WIRELESS COMMUNICATIONS

Andrea Goldsmith

Stanford University

Contents

1 Overview of Wireless Communications **1**

1.1 History of Wireless Communications 1
1.2 Wireless Vision 5
1.3 Technical Issues 8
1.4 Current Wireless Systems 10
1.4.1 Cellular Telephone Systems 10
1.4.2 Cordless Phones 14
1.4.3 Wireless LANs 15
1.4.4 Wide Area Wireless Data Services 16
1.4.5 Fixed Wireless Access 17
1.4.6 Paging Systems 17
1.4.7 Satellite Networks 18
1.4.8 Bluetooth 18
1.4.9 HomeRF 19
1.4.10 Other Wireless Systems and Applications 19
1.5 The Wireless Spectrum 20
1.5.1 Methods for Spectrum Allocation 20
1.5.2 Spectrum Allocations for Existing Systems 20
1.6 Standards 21

2 Path Loss and Shadowing **25**

2.1 Radio Wave Propagation 26
2.2 Transmit and Receive Signal Models 27
2.3 Free-Space Path Loss 29
2.4 Ray Tracing 30
2.4.1 Two-Ray Model 31
2.4.2 Dielectric Canyon (Ten-Ray Model) 34
2.4.3 General Ray Tracing 35
2.5 Simplified Path Loss Model 38
2.6 Empirical Path Loss Models 40
2.6.1 Okumura's Model 40
2.6.2 Hata Model 41
2.6.3 COST231 Extension to Hata Model 41
2.6.4 Walfisch/Bertoni Model 42
2.6.5 Piecewise Linear (Multi-Slope) Model 42
2.6.6 Indoor Propagation Models 43

2.7 Shadow Fading . . . 44
2.8 Combined Path Loss and Shadowing . . . 47
2.9 Outage Probability under Path Loss and Shadowing . . . 47
2.10 Cell Coverage Area . . . 48

3 Statistical Multipath Channel Models **63**
3.1 Time-Varying Channel Impulse Response . . . 63
3.2 Narrowband fading models . . . 68
3.2.1 Autocorrelation, Cross Correlation, and Power Spectral Density . . . 69
3.2.2 Envelope and Power Distributions . . . 74
3.2.3 Level Crossing Rate and Average Fade Duration . . . 76
3.2.4 Finite State Markov Models . . . 78
3.3 Wideband Fading Models . . . 79
3.3.1 Power Delay Profile . . . 82
3.3.2 Coherence Bandwidth . . . 84
3.3.3 Doppler Power Spectrum and Channel Coherence Time . . . 86
3.3.4 Transforms for Autocorrelation and Scattering Functions . . . 87
3.4 Discrete-Time Model . . . 88
3.5 Spatio-Temporal Models . . . 89

4 Capacity of Wireless Channels **97**
4.1 Introduction . . . 97
4.2 Capacity in AWGN . . . 98
4.3 Capacity of Flat-Fading Channels . . . 99
4.3.1 Channel and System Model . . . 99
4.3.2 Channel Distribution Information (CDI) Known . . . 100
4.3.3 Channel Side Information at Receiver . . . 101
4.3.4 Channel Side Information at the Transmitter and Receiver . . . 104
4.3.5 Capacity with Receiver Diversity . . . 110
4.3.6 Capacity Comparisons . . . 110
4.4 Capacity of Frequency-Selective Fading Channels . . . 113
4.4.1 Time-Invariant Channels . . . 113
4.4.2 Time-Varying Channels . . . 115

5 Digital Modulation and Detection **125**
5.1 Signal Space Analysis . . . 126
5.1.1 Signal and System Model . . . 126
5.1.2 Geometric Representation of Signals . . . 128
5.1.3 Receiver Structure and Sufficient Statistics . . . 130
5.1.4 Decision Regions and the Maximum Likelihood Decision Criterion . . . 133
5.1.5 Error Probability and the Union Bound . . . 135
5.2 Passband Modulation Principles . . . 140
5.3 Amplitude and Phase Modulation . . . 140
5.3.1 Pulse Amplitude Modulation (MPAM) . . . 142
5.3.2 Phase Shift Keying (MPSK) . . . 143
5.3.3 Quadrature Amplitude Modulation (MQAM) . . . 145
5.3.4 Differential Modulation . . . 146

5.3.5 Constellation Shaping . . . 149
5.3.6 Quadrature Offset . . . 150
5.4 Frequency Modulation . . . 150
5.4.1 Frequency Shift Keying (FSK) and Minimum Shift Keying (MSK) . . . 151
5.4.2 Continuous-Phase FSK (CPFSK) . . . 152
5.4.3 Noncoherent Detection of FSK . . . 153
5.5 Pulse Shaping . . . 154
5.6 Symbol Synchronization and Carrier Phase Recovery . . . 157
5.6.1 Receiver Structure with Phase and Timing Recovery . . . 157
5.6.2 Maximum Likelihood Phase Estimation . . . 159
5.6.3 Maximum-Likelihood Timing Estimation . . . 161

6 Performance of Digital Modulation over Wireless Channels 171
6.1 AWGN Channels . . . 171
6.1.1 Signal-to-Noise Power Ratio and Bit/Symbol Energy . . . 171
6.1.2 Error Probability for BPSK and QPSK . . . 172
6.1.3 Error Probability for MPSK . . . 174
6.1.4 Error Probability for MPAM and MQAM . . . 175
6.1.5 Error Probability for FSK and CPFSK . . . 177
6.1.6 Error Probability Approximation for Coherent Modulations . . . 178
6.1.7 Error Probability for Differential Modulation . . . 178
6.2 Alternate Q Function Representation . . . 180
6.3 Fading . . . 180
6.3.1 Outage Probability . . . 181
6.3.2 Average Probability of Error . . . 182
6.3.3 Moment Generating Function Approach to Average Error Probability . . . 184
6.3.4 Combined Outage and Average Error Probability . . . 188
6.4 Doppler Spread . . . 189
6.5 Intersymbol Interference . . . 191

7 Diversity 203
7.1 Realization of Independent Fading Paths . . . 203
7.2 Diversity System Model . . . 204
7.3 Selection Combining . . . 206
7.4 Threshold Combining . . . 208
7.5 Maximal Ratio Combining . . . 211
7.6 Equal-Gain Combining . . . 212
7.7 Moment Generating Functions in Diversity Analysis . . . 214
7.7.1 Diversity Analysis for MRC . . . 214
7.7.2 Diversity Analysis for EGC and SC . . . 218
7.7.3 Diversity Analysis for Noncoherent and Differentially Coherent Modulation . . . 218
7.8 Transmitter Diversity . . . 218

8 Coding for Wireless Channels 225
8.1 Code Design Considerations . . . 225
8.2 Linear Block Codes . . . 226
8.2.1 Binary Linear Block Codes . . . 227

8.2.2 Generator Matrix . . . 228
8.2.3 Parity Check Matrix and Syndrome Testing . . . 230
8.2.4 Cyclic Codes . . . 231
8.2.5 Hard Decision Decoding (HDD) . . . 234
8.2.6 Probability of Error for HDD in AWGN . . . 235
8.2.7 Probability of Error for SDD in AWGN . . . 238
8.2.8 Common Linear Block Codes . . . 240
8.2.9 Nonbinary Block Codes: the Reed Solomon Code . . . 240
8.2.10 Block Coding and Interleaving for Fading Channels . . . 241
8.3 Convolutional Codes . . . 244
8.3.1 Code Characterization: Trellis Diagrams . . . 244
8.3.2 Maximum Likelihood Decoding . . . 246
8.3.3 The Viterbi Algorithm . . . 249
8.3.4 Distance Properties . . . 250
8.3.5 State Diagrams and Transfer Functions . . . 251
8.3.6 Error Probability for Convolutional Codes . . . 253
8.3.7 Convolutional Coding and Interleaving for Fading Channels . . . 255
8.4 Concatenated Codes . . . 256
8.5 Turbo Codes . . . 257
8.6 Low Density Parity Check Codes . . . 259
8.7 Coded Modulation . . . 260
8.7.1 Coded Modulation for AWGN Channels . . . 260
8.7.2 Coded Modulation with Interleaving for Fading Channels . . . 264
8.8 Unequal Error Protection Codes . . . 264
8.9 Joint Source and Channel Coding . . . 266

9 Adaptive Modulation **277**
9.1 Introduction . . . 277
9.2 System Model . . . 278
9.3 Variable-Rate Variable-Power MQAM . . . 280
9.4 Constellation Restriction . . . 283
9.4.1 Optimal Adaptation . . . 283
9.4.2 Suboptimal Policies . . . 286
9.5 Simulation Results . . . 287
9.6 Channel Estimation Error and Delay . . . 289
9.7 Coding Issues and Capacity Revisited . . . 291

10 Multiple Antenna Systems **297**
10.1 Multiple Input Multiple Output (MIMO) Systems . . . 297
10.1.1 The Narrowband Multiple Antenna System Model . . . 297
10.1.2 Transmit Precoding and Receiver Shaping . . . 298
10.1.3 Parallel Decomposition of the MIMO Channel . . . 299
10.1.4 MIMO Channel Capacity . . . 300
10.1.5 Beamforming . . . 300
10.2 Space-time codes . . . 302
10.3 Smart Antennas . . . 302

11 Equalization **309**
11.1 Equalizer Types . . . 310
11.2 Folded Spectrum and ISI-Free Transmission . . . 311
11.3 Linear Equalizers . . . 313
11.3.1 Zero Forcing (ZF) Equalizers . . . 314
11.3.2 Minimum Mean Square Error (MMSE) Equalizer . . . 315
11.4 Maximum Likelihood Sequence Estimation . . . 317
11.5 Decision-Feedback Equalization . . . 318
11.6 Equalizer Training and Tracking . . . 319

12 Multicarrier Modulation **325**
12.1 Orthogonal Frequency Division Multiplexing (OFDM) . . . 326
12.2 Discrete Implementation of OFDM (Discrete Multitone) . . . 329
12.3 Fading across Subcarriers . . . 330
12.3.1 Frequency Equalization . . . 330
12.3.2 Precoding . . . 330
12.3.3 Adaptive Loading . . . 331
12.3.4 Coding across Subchannels . . . 332

13 Spread Spectrum and RAKE Receivers **337**
13.1 Spread Spectrum Modulation . . . 337
13.2 Pseudorandom (PN) Sequences (Spreading Codes) . . . 338
13.3 Direct Sequence Spread Spectrum . . . 340
13.4 RAKE receivers . . . 343
13.5 Spread Spectrum Multiple Access . . . 344
13.5.1 Spreading Codes for Multiple Access . . . 344
13.5.2 Broadcast Channels . . . 345
13.5.3 Multiple Access Channels . . . 348
13.5.4 Multiuser Detection . . . 351
13.6 Frequency-Hopping . . . 351

14 Multiuser Systems **355**
14.1 Multiuser Channels: Broadcast and Multiple Access . . . 355
14.2 Multiple Access . . . 356
14.2.1 Frequency Division . . . 356
14.2.2 Time-Division . . . 357
14.2.3 Code-Division . . . 357
14.2.4 Standards Debate . . . 358
14.3 Broadcast Channel Capacity Region . . . 358
14.3.1 The AWGN Broadcast Channel Model . . . 359
14.3.2 Capacity Region in AWGN under TD, FD, and CD . . . 359
14.3.3 Fading Broadcast Channel Capacity . . . 362
14.4 Multiple Access Channel Capacity Region . . . 367
14.4.1 The AWGN Multiple Access Channel . . . 367
14.4.2 Fading Multiaccess Channels . . . 368
14.5 Random Access . . . 369
14.6 Scheduling . . . 371

14.7 Power Control 372

15 Cellular Systems and Infrastructure-Based Wireless Networks 377
15.1 Cellular System Design 378
15.2 Frequency Reuse in Cellular Systems 378
15.2.1 Frequency Reuse in Code-Division Systems 378
15.2.2 Frequency Reuse in Time and Frequency Division Systems 379
15.3 Dynamic Resource Allocation in Cellular Systems 379
15.4 Area Spectral Efficiency 381
15.5 Interference Model 382
15.5.1 Reuse Distance, Multicell Capacity, and Area Efficiency 382
15.5.2 Efficiency Calculations 383
15.6 Power Control Impact on Interference 387
15.7 Interference Mitigation 389

16 Ad-Hoc Wireless Networks 393
16.0.1 Applications 396
16.0.2 Cross Layer Design 401
16.1 Link Design Issues 404
16.1.1 Fundamental Capacity Limits 404
16.1.2 Coding 405
16.1.3 Multiple Antennas 405
16.1.4 Power control 406
16.1.5 Adaptive Resource Allocation 406
16.2 Medium Access Control Design Issues 407
16.3 Network Design Issues 408
16.3.1 Neighbor Discovery and Network Connectivity 408
16.4 Routing 409
16.4.1 Scalability and Distributed Protocols 410
16.4.2 Network Capacity 411
16.5 Application Design Issues 411
16.5.1 Adaptive QoS 411
16.5.2 Application Adaptation and Cross Layer Design Revisited 412

Chapter 1

Overview of Wireless Communications

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

1.1 History of Wireless Communications

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

wireless networking.

Early radio systems transmitted analog signals. Today most radio systems transmit digital signals composed of binary bits, where the bits are obtained directly from a data signal or by digitizing an analog voice or music signal. A digital radio can transmit a continuous bit stream or it can group the bits into packets. The latter type of radio is called a *packet radio* and is characterized by bursty transmissions: the radio is idle except when it transmits a packet. The first network based on packet radio, ALOHANET, was developed at the University of Hawaii in 1971. This network enabled computer sites at seven campuses spread out over four islands to communicate with a central computer on Oahu via radio transmission. The network architecture used a star topology with the central computer at its hub. Any two computers could establish a bi-directional communications link between them by going through the central hub. ALOHANET incorporated the first set of protocols for channel access and routing in packet radio systems, and many of the underlying principles in these protocols are still in use today. The U.S. military was extremely interested in the combination of packet data and broadcast radio inherent to ALOHANET. Throughout the 70's and early 80's the Defense Advanced Research Projects Agency (DARPA) invested significant resources to develop networks using packet radios for tactical communications in the battlefield. The nodes in these ad hoc wireless networks had the ability to self-configure (or reconfigure) into a network without the aid of any established infrastructure. DARPA's investment in ad hoc networks peaked in the mid 1980's, but the resulting networks fell far short of expectations in terms of speed and performance. DARPA has continued work on ad hoc wireless network research for military use, but many technical challenges in terms of performance and robustness remain. Packet radio networks have also found commercial application in supporting wide-area wireless data services. These services, first introduced in the early 1990's, enable wireless data access (including email, file transfer, and web browsing) at fairly low speeds, on the order of 20 Kbps. The market for these wide-area wireless data services is relatively flat, due mainly to their low data rates, high cost, and lack of "killer applications". Next-generation cellular services are slated to provide wireless data in addition to voice, which will provide stiff competition to these data-only services.

The introduction of wired Ethernet technology in the 1970's steered many commercial companies away from radio-based networking. Ethernet's 10 Mbps data rate far exceeded anything available using radio, and companies did not mind running cables within and between their facilities to take advantage of these high rates. In 1985 the Federal Communications Commission (FCC) enabled the commercial development of wireless LANs by authorizing the public use of the Industrial, Scientific, and Medical (ISM) frequency bands for wireless LAN products. The ISM band was very attractive to wireless LAN vendors since they did not need to obtain an FCC license to operate in this band. However, the wireless LAN systems could not interfere with the primary ISM band users, which forced them to use a low power profile and an inefficient signaling scheme. Moreover, the interference from primary users within this frequency band was quite high. As a result these initial LAN systems had very poor performance in terms of data rates and coverage. This poor performance, coupled with concerns about security, lack of standardization, and high cost (the first network adaptors listed for $1,400 as compared to a few hundred dollars for a wired Ethernet card) resulted in weak sales for these initial LAN systems. Few of these systems were actually used for data networking: they were relegated to low-tech applications like inventory control. The current generation of wireless LANS, based on the IEEE 802.11b and 802.11a standards, have better performance, although the data rates are still relatively low (effective data rates on the order of 2 Mbps for 802.11b and around 10 Mbps for 802.11a) and the coverage area is still small (100-500 feet). Wired Ethernets today offer data rates of 100 Mbps, and the performance gap between wired and wireless LANs is likely to increase over time without additional spectrum allocation. Despite the big data rate differences, wireless LANs are becoming the prefered Internet access method in many

homes, offices, and campus environments due to their convenience and freedom from wires. However, most wireless LANs support applications that are not bandwidth-intensive (email, file transfer, web browsing) and typically have only one user at a time accessing the system. The challenge for widespread wireless LAN acceptance and use will be for the wireless technology to support many users simultaneously, especially if bandwidth-intensive applications become more prevalent.

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

A solution to this capacity problem emerged during the 50's and 60's when researchers at AT&T Bell Laboratories developed the cellular concept [1]. Cellular systems exploit the fact that the power of a transmitted signal falls off with distance. Thus, the same frequency channel can be allocated to users at spatially-separate locations with minimal interference between the users. Using this premise, a cellular system divides a geographical area into adjacent, non-overlapping, "cells". Different channel sets are assigned to each cell, and cells that are assigned the same channel set are spaced far enough apart so that interference between the mobiles in these cells is small. Each cell has a centralized transmitter and receiver (called a base station) that communicates with the mobile units in that cell, both for control purposes and as a call relay. All base stations have high-bandwidth connections to a mobile telephone switching office (MTSO), which is itself connected to the public-switched telephone network (PSTN). The handoff of mobile units crossing cell boundaries is typically handled by the MTSO, although in current systems some of this functionality is handled by the base stations and/or mobile units.

The original cellular system design was finalized in the late 60's. However, due to regulatory delays from the FCC, the system was not deployed until the early 80's, by which time much of the original technology was out-of-date. The explosive growth of the cellular industry took most everyone by surprise, especially the original inventors at AT&T, since AT&T basically abandoned the cellular business by the early 80's to focus on fiber optic networks. The first analog cellular system deployed in Chicago in 1983 was already saturated by 1984, at which point the FCC increased the cellular spectral allocation from 40 MHz to 50 MHz. As more and more cities became saturated with demand, the development of digital cellular technology for increased capacity and better performance became essential.

The second generation of cellular systems are digital. In addition to voice communication, these systems provide email, voice mail, and paging services. Unfortunately, the great market potential for cellular phones led to a proliferation of digital cellular standards. Today there are three different digital cellular phone standards in the U.S. alone, and other standards in Europe and Japan, none of which are compatible. The fact that different cities have different incompatible standards makes roaming throughout the U.S. using one digital cellular phone impossible. Most cellular phones today are dual-mode: they incorporate one of the digital standards along with the old analog standard, since only the analog standard provides universal coverage throughout the U.S. More details on today's digital cellular systems will be given in Section 15.

Radio paging systems are another example of an extremely successful wireless data network, with 50 million subscribers in the U.S. alone. However, their popularity is starting to wane with the widespread penetration and competitive cost of cellular telephone systems. Paging systems allow coverage over very wide areas by simultaneously broadcasting the pager message at high power from multiple base stations or

satellites. These systems have been around for many years. Early radio paging systems were analog 1 bit messages signaling a user that someone was trying to reach him or her. These systems required callback over the regular telephone system to obtain the phone number of the paging party. Recent advances now allow a short digital message, including a phone number and brief text, to be sent to the pagee as well. In paging systems most of the complexity is built into the transmitters, so that pager receivers are small, lightweight, and have a long battery life. The network protocols are also very simple since broadcasting a message over all base stations requires no routing or handoff. The spectral inefficiency of these simultaneous broadcasts is compensated by limiting each message to be very short. Paging systems continue to evolve to expand their capabilities beyond very low-rate one-way communication. Current systems are attempting to implement "answer-back" capability, i.e. two-way communication. This requires a major change in the pager design, since it must now transmit signals in addition to receiving them, and the transmission distances can be quite large. Recently many of the major paging companies have teamed up with the palmtop computer makers to incorporate paging functions into these devices [2]. This development indicates that short messaging without additional functionality is no longer competitive given other wireless communication options.

Commercial satellite communication systems are now emerging as another major component of the wireless communications infrastructure. Satellite systems can provide broadcast services over very wide areas, and are also necessary to fill the coverage gap between high-density user locations. Satellite mobile communication systems follow the same basic principle as cellular systems, except that the cell base stations are now satellites orbiting the earth. Satellite systems are typically characterized by the height of the satellite orbit, low-earth orbit (LEOs at roughly 2000 Km. altitude), medium-earth orbit (MEOs at roughly 9000 Km. altitude), or geosynchronous orbit (GEOs at roughly 40,000 Km. altitude). The geosynchronous orbits are seen as stationary from the earth, whereas the satellites with other orbits have their coverage area change over time. The 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 delay-constrained applications like voice. However, satellites at these orbits tend to have larger coverage areas, so fewer satellites (and dollars) are necessary to provide wide-area or global coverage.

The concept of using geosynchronous satellites for communications was first suggested by the science fiction writer Arthur C. Clarke in 1945. However, the first deployed satellites, the Soviet Union's Sputnik in 1957 and the Nasa/Bell Laboratories' Echo-1 in 1960, were not geosynchronous due to the difficulty of lifting a satellite into such a high orbit. The first GEO satellite was launched by Hughes and Nasa in 1963 and from then until recently GEOs dominated both commercial and government satellite systems. The trend in current satellite systems is to use lower orbits so that lightweight handheld devices can communicate with the satellite [3]. Inmarsat is the most well-known GEO satellite system today, but most new systems use LEO orbits. These LEOs provide global coverage but the link rates remain low due to power and bandwidth constraints. These systems allow calls any time and anywhere using a single communications device. The services provided by satellite systems include voice, paging, and messaging services, all at fairly low data rates [3, 4]. The LEO satellite systems that have been deployed are not experiencing the growth they had anticipated, and one of the first systems (Iridium) was forced into bankruptcy and went out of business.

A natural area for satellite systems is broadcast entertainment. Direct broadcast satellites operate in the 12 GHz frequency band. These systems offer hundreds of TV channels and are major competitors to cable. Satellite-delivered digital radio is an emerging application in the 2.3 GHz frequency band. These systems offer digital audio broadcasts nationwide at near-CD quality. Digital audio broadcasting is also quite popular in Europe.

1.2 Wireless Vision

The vision of wireless communications supporting information exchange between people or devices is the communications frontier of the next century. This vision will allow people to operate a virtual office anywhere in the world using a small handheld device - with seamless telephone, modem, fax, and computer communications. Wireless networks will also be used to connect together palmtop, laptop, and desktop computers anywhere within an office building or campus, as well as from the corner cafe. In the home these networks will enable a new class of intelligent home electronics that can interact with each other and with the Internet in addition to providing connectivity between computers, phones, and security/monitoring systems. Such smart homes can also help the elderly and disabled with assisted living, patient monitoring, and emergency response. Video teleconferencing will take place between buildings that are blocks or continents apart, and these conferences can include travelers as well, from the salesperson who missed his plane connection to the CEO off sailing in the Caribbean. Wireless video will be used to create remote classrooms, remote training facilities, and remote hospitals anywhere in the world. Wireless sensors have an enormous range of both commercial and military applications. Commercial applications include monitoring of fire hazards, hazardous waste sites, stress and strain in buildings and bridges, or carbon dioxide movement and the spread of chemicals and gasses at a disaster site. These wireless sensors will self-configure into a network to process and interpret sensor measurements and then convey this information to a centralized control location. Military applications include identification and tracking of enemy targets, detection of chemical and biological attacks, and the support of unmanned robotic vehicles. Finally, wireless networks enable distributed control systems, with remote devices, sensors, and actuators linked together via wireless communication channels. Such networks are imperative for coordinating unmanned mobile units and greatly reduce maintenance and reconfiguration costs over distributed control systems with wired communication links, for example in factory automation.

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

wireless networks, which have much lower data rates and higher BERs. Therefore, at least in the near future, wireless systems will continue to be fragmented, with different protocols tailored to support the requirements of different applications.

Will there be a large demand for all wireless applications, or will some flourish while others vanish? Companies are investing large sums of money to build multimedia wireless systems, yet the only highly profitable wireless application so far is voice. Experts have been predicting a huge market for wireless data services and products for the last 10 years, but the market for these products remains relatively small with sluggish growth. To examine the future of wireless data, it is useful to see the growth of various communication services over the last five years, as shown in Figure 1.1. In this figure we see that cellular and paging subscribers are growing exponentially. This growth is exceeded only by the growing demand for Internet access, driven by web browsing and email exchange. The number of laptop and palmtop computers is also growing steadily. These trends indicate that people want to communicate while on the move. They also want to take their computers wherever they go. It is therefore reasonable to assume that people want the same data communications capabilities on the move as they enjoy in their home or office. Yet the demand for high-speed wireless data has not materialized, except for relatively stationary users accessing the network via a wireless LAN. Why the discrepancy? Perhaps the main reason for the lack of enthusiasm in wireless data for highly mobile users is the high cost and poor performance of today's systems, along with a lack of "killer applications" for mobile users beyond voice and low-rate data.

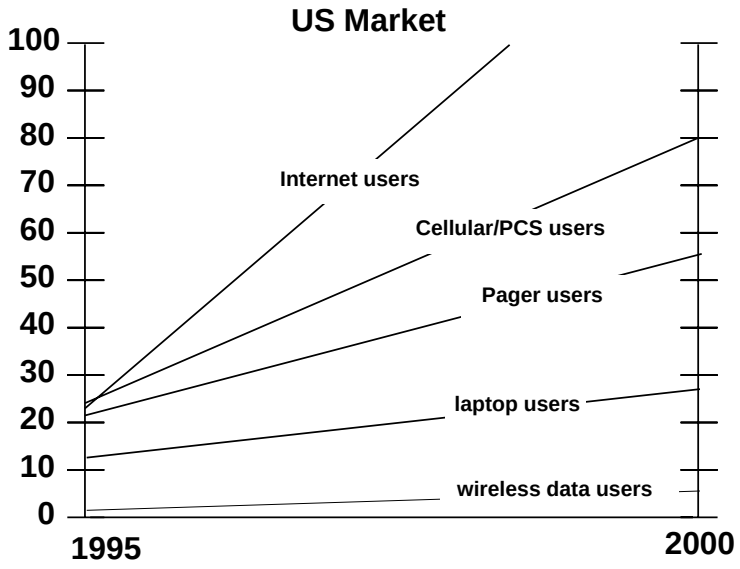


Figure 1.1: Growth of Wireless Communications Markets

Consider the performance gap between wired and wireless networks for both local and wide-area networks, as shown in Figures 1.2 and 1.3. Wired local-area networks have data rates that are two orders of magnitude higher than their wireless counterparts. ATM is promising 100,000 Kbps for wired wide-area networks, while today's wide-area wireless data services provide only tens of Kbps. Moreover, the performance gap between wired and wireless networks appears to be growing. Thus, the most formidable obstacle to the growth of wireless data systems is their performance. Many technical challenges must be overcome to improve wireless network performance such that users will accept this performance in exchange for mobility.

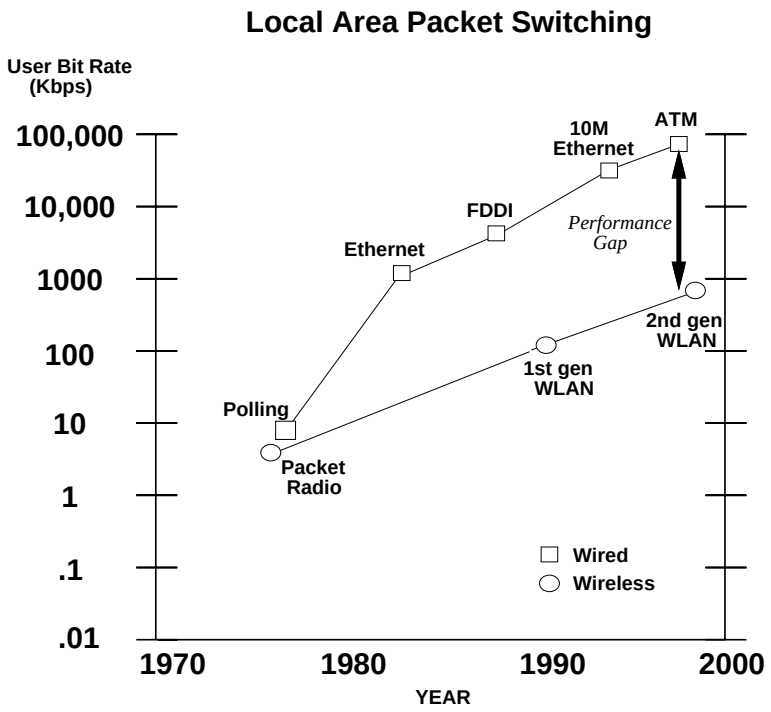


Figure 1.2: Performance Gap for Local Area Networks

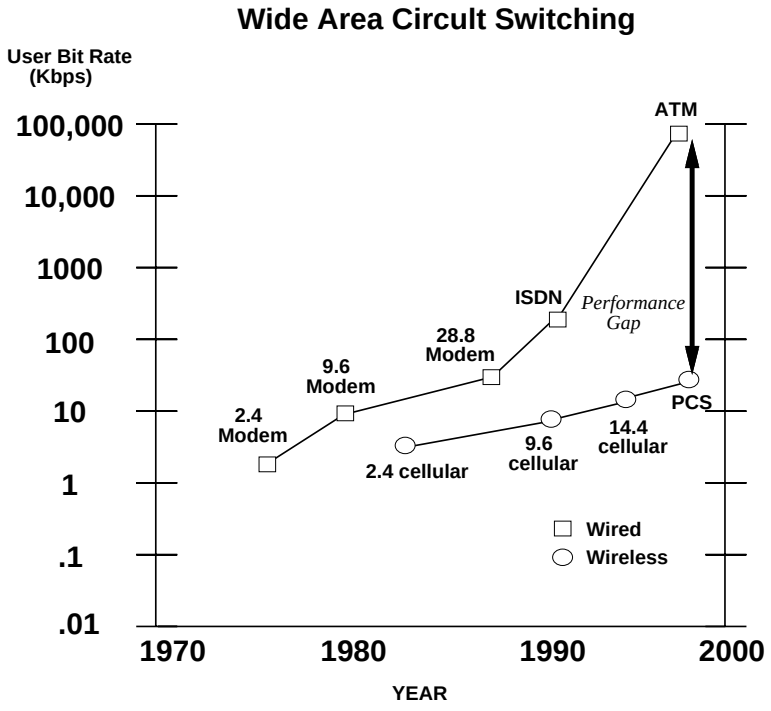


Figure 1.3: Performance Gap for Wide Area Networks

1.3 Technical Issues

The technical problems that must be solved to make the wireless vision a reality extend across all levels of the system design. At the hardware level the terminal must have multiple modes of operation to support the different applications and media. Desktop computers currently have the capability to process voice, image, text, and video data, but breakthroughs in circuit design are required to implement multimode operation in a small, lightweight, handheld device. Since most people don't want to carry around a twenty pound battery, the signal processing and communications hardware of the portable terminal must consume very little power, which will impact higher levels of the system design. Many of the signal processing techniques required for efficient spectral utilization and networking demand much processing power, precluding the use of low power devices. Hardware advances for low power circuits with high processing ability will relieve some of these limitations. However, placing the processing burden on fixed sites with large power resources has and will continue to dominate wireless system designs. The associated bottlenecks and single points-of-failure are clearly undesirable for the overall system. Moreover, in some applications (e.g. sensors) network nodes will not be able to recharge their batteries. In this case the finite battery energy must be allocated efficiently across all layers of the network protocol stack [5]. The finite bandwidth and random variations of the communication channel will also require robust compression schemes which degrade gracefully as the channel degrades.

The wireless communication channel is an unpredictable and difficult communications medium. First of all, the radio spectrum is a scarce resource that must be allocated to many different applications and systems. For this reason spectrum is controlled by regulatory bodies both regionally and globally. In the U.S. spectrum is allocated by the FCC, in Europe the equivalent body is the European Telecommunications Standards Institute (ETSI), and globally spectrum is controlled by the International Telecommunications Union (ITU). A regional or global system operating in a given frequency band must obey the restrictions for that band set forth by the corresponding regulatory body as well as any standards adopted for that spectrum. Spectrum can also be very expensive since in most countries, including the U.S., spectral licenses are now auctioned to the highest bidder. In the 2 GHz spectral auctions of the early 90s, companies spent over nine billion dollars for licenses, and the recent auctions in Europe for 3G spectrum garnered over 100 billion dollars. The spectrum obtained through these auctions must be used extremely efficiently to get a reasonable return on its investment, and it must also be reused over and over in the same geographical area, thus requiring cellular system designs with high capacity and good performance. At frequencies around several Gigahertz wireless radio components with reasonable size, power consumption, and cost are available. However, the spectrum in this frequency range is extremely crowded. Thus, technological breakthroughs to enable higher frequency systems with the same cost and performance would greatly reduce the spectrum shortage, although path loss at these higher frequencies increases, thereby limiting range.

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

Wireless networking is also a significant challenge [23, 24, 25, 26]. The network must be able to

locate a given user wherever it is amongst millions of globally-distributed mobile terminals. It must then route a call to that user as it moves at speeds of up to 100 mph. The finite resources of the network must be allocated in a fair and efficient manner relative to changing user demands and locations. Moreover, there currently exists a tremendous infrastructure of wired networks: the telephone system, the Internet, and fiber optic cable, which should be used to connect wireless systems together into a global network. However, wireless systems with mobile users will never be able to compete with wired systems in terms of data rate and reliability. The design of protocols to interface between wireless and wired networks with vastly different performance capabilities remains a challenging topic of research.

Perhaps the most significant technical challenge in wireless network design is an overhaul of the design process itself. Wired networks are mostly designed according to the layers of the OSI model: each layer is designed independent from the other layers with baseline mechanisms to interface between layers. This methodology greatly simplifies network design, although it leads to some inefficiency and performance loss due to the lack of a global design optimization. However, the large capacity and good reliability of wired network links make it easier to buffer high-level network protocols from the lower level protocols for link transmission and access, and the performance loss resulting from this isolated protocol design is fairly low. However, the situation is very different in a wireless network. Wireless links can exhibit very poor performance, and this performance along with user connectivity and network topology changes over time. In fact, the very notion of a wireless link is somewhat fuzzy due to the nature of radio propagation. The dynamic nature and poor performance of the underlying wireless communication channel indicates that high-performance wireless networks must be optimized for this channel and must adapt to its variations as well as to user mobility. Thus, these networks will require an integrated and adaptive protocol stack across all layers of the OSI model, from the link layer to the application layer.

In summary, technological advances in the following areas are needed to implement the wireless vision outlined above:

- Measurements and models for wireless indoor and outdoor channels.
- Hardware for low-power handheld computer and communication terminals.
- Techniques to mitigate wireless channel impairments and to improve the quality and spectral efficiency of communication over wireless channels.
- Better means of sharing the limited spectrum to accommodate the different wireless applications.
- Protocols for routing and mobility management which support users on the move.
- An architecture to connect the various wireless subnetworks together and to the backbone wireline network.
- An integrated and adaptive protocol stack for wireless networks that extends across all layers of the OSI model.

Given these requirements, the field of wireless communications draws from many areas of expertise, including physics, communications, signal processing, network theory and design, software design, and hardware design. Moreover, given the fundamental limitations of the wireless channels and the explosive demand for its utilization, communication between these interdisciplinary groups is necessary to implement the most rudimentary shell of the wireless vision depicted above.

We now give an overview of the wireless systems in operation today. It will be clear from this overview that the wireless vision remains a distant goal, with many challenges remaining before it will be realized. Many of these challenges will be examined in detail in later chapters.

1.4 Current Wireless Systems

1.4.1 Cellular Telephone Systems

Cellular telephone systems, also referred to as Personal Communication Systems (PCS), are extremely popular and lucrative worldwide: these systems have sparked much of the optimism about the future of wireless networks. Cellular telephone systems are designed to provide two-way voice communication at vehicle speeds with regional or national coverage. Cellular systems were initially designed for mobile terminals inside vehicles with antennas mounted on the vehicle roof. Today these systems have evolved to support lightweight handheld mobile terminals operating inside and outside buildings at both pedestrian and vehicle speeds.

The basic premise behind cellular system design is frequency reuse, which exploits path loss to reuse the same frequency spectrum at spatially-separated locations. Specifically, the coverage area of a cellular system is divided into nonoverlapping *cells* where some set of channels is assigned to each cell. This same channel set is used in another cell some distance away, as shown in Figure 1.4, where f_i denotes the channel set used in a particular cell. Operation within a cell is controlled by a centralized base station, as described in more detail below. The interference caused by users in different cells operating on the same channel set is called intercell interference. The spatial separation of cells that reuse the same channel set, the *reuse distance*, should be as small as possible to maximize the spectral efficiency obtained by frequency reuse. However, as the reuse distance decreases, intercell interference increases, due to the smaller propagation distance between interfering cells. Since intercell interference must remain below a given threshold for acceptable system performance, reuse distance cannot be reduced below some minimum value. In practice it is quite difficult to determine this minimum value since both the transmitting and interfering signals experience random power variations due to path loss, shadowing, and multipath. In order to determine the best reuse distance and base station placement, an accurate characterization of signal propagation within the cells is needed. This characterization is usually obtained using detailed analytical models, sophisticated computer-aided modeling, or empirical measurements.

Initial cellular system designs were mainly driven by the high cost of base stations, approximately one million dollars apiece. For this reason early cellular systems used a relatively small number of cells to cover an entire city or region. The cell base stations were placed on tall buildings or mountains and transmitted at very high power with cell coverage areas of several square miles. These large cells are called macrocells. Signals propagated out from base stations uniformly in all directions, so a mobile moving in a circle around the base station would have approximately constant received power. This circular contour of constant power yields a hexagonal cell shape for the system, since a hexagon is the closest shape to a circle that can cover a given area with multiple nonoverlapping cells.

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

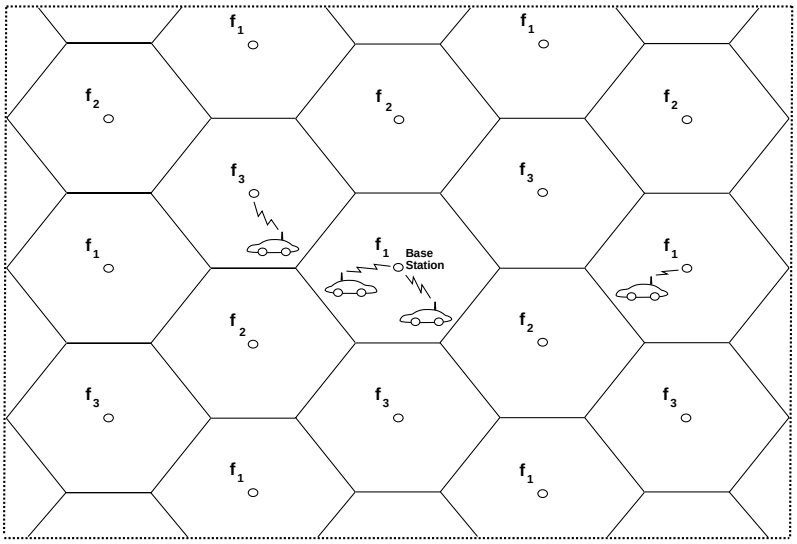


Figure 1.4: Cellular Systems.

general propagation models for small cells, since signal propagation in these cells is highly dependent on base station placement and the geometry of the surrounding reflectors. In particular, a hexagonal cell shape is not a good approximation to signal propagation in microcells. Microcellular systems are often designed using square or triangular cell shapes, but these shapes have a large margin of error in their approximation to microcell signal propagation [7].

All base stations in a city 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 network, allocating channels within each cell, coordinating handoffs between cells when a mobile traverses a cell boundary, and routing calls to and from mobile users in conjunction with the public switched telephone network (PSTN). A new user located in a given cell requests a channel by sending a call request to the cell's base station over a separate control channel. The request is relayed to the MTSO, which accepts the call request if a channel is available in that cell. If no channels are available then the call request is rejected. A call handoff is initiated when the base station or the mobile in a given cell detects that the received signal power for that call is approaching a given minimum threshold. In this case the base station informs the MTSO that the mobile requires a handoff, and the MTSO then queries surrounding base stations to determine if one of these stations can detect that mobile's signal. If so then the MTSO coordinates a handoff between the original base station and the new base station. If no channels are available in the cell with the new base station then the handoff fails and the call is terminated. False handoffs may also be initiated if a mobile is in a deep fade, causing its received signal power to drop below the minimum threshold even though it may be nowhere near a cell boundary.

Cellular telephone systems have recently moved from analog to digital technology. Digital technology has many advantages over analog. The components are cheaper, faster, smaller, and require less power. Voice quality is improved due to error correction coding. Digital systems also have higher capacity than analog systems since they are not limited to frequency division for multiple access, and they can take advantage of advanced compression techniques and voice activity factors. In addition, encryption

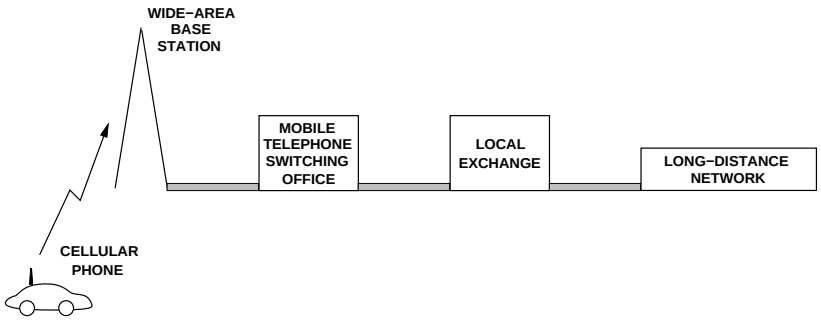


Figure 1.5: Current Cellular Network Architecture

techniques can be used to secure digital signals against eavesdropping. All cellular systems being deployed today are digital, and these systems provide voice mail, paging, and email services in addition to voice. Due to their lower cost and higher efficiency, service providers have used aggressive pricing tactics to encourage user migration from analog to digital systems. Since they are relatively new, digital systems do not always work as well as the old analog ones. Users experience poor voice quality, frequent call dropping, short battery life, and spotty coverage in certain areas. System performance will certainly improve as the technology and networks mature. However, it is unlikely that cellular phones will provide the same quality as wireline service any time soon. The great popularity of cellular systems indicates that users are willing to tolerate inferior voice communications in exchange for mobility.

Spectral sharing in digital cellular can be done using frequency-division, time-division, code-division (spread spectrum), or hybrid combinations of these techniques (see Chapter 14). In time-division the signal occupies the entire frequency band, and is divided into time slots t_i which are reused in distant cells [8]. Time division is depicted by Figure 1.4 if the f_is are replaced by t_is. Time-division is more difficult to implement than frequency-division since the users must be time-synchronized. However, it is easier to accommodate multiple data rates with time-division since multiple timeslots can be assigned to a given user. Spectral sharing can also be done using code division, which is commonly implemented using either direct-sequence or frequency-hopping spread spectrum [9]. In direct-sequence each user modulates its data sequence by a different pseudorandom chip sequence which is much faster than the data sequence. In the frequency domain, the narrowband data signal is convolved with the wideband chip signal, resulting in a signal with a much wider bandwidth than the original data signal - hence the name spread spectrum. In frequency hopping the carrier frequency used to modulate the narrowband data signal is varied by a pseudorandom chip sequence which may be faster or slower than the data sequence. Since the carrier frequency is hopped over a large signal bandwidth, frequency-hopping also spreads the data signal to a much wider bandwidth. Typically spread spectrum signals are superimposed onto each other within the same signal bandwidth. A spread spectrum receiver can separate each of the distinct signals by separately decoding each spreading sequence. However, since the codes are semi-orthogonal, the users within a cell interfere with each other (intracell interference), and codes that are reused in other cells also cause interference (intercell interference). Both the intracell and intercell interference power is reduced by the spreading gain of the code. Moreover, interference in spread spectrum systems can be further reduced through multiuser detection and interference cancellation.

In the U.S. the standards activities surrounding the second generation of digital cellular systems provoked a raging debate on multiple access for these systems, resulting in several incompatible standards [10, 11, 12]. In particular, there are two standards in the 900 MHz (cellular) frequency band: IS-54, which uses a combination of TDMA and FDMA, and IS-95, which uses semi-orthogonal CDMA [13, 14]. The

spectrum for digital cellular in the 2 GHz (PCS) frequency band was auctioned off, so service providers could use an existing standard or develop proprietary systems for their purchased spectrum. The end result has been three different digital cellular standards for this frequency band: IS-136 (which is basically the same as IS-54 at a higher frequency), IS-95, and the European digital cellular standard GSM, which uses a combination of TDMA and slow frequency-hopping. The digital cellular standard in Japan is similar to IS-54 and IS-136 but in a different frequency band, and the GSM system in Europe is at a different frequency than the GSM systems in the U.S. This proliferation of incompatible standards in the U.S. and abroad makes it impossible to roam between systems nationwide or globally without using multiple phones (and phone numbers).

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

The third generation of cellular phones is based on a wideband CDMA standard developed within the auspices of the International Telecommunications Union (ITU) [15]. The standard, initially called International Mobile Telecommunications 2000 (IMT-2000), provides different data rates depending on mobility and location, from 384 Kbps for pedestrian use to 144 Kbps for vehicular use to 2 Mbps for indoor office use. The 3G standard is incompatible with 2G systems, so service providers must invest in a new infrastructure before they can provide 3G service. The first 3G systems were deployed in Japan, where they have experienced limited success with a somewhat slower growth than expected. One reason that 3G services came out first in Japan is the process of 3G spectrum allocation, which in Japan was awarded without much up-front cost. The 3G spectrum in both Europe and the U.S. is allocated based on auctioning, thereby requiring a huge initial investment for any company wishing to provide 3G service. European companies collectively paid over 100 billion dollars in their 3G spectrum auctions. There has been much controversy over the 3G auction process in Europe, with companies charging that the nature of the auctions caused enormous overbidding and that it will be very difficult if not impossible to reap a profit on this spectrum. A few of the companies have already decided to write off their investment in 3G spectrum and not pursue system buildout. In fact 3G systems have not yet come online in Europe, and it appears that data enhancements to 2G systems may suffice to satisfy user demands. However, the 2G spectrum in Europe is severely overcrowded, so users will either eventually migrate to 3G or regulations will change so that 3G bandwidth can be used for 2G services (which is not currently allowed in Europe). 3G development in the U.S. has lagged far behind that of Europe. The available 3G spectrum in the U.S. in only about half that available in Europe. Due to wrangling about which parts of the spectrum will be used, the spectral auctions have been delayed. However, the U.S. does allow the 1G and 2G spectrum to be used for 3G, and this flexibility may allow a more gradual rollout and investment than the more restrictive 3G requirements in Europe. It appears that delaying 3G in the U.S. will allow U.S. service providers to learn from the mistakes and successes in Europe and Japan.

Efficient cellular system designs are *interference-limited*, i.e. the interference dominates the noise floor since otherwise more users could be added to the system. As a result, any technique to reduce interference in cellular systems leads directly to an increase in system capacity and performance. Some methods for interference reduction in use today or proposed for future systems include cell sectorization [6], directional and smart antennas [19], multiuser detection [20], and dynamic channel and resource allocation [21, 22].

1.4.2 Cordless Phones

Cordless telephones first appeared in the late 1970's and have experienced spectacular growth ever since. Roughly half of the phones in U.S. homes today are cordless. Cordless phones were originally designed to provide a low-cost low-mobility wireless connection to the PSTN, i.e. a short wireless link to replace the cord connecting a telephone base unit and its handset. Since cordless phones compete with wired handsets, their voice quality must be similar: initial cordless phones had poor voice quality and were quickly discarded by users. The first cordless systems allowed only one phone handset to connect to each base unit, and coverage was limited to a few rooms of a house or office. This is still the main premise behind cordless telephones in the U.S. today, although these phones now use digital technology instead of analog. In Europe and the Far East digital cordless phone systems have evolved to provide coverage over much wider areas, both in and away from home, and are similar in many ways to today's cellular telephone systems.

Digital cordless phone systems in the U.S. today consist of a wireless handset connected to a single base unit which in turn is connected to the PSTN. These cordless phones impose no added complexity on the telephone network, since the cordless base unit acts just like a wireline telephone for networking purposes. The movement of these cordless handsets is extremely limited: a handset must remain within range of its base unit. There is no coordination with other cordless phone systems, so a high density of these systems in a small area, e.g. an apartment building, can result in significant interference between systems. For this reason cordless phones today have multiple voice channels and scan between these channels to find the one with minimal interference. Spread spectrum cordless phones have also been introduced to reduce interference from other systems and narrowband interference.

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

Another evolution of the cordless telephone designed primarily for office buildings is the European DECT system. The main function of DECT is to provide local mobility support for users in an in-building

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

The most recent advance in cordless telephone system design is the Personal Handyphone System (PHS) in Japan. The PHS system is quite similar to a cellular system, with widespread base station deployment supporting handoff and call routing between base stations. With these capabilities PHS does not suffer from the main limitations of the CT-2 system. Initially PHS systems enjoyed one of the fastest growth rates ever for a new technology. In 1997, two years after its introduction, PHS subscribers peaked at about 7 million users, and has declined slightly since then due mainly to sharp price cutting by cellular providers. The main difference between a PHS system and a cellular system is that PHS cannot support call handoff at vehicle speeds. This deficiency is mainly due to the dynamic channel allocation procedure used in PHS. Dynamic channel allocation greatly increases the number of handsets that can be serviced by a single base station, thereby lowering the system cost, but it also complicates the handoff procedure. It is too soon to tell if PHS systems will go the same route as CT-2. However, it is clear from the recent history of cordless phone systems that to extend the range of these systems beyond the home requires either the same functionality as cellular systems or a significantly reduced cost.

1.4.3 Wireless LANs

Wireless LANs provide high-speed data within a small region, e.g. a campus or small building, as users move from place to place. Wireless devices that access these LANs are typically stationary or moving at pedestrian speeds. Nearly all wireless LANs in the United States use one of the ISM frequency bands. The appeal of these frequency bands, located at 900 MHz, 2.4 GHz, and 5.8 GHz, is that an FCC license is not required to operate in these bands. However, this advantage is a double-edged sword, since many other systems operate in these bands for the same reason, causing a great deal of interference between systems. The FCC mitigates this interference problem by setting a limit on the power per unit bandwidth for ISM-band systems. Wireless LANs can have either a star architecture, with wireless access points or hubs placed throughout the coverage region, or a peer-to-peer architecture, where the wireless terminals self-configure into a network.

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

The second generation of wireless LANs in the United States operate in the 2.4 GHz ISM band. A wireless LAN standard for this frequency band, the IEEE 802.11b standard, was developed to avoid some of the problems with the proprietary first generation systems. The standard specifies frequency hopped spread spectrum with data rates of around 1.6 Mbps (raw data rates of 11 Mbps) and a range

of approximately 500 ft. The network architecture can be either star or peer-to-peer. Many companies have developed products based on the 802.11b standard, and these products are constantly evolving to provide higher data rates and better coverage at very low cost. The market for 802.11b wireless LANs is growing, and computer manufacturers have begun integrating the cards directly into their laptops. Many companies and universities have installed 802.11b base stations throughout their locations, and even local coffee houses are installing these base stations to offer wireless access to customers. Optimism is high that the wireless LAN market is poised to take off (although this prediction has been made every year since the inception of wireless LANs).

Perhaps the main impediment to the ultimate success of the 802.11b wireless LANs is the newest wireless LAN standard, IEEE 802.11a. This wireless LAN operates in the 5 GHz unlicensed band, which has much more spectrum and less interference than the 2.4 GHz band. The 802.11a standard is based on OFDM modulation and provides on the order of 50 Mbps data rates. There was some initial concern that 802.11a systems would be significantly more expensive than 802.11b systems, but in fact they are becoming quite competitive in price.

In Europe wireless LAN development revolves around the HIPERLAN (high performance radio LAN) standards. The first HIPERLAN standard, HIPERLAN Type 1, is similar to the IEEE 802.11a wireless LAN standard and promises data rates of 20 Mbps at a range of 50 meters (150 feet). This system operates in the 5 GHz band. Its network architecture is peer-to-peer, and the channel access mechanism uses a variation of ALOHA with prioritization based on the lifetime of packets. The next generation of HIPERLAN, HIPERLAN Type 2, is still under development, but the goal is to provide data rates on the order of 54 Mbps with a similar range, and also to support access to cellular, ATM, and IP networks. HIPERLAN Type 2 is also supposed to include support for Quality-of-Service (QoS), however it is not yet clear how and to what extent this will be done.

1.4.4 Wide Area Wireless Data Services

Wide area wireless data services provide low rate wireless data to high-mobility users over a very large coverage area. In these systems a given geographical region is serviced by base stations mounted on towers, rooftops, or mountains. The base stations can be connected to a backbone wired network or form a multihop ad hoc network. Initial data rates for these systems were below 10 Kbps but gradually increased to 20 Kbps. There are two main players in wide area wireless data services: Motient and Bell South Mobile Data (formerly RAM Mobile Data). Metricom provided a similar service with a network architecture consisting of a large network of small inexpensive base stations with small coverage areas. The increased efficiency of the small coverage areas allowed for higher data rates in Metricom, 76 Kbps, than in the other wide-area wireless data systems. However, the high infrastructure cost for Metricom eventually forced it into bankruptcy, and the system was shut down. Some of the infrastructure was bought and is operating in a few araas as Ricochet.

The cellular digital packet data (CDPD) system is a wide area wireless data service overlayed on the analog cellular telephone network. CDPD shares the FDMA voice channels of the analog systems, since many of these channels are idle due to the growth of digital cellular. The CDPD service provides packet data transmission at rates of 19.2 Kbps, and is available throughout the U.S. However, since newer generations of cellular systems also provide data services, CDPD will likely be replaced by these newer services.

All of these wireless data services have failed to grow as rapidly or to attract as many subscribers as initially predicted, especially in comparison with the rousing success of wireless voice systems. There is disagreement on why these systems have experienced such anemic growth. Data rates for these systems are clearly low, especially in comparison with their wireline counterparts. Pricing for these services also

remains high. There is a perceived lack of "killer applications" for wireless data: while voice communication on the move seems essential for a large part of the population, most people can wait until they have access to a phone line or wired network for data exchange. This may change with the proliferation of laptop and palmtop computers and the explosive demand for constant Internet access and email exchange. Optimists point to these factors as the drivers for wireless data but, as with wireless LANs, wide area wireless data services have been the pot of gold around the corner for many years yet have so far failed to deliver on these high expectations.

1.4.5 Fixed Wireless Access

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

1.4.6 Paging Systems

Paging systems provide very low rate one-way data services to highly mobile users over a very wide coverage area. Paging systems have experienced steady growth for many years and currently serve about 56 million customers in the United States. However, the popularity of paging systems is declining as cellular systems become cheaper and more ubiquitous. In order to remain competitive paging companies have slashed prices, and few of these companies are currently profitable. To reverse their declining fortunes, a consortium of paging service providers have recently teamed up with Microsoft and Compaq to incorporate paging functionality and Internet access into palmtop computers [2].

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

Some paging services today offer rudimentary (1 bit) answer-back capabilities from the handheld paging device. However, the requirement for two-way communication destroys the asymmetrical link

advantage so well exploited in paging system design. A paging handset with answer-back capability requires a modulator and transmitter with sufficient power to reach the distant base station. These requirements significantly increase the size and weight and reduce the usage time of the handheld pager. This is especially true for paging systems with satellite base stations, unless terrestrial relays are used.

1.4.7 Satellite Networks

Satellite systems provide voice, data, and broadcast services with widespread, often global, coverage to high-mobility users as well as to fixed sites. Satellite systems have the same basic architecture as cellular systems, except that the cell base-stations are satellites orbiting the earth. Satellites are characterized by their orbit distance from the earth. There are three main types of satellite orbits: low-earth orbit (LEOs) at 500-2000 Kms, medium-earth orbit (MEO) at 10,000 Kms, and geosynchronous orbit (GEO) at 35,800 Kms. A geosynchronous satellite has a large coverage area that is stationary over time, since the earth and satellite orbits are synchronous. Satellites with lower orbits have smaller coverage areas, and these coverage areas change over time so that satellite handoff is needed for stationary users or fixed point service.

Since geosynchronous satellites have such large coverage areas just a handful of satellites are needed for global coverage. However, geosynchronous systems have several disadvantages for two-way communication. It takes a great deal of power to reach these satellites, so handsets are typically large and bulky. In addition, there is a large round-trip propagation delay: this delay is quite noticeable in two-way voice communication. Recall also from Section 15 that high-capacity cellular systems require small cell sizes. Since geosynchronous satellites have very large cells, these systems have small capacity, high cost, and low data rates, less than 10 Kbps. The main geosynchronous systems in operation today are the global Inmarsat system, MSAT in North America, Mobilesat in Australia, and EMS and LLM in Europe.

The trend in current satellite systems is to use the lower LEO orbits so that lightweight handheld devices can communicate with the satellites and propagation delay does not degrade voice quality. The best known of these new LEO systems are Globalstar and Teledesic. Globalstar provides voice and data services to globally-roaming mobile users at data rates under 10 Kbps. The system requires roughly 50 satellites to maintain global coverage. Teledesic uses 288 satellites to provide global coverage to fixed-point users at data rates up to 2 Mbps. Teledesic is set to be deployed in 2005. The cell size for each satellite in a LEO system is much larger than terrestrial macrocells or microcells, with the corresponding decrease in capacity associated with large cells. Cost of these satellites, to build, to launch, and to maintain, is also much higher than that of terrestrial base stations, so these new LEO systems are unlikely to be cost-competitive with terrestrial cellular and wireless data services. Although these LEO systems can certainly complement these terrestrial systems in low-population areas, and are also appealing to travelers desiring just one handset and phone number for global roaming, it remains to be seen if there are enough such users willing to pay the high cost of satellite services to make these systems economically viable. In fact, Iridium, the largest and best-known of the LEO systems, was forced into bankruptcy and disbanded.

1.4.8 Bluetooth

Bluetooth is a cable-replacement RF technology for short range connections between wireless devices. The Bluetooth standard is based on a tiny microchip incorporating a radio transceiver that is built into digital devices. The transceiver takes the place of a connecting cable for devices such as cell phones, laptop and palmtop computers, portable printers and projectors, and network access points. Bluetooth is mainly for short range communications, e.g. from a laptop to a nearby printer or from a cell phone to a wireless headset. Its normal range of operation is 10 m (at 1 mW transmit power), and this range

can be increased to 100 m by increasing the transmit power to 100 mW. The system operates in the unregulated 2.4 GHz frequency band, hence it can be used worldwide without any licensing issues. The Bluetooth standard provides 1 data channel at 721 Kbps and up to three voice channels at 56 Kbps for an aggregate bit rate of 1 Mbps. Networking is done via a packet switching protocol based on frequency hopping at 1600 hops per second.

The Bluetooth standard was developed jointly by 3 Com, Ericsson, Intel, IBM, Lucent, Microsoft, Motorola, Nokia, and Toshiba. The standard has now been adopted by over 1300 manufacturers, and products compatible with Bluetooth are starting to appear on the market now. Specifically, the following products all use Bluetooth technology: a wireless headset for cell phones (Ericsson), a wireless USB or RS232 connector (RTX Telecom, Adayma), wireless PCMCIA cards (IBM), and wireless settop boxes (Eagle Wireless), to name just a few. More details on Bluetooth, including Bluetooth products currently available or under development, can be found at the website http://www.bluetooth.com.

