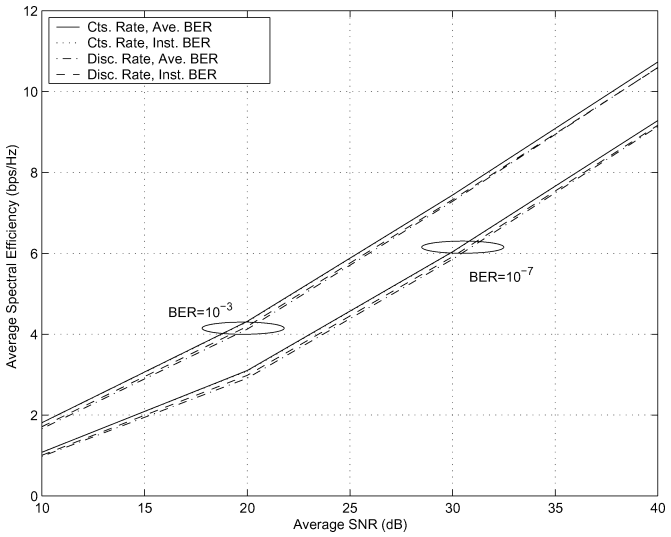

Figure 9.17: Spectral efficiency for different adaptation constraints.

for the Lagrangian λ_1. Under this assumption we can solve for the suboptimal rate region boundaries as

$$\gamma_{j-1} = \frac{f(k_j)}{k_j}\rho,\ 1 \leq j \leq N-1, \tag{9.88}$$

for some constant ρ. The constants λ_1 and ρ are found numerically such that the average power constraint

$$\sum_{j=1}^{N-1} \int_{\gamma_{j-1}}^{\gamma_j} \frac{P_j(\gamma)}{\bar{P}} p(\gamma)\, d\gamma = 1 \tag{9.89}$$

and the BER constraint (9.72) are satisfied. Note that the region boundaries (9.88) are suboptimal because $P(\gamma)$ is not necessarily continuous at the boundary regions, so these boundaries yield a suboptimal spectral efficiency.

In Figure 9.17 we plot average spectral efficiency for adaptive MQAM under both continuous and discrete rate adaptation, showing both average and instantaneous BER targets for a Rayleigh fading channel. The adaptive policies are based on the BER approximation (9.7) with a target BER of either 10^{-3} or 10^{-7}. For the discrete-rate cases we assume that six different MQAM signal constellations are available (seven fading regions), given by $\mathcal{M} = \{0, 4, 16, 64, 256, 1024, 4096\}$. We see in this figure that the spectral efficiencies of all four policies under the same instantaneous or average BER target are very close to each other. For discrete-rate adaptation, the spectral efficiency with an instantaneous BER target is slightly higher than with an average BER target even though the latter case is more constrained: that is because the efficiency with an average BER target is calculated using suboptimal rate region boundaries, which leads to a slight efficiency degradation.

9.5 Adaptive Techniques in Combined Fast and Slow Fading

In this section we examine adaptive techniques for composite fading channels consisting of both fast and slow fading (shadowing). We assume the fast fading changes too quickly to accurately measure and feed back to the transmitter, so the transmitter adapts only to the slow fading. The instantaneous SNR γ has distribution $p(\gamma|\bar{\gamma})$, where $\bar{\gamma}$ is a short-term average over the fast fading. This short-term average varies slowly because of shadowing and has a distribution $p(\bar{\gamma})$, where the average SNR relative to this distribution is $\bar{\bar{\gamma}}$. The transmitter adapts only to the slow fading $\bar{\gamma}$, hence its rate $k(\bar{\gamma})$ and power $P(\bar{\gamma})$ are functions of $\bar{\gamma}$. The power adaptation is subject to a long-term average power constraint over both the fast and slow fading:

$$\int_0^\infty P(\bar{\gamma}) p(\bar{\gamma})\, d\bar{\gamma} = \bar{P}. \tag{9.90}$$

As before, we approximate the instantaneous probability of bit error by the general form (9.44). Since the power and rate are functions of $\bar{\gamma}$, the conditional BER, conditioned on $\bar{\gamma}$, is

$$P_b(\gamma \mid \bar{\gamma}) \approx c_1 \exp\left[\frac{-c_2 \gamma P(\bar{\gamma})/\bar{P}}{2^{c_3 k(\bar{\gamma})} 3 - c_4}\right]. \tag{9.91}$$