1.4.9 HomeRF

HomeRF is a working group developing an open industry standard for wireless digital communication between PCs, PDAs, intelligent home appliances and consumer electronic devices anywhere in and around the home. The working group was initiated by Intel, HP, Microsoft, Compaq, and IBM. The main component of the HomeRF protocol is its Shared Wireless Access Protocol (SWAP), which operates in the unregulated 2.4 GHz frequency band (same band as Bluetooth).

The SWAP protocol is designed to carry both voice and data traffic and to interoperate with the PSTN and the Internet. The bandwidth sharing is enabled by frequency hopped spread spectrum at 50 hops/sec, however it also supports a TDMA service for delivery of interactive voice and other time-critical services, and a CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance) service for high speed packet data. The transmit power for HomeRF is specified at 100 mW which provides a data rate of 1-2 Mbps. However, in August 2000 the FCC okayed a five-fold increase in the HomeRF bandwidth, which will increase data rates to 10 Mbps. The range of HomeRF covers a typical home and backyard. Compaq and Intel released products in the spring of 2000 based on HomeRF in the $100 range, and other products in this price range are expected soon. More details on HomeRF can be found at http://www.homerf.org.

1.4.10 Other Wireless Systems and Applications

Many other commercial systems using wireless technology are on the market today. Remote sensor networks that collect data from unattended sensors and transmit this data back to a central processing location are being used for both indoor (equipment monitoring, climate control) and outdoor (earthquake sensing, remote data collection) applications. Satellite systems that provide vehicle tracking and dispatching (OMNITRACs) are very successful. Satellite navigation systems (the Global Positioning System or GPS) are also widely used for both military and commercial purposes. A wireless system for Digital Audio Broadcasting (DAB) has been available in Europe for quite some time and has recently been introduced in the U.S. as satellite radio. New systems and standards are constantly being developed and introduced, and this trend seems to be accelerating.

1.5 The Wireless Spectrum

1.5.1 Methods for Spectrum Allocation

Most countries have government agencies responsible for allocating and controlling the use of the radio spectrum. In the United States spectrum allocation is controlled by the Federal Communications Commission (FCC) for commercial use and by the Office of Spectral Management (OSM) for military use. The government decides how much spectrum to allocate between commercial and military use. Historically the FCC allocated spectral blocks for specific uses and assigned licenses to use these blocks to specific groups or companies. For example, in the 1980s the FCC allocated frequencies in the 800 MHz band for analog cellular phone service, and provided spectral licenses to two companies in each geographical area based on a number of criteria. While the FCC still typically allocates spectral blocks for specific purposes, over the last decade they have turned to spectral auctions for assigning licenses in each block to the highest bidder. While some argue that this market-based method is the fairest way for the government to allocate the limited spectral resource, and it provides significant revenue to the government besides, there are others who believe that this mechanism stifles innovation, limits competition, and hurts technology adoption. Specifically, the high cost of spectrum dictates that only large conglomerates can purchase it. Moreover, the large investment required to obtain spectrum delays the ability to invest in infrastructure for system rollout and results in very high initial prices for the end user. The 3G spectral auctions in Europe, in which several companies have already defaulted, have provided fuel to the fire against spectral auctions.

In addition to spectral auctions, the FCC also sets aside specific frequency bands that are free to use according to a specific set of etiquette rules. The rules may correspond to a specific communications standard, power levels, etc. The purpose of these "free bands" is to encourage innovation and low-cost implementation. Two of the most important emerging wireless systems, 802.11b wireless LANs and Bluetooth, co-exist in the free National Information Highway (NIH) band set aside at 2.5 GHz. However, one difficulty with free bands is that they can be killed by their own success: if a given system is widely used in a given band, it will generate much interference to other users colocated in that band. Satellite systems cover large areas spanning many countries and sometimes the globe. For wireless systems that span multiple countries, spectrum is allocated by the International Telecommunications Union Radio Communications group (ITU-R). The standards arm of this body, ITU-T, adopts telecommunication standards for global systems that must interoperate with each other across national boundaries.

1.5.2 Spectrum Allocations for Existing Systems

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

As discussed in the previous section, spectrum is allocated either in licensed bands (which the FCC assigns to specific operators) or in unlicensed bands (which can be used by any operator subject to certain operational requirements). The following table shows the licensed spectrum allocated to major commercial wireless systems in the U.S. today.

AM Radio	535-1605 KHz
FM Radio	88-108 MHz
Broadcast TV (Channels 2-6)	54-88 MHz
Broadcast TV (Channels 7-13)	174-216 MHz
Broadcast TV (UHF)	470-806 MHz
3G Broadband Wireless	746-764 MHz, 776-794 MHz
3G Broadband Wireless	1.7-1.85 MHz, 2.5-2.69 MHz
Analog and 2G Digital Cellular Phones	806-902 MHz
Personal Communications Service (2G Cell Phones)	1.85-1.99 GHz
Wireless Communications Service	2.305-2.32 GHz, 2.345-2.36 GHz
Satellite Digital Radio	2.32-2.325 GHz
Multichannel Multipoint Distribution Service (MMDS)	2.15-2.68 GHz
Digital Broadcast Satellite (Satellite TV)	12.2-12.7 GHz
Digital Electronic Message Service (DEMS)	24.25-24.45 GHz, 25.05-25.25 GHz
Teledesic	18.8-19.3 GHz, 28.6-29.1 GHz
Local Multipoint Distribution Service (LMDS)	27.5-29.5 GHz, 31-31.3 GHz
Fixed Wireless Services	38.6-40 GHz

Note that digital TV is slated for the same bands as broadcast TV. By 2006 all broadcasters are expected to switch from analog to digital transmission. Also, the 3G broadband wireless spectrum is currently allocated to UHF TV stations 60-69, but is slated to be reallocated for 3G. Both analog and 2G digital cellular services occupy the same cellular band at 800 MHz, and the cellular service providers decide how much of the band to allocate between digital and analog service.

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

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

ISM Band I has licensed users transmitting at high power that interfere with the unlicensed users. Therefore, the requirements for unlicensed use of this band is highly restrictive and performance is somewhat poor. The NII bands were set aside recently to provide a total of 300 MHz of spectrum with very few restrictions. It is expected that many new applications will take advantage of this large amount of unlicensed spectrum.

1.6 Standards

Communication systems that interact with each other require standardization. Standards are typically decided on by national or international committees: in the U.S. the TIA plays this role. These committees adopt standards that are developed by other organizations. The IEEE is the major player for standards

development in the United States, while ETSI plays this role in Europe. Both groups follow a lengthy process for standards development which entails input from companies and other interested parties, and a long and detailed review process. The standards process is a large time investment, but companies participate since if they can incorporate their ideas into the standard, this gives them an advantage in developing the resulting system. In general standards do not include all the details on all aspects of the system design. This allows companies to innovate and differentiate their products from other standardized systems. The main goal of standardization is for systems to interoperate with other systems following the same standard.

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

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

Bibliography

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

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

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

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

[5] A. J. Goldsmith and S. B. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," IEEE Wireless Communications Magazine, Aug. 2002.

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

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

[8] D. D. Falconer, F. Adachi, B. Gudmundson, "Time division multiple access methods for wireless personal communications," *IEEE Commun. Mag.*, pp.50-57, Jan. 1995.

[9] A. J. Viterbi, *CDMA Principles of Spread Spectrum Communications,* Addison-Wesley, 1995.

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

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

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

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

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

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

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

[17] *IEEE Personal Communications Magazine: Special Issue on Wireless ATM*, August 1996.

[18] K. Pahlavan and A. H. Levesque. *Wireless Information Networks.* New York, NY: John Wiley & Sons, Inc., 1995.

[19] *IEEE Pers. Commun. Mag: Special Issue on Smart Antennas*, February 1998.

[20] S. Verdú. *Multiuser Detection.* Cambridge, U.K.: Cambridge University Press, 1998.

[21] I. Katzela and M. Naghshineh. Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey. *IEEE Pers. Commun. Mag.*, pages 10–22, June 1996.

[22] G. Pottie. System design choices in personal communications. *IEEE Pers. Commun. Mag.*, pages 50–67, October 1995.

[23] P. Bhagwat, C. Perkins, and S. Tripathi, "Network layer mobility: an architecture and survey," *IEEE Pers. Commun. Mag.,* pp. 54-64, June 1996.

[24] A. Nakajima, "Intelligent network architecture for mobile multimedia communication," *IEICE Trans. Commun.,* pp. 1073-1082, Sep. 1994.

[25] D. Raychaudhuri, "Wireless ATM networks: architecture, system design and prototyping," *IEEE Pers. Commun. Mag.,* pp. 42-49, August 1996.

[26] E. Ayanoglu, K. Y. Eng, and M. J. Karol, "Wireless ATM: limits, challenges, and proposals," *IEEE Pers. Commun. Mag.,* pp. 18-34, Aug. 1996.

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

[28] R. Kohno, R. Meidan, and L. B. Milstein, "Spread spectrum access methods for wireless communications," *IEEE Commun. Mag.,* pp. 58–67, Jan. 1995.

Chapter 2

Path Loss and Shadowing

The wireless radio channel poses a severe challenge as a medium for reliable high-speed communication. It is not only susceptible to noise, interference, and other channel impediments, but these impediments change over time in unpredictable ways due to user movement. In this chapter we will characterize the variation in received signal power over distance due to path loss and shadowing. Path loss is caused by dissipation of the power radiated by the transmitter as well as effects of the propagation channel. Path loss models generally assume that path loss is the same at a given transmit-receive distance[1]. Shadowing is caused by obstacles between the transmitter and receiver that absorb power. When the obstacle absorbs all the power, the signal is blocked. Variation due to path loss occurs over very large distances (100-1000 meters), whereas variation due to shadowing occurs over distances proportional to the length of the obstructing object (10-100 meters in outdoor environments and less in indoor environments). Since variations due to path loss and shadowing occur over relatively large distances, this variation is sometimes refered to as **large-scale propagation effects** or **local mean attenuation**. Chapter 3 will deal with variation due to the constructive and destructive addition of multipath signal components. Variation due to multipath occurs over very short distances, on the order of the signal wavelength, so these variations are sometimes refered to as **small-scale propagation effects** or **multipath fading**. Figure 2.1 illustrates the ratio of the received-to-transmit power in dB versus log-distance for the combined effects of path loss, shadowing, and multipath.

After a brief introduction and 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 ray tracing propagation models. These models are used to approximate wave propagation according to Maxwell's equations, and are accurate models when the number of multipath components is small and the physical environment is known. Ray tracing models depend heavily on the geometry and dielectric properties of the region through which the signal propagates. We therefore also present some simple generic models with a few parameters that are commonly used in practice for system analysis and "back-of-the-envelope" system design. When the number of multipath components is large, or the geometry and dielectric properties of the propagation environment are unknown, statistical models must be used. These statistical multipath models will be described in Chapter 3.

While this chapter gives a brief overview of channel models for path loss and shadowing, comprehensive coverage of channel and propagation models at different frequencies of interest merits a book in its own right, and in fact there are several excellent texts on this topic [3, 4]. Channel models for specialized systems, e.g. multiple antenna (MIMO) and ultrawideband (UWB) systems, can be found in [62, 63].

[1]This assumes that the path loss model does not include shadowing effects

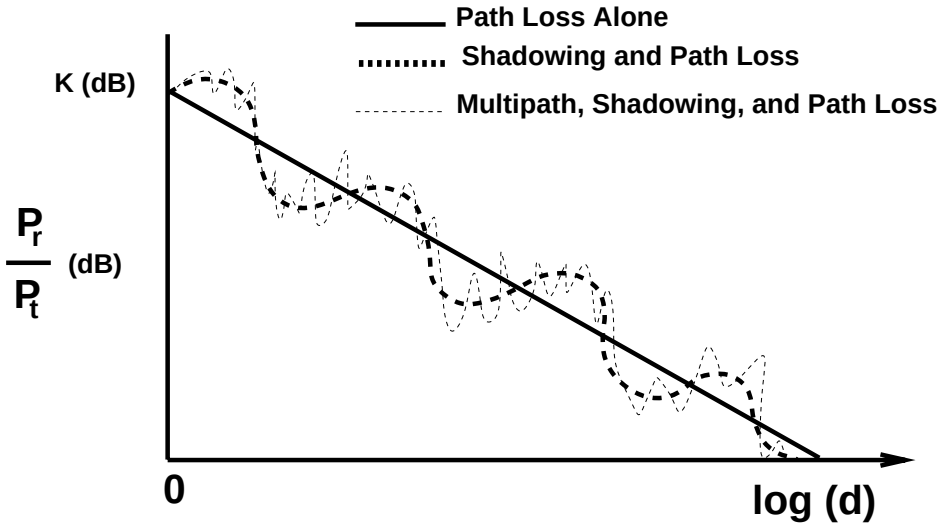


Figure 2.1: Path Loss, Shadowing and Multipath versus Distance.

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 the electromagnetic theory of light and 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. The work of Maxwell and Hertz initiated the field of radio communications: in 1894 Oliver Lodge used these principles to build the first wireless communication system, however 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 18 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 done by Reginald Fessenden in 1906 using a form of amplitude modulation, which got around 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 requires the calculation of the Radar Cross Section (RCS) of large and complex structures. Since these calculations are difficult, and many times the necessary parameters are not available, approximations have been developed to characterize signal propagation without resorting to Maxwell's equations.

The most common approximations use ray-tracing techniques. These techniques approximate the propagation of electromagnetic waves by representing the wavefronts as simple particles: the model determines the reflection and refraction effects on the wavefront but ignores the more complex scattering phenomenon predicted by Maxwell's coupled differential equations. 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, rural roads, and over water. We next consider more complex models with additional reflected, scattered, or diffracted components. Many propagation environments are not accurately reflected with 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 makes 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. Statistical models are most accurate in environments with fairly regular geometries and uniform dielectric properties. 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 [1].

2.2 Transmit and Receive Signal Models

Our models are developed mainly for signals in the UHF and SHF bands, from .3-3 GHz and 3-30 GHz, respectively. This range of frequencies is quite favorable for wireless system operation due to its propagation characteristics and relatively small required antenna size. We assume the transmission distances on the earth are small enough so as not to be affected by the earth's curvature.

All transmitted and received signals we consider are real. That is because modulators and demodulators are built using oscillators that generate real sinusoids (not complex exponentials). Thus, the transmitted signal output from a modulator is a real signal. Similarly, the demodulator only extracts the real part of the signal at its input. However, we typically model channels as having a complex frequency response due to the nature of the Fourier transform. As a result, real modulated and demodulated signals are often represented as the real part of a complex signal to facilitate analysis. This model gives rise to the complex baseband representation of bandpass signals, which we use for our transmitted and received signals. More details on the complex baseband representation for bandpass signals and systems can be found in Appendix A.

We model the transmitted signal as

$$\begin{aligned} s(t) &= \Re\left\{u(t)e^{j(2\pi f_c t+\phi_0)}\right\} \\ &= \Re\{u(t)\}\cos(2\pi f_c t+\phi_0) - \Im\{u(t)\}\sin(2\pi f_c t+\phi_0) \\ &= x(t)\cos(2\pi f_c t+\phi_0) - y(t)\sin(2\pi f_c t+\phi_0), \end{aligned} \tag{2.1}$$

where $u(t) = x(t) + jy(t)$ is a complex baseband signal with in-phase component $x(t) = \Re\{u(t)\}$, quadrature component $y(t) = \Im\{u(t)\}$, bandwidth B, and power P_u. The signal $u(t)$ is called the **complex envelope** or **complex lowpass equivalent signal** of $s(t)$. We call $u(t)$ the complex envelope of $s(t)$ since the magnitude of $u(t)$ is the magnitude of $s(t)$ and the phase of $u(t)$ is the phase of $s(t)$ relative to the carrier frequency f_c and initial phase offset ϕ_0. This is a common representation for bandpass signals with bandwidth $B << f_c$, as it allows signal manipulation via $u(t)$ irrespective of the carrier frequency and phase. The power in the transmitted signal $s(t)$ is $P_t = P_u/2$.

The received signal will have a similar form:

$$r(t) = \Re\left\{v(t)e^{j(2\pi f_c t+\phi_0)}\right\}, \tag{2.2}$$

where the complex baseband signal $v(t)$ will depend on the channel through which $s(t)$ propagates. In particular, as discussed in Appendix A, if $s(t)$ is transmitted through a time-invariant channel then $v(t) = u(t) * c(t)$, where $c(t)$ is the equivalent lowpass channel impulse response for the channel. Time-varying channels will be treated in Chapter 3. The received signal may have a Doppler shift of $f_D = v\cos\theta/\lambda$ associated with it, where θ is the arrival angle of the received signal relative to the direction of motion, v is the receiver velocity, and $\lambda = c/f_c$ is the signal wavelength ($c = 3 \times 10^8$ m/s is the speed of light). The geometry associated with the Doppler shift is shown in Fig. 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 $\Delta\phi = 2\pi v\Delta t\cos\theta/\lambda$. The Doppler frequency is then obtained from the relationship between signal frequency and phase:

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

If the receiver is moving towards the transmitter, i.e. $-\pi/2 \leq \theta \leq \pi/2$, then the Doppler frequency is positive, otherwise it is negative. We will ignore the Doppler term in the free-space and ray tracing models of this chapter, since for typical urban vehicle speeds (60 mph) and frequencies (around 1 GHz), it is less than 70 Hz [2]. However, we will include Doppler effects in Chapter 3 on statistical fading models.

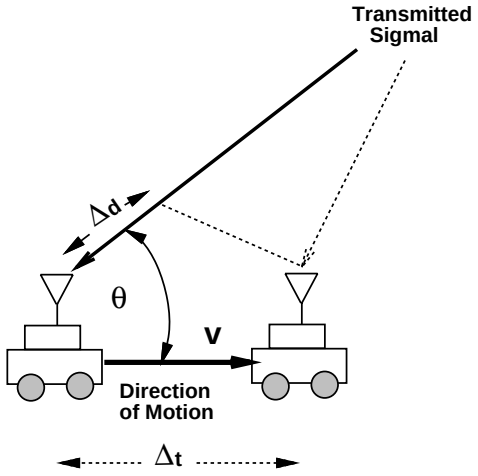


Figure 2.2: Geometry Associated with Doppler Shift.

Suppose $s(t)$ 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 dB value of the linear path loss or, equivalently, the difference in dB between the transmitted and received signal power:

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

In general path loss is a nonnegative number since the channel does not contain active elements, and thus can only attenuate the signal. The **path gain** in dB is defined as the negative of the path loss: $P_G = -P_L = 10\log_{10}(P_r/P_t)$, which is generally a negative number. With shadowing the received power will include the effects of path loss and an additional random component due to blockage from objects, as we discuss in Section 2.7.

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 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. Free-space path loss introduces a complex scale factor [3], resulting in the received signal

$$r(t) = \Re\left\{\frac{\lambda\sqrt{G_l}e^{-j(2\pi d/\lambda)}}{4\pi d}u(t)e^{j2\pi f_c t}\right\} \tag{2.6}$$

where $\sqrt{G_l}$ is the product of the transmit and receive antenna field radiation patterns in the LOS direction. The phase shift $e^{-j(2\pi d/\lambda)}$ is due to the distance d the wave travels, and can also be written in terms of the associated delay $\tau = d/c$ as $e^{-j(2\pi f_c\tau)}$, where c is the speed of light.

The power in the transmitted signal $s(t)$ is P_t, so the ratio of received to transmitted power computed from (2.6) is

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

Thus, the received signal power falls off inversely proportional 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 models, the received signal power falls off more quickly relative to this distance. The received signal power is also proportional to the square of the signal wavelength, so as the carrier frequency increases, the received power decreases. This dependence of received power on the signal wavelength λ is due to the effective area of the receive antenna [3]. However, the antenna gain of highly directional antennas can increase with frequency, so that receive power may actually increase with frequency for highly directional links. The received power can be expressed in dBm as

$$P_r\,(\text{dBm}) = P_t\,(\text{dBm}) + 10\log_{10}(G_l) + 20\log_{10}(\lambda) - 20\log_{10}(4\pi) - 20\log_{10}(d). \tag{2.8}$$

Free-space path loss is defined as the path loss of the free-space model:

$$P_L\,(\text{dB}) = 10\log_{10}\frac{P_t}{P_r} = -10\log_{10}\frac{G_l\lambda^2}{(4\pi d)^2}. \tag{2.9}$$

The **free-space path gain** is thus

$$P_G = -P_L = 10\log_{10}\frac{G_l\lambda^2}{(4\pi d)^2}. \tag{2.10}$$

Example 2.1: Consider an indoor wireless LAN with $f_c = 900$ MHz, cells of radius 10 m, and nondirectional antennas. Under the free-space path loss model, what transmit power is required at the access

point such that all terminals within the cell receive a minimum power of 10 μW. How does this change if the system frequency is 5 GHz?

Solution: We must find the 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:

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

Substituting in $G_l = 1$ (nondirectional antennas), $\lambda = c/f_c = .33$ m, $d = 100$ m, and $P_t = 10\mu$W yields $P_t = 145.01$W $= 21.61$ dBW (Recall that P Watts equals $10\log_{10}[P]$ dbW, dB relative to one Watt, and $10\log_{10}[P/.001]$ dBm, dB relative to one milliwatt). At 5 GHz only $\lambda = .06$ changes, so $P_t = 4.39$ KW $= 36.42$ dBW.

2.4 Ray Tracing

In a typical urban or indoor environment, a radio signal transmitted from a fixed source will encounter multiple objects in the environment that produce reflected, diffracted, or scattered copies of the transmitted signal, as shown in Figure 2.3. These additional copies of the transmitted signal, called multipath signal components, can be attenuated in power, delayed in time, and shifted in phase and/or frequency from the LOS signal path at the receiver. The multipath and transmitted signal are summed together at the receiver, which often produces distortion in the received signal relative to the transmitted signal.

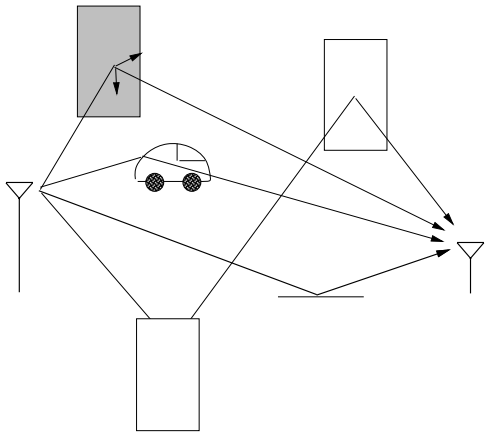


Figure 2.3: Reflected, Diffracted, and Scattered Wave Components

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 reflection, diffraction, and scattering effects 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 all the scatterers are large relative to a wavelength and fairly smooth. Comparison of the ray tracing method with empirical data shows it to accurately model received signal power in rural areas [9], along city streets where both the transmitter and receiver are close to the ground [7, 6, 9], or in indoor environments with appropriately adjusted diffraction coefficients [8]. Propagation effects besides received power variations, such as the delay spread of the multipath, are not always well-captured with ray tracing techniques [10].

If the transmitter, receiver, and reflectors are all immobile then the impact of the multiple received signal paths, and their delays relative to the LOS path, are fixed. However, if the source 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 very large or the reflector surfaces are not smooth 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 [12], 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. Computer programs based on ray tracing such as Lucent's Wireless Systems Engineering software (WiSE), Wireless Valley's SitePlanner®, and Marconi's Planet® *EV* are widely used for system planning in both indoor and outdoor environments. In these programs computer graphics are combined with aerial photographs (outdoor channels) or architectural drawings (indoor channels) to obtain a 3D geometric picture of the environment [1].

The following sections describe several ray tracing models of increasing complexity. We start with a simple two-ray model that predicts signal variation resulting from a ground reflection interfering with the LOS path. This model characterizes signal propagation in isolated areas with few reflectors, such as rural roads or highways. It is not typically a good model for indoor environments. We then present a ten-ray reflection model that predicts the variation of a signal propagating along a straight street or hallway. Finally, we describe a general model that predicts signal propagation for any propagation environment. The two-ray model only requires information about the antenna heights, while the ten-ray model requires antenna height and street/hallway width information, and the general model requires these parameters as well as detailed information about the geometry and dielectric properties of the reflectors, diffractors, and scatterers in the environment.

2.4.1 Two-Ray Model

The two-ray model is used when a single ground reflection dominates the multipath effect, as illustrated in Figure 2.4. The received signal consists of two components: the LOS component or ray, which is just 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). The reflected ray is shown in Figure 2.4 by the segments x and x'. If we ignore the effect of surface wave attenuation[2] then, by superposition, the received signal for the two-ray model is

$$r_{2ray}(t) = \Re\left\{\frac{\lambda}{4\pi}\left[\frac{\sqrt{G_l}u(t)e^{j(2\pi l/\lambda)}}{l} + \frac{R\sqrt{G_r}u(t-\tau)e^{-j2\pi(x+x')/\lambda}}{x+x'}\right]e^{j(2\pi f_c t+\phi_0)}\right\}, \tag{2.11}$$

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

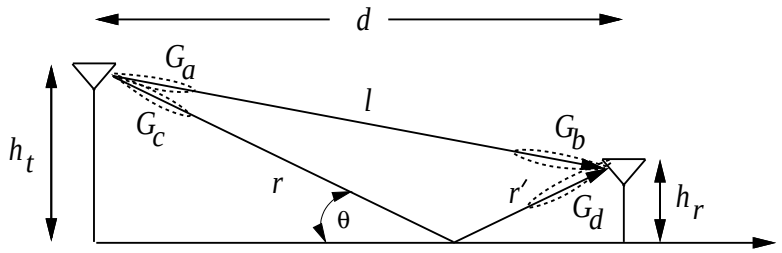


Figure 2.4: Two-Ray Model.

where $\tau = (x + x' - l)/c$ is the time delay of the ground reflection relative to the LOS ray, $\sqrt{G_l} = \sqrt{G_a G_b}$ is the product of the transmit and receive antenna field radiation patterns in the LOS direction, R is the ground reflection coefficient, and $\sqrt{G_r} = \sqrt{G_c G_d}$ is the product of the transmit and receive antenna field radiation patterns corresponding to the rays of length x and x', respectively. The **delay spread** of the two-ray model equals the delay between the LOS ray and the reflected ray: $(x + x' - l)/c$.

If the transmitted signal is narrowband relative to the delay spread ($\tau << B_u^{-1}$) then $u(t) \approx u(t-\tau)$. Thus, the received power of the two-ray model for narrowband transmission is

$$P_r = P_t \left[\frac{\lambda}{4\pi}\right]^2 \left| \frac{\sqrt{G_l}}{l} + \frac{R\sqrt{G_r} e^{j\Delta\phi}}{r + r'} \right|^2, \tag{2.12}$$

where $\Delta\phi = 2\pi(r' + r - l)/\lambda$ is the phase difference between the two received signal components. Equation (2.12) has been shown to agree very closely with empirical data [13]. If d denotes the horizontal separation of the antennas, h_t denotes the transmitter height, and h_r denotes the receiver height, then using geometry we can show that

$$x + x' - l = \sqrt{(h_t + h_r)^2 + d^2} - \sqrt{(h_t - h_r)^2 + d^2}. \tag{2.13}$$

When d is very large compared to $h_t + h_r$ we can use a Taylor series approximation in (2.13) to get

$$\Delta\phi = \frac{2\pi(x + x' - l)}{\lambda} \approx \frac{4\pi h_t h_r}{\lambda d}. \tag{2.14}$$

The ground reflection coefficient is given by [2, 14]

$$R = \frac{\sin\theta - Z}{\sin\theta + Z}, \tag{2.15}$$

where

$$Z = \begin{cases} \sqrt{\epsilon_r - \cos^2\theta}/\epsilon_r & \text{for vertical polarization} \\ \sqrt{\epsilon_r - \cos^2\theta} & \text{for horizontal polarization} \end{cases}, \tag{2.16}$$

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).

We see from Figure 2.4 and (2.15) that for asymptotically large d, $x + x' \approx l \approx d$, $\theta \approx 0$, $G_l \approx G_r$, and $R \approx -1$. Substituting these approximations into (2.12) yields that, in this asymptotic limit, the received signal power is approximately

$$P_r \approx \left[\frac{\lambda\sqrt{G_l}}{4\pi d}\right]^2 \left[\frac{4\pi h_t h_r}{\lambda d}\right]^2 P_t = \left[\frac{\sqrt{G_l} h_t h_r}{d^2}\right]^2 P_t, \tag{2.17}$$

or, equivalently, the dB attenuation is given by

$$P_r \text{ (dBm)} = P_t \text{ (dBm)} + 10\log_{10}(G_l) + 20\log_{10}(h_t h_r) - 40\log_{10}(d). \tag{2.18}$$

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. The received signal becomes independent of λ since the cancellation of the two multipath rays changes the effective area of the receive antenna. A plot of (2.18) as a function of distance is illustrated in Figure 2.5 for $f = 900$MHz, R=-1, $h_t = 50$m, $h_r = 2$m, $G_l = 1$, $G_r = 1$ and transmit power normalized so that the plot starts at 0 dBm. This plot can be separated into three segments. For small distances $d < h_t$, the path loss is roughly flat and proportional to $1/(d^2 + h_t^2)$ since, at these small distances, the distance between the transmitter and receiver is $D = \sqrt{d^2 + (h_t - h_r)^2}$ and thus $1/D^2 \approx 1/(d^2 + h_t^2)$ for $h_t >> h_r$, which is typically the case. Then, for distances bigger than h_t and up to a certain critical distance d_c, the wave experiences constructive and destructive interference of the two rays, resulting in a wave pattern with a sequence of maxima and minima. These maxima and minima are also refered to as small-scale or 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 to d^{-4}. At this critical distance the signal components only combine destructively, so they are out of phase by at least π. An approximation for d_c can be obtained by setting $\Delta\phi = \pi$ in (2.14), obtaining $d_c = 4h_t h_r/\lambda$, which is also shown in the figure. The power falloff with distance in the two-ray model can be approximated by a piecewise linear model with three segments, which is also shown in Figure 2.5 slightly offset from the actual power falloff curve for illustration purposes. In the first segment power falloff is constant and proportional to $1/(d^2 + h_t^2)$, 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 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 2.5 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 urban microcell ($h_t = 10$m, $h_r = 3$ m) and an indoor microcell ($h_t = 3$m, $h_r = 2$ m) for $f_c = 2$ GHz.

Solution: $d_c = 4h_t h_r/\lambda = 800$ meters for the urban microcell and 160 meters for the indoor system. A cell radius of 800 m in an urban microcell system is a bit large: urban microcells today are on the order of 100 m to maintain large capacity. However, if we used a cell size of 800 m under these system parameters, signal power would fall off as d^2 inside the cell, and interference from neighboring cells would fall off as d^4, and thus would be greatly reduced. Similarly, 160m is quite large for the cell radius of an indoor system, as there would typically be many walls the signal would have to go through for an indoor cell radius of that size. So an indoor system would typically have a smaller cell radius, on the order of 10-20 m.

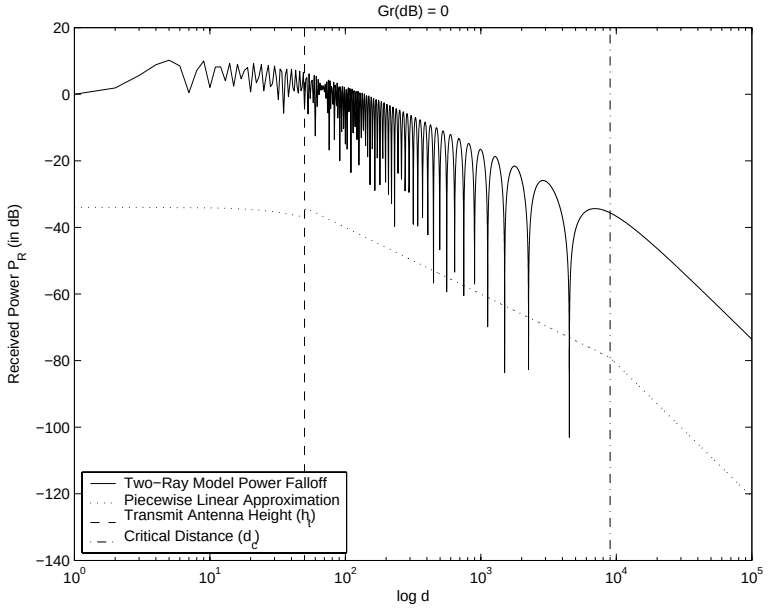


Figure 2.5: Received Power versus Distance for Two-Ray Model.

If we average out the local maxima and minima in (2.12), the resulting average power loss in dB versus log-distance can be approximated by dividing the power loss curve of Fig. 2.5 into three straight-line segments as follows. For $d < h_t$ the average power falloff with distance is constant. For $h_t < d < d_c$ the average power falloff with distance corresponds to free space where power falls off proportional to distance squared. For $d > d_c$, the power falloff with distance is approximated by the fourth-power law in (2.17). This three-segment model is a special case of the piecewise linear model, described in more detail in Section 2.6.5. The three straight-line segments that approximate the two-ray model, corresponding to constant power, free-space power falloff, and the fourth-power law, are also illustrated in Fig. 2.5.

2.4.2 Dielectric Canyon (Ten-Ray Model)

We now examine a model for urban area transmissions developed by Amitay [7]. This model assumes rectilinear streets[3] with buildings along both sides of the street and transmitter and receiver antenna heights that are well below the tops of the buildings. The building-lined streets act as a dielectric canyon to the propagating signal. 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 three reflections can generally be ignored. When the street layout is relatively straight, back reflections are usually negligible also. Experimental data show that a model of ten reflection rays closely approximates signal propagation through the dielectric canyon [7]. The ten rays incorporate all paths with one, two, or three reflections: specifically, there is the LOS, the ground-reflected (GR), the single-wall (SW) reflected, the double-wall (DW) reflected, the triple-wall (TW) reflected, the wall-ground (WG) reflected and the ground-wall (GW) 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 2.6.

[3] A rectilinear city is flat, with linear streets that intersect at 90^o angles, as in midtown Manhattan.

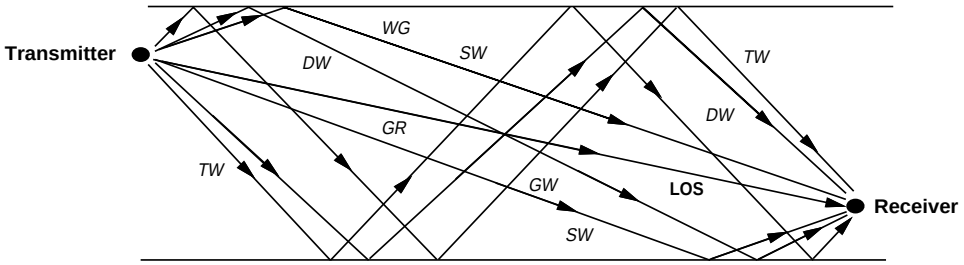


Figure 2.6: Overhead View of the Ten-Ray Model.

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

$$r_{10ray}(t) = \Re\left\{ \frac{\lambda}{4\pi}\left[\frac{\sqrt{G_l}u(t)e^{j(2\pi l)/\lambda}}{l} + \sum_{i=1}^{9} \frac{R_i\sqrt{G_{x_i}}u(t-\tau_i)e^{j(2\pi x_i)/\lambda}}{x_i} \right] e^{j(2\pi f_c t+\phi_0)}, \right\} \tag{2.19}$$

where x_i denotes the path length of the ith reflected ray, $\tau_i = (x_i - l)/c$, and $\sqrt{G_{x_i}}$ is the product of the transmit and receive antenna gains corresponding to the ith ray. For each reflection path, the coefficient R_i is either a single reflection coefficient given by (2.15) or, if the path corresponds to multiple reflections, the product of the reflection coefficients corresponding to each reflection. The dielectric constants used in (2.15) are approximately the same as the ground dielectric, so $\epsilon_r = 15$ is used for all the calculations of R_i. If we again assume a narrowband model such that $u(t) \approx u(t-\tau_i)$ for all i, then the received power corresponding to (2.19) is

$$P_r = P_t\left[\frac{\lambda}{4\pi}\right]^2 \left| \frac{\sqrt{G_l}}{l} + \sum_{i=1}^{9} \frac{R_i\sqrt{G_{x_i}}e^{j\Delta\phi_i}}{x_i} \right|^2, \tag{2.20}$$

where $\Delta\phi_i = 2\pi(x_i - l)/\lambda$.

Power falloff with distance in both the ten-ray model (2.20) and urban empirical data [13, 47, 48] for transmit antennas both above and below the building skyline is typically proportional to d^{-2}, even at relatively large distances. Moreover, this 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 (the two-ray model), which decays as d^{-4}. Other empirical studies [15, 49, 50] have obtained power falloff with distance proportional to $d^{-\gamma}$, where γ lies anywhere between two and six.

2.4.3 General Ray Tracing

General Ray Tracing (GRT) can be used to predict field strength and delay spread for any building configuration and antenna placement [11, 34, 35]. For this model, the building database (height, location, and dielectric properties) and the transmitter and receiver locations relative to the buildings must be specified exactly. Since this information is site-specific, the GRT model is not used to obtain general theories about system performance and layout; rather, it explains the basic mechanism of urban propagation, and can be used to obtain delay and signal strength information for a particular transmitter and receiver configuration.

The GRT method uses geometrical optics to trace the propagation of the LOS and reflected signal components, as well as signal components from building 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 building locations and dielectric properties. 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, which are typically blocked from the LOS and reflecting rays, these other multipath components may dominate.

The propagation model for direct and reflected paths was outlined in the previous section. Diffraction occurs when the transmitted signal "bends around" an object in its path to the receiver, as shown in Figure 2.7. Diffraction results from many phenomena, including the curved surface of the earth, hilly or irregular terrain, building edges, or obstructions blocking the LOS path between the transmitter and receiver [14, 3, 1]. Diffraction can be accurately characterized using the geometrical theory of diffraction (GTD) [38], however 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 path loss exceeding 100 dB for some incident angles on the wedge [8, 35, 36, 37]. Although wedge diffraction simplifies the GTD, it still requires a numerical solution for path loss [38, 39] and is thus is not commonly used. Diffraction is most commonly modeled by the **Fresnel knife edge diffraction model** due to its simplicity. The geometry of this model is shown in Figure 2.7, 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 surfance roughness, which can lead to inaccuracies [36]. The geometry of Figure 2.7 indicates that the diffracted signal travels distance $d + d'$ resuling in a phase shift of $\phi = 2\pi(d + d')/\lambda$. The geometry of Figure 2.7 indicates that for h small relative to d and d', the signal must travel an additional distance relative to the LOS path of approximately

$$\Delta d = \frac{h^2}{2}\frac{d + d'}{dd'},$$

and the corresponding phase shift relative to the LOS path is approximately

$$\Delta\phi = \frac{2\pi\Delta d}{\lambda} = \frac{\pi}{2}v^2 \tag{2.21}$$

where

$$v = h\sqrt{\frac{2(d + d')}{\lambda dd'}} \tag{2.22}$$

is called the **Fresnel-Kirchoff diffraction parameter**. The path loss associated with knife-edge diffraction is generally a function of v. However, computing this diffraction path loss is fairly complex, requiring the use of Huygen's principle, Fresnel zones, and the complex Fresnel integral [3]. Moreover, the resulting diffraction loss cannot generally be found in closed form. Approximations for knife-edge diffraction path loss (in dB) relative to LOS path loss are given by Lee [14, Chapter 2] as

$$L(v)(dB) = \begin{cases} 20\log_{10}[0.5 - 0.62v] & -0.8 \le v < 0 \\ 20\log_{10}[0.5e^{-.95v}] & 0 \le v < 1 \\ 20\log_{10}[0.4 - \sqrt{.1184 - (.38 - .1v)^2}] & 1 \le v < 2.4 \\ 20\log_{10}[.225/v] & v > 2.4 \end{cases}. \tag{2.23}$$

A similar approximation can be found in [40]. The knife-edge diffraction model yields the following formula for the received diffracted signal:

$$r(t) = \Re\left\{L(v)\sqrt{G_d}u(t - \tau)e^{-j(2\pi(d+d'))/\lambda}e^{j(2\pi f_c t + \phi_0)}, \right\}, \tag{2.24}$$

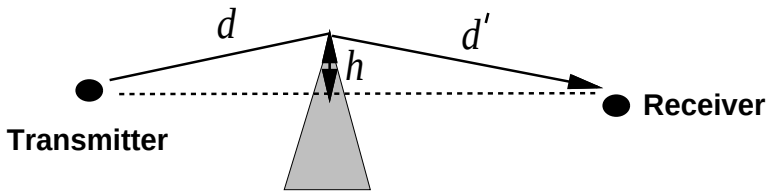


Figure 2.7: Knife-Edge Diffraction.

where $\sqrt{G_d}$ is the antenna gain and $\tau = \Delta d/c$ is the delay associated with the defracted ray relative to the LOS path.

In addition to the diffracted ray, there may also be multiple diffracted rays, or rays that are both reflected and diffracted. Models exist for including all possible permutations of reflection and diffraction [41]; however, the attenuation of the corresponding signal components is generally so large that these components are negligible relative to the noise.

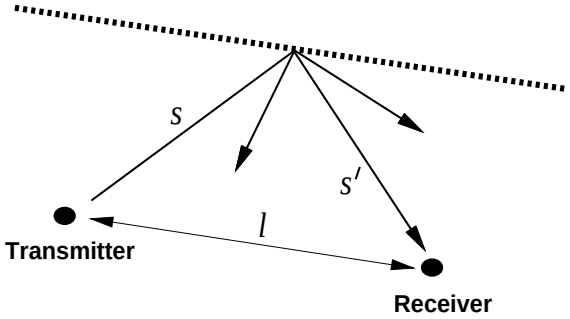


Figure 2.8: Scattering.

A scattered ray, shown in Figure 2.8 by the segments s' and s, has a path loss proportional to the product of s and s'. This multiplicative dependence is due to the additional spreading loss the ray experiences after scattering. The received signal due to a scattered ray is given by the bistatic radar equation [42]:

$$r(t) = \Re\left\{u(t-\tau)\frac{\lambda\sqrt{G_s\sigma}e^{-j(2\pi(s+s')/\lambda)}}{(4\pi)^{3/2}ss'}e^{j(2\pi f_c t+\phi_0)}\right\} \tag{2.25}$$

where $\tau = (s+s'-l)/c$ is the delay associated with the scattered ray, σ (in m^2) is the radar cross section of the scattering object, which depends on the roughness, size, and shape of the scatterer, and $\sqrt{G_s}$ is the antenna gain. The model assumes that the signal propagates from the transmitter to the scatterer based on free space propagation, and is then reradiated by the scatterer with transmit power equal to σ times the received power at the scatterer. From (2.25) the path loss associated with scattering is

$$P_r\,(\text{dBm}) = P_t\,(\text{dBm}) + 10\log_{10}(G_s) + 10\log_{10}(\sigma) - 30\log(4\pi) - 20\log s - 20\log(s'). \tag{2.26}$$

Empirical values of $10\log_{10}\sigma$ were determined in [43] for different buildings in several cities. Results from this study indicate that $\sigma = 10\log_{10}\sigma$ in dBm^2 ranges from -4.5dbm^2 to 55.7dbm^2, where dbm^2 denotes the dB value of the σ measurement with respect to one square meter.

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 diffusely scattered rays,

the total received signal is

$$\begin{aligned} r_{total}(t) &= \Re\left\{\left[\frac{\lambda}{4\pi}\right]\left[\frac{\sqrt{G_l}u(t)e^{j(2\pi l)/\lambda}}{l} + \sum_{i=1}^{N_r}\frac{R_{x_i}\sqrt{G_{x_i}}u(t-\tau_i)e^{-j(2\pi r_i/\lambda)}}{r_i}\right.\right. \\ &+ \sum_{j=1}^{N_d} L_j(v)\sqrt{G_{d_j}}u(t-\tau_j)e^{-j(2\pi(d_j+d'_j))/\lambda}e^{j(2\pi f_c t+\phi_0)}, \\ &+ \left.\left.\sum_{k=1}^{N_s}\frac{\sigma_k\sqrt{G_{s_k}}u(t-\tau_k)e^{j(2\pi(s_k+s'_k))/\lambda}}{s_k s'_k}\right]e^{j(2\pi f_c t+\phi_0)}\right\}, \end{aligned} \tag{2.27}$$

where τ_i,τ_j,τ_k is, respectively, the time delay of the given reflected, diffracted, or scattered ray normalized to the delay of the LOS ray, as defined above.

Any of these multipath components may have an additional attenuation factor if its propagation path is blocked by buildings or other objects. In this case, the attenuation factor of the obstructing object multiplies the component's path loss term in (2.27). This attenuation loss will vary widely, depending on the material and depth of the object [1, 44]. Models for random loss due to attenuation are described in Section 2.7.

2.5 Simplified Path Loss Model

The complexity of signal propagation makes it difficult to obtain a single model that characterizes path loss accurately across a range of different environments. Accurate path loss models can be obtained from complex ray tracing 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 tradeoff 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. Thus, the following simplified model for path loss as a function of distance is commonly used for system design:

$$P_r = P_t K\left[\frac{d_0}{d}\right]^{\gamma}. \tag{2.28}$$

The dB attenuation is thus

$$P_r\,(\text{dBm}) = P_t\,(\text{dBm}) + K\,(\text{dB}) - 10\gamma\log_{10}\left[\frac{d}{d_0}\right]. \tag{2.29}$$

In this approximation, K is a unitless constant which depends on the antenna characteristics and the average channel attenuation, d_0 is a reference distance for the antenna far-field, and γ is the path loss exponent. Due to scattering phenomena in the antenna near-field, the model (2.28) is generally only valid at transmission distances $d > d_0$, where d_0 is typically assumed to be 1-10 m indoors and 10-100 m outdoors. The value of $K < 1$ is sometimes set to the free space path loss at distance d_0:

$$K\,(\text{dB}) = -20\log_{10}(4\pi d_0/\lambda), \tag{2.30}$$

and this assumption is supported by empirical data for free-space path loss at a transmission distance of 100 m [32]. Alternatively, K can be determined by measurement at d_0 or optimized (alone or together with γ) to minimize the mean square error (MSE) between the model and empirical measurements [32].

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 the example below. Alternatively γ can be obtained from an empirically-based model that takes into account frequency and antenna height [32]. A table summarizing γ values for different indoor and outdoor environments and antenna heights at 900 MHz and 1.9 GHz taken from [28, 43, 32, 25, 24, 17, 20, 1] is given below. Path loss exponents at higher frequencies tend to be higher [29, 24, 23, 25] while path loss exponents at higher antenna heights tend to be lower [32]. Note that the wide range of empirical path loss exponents for indoor propagation may be due to attenuation caused by floors, objects, and partitions. These effects are discussed in more detail in Section 2.6.6.

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

Table 2.1: Typical Path Loss Exponents

Example 2.3: Consider the set of empirical measurements of P_r/P_t given in the table below for an indoor system at 2 GHz. Find the path loss exponent γ that minimizes the MSE between the simplified model (2.29) and the empirical dB power measurements, assuming that $d_0 = 1$ m and K is determined from the free space path loss formula at this d_0. Find the received power at 150 m for the simplified path loss model with this path loss exponent and a transmit power of 1 mW (0 dBm).

Distance from Transmitter	$M = P_r/P_t$
10 m	-70 dB
20 m	-75 dB
50 m	-90 dB
100 m	-110 dB
300 m	-125 dB

Table 2.2: Path Loss Measurements

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 $M_{\text{model}}(d_i) = K - 10\gamma \log_{10}(d)$ is the path loss based on (2.29) at d_i. Using the free space path loss formula, $K =$

$-20\log_{10}(4\pi)/.3333 = -31.54$ dB. Thus

$$\begin{aligned} F(\gamma) &= (-70 + 31.54 + 10\gamma)^2 + (-75 + 31.54 + 13.01\gamma)^2 + (-90 + 31.54 + 16.99\gamma)^2 \\ &+ (-110 + 31.54 + 20\gamma)^2 + (-125 + 31.54 + 24.77\gamma)^2 \\ &= 21676.3 - 11654.9\gamma + 1571.47\gamma^2. \end{aligned} \tag{2.31}$$

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

$$\frac{\partial F(\gamma)}{\partial \gamma} = -11654.9 + 3142.94\gamma = 0 \rightarrow \gamma = 3.71.$$

To find the received power at 150 m under the simplified path loss model with $K = -31.54$, $\gamma = 3.71$, and $P_t = 0$ dBm, we have $P_r(dBm) = P_t(dBm) + K(dB) - 10\gamma\log_{10}(d/d_0) = 0 - 31.54 - 10 * 3.71\log_{10}(150) = -112.27$ dBm. Clearly the measurements deviate from the simplified path loss model: this variation can be attributed to shadow fading, described in Section 2.7

2.6 Empirical Path Loss Models

Most mobile communication systems operate in complex propagation environments that cannot be accurately modeled by free-space path loss, ray tracing, or the simplified model. A number of path loss models have been developed over the years to predict path loss in typical wireless environments such as large urban macrocells, urban microcells, and, more recently, inside buildings. These models are mainly based on empirical measurements over a given distance in a given frequency range and a particular geographical area or building. However, applications of these models are not always restricted to environments in which the empirical measurements were made, which makes the accuracy of such empirically-based models applied to more general environments somewhat questionable. Nevertheless, many wireless systems use these models as a basis for performance analysis. In our discussion below we will begin with common models for urban macrocells, then describe more recent models for outdoor microcells and indoor propagation.

2.6.1 Okumura's Model

One of the most common models for signal prediction in large urban macrocells is Okumura's model [52]. 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, the upper end of which is higher than typical base stations today. The path loss formula of Okumura is given by

$$L_{50}\,(\text{dB}) = L_f + A_{mu}(f, d) - G(h_t) - G(h_r) - G_{AREA} \tag{2.32}$$

where d is the distance between transmitter and receiver, L_{50} is the median (50th percentile) value of propagation path loss, L_f is free space path loss, A_{mu} 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}

and G_{AREA} are obtained from Okumura's empirical plots [52, 1]. Okumura derived empirical formulas for $G(h_t)$ and $G(h_r)$ as

$$G(h_t) = 20\log_{10}(h_t/200), \quad 30m < h_t < 1000m \tag{2.33}$$

$$G(h_r) = \begin{cases} 10\log_{10}(h_r/3) & h_r \leq 3m \\ 20\log_{10}(h_r/3) & 3m < h_r < 10m \end{cases}. \tag{2.34}$$

Correction factors related to terrain are also developed in [52] that improve the model 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.

2.6.2 Hata Model

The Hata model [51] 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 since it is a closed-form formula and is not based on empirical curves for the different parameters. The standard formula for median path loss in urban areas under the Hata model is

$$L_{50,urban}(dB) = 69.55 + 26.16\log_{10}(f_c) - 13.82\log_{10}(h_{te}) - a(h_{re}) + (44.9 - 6.55\log_{10}(h_{te}))\log_{10}(d). \tag{2.35}$$

The parameters in this model are the same as under the Okumura model, and $a(h_{re})$ is a correction factor for the mobile antenna height based on the size of the coverage area. For small to medium sized cities, this factor is given by [51, 1]

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

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,

$$L_{50,suburban}(dB) = L_{50,urban}(dB) - 2[\log_{10}(f_c/28)]^2 - 5.4 \tag{2.36}$$

and

$$L_{50,rural}(dB) = L_{50,urban}(dB) - 4.78[\log_{10}(f_c)]^2 + 18.33\log_{10}(f_c) - K, \tag{2.37}$$

where K ranges from 35.94 (countryside) to 40.94 (desert). Hata's model does not provide for any path specific correction factors, as is available in the Okumura model. The Hata model well-approximates the Okumura model for distances $d > 1$ Km. Thus, it is a good model for first generation cellular systems, but does not model propagation well in current cellular systems with smaller cell sizes and higher frequencies. Indoor environments are also not captured with the Hata model.

2.6.3 COST231 Extension to Hata Model

The Hata model was extended by the European cooperative for scientific and technical research (EURO-COST) to 2 GHz as follows [53]:

$$L_{50,urban}(dB) = 46.3 + 33.9\log_{10}(f_c) - 13.82\log_{10}(h_t) - a(h_r) + (44.9 - 6.55\log_{10}(h_t))\log_{10}(d) + C_M, \tag{2.38}$$

where $a(h_r)$ is the same correction factor as before and C_M is 0 dB for medium sized cities and suburbs, and 3 dB for metropolitan areas. This model is referred to as the COST-231 extension to the Hata model, and is restricted to the following range of parameters: 1.5GHz $< f_c <$ 2 GHz, $30m < h_t <$ 200 m, 1m $< h_r <$ 10 m, 1Km $< d <$ 20 Km.

2.6.4 Walfisch/Bertoni Model

The COST extension to the Hata model does not consider the impact of diffraction from rooftops and buildings. A model for these effects was developed by Walfisch and Bertoni [54]. This model uses diffraction to predict average signal strength at street level. The model considers the path loss to be the product of three factors:

$$L = P_0 Q^2 P_1, \tag{2.39}$$

where P_0 is the free space path loss for omnidirectional antennas, Q^2 reflects the signal power reduction due to buildings that block the receiver at street level, and P_1 is based on the signal loss from the rooftop to the street due to diffraction. The model has been adopted for the IMT-2000 standard.

Other commonly used empirical models for path loss in macrocells include the Longley-Rice model, the Durkin model, and the Feuerstein model. Details of these models can be found in [1].

2.6.5 Piecewise Linear (Multi-Slope) Model

A common method for modeling path loss in outdoor microcells and indoor channels is a piecewise linear model of dB loss versus log-distance. This approximation is illustrated in Figure 2.9 for dB attenuation versus log-distance, where the dots represent hypothetical empirical measurements and the piecewise linear model represents an approximation to these measurements. A piecewise linear model with N segments must specify $N-1$ breakpoints $d_1, \ldots, d_{N-1}$ and the slopes corresponding to each segment $s_1, \ldots, s_N$. Different methods can be used to determine the number and location of breakpoints to be used in the model. Once these are fixed, the slopes corresponding to each segment can be obtained by linear regression. The piecewise linear model has been used to model path loss for outdoor channels in [16] and for indoor channels in [45].

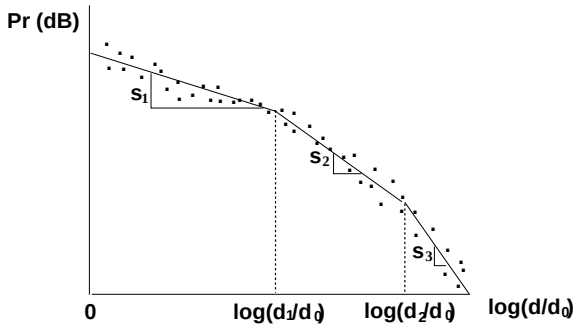


Figure 2.9: Piecewise Linear Model for Path Loss.

A special case of the piecewise 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_0 up to some critical distance d_c, after which point power falls off with path loss exponent γ_2:

$$P_r(dB) = \begin{cases} P_t + K - 10\gamma_1 \log_{10}(d/d_0) & d_0 \le d \le d_c \\ P_t + K - 10\gamma_1 \log_{10}(d_c/d_0) - 10\gamma_2 \log_{10}(d/d_c) & d > d_c \end{cases}. \tag{2.40}$$

The path loss exponents, K, and d_c are typically obtained via a regression fit to empirical data [32, 30]. The two-ray model described in Section 2.4 for $d > h_t$ can be approximated with the dual-slope model, with one breakpoint at the critical distance d_c and attenuation slope $s_1 = 20$ dB/decade and $s_2 = 40$ dB/decade.

The multiple equations in the dual-slope model can be captured with the following dual-slope approximation [15, 46]:

$$P_r = \frac{P_t K}{L(d)}, \tag{2.41}$$

where

$$L(d) \triangleq \left[\frac{d}{d_0}\right]^{\gamma_1} \sqrt[q]{1 + \left(\frac{d}{d_c}\right)^{(\gamma_1 - \gamma_2)q}}. \tag{2.42}$$

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_c. This model can be extended to more than two regions [16].

2.6.6 Indoor Propagation 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, and 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. Thus, it is difficult to find generic models that can be accurately applied to determine path loss in a specific indoor setting.

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 passed through. Specifically, measurements in [17, 19, 24, 20] indicate that at 900 MHz the attenuation when the transmitter and receiver are separated by a single floor ranges from 10-20 dB, while subsequent floor attenuation is 6-10 dB per floor for the next three floors, and then a few dB per floor for more than four floors. At higher frequencies the attenuation loss per floor is typically larger [19, 18]. The attenuation per floor is thought to decrease as the number of attenuating floors increases due to 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 [1, 21, 22, 17, 23], and Table 2.3 indicates a few examples of partition losses measured at 900-1300 MHz from this data. The partition loss obtained by different researchers for the same partition type at the same frequency often varies widely, making it difficult to make generalizations about partition loss from a specific data set.

Partition Type	Partition Loss in dB
Cloth Partition	1.4
Double Plasterboard Wall	3.4
Foil Insulation	3.9
Concrete wall	13
Aluminum Siding	20.4
All Metal	26

Table 2.3: Typical Partition Losses

The experimental data for floor and partition loss can be incorporated into the simple path loss

model (2.28) as

$$P_r = P_t + K - 10\gamma \log_{10}\left[\frac{d}{d_0}\right] - \sum_{i=1}^{N_f} FAF_i - \sum_{i=1}^{N_p} PAF_i, \tag{2.43}$$

where γ is obtained from the path loss for a same floor measurement (e.g. from Table 2.1), FAF_i represents the floor attenuation factor (FAF) for the ith floor traversed by the signal, and PAF_i represents the partition attenuation factor (PAF) 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 where the transmitter is located outside the building is the building penetration loss. Measurements indicate that building penetration loss is a function of frequency, height, and the building materials. Building penetration loss on the ground floor typically range from 8-20 dB for 900 MHz to 2 GHz [25, 26, 3]. The penetration loss decreases slightly as frequency increases, and also decreases by about 1.4 dB per floor at floors above the ground floor. This increase is typically due to reduced clutter at higher floors and the higher likelihood of a line-of-sight path. The type and number of windows in a building also have a significant impact on penetration loss [27]. Measurements made behind windows have about 6 dB less penetration loss than measurements made behind exterior walls. Moreover, plate glass has an attenuation of around 6 dB, whereas lead-lined glass has an attenuation between 3 and 30 dB.

2.7 Shadow Fading

In addition to path loss, a signal will typically experience random variation due to blockage from objects in the signal path, giving rise to a random variation about the path loss at a given distance. In addition, changes in reflecting surfaces and scattering objects can also cause random variation about the path loss. Thus, a model for the random attenuation due to these effects is also needed. Since 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, statistical models are widely used to characterize this attenuation. The most common model for this additional attenuation is log-normal shadowing. This model has been confirmed empirically to accurately model the variation in path loss or received power in both outdoor and indoor radio propagation environments (see e.g. [32, 59].)