Since the transmitter does not adapt to the fast fading γ, we cannot require a given instantaneous BER. However, since the transmitter adapts to the shadowing, we can require a target average probability of bit error averaged over the fast fading for a fixed value of the shadowing. This short-term average for a given $\bar{\gamma}$ is obtained by averaging $P_b(\gamma \mid \bar{\gamma})$ over the fast fading distribution $p(\gamma \mid \bar{\gamma})$:

$$\bar{P}_b(\bar{\gamma}) = \int_0^\infty P_b(\gamma \mid \bar{\gamma}) p(\gamma \mid \bar{\gamma})\, d\gamma. \tag{9.92}$$

Using (9.91) in (9.92) and assuming Rayleigh fading for the fast fading, this becomes

$$\bar{P}_b(\bar{\gamma}) = \frac{1}{\bar{\gamma}} \int_0^\infty c_1 \exp\left[\frac{-c_2 \gamma P(\bar{\gamma})/\bar{P}}{2^{c_3 k(\bar{\gamma})} - c_4} - \frac{\gamma}{\bar{\gamma}}\right] d\gamma = \frac{c_1}{\frac{1.5 \bar{\gamma} P(\bar{\gamma})/\bar{P}}{2^{c_3 k(\bar{\gamma})} - c_4} + 1}. \tag{9.93}$$

For example, with MQAM modulation with the tight BER bound (9.7), equation (9.93) becomes

$$\bar{P}_b(\bar{\gamma}) = \frac{.2}{\frac{c_2\bar{\gamma}P(\bar{\gamma})/\bar{P}}{2^{k(\bar{\gamma})}-1}+1}. \tag{9.94}$$

We can now invert (9.93) to obtain the adaptive rate $k(\bar{\gamma})$ as a function of the target average BER $\bar{P}_b$ and the power adaptation $P(\bar{\gamma})$:

$$k(\bar{\gamma}) = \frac{1}{c_3}\log_2\left(c_4 + \frac{K\bar{\gamma}P(\bar{\gamma})}{\bar{P}}\right), \tag{9.95}$$

where

$$K = \frac{c_2}{c_1/\bar{P}_b - 1} \tag{9.96}$$

depends only on the target average BER and decreases as this target decreases. We maximize spectral efficiency by maximizing

$$\mathbf{E}[k(\bar{\gamma})] = \int_0^\infty \frac{1}{c_3}\log_2\left(c_4 + \frac{K\bar{\gamma}P(\bar{\gamma})}{\bar{P}}\right)p(\bar{\gamma})\,d\bar{\gamma} \tag{9.97}$$

subject to the average power constraint (9.90).

Let us assume that $c_4 > 0$. Then this maximization and the power constraint are in the exact same form as (9.11) with the fading γ replaced by the slow fading $\bar{\gamma}$. Thus, the optimal power adaptation also has the same water-filling form as (9.13) and is given by

$$\frac{P(\bar{\gamma})}{\bar{P}} = \begin{cases} 1/\bar{\gamma}_0 - c_4/\bar{\gamma}K & \bar{\gamma} \geq c_4\bar{\gamma}_0/K, \\ 0 & \bar{\gamma} < c_4\bar{\gamma}_0/K, \end{cases} \tag{9.98}$$

where the channel is not used when $\bar{\gamma} < c_4\bar{\gamma}_0/K$. The value of $\bar{\gamma}_0$ is determined by the average power constraint. Substituting (9.98) into (9.95) yields the rate adaptation

$$k(\bar{\gamma}) = \frac{1}{c_3}\log_2\left(\frac{K\bar{\gamma}}{\bar{\gamma}_0}\right), \tag{9.99}$$

and the corresponding average spectral efficiency is given by

$$\frac{R}{B} = \int_{c_4\gamma_0/K}^\infty \log_2(\frac{K\bar{\gamma}}{\bar{\gamma}_0})p(\bar{\gamma})\,d\bar{\gamma}. \tag{9.100}$$

Thus, we see that in a composite fading channel where rate and power are adapted only to the slow fading and for $c_4 > 0$ in (9.44), water-filling *relative to the slow fading* is the optimal power adaptation for maximizing spectral efficiency subject to an average BER constraint.