In the log-normal shadowing model the path loss ψ is assumed random with a log-normal distribution given by

$$p(\psi) = \frac{\xi}{\sqrt{2\pi}\sigma_{\psi_{dB}}\psi} \exp\left[-\frac{(10\log_{10}\psi - \mu_{\psi_{dB}})^2}{2\sigma^2_{\psi_{dB}}}\right], \ \psi > 0, \tag{2.44}$$

where $\xi = 10/\ln 10$, $\mu_{\psi_{dB}}$ is the mean of $\psi_{\text{dB}} = 10\log_{10}\psi$ in dB and $\sigma_{\psi_{dB}}$ is the standard deviation of ψ_{dB}, also in dB. Note that if the path loss is log-normal, then the received power and receiver SNR will also be log-normal since these are just constant multiples of ψ. The mean of ψ (the linear average path loss) can be obtained from (2.44) as

$$\mu_\psi = E[\psi] = \exp\left[\frac{\mu_{\psi_{dB}}}{\xi} + \frac{\sigma^2_{\psi_{dB}}}{2\xi^2}\right]. \tag{2.45}$$

The conversion from the linear mean (in dB) to the log mean (in dB) is derived from (2.45) as

$$10\log_{10}\mu_\psi = \mu_{\psi_{dB}} + \frac{\sigma^2_{\psi_{dB}}}{2\xi}. \tag{2.46}$$

Performance in log-normal shadowing is typically parameterized by the log mean $\mu_{\psi_{dB}}$, which is refered to as the **average dB path loss** and is in units of dB. The linear mean path loss in dB, $10\log_{10}\mu_{\psi}$, is refered to as the **average path loss**.

With a change of variables we see that the distribution of the dB value of ψ is Gaussian with mean $\mu_{\psi_{dB}}$ and standard deviation $\sigma_{\psi_{dB}}$:

$$p(\psi_{\text{dB}}) = \frac{1}{\sqrt{2\pi}\sigma_{\psi_{dB}}}\exp\left[-\frac{(\psi_{\text{dB}}-\mu_{\psi_{dB}})^2}{2\sigma^2_{\psi_{dB}}}\right]. \tag{2.47}$$

The log-normal distribution is defined by two paramenters: $\mu_{\psi_{dB}}$ and $\sigma_{\psi_{dB}}$. Since blocking objects cause signal attenuation, $\mu_{\psi_{dB}}$ is always nonnegative. However, in some cases the average attenuation due to both path loss and shadowing is incorporated into the path loss model. For example, piecewise linear path loss models based on empirical data will incorporate the average shadowing associated with the measurements into the path loss model. In this case the shadowing model superimposed on the simplified path loss model should have $\mu_{\psi_{dB}} = 0$. However, if the path loss model does not incorporate average attenuation due to shadowing or if the shadowing model incorporates path loss via its mean, then $\mu_{\psi_{dB}}$ as well as $\sigma_{\psi_{dB}}$ will be positive, and must be obtained from an analytical model, simulation, or empirical measurements.

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 dB values of the empirical measurements. Specifically, given empirical (linear) path loss measurements $\{p_i\}_{i=1}^N$, should the mean path loss be determined as $\mu_{\psi} = \frac{1}{N}\sum_{i=1}^N p_i$ or as $\mu_{\psi_{dB}} = \frac{1}{N}\sum_{i=1}^N 10\log_{10} p_i$. 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 will see below, 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 smaller estimation error [61]. Finally, as we saw in Section 2.6.5, power falloff with distance models are often obtained by a piece-wise linear approximation to empirical measurements of dB power versus the log of distance [1].

Most empirical studies for outdoor channels support a standard deviation $\sigma_{\psi_{dB}}$ ranging from five to twelve dB in macrocells and four to thirteen dB in microcells [2, 15, 33, 55, 5]. The mean power $\mu_{\psi_{dB}}$ depends on the path loss and building properties in the area under consideration. The mean power $\mu_{\psi_{dB}}$ varies with distance due to path loss and the fact that average attenuation from objects increases with distance due 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

$$s(d) = ce^{-\alpha d}, \tag{2.48}$$

where c is an adjustment constant and α is an attenuation constant that depends on the object's materials and dielectric properties. If we assume that α is approximately equal for all blocking objects, and that the ith blocking object has depth d_i, then the attenuation of a signal as it propagates through this region is

$$s(d_t) = ce^{-\alpha\sum_i d_i} = ce^{-\alpha d_t}, \tag{2.49}$$

where $d_t = \sum_i d_i$ is the sum of the object depths through which the signal travels. If there are many objects between the transmitter and receiver, then we can approximate d_t by a Gaussian random variable.

Thus, $\log s(d_t) = \log c - \alpha d_t$ will have a Gaussian distribution with mean μ and standard deviation σ. The value of σ will depend on the environment and, as mentioned earlier, empirical measurements for σ range between four and twelve dB.

Example 2.4:

In Example 1.3 we found that the exponent for the simplified path loss model that best fit the measurements in Table 2.2 was $\gamma = 3.71$. Assuming the simplified path loss model with this exponent and the same $K = -31.54$ dB, find $\sigma^2_{\psi_{dB}}$, 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

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

where $M_{\text{measured}}(d_i)$ is the path loss measurement in Table 2.2 at distance d_i and $M_{\text{model}}(d_i) = K - 35.6\log_{10}(d)$. Thus

$$\begin{aligned}\sigma^2_{\psi_{dB}} &= \frac{1}{5}\Big[(-70 - 31.54 + 37.1)^2 + (-75 - 31.54 + 48.27)^2 + (-90 - 31.54 + 63.03)^2 + (-110 - 31.54 + 74.2)^2 \\ &+ (-125 - 31.54 + 91.90)^2\Big] \\ &= 13.29.\end{aligned}$$

Thus, the standard deviation of shadow fading on this path is $\sigma_{\psi_{dB}} = 3.65$ dB. Note that the bracketed term in the above expression equals the MMSE formula (2.31) from Example 1.3 with $\gamma = 3.71$.

Extensive measurements have been taken to characterize the empirical autocorrelation of shadowing for different environments at different frequencies, e.g. [55, 56, 60, 57, 58]. The most common analytical model for autocorrelation, first proposed by Gudmundson [55] based on empirical measurements, assumes the shadowing $\psi(d)$ is a first-order autoregressive process where the autocorrelation between shadow fading at two points separated by distance δ is given by

$$A(\delta) = \mathrm{E}[(\psi_{dB}(d) - \mu_{\psi_{dB}})(\psi_{dB}(d+\delta) - \mu_{\psi_{dB}})] = \sigma^2_{\psi_{dB}}\rho_D^{\delta/D}, \tag{2.50}$$

where ρ_D is the correlation between two points separated by a fixed distance D. This correlation must be obtained empirically, and varies with the propagation environment and carrier frequency. Measurements indicate that for suburban macrocells with $f_c = 900$ MHz, $\rho_D = .82$ for $D = 100$ m and for urban microcells with $f_c \approx 2$ GHz, $\rho_D = .3$ for $D = 10$ m [55, 57]. This model can be simplified and its empirical dependence removed by setting $\rho_D = 1/e$ for distance $D = X_c$, which yields

$$A(\delta) = \sigma^2_{\psi_{dB}} e^{-\delta/X_c}. \tag{2.51}$$

The **decorrelation distance** X_c in this model is the distance at which the signal autocorrelation 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 to 100 m [57, 60]. For users moving at velocity v, the shadowing decorrelation in time τ is obtained by substituting $v\tau = \delta$ in (2.50) or (2.51). Autocorrelation

relative to angular spread, which is useful for the multiple antenna systems treated in Chapter 10, has been investigated in [57, 56].

The first-order autoregressive correlation model (2.50) and its simplified form (2.51) are easy to analyze and to simulate. Specifically, one can simulate ψ_{dB} by first generating a white noise process and then passing it through a first order filter with a pole at $\rho_D^{-\delta/D}$ for autocorrelation (2.50) or at $e^{-\delta/X_c}$ for autocorrelation (2.51). The filter output will produce a shadowing random process with the desired autocorrelation properties [55, 5].

2.8 Combined Path Loss and Shadowing

Models for path loss and shadowing are typically superimposed to capture power falloff versus distance along with the random attenuation about this path loss from shadowing. In this combined model, average path loss ($\mu_{\psi_{dB}}$) is captured by the path loss model and shadow fading creates variations about this mean, as illustrated by the path loss and shadowing curve in Figure 2.1. Specifically, this curve plots the combination of the simplified path loss model (2.28) and the log-normal shadowing random process defined by (2.47) and (2.51). For this combined model the ratio of received to transmitted power in dB is given by:

$$\frac{P_r}{P_t}(dB) = 10\log_{10} K - 10\gamma \log_{10}\frac{d}{d_0} + \psi_{dB}, \tag{2.52}$$

where ψ_{dB} is a Gauss-distributed random variable with mean zero and variance $\sigma^2_{\psi_{dB}}$. In (2.52) and as shown in Figure 2.1, the path loss decreases linearly relative to $\log_{10} d$ 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.

The prior examples 2.3 and 2.4 illustrate the combined model for path loss and log-normal shadowing based on the measurements in Table 2.2, where path loss obeys the simplified path loss model with $K = -31.54$ dB and path loss exponent $\gamma = 3.71$ and shadowing obeys the log normal model with mean given by the path loss model and standard deviation $\sigma_{\psi_{dB}} = 3.65$ dB.

2.9 Outage Probability under Path Loss and Shadowing

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} below which performance becomes unacceptable (e.g. the voice quality in a cellular system is too poor to understand). 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** $p_{out}(P_{min}, d)$ under path loss and shadowing to be the probability that the received power at a given distance d, $P_r(d)$, falls below P_{min}: $p_{out}(P_{min}, d) = p(P_r(d) < P_{min})$. For the combined path loss and shadowing model of Section 2.8 this becomes

$$p(P_r(d) \le P_{min}) = 1 - Q\left(\frac{P_{min} - (P_t + 10\log_{10} K - 10\gamma \log_{10}(d/d_0))}{\sigma_{\psi_{dB}}}\right), \tag{2.53}$$

where the Q function is defined as the probability that a Gaussian random variable x with mean zero and variance one is bigger than z:

$$Q(x) \triangleq p(x > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \tag{2.54}$$

The conversion between the Q function and complementary error function is

$$Q(z) = \frac{1}{2} erfc\left(\frac{z}{\sqrt{2}}\right). \tag{2.55}$$

We will omit the parameters 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 combined path loss and shadowing models of Examples 2.3 and 2.4, assuming a transmit power of $P_t = 100$ mW and minimum power requirement $P_{min} = -110.5$ dBm.

Solution We have $P_t = 10$ mW $= 10$ dBm.

$$\begin{aligned} P_{out}(-110.5\text{dBm}, 150m) &= p(P_r(150m) < -110.5\text{dBm}) \\ &= 1 - Q\left(\frac{P_{min} - (P_t + 10\log_{10} K - 10\gamma \log_{10}(d/d_0))}{\sigma_{\psi_{dB}}}\right). \\ &= 1 - Q\left(\frac{-110 - (10 - 31.54 - 37.1\log_{10}[150])}{3.65}\right) \\ &= .049. \end{aligned}$$

An outage probabilities of 5% is a typical target in wireless system designs.

2.10 Cell Coverage Area

The **cell coverage area** in a cellular system is defined as the percentage of area within a cell that has received power above a given minimum. Consider a base station inside a circular cell of a given radius R. All mobiles within the cell require some minimum received SNR for acceptable performance. Assuming some reasonable noise and interference model, 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 $\overline{P}_R$, where the average is computed based on path loss alone. However, random 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 2.10, where we show contours of constant received power based on a fixed transmit power at the base station for path loss alone and for combined path loss and shadowing. For path loss alone constant power contours form a circle around the base station, since path loss is the same at a uniform distance from the base station. For combined path loss and shadowing the contours form an amoeba-like shape due to the random variations about the circular path loss contour caused by shadowing. The constant power contours for combined path loss and shadowing indicate the challenge 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, the base station must either transmit extra power to insure 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 not meet their minimum received power requirement. In fact, since the Gaussian distribution has infinite tails, there is a nonzero probability that *any* mobile within the cell will have a received power that falls

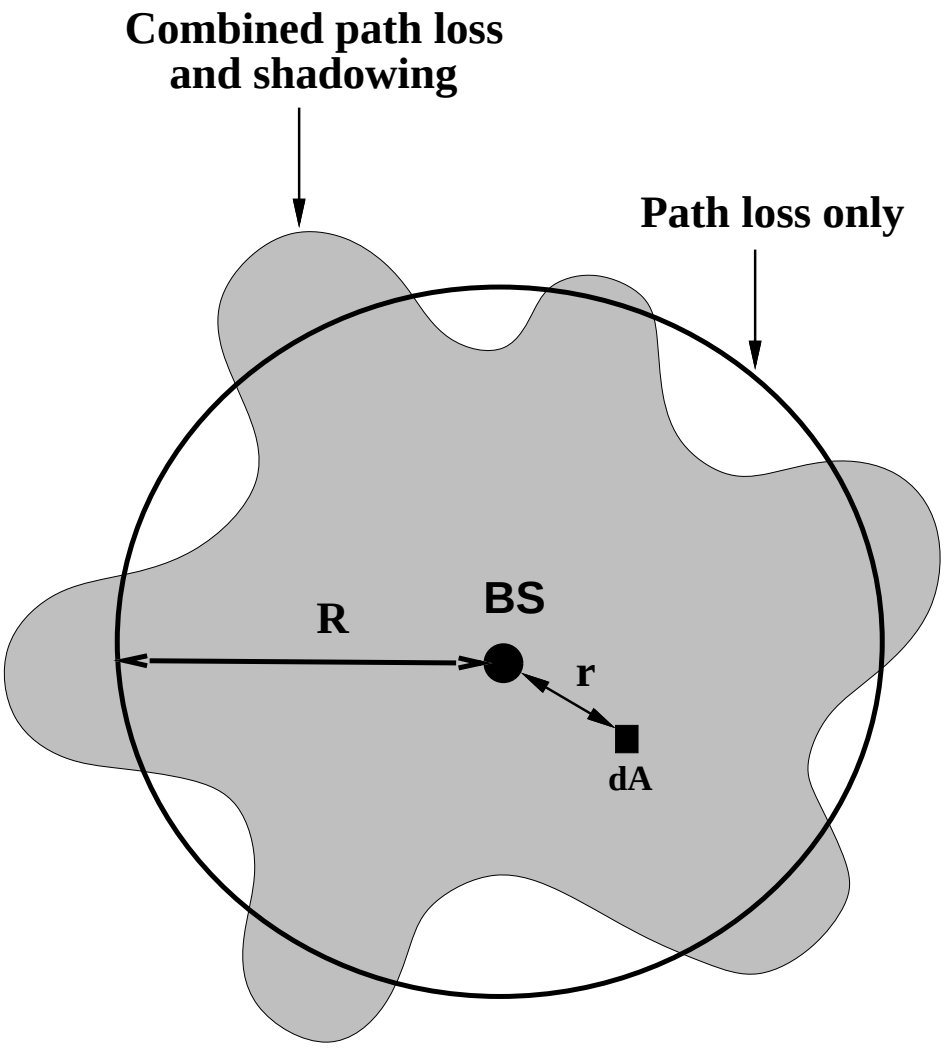


Figure 2.10: Contours of Constant Received Power.

below the minimum target, even if the mobile is close to the base station. This makes sense intuitively since a mobile may be in a tunnel or blocked by a large building, regardless of its proximity to the base station.

We now compute cell coverage area under path loss and shadowing. 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 (BS) in the cell, as shown in Figure 2.10. Let $P_r(r)$ be the received power in dA from combined path loss and shadowing, and let $P_A = p(P_r(r) > P_{min})$ in dA. Then the cell coverage area is given by

$$C = \frac{1}{\pi R^2} \int_{\text{cell area}} P_A dA = \frac{1}{\pi R^2} \int_0^{2\pi} \int_0^R P_A r dr d\theta. \tag{2.56}$$

The **outage probability of the cell** is defined as the percentage of area within the cell that does not meet its minimum power requirement P_{min}, i.e. $p_{out}^{cell} = 1 - C$.

Given the log-normal distribution for the shadowing,

$$p(P_R(r) \geq P_{min}) = Q\left(\frac{P_{min} - (P_t + 10\log_{10} K - 10\gamma \log_{10}(r/d_0))}{\sigma_{\psi_{dB}}}\right) = 1 - p_{out}(P_{min}, r). \tag{2.57}$$

where p_{out} is the outage probability defined in (2.53) with $d = r$. Locations within the cell with received power below P_{min} are said to be outage locations.

Combining (2.56) and (2.57) we get[4]

$$C = \frac{2}{R^2}\int_0^R rQ\left(a + b\ln\frac{r}{R}\right)dr, \tag{2.58}$$

where

$$a = \frac{P_{min} - \overline{P}_r(R)}{\sigma_{\psi_{dB}}} \quad b = \frac{10\gamma\log_{10}(e)}{\sigma_{\psi_{dB}}}, \tag{2.59}$$

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. This integral yields a closed-form solution for C in terms of a and b:

$$C = Q(a) + \exp\left(\frac{2-2ab}{b^2}\right)Q\left(\frac{2-ab}{b}\right). \tag{2.60}$$

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 coverage area simplifies to

$$C = \frac{1}{2} + \exp\left(\frac{2}{b^2}\right)Q\left(\frac{2}{b}\right). \tag{2.61}$$

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

Example 2.6:

Find the coverage area for a cell with the combined path loss and shadowing models of Examples 1.3 and 1.4, a cell radius of 200 m, a base station transmit power of $P_t = 100$ mW, and a minimum received power requirement of $P_{min} = -110$ dBm and of $P_{min} = -120$ dBm.

Solution We first consider $P_{min} = -110$ and check if $a = 0$ to determine whether to use the full formula (2.60) or the simplified formula (2.61). We have $\overline{P}_r(R) = P_t + K - 10\gamma\log_{10}(150) = 2 - 31.54 - 37.1\log_{10}[120] = -114.9$dBm $\neq -110$ dBm, so we use (2.60). Evaluating a and b from (2.59) yields $a = (-110 + 114.9)/3.65 = 1.34$ and $b = 37.1 * .434/3.65 = 4.41$. Substituting these into (2.60) yields

$$C = Q(1.34) + \exp\left(\frac{2 - 2(1.34 * 4.41)}{4.41^2}\right) = .58,$$

which would be a very low coverage value for an operational cellular system (lots of unhappy customers). Now considering the less stringent received power requirement $P_{min} = -120$ dBm yields $a = (-120 +$

[4]Recall that (2.57) is generally only valid 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 coverage area, since d_0 is typically very small compared to R and the outage probability for $r < d_0$ is negligible.

114.9)/3.65 = −1.397 and the same $b = 4.41$. Substituting these values into (2.60) yields $C = .986$, a much more acceptable value for coverage area.

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

Solution: For $P_{min} = \overline{P}_r(R)$ we have $a = 0$ so coverage is given by the formula (2.61). The coverage area thus depends only on the value for $b = 10\gamma \log_{10}[e]/\sigma_{\psi_{dB}}$, which in turn depends only on the ratio $\gamma/\sigma_{\psi_{dB}}$. The following table contains coverage area evaluated from (2.61) for the different γ and $\sigma_{\psi_{dB}}$ values.

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

Table 2.4: Coverage Area for Different γ and $\sigma_{\psi_{dB}}$

Not surprisingly, for fixed γ the coverage area increases as $\sigma_{\psi_{dB}}$ decreases: that is because a smaller $\sigma_{\psi_{dB}}$ means less variation about the mean path loss, and since with no shadowing we have 100% coverage (since $P_{min} = \overline{P}_r(R)$), we expect that as $\sigma_{\psi_{dB}}$ decreases to zero, coverage area increases to 100%. It is a bit more puzzling that for a fixed $\sigma_{\psi_{dB}}$ coverage area increases as γ increases, since a larger γ implies that received signal power falls off more quickly. But recall that we have set $P_{min} = \overline{P}_r(R)$, 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 coverage area increases with path loss exponent under this assumption is that, as γ increases, the transmit power must increase to satisfy $P_{min} = \overline{P}_r(R)$. This results in higher average power throughout the cell, resulting in a higher coverage area.

Bibliography

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

[2] W.C. Jakes, Jr., *Microwave Mobile Communications.* New York: Wiley, 1974. Reprinted by IEEE Press.

[3] D. Parsons, *The Mobile Radio Propagation Channel.* New York: Halsted Press (Division of Wiley). 1992.

[4] M. Pätzold, *Mobile Fading Channels.* New York: Wiley. 2002.

[5] G. Stuber, *Principles of Mobile Communications,* 2nd Ed., Boston: Kluwer Academic Press.

[6] J.W. McKown and R.L. Hamilton, Jr., "Ray tracing as a design tool for radio networks," *IEEE Network* , Vol. 5, No. 6, pp. 27–30, Nov. 1991.

[7] N. Amitay, "Modeling and computer simulation of wave propagation in lineal line-of-sight microcells," *IEEE Trans. Vehic. Technol.*, Vol VT-41, No. 4, pp. 337–342, Nov. 1992.

[8] , K. A. Remley, H. R. Anderson, and A. Weisshar, "Improving the accuracy of ray-tracing techniques for indoor propagation modeling," *IEEE Trans. Vehic. Technol.*, pp. 2350–2358, Nov. 2000.

[9] "Concepts and results for 3D digital terrain-based wave propagation models: an overview," *IEEE J. Select. Areas Commun.* pp. 1002–1012, Sept. 1993.

[10] "Applicability of ray-tracing techniques for prediction of outdoor channel characteristics," *IEEE Trans. Vehic. Technol.*, pp. 2336–2349, Nov. 2000.

[11] 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 Vehic. Technol. Conf.*, pp. 932–935, May 1992.

[12] A. Domazetovic, L.J. Greenstein, N. Mandayan, and I. Seskar, "A new modeling approach for wireless channels with predictable path geometries," *Proc. IEEE Vehic. Technol. Conf,* Sept. 2002.

[13] 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. Vehic. Technol. Conf.*, Vol VT-40, No. 1, pp. 203–210, Feb. 1991.

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

[15] J.-E. Berg, R. Bownds, and F. Lotse, "Path loss and fading models for microcells at 900 MHz," *Vehic. Technol. Conf. Rec.*, pp. 666–671, May 1992.

[16] E. McCune and K. Feher, "Closed-form propagation model combining one or more propagation constant segments," *Proc. IEEE Vehic. Technol. Conf.*, pp. 1108–1112, May 1997.

[17] S. Y Seidel and T. S. Rappaport, "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings," *IEEE Transactions on Antennas and Propagation*", pp. 207–217, Feb. 1992.

[18] S. Y. Seidel, T. S. Rappaport, M.J. Feuerstein, K.L. Blackard, L. Grindstaff, "The impact of surrounding buildings on propagation for wireless in-building personal communications system design," *Proceedings: IEEE Vehicular Technology Conference,* pp. 814–818, May 1992.

[19] A.J. Motley and J.M.P. Keenan, "Personal communication radio coverage in buildings at 900 MHz and 1700 MHz," *Electronic Letters*, pp. 763–764, June 1988.

[20] F.C. Owen and C.D. Pudney, "Radio propagation for digital cordless telephones at 1700 MHz and 900 MHz," *Electronic Letters,* pp. 52-53, Sept. 1988.

[21] 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," *Proceedings: IEEE Vehicular Technology Conference,* pp. 24-28, Sept. 2002.

[22] L.-S. Poon and H.-S. Wang, "Propagation characteristic measurement and frequency reuse planning in an office building," *Proceedings: IEEE Vehicular Technology Conference,* pp. 1807–1810, June 1994.

[23] G. Durgin, T.S. Rappaport, and H. Xu, "Partition-based path loss analysis for in-home and residential areas at 5.85 GHz," *Proceedings: IEEE Globecom Conference,* pp. 904–909, Nov. 1998.

[24] A. F. Toledo and A.M.D. Turkmani, "Propagation into and within buildings at 900, 1800, and 2300 MHz," *Proc. IEEE Vehicular Technology Conference,* pp. 633-636, May 1992.

[25] 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 Personal Communications Magazine,* pp. 40–47, April 1998.

[26] R. Hoppe, G. Wölfle, and F.M. Landstorfer, "Measurement of building penetration loss and propagation models for radio transmission into buildings," *Proc. IEEE Vehicular Technology Conference,* pp. 2298–2302, April 1999.

[27] E.H. Walker, "Penetration of radio signals into buildings in cellular radio environments," *Bell Systems Technical Journal*, Sept. 1983.

[28] W.C.Y. Lee, *Mobile Communication Design Fundamentals,* Indianapolis, IN: Sams, 1986.

[29] D.M.J. Devasirvathan, R.R. Murray, and D.R. Woiter, "Time delay spread measurements in a wireless local loop test bed," *Proceedings: IEEE Vehicular Technology Conference*," pp. 241–245, May 1995.

[30] 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 Transactions on Vehicular Technology,* pp. 487–498, Aug. 1994.

[31] S.T.S. Chia, "1.7 GHz propagation measurement for highway microcells," *Electronic Letters*, pp. 1279–1280, Aug. 1990.

[32] 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 Journal on Selected Areas in Communications*, pp. 1205–1211, July 1999.

[33] A.J. Goldsmith and L.J. Greenstein, "A measurement-based model for predicting coverage areas of urban microcells," *IEEE J. Selected Areas Commun.*, Vol. SAC-11, No. 7, pp. 1013–1023, Sept. 1993.

[34] F. Ikegami, S. Takeuchi, and S. Yoshida, "Theoretical prediction of mean field strength for urban mobile radio," *IEEE Trans. Antennas Propagat.*, Vol. AP-39, No. 3, pp. 299–302, March 1991.

[35] M.C. Lawton and J.P. McGeehan, "The application of GTD and ray launching techniques to channel modeling for cordless radio systems," *Vehic. Technol. Conf. Rec.*, pp. 125–130, May 1992.

[36] R.J. Luebbers, "Finite conductivity uniform GTD versus knife edge diffraction in prediction of propagation path loss," *IEEE Trans. Antennas Propagat.*, Vol. AP-32, No. 1, pp. 70–76, Jan. 1984.

[37] C. Bergljung and L.G. Olsson, "Rigorous diffraction theory applied to street microcell propagation," *Globecom Conf. Rec.*, pp. 1292–1296, Dec. 1991.

[38] J.B Keller, "Geometrical theory of diffraction," *J. Opt. Soc. Amer.,* pp. 116-130, 1962.

[39] R.G. Kouyoumjian and P.H. Pathak, "A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface," *Proc. IEEE*, pp. 1448–1461, Nov. 1974.

[40] G.K. Chan, "Propagation and coverage prediction for cellular radio systems," *IEEE Trans. Vehic. Technol.*, Vol VT-40, No. 4, pp. 665–670, Nov. 1991.

[41] K.C. Chamberlin and R.J. Luebbers, "An evaluation of Longley-Rice and GTD propagation models," *IEEE Trans. Antennas Propagat.*, vol AP-30, No. 11, pp. 1093–1098, Nov. 1982.

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

[43] 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. Vehic. Technol.*, Vol VT-40, No. 4, pp. 721–730, Nov. 1991.

[44] S.T.S. Chia, "1700 MHz urban microcells and their coverage into buildings," *IEE Antennas Propagat. Conf. Rec.*, pp. 504–511, York, U.K., April 1991.

[45] D. Akerberg, "Properties of a TDMA Picocellular Office Communication System," *Proc: IEEE Globecom*, pp. 1343–1349, Dec. 1988.

[46] P. Harley, "Short distance attenuation measurements at 900 MHz and 1.8 GHz using low antenna heights for microcells," *IEEE J. Selected Areas Commun.*, Vol. SAC-7, No. 1, pp. 5–11, Jan. 1989.

[47] J.-F. Wagen, "Signal strength measurements at 881 MHz for urban microcells in downtown Tampa," *Globecom Conf. Rec.*, pp. 1313–1317, Dec. 1991.

[48] R.J.C. Bultitude and G.K. Bedal, "Propagation characteristics on microcellular urban mobile radio channels at 910 MHz," *IEEE J. Selected Areas Commun.*, Vol. SAC-7, No. 1, pp. 31–39, Jan. 1989.

[49] J.H. Whitteker, "Measurements of path loss at 910 MHz for proposed microcell urban mobile systems," *IEEE Trans. Vehic. Technol.*, Vol VT-37, No. 6, pp. 125–129, Aug. 1988.

[50] H. Börjeson, C. Bergljung, and L.G. Olsson, "Outdoor microcell measurements at 1700 MHz.," *Vehic. Technol. Conf. Rec.*, pp. 927–931, May 1992.

[51] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Vehic. Technol.*, Vol VT-29, No. 3, pp. 317–325, Aug. 1980.

[52] T. Okumura, E. Ohmori, and K. Fukuda, "Field strength and its variability in VHF and UHF land mobile service," *Review Electrical Communication Laboratory,* Vol. 16, No. 9-10, pp. 825–873, Sept.-Oct. 1968.

[53] 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," Revision 2, The Hague, Sept. 1991.

[54] J. Walfisch and H.L. Bertoni, "A theoretical model of UHF propagation in urban environments," *IEEE Trans. Antennas and Propagation*, pp. 1788-1796, Oct. 1988.

[55] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electr. Lttrs.*, Vol. 27, pp. 2145–2146, Nov. 7, 1991.

[56] 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 Journal Selected Areas Communications*, pp. 523–531, April 2002.

[57] J. Weitzen and T. Lowe, "Measurement of angular and distance correlation properties of log-normal shadowing at 1900 MHz and its application to design of PCS systems," *IEEE Transactions on Vehicular Technology*, pp. 265–273, March 2002.

[58] W. Turin, R. Jana, S.S Ghassemzadeh, V. W. Rice, V. Tarokh, "Autoregressive modeling of an indoor UWB channel," *Proc. IEEE Conf. UWB Syst. Technol.*, pp. 71–74, May 2002.

[59] S.S. Ghassemzadeh, L.J. Greenstein, A. Kavcic, T. Sveinsson, V. Tarokh, "Indoor path loss model for residential and commercial buildings," *Proc. Vehic. Technol. Conf.*, pp. 3115–3119, Oct. 2003.

[60] M. Marsan and G.C. Hess, "Shadow variability in an urban land mobile radio environment," *Electronics Letters*, pp. 646–648, May 1990.

[61] 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 Transactions on Vehicular Technology*, Vol. 43, No. 3, pp. 439-446, Aug. 1994.

[62] *IEEE Journal Select. Areas Commun.* Special Issue on Channel and Propagation Modeling for Wireless Systems Design, April 2002 and Aug. 2002.

[63] *IEEE Journal Select. Areas Commun.* Special Issue on Ultra-Wideband radio in multiaccess wireless communications, Dec. 2002.

Chapter 2 Problems

1. Under a 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 ($G_l = 1$) and a carrier frequency $f = 5$ GHz, assuming a distance $d = 10m$. Repeat for $d = 100$m.

2. For a two-path propagation model with transmitter-receiver separation $d = 100$ m, $h_t = 10$ m, and $h_r = 2$ m, find the delay spread between the two signals.

3. For the two ray model, show how a Taylor series approximation applied to (2.13) results in the approximation
$$\Delta\phi = \frac{2\pi(r' + r - l)}{\lambda} \approx \frac{4\pi h_t h_r}{\lambda d}.$$

4. For the two-ray path loss 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-path model for a large macrocell in a suburban area with the base station mounted on a tower or building ($h_t = 20$m), the receivers at height $h_r = 3$m, and $f_c = 2GHz$. 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-path 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-path model with a LOS component and ground reflection?

7. Consider a two-path channel with impulse response $h(t) = \alpha_1\delta(\tau) + \alpha_2\delta(\tau - .022\mu\text{sec})$. 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 meters above the ground and 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 the ground wave cancellation, as will be illustrated in this problem. Plot the dB power ($10\log_{10} P_r$) versus log distance ($\log_{10} d$) for the two-ray model with the parameters $f = 900$MHz, R=-1, $h_t = 50$m, $h_r = 2$m, $G_l = 1$, and the following values for G_r: $G_r = 1, .316, .1$, and .01 (i.e. $G_r = 0, -5, -10$, and -20 dB, respectively). Each of the 4 plots should range in distance from $d = 1$m to $d = 100,000$m. 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 0dB. 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 with slope 0, 2, and 4, respectively). Note that at large distances it becomes increasingly difficult to have $G_r << G_l$ since it requires extremely precise angular directivity in the antennas.

9. What average power falloff with distance would you expect for the 10-ray model and why?

10. For the 10-ray model, assume the transmitter and receiver are in the middle of a street of width 20 m and are at the same height. The transmitter-receiver separation is 500 m. Find the delay spread for this model.

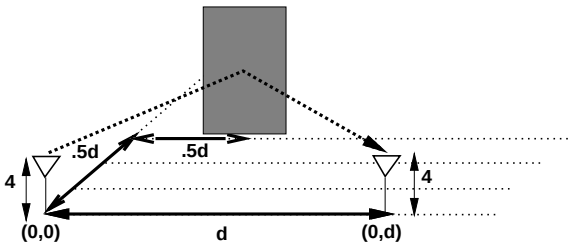


Figure 2.11: System with Scattering

11. Consider a system with a transmitter, receiver, and scatterer as shown in Figure 2.11. 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, i.e. on such a grid the transmitter is located at $(0, 0)$, the receiver is located at $(0, d)$ and the scatterer is located at $(.5d, .5d)$. Assume a radar cross section of 20 dBm^2. 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 just reflected with reflection coefficient $R = -1$.

12. Under what conditions is the simplified path loss model (2.28) the same as the free space path loss model (2.7).

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

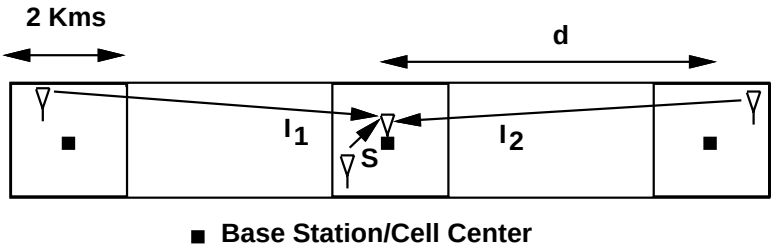
14. This problem shows how different propagation models can lead to very different SNRs (and therefore different link performance) for a given system design. Consider a linear cellular system using frequency division, as might operate along a highway or rural road, as shown in the figure below. Each 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 which are two kilometers 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 power (S/I) from mobiles to the base station) is greater than 20 dB. You can ignore all interferers except from the two nearest cells operating at the same frequency.

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

    (b) Propagation for both signal and interference follow the simplified path loss model (2.28) with $d_0 = 100$m, $K = 1$, and $\gamma = 3$.

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

15. 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 4 environments.



16. (Computer plots) Find parameters for a piecewise linear model with three segments to approximate the two-path model path loss (2.4) over distances between 10 and 1000 meters, assuming $h_t = 10$m and $h_r = 2$ m. Plot the path loss and the piecewise linear approximation using these parameters over this distance range.

17. 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 3 floors with attenuation 15 dB, 10 dB, and 6 dB, respectively, as well as 2 double plasterboard walls. Assume a reference distance $d_0 = 1$ and constant $K = 0$ dB.

18. The following table lists a set of empirical path loss measurements.

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

    (a) Find the parameters of a simplified path loss model plus log normal shadowing that best fit this data.

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

    (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 nonoutage.

19. Consider a cellular system where propagation follows free space path loss plus log normal shadowing with $\sigma = 6$ 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 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 -10 dBm. Find the maximum cell size so that a mobile on the cell boundary will have acceptable voice quality 90% of the time.

20. In this problem we will simulate the log normal fading process over distance based on the autocorrelation model (2.51). 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 $X_c = 20$ m and plot

the resulting log normal fading process over a distance d ranging from 0 to 200 m, sampling the process every meter. You should normalize your plot about 0 dB, since the mean of the log normal shadowing is captured by path loss.

21. In this problem we will explore the impact of different log-normal shadowing parameters on outage probability. Consider a a cellular system where the received signal power is distributed according to a log-normal distribution with mean μ dB and standard deviation σ_ψ dB. Assume the received signal power must be above 10 dBm for acceptable performance.

    (a) What is the outage probability when the log-normal distribution has $\mu_\psi = 15$dB and $\sigma_\psi = 8$dB?

    (b) For $\sigma_\psi = 4$dB, what value of μ_ψ is required such that the outage probability is less than 1%, a typical value for high-quality PCS systems?

    (c) Repeat (b) for $\sigma_\psi = 12$dB.

    (d) One proposed technique to reduce 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, which is typically the case for CDMA systems. Explain why this might reduce outage probability.

22. Find the coverage area for a microcellular system where path loss follows the simplied 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.

23. Consider a cellular system where path loss follows the simplied model with $\gamma = 6$, and 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 nonoutage, find the cell coverage area.

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

Chapter 3

Statistical Multipath Channel Models

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

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

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

3.1 Time-Varying Channel Impulse Response

Let the transmitted signal be as in Chapter 2:

$$s(t) = \Re\left\{u(t)e^{j(2\pi f_c t+\phi_0)}\right\} = \Re\{u(t)\}\cos(2\pi f_c t + \phi_0) - \Im\{u(t)\}\sin(2\pi f_c t + \phi_0), \tag{3.1}$$

where $u(t)$ is the complex envelope of $s(t)$ with bandwidth B_u, f_c is its carrier frequency, and ϕ_0 is the oscillator phase offset. The corresponding received signal is the sum of the line-of-sight (LOS) path and all resolvable multipath components:

$$r(t) = \Re\left\{\sum_{n=0}^{N(t)} \alpha_n(t)u(t-\tau_n(t))e^{j2\pi[f_c(t-\tau_n(t))+\phi_{D_n}+\phi_0]}\right\}, \tag{3.2}$$

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

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

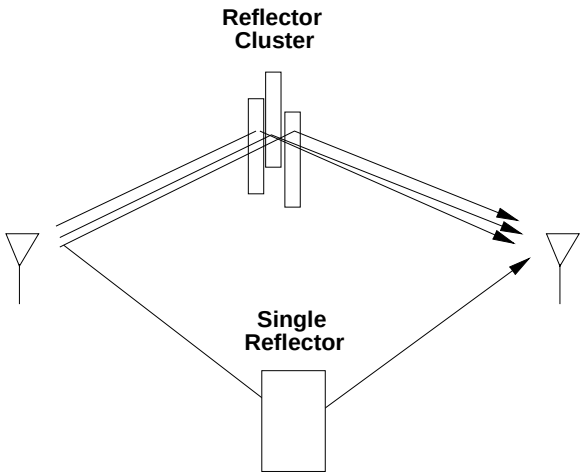


Figure 3.1: A Single Reflector and A Reflector Cluster.

Since the parameters $\alpha_n(t)$, $\tau_n(t)$, and $\phi_{D_n}(t)$) associated with each resolvable multipath component

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

change over time, they are characterized as random processes which we assume to be both stationary and ergodic. Thus, the received signal is also a stationary and ergodic random process. For wideband channels, where each term in (3.2) corresponds to a single reflector, these parameters change slowly as the propagation environment changes. For narrowband channels, where each term in (3.2) results from the sum of nonresolvable multipath components, the parameters can change quickly, on the order of a signal wavelength, due to constructive and destructive addition of the different components.

We can simplify $r(t)$ by letting

$$\phi_n(t) = 2\pi f_c \tau_n(t) - \phi_{D_n} - \phi_0. \tag{3.3}$$

We can then rewrite the received signal as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and is also shown in Figure 3.3.

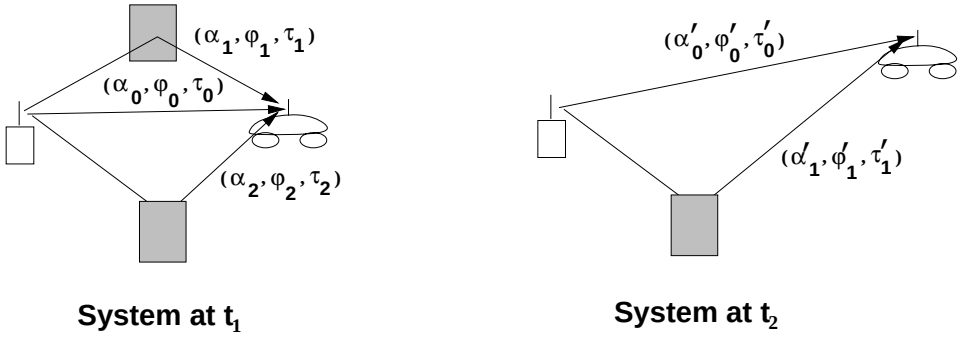


Figure 3.2: System Multipath at Two Different Measurement Times.

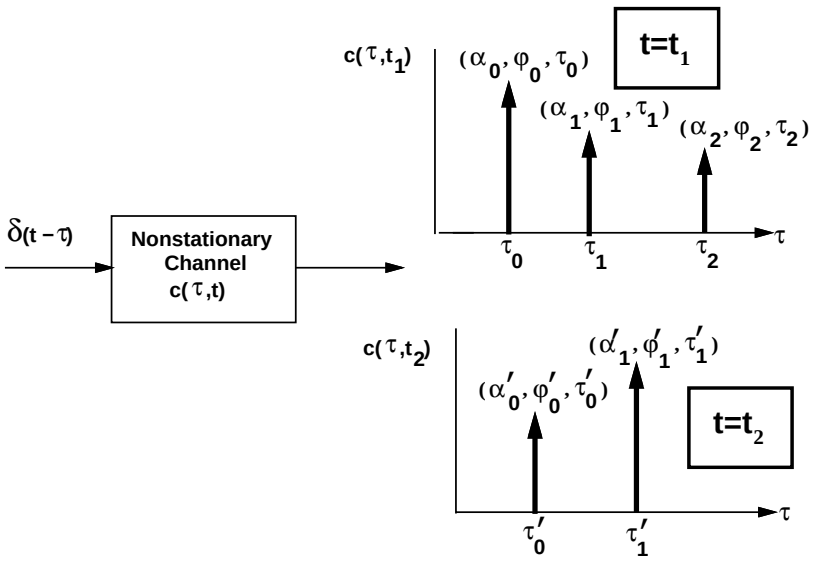


Figure 3.3: Response of Nonstationary Channel.

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

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

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

Example 3.1: Consider a wireless LAN operating in a factory near a conveyor belt. The transmitter and receiver have a LOS path between them with gain α_0, phase ϕ_0 and delay τ_0. Every T_0 seconds a metal item comes down the conveyor belt, creating an additional reflected signal path in addition to the LOS path with gain α_1, phase ϕ_1 and delay τ_1. Find the time-varying impulse response $c(\tau, t)$ of this channel.
Solution: For $t \neq nT_0$, $n = 1, 2, \ldots$ the channel impulse response corresponds to just the LOS path. For $t = nT_0$ the channel impulse response has both the LOS and reflected paths. Thus, $c(\tau, t)$ is given by

$$c(\tau, t) = \begin{cases} \alpha_0 e^{j\phi_0} \delta(\tau - \tau_0) & t \neq nT_0 \\ \alpha_0 e^{j\phi_0} \delta(\tau - \tau_0) + \alpha_1 e^{j\phi_1} \delta(\tau - \tau_1) & t = nT_0 \end{cases} \tag{3.11}$$

Note that for typical carrier frequencies, the nth multipath component will have $f_c\tau_n(t) >> 1$. For example, with $f_c = 1$ GHz and $\tau_n = 50$ ns (a typical value for an indoor system), $f_c\tau_n = 50 >> 1$. Outdoor wireless systems have multipath delays much greater than 50 ns, so this property also holds for these systems. If $f_c\tau_n(t) >> 1$ then a small change in the path delay $\tau_n(t)$ can lead to a very large phase change in the nth multipath component with phase $\phi_n(t) = 2\pi(f_c\tau_n(t) - \phi_{D_n} - \phi_0)$. Rapid phase

changes in each multipath component gives rise to constructive and destructive addition of the multipath components comprising the received signal, which in turn causes rapid variation in the received signal strength. This phenomenon, called *fading*, will be discussed in more detail in subsequent sections.

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

3.2 Narrowband fading models

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

$$r(t) = \Re\left\{u(t)e^{j2\pi f_c t}\left(\sum_n \alpha_n(t)e^{-j\phi_n(t)}\right)\right\}. \tag{3.12}$$

Equation (3.12) differs from the original transmitted signal by the complex scale factor in parentheses. This scale factor is independent of the transmitted signal $s(t)$ or, equivalently, the baseband signal $u(t)$, as long as the narrowband assumption $T_m << 1/B$ is satisfied. In order to characterize the random scale

factor caused by the multipath we choose $s(t)$ to be an unmodulated carrier:

$$s(t) = \Re\{e^{j2\pi f_c t}\} = \cos 2\pi f_c t, \tag{3.13}$$

which is narrowband for any T_m.

With this assumption the received signal becomes

$$r(t) = \Re\left\{\left[\sum_{n=0}^{N(t)} \alpha_n(t) e^{-j\phi_n(t)}\right] e^{j2\pi f_c t}\right\} = r_I(t)\cos 2\pi f_c t + r_Q(t)\sin 2\pi f_c t, \tag{3.14}$$

where the in-phase and quadrature components are given by

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

and

$$r_Q(t) = \sum_{n=1}^{N(t)} \alpha_n(t)\sin\phi_n(t). \tag{3.16}$$

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

3.2.1 Autocorrelation, Cross Correlation, and Power Spectral Density

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

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

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

where the second equality follows from the independence of α_n and ϕ_n and the last equality follows from the uniform distribution on ϕ_n. Similarly we can show that $\mathrm{E}[r_Q(t)] = 0$. Thus, the received signal also has $\mathrm{E}[r(t)] = 0$, i.e. it is a zero-mean Gaussian process. When there is a dominant LOS component in

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

the channel the phase of the received signal is dominated by the phase of the LOS component, which can be determined at the receiver, so the assumption of a random uniform phase no longer holds.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

component has the same received power, so $\mathrm{E}[\alpha_n^2] = P/N$, where P is the total received power. Then (3.21) becomes

$$A_{r_I}(\tau) = \frac{.5P}{N} \sum_{n=1}^{N} \cos(2\pi v\tau \cos n\Delta\theta/\lambda). \tag{3.24}$$

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

$$A_{r_I}(\tau) = \frac{.5P}{2\pi} \sum_{n=1}^{N} \cos(2\pi v\tau \cos n\Delta\theta/\lambda)\Delta\theta. \tag{3.25}$$

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

$$A_{r_I}(\tau) = \frac{.5P}{2\pi} \int \cos(2\pi v\tau \cos\theta/\lambda)d\theta = .5PJ_0(2\pi f_D\tau), \tag{3.26}$$

where

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

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

$$A_{r_I,r_Q}(\tau) = \frac{.5P}{2\pi} \int \sin(2\pi v\tau \cos\theta/\lambda)d\theta = 0. \tag{3.27}$$

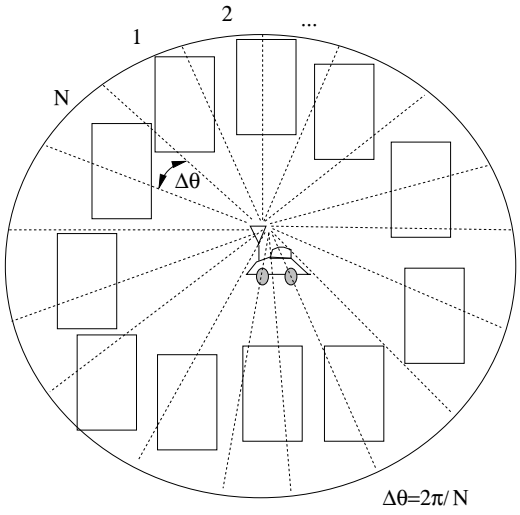


Figure 3.4: Dense Scattering Environment

A plot of $J_0(2\pi f_D\tau)$ is shown in Figure 3.5. There are several interesting observations from this plot. First we see that the autocorrelation is zero for $f_D\tau \approx .4$ or, equivalently, for $v\tau \approx .4\lambda$. Thus, the signal decorrelates over a distance of approximately one half wavelength, under the uniform θ_n assumption. This approximation is commonly used as a rule of thumb to determine many system parameters of interest.

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

For example, we will see in Chapter 7 that obtaining independent fading paths can be exploited by antenna diversity to remove some of the negative effects of fading. The antenna spacing must be such that each antenna receives an independent fading path and therefore, based on our analysis here, an antenna spacing of $.4\lambda$ should be used. Another interesting characteristic of this plot is that the signal recorrelates after it becomes uncorrelated. Thus, we cannot assume that the signal remains independent from its initial value at $d = 0$ for separation distances greater than $.4\lambda$. As a result, a Markov model is not completely accurate for Rayleigh fading, because of this recorrelation property. However, in many system analyses a correlation below .5 does not significantly degrade performance relative to uncorrelated fading [8, Chapter 9.6.5]. For such studies the fading process can be modeled as Markov by assuming that once the correlation function falls below .5, i.e. the separation distance is greater than a half wavelength, the signal has become decorrelated.

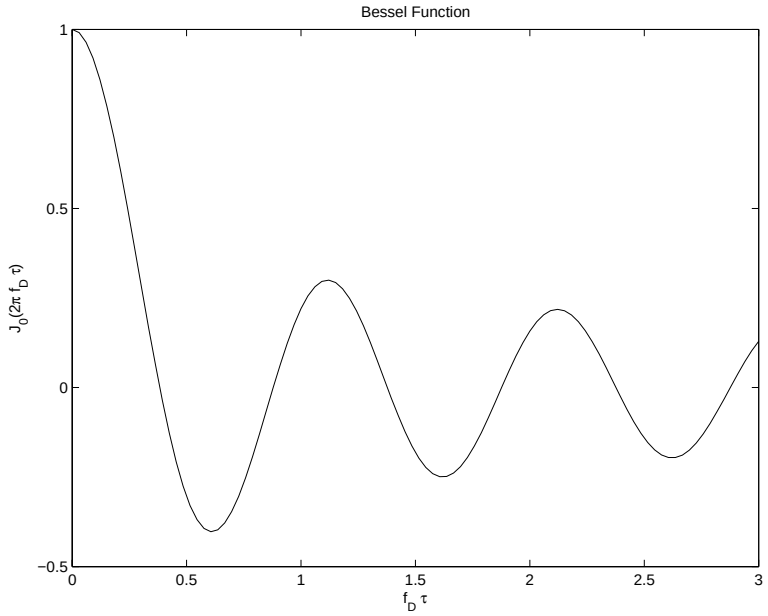


Figure 3.5: Bessel Function versus $f_d\tau$

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

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

This PSD is shown in Figure 3.6.

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

$$S_r(f) = \mathcal{F}[A_r(\tau)] = .5[S_{r_I}(f-f_c) + S_{r_I}(f+f_c)] = \begin{cases} \frac{P_r}{4\pi f_D} \frac{1}{\sqrt{1-\left(\frac{f-f_c}{f_D}\right)^2}} & |f-f_c| \le f_D \\ 0 & \text{else} \end{cases}, \tag{3.29}$$

Note that this PSD integrates to one. Since the PSD models the power density associated with multipath components as a function of their Doppler frequency, it can be viewed as the distribution (pdf) of the

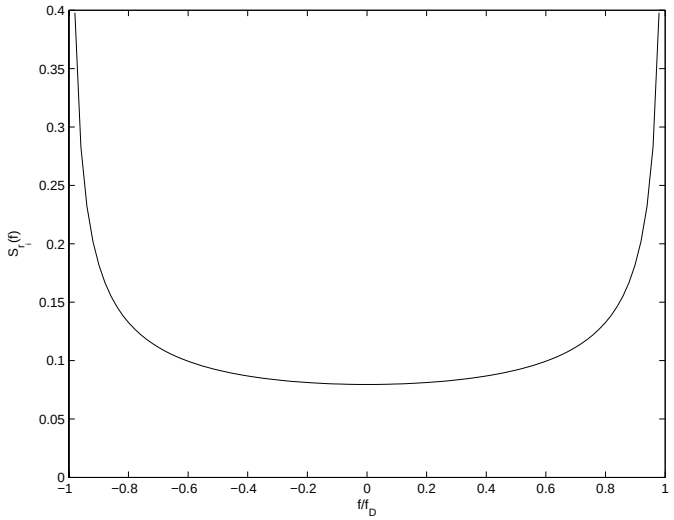


Figure 3.6: In-Phase and Quadrature PSD: $S_{r_I}(f) = S_{r_Q}(f)$

random frequency due to Doppler associated with multipath. We see from (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 similar environments 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 the PSD corresponds to the pdf of the random Doppler frequency $f_D(\theta)$. To see this, on the left in Figure 3.7 we plot $f_D(\theta) = f_D \cos(\theta), f_D = 2\pi v/\lambda$ along with a dotted line straight-line segment approximation $\underline{f}_D(\theta)$ to $f_D(\theta)$. On the right in this figure we plot PSD $S_{r_i}(f)$ along with a dotted line straight line segment approximation to it $\underline{S}_{r_i}(f)$. We see that $\cos(\theta) \approx \pm 1$ for a relatively large range of θ values. Thus, multipath components with angles of arrival in this range of values have Doppler frequency $f_D(\theta) \approx \pm f_D$, so the power associated with all of these multipath components will add together in the PSD at $f \approx f_D$. This is shown in our approximation by the fact that the segments where $\underline{f}_D(\theta) = \pm f_D$ on the left lead to delta functions at $\pm f_D$ in the pdf approximation $\underline{S}_{r_i}(f)$ on the right. The segments where $\underline{f}_D(\theta)$ has uniform slope on the left lead to the flat part of $\underline{S}_{r_i}(f)$ on the right, since there is one multipath component contributing power at each angular increment. Formulas for the autocorrelation and PSD in nonuniform scattering, corresponding to more typical microcell and indoor environments, can be found in [5, Chapter 1], [11, Chapter 2].

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

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

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

We have now completed our model for the three characteristics of power versus distance exhibited

Figure 3.7: Cosine and PSD Approximation by Straight Line Segments

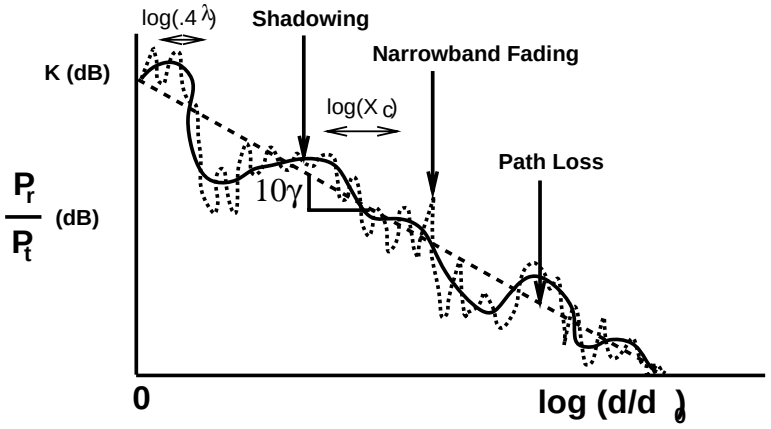


Figure 3.8: Combined Path Loss, Shadowing, and Narrowband Fading.

in narrowband wireless channels. These characteristics are illustrated in Figure 3.8, adding narrowband fading to the path loss and shadowing models developed in Chapter 2. In this figure we see the decrease in signal power due to path loss decreasing as d^{γ} with γ the path loss exponent, the more rapid variations due to shadowing which change on the order of the decorrelation distance X_c, and the very rapid variations due to multipath fading which change on the order of half the signal wavelength. If we blow up a small segment of this figure over distances where path loss and shadowing are constant we obtain Figure 3.9, where we show dB fluctuation in received power versus linear distance $d = vt$ (not log distance). In this figure the average received power P_r is normalized to 0 dBm. A mobile receiver traveling at fixed velocity v would experience the received power variations over time illustrated in this figure.

3.2.2 Envelope and Power Distributions

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

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

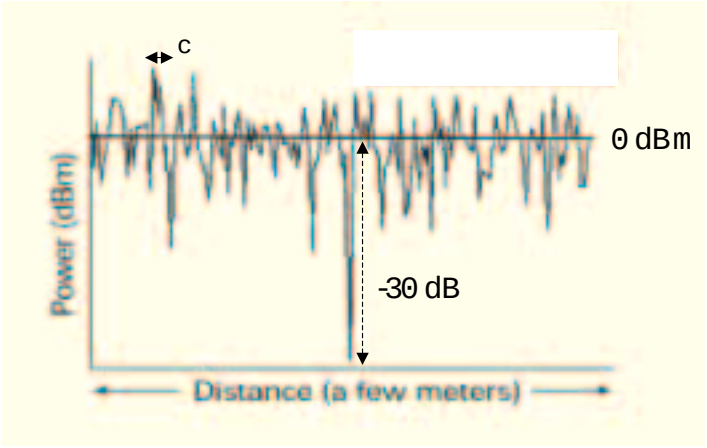


Figure 3.9: Narrowband Fading.

is Rayleigh-distributed with distribution

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2.3 Level Crossing Rate and Average Fade Duration

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

Consider the fading process shown in Figure 3.10. The expected amount of time the signal envelope spends in the interval $(Z, Z + dz)$ with envelope slope in the range $[\dot{z}, \dot{z} + d\dot{z}]$ over time duration dt is

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

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

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

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

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

$$L_Z = \frac{N_Z}{T} = \int_0^\infty \dot{z}p(Z, \dot{z})d\dot{z}. \tag{3.42}$$

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

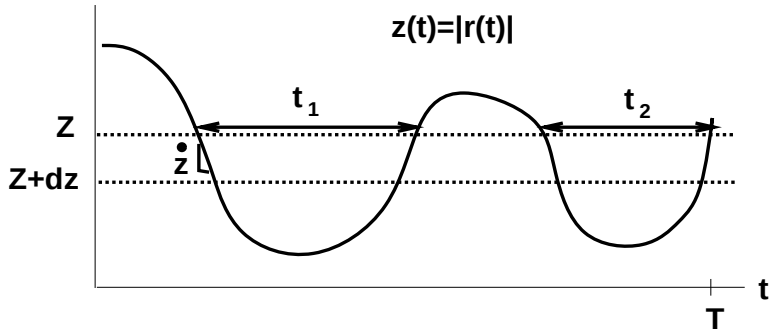


Figure 3.10: Level Crossing Rate and Fade Duration for Fading Process.

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

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

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

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

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

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

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

Thus, for T sufficiently large the average fade duration is

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

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

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

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

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

Example 3.3:

Consider a voice system with acceptable BER when the received signal power is at or above 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 the average, so $P_0 = P_r$ and thus $\rho = 1$. We require

$$\overline{t}_Z = \frac{e-1}{f_D\sqrt{2\pi}} \leq t = .060$$

and thus $f_D \geq (e-1)/(.060\sqrt{2\pi}) = 11$ Hz.

3.2.4 Finite State Markov Models

The complex mathematical characterization of flat-fading described in the previous subsections can be difficult to incorporate into wireless performance analysis such as the packet error probability. Therefore, simpler models that capture the main features of flat-fading channels are needed for these analytical

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

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

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

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

3.3 Wideband Fading Models

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

There are several techniques to mitigate the distortion due to multipath delay spread, including equalization, multicarrier modulation, and spread spectrum, which are discussed in later chapters. Systems with linear modulation typically use equalization to mitigate multipath distortion when $T << T_m$: equalization is not necessary in such systems if $T >> T_m$, but this can place significant constraints on

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

data rate. Multicarrier modulation and spread spectrum actually change the characteristics of the transmitted signal to mostly avoid intersymbol interference, however they still experience multipath distortion due to frequency-selective fading, which is described in Section 3.3.2.