Our derivation has assumed that the fast fading is Rayleigh; however, it can be shown [34] that, with $c_4 > 0$ in (9.44), the optimal power and rate adaptation for any fast fading distribution have the same form. Since we have assumed $c_4 > 0$ in (9.44), the positivity constraint on power dictates the cutoff value below which the channel is not used. As we saw in Section 9.4.1, when $c_4 \leq 0$ the positivity constraint on rate dictates this cutoff, and the optimal power adaptation becomes inverse water-filling for $c_4 < 0$ and on–off power adaptation for $c_4 = 0$.

Chapter 9 Problems

1. Find the average SNR required to achieve an average BER of $\bar{P}_b = 10^{-3}$ for 8-PSK modulation in Rayleigh fading. What is the spectral efficiency of this scheme, assuming a symbol time of $T_s = 1/B$?

2. Consider a truncated channel inversion variable-power technique for Rayleigh fading with average SNR of 20 dB. What value of σ corresponds to an outage probability of .1? Find the maximum size MQAM constellation that can be transmitted under this policy so that, in non-outage, $P_b \approx 10^{-3}$.

3. Find the power adaptation for QPSK modulation that maintains a fixed $P_b = 10^{-3}$ in non-outage for a Rayleigh fading channel with $\bar{\gamma} = 20$ dB. What is the outage probability of this system?

4. Consider a variable-rate MQAM modulation scheme with just two constellations, $M = 4$ and $M = 16$. Assume a target P_b of approximately 10^{-3}. If the target cannot be met then no data is transmitted.

   (a) Using the BER bound (9.7), find the range of γ-values associated with the three possible transmission schemes (no transmission, 4-QAM, and 16-QAM) where the BER target is met. What is the cutoff γ_0 below which the channel is not used?

   (b) Assuming Rayleigh fading with $\bar{\gamma} = 20$ dB, find the average data rate of the variable-rate scheme.

   (c) Suppose that, instead of suspending transmission below γ_0, BPSK is transmitted for $0 \leq \gamma \leq \gamma_0$. Using the loose bound (9.6), find the average probability of error for this BPSK transmission.

5. Consider an adaptive modulation and coding scheme consisting of three modulations: BPSK, QPSK, and 8-PSK, along with three block codes of rate 1/2, 1/3, and 1/4. Assume that the first code provides roughly 3 dB of coding gain for each modulation type, the second code 4 dB, and the third code 5 dB. For each possible value of SNR $0 \leq \gamma \leq \infty$, find the combined coding and modulation with the maximum data rate for a target BER of 10^{-3} (you can use any reasonable approximation for modulation BER in this calculation, with SNR increased by the coding gain). Find the average data rate of the system for a Rayleigh fading channel with average SNR of 20 dB, assuming no transmission if the target BER cannot be met with any combination of modulation and coding.

6. Show that the spectral efficiency given by (9.11) with power constraint (9.12) is maximized by the water-filling power adaptation (9.13). Do this by setting up the Lagrangian equation, differentiating it, and solving for the maximizing power adaptation. Also show that, with this power adaptation, the rate adaptation is as given in (9.16).

7. In this problem we compare the spectral efficiency of nonadaptive techniques with that of adaptive techniques.

   (a) Using the tight BER bound for MQAM modulation given by (9.7), find an expression for the average probability of bit error in Rayleigh fading as a function of M and $\bar{\gamma}$.

   (b) Based on the expression found in part (a), find the maximum constellation size that can be transmitted over a Rayleigh fading channel with a target average BER of 10^{-3}, assuming $\bar{\gamma} = 20$ dB.

   (c) Compare the spectral efficiency of part (b) with that of adaptive modulation shown in Figure 9.3 for the same parameters. What is the spectral efficiency difference between the adaptive and nonadaptive techniques?

8. Consider a Rayleigh fading channel with an average SNR of 20 dB. Assume a target BER of 10^{-4}.

(a) Find the optimal rate and power adaptation for variable-rate variable-power MQAM as well as the cutoff value γ_0/K below which the channel is not used.

(b) Find the average spectral efficiency for the adaptive scheme derived in part (a).

(c) Compare your answer in part (b) to the spectral efficiency of truncated channel inversion, where γ_0 is chosen to maximize this efficiency.