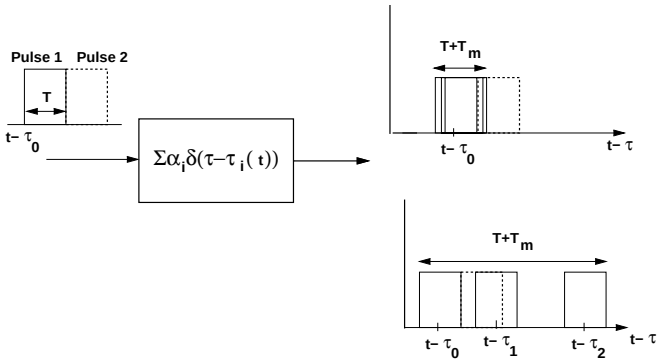


Figure 3.11: Multipath Resolution.

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

Although the approximation in (3.12) no longer applies when the signal bandwidth is large relative to the inverse of the multipath delay spread, if the number of multipath components is large and the phase of each component is uniformly distributed then the received signal will still be a zero-mean complex Gaussian process with a Rayleigh-distributed envelope. However, wideband fading differs from narrowband fading in terms of the resolution of the different multipath components. Specifically, for narrowband signals, the multipath components have a time resolution that is less than the inverse of the signal bandwidth, so the multipath components characterized in Equation (3.6) combine at the receiver to yield the original transmitted signal with amplitude and phase characterized by random processes. These random processes are characterized by their autocorrelation or PSD, and their instantaneous distributions, as discussed in Section 3.2. However, with wideband signals, the received signal experiences distortion due to the delay spread of the different multipath components, so the received signal can no longer be characterized by just the amplitude and phase random processes. The effect of multipath on wideband signals must therefore take into account both the multipath delay spread and the time-variations associated with the channel.

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

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

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

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

the deterministic scattering function $S_c(\tau, \rho)$ captures the Doppler characteristics of the channel via the frequency parameter ρ.

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

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

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

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

$$A_c(\tau_1, \tau_2; \Delta t) = .5E[c^*(\tau_1; t)c(\tau_2; t + \Delta t)]. \tag{3.51}$$

Moreover, in practice the channel response associated with a given multipath component of delay τ_1 is uncorrelated with the response associated with a multipath component at a different delay $\tau_2 \neq \tau_1$, since the two components are caused by different scatterers. We say that such a channel has uncorrelated scattering (US). We abbreviate channels that are WSS with US as WSSUS channels. The WSSUS channel model was first introduced by Bello in his landmark paper [16], where he also developed two-dimensional transform relationships associated with this autocorrelation. These relationships will be discussed in Section 3.3.4. Incorporating the US property into (3.51) yields

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

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

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

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

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

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

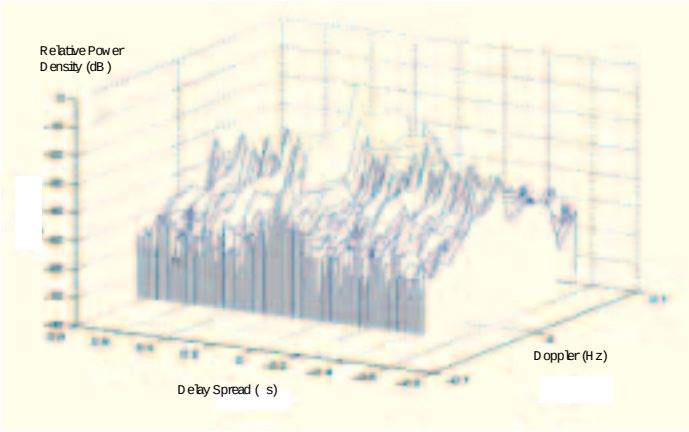


Figure 3.12: Scattering Function.

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

3.3.1 Power Delay Profile

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

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

and

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

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

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

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

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

While $\mu_{T_m} \approx \sigma_{T_m}$ in most channels with a large number of scatterers, a channel with no LOS component and a small number of multipath components with approximately the same large delay will have $\mu_{T_m} >> \sigma_{T_m}$. In this case the large value of μ_{T_m} is a misleading metric of delay spread, since in fact all copies of the transmitted signal arrive at rougly the same time. That is why rms delay spread is used more often than mean delay spread to characterize time-spreading in multipath channels.

Example 3.4:

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

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

Show that the average delay spread (3.54) is $\mu_{T_m} = \overline{T}_m$ and find the rms delay spread (3.55).
Solution: It is easily shown that $A_c(\tau)$ integrates to one. The average delay spread is thus given by

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

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

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

Example 3.5:

Consider a wideband channel with multipath intensity profile

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

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

Solution: The average delay spread is

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

The rms delay spread is

$$\sigma_{T_m} = \frac{\int_0^{20*10^{-6}} (\tau - \mu_{T_m})^2 e^{-\tau} d\tau}{\int_0^{20*10^{-6}} e^{-\tau} d\tau} = 5.76 \ \mu\text{sec}.$$

We see in this example that the mean delay spread is roughly twice its rms value, in contrast to the prior example where they were equal. This is due to the fact that in this example all multipath components have delay spread less than 20 μsec, which limits the rms spread. When the average and rms delay spreads differ significantly we use rms delay spread to determine ISI effects. Specifically, 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 $T_s = 57.6$ μsec or a symbol rate of $R_s = 1/T_s = 17.35$ Kilosymbols per second. This is a highly constrained symbol rate for many wireless systems. Specifically, for binary modulations where the symbol rate equals the data rate (bits per second, or bps), high-quality voice requires on the order of 32 Kbps and high-speed data can requires on the order of 10-100 Mbps.

3.3.2 Coherence Bandwidth

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

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

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

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

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

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

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

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

If we define $A_C(\Delta f) \triangleq A_C(\Delta f; 0)$ then from (3.59),

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

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

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

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

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

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

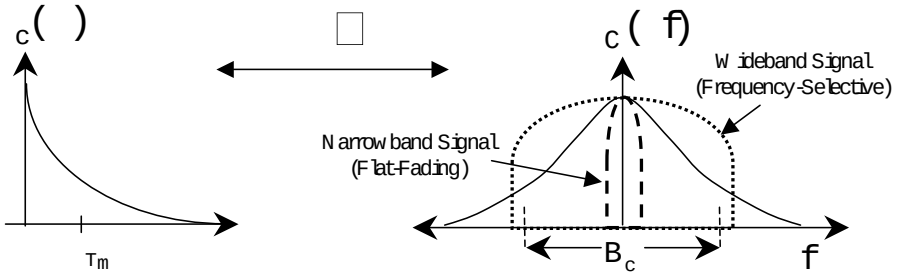


Figure 3.13: Power Delay Profile, RMS Delay Spread, and Coherence Bandwidth, and Coherence Bandwidth.

systems than in outdoor systems due to their lower delay spread values, which is why indoor systems tend to have higher data rates than outdoor systems.

3.3.3 Doppler Power Spectrum and Channel Coherence Time

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

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

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

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

where $A_C(\Delta t) \stackrel{\triangle}{=} A_C(\Delta f = 0; \Delta t)$. Note that $A_C(\Delta t)$ is an autocorrelation function defining how the channel impulse response decorrelates over time. In particular $A_C(\Delta t = T) = 0$ indicates that observations of the channel impulse response at times separated by T are uncorrelated and therefore independent, since the channel is a Gaussian random process. We define the **channel coherence time** T_c to be the range of values over which $A_C(\Delta t)$ is approximately nonzero. Thus, the time-varying channel decorrelates after approximately T_c seconds. The function $S_C(\rho)$ is called the **Doppler power spectrum** of the channel: as the Fourier transform of an autocorrelation it gives the PSD of the received signal as a function of Doppler ρ. The maximum ρ value for which $|S_C(\rho)|$ is greater than zero is called the **Doppler spread** of the channel, and is denoted by B_D. By the Fourier transform relationship between $A_C(\Delta t)$ and $S_C(\rho)$, $B_D \approx 1/T_c$. If the transmitter and reflectors are all stationary and the receiver is moving with velocity v, then $B_D \leq v/\lambda = f_D$. Recall that in the narrowband fading model samples became independent at time $\Delta t = .4/f_D$, so in general $B_D \approx k/T_c$ where k depends on the shape of $S_c(\rho)$. We illustrate the Doppler power spectrum $S_C(\rho)$ and its inverse Fourier transform $A_C(\Delta_t)$ in Figure 3.14.

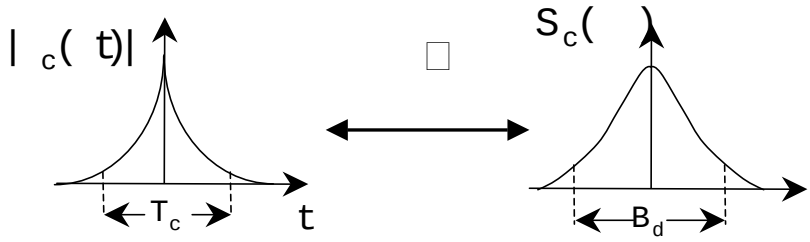


Figure 3.14: Doppler Power Spectrum, Doppler Spread, and Coherence Time.

Example 3.7:

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

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

3.3.4 Transforms for Autocorrelation and Scattering Functions

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

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

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

Empirical measurements of the scattering function for a given channel are often used to approximate empirically the channel's delay spread, coherence bandwidth, Doppler spread, and coherence time. A channel with empirical scattering function $S_c(\tau;\rho)$ approximates the delay spread T_m by computing the empirical power delay profile $A_c(\tau)$ from $A_c(\tau,\Delta t) = \mathcal{F}_\rho^{-1}[S_c(\tau;\rho)]$ with $\Delta t = 0$ and then characterizing the delay spread from this power delay profile, e.g. using the rms delay spread (3.55). The coherence bandwidth can then be approximated as $B_c \approx 1/T_m$. 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 $T_c \approx 1/B_D$.

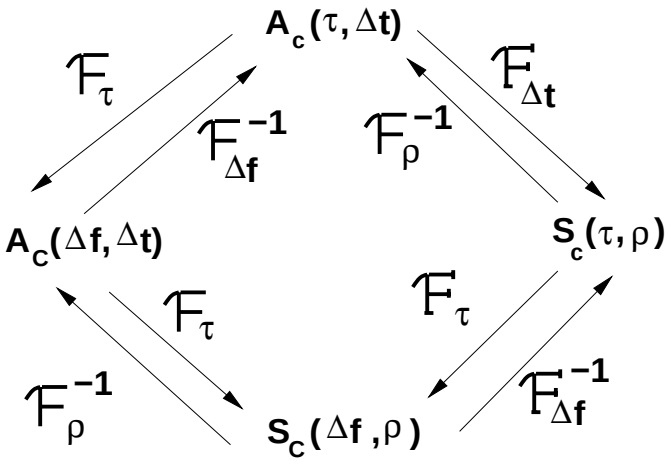


Figure 3.15: Fourier Transform Relationships

3.4 Discrete-Time Model

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

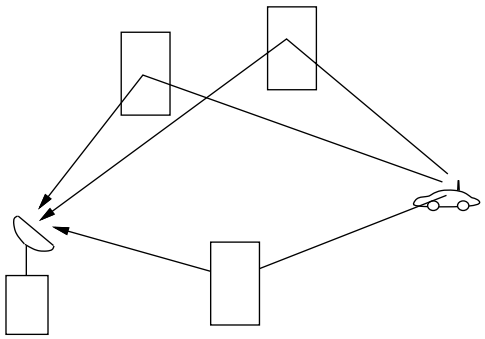


Figure 3.16: Point Scatterer Channel Model

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

$$c(\tau; t) = \sum_{n=1}^{N} \alpha_n(t) e^{-j\phi_n(t)} \delta(\tau - \tau_n(t)). \tag{3.64}$$

The statistics of the received signal for a given t are thus given by the statistics of $\{\tau_n\}_1^N, \{\alpha_n\}_1^N$, and $\{\phi_n\}_1^N$. The model can be further simplified using a discrete time approximation as follows: For a fixed t, the time axis is divided into M equal intervals of duration T such that $MT \geq T_m$, where T_m is the

delay spread of the channel, which is derived empirically. The subpaths are restricted to lie in one of the M time interval bins, as shown in Figure 3.17. The multipath spread of this discrete model is $T_m = MT$, and the resolution between paths is T. This resolution is based on the transmitted signal bandwidth: $T \approx 1/B$. The statistics for the nth bin are that r_n, $1 \leq n \leq M$, is a binary indicator of the existance of a multipath component in the nth bin: so r_n is one if there is a multipath component in the nth bin and zero otherwise. If $r_n = 1$ then (a_n, θ_n), the amplitude and phase corresponding to this multipath component, follow an empirically determined distribution. This distribution is obtained by sample averages of (a_n, θ_n) for each n at different locations in the propagation environment. The empirical distribution of (a_n, θ_n) and (a_m, θ_m), $n \neq m$, is generally different, it may correspond to the same family of fading but with different parameters (e.g. Ricean fading with different K factors), or it may correspond to different fading distributions altogether (e.g. Rayleigh fading for the nth bin, Nakagami fading for the mth bin).

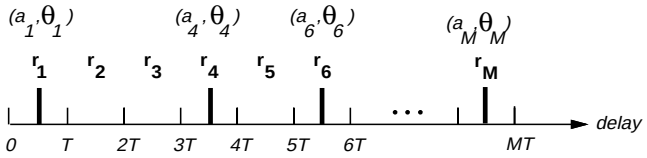


Figure 3.17: Discrete Time Approximation

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

3.5 Spatio-Temporal Models

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

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

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

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

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

where $\psi_{n,i} = [x_i \cos\theta_n(t) + y_i \sin\theta_n(t)]2\pi/\lambda$ for (x_i, y_i) the antenna location relative to the origin and $\theta_n(t)$ the AOA of the multipath relative to the origin of the antenna array. Details for the distribution of

the AOA for different propagation environments along with the corresponding correlations across antenna elements can be found in [24]

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 3 Problems

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(c) Turn in your well-commented code.

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

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

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

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

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

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

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

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

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

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

16. Consider a wideband channel characterized by the autocorrelation function

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

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

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

(b) Sketch the scattering function of this channel.

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

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

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

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

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

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

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

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

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

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

Chapter 4

Capacity of Wireless Channels

4.1 Introduction

The growing demand for wireless communication makes it important to determine the capacity limits of these channels. These capacity limits dictate the maximum data rates that can be achieved without any constraints on delay or complexity of the encoder and decoder. Channel capacity was pioneered by Claude Shannon in the late 1940s, where he developed a mathematical theory of communication based on the notion of mutual information between the input and output of a channel [1, 2, 3]. Shannon defined capacity as the mutual information maximized over all possible input distributions. The significance of this mathematical construct was Shannon's coding theorem and converse, which proved that a code did exist that could achieve a data rate close to capacity with negligible probability of error, and that any data rate higher than capacity could not be achieved without an error probability bounded away from zero. Shannon's ideas were quite revolutionary at the time, given the high data rates he predicted were possible on telephone channels and the notion that coding could reduce error probability without reducing data rate or causing bandwidth expansion. In time sophisticated modulation and coding technology validated Shannon's theory such that on telephone lines today, we achieve data rates very close to Shannon capacity with very low probability of error. These sophisticated modulation and coding strategies are treated in Chapters 5 and 8, respectively.

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

We will first investigate flat-fading channel capacity when there is no information about the channel fading (its value or distribution) at either the receiver or the transmitter. This is the "worst-case" scenario in terms of system design, and the capacity of this channel for most fading distributions (e.g. Rayleigh, Ricean, and Nakagami fading) is zero under this assumption. We then consider a flat-fading channel where only the fading distribution is known at the transmitter and receiver. Capacity under this assumption

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

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

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

4.2 Capacity in AWGN

Consider a discrete-time additive white Gaussian noise (AWGN) channel with channel input/output relationship $y[i] = x[i] + n[i]$, where $x[i]$ is the channel input at time i, $y[i]$ is the corresponding channel output, and $n[i]$ is a white Gaussian noise random process. Assume a channel bandwidth B and transmit power S. The channel SNR, the power in $x[i]$ divided by the power in $n[i]$, is constant and given by $\gamma = S/(N_0B)$, where N_0 is the power spectral density of the noise. The capacity of this channel is given by Shannon's well-known formula [1]:

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

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

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

where the sum is taken over all possible input and output pairs (x,y). The log function is typically with respect to base 2, in which case the units of mutual information are bits per second. Shannon proved that channel capacity equals the mutual information of the channel maximized over all possible input distributions:

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

For the AWGN channel, the maximizing input distribution is Gaussian, which results in the channel capacity given by (4.1). For channels with memory, mutual information and channel capacity are defined

relative to input and output sequences X^n and Y^n. More details on channel capacity, mutual information, and the coding theorem and converse can be found in [2, 5, 6].

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

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

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

Solution: The received SNR is $\gamma = P_r(d)/(N_0B) = .1 * .1^3/(10^{-6} * 30 * 10^3) = 33 = 15$ dB for $d = 100$ m and $\gamma = .1 * .01^3/(10^{-6} * 30 * 10^3) = .033 = -15$ dB for $d = 1000$ m. The corresponding capacities are $C = B\log_2(1+\gamma) = 30000\log_2(1+33) = 152.6$ Kbps for $d = 100$m and $C = 30000\log_2(1+.033) = 1.4$ Kbps for $d = 1000$ m. Note the significant decrease in capacity at farther distances, due to the path loss exponent of 3, which greatly reduces received power as distance increases.

4.3 Capacity of Flat-Fading Channels

4.3.1 Channel and System Model

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

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

during the transmission of the codeword.

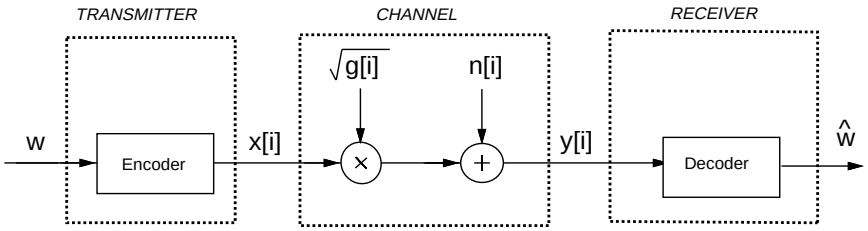


Figure 4.1: Flat-Fading Channel and System Model.

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

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

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

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

Transmitter and receiver CSI allows the transmitter to adapt both its power and rate to the channel gain at time i, and leads to the highest capacity of the three scenarios. Note that since the instantaneous SNR $\gamma[i]$ is just $g[i]$ multipled by the constant $\overline{S}/(N_0B)$, 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 [18, 19].

4.3.2 Channel Distribution Information (CDI) Known

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

The capacity-acheiving input distribution and corresponding fading channel capacity under CDI is 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 exponential and changes independently with each channel use. The optimal input distribution for this channel was shown in [8] to be discrete with a finite number of mass points, one of which is located at zero. This optimal distribution and its corresponding capacity must be found numerically. The lack of closed-form solutions for capacity or the optimal input distribution is somewhat surprising given the fact that the fading follows the most common fading distribution and has no correlation structure. For flat-fading channels that are not necessarily Rayleigh or i.i.d. upper and lower bounds on capacity have been determined in [11], and these bounds are tight at high SNRs.

Finite State Markov Channels for fading channels were discussed in Chapter 3.2.4. This model approximates the fading correlation as a Markov process. While the Markov nature of the fading dictates that the fading at a given time depends only on fading at the previous time sample, it turns out that the receiver must decode all past channel outputs jointly with the current output for optimal (i.e. capacity-achieving) decoding. This significantly complicates capacity analysis. The capacity of Finite State Markov channels has been derived for i.i.d. inputs in [9, 17] and for general inputs in [10]. Channel capacity in both cases depends on the limiting distribution of the channel conditioned on all past inputs and outputs, which can be computed recursively. As with the i.i.d. Rayleigh fading channel, the complexity of the capacity analysis along with the final result for this relatively simple fading model is very high, indicating the difficulty of obtaining the capacity and related design insights into channels when only CDI is available.

4.3.3 Channel Side Information at Receiver

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

Shannon (Ergodic) Capacity

Shannon capacity of a fading channel with receiver CSI for an average power constraint $\overline{S}$ can be obtained from results in [22] as

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

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

By Jensen's inequality,

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

where $\overline{\gamma}$ is the average SNR on the channel. Thus we see that the Shannon capacity of a fading channel with receiver CSI only is less than the Shannon capacity of an AWGN channel with the same average SNR. In other words, fading reduces Shannon capacity when only the receiver has CSI. Moreover, without transmitter side information, 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 [19].

Example 4.2: Consider a flat-fading channel with i.i.d. channel gain $g[i]$ which can take on three possible values: $g_1 = .025$ with probability $p_1 = .1$, $g_2 = .25$ with probability $p_2 = .5$, and $g_3 = 1$ with probability $p_3 = .4$. The transmit power is 10 mW, the noise spectral density is $N_0 = 10^{-9}$ W/Hz, and the channel bandwidth is 30 KHz. Assume the receiver has knowledge of the instantaneous value of $g[i]$ but the transmitter does not. Find the Shannon capacity of this channel and compare with the capacity of an AWGN channel with the same average SNR.

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

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

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

Capacity with Outage

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

Capacity with outage allows bits sent over a given transmission burst to be decoded at the end of the burst with some probability that these bits will be decoded incorrectly. Specifically, the transmitter fixes a minimum received SNR γ_{min} and encodes for a data rate $C = B\log_2(1+\gamma_{min})$. The data is correctly received if the instantaneous received SNR is greater than or equal to γ_{min} [11, 12]. If the received SNR is below γ_{min} then the bits received over that transmission burst cannot be decoded correctly, and the

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

receiver declares an outage. The probability of outage is thus $p_{out} = p(\gamma < \gamma_{min})$. The average rate correctly received over many transmission bursts is $C_o = (1 - p_{out})B\log_2(1 + \gamma_{min})$ since data is only correctly received on $1 - p_{out}$ transmissions. The value of γ_{min} is typically a design parameter based on the acceptable outage probability. Capacity with outage is typically characterized by a plot of capacity versus outage, as shown in Figure 4.2. In this figure we plot the normalized capacity $C/B = \log_2(1+\gamma_{min})$ as a function of outage probability $p_{out} = p(\gamma < \gamma_{min})$ for a Rayleigh fading channel (γ exponential) with $\overline{\gamma} = 20$ dB. We see that capacity approaches zero for small outage probability, due to the requirement to correctly decode bits transmitted under severe fading, and increases dramatically as outage probability increases. Note, however, that these high capacity values for large outage probabilities have higher probability of incorrect data reception. The average rate correctly received can be maximized by finding the γ_{min} that maximizes C_o (or, equivalently, the maximizing p_{out}).

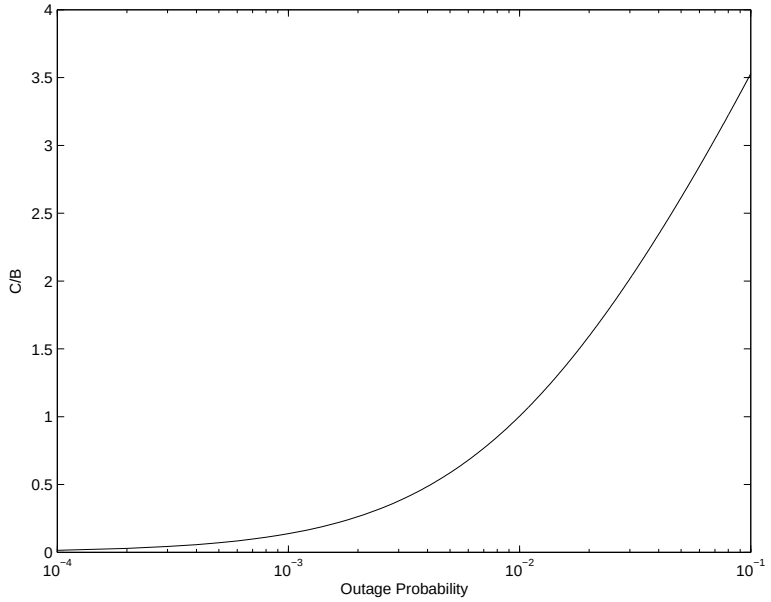


Figure 4.2: Normalized Capacity (C/B) versus Outage Probability.

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

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

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

4.3.4 Channel Side Information at the Transmitter and Receiver

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

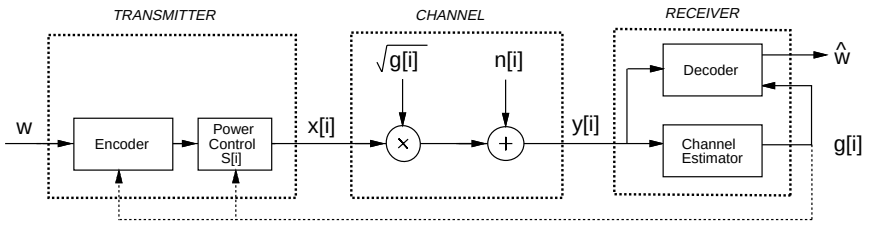


Figure 4.3: System Model with Transmitter and Receiver CSI.

Shannon Capacity

We now consider the Shannon capacity when the channel power gain $g[i]$ is known to both the transmitter and receiver at time i. The Shannon capacity of a time-varying channel with side information about the channel state at both the transmitter and receiver was originally considered by Wolfowitz for the following model. Let $s[i]$ be a stationary and ergodic stochastic process representing the channel state, which takes values on a finite set $\mathcal{S}$ of discrete memoryless channels. Let C_s denote the capacity of a particular channel $s \in \mathcal{S}$, and $p(s)$ denote the probability, or fraction of time, that the channel is in state s. The capacity of this time-varying channel is then given by Theorem 4.6.1 of [20]:

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

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

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

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

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

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

With this additional constraint, we cannot apply (4.7) directly to obtain the capacity. However, we expect that the capacity with this average power constraint will be the average capacity given by (4.7) with the power optimally distributed over time. This motivates defining the fading channel capacity with average power constraint (4.8) as

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

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

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

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

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

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

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

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

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

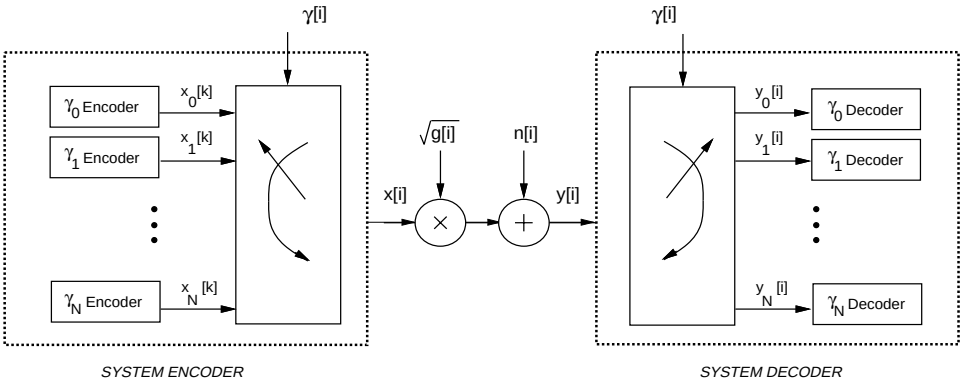


Figure 4.4: Multiplexed Coding and Decoding.

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

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

The multiplexing nature of the capacity-achieving coding strategy indicates that (4.13) is achieved with a time-varying data rate, where the rate corresponding to instantaneous SNR γ is $B \log_2(\gamma/\gamma_0)$. Since γ_0 is constant, this means that as the instantaneous SNR increases, the data rate sent over the channel for that instantaneous SNR also increases. Note that this multiplexing strategy is not the only way to achieve capacity (4.13): it can also be achieved by adapting the transmit power and sending at a fixed rate [21]. We will see in Section 4.3.6 that for Rayleigh fading this capacity can exceed that of an AWGN channel with the same average power, in contrast to the case of receiver CSI only, where fading always decreases capacity.

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

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

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

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

Note that this expression only depends on the distribution $p(\gamma)$. The value for γ_0 cannot be solved for in closed form for typical continuous pdfs $p(\gamma)$ and thus must be found numerically [13].

Since γ is time-varying, the maximizing power adaptation policy of (4.12) is a "water-filling" formula in time, as illustrated in Figure 4.5. This curve shows how much power is allocated to the channel for instantaneous SNR $\gamma(t) = \gamma$. The water-filling terminology refers to the fact that the line $1/\gamma$ sketches out the bottom of a bowl, and power is poured into the bowl to a constant water level of $1/\gamma_0$. The amount of power allocated for a given γ equals $1/\gamma_0 - 1/\gamma$, the amount of water between the bottom of the bowl ($1/\gamma$) and the constant water line ($1/\gamma_0$). The intuition behind water-filling is to take advantage

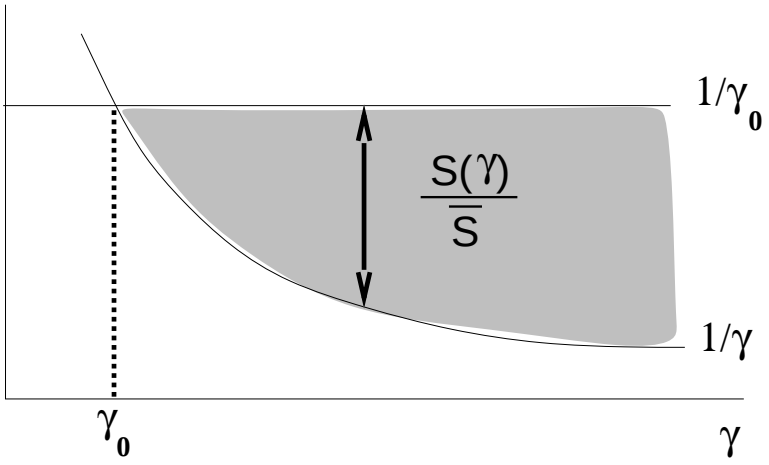


Figure 4.5: Optimal Power Allocation: Water-Filling.

of good channel conditions: when channel conditions are good (γ large) more power and a higher data rates is sent over the channel. As channel quality degrades (γ small) less power and rate are sent over the channel. If the instantaneous channel SNR falls below the cutoff value, the channel is not used. Adaptive modulation and coding techniques that follow this same principle were developed in [15, 16] and are discussed in Chapter 9.

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

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

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

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

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

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

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

channel model yields

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

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

$$\sum_{i=2}^{3}\frac{p(\gamma_i)}{\gamma_0}-\sum_{i=2}^{3}\frac{p(\gamma_i)}{\gamma_i}=1\Rightarrow\frac{.9}{\gamma_0}=1-\sum_{i=2}^{3}\frac{p(\gamma_i)}{\gamma_i}=1+\left(\frac{.5}{83.33}+\frac{.4}{333.33}\right)=.89.$$

Solving for γ_0 yields $\gamma_0 = 1/.89 = 1.12$. So by assuming the weakest channel with SNR γ_1 is not used, we obtain a consistent value for γ_0 with $\gamma_1 < \gamma_0 \leq \gamma_2$. The capacity of the channel then becomes

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

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

Zero-Outage Capacity and Channel Inversion

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

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

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

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

Zero-outage capacity can exhibit a large data rate reduction relative to Shannon capacity in extreme fading environments. For example, in Rayleigh fading $\mathbf{E}[1/\gamma]$ is infinite, and thus the zero-outage capacity given by (4.18) is zero. Channel inversion is common in spread spectrum systems with near-far interference

imbalances [23]. It is also the simplest scheme to implement, since the encoder and decoder are designed for an AWGN channel, independent of the fading statistics.

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

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

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

we have $C = 30000\log_2(1 + 1/.1272) = 94,43$ 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 zero-outage capacity may be significantly smaller than Shannon capacity on a fading channel is the requirement to maintain a constant data rate in all fading states. By suspending transmission in particularly bad fading states (outage channel states), we can maintain a higher constant data rate in the other states and thereby significantly increase capacity. The **outage capacity** is defined as the maximum data rate that can be maintained in all nonoutage channel states times the probability of nonoutage. Outage capacity is achieved with a **truncated channel inversion** policy for power adaptation that only compensates for fading above a certain cutoff fade depth γ_0:

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

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

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

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

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

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

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

This maximum outage capacity will still be less than Shannon capacity (4.13) since truncated channel inversion is a suboptimal transmission strategy. However, the transmit and receive strategies associated

with inversion or truncated inversion may be easier to implement or have lower complexity than the water-filling schemes associated with Shannon capacity.

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

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

4.3.5 Capacity with Receiver Diversity

Receiver diversity is a well-known technique to improve the performance of wireless communications in fading channels. The main advantage of receiver diversity is that it mitigates the fluctuations due to fading so that the channel appears more like an AWGN channel. More details on receiver diversity and its performance will be given in Chapter 7. Since receiver diversity mitigates the impact of fading, an interesting question is whether it also increases the capacity of a fading channel. The capacity calculation under diversity combining first requires that the distribution of the received SNR $p(\gamma)$ under the given diversity combining technique be obtained. Once this distribution is known it can be substituted into any of the capacity formulas above to obtain the capacity under diversity combining. The specific capacity formula used depends on the assumptions about channel side information, e.g. for the case of perfect transmitter and receiver CSI the formula (4.13) would be used. Capacity under both maximal ratio and selection combining diversity for these different capacity formulas was computed in [26]. It was found that, as expected, the capacity with perfect transmitter and receiver CSI is bigger than with receiver CSI only, which in turn is bigger than with channel inversion. The performance gap of these different formulas decreases as the number of antenna branches increases. This trend is expected, since a large number of antenna branches makes the channel look like AWGN, for which all of the different capacity formulas have roughly the same performance.

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

4.3.6 Capacity Comparisons

In this section we compare capacity with transmitter and receiver CSI for different power allocation policies along with the capacity under receiver CSI only. Figures 4.6, 4.7, and 4.8 show plots of the

different capacities (4.4), 4.9), (4.18), and (4.22) as a function of average received SNR for log-normal fading (σ=8 dB standard deviation), Rayleigh fading, and Nakagami fading (with Nakagami parameter $m = 2$). We will see in Chapter 7 that Nakagami fading with $m = 2$ is equivalent to Rayleigh fading with two-antenna receiver diversity. The capacity in AWGN for the same average power is also shown for comparison. Note that the capacity in log-normal fading is plotted relative to average dB SNR (μ_{dB}), not average SNR in dB ($10\log_{10}\mu$): the relation between these values, as given by (2.46) in Chapter 2, is $10\log_{10}\mu = \mu_{dB} + \sigma_{dB}^2\ln(10)/20$.

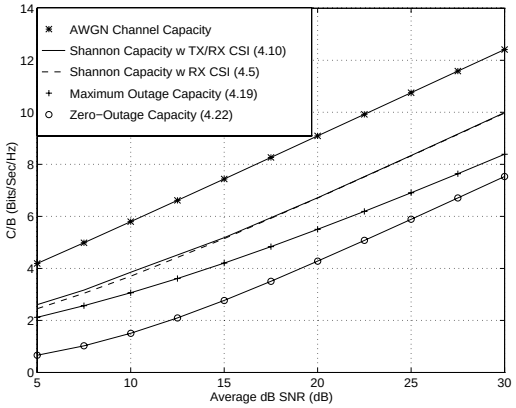


Figure 4.6: Capacity in Log-Normal Shadowing.

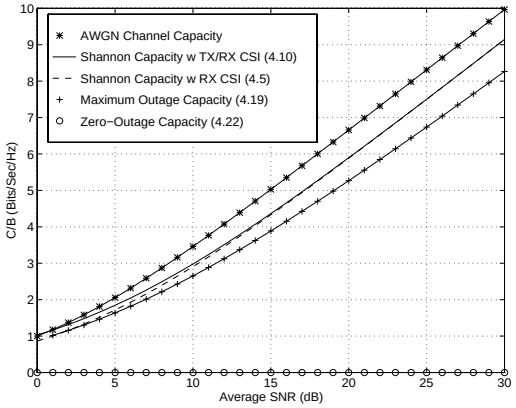


Figure 4.7: Capacity in Rayleigh Fading.

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

The severity of the fading is indicated by the Nakagami parameter m, where $m = 1$ for Rayleigh

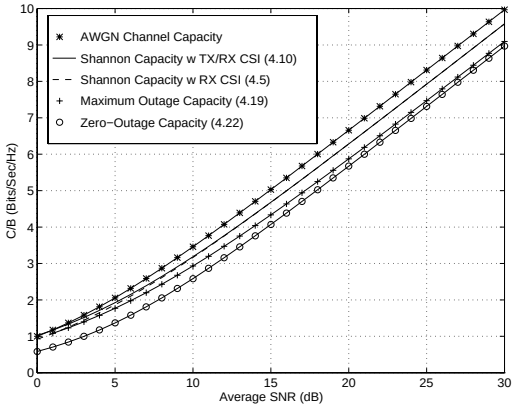


Figure 4.8: Capacity in Nakagami Fading ($m = 2$).

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 policies also decreases, and their respective capacities approach that of the AWGN channel.

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

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

In general, Shannon capacity analysis does not show how to design adaptive or nonadaptive techniques for real systems. Achievable rates for adaptive trellis-coded MQAM have been investigated in [16], where a simple 4-state trellis code combined with adaptive six-constellation MQAM modulation was shown to achieve rates within 7 dB of the Shannon capacity (4.9) in Figures 4.6 and 4.7. More complex codes further close the gap to the Shannon limit of fading channels with transmitter adaptation.

4.4 Capacity of Frequency-Selective Fading Channels

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

4.4.1 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 S. When the channel is time-invariant it is typically assumed that $H(f)$ is known at both the transmitter and receiver: capacity of time-invariant channels under different assumptions of this channel knowledge are discussed in [12].

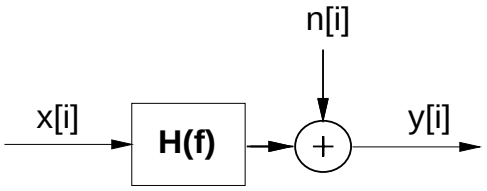


Figure 4.9: Time-Invariant Frequency-Selective Fading Channel.

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

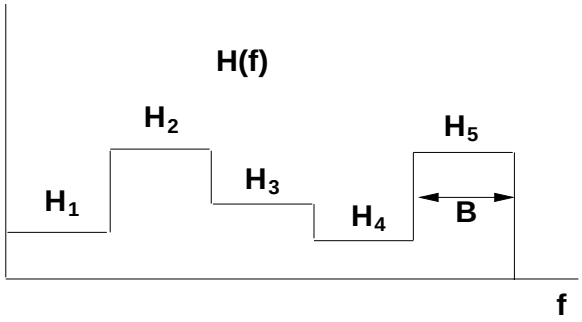


Figure 4.10: Block Frequency-Selective Fading

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

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

Note that this is similar to the capacity and optimal power allocation for a flat-fading channel, with power and rate changing over frequency in a deterministic way rather than over time in a probabilistic

way. The optimal power allocation is found via the same Lagrangian technique used in the flat-fading case, which leads to the water-filling power allocation

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

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

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

The capacity then becomes

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

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

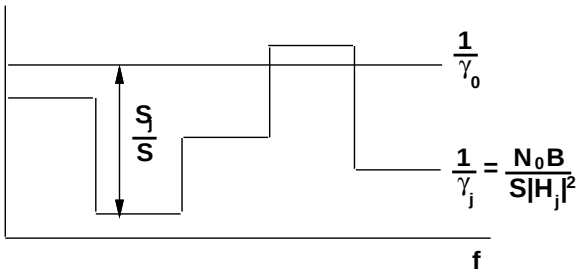


Figure 4.11: Water-Filling in Block Frequency-Selective Fading

When $H(f)$ is continuous the capacity under power constraint S is similar to the case of the block-fading channel, with some mathematical intricacies needed to show that the channel capacity is given by [5]

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

The equation inside the integral can be thought of as the incremental capacity associated with a given frequency f over the bandwidth df with power allocation $S(f)$. The optimal power allocation over frequency, $S(f)$, found via the Lagrangian technique applied to (4.27), is again water-filling:

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

This results in channel capacity

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

Example 4.7: Consider a time-invariant block fading frequency-selective fading channel consisting of

three subchannels of bandwidth $B = 1$ MHz. The frequency response associated with each channel is $H_1 = 1$, $H_2 = 2$ and $H_3 = 3$. The transmit power constraint is $S = 10$ mW and the noise power spectral density is $N_0 = 10^{-9}$ W/Hz. Find the Shannon capacity of this channel and the optimal power allocation that achieves this capacity.

Solution: We first first find $\gamma_j = |H_j|^2 S/(N_b)$ for each subchannel, yielding $\gamma_1 = 10$, $\gamma_2 = 40$ and $\gamma_3 = 90$. The cutoff γ_0 must satisfy (4.26). Assuming all subchannels are allocated power, this yields

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

Since the cutoff γ_0 is less than γ_j for all j, our assumption that all subchannels are allocated power is consistent, so this is the correct cutoff value. The corresponding capacity is $C = \sum_{j=1}^{3} B \log_2(\gamma_j/\gamma_0) = 1000000(\log_2(10/2.64) + \log_2(40/2.64) + \log_2(90/2.64)) = 7.61$ Mbps.

4.4.2 Time-Varying Channels

The time-varying frequency-selective fading channel is similar to the model shown in Figure 4.9, except that $H(f) = H(f, i)$, i.e. the channel varies over both frequency and time. It is difficult to determine the capacity of time-varying frequency-selective fading channels, even when the instantaneous channel $H(f, i)$ is known perfectly at the transmitter and receiver, due to the random effects of self-interference (ISI). In the case of transmitter and receiver side information, the optimal adaptation scheme must consider the effect of the channel on the past sequence of transmitted bits, and how the ISI resulting from these bits will affect future transmissions [25]. The capacity of time-varying frequency-selective fading channels is in general unknown, however upper and lower bounds and limiting formulas exist [25, 33].

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

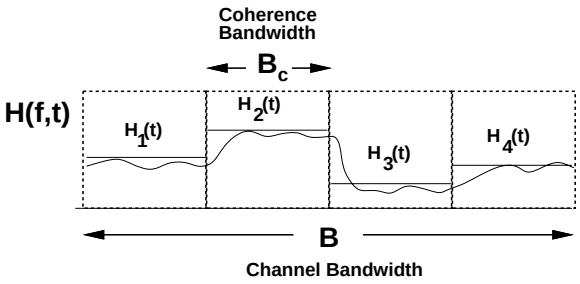


Figure 4.12: Channel Division in Frequency-Selective Fading

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

both time and frequency:

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

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

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

Define $\gamma_j[i] = |H_j[i]|^2\overline{S}/(N_0B)$ to be the instantaneous SNR on the jth subchannel at time i assuming the total power $\overline{S}$ is allocated to that time and frequency. We allow the power $S_j(\gamma_j)$ 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):

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

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

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

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

$$\frac{\partial J(S_j(\gamma_j))}{\partial S_j(\gamma_j)} = \left[\left(\frac{1}{1 + \gamma_j S(\gamma_j)/\overline{S}}\right)\frac{\gamma_j}{\overline{S}} - \lambda\right] p(\gamma_j) = 0. \tag{4.33}$$

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

$$\frac{S_j(\gamma_j)}{\overline{S}} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_j} & \gamma_j \ge \gamma_0 \\ 0 & \gamma_j < \gamma_0 \end{cases}, \tag{4.34}$$

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

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

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

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

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

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[26] M.-S. Alouini and A. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," IEEE Transactions on Vehicular Technology, pp. 1165 -1181, July 1999.

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

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

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

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

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

Chapter 4 Problems

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

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

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

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

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

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

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

   (a) Find the Shannon capacity of this channel.

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

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

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

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

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

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

   (a) Compute the distribution of the received SNR.

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

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

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

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

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

   (b) Compute the Shannon capacity of this channel.

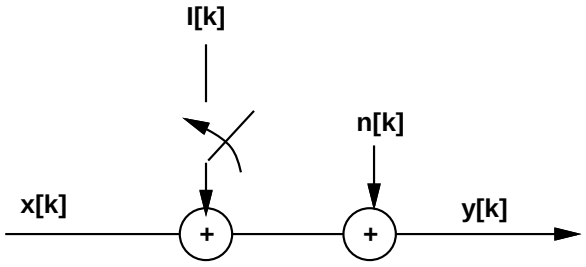
   (c) Compare your answer in part (b) with the channel capacity in AWGN with the same average SNR.

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

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

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

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



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

channel bandwidth B is 10 KHz (receiver noise power is N_oB), and the interference power (when on) is 9 mW.

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 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 data rates, powerful error correction techniques, resistance to channel impairments, more efficient multiple access strategies, and better security and privacy. Specifically, high level modulation techniques such as MQAM allow much higher data rates in digital modulation as compared to analog modulation with the same signal bandwidth. 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 ISI. Spread spectrum techniques applied to digital modulation can remove or combine multipath, resist interference, and detect multiple users simultaneously. 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 communications 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 determining 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)
- low power/cost implementation

Often these are conflicting requirements, and the choice of modulation is based on finding the technique that achieves the best tradeoff between these requirements.

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 general tradeoff of 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 technique. 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 are more susceptible to noise, fading, and hardware imperfections. Finally, the simplest demodulators require a coherent phase reference with respect to the transmitted signal. This coherent reference may be difficult to obtain or significantly increase receiver complexity. Thus, modulation techniques that do no require a coherent phase reference 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.

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 $m_i = \{\mathbf{b}_1, \ldots, \mathbf{b}_K\} \in \mathcal{M}$, 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 $[0, T)$. Since the

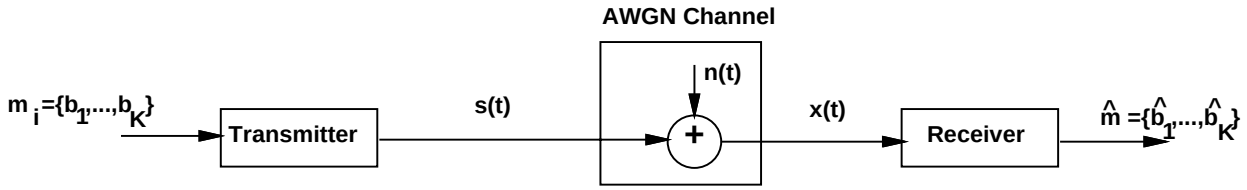


Figure 5.1: Communication System Model

channel is analog, the message must be embedded into an analog signal for channel transmission. Thus, each message $m_i \in \mathcal{M}$ is mapped to a unique analog signal $s_i(t) \in \mathcal{S} = \{s_1(t), \ldots, s_M(t)\}$ 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, each signal $s_i(t) \in \mathcal{S}$ 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 the 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 $s_i(t)$.

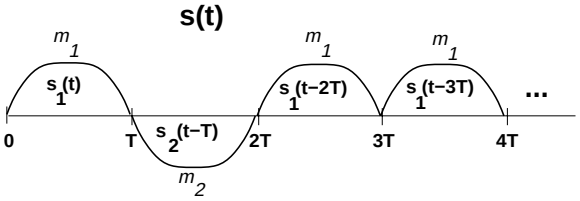


Figure 5.2: Transmitted Signal for a Sequence of Messages

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 $x(t) = s(t) + n(t)$. Given $x(t)$ the receiver must determine the best estimate of which $s_i(t) \in \mathcal{S}$ 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 $m_i(t) \in \mathcal{M}$ and the receiver then outputs this best estimate $\hat{m} = \{\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_K\} \in \mathcal{M}$ of the transmitted bit sequence.

The goal of the receiver design in estimating the transmitted message is to minimize the probability of message error:

$$P_e = \sum_{i=1}^{M} p(\hat{m} \neq m_i | m_i \text{ sent}) p(m_i \text{ sent}) \tag{5.2}$$

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, and methods to combat this performance degradation 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 [2, 3], it can be shown that any set of M real energy signals $\mathcal{S} = (s_1(t), \ldots, s_M(t))$ defined on $[0, T)$ 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 the basis functions span the signal associated with the signals in $\mathcal{S}$. Thus, we can write each $s_i(t) \in \mathcal{S}$ in terms of its **basis function representation** as

$$s_i(t) = \sum_{j=1}^{N} s_{ij}\phi_j(t), \quad 0 \leq t < T, \tag{5.3}$$

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 $N < M$.

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 only get an approximation to (5.5), since

$$\int_0^T \phi_i^2(t)dt = \frac{2}{T}\int_0^T .5[1 + \cos(2\pi f_c t)]dt = 1 + \frac{\sin(4\pi f_c T)}{4\pi f_c T}. \tag{5.8}$$

The numerator in the second term of (5.8) is bounded by one and for $f_c T >> 1$ the denominator of this term is very large. Thus, this second term can be neglected. Similarly,

$$\int_0^T \phi_i(t)\phi_j(t)dt = \frac{2}{T}\int_0^T .5\sin(2\pi f_c t)dt = \frac{-\cos(4\pi f_c T)}{4\pi f_c T} \approx 0, \tag{5.9}$$

where the approximation is taken as an equality for $f_c T >> 1$.

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 complex baseband representation of $s_i(t)$ in terms of its in-phase and quadrature components with an extra factor of $\sqrt{2/T}$:

$$s_i(t) = s_{i1}\sqrt{\frac{2}{T}}\cos(2\pi f_c t) + s_{i2}\sqrt{\frac{2}{T}}\sin(2\pi f_c t). \tag{5.10}$$

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, i.e. we must have

$$\int_0^T g^2(t)\cos^2(2\pi f_c t)dt = 1 \tag{5.12}$$

and

$$\int_0^T g^2(t)\cos(2\pi f_c t)\sin(2\pi f_c t) = 0, \tag{5.13}$$

where the equalities may be approximations for $f_c T >> 1$ as in (5.8) and (5.9) above. If the bandwidth of $g(t)$ satisfies $B << f_c$ then $g^2(t)$ is roughly constant over T_c, so (5.13) is approximately true since the sine and cosine functions are orthogonal over one period $T_c = 1/f_c$. 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 $s_2 = -\alpha\sqrt{T/2}$.

We denote the coefficients $\{s_{ij}\}$ as a vector $\mathbf{s}_i = (s_{i1}, \ldots, s_{iN}) \in \mathcal{R}^N$ which is called 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 $\mathbf{s_i} \in \mathcal{R}^{\mathbf{N}}$ 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 $\mathbf{s_i} \in \mathcal{R}^{\mathbf{2}}$ with the ith axis of $\mathcal{R}^2$ 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 $\mathcal{R}^2$. This greatly simplifies the analysis of the system performance as well as the derivation of the 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 $\mathcal{R}^N$. The length of a vector in $\mathcal{R}^N$ is defined as

$$||\mathbf{s}_i|| = \sqrt{\sum_{j=1}^{N} s_{ij}^2}. \tag{5.14}$$

The distance between two signal constellation points $\mathbf{s}_i$ and $\mathbf{s}_k$ is thus

$$||\mathbf{s}_i - \mathbf{s}_k|| = \sqrt{\sum_{j=1}^{N}(s_{ij} - s_{kj})^2} = \sqrt{\int_0^T (s_i(t) - s_k(t))^2 dt}, \tag{5.15}$$

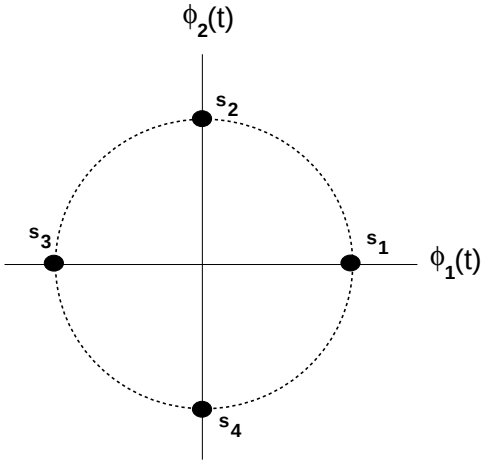


Figure 5.3: Signal Space Representation

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 $< s_i(t), s_k(t) >$ between two real signals $s_i(t)$ and $s_k(t)$ on the interval $[0, T]$ is

$$< s_i(t), s_k(t) >= \int_0^T s_i(t)s_k(t)dt. \tag{5.16}$$

Similarly, the inner product $< \mathbf{s}_i, \mathbf{s}_k >$ between two real vectors is

$$< \mathbf{s}_i\mathbf{s}_k >= \mathbf{s}_i^T\mathbf{s}_k = \int_0^T s_i(t)s_k(t)dt =< s_i(t), s_k(t) >, \tag{5.17}$$

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 $x(t) = s_i(t) + n(t), 0 \leq t < T$, 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 $x(t)$ over each time interval into a vector, as it allows us to work in finite-dimensional vector space to estimate the transmitted signal. However, this conversion should not compromise the estimation accuracy. For this conversion, 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 $x(t)$ as

$$\sum_{j=1}^{N}(s_{ij}+n_j)\phi_j(t)+n_r(t)=\sum_{j=1}^{N}x_j\phi_j(t)+n_r(t), \tag{5.20}$$

where $x_j = s_{ij}+n_j$ 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 $x(t)$ 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 $\mathbf{x} = (x_1, \ldots, x_N)$ alone. In other words, $\mathbf{x} = (x_1, \ldots, x_N)$ is a **sufficient statistic** for $x(t)$ in the optimal detection of the transmitted messages.

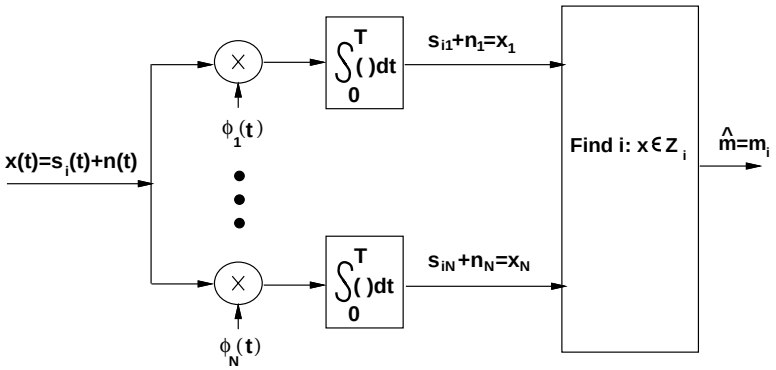


Figure 5.4: Receiver Structure for Signal Detection in AWGN.

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 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 $\phi_2(t)$. The vector space in the figure shows the projection of the received signal onto each of these basis functions. 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 $\mathbf{x} + n_r$. From the figure it appears that projecting $\mathbf{x} + n_r$ onto $\mathbf{x}$ will not compromise the detection of which constellation $\mathbf{s}_i$ was transmitted, since n_r lies in a space orthogonal to the space where $\mathbf{s}_i$ lies. We now proceed to show mathematically why this intuition is correct.

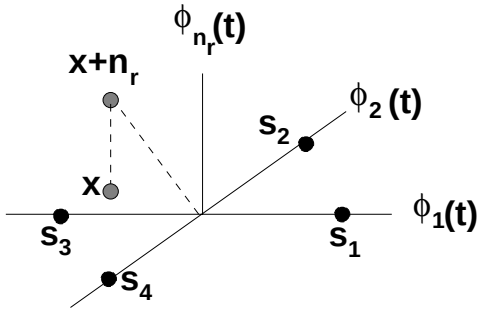


Figure 5.5: Projection of Received Signal onto Received Vector $\mathbf{x}$.

Let us first examine the distribution of $\mathbf{x}$. Since $n(t)$ is a Gaussian random process, if we condition on the transmitted signal $s_i(t)$ then the channel output $x(t) = s_i(t) + n(t)$ is also a Gaussian random

process and $\mathbf{x} = (x_1, \ldots, x_N)$ is a Gaussian random vector. Recall that $x_j = s_{ij} + n_j$. Thus, conditioned on a transmitted constellation $\mathbf{s}_i$, we have that

$$\mu_{x_j|\mathbf{s}_i} = E[x_j|\mathbf{s}_i] = E[s_{ij} + n_j|s_{ij}] = s_{ij} \tag{5.21}$$

since $n(t)$ has zero mean, and

$$\sigma_{x_j|\mathbf{s}_i} = E[x_j - \mu_{x_j|\mathbf{s}_i}]^2 = E[= s_{ij} + n_j - s_{ij}|s_{ij}]^2 = E[n_j^2]. \tag{5.22}$$

Moreover,

$$\begin{aligned}
\mathrm{Cov}[x_j x_k|\mathbf{s}_i] &= E[(x_j - \mu_{x_j})(x_k - \mu_{x_k})|\mathbf{s}_i] \\
&= E[n_j n_k] \\
&= E\left[\int_0^T n(t)\phi_j(t)dt[\int_0^T n(\tau)\phi_k(\tau)d\tau\right] \\
&= \int_0^T \int_0^T E[n(t)n(\tau)]\phi_j(t)\phi_k(\tau)dtd\tau \\
&= \int_0^T \int_0^T \frac{N_0}{2}\delta(t-\tau)\phi_j(t)\phi_k(\tau)dtd\tau \\
&= \frac{N_0}{2}\int_0^T \phi_j(t)\phi_k(t)dt \\
&= \begin{cases} N_0/2 & i = j \\ 0 & i \neq j \end{cases}
\end{aligned} \tag{5.23}$$

where the last equality follows from the orthogonality of the basis functions. Thus, conditioned on the transmitted constellation $\mathbf{s}_i$, the x_js are uncorrelated and, since they are Gaussian, they are also independent. Moreover $E[n_j^2] = N_0/2$.

We have shown that, conditioned on the transmitted constellation $\mathbf{s}_i$, x_j is a Gauss-distributed random variable that is independent of $x_k, k \neq j$ and has mean s_{ij} and variance $N_0/2$. Thus, the conditional distribution of $\mathbf{x}$ is given by

$$p(\mathbf{x}|\mathbf{s}_i \text{ sent}) = \prod_{j=1}^N p(x_j|m_i) = \frac{1}{(\pi N_0)^{N/2}}\exp\left[-\frac{1}{N_0}\sum_{j=1}^N (x_j - s_{ij})^2\right]. \tag{5.24}$$

It is also straightforward to show that $E[x_j n_r(t)|\mathbf{s}_i] = 0$ for any $t, 0 \leq t < T$. Thus, since x_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 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 $x(t)$. To minimize $P_e = p(\hat{m} \neq m_i|x(t)) = 1 - p(\hat{m} = m_i|x(t))$ we maximize $p(\hat{m} = m_i|x(t))$. Therefore, the receiver output $\hat{m}$ given received signal $x(t)$ should correspond to the message m_i that maximizes $p(m_i \text{ sent}|x(t))$. Since there is a one-to-one mapping between messages and signal constellation points, this is equivalent to maximizing $p(\mathbf{s}_i \text{ sent}|x(t))$. Recalling that $x(t)$ is completely described by $\mathbf{x} = (x_1, \ldots, x_N)$ and $n_r(t)$, we have

$$p(\mathbf{s}_i \text{ sent}|x(t)) = p((s_{i1}, \ldots, s_{iN}) \text{ sent}|(x_1, \ldots, x_N, n_r(t))$$

$$
\begin{aligned}
&= \frac{p((s_{i1},\ldots,s_{iN})\text{ sent},(x_1,\ldots,x_N),n_r(t))}{p((x_1,\ldots,x_N),n_r(t))} \\
&= \frac{p((s_{i1},\ldots,s_{iN})\text{ sent},(x_1,\ldots,x_N))p(n_r(t)}{p(x_1,\ldots,x_N)p(n_r(t))} \\
&= p((s_{i1},\ldots,s_{iN})\text{ sent}|(x_1,\ldots,x_N)), \qquad (5.25)\\
& \qquad (5.26)
\end{aligned}
$$

where the third equality follows from the fact that the $n_r(t)$ is independent of both $(x_1,\ldots,x_N)$ and of $(s_{i1},\ldots,s_{iN})$. This analysis shows that $(x_1,\ldots,x_N)$ is a sufficient statistic for $x(t)$ 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 $\mathbf{x}$ is a sufficient statistic for the received signal $x(t)$, we call $\mathbf{x}$ the **received vector** associated with $x(t)$.

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 $1-P_e = p(\hat{m}\text{ sent}|\mathbf{x})$. In other words, given a received vector $\mathbf{x}$, the optimal receiver selects $\hat{m}=m_i$ corresponding to the constellation $\mathbf{s}_i$ that satisfies $p(\mathbf{s}_i\text{ sent}|\mathbf{x}) > p(\mathbf{s}_j\text{ sent}|\mathbf{x})\forall j\neq i$. Let us define a set of **decisions regions** $(Z_1,\ldots,Z_M)$ that are subsets of the signal space $\mathcal{R}^N$ by

$$Z_i = (\mathbf{x} : p(\mathbf{s}_i\text{ sent}|\mathbf{x}) > p(\mathbf{s}_j\text{ sent}|\mathbf{x})\forall j\neq i). \qquad (5.27)$$

Clearly these regions do not overlap. Moreover, they partition the signal space assuming there is no $\mathbf{x}\in\mathcal{R}^N$ for which $p(\mathbf{s}_i\text{ sent}|\mathbf{x}) = p(\mathbf{s}_j\text{ sent}|\mathbf{x})$. If such points exist then the signal space is partitioned with decision regions by arbitrarily assigning such points to either decision region Z_i or Z_j. Once the signal space has been partitioned by decision regions, then for a received vector $\mathbf{x}\in Z_i$ the optimal receiver outputs the message estimate $\hat{m}=m_i$. Thus, the receiver processing consists of computing the received vector $\mathbf{x}$ from $x(t)$, finding which decision region Z_i contains $\mathbf{x}$, 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 $\mathbf{x}$ lies in region Z_1, so the receiver will output the message m_1 as the best message estimate given received vector $\mathbf{x}$.

We now examine the decision regions in more detail. We will abbreviate $p(\mathbf{s}_i\text{ sent}|\mathbf{x}\text{ received})$ as $p(\mathbf{s}_i|\mathbf{x})$ and $p(\mathbf{s}_i\text{ sent})$ as $p(\mathbf{s}_i)$. By Bayes rule,

$$p(\mathbf{s}_i|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{s}_i)p(\mathbf{s}_i)}{p(\mathbf{x})}. \qquad (5.28)$$

To minimize error probability, the receiver output $\hat{m}=m_i$ corresponds to the constellation $\mathbf{s}_i$ that maximizes $p(\mathbf{s}_i|\mathbf{x})$, i.e. $\mathbf{s}_i$ must satisfy

$$\arg\max_{\mathbf{s}_i}\frac{p(\mathbf{x}|\mathbf{s}_i)p(\mathbf{s}_i)}{p(\mathbf{x})} = \arg\max_{\mathbf{s}_i} p(\mathbf{x}|\mathbf{s}_i)p(\mathbf{s}_i), i=1,\ldots,M, \qquad (5.29)$$

where the second equality follows from the fact that $p(\mathbf{x})$ is not a function of $\mathbf{s}_i$. Assuming equally likely messages ($p(\mathbf{s}_i)=p(1/M)$), the receiver output $\hat{m}=m_i$ corresponding to the constellation $\mathbf{s}_i$ that satisfies

$$\arg\max_{\mathbf{s}_i} p(\mathbf{x}|\mathbf{s}_i), i=1,\ldots,M. \qquad (5.30)$$

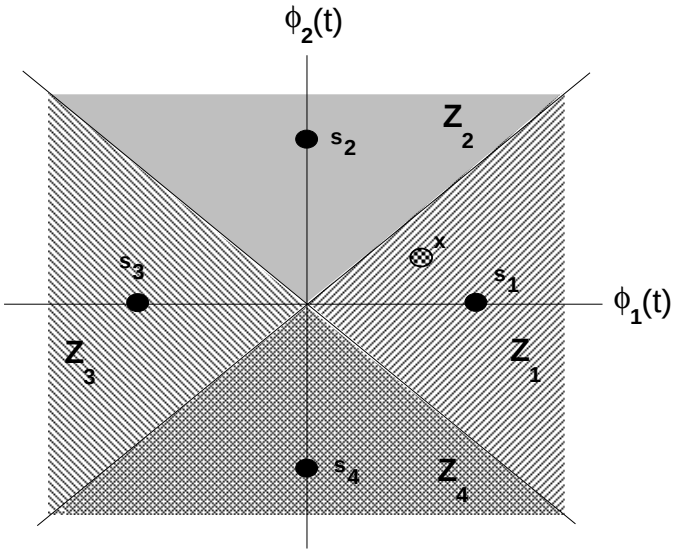


Figure 5.6: Decision Regions

Let us define the likelihood function associated with our receiver as

$$L(\mathbf{s}_i) = p(\mathbf{x}|\mathbf{s}_i). \tag{5.31}$$

Given a received vector $\mathbf{x}$, 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 the log likelihood function, defined as $l(\mathbf{s}_i) = \log L(\mathbf{s}_i)$. Using (5.24) for $L(\mathbf{s}_i) = p(\mathbf{x}|\mathbf{s}_i)$ then yields

$$l(\mathbf{s}_i) = -\frac{1}{N_0}\sum_{j=1}^{N}(x_j - s_{ij})^2 = ||\mathbf{x} - \mathbf{s}_i||^2. \tag{5.32}$$

Thus, the log likelihood function $l(\mathbf{s}_i)$ depends only on the distance between the received vector $\mathbf{x}$ and the constellation point $\mathbf{s}_i$.

The maximum likelihood receiver is implemented using the structure shown in Figure 5.4. First $\mathbf{x}$ is computed from $x(t)$, and then the signal constellation closest to $\mathbf{x}$ is determined as the constellation point $\mathbf{s}_i$ satisfying

$$\arg\max_{\mathbf{s}_i} \ -\frac{1}{N_0}\sum_{j=1}^{N}(x_j - s_{ij})^2 = \arg\max_{\mathbf{s}_i} \ -\frac{1}{N_0}||\mathbf{x} - \mathbf{s}_i||^2. \tag{5.33}$$

This $\mathbf{s}_i$ is determined from the decision region Z_i that contains $\mathbf{x}$, where Z_i is defined by

$$Z_i = (\mathbf{x} : ||\mathbf{x} - \mathbf{s}_i|| < ||\mathbf{x} - \mathbf{s}_j|| \ \forall j = 1, \ldots, M, j \neq i) \quad i = 1, \ldots, M. \tag{5.34}$$

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 very simple to implement since 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: to minimize error probability the decision regions Z_i must be modified to take into account the message probabilities, as indicated in (5.28).

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 function. 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, the output SNR is maximized. One can also show that the sampled matched filter outputs $(x_1, \ldots, x_n)$ in Figure 5.7 are the same as the $(x_1, \ldots, x_n)$ in Figure 5.4, so the two receivers are equivalent.

Example 5.2:
For BPSK modulation, find decision regions Z_1 and Z_2 corresponding to constellations $s_1 = A$ and $s_2 = -A$.

Solution: The signal space is one-dimensional, so $\mathbf{x} \in \mathcal{R}$. By (5.34) the decision region $Z_1 \subset \mathcal{R}$ is defined by

$$Z_1 = (\mathbf{x} : ||\mathbf{x} - A|| < ||\mathbf{x} - (-A)||) = (\mathbf{x} : \mathbf{x} > 0).$$

Thus, Z_1 contains all positive numbers on the real line. Similarly

$$Z_2 = (\mathbf{x} : ||\mathbf{x} - (-A)|| < ||\mathbf{x} - A||) = (\mathbf{x} : \mathbf{x} < 0).$$

So Z_2 contains all negative numbers on the real line. For $\mathbf{x} = 0$ the distance is the same to $s_1 = A$ and $s_2 = -A$ so we arbitrarily assign $\mathbf{x} = 0$ to Z_2.

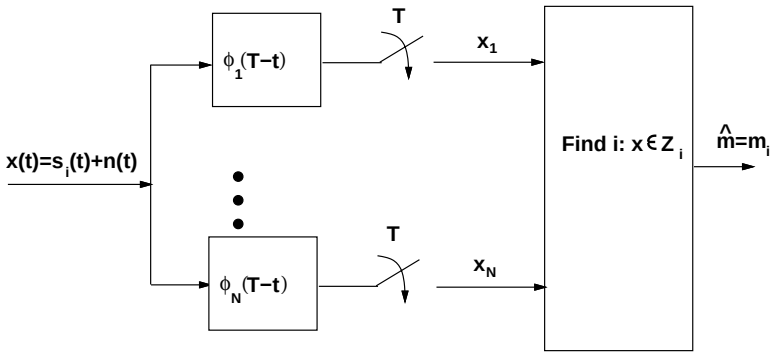


Figure 5.7: Matched Filter Receiver Structure.

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

$$P_e = \sum_{i=1}^{M} p(\mathbf{x} \notin Z_i | m_i \text{ sent}) p(m_i \text{ sent})$$

$$
\begin{aligned}
&= \frac{1}{M}\sum_{i=1}^{M} p(\mathbf{x} \notin Z_i | m_i \text{ sent}) \\
&= 1 - \frac{1}{M}\sum_{i=1}^{M} p(\mathbf{x} \in Z_i | m_i \text{ sent}) \\
&= 1 - \frac{1}{M}\sum_{i=1}^{M} \int_{Z_i} p(\mathbf{x}|m_i) d\mathbf{x} \\
&= 1 - \frac{1}{M}\sum_{i=1}^{M} \int_{Z_i} p(\mathbf{x} = \mathbf{s}_i + \mathbf{n}|\mathbf{s}_i) d\mathbf{n}. \\
&= 1 - \frac{1}{M}\sum_{i=1}^{M} \int_{Z_i - \mathbf{s}_i} p(\mathbf{n}) d\mathbf{n} \qquad (5.35)
\end{aligned}
$$

The integrals in (5.35) are over the N-dimensional subset $Z_i \subset \mathcal{R}^N$. We illustrate this error probability calculation in Figure 5.8, where the constellation points $s_1, \ldots, s_8$ are equally spaced around a circle with minimum separation d_{min}. The probability of correct reception assuming the first symbol is sent, $p(\mathbf{x} \in Z_1 | m_1 \text{ sent})$, corresponds to the probability $p(\mathbf{x} = \mathbf{s}_1 + \mathbf{n}|\mathbf{s}_1)$ that when noise is added to the transmitted constellation $\mathbf{s}_1$, the resulting vector $\mathbf{x} = \mathbf{s}_1 + \mathbf{n}$ remains in the Z_1 region shown by the shaded area.

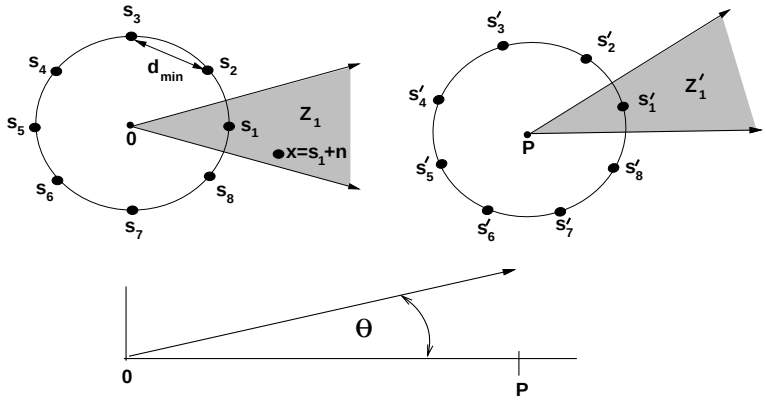


Figure 5.8: Error Probability Integral and Its Rotational/Shift Invariance

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 θ and axis shift of P relative to the constellation on the left side. Thus, $\mathbf{s}'_i = \mathbf{s}_i e^{j\theta} + P$. 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, thus the polar representation $\mathbf{n} = \mathbf{r}e^{j\theta}$ has θ 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 $P \in \mathcal{R}^{\mathcal{N}}$, the decision regions defined by (5.34) are also shifted by P. Let $(\mathbf{s}_i, Z_i)$ denote a constellation point and corresponding decision region before the shift and $(\mathbf{s}'_i, Z'_i)$ denote the corresponding constellation point and decision region after the shift. It is then straightforward to show that $p(\mathbf{x} = \mathbf{s}_i + \mathbf{n} \in Z_i | s_i) = p(\mathbf{x}' = \mathbf{s}'_i + \mathbf{n} \in Z'_i | \mathbf{s}'_i)$. Thus, the error probability after an axis shift of the constellation points will remain unchanged.

While (5.35) gives an exact solution to the probability of error, we cannot solve for this error prob-

ability 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 $||\mathbf{x} - \mathbf{s}_k|| < ||\mathbf{x} - \mathbf{s}_i||$ given that the constellation point $\mathbf{s}_i$ was sent. If the event A_{ik} occurs, then the constellation will be decoded in error since the transmitted constellation $\mathbf{s}_i$ is not the closest constellation point to the received vector $\mathbf{x}$. 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 $||\mathbf{x} - \mathbf{s}_l|| < ||\mathbf{x} - \mathbf{s}_k|| < ||\mathbf{x} - \mathbf{s}_i||$. The constellation is decoded correctly if $||\mathbf{x} - \mathbf{s}_i|| < ||\mathbf{x} - \mathbf{s}_k|| \; \forall k \neq i$. Thus

$$P_e(m_i \text{ sent}) = p\left(\bigcup_{\substack{k=1 \\ k\neq i}}^{M} A_{ik}\right) \leq \sum_{\substack{k=1 \\ k\neq i}}^{M} p\left(A_{ik}\right), \tag{5.36}$$

where the inequality follows from the union bound on probability.

Let us now consider $p(A_{ik})$ more closely. We have

$$\begin{aligned} p(A_{ik}) &= p(||\mathbf{s}_k - \mathbf{x}|| < ||\mathbf{s}_i - \mathbf{x}|| \; | \mathbf{s}_i \text{ sent}) \\ &= p(||\mathbf{s}_k - (\mathbf{s}_i + \mathbf{n})|| < ||\mathbf{s}_i - (\mathbf{s}_i + \mathbf{n})||) \\ &= p(||\mathbf{n} + \mathbf{s}_i - \mathbf{s}_k|| < ||\mathbf{n}||), \end{aligned} \tag{5.37}$$

i.e. 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 $\mathbf{s}_i - \mathbf{s}_j$, 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 and variance $N_0/2$. The event A_{ik} occurs if n is closer to $\mathbf{s}_i - \mathbf{s}_k$ than to zero, i.e. if $n > d_{ik}/2$, where $d_{ik} = ||\mathbf{s}_i - \mathbf{s}_k||$ equals the distances between constellation points $\mathbf{s}_i$ and $\mathbf{s}_k$. Thus,

$$p(A_{ik}) = p(n > d_{ik}/2) = \int_{d_{ik}/2}^{\infty} \frac{1}{\sqrt{\pi N_o}} \exp\left[\frac{-v^2}{N_0}\right] dv = Q\left(\frac{d_{ij}}{\sqrt{2N_0}}\right). \tag{5.38}$$

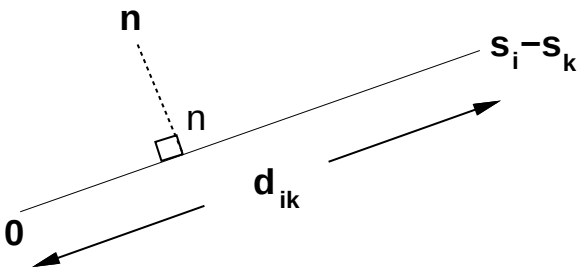


Figure 5.9: Noise Projecion

Substituting into (5.36) we get

$$P_e(m_i \text{ sent}) \leq \sum_{\substack{k=1 \\ k\neq i}}^{M} Q\left(\frac{d_{ij}}{\sqrt{2N_0}}\right). \tag{5.39}$$

Summing over all possible messages yields the **union bound**

$$P_e = \sum_{i=1}^{M} p(m_i) P_e(m_i \text{ sent}) \leq \frac{1}{M} \sum_{i=1}^{M} \sum_{\substack{k=1 \\ k \neq i}}^{M} Q\left(\frac{d_{ij}}{\sqrt{2N_0}}\right), \tag{5.40}$$

where the Q function, $Q(z)$, is defined as the probability that a Gaussian random variable x with mean 0 and variance 1 is bigger than z:

$$Q(z) = p(x > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \tag{5.41}$$

The Q function cannot be solved for in closed form. It can be obtained from the complementary error function as

$$Q(z) = \frac{1}{2} \text{erfc}\left(\frac{z}{\sqrt{2}}\right). \tag{5.42}$$

Defining the *minimum distance* of the constellation as $d_{min} = \min_{i,k} d_{ik}$, we can simplify (5.40) with the looser bound

$$P_e \leq (M-1) Q\left(\frac{d_{min}}{\sqrt{2N_0}}\right). \tag{5.43}$$

Using a well-known bound for the Q function yields a closed-form bound

$$P_e \leq \frac{M-1}{\sqrt{\pi}} \exp\left[\frac{-d_{min}^2}{4N_0}\right]. \tag{5.44}$$

Finally, P_e is sometimes approximated as the probability of error associated with constellations at the minimum distance d_{min} multiplied by the number of neighbors at this distance $M_{d_{min}}$:

$$P_e \approx M_{d_{min}} Q\left(\frac{d_{min}}{\sqrt{2N_0}}\right). \tag{5.45}$$

This approximation is called the **nearest neighbor approximation** to P_s. 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 (5.43) since $M \geq M_{d_{min}}$, and will also be slightly less than the union bound (5.40), since this 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 $x > y$, $Q(x) << Q(y)$ due to the exponential falloff of the Gaussian distribution in (5.41). 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 (5.45) is made in [4] and also referenced in [5]. Moreover, [4] indicates that (5.45) 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 bound 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 $\mathcal{R}^2$ 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 (5.43), closed

form bound (5.44), and nearest neighbor approximation (5.45) 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 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, $P_e(m_i \text{ sent}) = P_e(m_j \text{ sent}), j \neq i$, so the union bound simplifies to

$$P_e \leq \sum_{j=2}^{4} Q\left(\frac{d_{1j}}{\sqrt{2N_0}}\right) = 2Q(A/\sqrt{N_0}) + Q(\sqrt{2}A/\sqrt{N_0}) = 2Q(4) + Q(\sqrt{32}) = 3.1679 * 10^{-5}.$$

The looser bound yields

$$P_e \leq 3Q(4) = 9.5014 * 10^{-5}$$

which is roughly a factor of 3 looser than the union bound. The closed-form bound yields

$$P_e \leq \frac{3}{\pi}\exp\left[\frac{-.5A^2}{N_0}\right]. = 3.2034 * 10^{-4},$$

which differs from the union bound by about an order of magnitude. Finally, the nearest neighbor approximation yields

$$P_e \approx 2Q(4) = 3.1671 * 10^{-5},$$

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 (5.43) is exact:

$$P_b = Q\left(\frac{d_{min}}{\sqrt{2N_0}}\right). \tag{5.46}$$

The minimum distance squared in (5.44) and (5.46) 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: $P_e = p(\hat{m} \neq m_i | m_i \text{ sent})$, where m_i corresponds to a message with $\log_2 M$ bits. However, system designers are typically more interested in the bit error probability than in the symbol error probability, since 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 $\log_2 M$ possible bit sequences to messages $m_i, i = 1, \ldots, M$ so that a symbol 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, assuming that mistaking a signal constellation for a constellation other than its nearest neighbors has a very low probability, we can make the approximation

$$P_b \approx \frac{P_e}{\log_2 M}. \tag{5.47}$$

The most common form of mapping with the property is called Gray coding, which 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 below.

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

$$s(t) = \alpha(t)\cos[2\pi(f_c + f(t))t + \theta(t) + \phi_0] = \alpha(t)\cos(2\pi f_c t + \phi(t) + \phi_0), \tag{5.48}$$

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 (5.48) in terms of its in-phase and quadrature components as:

$$s(t) = \alpha(t)\cos\phi(t)\cos(2\pi f_c t+\phi_0) - \alpha(t)\sin\phi(t)\sin(2\pi f_c t+\phi_0) = s_I(t)\cos(2\pi f_c t+\phi_0) - s_Q(t)\sin(2\pi f_c t+\phi_0), \tag{5.49}$$

where $s_I(t) = \alpha(t)\cos\phi(t)$ is called the in-phase component of $s(t)$ and $s_Q(t) = \alpha(t)\sin\phi(t)$ is called its quadrature component. We can also write $s(t)$ in its complex baseband representation as

$$s(t) = \Re\{u(t)e^{j(2\pi f_c t+\phi_0)}\}, \tag{5.50}$$

where $u(t) = s_I(t) + js_Q(t)$. This representation, described in more detail in Appendix A, is useful since 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 $s(t) = s_I(t)\cos(2\pi f_c t + \phi_0) - s_Q(t)\sin(2\pi f_c t + \phi_0)$ 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. 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 $.5 \le K_g \le 1$, 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 $1/T_s$. Since the pulse shape $g(t)$ is fixed, the signal constellation for amplitude and phase modulation is defined based on the constellation point: $(s_{i1}, s_{i2}) \in \mathcal{R}^2, i = 1, \ldots, M$. The complex baseband representation of $s(t)$ is

$$s(t) = \Re\{u(t)e^{j(2\pi f_c t+\phi_0)}\} \tag{5.51}$$

where $u(t) = s_I(t) + js_Q(t) = (s_{i1} + js_{i2})g(t)$. The constellation point $\mathbf{s}_i = (s_{i1}, s_{i2})$ 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 $K = \log_2 M$, signal constellation $(s_{i1}, s_{i2}) \in \mathcal{R}^2, i = 1, \ldots, M$, and choice of pulse shape $g(t)$ determines the digital modulation design. The pulse shape $g(t)$ is designed to improve spectral efficiency and combat ISI, as discussed in Section 5.5 below.

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], which is called **coherent detection**. If $\phi - \phi_0 = \Delta\phi \neq 0$ then the in-phase branch will have an unwanted term associated with the quadrature branch and vice versa, i.e. $x_1 = s_{i1}\cos(\Delta\phi) + s_{i2}\sin(\Delta\phi) + n_1$ and $x_2 = s_{i1}\sin(\Delta\phi) + s_{i2}\cos(\Delta\phi) + n_2$, 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.

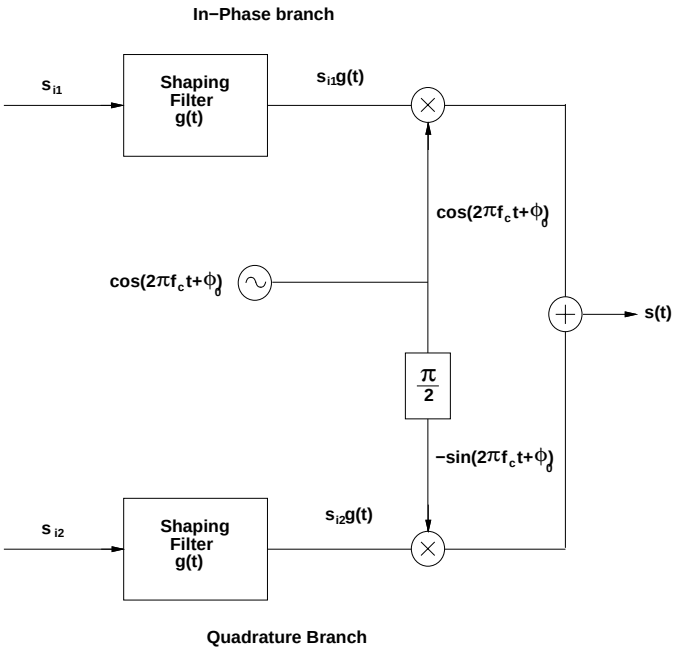


Figure 5.10: Amplitude/Phase Modulator.

[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.

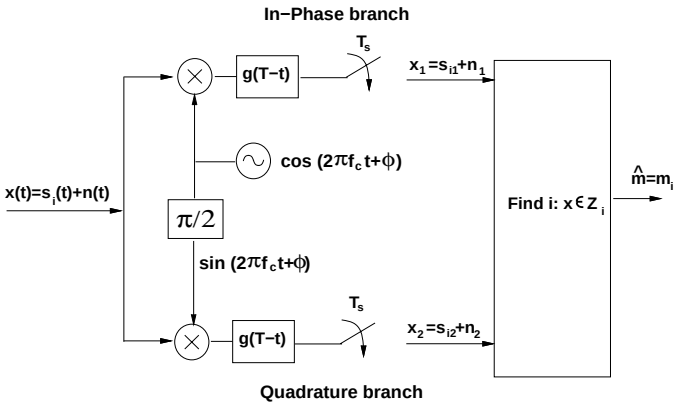


Figure 5.11: Amplitude/Phase Demodulator (Coherent: $\phi = \phi_0$).

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

$$s_i(t) = \Re\{A_i g(t) e^{j2\pi f_c t}\} = A_i g(t)\cos(2\pi f_c t), \quad 0 \le t \le T_s >> 1/f_c, \tag{5.52}$$

where $A_i = (2i-1-M)d, i = 1, 2, \ldots, M$ defines the signal constellation, parameterized by the distance d which is typically a function of the signal energy, and $g(t)$ is the pulse shape satisfying (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.

Over each symbol period the MPAM signal associated with the ith constellation has energy

$$E_{s_i} = \int_0^{T_s} s_i^2(t)dt = \int_0^{T_s} A_i^2 g^2(t)\cos^2(2\pi f_c t)dt = A_i^2 \tag{5.53}$$

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

$$\overline{E_s} = \frac{1}{M}\sum_{i=1}^{M} A_i^2. \tag{5.54}$$

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), this results in only a single bit error in the sequence of K bits.

Example 5.4:

[2]Recall from (5.8) that (5.12) and therefore (5.53) are not necessarily exact equalities, but very good approximations for $f_c T_s >> 1$.

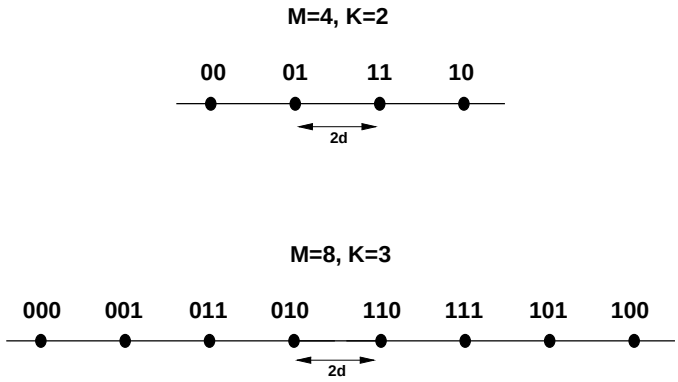


Figure 5.12: Gray Encoding for MPAM.

For $g(t) = \sqrt{2/T_s}, 0 \leq t < T_s$ a rectangular pulse shape, find the average energy of 4PAM modulation.

Solution: For 4PAM the A_i values are $A_i = \{-3d, -d, d, 3d\}$, so the average energy is

$$\overline{E_s} = \frac{d^2}{8}(9 + 1 + 1 + 9) = 2.5d^2.$$

The decision regions $Z_i, i = 1, \ldots, M$ associated with the 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

$$Z_i = \begin{cases} (-\infty, A_i + d) & i = 1 \\ [A, A_i) & 2 \leq i \leq M - 1 \\ [A_i - d, \infty) & i = M \end{cases}$$

From (5.52) 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 x to a decision region Z_i and outputs the corresponding bit sequence $\hat{m} = m_i = \{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$.

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 is given by

$$\begin{aligned} s_i(t) &= \Re\{Ag(t)e^{j2\pi(i-1)/M}e^{j2\pi f_c t}\}, \quad 0 \leq t \leq T_s \\ &= Ag(t)\cos\left[2\pi f_c t + \frac{2\pi(i-1)}{M}\right] \\ &= Ag(t)\cos\left[\frac{2\pi(i-1)}{M}\right]\cos 2\pi f_c t - Ag(t)\sin\left[\frac{2\pi(i-1)}{M}\right]\sin 2\pi f_c t. \end{aligned} \tag{5.55}$$

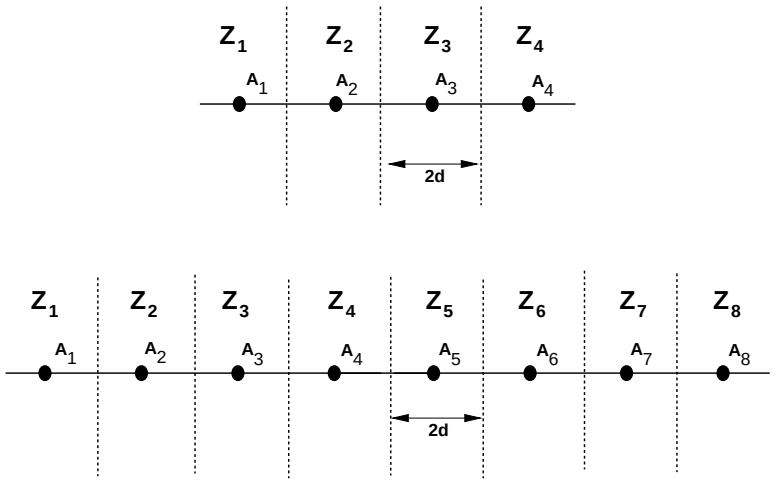


Figure 5.13: Decision Regions for MPAM

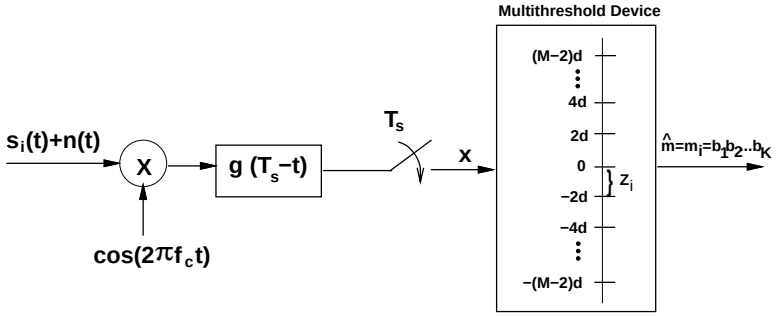


Figure 5.14: Coherent Demodulator for MPAM

Thus, the constellation points or symbols (s_{i1}, s_{i2}) are given by $s_{i1} = A\cos[\frac{2\pi(i-1)}{M}]$ and $s_{i2} = A\sin[\frac{2\pi(i-1)}{M}]$ for $i = 1, \ldots, M$. The pulse shape $g(t)$ satisfies (5.12) and (5.13), and $\theta_i = \frac{2\pi(i-1)}{M}, i = 1, 2, \ldots, M = 2^K$ are the different phases in the signal constellation points that convey the information bits. The minimum distance between constellation points is $d_{min} = A\sin(2\pi/M)$, where A is typically a function of the signal energy. 2PSK is often referred to as binary PSK or BPSK, while 4PSK is often called quadrature phase shift keying (QPSK), and is the same as MQAM with $M = 4$ which is defined below.

All possible transmitted signals $s_i(t)$ have equal energy:

$$E_{s_i} = \int_0^{T_s} s_i^2(t)dt = A^2 \tag{5.56}$$

Note that for $g(t) = \sqrt{2/T_s}, 0 \le t \le T_s$, 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 make MPSK nonconstant 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, as illustrated in 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 $M = 8$ are shown in Figure 5.16. If

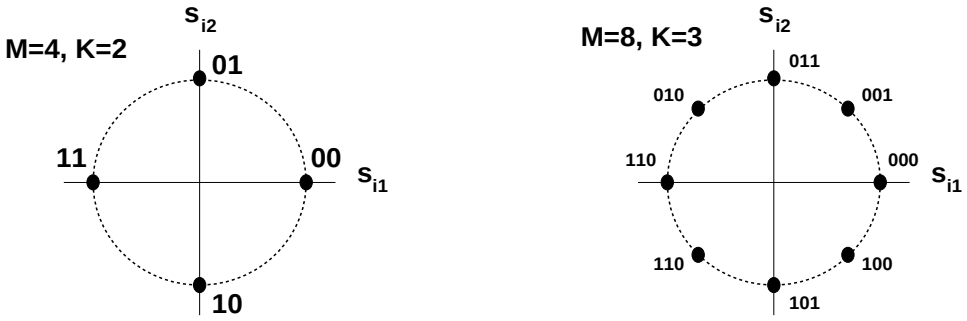


Figure 5.15: Gray Encoding for MPSK.

we represent $\mathbf{x} = re^{j\theta} \in \mathcal{R}^2$ in polar coordinates then these decision regions for any M are defined by

$$Z_i = \{re^{j\theta} : 2\pi(i - .5)/M \leq \theta < 2\pi(i + .5)/M\}. \tag{5.57}$$

From (5.55) 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 $Z_1 = (\mathbf{x} : \mathbf{x} > 0)$ and $Z_2 = (\mathbf{x} : \mathbf{x} \leq 0)$. 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 x to the positive or negative half of the real line, and 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 point $s_1 = A$ and the message corresponding to a bit value of 0, $m_2 = 0$, is mapped to the constellation point $s_2 = -A$.

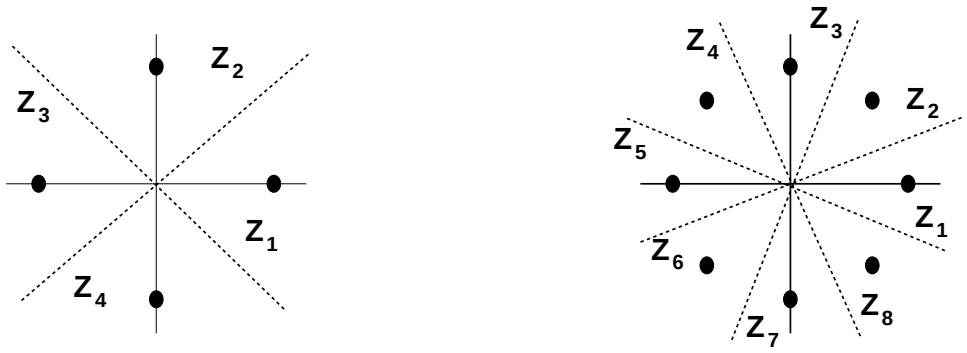


Figure 5.16: Decision Regions for MPSK

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.

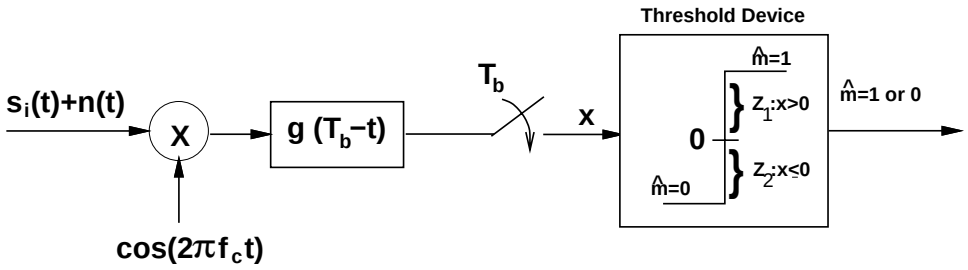


Figure 5.17: Coherent Demodulator for BPSK.

The transmitted signal is given by

$$s_i(t) = \Re\{A_i e^{j\theta_i} g(t) e^{j2\pi f_c t}\} = A_i \cos(\theta_i) g(t) \cos(2\pi f_c t) - A_i \sin(\theta_i) g(t) \sin(2\pi f_c t), \quad 0 \le t \le T_s. \quad (5.58)$$

where the pulse shape $g(t)$ satisfies (5.12) and (5.13). The energy in $s_i(t)$ is

$$E_{s_i} = \int_0^{T_s} s_i^2(t) = A_i^2, \quad (5.59)$$

the same as for MPAM. The distance between any pair of symbols in the signal constellation is

$$d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j|| = \sqrt{(s_{i1} - s_{k1})^2 + (s_{i2} - s_{k2})^2}. \quad (5.60)$$

For square signal constellations, where s_{i1} and s_{i2} take values on $(2i-1-L)d, i = 1, 2, \ldots, L = 2^l$, 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 4QAM and 16QAM, which are shown in Figure 5.18 below. These square constellations have $M = 2^{2l} = L^2$ constellation points, which are used to send $2l$ bits/symbol, or l bits per dimension. It can be shown that the average power of a square signal constellation with l bits per dimension, S_l, is proportional to $4^l/3$, and it follows that the average power for one more bit per dimension $S_{l+1} \approx 4S_l$. 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 (5.58) 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. Thus, these modulation techniques require coherent demodulation, i.e. 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 dicussed in more detail in Section 5.6. Due to the difficulties as well as the cost and complexity associated with carrier phase recovery, differential modulation techniques, which do

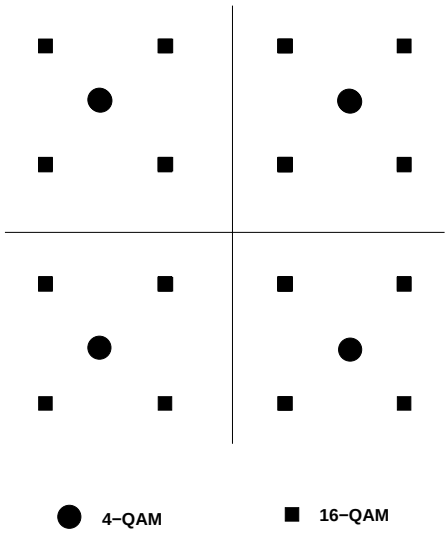


Figure 5.18: 4QAM and 16QAM Constellations.

not require a coherent phase reference at the receiver, are generally preferred to coherent modulation for wireless applications.

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* 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 $\theta(k-1) = e^{j\theta_i}$, $\theta_i = 0, \pi$, 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 4PSK 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, the corresponding symbol would have phase $\theta(k) = e^{j\theta_i}$, i.e. 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 $\theta(k) = e^{j(\theta_i - \pi/2)}$, 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 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

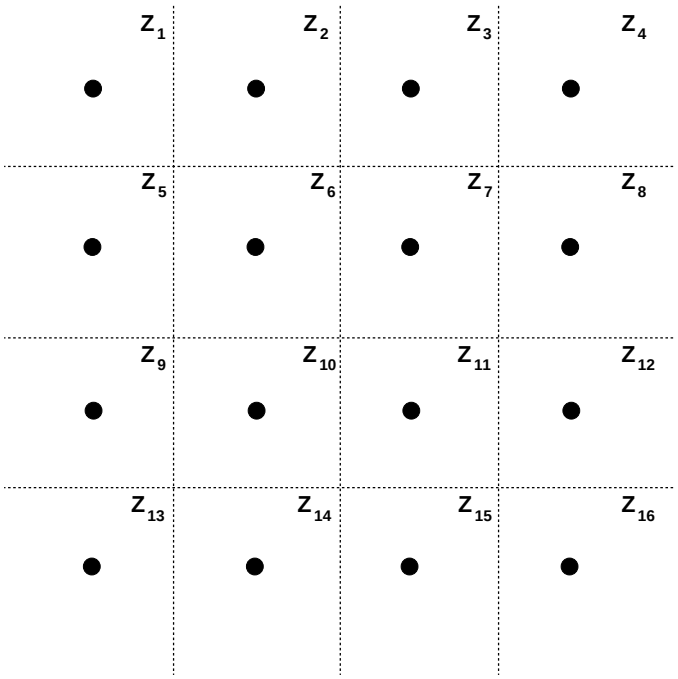


Figure 5.19: Decision Regions for MQAM with $M = 16$

either 01 or 10, i.e. 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 π/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 D-MPSK, and for BPSK and QPSK this becomes DPSK and D-QPSK, respectively.

Bit Sequence	Phase Transition
00	0
01	$\pi/2$
10	$-\pi/2$
11	π

Table 5.1: Mapping for D-QPSK with Gray Encoding

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, re-

sults 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 $\mathbf{s}(k) = Ae^{j\theta(k)+\phi_0}$ The received vector associated with the sampler outputs is

$$\mathbf{z}(k) = x_1(k) + jx_2(k) = Ae^{j\theta(k)+\phi_0} + n(k), \tag{5.61}$$

where $n(k)$ is complex white Gaussian noise. The received vector at the previous time sample $k-1$ is thus

$$\mathbf{z}(k-1) = x_1(k-1) + jx_2(k-1) = Ae^{j\theta(k-1)+\phi_0} + n(k-1). \tag{5.62}$$

The phase difference between $\mathbf{z}(k)$ and $\mathbf{z}(k-1)$ dictates which symbol was transmitted. Consider

$$\mathbf{z}(k)\mathbf{z}^*(k-1) = A^2e^{j(\theta(k)-\theta(k-1)} + Ae^{j\theta(k)+\phi_0}n^*(k-1) + Ae^{-j\theta(k-1)+\phi_0}n(k) + n(k)n^*(k-1). \tag{5.63}$$

In the absence of noise ($n(k) = n(k-1) = 0$) only the first term in (5.63) 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.

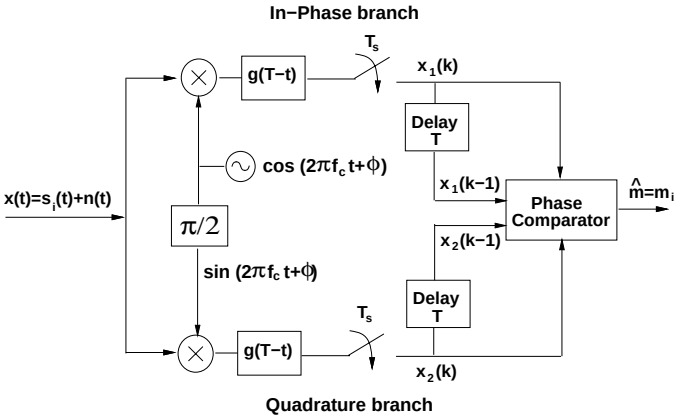


Figure 5.20: Differential PSK Demodulator.

Differential modulation is less sensitive to a random drift in the carrier phase. However, if the channel has a nonzero Doppler frequency, the signal phase can decorrelate between symbol times, making the previous symbol a very 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.

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 expense of increased complexity in the constellation map [17]. The optimal constellation shape is a sphere in N-dimensional space, which must be mapped to a sequence of constellations in 2-dimensional space in order to be generated by the modulator shown in Figure 5.10. The general conclusion in [17] 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, as well as constellations that allow a noninteger number of bits per symbol, can be found in [17].

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 degrees, 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. These abrupt transitions are avoided by offsetting the quadrature branch pulse $g(t)$ by half a symbol period, as shown in Figure 5.21. This offset makes the signal less sensitive to distortion during symbol transitions.

Phase modulation with phase offset is usually abbreviated as O-MPSK, where the O indicates the offset. For example, QPSK modulation with quadrature offset is referred to as O-QPSK. O-QPSK has the same spectral properties as QPSK for linear amplification, but has higher spectral efficiency under nonlinear amplification, since the maximum phase transition of the signal is 90 degrees, 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 degree phase shift used in the IS-54 standard for digital cellular is $\pi/4$-QPSK [12]. This technique allows for a maximum phase transition of 135 degrees, versus 90 degrees for offset QPSK and 180 degrees for QPSK. Thus, $\pi/4$-QPSK does not have as good spectral properties as O-QPSK under nonlinear amplification. However, $\pi/4$-QPSK can be differentially encoded, eliminating 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, each symbol time $K = \log_2 M$ bits are encoded into the frequency of the transmitted signal $s(t), 0 \le t < T_s$, 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 orthogonality of the basis functions requires a minimum separation between different carrier frequencies of $\Delta f = \min_{ij} |f_j - f_i| = .5/T_s$.

Since frequency modulation encodes information in the signal frequency, the transmitted signal $s(t)$ has a constant envelope A. Because the signal is constant envelope, nonlinear amplifiers can be used with high power efficiency, and 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

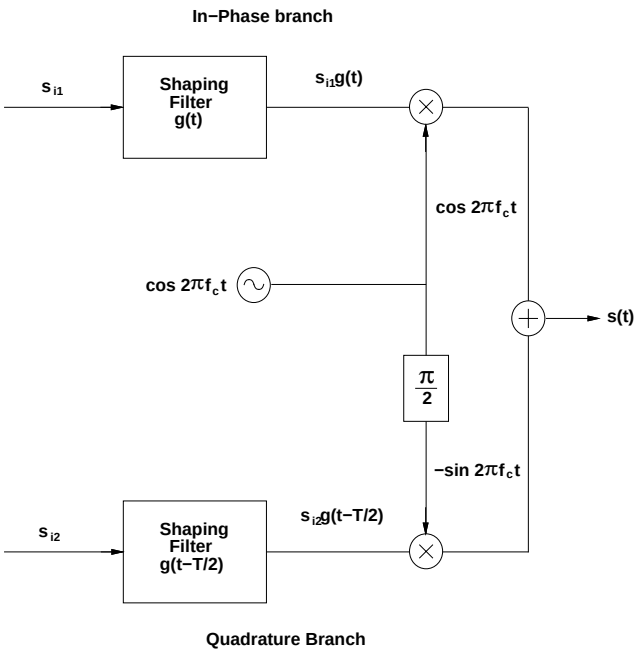


Figure 5.21: Modulator with Quadrature Offset.

and phase modulation techniques described in Section 5.3.

In it's 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 that the jth carrier signal be matched in phase to the jth carrier signal at the transmitter, 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 that 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 below.

5.4.1 Frequency Shift Keying (FSK) and Minimum Shift Keying (MSK)

In MFSK the modulated signal is given

$$s_i(t) = A\cos[2\pi f_c t + 2\pi\alpha_i \Delta f_c t + \phi_i], \quad 0 \leq t < T_s, \tag{5.64}$$

where $\alpha_i = (2i - 1 - M), i = 1, 2, \ldots, M = 2^K$. The minimum frequency separation between FSK carriers is thus $2\Delta f_c$. MFSK consists of M basis functions $\phi_i(t) = \sqrt{2/T_s}\cos[2\pi f_c t + 2\pi\alpha_i \Delta f_c t + \phi_i]$, where the $\sqrt{2/T_s}$ is a normalization factor to insure that $\int_0^{T_s} \phi_i^2(t) = 1$. Over a 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 $f_i = f_c + \alpha_i \Delta f_c$ and the modulator switches between these different oscillators each symbol time T_s. However, with this implementation there will be a discontinuous phase

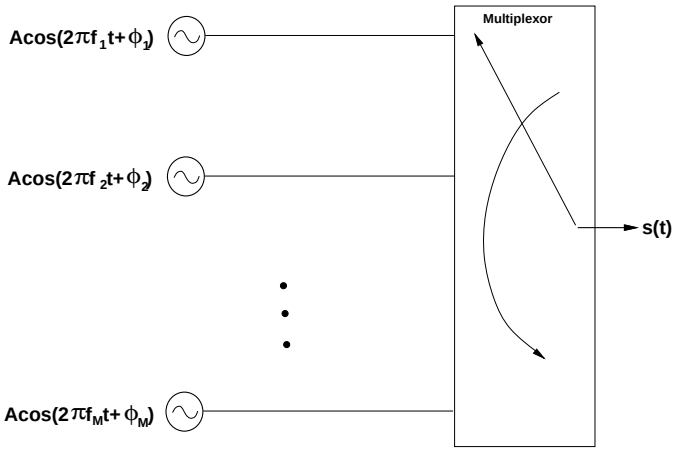


Figure 5.22: Frequency Modulator.

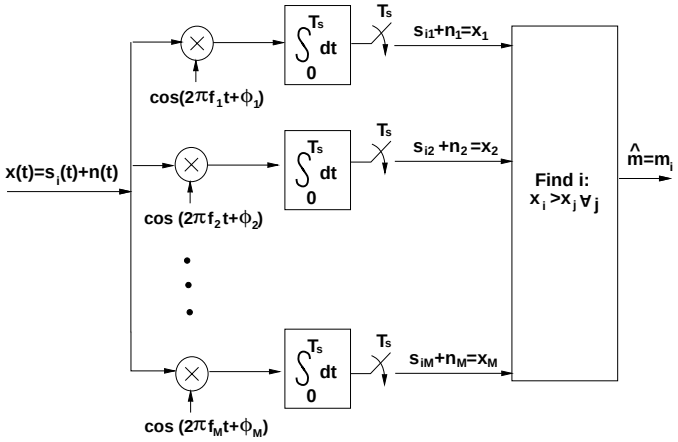


Figure 5.23: Frequency Demodulator (Coherent)

transition at the switching times due to phase offsets between the oscillators. This discontinuous phase leads to a spectral broadening, which is undesirable. An FSK modulator that maintains continuous phase is discussed in the next section. Coherent detection of MFSK uses the standard structure of Figure 5.4. 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 FSK where the minimum frequency separation is $2\Delta f_c = .5/T_s$. Note that this is the minimum frequency separation so that $< s_i(t), s_j(t) >= 0$ over a symbol time, for $i \neq j$. Since signal orthogonality is required for demodulation, $2\Delta f_c = .5/T_s$ is the minimum possible frequency separation in FSK, and therefore it occupies the minimum bandwidth.

5.4.2 Continuous-Phase FSK (CPFSK)

A better way to generate MFSK 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

$$s_i(t) = A\cos\left[2\pi f_c t + 2\pi\beta\int_{-\infty}^{t} u(\tau)d\tau\right] = A\cos[2\pi f_c t + \theta(t)], \tag{5.65}$$

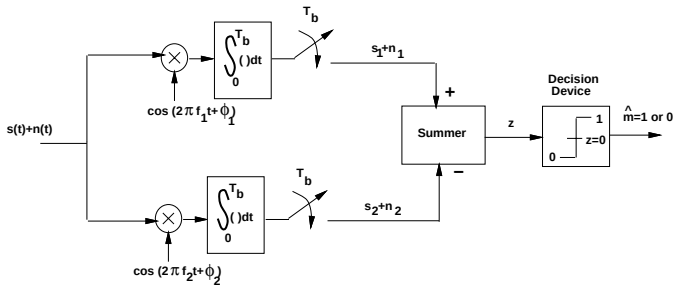


Figure 5.24: Demodulator for FSK

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 [1], for β small the transmission bandwidth of $s(t)$ is approximately

$$B_s \approx M\Delta f_c + 2B_g, \tag{5.66}$$

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 occupancy of a CPFSK-modulated signal is larger than that of a linearly modulated signal by $M\Delta f_c \geq .5M/T_s$. The spectral efficiency penalty of CPFSK relative to linear modulation increases with data rate, in particular with the number of 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 therefore it is optimal to detect all symbols simultaneously. However, sequence detection can be impractical due to the memory and computational requirements associated with making decisions based on sequences of symbols. Details on both symbol-by-symbol and sequence detectors for coherent demodulation of CPFSK can be found in [10, Chapter 5.3].

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 by detecting the energy of the signal at each frequency and, if the ith branch has the highest energy of all branches, then the receiver outputs message m_i. The modified receiver is shown in Figure 5.25.

Suppose the transmitted signal corresponds to frequency f_i:

$$s(t) = A\cos(2\pi f_i t + \phi_i) = A\cos(\phi_i)\cos(2\pi f_i t) - A\sin(\phi_i)\sin(2\pi f_i t), \quad 0 \leq t < T_s. \tag{5.67}$$

The phase ϕ_i represents the phase offset between the transmitter and receiver oscillators at frequency f_i. The coherent receiver in Figure 5.23 only detects 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 carrier frequency $f_j, j = 1, \ldots, M$, 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

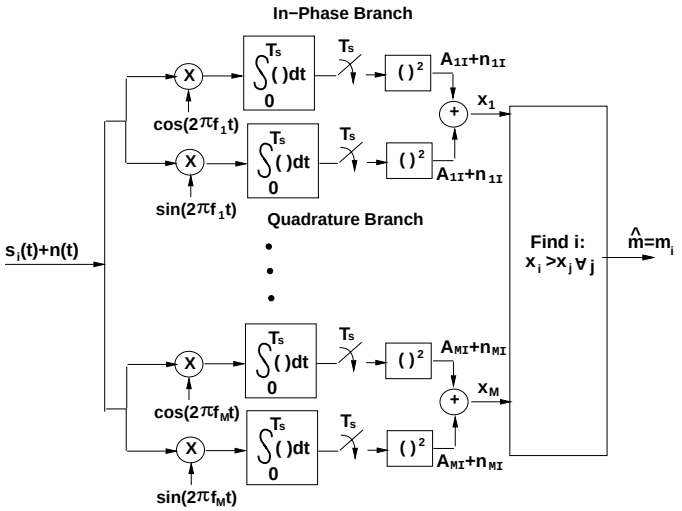


Figure 5.25: Noncoherent FSK Demodulator

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 $i = j$, $A_{jI} = A^2 \cos(\phi_i)$ and $A_{jQ} = A^2 \sin(\phi_i)$. 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 $A^2 \cos(\phi_i) + A^2 \sin(\phi_i) = A^2$, 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. A similar structure where each branch consists of a filter matched to the carrier frequency followed by an envelope detector and sampler can also be used [2, Chapter 6.8]. Note that the noncoherent receiver of Figure 5.25 still requires accurate synchronization for sampling. Synchronization issues are discussed in Section 5.6.

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 very high spectral sidelobes, which means that signals must use a larger bandwidth to eliminate some of the adjacent channel sidelobe energy. Pulse shaping is a method to reduce sidelobe energy relative to a rectangular pulse, however the shaping must be done in such a way that intersymbol interference (ISI) 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 $g^*(-t)$, so 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 11 we 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 equals zero at the ideal sampling point associated with past or future symbols:

$$p(kT_s) = \begin{cases} p_0 = p(0) & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{5.68}$$

In the frequency domain this translates to

$$\sum_{l=-\infty}^{\infty} P(f + l/T_s) = p_0 T_s. \tag{5.69}$$

The following pulse shapes all satisfy the Nyquist criterion.

1. Rectangular pulses: $g(t) = \sqrt{2/T_s}, 0 \le t \le T_s$, which yields the triangular effective pulse shape

$$p(t) = \begin{cases} 2 + 2t/T_s & -T_s \le t < 0 \\ 2 - 2t/T_s & 0 \le t < T_s \\ 0 & \text{else} \end{cases}$$

   This pulse shape leads to constant envelope signals in MPSK, but has lousy spectral properties due to its high sidelobes.

2. Cosine pulses: $p(t) = \sin \pi t/T_s, 0 \le t \le T_s$. Cosine pulses are mostly used in MSK modulation, where the quadrature branch of the PSK modulation has its pulse shifted 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 $z(t)$ is first specified relative to its Fourier Transform:

$$P(f) = \begin{cases} T_s & 0 \le |f| \le (1-\beta)/2T_s \\ \frac{T_s}{2}\left[1 - \sin\frac{\pi T_s}{\beta}\left(f - \frac{1}{2T_s}\right)\right] & (1-\beta)/2T_s \le |f| \le (1+\beta)/2T_s \end{cases},$$

   where β is defined as the rolloff factor, which determines the rate of spectral rolloff, as shown in Figure 5.26. Setting $\beta = 0$ yields a rectangular pulse. The pulse $f(t)$ in the time domain corresponding to $F(f)$ is

$$p(t) = \frac{\sin \pi t/T_s}{\pi t/T_s} \frac{\cos \beta \pi t/T_s}{1 - 4\beta^2 t^2/T_s^2}.$$

Both time and frequency domain properties of the Raised Cosine pulse are shown in Figure 5.26. 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 intersymbol 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 it decays 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 intersymbol interference 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 FSK is the Gaussian pulse shape, defined as

$$g(t) = \frac{\sqrt{\pi}}{\alpha} e^{-\pi^2 t^2/\alpha^2}, \tag{5.70}$$

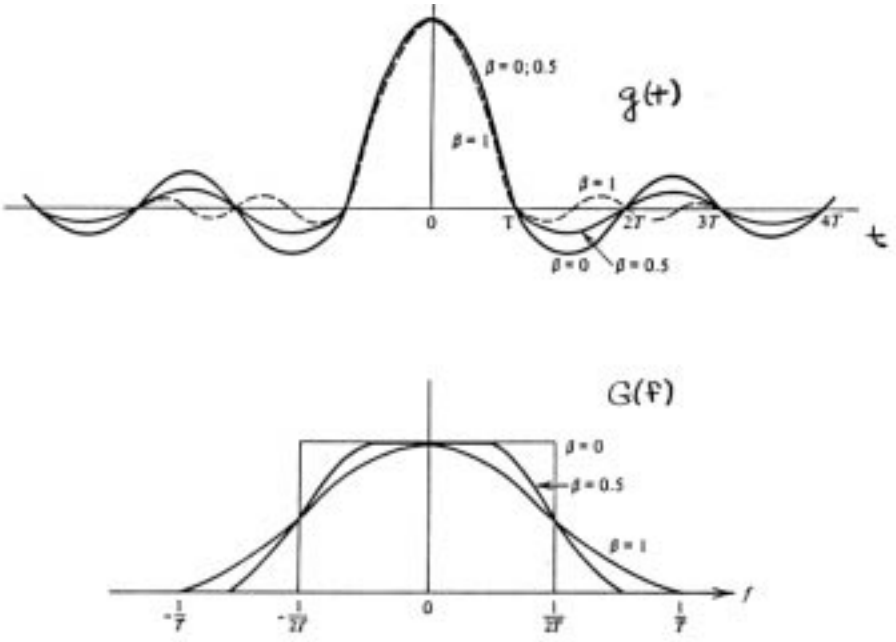


Figure 5.26: Spectral and Time-Domain Properties of the Raised Cosine Pulse.

where α is a parameter that dictates spectral efficiency. The spectrum of $g(t)$, which dictates the spectrum of the FSK signal, is given by

$$G(f) = e^{-\alpha^2 f^2}. \tag{5.71}$$

The parameter α is related to the 3dB bandwidth of $g(t)$,B_z, by

$$\alpha = \frac{\sqrt{-\ln \sqrt{.5}}}{B_z}. \tag{5.72}$$

Clearly making α large 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 are constant amplitude, and a high spectal efficiency since the Gaussian pulse shape has good spectral properties for large α. For this reason GMSK is used in the GSM standard for digital cellular systems. Although this is a good choice for voice modulation, it is not necessarily a good choice for data. The Gaussian pulse shape does not satisfy the Nyquist criterion, and therefore 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 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 minimum B_g, thereby decreasing the spectral efficiency of GMSK for data transmission. ISI mitigation techniques such as equalization can be used to reduce the ISI in this case so that a tolerable BER is possible without significantly compromising spectral efficiency.

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 above.

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 to study this problem and to develop 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 since 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 parameter estimation of carrier phase and symbol timing when 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], and 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 timing and carrier phase, especially in wideband systems like UWB. Multicarrier and spread spectrum systems have addition considerations related to synchronization and carrier recovery which will be discussed in Chapters 12 and 13, respectively.