9. Consider a discrete time-varying AWGN channel with four channel states. Assuming a fixed transmit power $\bar{P}$, the received SNR associated with each channel state is $\gamma_1 = 5$ dB, $\gamma_2 = 10$ dB, $\gamma_3 = 15$ dB, and $\gamma_4 = 20$ dB, respectively. The probabilities associated with the channel states are $p(\gamma_1) = .4$ and $p(\gamma_2) = p(\gamma_3) = p(\gamma_4) = .2$. Assume a target BER of 10^{-3}.

(a) Find the optimal power and rate adaptation for continous-rate adaptive MQAM on this channel.

(b) Find the average spectral efficiency with this optimal adaptation.

(c) Find the truncated channel inversion power control policy for this channel and the maximum data rate that can be supported with this policy.

10. Consider a Rayleigh fading channel with an average received SNR of 20 dB and a required BER of 10^{-3}. Find the spectral efficiency of this channel using truncated channel inversion, assuming the constellation is restricted to size $0, 2, 4, 16, 64$, or 256.

11. Consider a Rayleigh fading channel with an average received SNR of 20 dB, a Doppler frequency of 80 Hz, and a required BER of 10^{-3}.

(a) Suppose you use adaptive MQAM modulation on this channel with constellations restricted to size $0, 2, 4, 16$, and 64. Using $\gamma_K^* = .1$, find the fading regions R_j associated with each of these constellations. Also find the average spectral efficiency of this restricted adaptive modulation scheme and the average time spent in each region R_j. If the symbol rate is $T_s = B^{-1}$, over approximately how many symbols is each constellation transmitted before a change in constellation size is needed?

(b) Find the exact BER of your adaptive scheme using (9.33). How does it differ from the target BER?

12. Consider a Rayleigh fading channel with an average received SNR of 20 dB, a signal bandwidth of 30 kHz, a Doppler frequency of 80 Hz, and a required BER of 10^{-3}.

(a) Suppose the estimation error $\varepsilon = \hat{\gamma}/\gamma$ in a variable-rate variable-power MQAM system with a target BER of 10^{-3} is uniformly distributed between .5 and 1.5. Find the resulting average probability of bit error for this system.

(b) Find an expression for the average probability of error in a variable-rate variable-power MQAM system in which the SNR estimate $\hat{\gamma}$ available at the transmitter is both a delayed and noisy estimate of γ : $\hat{\gamma}(t) = \gamma(t - \tau) + \gamma_\varepsilon(t)$. What joint distribution is needed to compute this average?

13. Consider an adaptive trellis-coded MQAM system with a coding gain of 3 dB. Assume a Rayleigh fading channel with an average received SNR of 20 dB. Find the optimal adaptive power and rate policy for this system and the corresponding average spectral efficiency. Assume a target BER of 10^{-3}.

14. In Chapter 6, a bound on P_b for nonrectangular MQAM was given as

$$P_b \approx \frac{4}{\log_2 M} Q\left(\sqrt{\frac{3\gamma}{(M-1)}}\right).$$

Find values for c_1, c_2, c_3, and c_4 for the general BER form (9.44) to approximate this bound with $M = 8$. Any curve approximation technique is acceptable. Plot both BER formulas for $0 \leq \gamma \leq 30$ dB.

15. Show that the average spectral efficiency $\mathbf{E}[k(\gamma)]$ for $k(\gamma)$ given by (9.45) with power constraint $\bar{P}$ is maximized by the power adaptation (9.48).

16. In this problem we investigate the optimal adaptive modulation for MPSK modulation based on the three BER bounds (9.52), (9.53), and (9.54). We assume a Rayleigh fading channel (so that γ is exponentially distributed with $\bar{\gamma} = 30$ dB) and a target BER of $P_b = 10^{-7}$.

    (a) The cutoff fade depth γ_0 must satisfy

    $$\int_{\gamma_0/K}^{\infty} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma K}\right) p(\gamma)\, d\gamma \leq 1$$

    for K given by (9.49). Find the cutoff value γ_0 corresponding to the power adaptation for each of the three bounds.

    (b) Plot $P(\gamma)/\bar{P}$ and $k(\gamma)$ as a function of γ for Bounds 1, 2, and 3 for γ ranging from 0 dB to 30 dB. Also state whether the cutoff value below which the channel is not used is based on the power or rate positivity constraint.

    (c) How does the power adaptation associated with the different bounds differ at low SNRs? What about at high SNRs?