The importance of synchronization and carrier phase estimation cannot be overstated: without it wireless systems could not function. Moreover, as data rates increase and channels become more complex by adding addition degrees of freedom (e.g. multiple antennas), the task of receiver synchronizaton and phase recovery becomes even more complex and challenging. Techniques for synchronization and carrier recovery have been developed and analyzed extensively for many years, and these techniques continually evolve 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 as well as performance analysis of these techniques can be found in [18, 19], and more condensed treatments can be found in [6, Chapter 6],[20].

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 5.27. 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, and the resulting circuit complexity motivates the need for the noncoherent demodulators described in Section 5.4.3. We see in Figure 5.27 that the carrier phase and timing recovery circuits operate directly on the received signal prior to demodulation.

[3]That is why delay spread is typically characterized by its rms value about its mean, as discussed in more detail in Chapter 2.

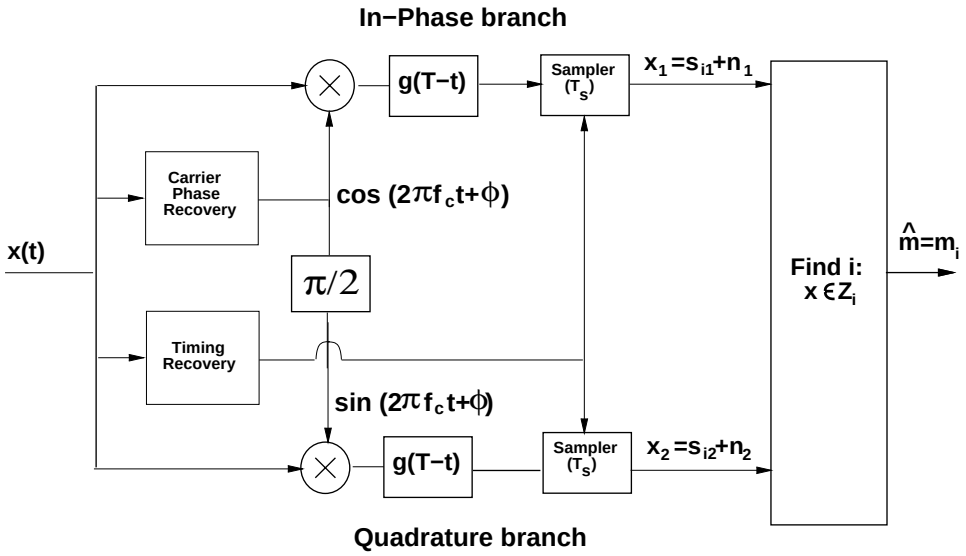


Figure 5.27: Receiver Structure with Carrier and Timing Recovery.

Assuming an AWGN channel, the received signal $x(t)$ is a delayed version of the transmitted signal $s(t)$ plus AWGN $n(t)$: $x(t) = s(t-\tau)+n(t)$, where τ is the random propagation delay. Using the complex baseband form we have $s(t) = \Re[u(t)e^{j(2\pi f_c t+\phi_0)}]$ and thus

$$x(t) = \Re\left\{\left(u(t-\tau)e^{j\phi} + z(t)\right)e^{j2\pi f_c t}\right\}, \tag{5.73}$$

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 $\theta = (\phi, \tau)$. Then we can express the received signal in terms of θ as

$$x(t) = s(t;\theta) + n(t). \tag{5.74}$$

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 done initially over this interval and thereafter estimation is performed continually by updating the initial estimatre using tracking loops. Our development below focuses just on the initial parameter estimation over T_0: discussion of parameter tracking can be found in [18, 19].

There are two common estimation methods for signal parameters in noise, the maximum-likelihood criterion (ML), discussed in Section 5.1.4 in the context of receiver design, and the maximum a posteriori (MAP) criterion. The ML criterion choses the estimate $\hat{\theta}$ that maximizes $p(x(t)|\theta)$ over the observation interval T_0, whereas the MAP criterion assumes some probability distribution on θ, $p(\theta)$, and choses the estimate $\hat{\theta}$ that maximizes

$$p(\theta|x(t)) = \frac{p(x(t)|\theta)p(\theta)}{p(x(t))}$$

over T_0. We assume that there is no prior knowledge of $\hat{\theta}$, so that $p(\theta)$ becomes uniform and therefore the MAP and ML criteria are equivalent.

To characterize the distribution $p(x(t)|\theta), 0 \leq t < T_0$, let us expand $x(t)$ over the observation interval

along a set of orthonormal basis functions $\{\phi_k(t)\}$ as

$$x(t) = \sum_{k=1}^{K} x_k \phi_k(t), 0 \leq t < T_0.$$

Since $n(t)$ is white with zero mean and power spectral density $N_0/2$, the pdf of the vector $\mathbf{x} = (x_1, \ldots, x_K)$ conditioned on the unknown parameter θ is given by

$$p(\mathbf{x}|\theta) = \left(\frac{1}{\sqrt{\pi N_0}\sigma}\right)^K \exp\left[-\sum_{k=1}^{K} \frac{(x_k - s_k(\theta))^2}{N_0}\right], \tag{5.75}$$

where by the basis expansion

$$x_k = \int_{T_0} x(t)\phi_k(t),$$

and we define

$$s_k(\theta) = \int_{T_0} s(t;\theta)\phi_k(t).$$

We can show that

$$\sum_{k=1}^{K} [x_k - s_k(\theta)]^2 = \int_{T_0} [x(t) - s(t;\theta)]^2 dt. \tag{5.76}$$

Using this in (5.75) yields that maximizing $p(\mathbf{x}|\theta)$ is equivalent to maximizing the **likelihood function**

$$\Lambda(\theta) = \exp\left[-\frac{1}{N_0}\int_{T_0} [x(t) - s(t;\theta)]^2\right] dt. \tag{5.77}$$

Maximization of the likelihood function (5.77) results in the joint ML estimate of the carrier phase and symbol timing. ML estimation of the carrier phase and symbol timing can also be done separately. In subsequent sections we will discuss the separate estimation of carrier phase and symbol timing in more detail. Techniques for joint estimation are more complex: details of such techniques can be found in [18, Chapters 8-9],[6, Chapter 6.4].

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

$$\begin{aligned} \Lambda(\phi) &= \exp\left[-\frac{1}{N_0}\int_{T_0} [x(t) - s(t;\phi)]^2 dt\right] \\ &= \exp\left[-\frac{1}{N_0}\int_{T_0} x^2(t)dt + \frac{2}{N_0}\int_{T_0} x(t)s(t;\phi)dt - \frac{1}{N_0}s^2(t;\phi)dt.\right] \end{aligned} \tag{5.78}$$

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. With these observations we see that the $\hat{\phi}$ that maximizes (5.78) also maximizes

$$\Lambda'(\phi) = \int_{T_0} x(t)s(t;\phi)dt. \tag{5.79}$$

We can solve directly for the maximizing $\hat{\phi}$ in the simple case where the received signal is just an unmodulated carrier plus noise: $x(t) = A\cos(2\pi f_c t + \phi) + n(t)$. Then $\hat{\phi}$ must maximize

$$\Lambda'(\phi) = \int_{T_0} x(t)\cos(2\pi f_c t + \phi)dt. \tag{5.80}$$

Differentiating $\Lambda'(\phi)$ relative to ϕ and setting it to zero yields that $\hat{\phi}$ satisfies

$$\int_{T_0} x(t)\sin(2\pi f_c t + \hat{\phi})dt = 0. \tag{5.81}$$

Solving (5.81) for $\hat{\phi}$ yields

$$\hat{\phi} = -\tan^{-1}\left[\frac{\int_{T_0} x(t)\sin(2\pi f_c t)dt}{\int_{T_0} x(t)\cos(2\pi f_c t)dt}\right]. \tag{5.82}$$

While we can build a circuit to compute (5.82) from the received signal $x(t)$, in practice carrier phase recovery is done using a phase lock loop to satisfy (5.81), as shown in Figure 5.21. In this figure the integrator input in the absence of noise is given by $e(t) = x(t)\sin(2\pi f_c t + \hat{\phi})$, and the integrator output is

$$z(t) = \int_{T_0} x(t)\sin(2\pi f_c t + \hat{\phi})dt,$$

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 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 5.21 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 $e(t) = A\cos(2\pi f_c t + \phi)\sin(2\pi f_c t + \hat{\phi}) = .5A\sin(\hat{\phi} - \phi) + .5A\sin(2\pi f_c t + \phi + \hat{\phi})$. The above discussion of the PLL operation assumes that $\hat{\phi} \approx \phi$ since otherwise the polarity of $z(t)$ may not indicate the correct phase adjustment, i.e. 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 $z(t) = 0$, 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 [6, 18].

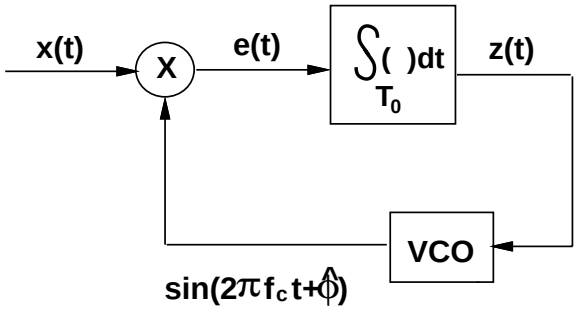


Figure 5.28: Phase Lock Loop for Carrier Phase Recovery (Unmodulated Carrier)

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 assumed known or it is treated as random such that the phase estimate is averaged over the data statistics. The first scenario is refered to as **decision-directed** parameter estimation, and this scenario typically results from sending a known training sequence. The second scenario is refered 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** since 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 tradeoffs in performance and complexity. A more comprehensive discussion of design and performance of carrier phase recovery be found in [18],[6, Chapter 6.2.4-6.2.5].

5.6.3 Maximum-Likelihood Timing Estimation

In this section we derive the maximum likelihood delay τ assuming the carrier phase is known. Since we assume that the phase ϕ is known, the timing recovery will not affect downconversion by the carrier shown in Figure 5.27. Thus, it suffices to consider timing estimation for the in-phase or quadrature baseband equivalent signals of $x(t)$ and $s(t;\tau)$. We denote the in-phase and quadrature components for $x(t)$ as $x_I(t)$ and $x_Q(t)$ and for $s(t;\tau)$ as $s_I(t;\tau)$ and $s_Q(t;\tau)$. We focus on the in-phase branch as the timing recovered from this branch can be used for the quadrature branch. The baseband equivalent 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 baseband equivalent received signal is $x_I(t) = s_I(t;\tau) + n_I(t)$. 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).

The likelihood function (5.77) with known phase ϕ has a similar form as (5.78), the case of known delay:

$$\begin{aligned} \Lambda(\tau) &= \exp\left[-\frac{1}{N_0}\int_{T_0}[x_I(t) - s_I(t;\tau)]^2 dt\right]. \\ &= \exp\left[-\frac{1}{N_0}\int_{T_0} x_I^2(t)dt + \frac{2}{N_0}\int_{T_0} x_I(t)s_I(t;\tau)dt - \frac{1}{N_0}s_I^2(t;\tau)dt.\right] \end{aligned} \tag{5.84}$$

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

$$\Lambda'(\tau) = \int_{T_0} x_I(t)s_I(t;\tau)dt = \sum_k s_I(k)\int_{T_0} x(t)g(t - kT_s - \tau)dt = \sum_k s_I(k)z_k(\tau), \tag{5.85}$$

where

$$z_k(\tau) = \int_{T_0} x(t)g(t - kT_s - \tau)dt. \tag{5.86}$$

Differentiating (5.85) relative to τ and setting it to zero yields that the timing estimate $\hat{\tau}$ must satisfy

$$\sum_k s_I(k)\frac{\partial}{\partial\tau}z_k(\tau) = 0. \tag{5.87}$$

For decision-directed estimation, (5.87) gives rise to the estimator shown in Figure 5.22. 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.

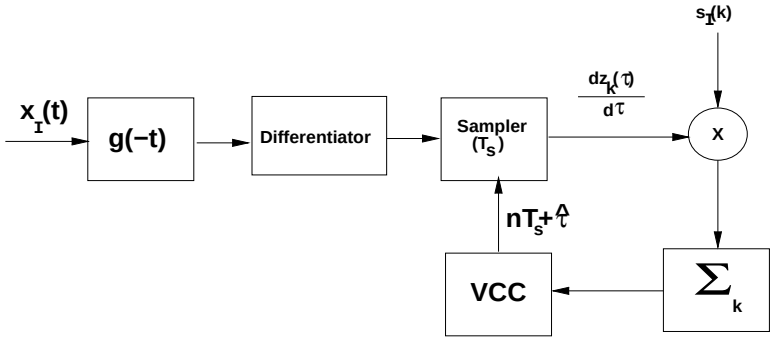


Figure 5.29: Decision-Directed Timing Estimation

One structure for non decision-directed timing estimation is the **early-late gate synchronizer** shown in Figure 5.23. 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 that fact that its maximum value is at $\tau = 0$. The input to the sampler in the upper branch of Figure 5.23 is proportional to the autocorrelation $R_g(\hat{\tau} - \tau + \delta) = \int_0^{T_s} g(t-\tau)g(t-\hat{\tau}+\delta)$ and the input to the sampler in the lower branch is proportional to the autocorrelation $R_g(\hat{\tau} - \tau - \delta) = \int_0^{T_s} g(t-\tau)g(t-\hat{\tau}-\delta)$. 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 (VCC) 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 $\hat{\tau}$.

More details on these and other structures for decision-directed and non decision-directed timing estimation as well as their performance tradeoffs can be found in [18],[6, Chapter 6.2.4-6.2.5].

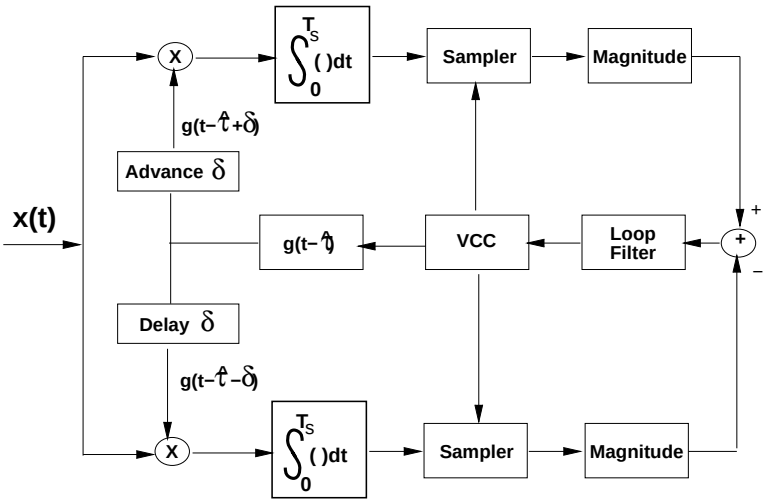


Figure 5.30: Early-Late Gate Synchronizer

Bibliography

[1] S. Haykin, *An Introduction to Analog and Digital Communications*. New York: Wiley, 1989.

[2] S. Haykin, *Communication Systems*. New York: Wiley, 2002.

[3] J. Proakis and M. Salehi, *Communication Systems Engineering*. Prentice Hall, 2002.

[4] M. Fitz, "Further results in the unified analysis of digital communication systems," *IEEE Trans. on Commun.* March 1992.

[5] R. Ziemer, "An overview of modulation and coding for wireless communications," *IEEE Trans. on Commun.*, 1993.

[6] J.G. Proakis, *Digital Communications*. 4th Ed. New York: McGraw-Hill, 2001.

[7] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection,* Prentice Hall: 1995.

[8] T.S. Rappaport, *Wireless Communications - Principles and Practice*, IEEE Press, 1996.

[9] G. L. Stuber, *Principles of Mobile Communications*, Kluwer Academic Publishers, 1996.

[10] J.M. Wozencraft and I.M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[11] J. C.-I. Chuang, "The effects of time delay spread on portable radio communications channels with digital modulation," *IEEE J. Select. Areas Commun.*, June 1987.

[12] A. Mehrotra, *Cellular Radio Performance Engineering* Norwood, MA : Artech House, 1994.

[13] S. Lin and D.J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*. Englewood Cliffs, NJ: Prentice Hall, 1983.

[14] G. Ungerboeck. "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, Vol. IT-28, No. 1, pp. 55–67, Jan. 1982.

[15] G.D. Forney, Jr., "Coset codes - Part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory*, Vol. IT-34, No. 5, pp. 1123–1151, Sept. 1988.

[16] G. Ungerboeck. "Trellis-coded modulation with redundant signal sets, Part I: Introduction and Part II: State of the art." *IEEE Commun. Mag.*, Vol. 25, No. 2, pp. 5–21, Feb. 1987.

[17] G.D. Forney, Jr., and L.-F. Wei, "Multidimensional constellations - Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Selected Areas Commun.*, Vol. SAC-7, No. 6, pp. 877–892, Aug. 1989.

[18] U. Mengali and A. N. D'Andrea, *Synchronization Techniques for Digital Receivers.* New York: Plenum Press, 1997.

[19] H. Meyr, M. Moeneclaey, and S.A. Fechtel, *Digital Communication Receivers, Vol. 2, Synchronization, Channel Estimation, and Signal Processing.* New York: Wiley, 1997.

[20] L.E. Franks, "Carrier and bit synchronization in data communication - A tutorial review," *IEEE Trans. Commun.* pp. 1007–1121, Aug. 1980.

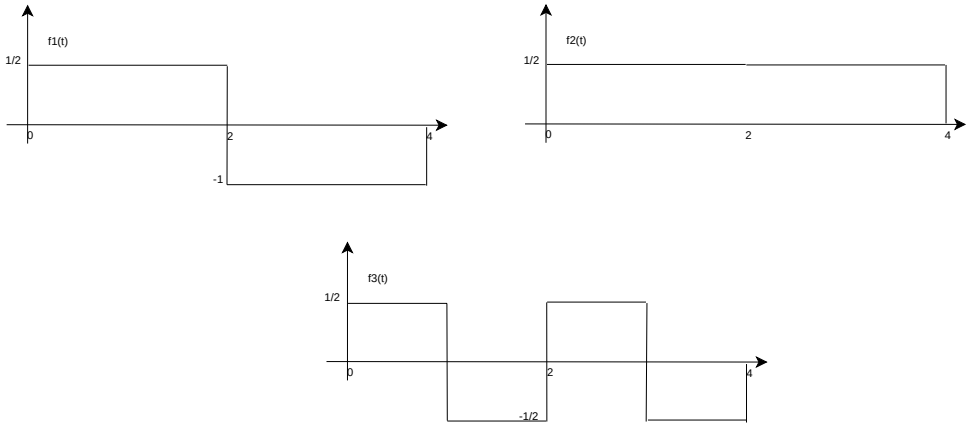
Chapter 5 Problems

1. Show using properties of orthonormal basis functions that if $s_i(t)$ and $s_j(t)$ have constellation points $\mathbf{s}_i$ and $\mathbf{s}_i$, respectively, then
$$||\mathbf{s}_i - \mathbf{s}_j||^2 = \int_0^T (s_i(t) - s_j(t))^2 dt.$$

2. Find an alternate 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 $s_m(t)$, $1 \le m \le M$, $0 \le t \le T$, all of which have the same energy $\mathcal{E}$. Define a new set of M waveforms as
$$s'_m(t) = s_m(t) - \frac{1}{M}\sum_{i=1}^{M} s_i(t), \quad 1 \le m \le M, \quad 0 \le t \le T$$
Show that the M signal waveforms $\{s'_m(t)\}$ have equal energy, given by
$$\mathcal{E}' = (M-1)\mathcal{E}/\mathcal{M}$$
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 below



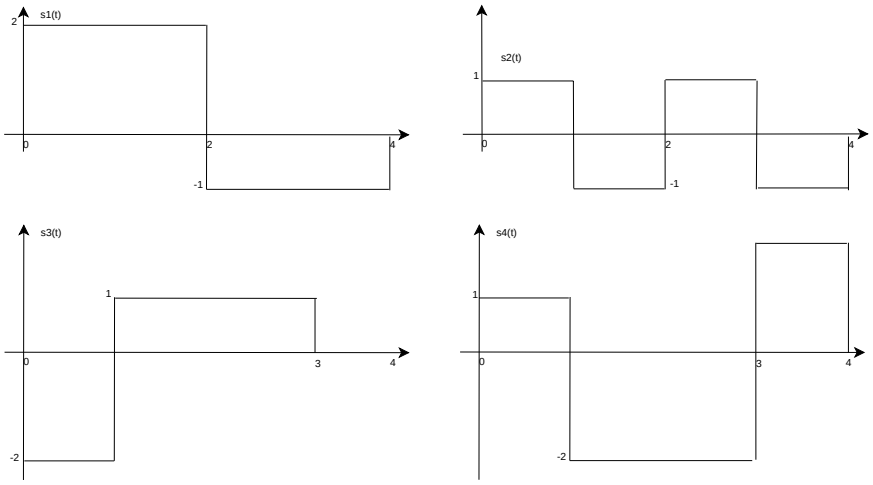
   (a) Show that these waveforms are orthonormal.

   (b) Express the waveform $x(t)$ as a linear combination of $\{\phi_i(t)\}$ and find the coefficients, where $x(t)$ is given as
$$x(t) = \begin{cases} -1 & (0 \le t \le 1) \\ 1 & (1 \le t \le 3) \\ -1 & (3 \le t \le 4) \end{cases}$$

5. Consider the four signal waveforms as shown in the figure below

   (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, i.e. $p(m_i) = p_i, i = 1, \ldots, M$, 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 $s_4 = (0, A_c)$, with $p(s_1) = p(s_3) = .2$ and $p(s_1) = p(s_3) = .3$

7. Show that the remainder noise term $n_r(t_k)$ is independent of the correlator outputs x_i for all i, i.e. show that $E[n_r(t_k)x_i] = 0, \forall\ i$. Thus, since x_j conditioned on $\mathbf{s}_i$ and $n_r(t)$ are Gaussian and uncorrelated, they are independent.

8. Show that if a given input signal is passed through a filter matched to that signal, the output SNR is maximized.

9. Find the matched filters $g(T-t), 0 \leq t \leq T$ and plot $\int_0^T g(t)g(T-t)dt$ for the following waveforms:

   (a) Rectangular pulse: $g(t) = \sqrt{\frac{2}{T}}$

   (b) Sinc pulse: $g(t) = \text{sinc}(t)$.

   (c) Gaussian pulse: $g(t) = \frac{\sqrt{\pi}}{\alpha} e^{-\pi^2 t^2/\alpha^2}$

10. Show that the ML receiver of Figure 5.4 is equivalent to the matched filter receiver of Figure 5.7

11. Compute the three bounds (5.40), (5.43), (5.44), and the approximation (5.45) 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 $A_c/\sqrt{N_0} = 4$

12. Find the input to each branch of the decision device in Figure 5.10 if the transmit carrier phase ϕ_0 differs from the receiver carrier phase ϕ by $\Delta\phi$.

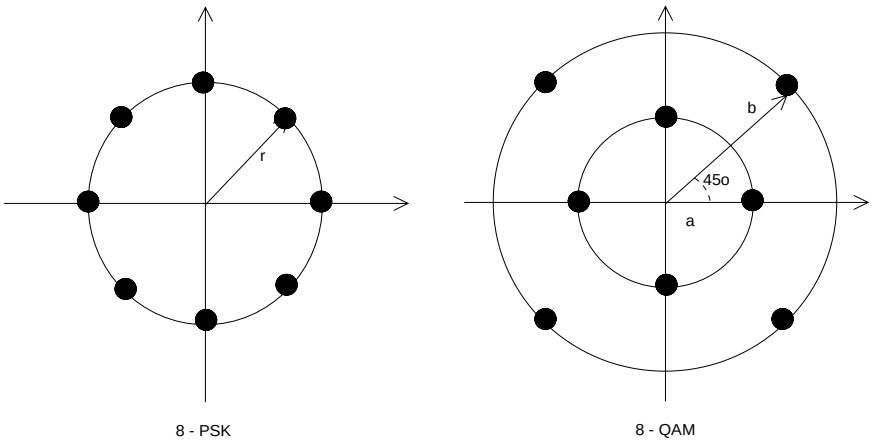
13. Consider a 4-PSK constellation with $d_{\min} = \sqrt{2}$. What is the additional energy required to send one extra bit (8-PSK) while keeping the same minimum distance (and consequently the same bit error probability)?

14. Show that the average power of a square signal constellation with l bits per dimension, S_l, is proportional to $4^l/3$ and that the average power for one more bit per dimension $S_{l+1} \approx 4S_l$. Find S_l for $l = 2$ and compute the average energy of MPSK and MPAM constellations with the same number of bits per symbol.

15. For MPSK with differential modulation, let $\Delta\phi$ denote the phase drift of the channel over a symbol time T_s. In the absence of noise, how large must $\Delta\phi$ be to make a detection error?

16. Find the Gray encoding of bit sequences to phase transitions in differential 8PSK. Then find the sequence of symbols transmitted using differential 8PSK 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 $\mathbf{s}(k-1) = Ae^{j\pi/4}$.

17. Consider the octal signal point constellation in the figure shown below



    (a) The nearest neighbor signal points in the 8-QAM signal constellation are s eparated in distance by 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.

    (c) Determine the average transmitter powers for the two signal constellations and compare the two powers. What is the relative power advantage of one constel lation 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 const ellation such that nearest (adjacent) points differ in only one bit position?

    (e) Determine the symbol rate if the desired bit rate is 90 Mbps.

18. The π/-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 the last symbol transmitted on the in-phase branch had a phase of π and the last symbol transmitted on the quadrature branch had a phase of $-3\pi/4$.

19. 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 $\Delta f = \min_{ij} |f_j - f_i| = .5/T_s$

20. Show that the Nyquist criterion for zero ISI pulses given by (5.68) is equivalent to the frequency domain condition (5.69).

21. Show that the Gaussian pulse shape does not satisfy the Nyquist criterion.

Chapter 6

Performance of Digital Modulation over Wireless Channels

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

6.1 AWGN Channels

In this section we define the signal-to-noise power ratio (SNR) and its relation to energy-per-bit (E_b) and energy-per-symbol (E_s). We then examine the error probability on AWGN channels for different modulation techniques as parameterized by these energy metrics. Our analysis uses the signal space concepts of Chapter 5.1.

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

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

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

density $N_0/2$, the total noise power within the bandwidth $2B$ is $N = N_0/2 \times 2B = N_0 B$. So the received SNR is given by

$$\text{SNR} = \frac{P_r}{N_0 B}.$$

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

The quantities $\gamma_s = E_s/N_0$ and $\gamma_b = E_b/N_0$ are sometimes called the SNR per symbol and the SNR per bit, respectively. For performance specification, we are interested in the bit error probability P_b as a function of γ_b. However, for M-aray signaling (e.g. MPAM and MPSK), the bit error probability depends on both the symbol error probability and the mapping of bits to symbols. Thus, we typically compute the symbol error probability P_s as a function of γ_s based on the signal space concepts of Chapter 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-aray signaling lead to the approximations

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

and

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

6.1.2 Error Probability for BPSK and QPSK

We first consider BPSK modulation with coherent detection and perfect recovery of the carrier frequency and phase. With binary modulation each symbol corresponds to one bit, so the symbol and bit error rates are the same. The transmitted signal is $s_1(t) = Ag(t)\cos(2\pi f_c t)$ to sent a 0 bit and $s_2(t) = -Ag(t)\cos(2\pi f_c t)$ to send a 1 bit. From (5.46) we have that the probability of error is

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

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

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

from (5.56). Thus, the signal constellation for BPSK in terms of energy-per-bit is given by $\mathbf{s}_0 = \sqrt{E_b}$ and $\mathbf{s}_1 = -\sqrt{E_b}$. This yields the minimum distance $d_{min} = 2A = 2\sqrt{E_b}$. Substituting this into (6.4) yields

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

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

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

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

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

From Example 5.1.5, the union bound (5.40) on P_s for QPSK is

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

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

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

The closed form bound (5.44) becomes

$$P_s \leq \frac{3}{\pi}\exp\left[\frac{-.5A^2}{N_0}\right] = \frac{3}{\pi}\exp[-\gamma_s/4]. \tag{6.11}$$

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

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

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

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

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

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

The exact symbol error probability P_s is

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

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

$$P_s \approx 2Q(\sqrt{\gamma_s/2}) = 2Q(\sqrt{5.012}) = 2.517 * 10^{-2},$$

which is more than an order of magnitude larger than the exact P_s. Thus, the nearest neighbor approximation can be quite loose.

6.1.3 Error Probability for MPSK

The signal constellation for MPSK has $s_{i1} = A\cos[\frac{2\pi(i-1)}{M}]$ and $s_{i2} = A\sin[\frac{2\pi(i-1)}{M}]$ for $i = 1, \ldots, M$. The symbol energy is $E_s = A^2$, so $\gamma_s = A^2/N_0$. From (5.57), for the received vector $\mathbf{x} = re^{j\theta}$ represented in polar coordinates, an error occurs if the ith signal constellation point is transmitted and $\theta \notin (2\pi(i - 1 - .5)/M, 2\pi(i - 1 + .5)/M)$. The joint distribution of r and θ can be obtained through a bivariate transformation of the noise n_1 and n_2 on the in-phase and quadrature branches [4, Chapter 5.4], which yields

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

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

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

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

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

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

Each point in the MPSK constellation has two nearest neighbors at distance $d_{min} = \sqrt{2A^2}$. Thus, the nearest neighbor approximation (5.45) to P_s is given by

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

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

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

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

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

Example 6.2:
Compare the probability of bit error for 8PSK and 16PSK assuming $\gamma_b = 15$ dB and using the P_s approximation given in (6.18) along with the approximations (6.3) and (6.2).

Solution: From (6.2)we have that for 8PSK, $\gamma_s = (\log_2 8) \cdot 10^{15/10} = 94.87$. Substituting this into (6.18) yields

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

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

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

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

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

6.1.4 Error Probability for MPAM and MQAM

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

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

From (5.54) the average energy per symbol for MPAM is

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

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

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

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

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

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

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

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

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

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

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

The nearest neighbor approximation for nonrectangular constellations is

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

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

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

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

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

The nearest neighbor approximation is given by

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

which differs by roughly a factor of 2 from the exact value. The symbol error probability for 16PSK in the previous example is $P_s \approx 1.916 \cdot 10^{-7}$, which is roughly four times larger than the exact P_s for 16QAM. The larger P_s for MPSK versus MQAM with the same M and same γ_b is due to the fact that MQAM uses both amplitude and phase to encode data, whereas MPSK uses just the phase. Thus, for the same energy per symbol or bit, MQAM makes more efficient use of energy and thus has better performance.

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

6.1.5 Error Probability for FSK and CPFSK

Let us first consider the error probability of traditional binary FSK with the coherent demodulator of Figure 5.24. Since demodulation is coherent, we can neglect any phase offset in the carrier signals. The transmitted signal is defined by

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

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

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

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

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

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

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

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

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

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

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

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

6.1.6 Error Probability Approximation for Coherent Modulations

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

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

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

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

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

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

6.1.7 Error Probability for Differential Modulation

The probability of error for differential modulation is based on the phase difference associated with the phase comparator input of Figure 5.20. Specifically, the phase comparator extracts the phase of

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

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

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

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

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

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

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

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

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

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

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

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

In DPSK modulation we need only consider the in-phase branch of Figure 5.20 to make a decision, so we set $x = \Re\{\tilde{z}\}$ in our analysis. In particular, assuming a zero is transmitted, if $x = A + \Re\{\tilde{n}^*(k-1) + \tilde{n}(k)\} < 0$ then a decision error is made. This probability can be obtained by finding the characteristic or moment-generating function for x, taking the inverse Laplace transform to get the distribution of x, and then integrating over the decision region $x < 0$. This technique is very general and can be applied to a wide variety of different modulation and detection types in both AWGN and fading [19, Chapter 1.1]: we will use it later to compute the average probability of symbol error for linear modulations in fading both with and without diversity. In DPSK the characteristic function for x is obtained using the general quadratic form of complex Gaussian random variables [1, Appendix B][18, Appendix B], and the resulting bit error probability is given by

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

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

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

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

6.2 Alternate Q Function Representation

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

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

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

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

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

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

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

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

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

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

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

6.3 Fading

In AWGN the probability of symbol error depends on the received SNR or, equivalently, on γ_s. In a fading environment the received signal power varies randomly over distance or time due to shadowing and/or multipath fading. Thus, in fading γ_s is a random variables with distribution $p_{\gamma_s}(\gamma)$, and therefore

$P_s(\gamma_s)$ is also random. The performance metric for γ_s random depends on the rate of change of the fading. There are three different performance criteria that can be used to characterize the random variable P_s:

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

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

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

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

6.3.1 Outage Probability

The outage probability relative to γ_0 is defined as

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

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

In Rayleigh fading the outage probability becomes

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

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

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

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

Example 6.4: Determine the required $\overline{\gamma}_b$ for BPSK modulation in slow Rayleigh fading such that 95% of the time (or in space), $P_b(\gamma_b) < 10^{-4}$.

Solution: For BPSK modulation in AWGN the target BER is obtained at 8.5 dB, i.e. for $P_b(\gamma_b) = Q(\sqrt{2\gamma_b})$, $P_b(10^{.85}) = 10^{-4}$. Thus, $\gamma_0 = 8.5$ dB. Since we want $P_{out} = p(\gamma_b < \gamma_0) = .05$ we have

$$\overline{\gamma}_b = \frac{\gamma_0}{-\ln(1 - P_{out})} = \frac{10^{.85}}{-\ln(1 - .05)} = 21.4\,\text{dB}. \tag{6.49}$$

6.3.2 Average Probability of Error

The average probability of error is used as a performance metric when $T_s \approx T_c$. Thus, we can assume that γ_s is roughly constant over a symbol time. Then the averaged probability of error is computed by integrating the error probability in AWGN over the fading distribution:

$$\overline{P}_s = \int_0^\infty P_s(\gamma) p_{\gamma_s}(\gamma) d\gamma, \tag{6.50}$$

where $P_s(\gamma)$ is the probability of symbol error in AWGN with SNR γ, which can be approximated by the expressions in Table 6.1. For a given distribution of the fading amplitude r (i.e. Rayleigh, Rician, log-normal, etc.), we compute $p_{\gamma_s}(\gamma)$ by making the change of variable

$$p_{\gamma_s}(\gamma) d\gamma = p(r) dr. \tag{6.51}$$

For example, in Rayleigh fading the received signal amplitude r has the Rayleigh distribution

$$p(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \quad r \geq 0, \tag{6.52}$$

and the signal power is exponentially distributed with mean $2\sigma^2$. The SNR per symbol for a given amplitude r is

$$\gamma = \frac{r^2 T_s}{2\sigma_n^2}, \tag{6.53}$$

where $\sigma_n^2 = N_0/2$ is the PSD of the noise in the in-phase and quadrature branches. Differentiating both sides of this expression yields

$$d\gamma = \frac{r T_s}{\sigma_n^2} dr. \tag{6.54}$$

Substituting (6.53) and (6.54) into (6.52) and then (6.51) yields

$$p_{\gamma_s}(\gamma) = \frac{T_s \sigma_n^2}{\sigma^2} e^{-\gamma T_s \sigma_n^2/\sigma^2}. \tag{6.55}$$

Since the average SNR per symbol $\overline{\gamma}_s$ is just $\sigma^2/(T_s\sigma_n^2)$, we can rewrite (6.55) as

$$p_{\gamma_s}(\gamma) = \frac{1}{\overline{\gamma}_s} e^{-\gamma/\overline{\gamma}_s}, \tag{6.56}$$

which is exponential. For binary signaling this reduces to

$$p_{\gamma_b}(\gamma) = \frac{1}{\overline{\gamma}_b} e^{-\gamma/\overline{\gamma}_b}, \tag{6.57}$$

Integrating (6.6) over the distribution (6.57) yields the following average probability of error for BPSK in Rayleigh fading.

$$\text{BPSK:} \qquad \overline{P}_b = \frac{1}{2}\left[1 - \sqrt{\frac{\overline{\gamma}_b}{1+\overline{\gamma}_b}}\right] \approx \frac{1}{4\overline{\gamma}_b}, \tag{6.58}$$

where the approximation holds for large $\overline{\gamma}_b$. A similar integration of (6.31) over (6.57) yields the average probability of error for binary FSK in Rayleigh fading as

$$\text{Binary FSK:} \qquad \overline{P}_b = \frac{1}{2}\left[1 - \sqrt{\frac{\overline{\gamma}_b}{2+\overline{\gamma}_b}}\right] \approx \frac{1}{4\overline{\gamma}_b}. \tag{6.59}$$

Thus, the performance of BPSK and binary FSK converge at high SNRs. For noncoherent modulation, if we assume the channel phase is relatively constant over a symbol time, then we obtain the probability of error by again integrating the error probability in AWGN over the fading distribution. For DPSK this yields

$$\text{DPSK:} \qquad \overline{P}_b = \frac{1}{2(1+\overline{\gamma}_b)} \approx \frac{1}{2\overline{\gamma}_b}, \tag{6.60}$$

where again the approximation holds for large $\overline{\gamma}_b$. Note that in the limit of large $\overline{\gamma}_b$, there is an approximate 3 dB power penalty in using DPSK instead of BPSK. This was also observed in AWGN, and is the power penalty of differential detection. In practice the power penalty is somewhat smaller, since DPSK can correct for slow phase changes introduced in the channel or receiver, which are not taken into account in these error calculations.

If we use the general approximation $P_s \approx \alpha_M Q(\sqrt{\beta_M \gamma_s})$ then the average probability of symbol error in Rayleigh fading can be approximated as

$$\overline{P}_s \approx \int_0^\infty \alpha_M Q(\sqrt{\beta_M \gamma}) \cdot \frac{1}{\overline{\gamma}_s} e^{-\gamma/\overline{\gamma}_s} d\gamma_s. = \frac{\alpha_m}{2}\left[1 - \sqrt{\frac{.5\beta_M \overline{\gamma}_s}{1+.5\beta_M \overline{\gamma}_s}}\right]. \tag{6.61}$$

It is interesting to compare bit error probability of the different modulation schemes in AWGN and fading. For binary PSK, FSK, and DPSK, the bit error probability in AWGN decreases exponentially with increasing γ_b. However, in fading the bit error probability for all the modulation types decreases just linearly with increasing $\overline{\gamma}_b$. Similar behaivor occurs for nonbinary modulation. Thus, the power necessary to maintain a given P_b, particularly for small values, is much higher in fading channels than in AWGN channels. For example, in Figure 6.1 we plot the error probability of BPSK in AWGN and in flat Rayleigh fading. We see that it requires approximately 8 dB SNR to maintain a 10^{-3} bit error rate in AWGN while it requires approximately 24 dB SNR to maintain the same error rate in fading. A similar plot for the error probabilities of MQAM, based on the approximations (6.24) and (6.61), is shown in Figure 6.2. From these figures it is clear that to maintain low power requires some technique to remove the effects of fading. We will discuss some of these techniques, including diversity combining, spread spectrum, and RAKE receivers, in later chapters.

Rayleigh fading is one of the worst-case fading scenarios. In Figure 6.3 we show the average bit error probability of BPSK in Nakagami fading for different values of the Nakagami-m parameter. We see that as m increases, the fading decreases, and the average bit error probability converges to that of an AWGN channel.

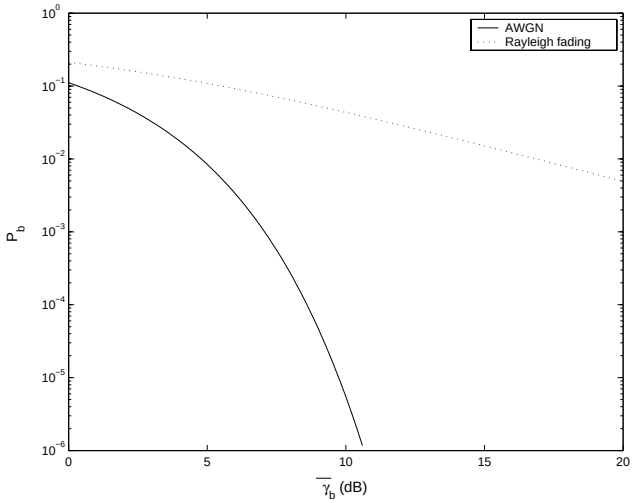


Figure 6.1: Average P_b for BPSK in Rayleigh Fading and AWGN.

6.3.3 Moment Generating Function Approach to Average Error Probability

The **moment generating function** (MGF) for a nonnegative random variable γ with pdf $p_\gamma(\gamma), \gamma \geq 0$, is defined as

$$\mathcal{M}_\gamma(s) = \int_0^\infty p_\gamma(\gamma) e^{s\gamma} d\gamma. \tag{6.62}$$

Note that this function is just the Laplace transform of the pdf $p_\gamma(\gamma)$ with the argument reversed in sign: $\mathcal{L}[p_\gamma(\gamma)] = \mathcal{M}_\gamma(-s)$. Thus, the MGF for most fading distributions of interest can be computed either in closed-form using classical Laplace transforms or through numerical integration. In particular, the MGF for common multipath fading distributions are as follows [19, Chapter 5.1].

Rayleigh:

$$\mathcal{M}_{\gamma_s}(s) = (1 - s\overline{\gamma}_s)^{-1}. \tag{6.63}$$

Ricean with factor K:

$$\mathcal{M}_{\gamma_s}(s) = \frac{1+K}{1+K-s\overline{\gamma}_s} \exp\left[\frac{Ks\overline{\gamma}_s}{1+K-s\overline{\gamma}_s}\right]. \tag{6.64}$$

Nakagami-m:

$$\mathcal{M}_{\gamma_s}(s) = \left(1 - \frac{s\overline{\gamma}_s}{m}\right)^{-m}. \tag{6.65}$$

As indicated by its name, the moments $E[\gamma^n]$ of γ can be obtained from $\mathcal{M}_\gamma(s)$ as

$$E[\gamma^n] = \frac{\partial^n}{\partial s^n} \left[\mathcal{M}_{\gamma_s}(s)\right|_{s=0}. \tag{6.66}$$

The MGF is a very useful tool in performance analysis of modulation in fading both with and without diversity. In this section we discuss how it can be used to simplify performance analysis of average probability of symbol error in fading. In the next chapter we will see that it also greatly simplifies analysis in fading channels with diversity.

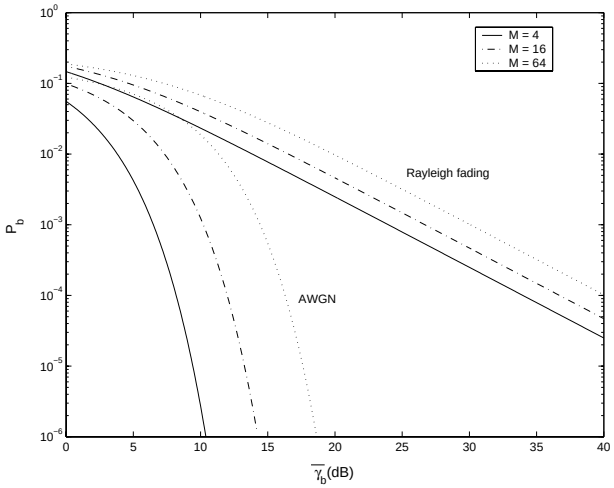


Figure 6.2: Average P_b for MQAM in Rayleigh Fading and AWGN.

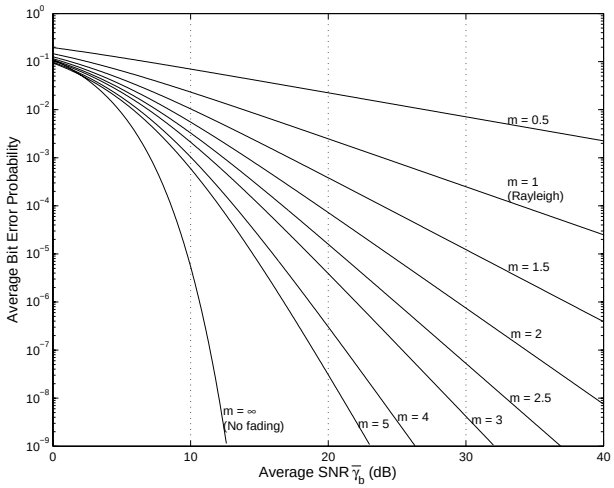


Figure 6.3: Average P_b for BPSK in Nakagami Fading.

The basic premise of the MGF approach for computing average error probability in fading is to express the probability of error P_s in AWGN for the modulation of interest either as an exponential function of γ_s,

$$P_s = c_1 \exp[-c_2 \gamma_s] \tag{6.67}$$

for constants c_1 and c_2, or as a finite range integral of such an exponential function:

$$P_s = \int_A^B c_1 \exp[-c_2(x)\gamma] dx, \tag{6.68}$$

where the constant $c_2(x)$ may depend on the integrand but the SNR γ does not and is not in the limits of integration either. These forms allow the average probability of error to be expressed in terms of the MGF for the fading distribution. Specifically, if $P_s = \alpha \exp[-\beta\gamma_s]$, then

$$\overline{P}_s = \int_0^\infty c_1 \exp[-c_2\gamma] p_{\gamma_s}(\gamma) d\gamma = c_1 \mathcal{M}_{\gamma_s}(-c_2). \tag{6.69}$$

Since DPSK is in this form with $c_1 = 1/2$ and $c_2 = 1$, we see that the average probability of bit error for DPSK in any type of fading is

$$\overline{P}_b = \frac{1}{2}\mathcal{M}_{\gamma_s}(-1), \tag{6.70}$$

where $\mathcal{M}_{\gamma_s}(s)$ is the MGF of the fading distribution. For example, using $\mathcal{M}_{\gamma_s}(s)$ for Rayleigh fading given by (6.63) with $s = -1$ yields $\overline{P}_b = [2(1 + \overline{\gamma}_b)]^{-1}$, which is the same as we obtained in (6.60). If P_s is in the integral form of (6.68) then

$$\overline{P}_s = \int_0^\infty \int_A^B c_1 \exp[-c_2(x)\gamma] dx p_{\gamma_s}(\gamma) d\gamma = c_1 \int_A^B \left[\int_0^\infty \exp[-c_2(x)\gamma] p_{\gamma_s}(\gamma) d\gamma\right] dx = c_1 \int_A^B \mathcal{M}_{\gamma_s}(-c_2(x)) dx. \tag{6.71}$$

In this latter case, the average probability of symbol error is a single finite-range integral of the MGF of the fading distribution, which can typically be found in closed form or easily evaluated numerically.

Let us now apply the MGF approach to specific modulations and fading distributions. In (6.33) we gave a general expression for P_s of coherent modulation in AWGN in terms of the Gaussian Q function. We now make a slight change of notation in (6.33) setting $\alpha = \alpha_M$ and $g = .5\beta_M$ to get

$$P_s(\gamma_s) = \alpha Q(\sqrt{2g\gamma_s}), \tag{6.72}$$

where α and g are constants that depend on the modulation. The notation change is to obtain the error probability as an exact MGF, as we now show.

Using the alternate Q function representation (6.43), we get that

$$P_s = \frac{\alpha}{\pi} \int_0^{\pi/2} \exp\left[\frac{-g\gamma}{\sin^2\phi}\right] d\phi, \tag{6.73}$$

which is in the desired form (6.68). Thus, the average error probability in fading for modulations with $P_s = \alpha Q(\sqrt{2g\gamma_s})$ in AWGN is given by

$$\begin{aligned}\overline{P}_s &= \frac{\alpha}{\pi} \int_0^\infty \int_0^{\pi/2} \exp\left[\frac{-g\gamma}{\sin^2\phi}\right] d\phi p_{\gamma_s}(\gamma) d\gamma \\ &= \frac{\alpha}{\pi} \int_0^{\pi/2} \left[\int_0^\infty \exp\left[\frac{-g\gamma}{\sin^2\phi}\right] p_{\gamma_s}(\gamma) d\gamma\right] d\phi = \frac{\alpha}{\pi} \int_0^{\pi/2} \mathcal{M}_{\gamma_s}\left(\frac{-g}{\sin^2\phi}\right) d\phi,\end{aligned} \tag{6.74}$$

where $\mathcal{M}_{\gamma_s}(s)$ is the MGF associated with the pdf $p_{\gamma_s}(\gamma)$ as defined by (6.62). Recall that Table 6.1 approximates the error probability in AWGN for many modulations of interest as $P_s \approx \alpha Q(\sqrt{2g\gamma_s})$, and thus (6.74) gives an approximation for the average error probability of these modulations in fading. Moreover, the exact average probability of symbol error for coherent MPSK can be obtained in a form similar to (6.74) by noting that Craig's formula for P_s of MPSK in AWGN given by (6.45) is in the desired form (6.68). Thus, the exact average probability of error for MPSK becomes

$$\begin{aligned}\overline{P}_s &= \int_0^\infty \frac{1}{\pi} \int_0^{(M-1)\pi/M} \exp\left[\frac{-g\gamma_s}{\sin^2\phi}\right] d\phi p_{\gamma_s}(\gamma) d\gamma \\ &= \frac{1}{\pi} \int_0^{\frac{(M-1)\pi}{M}} \left[\int_0^\infty \exp\left[\frac{-g\gamma_s}{\sin^2\phi}\right] p_{\gamma_s}(\gamma) d\gamma\right] d\phi = \frac{1}{\pi} \int_0^{\frac{(M-1)\pi}{M}} \mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right) d\phi,\end{aligned} \tag{6.75}$$

where $g = \sin^2(\pi/M)$ depends on the size of the MPSK constellation. The MGF $\mathcal{M}_{\gamma_s}(s)$ for Rayleigh, Rician, and Nakagami-m distributions were given, respectively, by (6.63), (6.64), and (6.65) above. Substituting $s = -g/\sin^2\phi$ in these expressions yields

Rayleigh:

$$\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right) = \left(1+\frac{g\,\overline{\gamma}_s}{\sin^2\phi}\right)^{-1}. \tag{6.76}$$

Ricean with factor K:

$$\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right) = \frac{(1+K)\,\sin^2\phi}{(1+K)\sin^2\phi + g\,\overline{\gamma}_s}\;\exp\left(-\frac{K\,g\,\overline{\gamma}_s}{(1+K)\sin^2\phi + g\,\overline{\gamma}_s}\right). \tag{6.77}$$

Nakagami-m:

$$\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right) = \left(1+\frac{g\,\overline{\gamma}_s}{m\,\sin^2\phi}\right)^{-m}. \tag{6.78}$$

All of these functions are simple trigonometrics and are therefore easy to integrate over the finite range in (6.74) or (6.75).

Example 6.5: Use the MGF technique to find an expression for the average probability of error for BPSK modulation in Nakagami fading.

Solution: We use the fact that for an AWGN channel BPSK has $P_b = Q(\sqrt{2\gamma_b})$, so $\alpha = 1$ and $g = 1$ in (6.72). The moment generating function for Nakagami-m fading is given by (6.78), and substituting this into (6.74) with $\alpha = g = 1$ yields

$$\overline{P}_b = \frac{1}{\pi}\int_0^{\pi/2}\left(1+\frac{\overline{\gamma}_b}{m\sin^2\phi}\right)^{-m} d\phi.$$

From (6.23) we see that the exact probability of symbol error for MQAM in AWGN contains both the Q function and its square. Fortunately, an alternate form of $Q^2(z)$ derived in [8] allows us to apply the same techniques used above for MPSK to MQAM modulation. Specifically, an alternate representation of $Q^2(z)$ is derived in [8] as

$$Q^2(z) = \frac{1}{\pi}\int_0^{\pi/4}\exp\left[\frac{-z^2}{2\sin^2\phi}\right] d\phi. \tag{6.79}$$

Note that this is identical to the alternate representation for $Q(z)$ given in (6.43) except that the upper limit of the integral is $\pi/4$ instead of $\pi/2$. Thus we can write (6.23) in terms of the alternate representations for $Q(z)$ and $Q^2(z)$ as

$$P_s(\gamma_s) = \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\exp\left(-\frac{g\gamma_s}{\sin^2\phi}\right)d\phi - \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\exp\left(-\frac{g\gamma_s}{\sin^2\phi}\right)d\phi, \tag{6.80}$$

where $g = 1.5/(M-1)$ is a function of the size of the MQAM constellation. Then the average probability of symbol error in fading becomes

$$\begin{aligned}
\overline{P}_s &= \int_0^\infty P_s(\gamma)p_{\gamma_s}(\gamma)d\gamma \\
&= \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\int_0^\infty \exp\left(-\frac{g\gamma}{\sin^2\phi}\right)p_{\gamma_s}(\gamma)d\gamma d\phi - \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\int_0^\infty \exp\left(-\frac{g\gamma}{\sin^2\phi}\right)p_{\gamma_s}(\gamma)d\gamma d\phi \\
&= \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right)d\phi - \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right)d\phi.
\end{aligned} \tag{6.81}$$

Thus, the exact average probability of symbol error is obtained via two finite-range integrals of the MGF of the fading distribution, which can typically be found in closed form or easily evaluated numerically.

The MGF approach can also be applied to noncoherent and differential modulations. For example, consider noncoherent M-FSK, with P_s in AWGN given by (6.32), which is a finite sum of the desired form (6.67). Thus, in fading, the average symbol error probability of noncoherent M-FSK is given by

$$\begin{aligned}
\overline{P}_s &= \int_0^\infty \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \exp\left[\frac{-m\gamma}{m+1}\right] p_{\gamma_s}(\gamma) d\gamma \\
&= \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \left[\int_0^\infty \exp\left[\frac{-m\gamma}{m+1}\right] p_{\gamma_s}(\gamma) d\gamma\right] \\
&= \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \mathcal{M}_{\gamma_s}\left(-\frac{m}{m+1}\right). \qquad (6.82)
\end{aligned}$$

Finally, for differential MPSK, it can be shown [11] that the average probability of symbol error is given by

$$P_s = \frac{\sqrt{g_{psk}}}{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\exp[-\gamma_s(1-\sqrt{1-g_{psk}}\cos\theta)]}{1-\sqrt{1-g_{psk}}\cos\theta} d\theta \qquad (6.83)$$

for $g_{psk} = \sin^2(\pi/M)$, which is in the desired form (6.68). Thus we can express the average probability of symbol error in terms of the MGF of the fading distribution as

$$\overline{P}_s = \frac{\sqrt{g_{psk}}}{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\mathcal{M}_{\gamma_s}\left(-(1-\sqrt{1-g_{psk}}\cos\theta)\right)}{1-\sqrt{1-g_{psk}}\cos\theta} d\theta. \qquad (6.84)$$

A more extensive discussion of the MGF technique for finding average probability of symbol error for different modulations and fading distributions can be found in [19, Chapter 8.2].

6.3.4 Combined Outage and Average Error Probability

When the fading environment is a superposition of both fast and slow fading, i.e. log-normal shadowing and Rayleigh fading, a common performance metric is combined outage and average error probability, where outage occurs when the slow fading falls below some target value and the average performance in nonoutage is obtained by averaging over the fast fading. We use the following notation:

- Let $\overline{\overline{\gamma}}_s$ denote the average SNR per symbol due to shadowing and path loss.
- Let $\overline{\gamma}_s$ denote the (random) SNR per symbol due to shadowing and path loss with average value $\overline{\overline{\gamma}}_s$.
- Let γ_s denote the random SNR due to path loss, shadowing, and multipath.

With this notation we can specify an average error probability $\overline{P}_s$ with some probability $1 - P_{out}$. An outage is declared when the received SNR per symbol due to shadowing and path loss alone, $\overline{\gamma}_s$, falls below a given target value $\overline{\gamma}_{s_0}$. When not in outage ($\overline{\gamma}_s \geq \overline{\gamma}_{s_0}$), the average probability of error is obtained by averaging over the distribution of the fast fading conditioned on the mean SNR:

$$\overline{P}_s = \int_0^\infty P_s(\gamma_s) p(\gamma_s|\overline{\gamma}_s) d\gamma_s. \qquad (6.85)$$

The criterion used to determine the outage target $\overline{\gamma}_{s_0}$ is typically based on a given maximum average probability of error, i.e. $\overline{P}_s \leq \overline{P}_{s_0}$, where the target $\overline{\gamma}_{s_0}$ must then satisfy

$$\overline{P}_{s_0} = \int_0^\infty P_s(\gamma_s) p(\gamma_s | \overline{\gamma}_{s_0}) d\gamma_s. \tag{6.86}$$

Clearly whenever $\overline{\gamma}_s > \overline{\gamma}_{s_0}$, the average error probability will be below the target value.

Example 6.6:

Consider BPSK modulation in a channel with both log-normal shadowing ($\sigma = 8$ dB) and Rayleigh fading. The desired maximum average error probability is $\overline{P}_{b_0} = 10^{-4}$, which requires $\overline{\gamma}_{b_0} = 34$ dB. Determine the value of $\overline{\overline{\gamma}}_b$ that will insure $\overline{P}_b \leq 10^{-4}$ with probability $1 - P_{out} = .95$.

Solution: We must find $\overline{\overline{\gamma_b}}$, the average of γ_b in both the fast and slow fading, such that $p(\overline{\gamma_b} > \gamma_{b_0}) = 1 - P_{out}$. For log-normal shadowing we compute this as:

$$p(\overline{\gamma_b} > 34) = p\left(\frac{\overline{\gamma_b} - \overline{\overline{\gamma_b}}}{\sigma} \geq \frac{34 - \overline{\overline{\gamma_b}}}{\sigma}\right) = Q\left(\frac{34 - \overline{\overline{\gamma_b}}}{\sigma}\right) = 1 - P_{out}, \tag{6.87}$$

since $(\overline{\gamma_b} - \overline{\overline{\gamma_b}})/\sigma$ is a Gauss-distributed random variable with mean zero and standard deviation one. Thus, the value of $\overline{\overline{\gamma_b}}$ is obtained by substituting the values of P_{out} and σ in (6.87) and using a table of Q functions or an inversion program, which yields $(34 - \overline{\overline{\gamma_b}})/8 = -1.6$ or $\overline{\overline{\gamma}}_b = 46.8$ dB.

6.4 Doppler Spread

Doppler spread results in an irreducible error floor for modulation techniques using differential detection. This is due to the fact that in differential modulation the signal phase associated with one symbol is used as a phase reference for the next symbol. If the channel phase decorrelates over a symbol, then the phase reference becomes extremely noisy, leading to a high symbol error rate that is independent of received signal power. The phase correlation between symbols and therefore the degradation in performance are functions of the Doppler frequency $f_D = v/\lambda$ and the symbol time T_s.