17. Show that, under discrete rate adaptation for general M-ary modulation, the power adaptation that maintains a target instantaneous BER is given by (9.64). Also show that the region boundaries that maximize spectral efficiency–obtained using the Lagrangin given in (9.66) – are given by (9.68) and (9.69).

18. Show that, for general M-ary modulation with an average target BER, the Lagrangian (9.80) implies that the optimal power and BER adaptation must satisfy (9.82). Then show how (9.82) leads to BER adaptation given by (9.83), which in turn leads to the power adaptation given by (9.84) and (9.85). Finally, use (9.81) to show that the optimal rate region boundaries must satisfy (9.86).

19. Consider adaptive MPSK where the constellation is restricted to either no transmission or $M = 2, 4, 8, 16$. Assume the probability of error is approximated using (9.52). Find and plot the optimal discrete-rate continuous-power adaptation for $0 \leq \gamma \leq 30$ dB assuming a Rayleigh fading channel with $\bar{\gamma} = 20$ dB and a target P_b of 10^{-4}. What is the resulting average spectral efficiency?

20. We assume the same discrete-rate adaptive MPSK as in the previous problem, except now there is an average target P_b of 10^{-4} instead of an instantaneous target. Find the optimal discrete-rate continuous-power adaptation for a Rayleigh fading channel with $\bar{\gamma} = 20$ dB and the corresponding average spectral efficiency.

21. Consider a composite fading channel with fast Rayleigh fading and slow log-normal shadowing with an average dB SNR $\mu_{\psi_{\rm dB}} = 20$ dB (averaged over both fast and slow fading) and $\sigma_{\psi_{\rm dB}} = 8$ dB. Assume an adaptive MPSK modulation that adapts only to the shadowing, with a target average BER of 10^{-3}. Using the BER approximation (9.52), find the optimal power and rate adaptation policies as a function of the slow fading $\bar{\gamma}$ that maximize average spectral efficiency while meeting the average BER target. Also determine the average spectral efficiency that results from these policies.

22. In Section 9.5 we determined the optimal adaptive rate and power policies to maximize average spectral efficiency while meeting a target average BER in combined Rayleigh fading and shadowing. The derivation assumed the general bound (9.44) with $c_4 > 0$. For the same composite channel, find the optimal adaptive rate and power policies to maximize average spectral efficiency while meeting a target average BER assuming $c_4 < 0$. *Hint:* The derivation is similar to the case of continuous-rate adaptation using the second MPSK bound and results in the same channel inversion power control.

23. As in the previous problem, we again examine the adaptative rate and power policies to maximize average spectral efficiency while meeting a target average BER in combined Rayleigh fading and shadowing. In this problem we assume the general bound (9.44) with $c_4 = 0$. For the composite channel, find the optimal adaptive rate and power policies to maximize average spectral efficiency while meeting a target average BER assuming $c_4 = 0$. *Hint*: the derivation is similar to that of Section 9.4.1 for the third MPSK bound and results in the same on-off power control.

Bibliography

[1] J. F. Hayes, "Adaptive feedback communications," *IEEE Trans. Commun. Tech*., pp. 29–34, February 1968.

[2] J. K. Cavers, "Variable-rate transmission for Rayleigh fading channels," *IEEE Trans. Commun*., pp. 15–22, February 1972.

[3] S. Otsuki, S. Sampei, and N. Morinaga, "Square-QAM adaptive modulation/TDMA/TDD systems using modulation level estimation with Walsh function," *Elec. Lett*., pp. 169–71, February 1995.

[4] W. T. Webb and R. Steele, "Variable rate QAM for mobile radio," *IEEE Trans. Commun*., pp. 2223–30, July 1995.

[5] Y. Kamio, S. Sampei, H. Sasaoka, and N. Morinaga, "Performance of modulation-level-controlled adaptive-modulation under limited transmission delay time for land mobile communications," *Proc. IEEE Veh. Tech. Conf*., pp. 221–5, July 1995.

[6] B. Vucetic, "An adaptive coding scheme for time-varying channels," *IEEE Trans. Commun*., pp. 653–63, May 1991.

[7] M. Rice and S. B. Wicker, "Adaptive error control for slowly varying channels," *IEEE Trans. Commun*., pp. 917–26, February–April 1994.