The first analysis of the irreducible error floor due to Doppler was done by Bello and Nelin in [17]. In that work analytical expressions for the irreducible error floor of noncoherent FSK and DPSK due to Doppler are determined for a Gaussian Doppler power spectrum. However, these expressions are not in closed-form, so must be evaluated numerically. Closed-form expressions for the bit error probability of DPSK in fast Rician fading, where the channel decorrelates over a bit time, can be obtained using the MGF technique, with the MGF obtained based on the general quadratic form of complex Gaussian random variables [18, Appendix B] [1, Appendix B]. A different approach utilizing alternate forms of the Marcum Q function can also be used [19, Chapter 8.2.5]. The resulting average bit error probability for DPSK is

$$\overline{P}_b = \frac{1}{2}\left[\frac{1 + K + \overline{\gamma}_b(1 - \rho_C)}{1 + K + \overline{\gamma}_b}\right] \exp\left(-\frac{K\overline{\gamma}_b}{1 + K + \overline{\gamma}_b}\right), \tag{6.88}$$

where ρ_C is the channel correlation coefficient after a bit time T_b, K is the fading parameter of the Rician distribution, and $\overline{\gamma}_b$ is the average SNR per bit. For Rayleigh fading ($K = 0$) this simplifies to

$$\overline{P}_b = \frac{1}{2}\left[\frac{1 + \overline{\gamma}_b(1 - \rho_C)}{1 + \overline{\gamma}_b}\right]. \tag{6.89}$$

Letting $\overline{\gamma}_b \to \infty$ in (6.88) yields the irreducible error floor:

$$\text{DPSK:} \quad \overline{P}_{floor} = \frac{(1-\rho_C)e^{-K}}{2}. \tag{6.90}$$

A similar approach is used in [20] to bound the bit error probability of DQPSK in fast Rician fading as

$$P_b \leq \frac{1}{2}\left[1 - \sqrt{\frac{(\rho_C\overline{\gamma}_s/\sqrt{2})^2}{(\overline{\gamma}_s+1)^2 - (\rho_C\overline{\gamma}_s/\sqrt{2})^2}}\right] \exp\left[-\frac{(2-\sqrt{2})K\overline{\gamma}_s/2}{(\overline{\gamma}_s+1) - (\rho_C\overline{\gamma}_s/\sqrt{2})}\right], \tag{6.91}$$

where K is as before, ρ_C is the signal correlation coefficient after a symbol time T_s, and $\overline{\gamma}_s$ is the average SNR per symbol. Letting $\overline{\gamma}_s \to \infty$ yields the irreducible error floor:

$$\text{DQPSK:} \quad \overline{P}_{floor} = \frac{1}{2}\left[1 - \sqrt{\frac{(\rho_C/\sqrt{2})^2}{1-(\rho_C/\sqrt{2})^2}}\right] \exp\left[-\frac{(2-\sqrt{2})(K/2)}{1-\rho_C/\sqrt{2}}\right]. \tag{6.92}$$

As discussed in Chapter 3.2.1, the channel correlation $A_C(t)$ over time t equals the inverse Fourier transform of the Doppler power spectrum $S_C(f)$ as a function of Doppler frequency f. The correlation coefficient is thus $\rho_C = A_C(T)/A_C(0)$ evaluated at $T = T_s$ for DQPSK or $T = T_b$ for DPSK. Table 6.2, from [21], gives the value of ρ_C for several different Doppler power spectra models, where B_D is the Doppler spread of the channel. Assuming the uniform scattering model ($\rho_C = J_0(2\pi f_D T_b)$) and Rayleigh fading ($K = 0$) in (6.90) yields an irreducible error for DPSK of

$$P_{floor} = \frac{1 - J_0(2\pi f_D T_b)}{2} \approx .5(\pi f_D T_b)^2, \tag{6.93}$$

where $B_D = f_D = v/\lambda$ is the maximum Doppler in the channel. Note that in this expression, the error floor decreases with data rate $R = 1/T_b$. This is true in general for irreducible error floors of differential modulation due to Doppler, since the channel has less time to decorrelated between transmitted symbols. This phenomenon is one of the few instances in digital communications where performance improves as data rate increases.

Type	Doppler Power Spectrum $S_C(f)$	$\rho_C = A_C(T)/A_C(0)$
Rectangular	$\frac{S_0}{2B_D}$, $\lvert f\rvert < B_D$	$\text{sinc}(2B_DT)$
Gaussian	$\frac{S_0}{\sqrt{\pi}B_D}e^{-f^2/B_D^2}$	$e^{-(\pi B_D T)^2}$
Uniform Scattering	$\frac{S_0}{\pi\sqrt{B_D^2-f^2}}$, $\lvert f\rvert < B_D$	$J_0(2\pi B_D T)$
1st Order Butterworth	$\frac{S_0 B_D}{\pi(f^2+B_D^2)}$)	$e^{-2\pi B_D T}$

Table 6.2: Correlation Coefficients for Different Doppler Power Spectra Models.

A plot of (6.88), the error probability of DPSK in fast Rician fading, for uniform scattering ($\rho_C = J_0(2\pi f_D T_b)$) and different values of $f_D T_b$ is shown in Figure 6.4. We see from this figure that the error floor starts to dominate at $\overline{\gamma}_b = 15$ dB in Rayleigh fading ($K = 0$), and as K increases the value of $\overline{\gamma}_b$ where the error floor dominates also increases. We also see that increasing the data rate $R_b = 1/T_b$ by an order of magnitude increases the error floor by roughly two orders of magnitude.

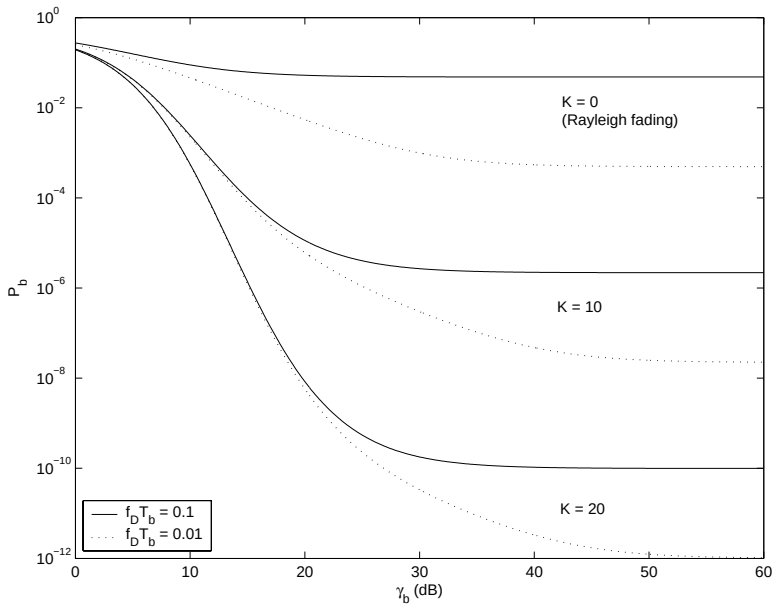


Figure 6.4: Average P_b for DPSK in Fast Rician Fading with Uniform Scattering.

Example 6.7:
Assume a Rayleigh fading channel with uniform scattering and a maximum Doppler of $f_D = 80$ Hertz. For what approximate range of data rates will the irreducible error floor of DPSK be below 10^{-4}.

Solution: We have $P_{floor} \approx .5(\pi f_D T_b)^2 < 10^{-4}$. Solving for T_b with $f_D = 80$ Hz, we get

$$T_b < \frac{\sqrt{2 \cdot 10^{-4}}}{\pi \cdot 80} = 5.63 \cdot 10^{-5},$$

which yields $R > 17.77$ Kbps.

Deriving analytical expressions for the irreducible error floor becomes intractable with more complex modulations, in which case simulations are often used. In particular, simulatons of the irreducible error floor for $\pi/4$ DQPSK with square root raised cosine filtering have been conducted since this modulation is used in the IS-54 TDMA standard [22, 23]. These simulation results indicate error floors between 10^{-3} and 10^{-4}. Surprisingly, in these simulations the error floor increases with vehicle speed, despite the fact that at higher vehicle speeds, the channel decorrelates less over a symbol time.

6.5 Intersymbol Interference

Frequency-selective fading give rise to ISI, where the received symbol over a given symbol period experiences interference from other symbols that have been delayed by multipath. Since increasing signal power also increases the power of the ISI, this interference gives rise to an irreducible error floor that is independent of signal power. The irreducible error floor is difficult to analyze, since it depends on the

ISI characteristics and the modulation format, and the ISI characteristics depend on the characteristics of the channel and the sequence of transmitted symbols.

The first extensive analysis of ISI degradation to symbol error probability was done by Bello and Nelin [24]. In that work analytical expressions for the irreducible error floor of coherent FSK and noncoherent DPSK are determined assuming a Gaussian delay profile for the channel. To simplify the analysis, only ISI associated with adjacent symbols was taken into account. Even with this simplification, the expressions are very complex and must be approximated for evaluation. The irreducible error floor can also be evaluated analytically based on the worst-case sequence of transmitted symbols or it can be averaged over all possible symbol sequences [25, Chapter 8.2]. These expressions are also complex to evaluate due to their dependence on the channel and symbol sequence characteristics. An approximation to symbol error probability with ISI can be obtained by treating the ISI as uncorrelated white Gaussian noise [28]. Then the SNR becomes

$$\hat{\gamma}_s = \frac{P_r}{N_0 B + I}, \tag{6.94}$$

where I is the power associated with the ISI. In a static channel the resulting probability of symbol error will be $P_s(\hat{\gamma}_s)$ where P_s is the probability of symbol error in AWGN. If both the transmitted signal and the ISI experience flat-fading, then $\hat{\gamma}_s$ will be a random variable with a distribution $p(\hat{\gamma}_s)$, and the average symbol error probability is then $\overline{P}_s = \int P_s(\hat{\gamma}_s)p(\hat{\gamma}_s)$. Note that $\hat{\gamma}_s$ is the ratio of two random variables: the received power P_r and the ISI power I, and thus the resulting distribution $p(\hat{\gamma}_s)$ may be hard to obtain or not in closed form.

Irreducible error floors due to ISI are often obtained by simulation, which can easily incorporate different channel models, modulation formats, and symbol sequence characteristics [26, 28, 27, 22, 23]. The most extensive simulations for determining irreducible error floor due to ISI were done by Chuang in [26]. In this work BPSK, DPSK, QPSK, OQPSK and MSK modulations were simulated for different pulse shapes and for channels with different power delay profiles, including a Gaussian, exponential, equal-amplitude two-ray, and empirical power delay profile. The results of [26] indicate that the irreducible error floor is more sensitive to the rms delay spread of the channel than to the shape of its power delay profile. Moreover, pulse shaping can significantly impact the error floor: in the raised cosine pulses discussed in Chapter 5.5, increasing β from zero to one can reduce the error floor by over an order of magnitude. An example of Chuang's simulation results is shown in Figure 6.5. This figure plots the irreducible bit error rate as a function of normalized rms delay spread $d = \sigma_{T_m}/T_s$ for BPSK, QPSK, OQPSK, and MSK modulation assuming a static channel with a Gaussian power delay profile. We see from this figure that for all modulations, we can approximately bound the irreducible error floor as $P_{floor} \leq d^2$ for $.02 \leq d \leq .1$. Other simulation results support this bound as well [28]. This bound imposes severe constraints on data rate even when symbol error probabilities on the order of 10^{-2} are acceptable. For example, the rms delay spread in a typical urban environment is approximately $\sigma_{T_m} = 2.5\mu$sec. To keep $\sigma_{T_m} < .1T_s$ requires that the data rate not exceed 40 Kbaud, which generally isn't enough for high-speed data applications. In rural environments, where multipath is not attenuated to the same degree as in cities, $\sigma_{T_m} \approx 25\mu$sec, which reduces the maximum data rate to 4 Kbaud.

Example 6.8:
Using the approximation $P_{floor} \leq (\sigma_{T_m}/T_s)^2$, find the maximum data rate that can be transmitted through a channel with delay spread $\sigma_{T_m} = 3\mu$ sec using either BPSK or QPSK modulation such that the probability of bit error P_b is less than 10^{-3}.

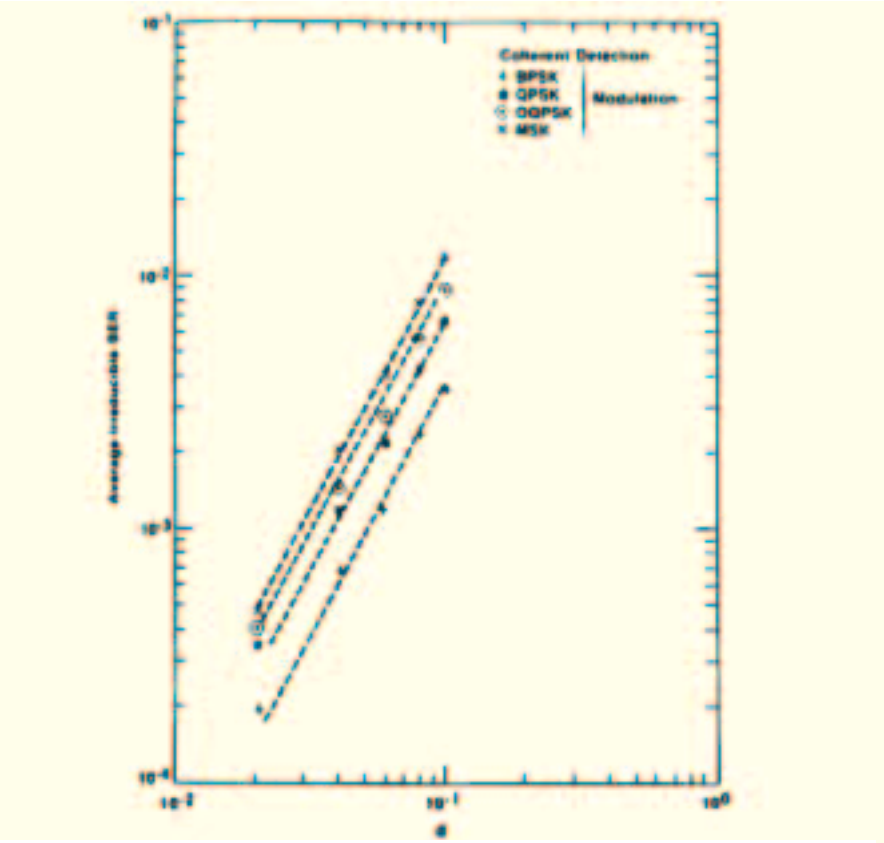


Figure 6.5: Irreducible error versus normalized rms delay spread $d = \sigma_{T_m}/T_s$ for Gaussian power delay profile (from [26] ©IEEE).

Solution: For BPSK, we set $P_{floor} = (\sigma_{T_m}/T_b)^2$, so we require $T_b \geq \sigma_{T_m}/\sqrt{P_{floor}} = 94.87\mu$secs, which leads to a data rate of $R = 1/T_b = 10.54$ Kbps. For QPSK, the same calculation yields $T_s \geq \sigma_{T_m}/\sqrt{P_{floor}} = 94.87\mu$secs. Since there are 2 bits per symbol, this leads to a data rate of $R = 2/T_s = 21.01$ Kbps. This indicates that for a given data rate, QPSK is more robust to ISI than BPSK, due to that fact that its symbol time is slower. This result is also true using the more accurate error floors associated with Figure 6.5 rather than the bound in this example.

Bibliography

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

[2] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection,* Prentice Hall: 1995.

[3] S. Haykin, *An Introduction to Analog and Digital Communications*. New York: Wiley, 1989.

[4] G. L. Stuber, *Principles of Mobile Communications*, Kluwer Academic Publishers, 1996.

[5] J. Craig, "New, simple and exact result for calculating the probability of error for two-dimensional signal constellations," Proc. Milcom 1991.

[6] F. S. Weinstein, "Simplified relationships for the probability distribution of the phase of a sine wave in narrow-band normal noise," *IEEE Trans. on Inform. Theory*, pp. 658–661, Sept. 1974.

[7] R. F. Pawula, "A new formula for MDPSK symbol error probability," *IEEE Commun. Letters*, pp. 271–272, Oct. 1998.

[8] M.K. Simon and D. Divsalar, "Some new twists to problems involving the Gaussian probability integral," *IEEE Trans. Commun.*, pp. 200-210, Feb. 1998.

[9] S. Rhodes, "Effect of noisy phase reference on coherent detection of offset-QPSK signals," *IEEE Trans. Commun.*, Vol 22, No. 8, pp. 1046–1055, Aug. 1974.

[10] N. R. Sollenberger and J. C.-I. Chuang, "Low-overhead symbol timing and carrier recovery for portable TDMA radio systems," *IEEE Trans. Commun.*, Vol 39, No. 10, pp. 1886–1892, Oct. 1990.

[11] R.F. Pawula, "on M-ary DPSK transmission over terrestrial and satellite channels," *IEEE Trans. Commun.*, Vol 32, No. 7, pp. 754–761, July 1984.

[12] W. Cowley and L. Sabel, "The performance of two symbol timing recovery algorithms for PSK demodulators," *IEEE Trans. Commun.*, Vol 42, No. 6, pp. 2345–2355, June 1994.

[13] S. Hinedi, M. Simon, and D. Raphaeli, "The performance of noncoherent orthogonal M-FSK in the presence of timing and frequency errors," *IEEE Trans. Commun.*, Vol 43, No. 2-4, pp. 922–933, Feb./March/April 1995.

[14] E. Grayver and B. Daneshrad, A low-power all-digital FSK receiver for deep space applications," *IEEE Trans. Commun.*, Vol 49, No. 5, pp. 911–921, May 2001.

[15] W.T. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation*, IEEE/Pentech Press, 1994.

[16] X. Tang, M.-S. Alouini, and A. Goldsmith, "Effects of channel estimation error on M-QAM BER performance in Rayleigh fading," *IEEE Trans. Commun.*, Vol 47, No. 12, pp. 1856–1864, Dec. 1999.

[17] P. A. Bello and B.D. Nelin, "The influence of fading spectrum on the bit error probabilities of incoherent and differentially coherent matched filter receivers," *IEEE Trans. Commun. Syst.*, Vol. 10, No. 2, pp. 160–168, June 1962.

[18] M. Schwartz, W.R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw Hill 1966, reprinted by Wily-IEEE Press, 1995.

[19] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels A Unified Approach to Performance Analysis*, Wiley 2000.

[20] P. Y. Kam, "Tight bounds on the bit-error probabilities of 2DPSK and 4DPSK in nonselective Rician fading," *IEEE Trans. Commun.*, pp. 860–862, July 1998.

[21] P. Y. Kam, "Bit error probabilities of MDPSK over the nonselective Rayleigh fading channel with diversity reception," *IEEE Trans. Commun.*, pp. 220–224, Feb. 1991.

[22] V. Fung, R.S. Rappaport, and B. Thoma, "Bit error simulation for $\pi/4$ DQPSK mobile radio communication using two-ray and measurement based impulse response models," *IEEE J. Select. Areas Commun.*, Vol. 11, No. 3, pp. 393–405, April 1993.

[23] S. Chennakeshu and G. J. Saulnier, "Differential detection of $\pi/4$-shifted-DQPSK for digital cellular radio," *IEEE Trans. Vehic. Technol.*, Vol. 42, No. 1, Feb. 1993.

[24] P. A. Bello and B.D. Nelin, "The effects of frequency selective fading on the binary error probabilities of incoherent and differentially coherent matched filter receivers," *IEEE Trans. Commun. Syst.*, Vol 11, pp. 170–186, June 1963.

[25] M. B. Pursley, *Introduction to Digital Communications*, Prentice Hall, 2005.

[26] J. Chuang, "The effects of time delay spread on portable radio communications channels with digital modulation," *IEEE J. Selected Areas Commun.*, Vol. SAC-5, No. 5, pp. 879–889, June 1987.

[27] C. Liu and K. Feher, "Bit error rate performance fo $\pi/4$ DQPSK in a frequency selective fast Rayleigh fading channel," *IEEE Trans. Vehic. Technol.*, Vol. 40, No. 3, pp. 558–568, Aug. 1991.

[28] S. Gurunathan and K. Feher, "Multipath simulation models for mobile radio channels," *Proc. IEEE Vehic. Technol. Conf.* pp. 131–134, May 1992.

Chapter 6 Problems

1. Consider a system in which data is transferred at a rate of 100 bits/sec over the channel.

   (a) Find the symbol duration if we use sinc pulse for signalling and the channel bandwidth is 10 kHz.

   (b) If the received SNR is 10 dB. Find the SNR per symbol and the SNR per bit if 4-QAM is used.

   (c) Find the SNR per symbol and the SNR per bit for 16-QAM and compare with these metrics for 4-QAM.

2. Consider BPSK modulation where the apriori probability of 0 and 1 is not the same. Specifically $p[s_n = 0] = 0.3$ and $p[s_n = 1] = 0.7$.

   (a) Find the probability of bit error P_b in AWGN assuming we encode a **1** as $s_1(t) = A\cos(2\pi f_c t)$ and a **0** as amplitude $s_2(t) = -A\cos(2\pi f_c t$, and the receiver structure is as shown in Figure 5.17.

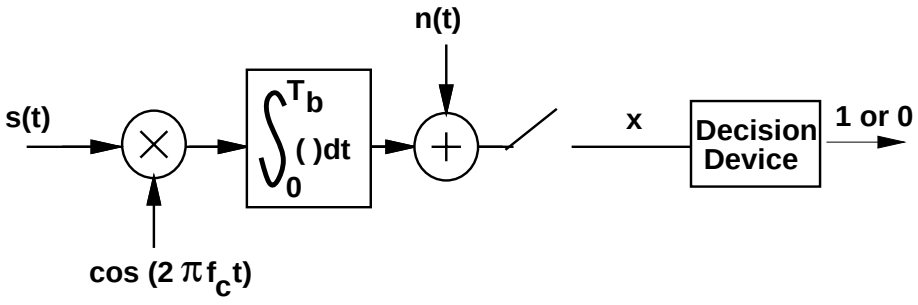
   (b) Suppose you can change the threshold value in the receiver of Figure 5.17. Find the threshold value that yields equal error probability regardless of which bit is transmitted, i.e. the threshold value that yields $p(\hat{m} = 0|m = 1)p(m = 1) = p(\hat{m} = 1|m = 0)p(m = 0)$.

   (c) Now suppose we change the modulation so that $s_1(t) = A\cos(2\pi f_c t)$ and $s_2(t) = -B\cos(2\pi f_c t$. Find A and B so that the receiver of Figure 5.17 with threshold at zero has $p(\hat{m} = 0|m = 1)p(m = 1) = p(\hat{m} = 1|m = 0)p(m = 0)$.

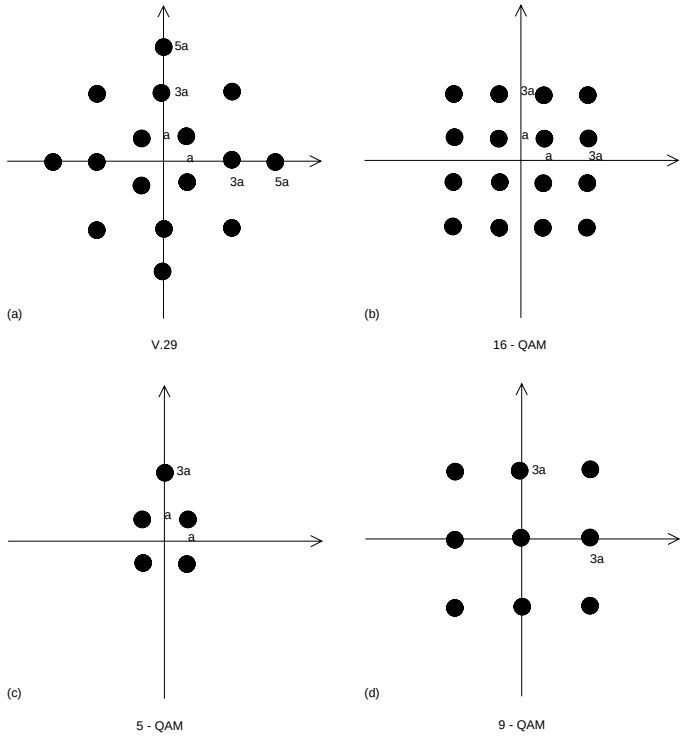
   (d) Compute and compare the expression for P_b in parts (a), (b) and (c) assuming $E_b/N_0 = 10$ dB. For which system is p_b minimized?

3. Consider a BPSK receiver where the demodulator has a phase offset of ϕ relative to the transmitted signal, so for a transmitted signal $s(t) = \pm g(t)\cos(2\pi f_c t)$, the carrier in the demodulator of Figure 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 where the receiver noise is added after the integrator, as shown in the figure below. The decision device outputs a "1" if its input $\mathbf{x}$ has Re$\mathbf{x} \geq 0$, and a "0" otherwise. Suppose the additive noise term $n(t) = 1.1e^{j\theta}$, where $p(\theta = n\pi/3) = 1/6$ for $n = 0, 1, 2, 3, 4, 5$. What is the probability of making a decision error in the decision device, i.e. outputting the wrong demodulated bit, assuming $A_c = \sqrt{2/T_b}$ and that information bits corresponding to a "1" ($s(t) = A_c\cos(2\pi f_c t)$) or a "0" ($s(t) = -A_c\cos(2\pi f_c t)$) are equally likely.



5. Find an approximation to P_s for the following signal constellations:



6. Plot the exact symbol error probability and the approximation from Table 6.1 of 16QAM with $0 \leq \gamma_s \leq 30$ dB. Does the error in the approximation increase or decrease with γ_s and why?

7. Plot the symbol error probability P_s for QPSK using the approximation in Table 6.1 and Craig's exact result for $0 \leq \gamma_s \leq 30$ dB. Does the error in the approximation increase or decrease with γ_s and why?

8. In this problem we derive an algebraic proof of the alternate representation of the Q-function (6.43) from its original representation (6.42). We will work with the complementary error function (erfc) for simplicity and make the conversion at the end. The erfc(x) function is traditionally defined by

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt. \tag{6.95}$$

The alternate representation is of this, corresonding to the alternate representation of the Q-function (6.43) is

$$\text{erfc}(x) = \frac{2}{\pi} \int_0^{\pi/2} e^{-x^2/\sin^2\theta} d\theta. \tag{6.96}$$

(a) Consider the integral

$$I_x(a) \triangleq \int_0^\infty \frac{e^{-at^2}}{x^2 + t^2} dt. \tag{6.97}$$

Show that $I_x(a)$ satisfies the following differential equation:

$$x^2 I_x(a) - \frac{\partial I_x(a)}{\partial a} = \frac{1}{2}\sqrt{\frac{\pi}{a}}. \tag{6.98}$$

(b) Solve the differential equation (6.98) and deduce that

$$I_x(a) \triangleq \int_0^\infty \frac{e^{-at^2}}{x^2+t^2}dt = \frac{\pi}{2x}e^{ax^2}\text{erfc}(x\sqrt{a}). \tag{6.99}$$

`Hint`: $I_x(a)$ is a function in two variables x and a. However, since all our manipulations deal with a only, you can assume x to be a constant while solving the differential equation.

(c) Setting $a = 1$ in (6.99) and making a suitable change of variables in the LHS of (6.99), derive the alternate representation of the erfc function :

$$\text{erfc}(x) = \frac{2}{\pi}\int_0^{\pi/2} e^{-x^2/\sin^2\theta}d\theta$$

(d) Convert this alternate representation of the erfc function to the alternate representation of the Q function.

9. Consider a communication system which uses BPSK signalling with average signal power of 100 Watts and the noise power at the receiver is 4 Watts. Can this system be used for transmission of data? Can it be used for voice? Now consider there is fading with an average SNR $\bar{\gamma}_b = 20$ dB. How does your answer to the above question change?

10. Consider a cellular system at 900 MHz with a transmission rate of 64 Kbps and multipath fading. Explain which performance metric, average probability of error or outage probability, is more appropriate and why for user speeds of 1 mph, 10 mph, and 100 mph.

11. Derive the expression for the moment generating function for Rayleigh fading, Eq. 6.37.

12. This problem illustrates why satellite systems that must compensate for shadow fading are going bankrupt. Consider a LEO satellite system orbiting 500 Km above the earth. Assume the signal follows a free space path loss model with no multipath fading or shadowing. The transmitted signal has a carrier frequency of 900 MHz and a bandwidth of 10 KHz. The handheld receivers have noise spectral density of 10^{-16} (total noise power is N_oB) mW/Hz. Assume nondirectional antennas (0 dB gain) at both the transmitter and receiver. Suppose the satellite must support users in a circular cell on the earth of radius 100 Km at a BER of 10^{-6}.

(a) For DPSK modulation what transmit power is needed such that all users in the cell meet the 10^{-6} BER target.

(b) Repeat part (a) assuming that the channel also experiences log normal shadowing with $\sigma = 8$ dB, and that users in a cell must have $P_b = 10^{-6}$ (for each bit) with probability 0.9.

13. In this problem we explore the power penalty involved in going to higher level signal modulations, i.e. from BPSK to 16PSK.

(a) Find the minimum distance between constellation points in 16PSK modulation as a function of signal energy E_s.

(b) Find α_M and β_M such that the symbol error probability of 16PSK in AWGN is approximately

$$P_s \approx \alpha_M Q\left(\sqrt{\beta_M \gamma_s}\right).$$

(c) Using your expression in part (b), find an approximation for the average symbol error probability of 16PSK in Rayleigh fading in terms of $\overline{\gamma}_s$.

(d) Convert the expressions for average symbol error probability of 16PSK in Rayleigh fading to expressions for average bit error probability assuming Gray coding.

(e) Find the approximate value of $\overline{\gamma}_b$ required to obtain a BER of 10^{-3} in Rayleigh fading for BPSK and 16PSK. What is the power penalty in going to the higher level signal constellation at this BER?

14. Find a closed-form expression for the average probability of error for DPSK modulation in Nakagami-m fading evalute for $m = 4$ and $\overline{\gamma}_b = 10$ dB.

15. The Nakagami distribution is parameterized by m, which ranges from $m = .5$ to $m = \infty$. The m parameter measures the ratio of LOS signal power to multipath power, so $m = 1$ corresponds to Rayleigh fading, $m = \infty$ corresponds to an AWGN channel with no fading, and $m = .5$ corresponds to fading that results in performance that is worse than with a Rayleigh distribution. In this problem we explore the impact of the parameter m on the performance of BPSK modulation in Nakagami fading.

    Plot the average bit error $\overline{P_b}$ of BPSK modulation in Nakagami fading with average SNR ranging from 0 to 20dB for m parameters $m = 1$ (Rayleigh), $m = 2$, and $m = 4$ (The Moment Generating Function technique of Section 6.3.3 should be used to obtain the average error probability). At an average SNR of 10 dB, what is the difference in average BER?

16. Assume a cellular system with log-normal shadowing plus Rayleigh fading. The signal modulation is DPSK. The service provider has determined that it can deal with an outage probability of .01, i.e. 1 in 100 customers are unhappy at any given time. In nonoutage the voice BER requirement is $\overline{P}_b = 10^{-3}$. Assume a noise power spectral density of $N_o = 10^{-16}$ mW/Hz, a signal bandwidth of 30 KHz, a carrier frequency of 900 MHz, free space path loss propagation with nondirectional antennas, and shadowing standard deviation of $\sigma = 6$ dB. Find the maximum cell size that can achieve this performance if the transmit power at the mobiles is limited to 100 mW.

17. In this problem we derive the probability of bit error for DPSK in fast Rayleigh fading. By symmetry, the probability of error is the same for transmitting a zero bit or a one bit. Let us assume that over time kT_b a zero bit is transmitted, so the transmitted symbol at time kT_b is the same as at time $k-1$: $\mathbf{s}(k) = \mathbf{s}(k-1)$. In fast fading the corresponding received symbols are $\mathbf{z}(k-1) = g_{k-1}\mathbf{s}(k-1) + n(k-1)$ and $\mathbf{z}(k) = g_k\mathbf{s}(k-1) + n(k)$, where g_{k-1} and g_k are the fading channel gains associated with transmissions over times $(k-1)T_b$ and kT_b.

    a) Show that the decision variable input to the phase comparator of Figure 5.20 to extract the phase difference is $\mathbf{z}(k)\mathbf{z}^*(k-1) = g_k g_{k-1}^* + g_k\mathbf{s}(k-1)n_{k-1}^* + g_{k-1}^* s_{k-1}^* n_k + n_k n_{k-1}^*$.

    Assuming a reasonable SNR, the last term $n_k n_{k-1}^*$ of this expression can be neglected. Neglecting this term and defining $\tilde{n}_k = s_{k-1}^* n_k$ and $\tilde{n}_{k-1} = s_{k-1}^* n_{k-1}$, we get a new random variable $\tilde{z} = g_k g_{k-1}^* + g_k\tilde{n}_{k-1}^* + g_{k-1}^*\tilde{n}_k$. Given that a zero bit was transmitted over time kT_b, an error is made if $x = \Re\{\tilde{z}\} < 0$, so we must determine the distribution of x. The characteristic function for x is the 2-sided Laplace transform of the pdf of x:

$$\Phi_X(s) = \int_{-\infty}^{\infty} p_X(s)e^{-sx}dx = E[e^{-sx}].$$

This function will have a left plane pole p_1 and a right plane pole p_2, so can be written as

$$\Phi_X(s) = \frac{p_1 p_2}{(s-p_1)(s-p_2)}.$$

The left plane pole p_1 corresponds to the pdf $p_X(x)$ for $x \geq 0$ and the right plane pole corresponds to the pdf $p_X(x)$ for $x < 0$

b) Show through partial fraction expansion that $\Phi_X(s)$ can be written as

$$\Phi_X(s) = \frac{p_1 p_2}{(p_1-p_2)} \frac{1}{(s-p_1)} + \frac{p_1 p_2}{(p_2-p_1)} \frac{1}{(s-p_2)}.$$

An error is made if $x = \Re\{\tilde{z}\} < 0$, so we need only consider the pdf $p_X(x)$ for $x < 0$ corresponding to the second term of $\Phi_X(s)$ in part b).

c) Show that the inverse Laplace transform of the second term of $\Phi_X(s)$ from part b) is

$$p_X(x) = \frac{p_1 p_2}{p_2 - p_1} e^{p_2 x}, \quad x < 0.$$

d) Use part c) to show that $P_b = -p_1/(p_2 - p_1)$.

In $x = \Re\{\tilde{z}\} = \Re\{g_k g_{k-1}^* + g_k \tilde{n}_{k-1}^* + g_{k-1}^* \tilde{n}_k .\}$ the channel gains g_k and g_{k-1} and noises $\tilde{n}_k$ and $\tilde{n}_{k-1}$ are complex Gaussian random variables. Thus, the poles p_1 and p_2 in $p_X(x)$ are derived using the general quadratic form of complex Gaussian random variables [1, Appendix B][18, Appendix B] as

$$p_1 = \frac{-1}{2(\overline{\gamma}_b[1+\rho_c])] + N_0)},$$

and

$$p_2 = \frac{1}{2(\overline{\gamma}_b[1-\rho_c])] + N_0)},$$

for ρ_C the correlation coefficient of the channel over the bit time T_b.

e) Find a general expression for P_b in fast Rayleigh fading using these values of p_1 and p_2 in the P_e expression from part d).

f) Show that this reduces to the average probability of error $\overline{P}_b = \frac{1}{2(1+\overline{\gamma}_b)}$ for a slowly fading channel that does not decorrelate over a bit time.

18. Plot the bit error probability for DPSK in fast Rayleigh fading for $\overline{\gamma}_b$ ranging from 0 to 60 dB and $\rho_C = J_0(2\pi B_D T)$ with $B_D T = .01, .001$, and $.0001$. For each value of $B_d T$, at approximately what value of $\overline{\gamma}_b$ does the error floor dominate the error probability/

19. Find the irreducible error floor for DQPSK modulation due to Doppler, assuming a Gaussian Doppler power spectrum with $B_D = 80$ Hz and Rician fading with $K = 2$.

20. Consider a wireless channel with an average delay spread of 100 nsec and a doppler spread of 80 Hz. Given the error floors due to doppler and ISI, for DQPSK modulation in Rayleigh fading and uniform scattering, approximately what range of data rates can be transmitted over this channel with a BER less than 10^{-4}.

21. Using the error floors of Figure 6.5, find the maximum data rate that can be transmitted through a channel with delay spread $\sigma_{T_m} = 3\mu$ sec using BPSK, QPSK, or MSK modulation such that the probability of bit error P_b is less than 10^{-3}.

Chapter 7

Diversity

In Chapter 6 we saw that both Rayleigh fading and log normal shadowing induce a very large power penalty on the performance of modulation over wireless channels. One of the most powerful techniques to mitigate the effects of fading is to use diversity-combining of independently fading signal paths. Diversity-combining uses 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 some way such that the fading of the resultant signal is reduced. For example, for a system with two antennas that experience independent fading, it is unlikely that both antennas experience deep fades at the same time. By selecting the strongest signal between the two antennas, called selection combining, we will obtain a much better signal than if we just had one antenna. This chapter focuses on the analysis of the most common forms of diversity: selection combining, threshold combining, equal-gain combining, and maximal-ratio combining. Other diversity techniques that have potential benefits over these schemes in terms of performance or complexity are discussed in [1, Chapter 9.10].

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. This requires coordination among the different base stations or access 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 discuss 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 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 space diversity, independent fading paths are realized without an increase in transmit signal power or bandwidth. The separation between antennas must be such that the fading amplitudes corresponding to each antenna are approximately independent. For example, from (3.26) in Chapter 3, in a uniform scattering environment with isotropic 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 (which is common at the base station if the system has cell sectorization), then the multipath is confined to a small angle relative to the LOS ray, which means that a larger antenna

separation is required to get independent fading samples [2].

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 deep fades. 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) since 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 angle of arrival of one of the multipath components (preferably the strongest one). **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. Spread spectrum techniques, discussed in Chapter 13, are sometimes described as providing frequency diversity since the channel gain varies across the bandwidth of the transmitted signal. However, this is not equivalent to sending the same information signal over indepedently fading paths. As discussed in Chapter 13.4, spread spectrum with RAKE reception does provide independently fading paths of the information signal and thus is a form of frequency 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 can also be achieved through coding and interleaving, as will be discussed in Chapter 8. Clearly time diversity can't be used for stationary applications, since the channel coherence time is infinite and thus fading is highly correlated over time.

7.2 Diversity System Model

A diversity system combines the independent fading paths to obtain a resultant signal that is then passed through a standard demodulator. The combining can be done in several ways which vary in complexity and overall performance. We will use space diversity as a reference to describe the 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.

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. Specifically, when all but one of the complex α_is are zero, only one path is passed to the combiner output. When more than one of the α_is is nonzero, the combiner adds together multiple paths, where each path may be weighted by different value. Combining more than one branch signal requires **co-phasing**, where the phase θ_i of the ith branch is removed through the multiplication by $\alpha_i = a_i e^{-j\theta_i}$ for some real-valued a_i. This phase removal requires coherent detection of each branch to determine its phase θ_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.

The multiplication by α_i can be performed either before detection (predetection) or after detection (post-detection) with essentially no difference in performance. Combining is typically performed post-detection, since the branch signal power and/or phase is required to determine the appropriate α_i value. Post-detection combining of multiple branches requires a dedicated receiver for each branch to determine the branch phase, which increases the hardware complexity and power consumption, particular for a large number of branches.

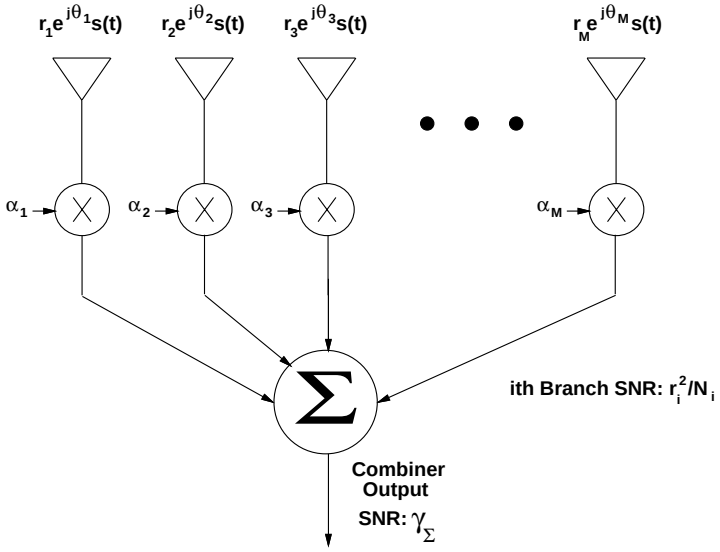


Figure 7.1: Linear Combiner.

The main purpose of diversity is to combine the independent fading paths so that the effects of fading are mitigated. The signal output from the combiner equals the original transmitted signal $s(t)$ multiplied by a random complex amplitude term $\alpha_\Sigma = \sum_i \alpha_i r_i e^{j\theta_i}$ that results from the path combining. This complex 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 shown in more detail below. Since the combiner output is fed into a standard demodulator for the transmitted signal $s(t)$, the performance of the diversity system in terms of $\overline{P}_s$ and P_{out} is as defined in Section 6.3.1, i.e.

$$\overline{P}_s = \int_0^\infty P_s(\gamma) p_{\gamma_\Sigma}(\gamma) d\gamma, \tag{7.1}$$

where $P_s(\gamma_\Sigma)$ is the probability of symbol error for demodulation of $s(t)$ in AWGN with SNR γ_Σ, and

$$P_{out} = p(\gamma_\Sigma \le \gamma_0) = \int_0^{\gamma_0} p_{\gamma_\Sigma}(\gamma) d\gamma, \tag{7.2}$$

for some target SNR value γ_0.

In the following subsections we will describe the different combining techniques and their performance in more detail. These techniques entail various tradeoffs between performance and complexity.

7.3 Selection Combining

In selection combining (SC), the combiner outputs the signal on the branch with the highest SNR r_i^2/N_i. This is equivalent to choosing the branch with the highest $r_i^2 + N_i$ if the noise $N_i = N$ is the same on all branches [1]. Since only one branch is used at a time, SC often 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 simultaneously and continuously monitor SNR on each branch. 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, so this technique can be used with either coherent or differential modulation.

For M branch diversity, the CDF of γ_Σ is given by

$$P_{\gamma_\Sigma}(\gamma) = p(\gamma_\Sigma < \gamma) = p(\max[\gamma_1, \gamma_2, \ldots, \gamma_M] < \gamma) = \prod_{i=1}^{M} p(\gamma_i < \gamma). \tag{7.3}$$

We obtain the pdf of γ_Σ by differentiating $P_{\gamma_\Sigma}(\gamma)$ relative to γ, and 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. The instantaneous SNR on the ith branch is therefore given by $\gamma_i = r_i^2/N$. Defining the average SNR on the ith branch as $\overline{\gamma}_i = \mathrm{E}[\gamma_i]$, the SNR distribution will be exponential:

$$p(\gamma_i) = \frac{1}{\overline{\gamma}_i} e^{-\gamma_i/\overline{\gamma}_i}. \tag{7.4}$$

From (6.47), the outage probability for a target γ_0 on the ith branch in Rayleigh fading is

$$P_{out}(\gamma_0) = 1 - e^{-\gamma_0/\overline{\gamma_i}}. \tag{7.5}$$

The outage probability of the selection-combiner for the target γ_0 is then

$$P_{out}(\gamma_0) == \prod_{i=1}^{M} p(\gamma_i < \gamma_0) = \prod_{i=1}^{M} \left[1 - e^{-\gamma_0/\overline{\gamma}_i}\right]. \tag{7.6}$$

If the average SNR for all of the branches are the same ($\overline{\gamma}_i = \overline{\gamma}$ for all i), then this reduces to

$$P_{out}(\gamma_0) = p(\gamma_\Sigma < \gamma_0) = \left[1 - e^{-\gamma_0/\overline{\gamma}}\right]^M. \tag{7.7}$$

Differentiating (7.7) relative to γ_0 yields the pdf for γ_Σ:

$$p_{\gamma_\Sigma}(\gamma) = p(\gamma_\Sigma = \gamma) = \frac{M}{\overline{\gamma}} \left[1 - e^{-\gamma/\overline{\gamma}}\right]^{M-1} e^{-\gamma/\overline{\gamma}}. \tag{7.8}$$

From (7.8) we see that the average SNR of the combiner output in i.i.d. Rayleigh fading is

$$\begin{aligned} \overline{\gamma}_\Sigma &= \int_0^\infty \gamma p_{\gamma_\Sigma}(\gamma) d\gamma \\ &= \int_0^\infty \frac{\gamma M}{\overline{\gamma}} \left[1 - e^{-\gamma/\overline{\gamma}}\right]^{M-1} e^{-\gamma/\overline{\gamma}} d\gamma \\ &= \overline{\gamma} \sum_{i=1}^{M} \frac{1}{i}. \end{aligned}$$

[1] In practice $r_i^2 + N_i$ is easier to measure than SNR since it just entails find the total power in the received signal.

Thus, the average SNR gain increases 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 SNR gain. This trend is also illustrated in Figure 7.2, which shows P_{out} versus $\overline{\gamma}/\gamma_0$ for different M in i.i.d. Rayleigh fading We see that there is dramatic improvement even with just two-branch selection combining: going from $M = 1$ to $M = 2$ at 1% outage probability there is 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 approximately 7 dB, and from three-branch to four-branch results in an additional reduction of approximately 4 dB. Clearly the power savings is most substantial going from no diversity to two-branch diversity, with diminishing returns as the number of branches is increased.

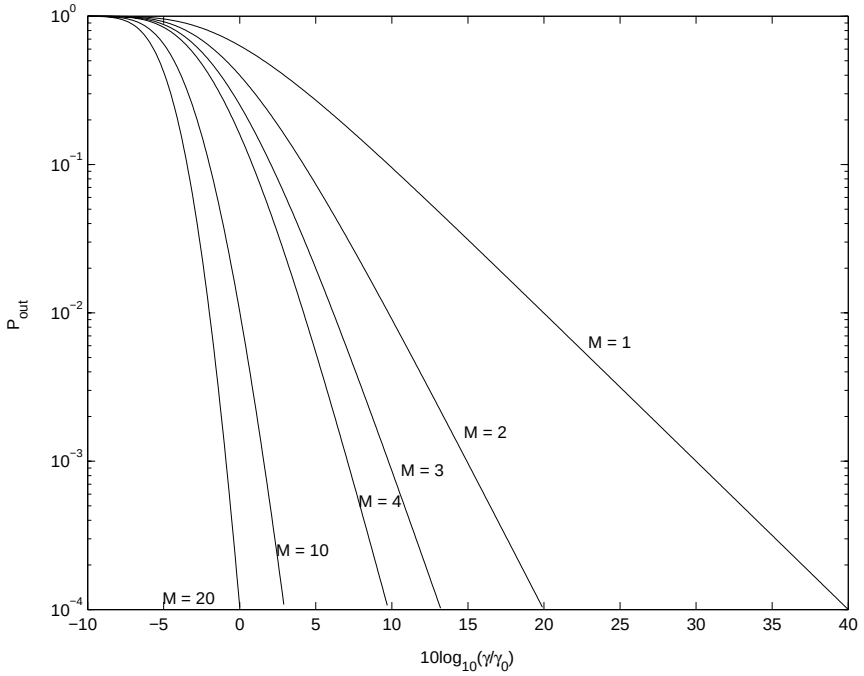


Figure 7.2: Outage Probability of Selection Combining in Rayleigh Fading.

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 $\overline{\gamma} = 15$ dB.

Solution: A BPSK modulated signal with $\gamma_b = 7$ dB has $P_b = 10^{-3}$. Thus, we have $\gamma_0 = 7$ dB. Substituting $\gamma_0 = 10^{.7}$ and $\overline{\gamma} = 10^{1.5}$ into (7.7) yields $P_{out} = .1466$ for $M = 1$, $P_{out} = .0215$ for $M = 2$, and $P_{out} = .0031$ for $M = 2$. We see that each additional branch reduces outage probability by almost an order of magnitude.

The average probability of symbol error is obtained from (7.1) 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. We plot $\overline{P}_b$ versus $\overline{\gamma}_b$ in i.i.d. Rayleigh fading, obtained by a numerical evaluation of $\int Q(\sqrt{2\gamma})p_{\gamma_\Sigma}(\gamma)$ for $p_{\gamma_\Sigma}(\gamma)$ given by (7.8), in Figure 7.3. Closed-form results do exist for differential modulation under i.i.d. Rayleigh fading on each branch [4, Chapter 6.1][1, Chapter 9.7]. For example, it can be shown that for DPSK with $p_{\gamma_\Sigma}(\gamma)$ given by (7.8) the average probability of symbol error is given by

$$\overline{P}_b = \int_0^\infty \frac{1}{2} e^\gamma p_{\gamma_\Sigma}(\gamma) d\gamma = \frac{M}{2} \sum_{m=0}^{M-1} (-1)^m \frac{\binom{M-1}{m}}{1+m+\overline{\gamma}}. \tag{7.9}$$

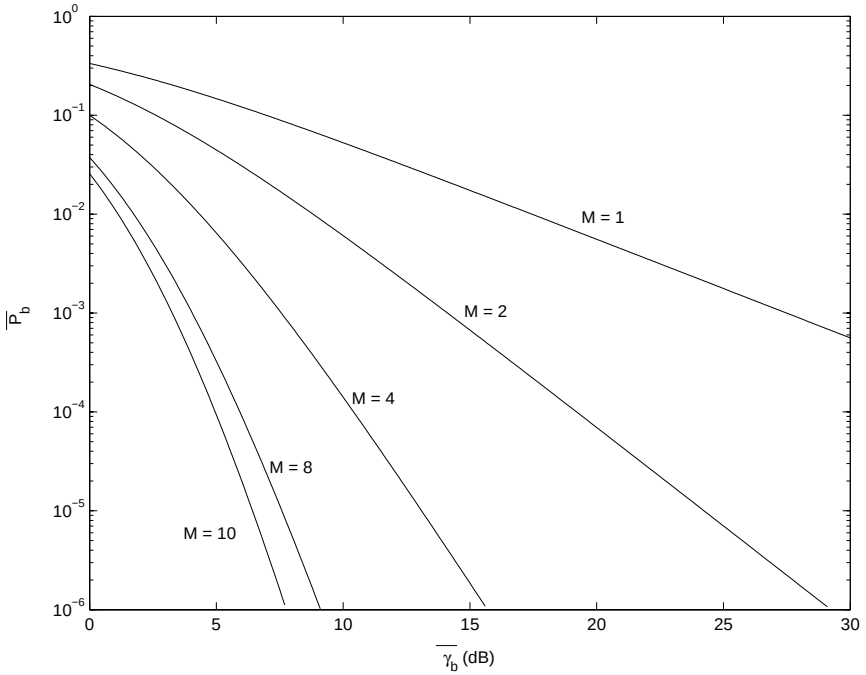


Figure 7.3: $\overline{P}_b$ of BPSK under SC in i.i.d. Rayleigh Fading.

In the above 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 0.5. Derivation of the exact performance degradation due to branch correlation can be found in [1, Chapter 9.7][2].

7.4 Threshold Combining

SC 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 with SNR above a given threshold γ_T. As in SC, since only one branch output is used at

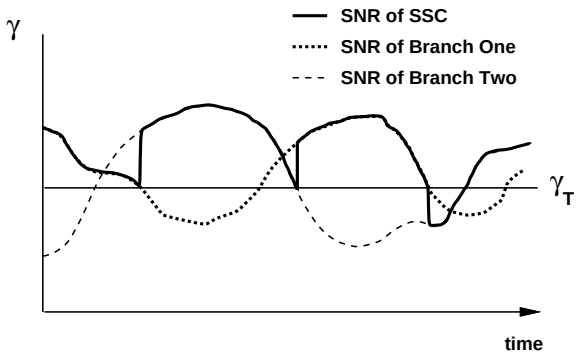


Figure 7.4: SNR of SSC Technique.

a time, co-phasing is not required. Thus, this technique can be used with either coherent or differential modulation.

Once a branch is chosen, as long as the SNR on that branch remains above the desired threshold, the combiner outputs that signal. 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 to decide 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 7.4. Since the SSC does not select the branch with the highest SNR, its performance is between that of no diversity and ideal SC.

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 pdf $p_{\gamma_i}(\gamma)$ of the individual branch SNRs as

$$P_{\gamma_\Sigma}(\gamma) = \begin{cases} P_{\gamma_1}(\gamma_T)P_{\gamma_2}(\gamma) & \gamma < \gamma_T \\ p(\gamma_T \leq \gamma_1 \leq \gamma) + P_{\gamma_1}(\gamma_T)P_{\gamma_2}(\gamma) & \gamma \geq \gamma_T. \end{cases} \tag{7.10}$$

For Rayleigh fading in each branch with $\overline{\gamma}_i = \overline{\gamma}, i = 1, 2$ this yields

$$P_{\gamma_\Sigma}(\gamma) = \begin{cases} 1 - e^{-\gamma_T/\overline{\gamma}} - e^{-\gamma/\overline{\gamma}} + e^{-(\gamma_T+\gamma)/\overline{\gamma}} & \gamma < \gamma_T \\ 1 - 2e^{-\gamma/\overline{\gamma}} + e^{-(\gamma_T+\gamma)/\overline{\gamma}} & \gamma \geq \gamma_T. \end{cases} \tag{7.11}$$

The outage probability P_{out} associated with a given γ_0 is obtained by evaluating $P_{\gamma_\Sigma}(\gamma)$ at $\gamma = \gamma_0$:

$$P_{out}(\gamma_0) = P_{\gamma_\Sigma}(\gamma_0) = \begin{cases} 1 - e^{-\gamma_T/\overline{\gamma}} - e^{-\gamma_0/\overline{\gamma}} + e^{-(\gamma_T+\gamma_0)/\overline{\gamma}} & \gamma < \gamma_T \\ 1 - 2e^{-\gamma_0/\overline{\gamma}} + e^{-(\gamma_T+\gamma_0)/\overline{\gamma}} & \gamma \geq \gamma_T. \end{cases} \tag{7.12}$$

The performance of SSC under other types of fading, as well as the effects of fading correlation, is studied in [1, Chapter 9.8],[6, 7]. In particular, it is shown in [1, Chapter 9.8] that for any fading distribution, SSC with an optimized threshold 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 $\gamma_T = 5, 7$, and 10 dB. Assume

the average branch SNR is $\overline{\gamma} = 15$ dB. Discuss how the outage proability 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, we have $\gamma_0 = 7$ dB. For $\gamma_T = 5$ dB, $\gamma_0 \geq \gamma_T$, so we use the second line of (7.12) to get

$$P_{out} = 1 - 2e^{-10^{.7}/10^{1.5}} + e^{-(10^{.5}+10^{1.5})/10^{1.5}} = .0654.$$

For $\gamma_T = 7$ dB, $\gamma_0 = \gamma_T$, so we again use the second line of (7.12) to get

$$P_{out} = 1 - 2e^{-10^{.7}/10^{1.5}} + e^{-(10^{.7}+10^{1.5})/10^{1.5}} = .0215.$$

For $\gamma_T = 10$ dB, $\gamma_0 < \gamma_T$, so we use the first line of (7.12) to get

$$P_{out} = 1 - e^{-10/10^{1.5}} - e^{-10^{.7}/10^{1.5}} + -e^{-(10+10^{.7})/10^{1.5}} = .0397.$$

We see that the outage probability is smaller for $\gamma_T = 7$ dB than for the other two values. At $\gamma_T = 5$ 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, which contributes to a large outage probability. At $\gamma_T = 10$ 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 indicates that the threshold γ_T that minimizes P_{out} is typically close to the target γ_0.

From Example 7.1, SC has $P_{out} = .0215$. Thus, $\gamma_t = 7$ 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 $P_{out} = .1466$ in Example 7.1.

We obtain the pdf of γ_Σ by differentiating (7.10) relative to γ. Then the average probability of error is obtained from (7.1) with $P_s(\gamma)$ the probability of symbol error in AWGN and $p_{\gamma_\Sigma}(\gamma)$ the pdf of the SSC output SNR. For most fading distributions and coherent modulations, this result cannot be obtained in closed-form and must be evaluated numerically or by approximation. However, for i.i.d. Rayleigh fading we can differentiate (7.11) to get

$$p_{\gamma_\Sigma}(\gamma) = \begin{cases} \left(1 - e^{-\gamma_T/\overline{\gamma}}\right)\frac{1}{\overline{\gamma}}e^{-\gamma/\overline{\gamma}} & \gamma < \gamma_T \\ \left(2 - e^{-\gamma_T/\overline{\gamma}}\right)\frac{1}{\overline{\gamma}}e^{-\gamma/\overline{\gamma}} & \gamma \geq \gamma_T. \end{cases} \tag{7.13}$$

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

$$\overline{P}_b = \int_0^\infty \frac{1}{2}e^{-\gamma}p_{\gamma_\Sigma}(\gamma)d\gamma = \frac{1}{2(1+\overline{\gamma})}\left(1 - e^{-\gamma_T/\overline{\gamma}} + e^{-\gamma_T}e^{-\gamma_T/\overline{\gamma}}.\right) \tag{7.14}$$

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 $\gamma_T = 5, 7$, and 10 dB. Assume the average branch SNR is $\overline{\gamma} = 15$ 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 (7.14) with $\overline{\gamma} = 15$ dB and $\gamma_T = 3, 7$, and 10 dB yields, respectively, $\overline{P}_b = .0029$, $\overline{P}_b = .0023$, $\overline{P}_b = .0042$. 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 (7.9) we have that with SC, $\overline{P}_b = .5(1 + 10^{1.5})^{-1} - .5(2 + 10^{1.5})^{-1} = 4.56 \cdot 10^{-4}$, which is roughly an order of magnitude less than with SSC and an optimized threshold. With no diversity, $\overline{P}_b = .5(1 + 10^{1.5})^{-1} = .0153$, which is roughly an order of magnitude worse than with two-branch SSC.

7.5 Maximal Ratio Combining

In SC and SSC, the output of the combiner equals the signal on one of the branches. In maximal ratio combining (MRC) the output is a weighted sum of all branches, so the α_is in Figure 7.1 are all nonzero. Since the signals are cophased, $\alpha_i = a_i e^{-j\theta_i}$, 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 in each branch yields a total noise power N_{tot} at the combiner output of $N_{tot} = \sum_{i=1}^{M} a_i^2 N$. Thus, the output SNR of the combiner is

$$\gamma_\Sigma = \frac{r^2}{N_{tot}} = \frac{1}{N} \frac{\left(\sum_{i=1}^{M} a_i r_i\right)^2}{\sum_{i=1}^{M} a_i^2}. \tag{7.15}$$

The goal is to chose the α_is to maximize γ_Σ. Intuitively, branches with a high SNR should be weighted more than branches with a low SNR, so the weights a_i^2 should be proportional to the branch SNRs r_i^2/N. We find the a_is that maximize γ_Σ by taking partial derivatives of (7.15) or using the Swartz inequality [2]. Solving for the optimal weights yields $a_i^2 = r_i^2/N$, and the resulting combiner SNR becomes $\gamma_\Sigma = \sum_{i=1}^{M} r_i^2/N = \sum_{i=1}^{M} \gamma_i$. Thus, the SNR of the combiner output is the sum of SNRs on each branch. The average combiner SNR increases linearly with the number of diversity branches M, in contrast to the diminishing returns associated with the average combiner SNR in SC given by (7.9).

To obtain the distribution of γ_Σ we take the product of the exponential moment generating or characteristic functions. Assuming i.i.d. Rayleigh fading on each branch with equal average branch SNR $\overline{\gamma}$, the distribution of γ_Σ is chi-squared with $2M$ degrees of freedom, expected value $\overline{\gamma}_\Sigma = M\overline{\gamma}$, and variance $2M\overline{\gamma}$:

$$p_{\gamma_\Sigma}(\gamma) = \frac{\gamma^{M-1} e^{-\gamma/\overline{\gamma}}}{\overline{\gamma}^M (M-1)!}, \quad \gamma \geq 0. \tag{7.16}$$

The corresponding outage probability for a given threshold γ_0 is given by

$$P_{out} = p(\gamma_\Sigma < \gamma_0) = \int_0^{\gamma_0} p_{\gamma_\Sigma}(\gamma) d\gamma = 1 - e^{-\gamma_0/\overline{\gamma}} \sum_{k=1}^{M} \frac{(\gamma_0/\overline{\gamma})^{k-1}}{(k-1)!}. \tag{7.17}$$

Figure 7.5 plots P_{out} for maximal ratio combining indexed by the number of diversity branches.