[8] S. M. Alamouti and S. Kallel, "Adaptive trellis-coded multiple-phased-shift keying for Rayleigh fading channels," *IEEE Trans. Commun*., pp. 2305–14, June 1994.

[9] T. Ue, S. Sampei, and N. Morinaga, "Symbol rate and modulation level controlled adaptive modulation/TDMA/TDD for personal communication systems," *Proc. IEEE Veh. Tech. Conf*., pp. 306–10, July 1995.

[10] H. Matsuoka, and Y. Kamio, "Symbol rate and modulation level controlled adaptive modulation/TDMA/TDD for personal communication systems," *Proc. IEEE Veh. Tech. Conf*., pp. 487–91, April 1996.

[11] S. Sampei, N. Morinaga, and Y. Kamio, "Adaptive modulation/TDMA with a BDDFE for 2 Mbit/s multimedia wireless communication systems," *Proc. IEEE Veh. Tech. Conf*., pp. 311–15, July 1995.

[12] S. T. Chung and A. J. Goldsmith, "Degrees of freedom in adaptive modulation: A unified view," *IEEE Trans. Commun*., pp. 1561–71, September 2001.

[13] A. Furuskar, S. Mazur, F. Muller, and H. Olofsson, "EDGE: Enhanced data rates for GSM and TDMA/136 evolution," *IEEE Wireless Commun. Mag*., pp. 56–66, June 1999.

[14] A. Ghosh, L. Jalloul, B. Love, M. Cudak, and B. Classon, "Air-interface for 1XTREME/lxEV-DV," *Proc. IEEE Veh. Tech. Conf*., pp. 2474–8, May 2001.

[15] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Commun. Mag*., pp. 54–64, January 2000.

[16] H. Sari, "Trends and challenges in broadband wireless access," *Proc. Sympos. Commun. Veh. Tech*., pp. 210–14, October 2000.

[17] K. M. Kamath and D. L. Goeckel, "Adaptive-modulation schemes for minimum outage probability in wireless systems," *IEEE Trans. Commun*., pp. 1632–5, October 2004.

[18] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun*., pp. 389–400, April 1988.

[19] G. J. Foschini and J. Salz, "Digital communications over fading radio channels," *Bell System Tech. J.,* pp. 429–56, February 1983.

[20] G. D. Forney, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff, and S. U. Quereshi, "Efficient modulation for band-limited channels," *IEEE J. Sel. Areas Commun*., pp. 632–47, September 1984.

[21] J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.

[22] M. Filip and E. Vilar, "Implementation of adaptive modulation as a fade countermeasure," *Internat. J. Sat. Commun*., pp. 181–91, 1994.

[23] C. C. Tan and N. C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Trans. Commun*., pp. 2032–40, December 2000.

[24] L. Kleinrock, *Queueing Systems*, vol. I: *Theory*, Wiley, New York, 1975.

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

[26] H. S. Wang and P.-C. Chang, "On verifying the first-order Markov assumption for a Rayleigh fading channel model," *IEEE Trans.Veh.Tech*., pp. 353–7, May 1996.

[27] X. Tang, M.-S. Alouini, and A. Goldsmith, " Effect of channel estimation error on M-QAM BER performance in Rayleigh fading," *IEEE Trans. Commun*., pp. 1856–64, December 1999.

[28] J. K. Cavers, "An analysis of pilot symbol assisted modulation for Rayleigh fading channels," *IEEE Trans.Veh.Tech*., pp. 686–93, November 1991.

[29] A. J. Goldsmith and L. J. Greenstein, "Effect of average power estimation error on adaptive MQAM modulation," *Proc. IEEE Internat. Conf. Commun*., pp. 1105–9, June 1997.

[30] D. L. Goeckel, "Adaptive coding for time-varying channels using outdated fading estimates," *IEEE Trans. Commun*., pp. 844–55, June 1999.

[31] A. Duel-Hallen, S. Hu, and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Proc. Mag*., pp. 62–75, May 2000.

[32] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer J. Wireless Pers. Commun*., pp. 119–43, May 2000.

[33] S.-G. Chua and A. J. Goldsmith, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun*., pp. 595–602, May 1998.

[34] S. Vishwanath, S. A. Jafar, and A. J. Goldsmith, "Adaptive resource allocation in composite fading environments," *Proc. IEEE Globecom Conf.*, pp. 1312–16, November 2001.