The average probability of symbol error is obtained from (7.1) with $P_s(\gamma)$ the probability of symbol error in AWGN for the signal modulation and $p_{\gamma_\Sigma}(\gamma)$ the pdf of γ_Σ. For BPSK modulation with i.i.d

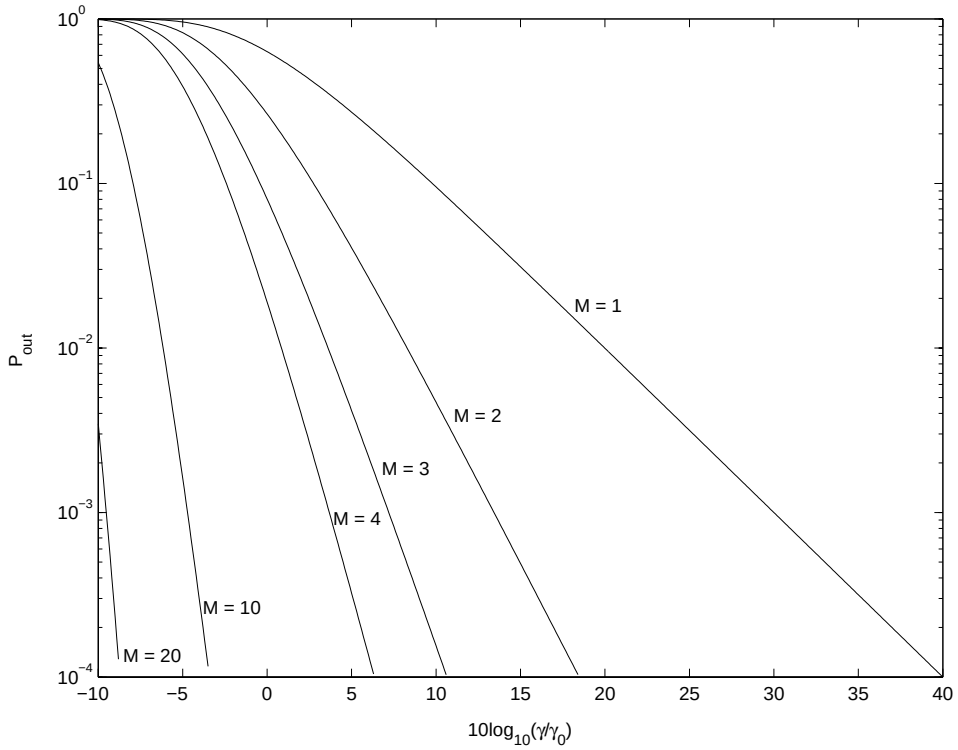


Figure 7.5: P_{out} for MRC with i.i.d. Rayleigh fading.

Rayleigh fading, where $p_{\gamma_\Sigma}(\gamma)$ is given by (7.16), it can be shown that [4, Chapter 6.3]

$$\overline{P}_b = \int_0^\infty Q(\sqrt{2\gamma}) p_{\gamma_\Sigma}(\gamma) d\gamma = \left(\frac{1-\Gamma}{2}\right)^M \sum_{m=0}^{M-1} \left(\begin{array}{c} M-1+m \\ m \end{array} \right) \left(\frac{1+\Gamma}{2}\right)^m, \tag{7.18}$$

where $\Gamma = \sqrt{\overline{\gamma}/(1+\overline{\gamma})}$. This equation is plotted in Figure 7.6. Comparing the outage probability for MRC in Figure 7.5 with that of SC in Figure 7.2 or the average probability of error for MRC in Figure 7.6 with that of SC in Figure 7.3 indicates that MRC has significantly better performance than SC. In Section 7.7 we will use a different analysis based on MGFs 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.

7.6 Equal-Gain Combining

MRC requires knowledge of the time-varying SNR on each branch, which can be very difficult to measure. A simpler technique is equal-gain combining, which co-phases the signals on each branch and then combines them with equal weighting, $\alpha_i = e^{-\theta_i}$. The SNR of the combiner output, assuming equal noise power N in each branch, is then given by

$$\gamma_\Sigma = \frac{1}{NM} \left(\sum_{i=1}^M r_i \right)^2. \tag{7.19}$$

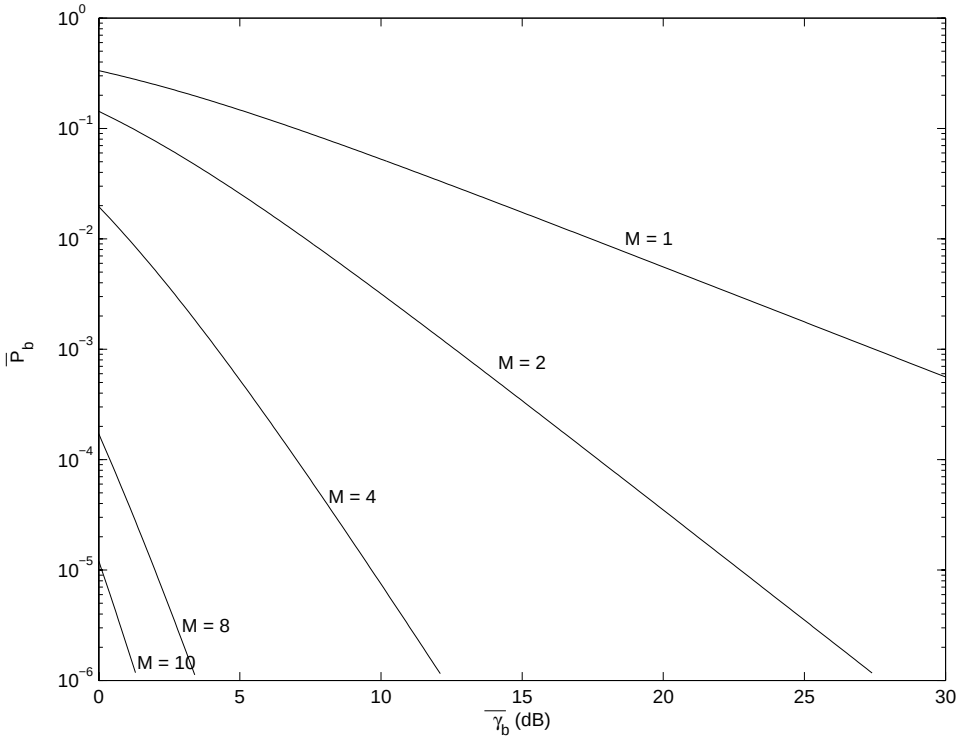


Figure 7.6: $\overline{P}_b$ for MRC with i.i.d. Rayleigh fading.

The pdf and CDF of γ_Σ do not exist in closed-form. For i.i.d. Rayleigh fading and two-branch diversity and average branch SNR $\overline{\gamma}$, an expression for the CDF in terms of the Q function can be derived as [8, Chapter 5.6][4, Chapter 6.4]

$$P_{\gamma_\Sigma}(\gamma) = 1 - e^{-2\gamma/\overline{\gamma}}\sqrt{\frac{\pi\gamma}{\overline{\gamma}}}e^{-\gamma/\overline{\gamma}}\left(1 - 2Q\left(\sqrt{2\gamma/\overline{\gamma}}\right)\right). \tag{7.20}$$

The resulting outage probability is given by

$$P_{out}(\gamma_0) = 1 - e^{-2\gamma_R} - \sqrt{\pi\gamma_R}e^{-\gamma_R}\left(1 - 2Q\left(\sqrt{2\gamma_R}\right)\right), \tag{7.21}$$

where $\gamma_R = \gamma_0/\overline{\gamma}$. Differentiating (7.20) relative to γ yields the pdf

$$p_{\gamma_\Sigma}(\gamma) = \frac{1}{\overline{\gamma}}e^{-2\gamma/\overline{\gamma}} + \sqrt{\pi}e^{-\gamma/\overline{\gamma}}\left(\frac{1}{\sqrt{4\gamma\overline{\gamma}}} - \frac{1}{\overline{\gamma}}\sqrt{\frac{\gamma}{\overline{\gamma}}}\right)\left(1 + 2Q(\sqrt{2\gamma/\overline{\gamma}})\right). \tag{7.22}$$

Substituting this into (7.1) for BPSK yields the average probability of bit error

$$\overline{P}_b = \int_0^\infty Q(\sqrt{2\gamma})p_{\gamma_\Sigma}(\gamma)d\gamma = .5\left(1 - \sqrt{1 - \left(\frac{1}{1+\overline{\gamma}}\right)^2}\right). \tag{7.23}$$

It is shown in [8, Chapter 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, Chapter 9].

Example 7.4: Compare the average probability of bit error of BPSK under MRC and EGC two-branch diversity with i.i.d. Rayleigh fading with average SNR of 10 dB on each branch.

Solution: From (7.18), under MRC we have

$$\overline{P}_b = \left(\frac{1-\sqrt{10/11}}{2}\right)^2 \left(2+\sqrt{10/11}\right) = 1.60 \cdot 10^{-3}.$$

From (7.23), under EGC we have

$$\overline{P}_b = .5\left(1-\sqrt{1-\left(\frac{1}{11}\right)^2}\right) = 2.07 \cdot 10^{-3}.$$

So we see that the performance of MRC and EGC are almost the same.

7.7 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 pdf $p_{\gamma_\Sigma}(\gamma)$ 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 (7.1) and (7.2), 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. The pdf for $p_{\gamma_\Sigma}(\gamma)$ is often in the form of an infinite-range integral, in which case the expressions for (7.1) and (7.2) become double integrals that can be difficult to evaluate numerically. Even when $p_{\gamma_\Sigma}(\gamma)$ is in closed form, the corresponding integrals (7.1) and (7.2 may not lead to closed-form solutions and may be difficult to evaluate numerically. A large body of work over many decades has addressed approximations and numerical techniques to compute the integrals associated with average probability of symbol error for different modulations, fading distributions, and combining techniques (see [9] and the references therein). Expressing the average error probability in terms of the MGF for γ_Σ instead of its pdf often eliminates these integration difficulties. Specifically, when the diversity fading paths that 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 and SC.

7.7.1 Diversity Analysis for MRC

The simplicity of using MGFs in the analysis of MRC stems from the fact that, as derived in Section 7.5, the combiner SNR γ_Σ is the sum of branch SNRS γ_i:

$$\gamma_\Sigma = \sum_{i=1}^{M} \gamma_i. \tag{7.24}$$

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 (6.67), or as a finite range integral of such a function, as in (6.68).

We first consider the case where P_s is in the form of (6.67). Then the average probability of symbol error under MRC is

$$\overline{P}_s = \int_0^\infty c_1 \exp[-c_2\gamma] p_{\gamma_\Sigma}(\gamma) d\gamma. \tag{7.25}$$

We assume that the branch SNRs are independent, so that their joint pdf becomes a product of the individual pdfs: $p_{\gamma_1,\ldots,\gamma_M}(\gamma_1,\ldots,\gamma_M) = p_{\gamma_1}(\gamma_1)\ldots p_{\gamma_M}(\gamma_M)$. Using this factorization and substituting $\gamma = \gamma_1 + \ldots + \gamma_M$ in (7.25) yields

$$\overline{P}_s = c_1 \underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{M-\text{fold}} \exp[-c_2(\gamma_1 + \ldots + \gamma_M)] p_{\gamma_1}(\gamma_1)\ldots p_{\gamma_M}(\gamma_M) d\gamma_1 \ldots d\gamma_M. \tag{7.26}$$

Now using the product forms $\exp[-\beta(\gamma_1 + \ldots + \gamma_M)] = \prod_{i=1}^M \exp[-\beta\gamma_i]$ and $p_{\gamma_1}(\gamma_1)\ldots p_{\gamma_M}(\gamma_M) = \prod_{i=1}^M p_{\gamma_i}(\gamma_i)$ in (7.26) yields

$$\overline{P}_s = c_1 \underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{M-\text{fold}} \prod_{i=1}^M \exp[-c_2\gamma_i] p_{\gamma_i}(\gamma_i) d\gamma_i. \tag{7.27}$$

Finally, switching the order of integration and multiplication in (7.27) yields our desired final form

$$\overline{P}_s = c_1 \prod_{i=1}^M \int_0^\infty \exp[-c_2\gamma_i] p_{\gamma_i}(\gamma_i) d\gamma_i = \alpha \prod_{i=1}^M M_{\gamma_i}(-\beta). \tag{7.28}$$

Thus, the average probability of symbol error is just the product of MGFs associated with the SNR on each branch.

Similary, when P_s is in the form of (6.68), we get

$$\overline{P}_s = \int_0^\infty \int_A^B c_1 \exp[-c_2(x)\gamma] dx p_{\gamma_\Sigma}(\gamma) d\gamma = \underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{M-\text{fold}} \int_A^B c_1 \prod_{i=1}^M \exp[-c_2(x)\gamma_i] p_{\gamma_i}(\gamma_i) d\gamma_i. \tag{7.29}$$

Again switching the order of integration and multiplication yields our desired final form

$$\overline{P}_s = c_1 \int_A^B \prod_{i=1}^M \int_0^\infty \exp[-c_2(x)\gamma_i] p_{\gamma_i}(\gamma_i) d\gamma_i = c_1 \int_A^B \prod_{i=1}^M M_{\gamma_i}(-c_2(x)) dx. \tag{7.30}$$

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 (7.28) and (7.30) 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 (6.67) with $c_1 = 1/2$ and $c_2 = 1$. Thus, from (7.28), the average probability of bit error in DPSK under M-fold MRC diversity is

$$\overline{P}_b = \frac{1}{2} \prod_{i=1}^M M_{\gamma_i}(-1), \tag{7.31}$$

where $M_{\gamma_i}(s)$ is the MGF of the fading distribution for the ith diversity branch, given by (6.63), (6.64), and (6.65) for, respectively, Rayleigh, Ricean, and Nakagami fading. Note that this reduces to the probability of average bit error without diversity given by (6.60) 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 $\overline{\gamma}_1 = 15$ dB and $\overline{\gamma}_2 = \overline{\gamma}_3 = 5$ dB. Compare with the case of no diversity with $\overline{\gamma} = 15$ dB.

Solution: From (6.63), $M_{\gamma_i}(s) = (1 - s\overline{\gamma}_i)^{-1}$ Using this MGF in (7.31) with $s = -1$ yields

$$\overline{P}_b = \frac{1}{2}\frac{1}{1+10^{1.5}}\left(\frac{1}{1+10^5}\right)^2 = 8.85 * 10^{-4}.$$

With no diversity we have

$$\overline{P}_b = \frac{1}{2(1+10^{1.5})} = 1.53 * 10^{-2}.$$

This indicates that additional diversity branches can significantly reduce average BER, even when the SNR on this 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 $\overline{\gamma}_1 = 15$ dB, Ricean fading in the second branch with $K = 3$ and $\overline{\gamma}_2 = 5$, and Nakagami fading in the third branch with $m = 4$ and $\overline{\gamma}_3 = 5$ dB. Compare with the results of the prior example.
Solution: From (6.64) and (6.65), for Nakagami fading $M_{\gamma_i}(s) = (1 - s\overline{\gamma}_i/m)^{-m}$ and for Riciean fading

$$\mathcal{M}_{\gamma_s}(s) = \frac{1+K}{1+K-s\overline{\gamma}_s}\exp\left[\frac{Ks\overline{\gamma}_s}{1+K-s\overline{\gamma}_s}\right].$$

Using these MGFs in (7.31) with $s = -1$ yields

$$\overline{P}_b = \frac{1}{2}\left(\frac{1}{1+10^{1.5}/2}\right)^2\frac{4}{4+10^{.5}}\exp[-3\cdot 10^{.5}/(4+10^{.5})]\left(\frac{1}{1+10^{.5}/4}\right)^4 = 6.9\cdot 10^{-5}$$

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 Ricean fading are a 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 just consists of multiplying together different functions in closed-form, whose result is then also in closed-form. Computing the pdf of the sum of random variables from different families involves the convolution of their pdfs, which rarely leads to a closed-form pdf.

For BPSK we see from (6.44) that P_b has the same form as (6.68) with the integration over ϕ where $c_1 = 1/\pi$, $A = 0$, $B = \pi/2$, and $c_2(\phi) = 1/\sin^2\phi$. Thus we obtain the average bit error probability for

BPSK with M-fold diversity as

$$\overline{P}_b = \frac{1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{M} M_{\gamma_i}\left(-\frac{1}{\sin^2\phi}\right)d\phi. \tag{7.32}$$

Similarly, if $P_s = \alpha Q(\sqrt{2g\gamma_s})$ then P_s has the same form as (6.68) with integration over ϕ, $c_1 = 1/\pi$, $A = 0$, $B = \pi/2$, and $c_2(\phi) = g/\sin^2\phi$, and the resulting average symbol error probability with M-fold diversity is given by

$$\overline{P}_s = \frac{1}{\pi}\int_0^{\pi/2}\prod_{i=1}^{M} M_{\gamma_i}\left(-\frac{g}{\sin^2\phi}\right)d\phi. \tag{7.33}$$

If the branch SNRs are i.i.d. then this simplifies to

$$\overline{P}_s = \frac{1}{\pi}\int_0^{\pi/2}\left(M_{\gamma}\left(-\frac{g}{\sin^2\phi}\right)\right)^{M} d\phi, \tag{7.34}$$

where $M_\gamma(s)$ is the common MGF for the branch SNRs. The probability of symbol error for MPSK in (6.45) is also in the form (6.68), leading to average symbol error probability

$$\overline{P}_s = \frac{1}{\pi}\int_0^{\frac{(M-1)\pi}{M}}\prod_{i=1}^{M} M_{\gamma_i}\left(-\frac{g}{\sin^2\phi}\right)d\phi. \tag{7.35}$$

where $g = \sin^2\left(\frac{\pi}{M}\right)$. For i.i.d. fading this simplifies to

$$\overline{P}_s = \frac{1}{\pi}\int_0^{\frac{(M-1)\pi}{M}}\left(M_{\gamma}\left(-\frac{g}{\sin^2\phi}\right)\right)^{M} d\phi. \tag{7.36}$$

Example 7.7: Find an expression for the average symbol error probability for 8PSK modulation for two-branch MRC combining, where each branch is Rayleigh fading with average SNR of 20 dB.

Solution The MGF for Rayleigh is $M_{\gamma_i}(s) = (1-s\overline{\gamma}_i)^{-1}$. Using this MGF in (7.36) with $s = -\sin^2\pi/8/\sin^2\phi$ and $\overline{\gamma} = 100$ yields

$$\overline{P}_s = \frac{1}{\pi}\int_0^{7\pi/8}\left(\frac{1}{1+\frac{100\sin^2\pi/8}{\sin^2\phi}}\right)^{2} d\phi.$$

This expression does not lead to a closed-form solution and so must be evaluated numerically, which results in $\overline{P}_s = 1.56\cdot 10^{-3}$.

We can use similar techniques to extend the derivation of the exact error probability for MQAM in fading, given by (7.37), to include MRC diversity. Specifically, we first integrate the expression for P_s in AWGN, expressed in (6.80) 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 used above in the MPSK derivation. The resulting average probability of symbol error for MQAM modulation with MRC combining is given by

$$\overline{P}_s = \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\prod_{i=1}^{M}\mathcal{M}_{\gamma_i}\left(-\frac{g}{\sin^2\phi}\right)\,d\phi - \frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\prod_{i=1}^{M}\mathcal{M}_{\gamma_i}\left(-\frac{g}{\sin^2\phi}\right)\,d\phi. \quad (7.37)$$

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 [19, Chapter 9.2].

7.7.2 Diversity Analysis for EGC and SC

MGFs are less useful in the analysis of EGC and SC than in MRC. The reason is that with MRC, $\gamma_\Sigma = \sum_i \gamma_i$, so $\exp[-c_2\gamma_\Sigma] = \prod_i \exp[-c_2\gamma_i]$ 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 pdf. This is illustrated in [19, Chapter 9.3.3], where the exact probability of symbol error for MPSK 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 pdf. The resulting average error probability, given by [19, Equation 9.78], is a finite-range integral over a sum of closed-form expressions, and is thus easily evaluated numerically.

7.7.3 Diversity Analysis for Noncoherent and Differentially Coherent Modulation

A similar approach to determining the average symbol error probability of noncoherent and differentially coherent modulations with diversity combining is presented in [10, 19]. 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 leads to BER expressions involving a single finite-range integral that can be readily evaluated numerically. More details on this approach can be found in [10] and [19].

7.8 Transmitter Diversity

In transmit diversity there are multiple antennas available at the transmitter, and the transmitted signal $s(t)$ is sent over the ith antenna with a branch weight α_i. Transmit diversity is desirable in systems such as cellular systems where more space, power, processing capability is available on the transmit side versus the receive side. The path gain associated with the ith antenna is $r_i e^{j\theta_i}$, and the signals transmitted over all antennas are added "in the air", which leads to a received signal given by

$$r(t) = \sum_{i=1}^{M} \alpha_i r_i e^{j\theta_i} s(t). \quad (7.38)$$

This system works the same as with receiver diversity assuming that the transmit weights α_i are the same as if the diversity was implemented at the receiver, and the analysis is then also identical. Thus, transmitter diversity provides the same diversity gains as 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 in transmitting back to the mobile, since under time-division the forward and reverse links are reciprocal.

Other forms of transmit diversity use space-time coding. A particularly simple and prevalent scheme for this was developed by Alamouti in [11]. In Alamouti's space-time code, two transmit antennas use a simple repetition code, which is decoded in the receiver using maximum-likelihood decoding. This scheme provides the same diversity gain as two-branch MRC in the receiver, but also requires knowledge of the channel gain and phase associated with each of the transmit antennas. More details on Alamouti's code and more sophisticated space-time codes are discussed in Chapter 10.2.

Bibliography

[1] M. Simon and M.-S. Alouini, *Digital Communication over Fading Channels A Unified Approach to Performance Analysis*. Wiley, 2000.

[2] W. Lee, *Mobile Communications Engineering*. New York: McGraw-Hill, 1982.

[3] J. Winters, "Signal acquisition and tracking with adaptive arrays in the digital mobile radio system is-54 with flat fading," *IEEE Trans. Vehic. Technol.*, vol. 43, pp. 1740–1751, Nov. 1993.

[4] G. L. Stuber, *Principles of Mobile Communications, 2nd Ed.* Kluwer Academic Publishers, 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. 1887–1895, Dec. 1979.

[6] A. Abu-Dayya and N. Beaulieu, "Switched diversity on microcellular ricean channels," *IEEE Trans. Vehic. Technol.*, pp. 970–976, Nov. 1994.

[7] A. Abu-Dayya and N. Beaulieu, "Analysis of switched diversity systems on generalized-fading channels," *IEEE Trans. Commun.*, pp. 2959–2966, Nov. 1994.

[8] M. Yacoub, *Principles of Mobile Radio Engineering*. CRC Press, 1993.

[9] M. K. Simon and M. -S. Alouini, "A unified approach to the performance analysis of digital communications over generalized fading channels," *Proc. IEEE*, vol. 86, pp. 1860–1877, September 1998.

[10] 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.*, vol. COM-46, pp. 1625–1638, December 1998.

[11] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, pp. 1451–1458, Oct. 1998.

[12] E. Visotsky and U. Madhow, "Space-time transmit precoding with imperfect feedback," *Proc. Intl. Symp. Inform. Theory*, pp. 357–366, June 2000.

[13] A. Paulraj, "Space-time modems for wireless personal communications," *IEEE Pers. Commun. Mag.*, vol. 5, pp. 36–48, Feb. 1998.

[14] R. Kohno, "Spatial and temporal communication theory using adaptive antenna array," *IEEE Pers. Commun. Mag.*, vol. 5, pp. 36–48, Feb. 1998.

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 $\overline{\gamma}_1 = 10$ dB, $\overline{\gamma}_2 = 15$ dB, and $\overline{\gamma}_3 = 20$ dB

2. Plot the pdf $p_{\gamma_\Sigma}(\gamma)$ given by (7.8) for the selection-combiner SNR in Rayleigh fading with M branch diversity assuming $M = 1, 2, 4, 8$, and 10. Assume each branch has 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 pdf changes with increasing M and why that leads to lower probability of error.

3. Derive the average probability of bit error for DPSK under SC with i.i.d. Rayleigh fading on each branch as given by (7.9).

4. Derive a general expression for the CDF of the SSC output SNR for branch statistics that are not i.i.d. and show that it reduces to (7.10) for i.i.d. branch statistics. Evaluate your expression assuming Rayleigh fading in each branch with different average SNRs $\overline{\gamma}_1$ and $\overline{\gamma}_2$.

5. Derive the average probability of bit error for DPSK under SSC with i.i.d. Rayleigh fading on each branch as given by (7.14).

6. Compare the average probability of bit error for DPSK under no diversity, SC, and SSC, assuming i.i.d. Rayleigh fading on each branch and an average branch SNR of 10 dB and of 20 dB. How does the relative performance change as the branch SNR increases.

7. Plot the average probability of bit error for DPSK under SSC with $M = 2, 3$, and 4, assuming i.i.d. Rayleigh fading on each branch and an average branch SNR ranging from 0 to 20 dB.

8. Show that the weights α_i that maximize γ_Σ under MRC are $\alpha_i^2 = r_i^2/N$ for N the common noise power on each branch. Also show that with these weights, $\gamma_\Sigma = \sum_i \gamma_i$.

9. Derive the average probability of bit error for BPSK under MRC with i.i.d. Rayleigh fading on each branch as given by (7.18).

10. Derive the average probability of bit error for BPSK under EGC with i.i.d. Rayleigh fading on each branch as given by (7.23).

11. Compare the average probability of bit error for BPSK modulation under no diversity, two-branch SC, two-branch SSC, two-branch EGC, and two-branch MRC. Assume i.i.d. Rayleigh fading on each branch with equal branch SNR of 10 dB and of 20 dB. How does the relative performance change as the branch SNR increases.

12. 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 to 20 dB. What is the maximum dB penalty of EGC as compared to MRC?

13. Compare the outage probability of BPSK modulation at $P_b = 10^{-3}$ under MRC and under EGC assuming two-branch diversity with i.i.d. Rayleigh fading on each branch and average branch SNR $\overline{\gamma}$=10 dB.

14. Compare the average probability of bit error for BPSK under MRC and under EGC assuming two-branch diversity with i.i.d. Rayleigh fading on each branch and average branch SNR $\overline{\gamma}$=10 dB.

15. Consider a fading channel with BPSK modulation, 3 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 $m = 2$ (Using the alternate Q function representation greatly simplifies this computation, at least for Nakagami fading).

16. Consider a fading channel with BPSK modulation, 3 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 $m = 2$ (Using the alternate Q function representation greatly simplifies this computation, at least for Nakagami fading).

17. 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, and your plots should have that average branch SNR ranging from 5 to 20 dB.

18. Plot the average probability of error as a function of branch SNR for an M-branch MRC system with 8PSK modulation for $M = 1, 2, 4, 8$. Assume each branch has Rayleigh fading with the same average SNR. Your plots should have an SNR that ranges from 5 to 20 dB.

19. Derive the average probability of symbol error for MQAM modulation under MRC diversity given by (7.37) from the probability of error in AWGN (6.80) by utilizing the alternate representation of Q and Q^2,

20. Compare the average probability of symbol error for 16PSK and 16QAM modulation, assuming three-branch MRC diversity with Rayleigh fading on the first branch and Ricean fading on the second and third branches with $K = 2$. Assume equal average branch SNRs of 10 dB.

21. Plot the average probability of error as a function of branch SNR for an M-branch MRC system with 16QAM modulation for $M = 1, 2, 4, 8$. Assume each branch has Rayleigh fading with the same average SNR. Your plots should have an SNR that ranges from 5 to 20 dB.

Chapter 8

Coding for Wireless Channels

Error correction codes provide the capability for 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. In this chapter we describe codes designed for errors introduced by AWGN channels and by fading channels. Fading channel codes are either designed specifically for fading channels or are based on using AWGN channel codes combined with interleaving. The basic idea behind coding and interleaving is to randomizes the location of errors that occur in bursts, since most codes designed for AWGN channels do not work well when there is a long sequence of errors. Thus, the interleaver disburses the location of errors occuring in bursts such that just a few simultaneous errors occur, which can typically be corrected by most AWGN codes.

We first discuss the basic design parameters in coding, including coding gain, rate penalty, and bandwidth expansion. We next describe block and convolutional codes. These coding methods have been around for many decades, but often require increased bandwidth or reduced data rate in exchange for their error correction capabilities. We will also discuss the use of block or convolutional codes in conjunction with a block or convolutional interleaver for use in fading channels. Concatenated codes and their evolution to turbo and low density parity check codes are covered next. These extremely powerful codes exhibit near-capacity performance with reasonable complexity levels, and are being implemented in current wireless standards. All of these coding techniques provide coding gain at a cost of increased bandwidth or reduced data rate. Trellis codes were invented in the late 1970s as a technique to obtain error correction without rate or bandwidth penalty through a joint design of the modulation and coding. We will discuss the basic design principle behind trellis and more general lattice codes for both AWGN and fading channels. We conclude the chapter with a discussion of unequal error protection codes and joint source and channel coding.

8.1 Code Design Considerations

The main reason to apply error correction coding in a wireless system is to reduce the probability of bit error P_b. The amount of P_b reduction is typically characterized by the coding gain of a code. Coding gain for a given code is defined as the amount that the SNR can be reduced under the coding technique for a given P_b. Coding gain is illustrated 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 some codes have negative coding gain at low SNRs, since the extra redundancy required in the code does not provide sufficient performance gain at these SNRs to yield a positive coding gain. Obviously, a system designer tries to avoid using a code with negative coding gain. However, since the coding gain varies with

SNR, it might not be possible to insure that any given code has a positive coding gain due to random fluctuations in the channel SNR.

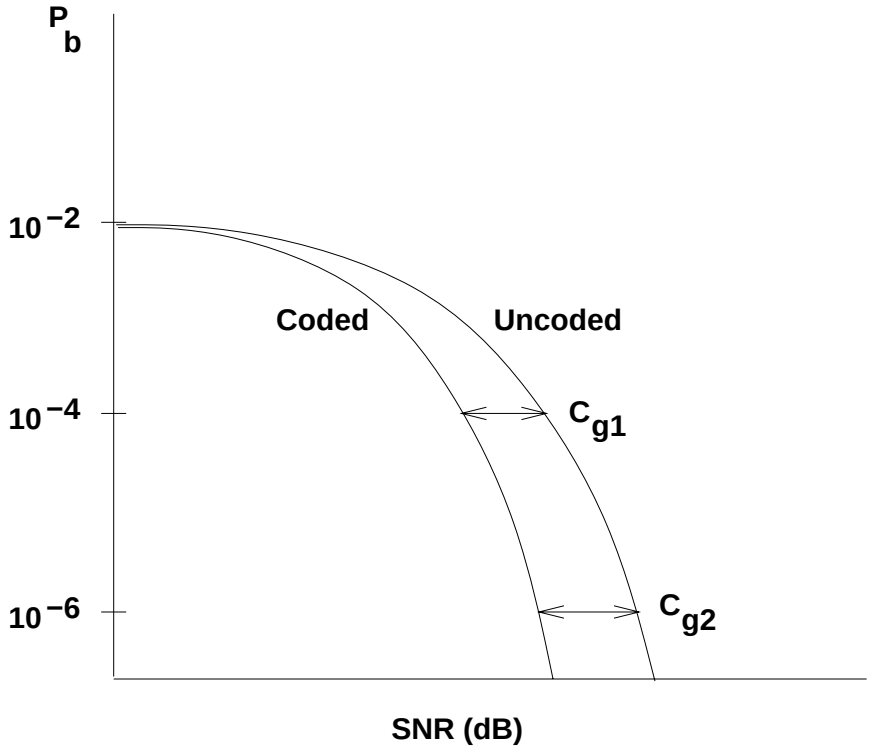


Figure 8.1: Coding Gain.

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 decreased data rate or increase in signal bandwidth. Specifically, 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 $\frac{k}{n}R_b$, i.e. coding decreases the data rate by the fraction k/n. However, 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 a joint design of the code and modulation to obtain coding gain without this bandwidth expansion, as discussed in more detail in Section 8.7.

8.2 Linear Block Codes

Linear block codes are conceptually simple codes that are basically an extension of 1-bit parity check codes for error detection. A 1-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 1s in a block is odd or even. Thus, if a single error occurs, either the parity bit is corrupted or the number of detected 1s 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 1s 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 will be discussed in more detail below. 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, such that 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/second, then the information rate associated with an (n, k) block code is $R_b = R_c R_s = \frac{k}{n} R_s$ bits/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, such that 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:

- $S_1 = \{000, 001, 100, 101\}$
- $S_2 = \{000, 100, 110, 111\}$
- $S_3 = \{001, 100, 101\}$

Solution: It is easily verified that S_1 is a subspace, since it contains the all-zero vector and the sum of any two tuples in S_1 is also in S_1. S_2 is not a subspace since it is not closed under addition, as $110 + 111 = 001 \notin S_2$. S_3 is not a subspace since, although it is closed under addition, 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 $\mathbf{C_m}(l)$. 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 1s 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} with 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 $0 \oplus 0 = 1 \oplus 1 = 0$, the Hamming distance between $\mathbf{C_i}$ and $\mathbf{C_j}$ is equal to the weight of $\mathbf{C_i} \oplus \mathbf{C_j}$. For example, with $\mathbf{C_i} = [00101]$ and $\mathbf{C_j} = [10011]$ as given above, $w(\mathbf{C_i}) = 2$, $w(\mathbf{C_j}) = 3$, and $d_{ij} = w(\mathbf{C_i} \oplus \mathbf{C_j}) = w([10110]) = 3$. Since the Hamming distance between any two codewords equals the weight of their sum, we can determine the minimum distance between all codewords in a code by just looking at the minimum distance between all codewords and the all zero codeword. 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 a linear 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

$$c_{ij} = u_{i1}g_{1j} + u_{i2}g_{2j} + \ldots + u_{ik}g_{kj}, \quad j = 1, \ldots, n, \tag{8.4}$$

where g is binary (0 or 1) and 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:

$$\mathbf{C_i} = u_{i1}\mathbf{g_1} + u_{i2}\mathbf{g_2} + \ldots + u_{ik}\mathbf{g}_k. \tag{8.7}$$

Since a linear (n, k) block code is a subspace of dimension k, the k row vectors $\{\mathbf{g_l}\}_{l=1}^{k}$ of 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

$$\mathbf{G} = [\mathbf{I_k}|\mathbf{P}] = \left[\begin{array}{cccc|cccc} 1 & 0 & \ldots & 0 & p_{11} & p_{12} & \cdots & p_{1(n-k)} \\ 0 & 1 & \ldots & 0 & p_{21} & p_{22} & \cdots & p_{2(n-k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & p_{k1} & p_{k2} & \cdots & p_{k(n-k)} \end{array}\right], \tag{8.8}$$

where $\mathbf{I_k}$ is a $k \times k$ identity matrix and $\mathbf{P}$ is a $k \times (n-k)$ matrix called a **parity check 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

$$p_j = u_{i1}p_{1j} + \ldots + u_{ik}p_{kj}, \;\; j = 1, \ldots, n-k. \tag{8.10}$$

Note that any generator matrix for an (n, k) linear block code can be reduced by row operations and column operations to a generator matrix with the systematic form.

Example 8.2: Systematic linear block codes are typically implemented with linear shift registers, with $n-k$ modulo-2 adders tied to the appropriate stages of the shift register. The resulting parity bits are appended to the end of the information bits to form the codeword. Find the linear shift register for generating a $(7, 4)$ binary code corresponding to 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 parity check matrix

$$\mathbf{P} = \left[\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right]. \tag{8.12}$$

Let P_{lj} denote the ljth element of $\mathbf{P}$. From (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 $p_3 = u_{i1}P_{13}+u_{i2}P_{23}+u_{i3}P_{33}++u_{i3}P_{43} = u_{i2}+u_{i3}$. The shift register implementation to generate these partiy bits is shown in the following figure. The codeword output is $[u_{i1}u_{i1}u_{i1}u_{i1}p_1p_1p_1]$, where the switch is in the down position to output the systematic bits $u_{ij}, j = 1, \ldots, 4$ of the code, and in the up position to output the parity bits $p_j, j = 1, 2, 3$ of the code.

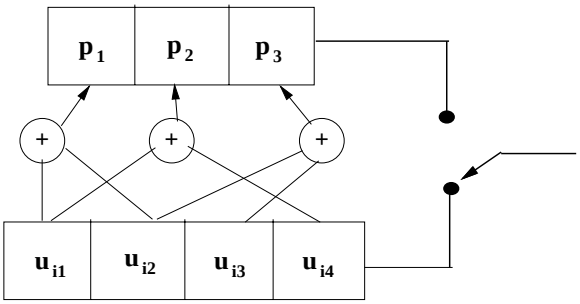


Figure 8.2: Shift register implementation of (7,4) binary code.

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

$$\mathbf{H} = [\mathbf{I_{n-k}}|\mathbf{P}^T]. \tag{8.13}$$

It is easily verified that $\mathbf{GH}^T = \mathbf{0_{k,n-k}}$, 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 $\mathbf{G}$: $\mathbf{C_i} = \mathbf{U_iG}$. Thus,

$$\mathbf{C_iH}^T = \mathbf{U_iGH}^T = \mathbf{0_{n-k}} \tag{8.14}$$

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, 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 vector indicating which codeword symbols were corrupted by the channel. We define the **syndrome** of $\mathbf{R}$ as

$$\mathbf{S} = \mathbf{RH}^T. \tag{8.16}$$

If $\mathbf{R}$ is a valid codeword, i.e. $\mathbf{R} = \mathbf{C_i}$ for some i, then $\mathbf{S} = \mathbf{C_iH}^T = \mathbf{0_{n-k}}$ 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 that 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, then 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

$$\mathbf{S} = \mathbf{R}\mathbf{H}^T = (\mathbf{C} + \mathbf{e})\mathbf{H}^T = \mathbf{C}\mathbf{H}^T + \mathbf{e}\mathbf{H}^T = \mathbf{0_{n-k}} + \mathbf{e}\mathbf{H}^T. \tag{8.17}$$

Since $\mathbf{S} = \mathbf{e}\mathbf{H}^T$ 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

$$\hat{\mathbf{C}} = \mathbf{R} + \hat{\mathbf{e}} = \mathbf{C} + \mathbf{e} + \hat{\mathbf{e}}. \tag{8.18}$$

When the most likely error pattern does occur, i.e. $\hat{\mathbf{e}} = \mathbf{e}$, then $\hat{\mathbf{C}} = \mathbf{C}$, i.e. 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 the 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 1s (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 upperbounded by

$$d_{min} \leq n - k + 1. \tag{8.19}$$

8.2.4 Cyclic Codes

Cyclic codes are a subclass of linear block codes where 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 $\mathbf{C}^{(1)}$ and equal to $\mathbf{C}^{(1)} = (c_{n-1} c_1 \ldots c_{n-2})$ 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

$$g(X) = g_0 + g_1 X + \ldots + g_{n-k} X^{n-k}, \tag{8.20}$$

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*

$$u(X) = u_0 + u_1 X + \ldots + u_{k-1} X^{k-1}. \tag{8.21}$$

The codeword associated with a given k-bit information sequence is obtained from the polynomial coefficients of the generator polynomial times the message polynomial, i.e. the codeword $C = (c_0 \ldots c_{n-1})$ is obtained from

$$c(X) = u(X)g(X) = c_0 + c_1 X + \ldots + c_{n-1} X^{n-1}. \tag{8.22}$$

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), i.e.

$$\frac{c(X)}{g(X)} = q(X) \tag{8.23}$$

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

$$\begin{array}{rl} & X^3+1 \\ X^3+X^2+1 \quad & \Big/\overline{X^{6\cdot}+X^5+X^2+1} \\ & X^6+X^5+X^3 \\ \hline & \quad X^3+X^2+1 \\ & \quad X^3+X^2+1 \\ \hline & \quad 0. \end{array} \tag{8.24}$$

Since $g(X)$ divides $c(X)$ with no remainder, it is a valid codeword. In fact, we have $c_1(X) = (1 + X^3)g(X) = u(X)g(X)$, 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 $g(X) = 1 + X^2 + X^3$, we have

$$\begin{array}{rl} & X^3+1 \\ X^3+X^2+1 \quad & \Big/\overline{X^{6\cdot}+X^5+X^2+1} \\ & X^6+X^5+X^3 \\ \hline & \quad X^2+1 \end{array} \tag{8.25}$$

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

$$X^{n-k}u(X) = u_0X^{n-k} + u_1X^{n-k+1} + \ldots + u_{k-1}X^{n-1}. \tag{8.26}$$

This shifts the message bits to the k rightmost digits of the codeword polynomial. If we next divide (8.26) by $g(X)$, we obtain

$$\frac{X^{n-k}u(X)}{g(X)} = q(X) + \frac{p(X)}{g(X)}, \tag{8.27}$$

where $q(X)$ is a polynomial of degree $k-1$ and $p(X)$ is a remainder polynomial of degree $n-k-1$. Multiplying (8.27) through by $g(X)$ we obtain

$$X^{n-k}u(X) = q(X)g(X) + p(X). \tag{8.28}$$

Adding $p(X)$ to both sides yields

$$p(X) + X^{n-k}u(X) = q(X)g(X). \tag{8.29}$$

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

$$p(X) = p_0 + p_1X + \ldots p_{n-k-1}X^{n-k-1}. \tag{8.30}$$

Combining (8.26) and (8.30) we get

$$p(X) + X^{n-k}u(X) = p_0 + p_1X + \ldots p_{n-k-1}X^{n-k-1} + u_0X^{n-k} + u_1X^{n-k+1} + \ldots + u_{k-1}X^{n-1}. \tag{8.31}$$

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 get 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 (8.31). The polynomial multiplications are trivial to implement, and the polynomial division is easily implemented with a feedback shift register [2, 1]. 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

$$c(X) = u(X)g(X). \tag{8.32}$$

The received codeword can also be written in polynomial form as

$$r(X) = c(X) + e(X) = u(X)g(X) + e(X) \tag{8.33}$$

where $e(X)$ is the error polynomial of degree $n-1$ 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** for the received codeword $s(X)$ is defined as the remainder when $r(X)$ is divided by $g(X)$, so $s(X)$ has degree $n-k-1$. But by (8.33), $e(X)$ is also the remainder when $r(X)$ is divided by $g(X)$. Therefore, the syndrome polynomial $s(X)$ is identical to the error polynomial $e(X)$. Moreover, we obtain the syndrome through a division circuit similar to the one used for generating the code. As stated above, this division circuit is typically implemented using a feedback shift register, resulting in a low-cost low-complexity implementation.

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, i.e. 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 decoding 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. When these distances are used in the channel decoder it is called soft-decision decoding. Soft decision decoding of linear block codes is treated in Section 8.2.7.

Hard decision decoding typically uses **minimum-distance** decoding. In minimum-distance decoding the n bits corresponding to a codeword are first demodulated, and the demodulator output is passed to the decoder. The decoder compares this received codeword to the 2^k possible codewords comprising the code, and decides in favor of the codeword that is closest in Hamming distance (differs in the least number of bits) to the received codeword. Mathematically, for a received codeword R the decoder uses the formula

$$\text{pick } \mathbf{C_j} \text{ s.t. } d(\mathbf{C_j}, \mathbf{R}) \leq d(\mathbf{C_i}, \mathbf{R}) \forall i \neq j. \tag{8.34}$$

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.

Maximum-likelihood decoding picks the transmitted codeword that has the highest probability of having produced the received codeword, i.e. given the received codeword $\mathbf{R}$, the maximum-likelihood decoder choses the codeword $\mathbf{C_j}$ as

$$\text{pick } \mathbf{C_j} \text{ s.t.} p(\mathbf{R}|\mathbf{C_j}) \geq p(\mathbf{R}|\mathbf{C_i}), i = 1, \ldots, 2^k. \tag{8.35}$$

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 (8.34) and the maximum-likelihood criterion (8.35) are equivalent. Once the maximum-likelihood codeword $\mathbf{C_i}$ is determined, it is decoded to the k bits that produce codeword $\mathbf{C_i}$.

Since maximum-likelihood detection of codewords is based on a distance decoding metric, we can best illustrate this process in signal space, as shown in Figure 8.3. The minimum Hamming distance between codewords, illustrated by the black dots in this figure, is d_{min}. Each codeword is centered inside a circle of radius $t = \lfloor .5(d_{min} - 1) \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer greater than or equal to x. The shaded dots represent received codewords where one or more bits differ from those of the transmitted codeword. The figure indicates that $\mathbf{C_1}$ and $\mathbf{C_2}$ differ by 3 bits, since there are two distinct codewords in between them.

Minimum distance decoding can be used to either detect or correct errors. Detected errors in a data block either cause 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 $2^n - 2^k$ nondetectable error patterns.

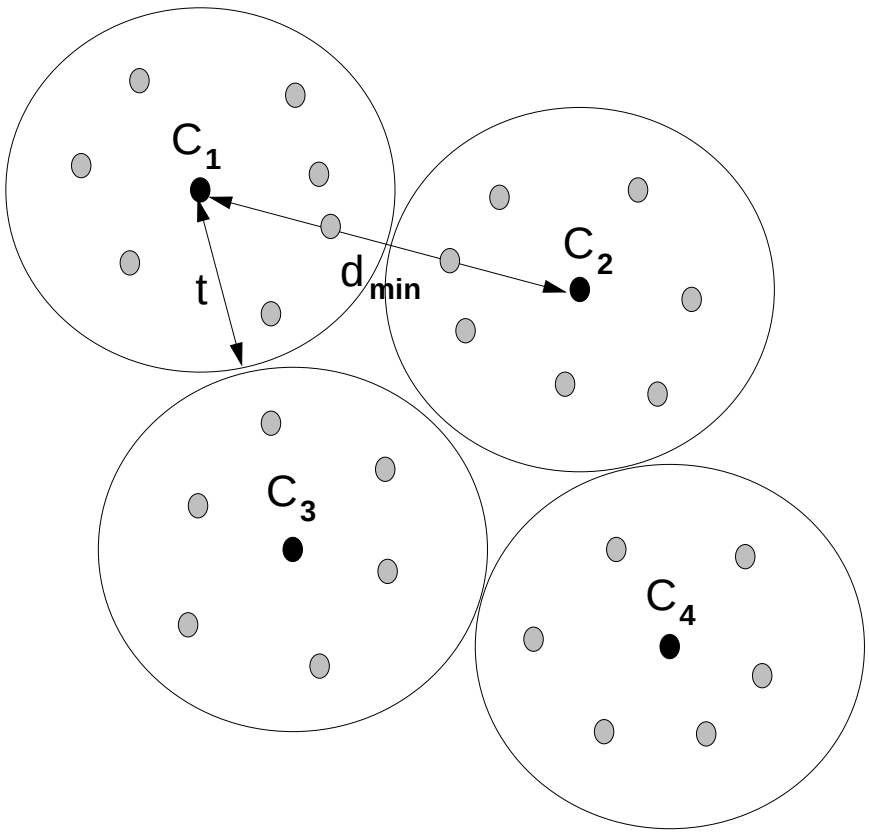


Figure 8.3: Maximum-Likelihood Decoding in Signal Space.

Example 8.4:

A $(5, 2)$ code has codewords $C_0 = [00000]$, $C_1 = [01011]$, $C_2 = [10101]$, and $C_3 = [11110]$. Suppose the all zero codeword 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 codewords, i.e. $\mathbf{e_1} = [01011]$, $\mathbf{e_2} = [10101]$, 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 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:

$$P_e \leq \sum_{j=t+1}^{n} \binom{n}{j} p^j (1-p)^{n-j}, \tag{8.36}$$

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, we have $p = \mathrm{Q}(\sqrt{2E_c/N_0})$, where E_c is the energy per codeword symbol and N_0 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 increases 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**, i.e. it performs worse than uncoded modulation. The bound (8.36) 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}:

$$P_e \geq \sum_{j=t+1}^{d_{min}} \binom{d_{min}}{j} p^j (1-p)^{d_{min}-j}. \tag{8.37}$$

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 (8.37), the probability of mistaking a given codeword for a nearest neighbor at distance d_{min}:

$$P_e \leq (2^k - 1) \sum_{j=t+1}^{d_{min}} \binom{d_{min}}{j} p^j (1-p)^{d_{min}-j}. \tag{8.38}$$

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 Chernoff 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] that the probability of decoding the all-zero codeword as the jth codeword with weight w_j is upper bounded by

$$P(w_j) \leq [4p(1-p)]^{w_j/2}. \tag{8.39}$$

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 get the upper bound

$$P_e \leq \sum_{j=2}^{2^k} [4p(1-p)]^{w_j/2}. \tag{8.40}$$

This bound requires the weight distribution $\{w_j\}_{j=1}^{2^k}$ for all codewords (other than the all-zero codeword corresponding to $j = 1$) in the code. A simpler, slightly looser upper bound is obtained from (8.40) by

using d_{min} instead of the individual codeword weights. This simplification yields the bound

$$P_e \leq (2^k - 1)[4p(1-p)]^{d_{min}/2}. \tag{8.41}$$

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 trellis codes and turbo trellis coded modulation, which will be discussed in Section 8.7.

The probability of bit error after decoding the received codeword in general depends on the particular code and decoder, in particular how bits are mapped to codewords, similar to the bit mapping procedure associated with non-binary modulation. This bit error probability is often approximated as [1]

$$P_b \approx \frac{1}{n} \sum_{j=t+1}^{n} j \begin{pmatrix} n \\ j \end{pmatrix} p^j (1-p)^{n-j}, \tag{8.42}$$

which, for $t = 1$, can be simplified to [1] $P_b \approx p - p(1-p)^{n-1}$. At high SNRs and with good bit-to-codeword mappings we can make the approximation that one codeword error corresponds to a single bit error. With this approximation $P_b = P_e/k$, since there are k bits per codeword.

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 (8.41), assuming the codeword symbols are transmitted over the channel using BPSK modulation with $E_b/N_0 = 10$ 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 $E_b/N_0 = 10$ dB=10, we have $E_c/N_0 = \frac{12}{24}10 = 5$. Thus, $p = \text{Q}(\sqrt{10}) = 7.82 \cdot 10^{-4}$. Using this value in (8.41) with $k = 12$ and $d_{min} = 8$ yields $P_e \leq 3.92 \cdot 10^{-7}$. Using the P_b approximation we get $P_b \approx \frac{1}{k}P_e = 3.27 \cdot 10^{-8}$. For uncoded modulation we have $P_b = Q(\sqrt{2E_b/N_0}) = Q(\sqrt{20}) = 3.87 \cdot 10^{-6}$. So we get over two orders of magnitude coding gain with this code. Note that the loose bound can be orders of magnitude away from the true error probability, as we will see in the next example, so this calculation may significantly underestimate the coding gain of the code.

Example 8.6: Consider a (5,2) linear block code with a minimum distance $d_{min} = 3$ (a Hamming code, discussed in Section 8.2.8, is one such code). Find the union bound (8.38) and looser bound (8.41) assuming the codeword symbols are transmitted over the channel using BPSK modulation with $E_b/N_0 = 10$ dB. Compare the probability of bit error associated with the union bound using the approximation $P_b = P_e/k$ with that of uncoded modulation. How does this comparison change if the SNR is fixed at 10 dB regardless of whether coding is used or not.

Solution: For $E_b/N_0 = 10$ dB=10, we have $E_c/N_0 = \frac{2}{5}10 = 4$. Thus, $p = \text{Q}(\sqrt{8}) = 2.34 \cdot 10^{-3}$. In the union bound (8.38) we then substitute this value for p, $d_{min} = 3$, $t = \lfloor .5(d_{min} - 1) \rfloor = 1$, and $k = 2$, which yields $P_e \leq 4.92 \cdot 10^{-5}$, and the simpler bound (8.41) with the same p, d_{min} and

k yields $P_e \leq 2.7 \cdot 10^{-3}$. Thus, the loose bound is two orders of magnitude looser than the union bound. Using the P_b approximation we get $P_b \approx \frac{1}{k} P_e = 2.46 \cdot 10^{-5}$. Uncoded modulation yields $P_b = \text{Q}(\sqrt{2E_b/N_0}) = \text{Q}(\sqrt{20}) = 3.87 \cdot 10^{-6}$. Comparing this with the union bound, we see that the code has negative gain relative to uncoded modulation, i.e. for a fixed E_b/N_0 we get a lower P_b without coding than with coding. This is because the (5,2) code is not that powerful, and we must spread the bit energy over the entire codeword. However, if the SNR is fixed at 10 dB for both coded and uncoded modulation, we have $E_c/N_0 = E_b/N_0 = 10$. Then $p = 3.87 \cdot 10^{-6}$ is the same as P_b for uncoded modulation. Using this value of p in (8.38) we get $P_e = 1.35 \cdot 10^{-9}$ and the corresponding $P_b = P_e/k = 5.40 \cdot 10^{-11}$, which is much less than in the uncoded case. Thus, the (5,2) code has powerful error correction capability if the transmit power is constant, which results in the same E_c for coded modulation as E_b for uncoded modulation.

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 0 if it is closer to $\sqrt{E_b}$ and as a 1 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 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 $E_c = \frac{k}{n} E_b$. If the jth codeword symbol is a 0, it will be received as $r_j = \sqrt{E_c} + n_j$ and if it is a 1, it will be received as $r_j = -\sqrt{E_c} + n_j$, where n_j is the AWGN noise 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 $\mathbf{C_i}, i = 1, \ldots, 2^k$ in the code, and the decoder choses the codeword $\mathbf{C_i}$ with the highest correlation metric. The correlation metric is defined as

$$C(\mathbf{R}, \mathbf{C_i}) = \sum_{j=1}^{n} (2c_{ij} - 1) r_j, \tag{8.43}$$

where c_{ij} denotes the jth coded bit in the codeword $\mathbf{C_i}$. If $c_{ij} = 1$, $2c_{ij} - 1 = 1$ and if $c_{ij} = 0$, $2c_{ij} - 1 = -1$. So 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. Thus, $C(\mathbf{R}, \mathbf{C_i})$ is large when most of the received symbols have a large magnitude and the same polarity as the corresponding symbols in C_i, is smaller when most of the received symbols have a small magnitude and the same polarity as the corresponding symbols in C_i, and is typically negative when most of the received symbols have a different polarity than the corresponding symbols in 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. Let us assume the all zero codeword $\mathbf{C_1}$ is transmitted and the corresponding received codeword is $\mathbf{R}$. To correctly decode $\mathbf{R}$, we must have that $C(\mathbf{R}, \mathbf{C_1}) > C(\mathbf{R}, \mathbf{C_i}), i = 2, \ldots, 2^k$. Let w_i denote the Hamming weight of the ith codeword $\mathbf{C_i}$, which equals the number of 1s in $\mathbf{C_i}$. Then conditioned on the transmitted

codeword $\mathbf{C_1}$, $C(\mathbf{R}, \mathbf{C_i})$ 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 $\mathbf{R}$. This complicates computation of the exact probability of codeword error, but we can simplify the computation using the union bound. Specifically, the probability $P_e(\mathbf{C_i})$ that $C(\mathbf{R}, \mathbf{C_1}) < C(\mathbf{R}, \mathbf{C_i})$ equals the probability that a Gauss-distributed random variable with mean $\sqrt{nE_c}(1 - 2w_i/n)$ and variance $N_0/2$ is greater than a Gauss-distributed random variable with mean $\sqrt{E_c}n$ and the same variance. This probability is given by:

$$P_e(\mathbf{C_i}) = Q\left(\sqrt{\frac{nE_c}{N_0}}(1 - (1 - 2w_i/n))\right) = Q\left(\sqrt{\frac{R_cE_b}{N_0}}(2w_i))\right) = Q(\sqrt{2R_c\gamma_b w_i}). \tag{8.44}$$

Then by the union bound the average probability of error is upper bounded by the sum of pairwise error probabilities relative to each $\mathbf{C_i}$:

$$P_e \leq \sum_{i=2}^{2^k} P_e(\mathbf{C_i}) = \sum_{i=2}^{2^k} Q(\sqrt{2\gamma_b R_c w_i}). \tag{8.45}$$

The computation of (8.45) requires the weight distribution $w_i, i = 2, \ldots, 2^k$ of the code. This bound can be simplified by noting that $w_i \geq d_{min}$, so

$$P_e \leq (2^k - 1)Q(\sqrt{2\gamma_b R_c d_{min}}). \tag{8.46}$$

A well-known bound on the Q function is $Q(\sqrt{2x}) < \exp[-x]$. Applying this bound to (8.45) yields

$$P_e \leq (2^k - 1)e^{-\gamma_b R_c d_{min}} < 2^k e^{-\gamma_b R_c d_{min}} = e^{-\gamma_b R_c d_{min} + k \ln 2}. \tag{8.47}$$

Comparing this bound with that of uncoded BPSK modulation

$$P_b = Q(\sqrt{2\gamma_b}) < e^{-\gamma_b}, \tag{8.48}$$

we get a dB coding gain of approximately

$$G_c = 10 \log_{10}[(\gamma_b R_c d_{min} - k \ln 2)/\gamma_b] = 10 \log_{10}[R_c d_{min} - k \ln 2/\gamma_b]. \tag{8.49}$$

Note that the coding gain depends on the code rate, the number of information bits per codeword, the minimum distance of the code, and the channel SNR. In particular, the coding gain decreases with γ_b, and becomes negative at sufficiently low SNRs. In general the performance of SDD is about 2 dB better than HDD [2, Chapter 8.1].

Example 8.7: Find the approximate coding gain of SDD over uncoded modulation for the (24,12) code with $d_{min} = 8$ considered in Example 8.2.6 above, with $\gamma_b = 10$ dB.

Solution: Setting $\gamma_b = 10$, $R_c = 12/24$, $d_{min} = 8$, and $k = 12$ in (8.49) 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, 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 therefore have probability of error given exactly by the right side of (8.36).

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 since the information bit rate is one 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, 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 codes are cyclic codes, and 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 block lengths, 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 (8.42) where, for coherent BPSK, we have

$$p = \mathbf{Q}\left(\sqrt{\frac{2E_c}{N_0}}\right) = \mathbf{Q}\left(\sqrt{2R_c\gamma_b}\right). \tag{8.50}$$

In this figure the BCH (127,36) code actually has a negative coding gain at low SNRs. This is not uncommon for powerful channel codes due to their reduced energy per symbol, as was discussed in Section 8.2.5.

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 bits 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 information bits can be mapped into one codeword symbol.

The most common nonbinary block code is the Reed Soloman (RS) code, used in a range of applications from magnetic recording to Cellular Digital Packet Data (CDPD). RS 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, a RS code can correct up to $t = .5(N - K)$ codeword symbol errors. In nonbinary codes the minimum distance between codewords is defined as the number of codeword symbols in which the codewords differ. RS 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.

Since nonbinary codes, and RS codes in particular, generate symbols corresponding to 2^k bits, they are well-matched to M-ary modulation techniques for $M = 2^k$. 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

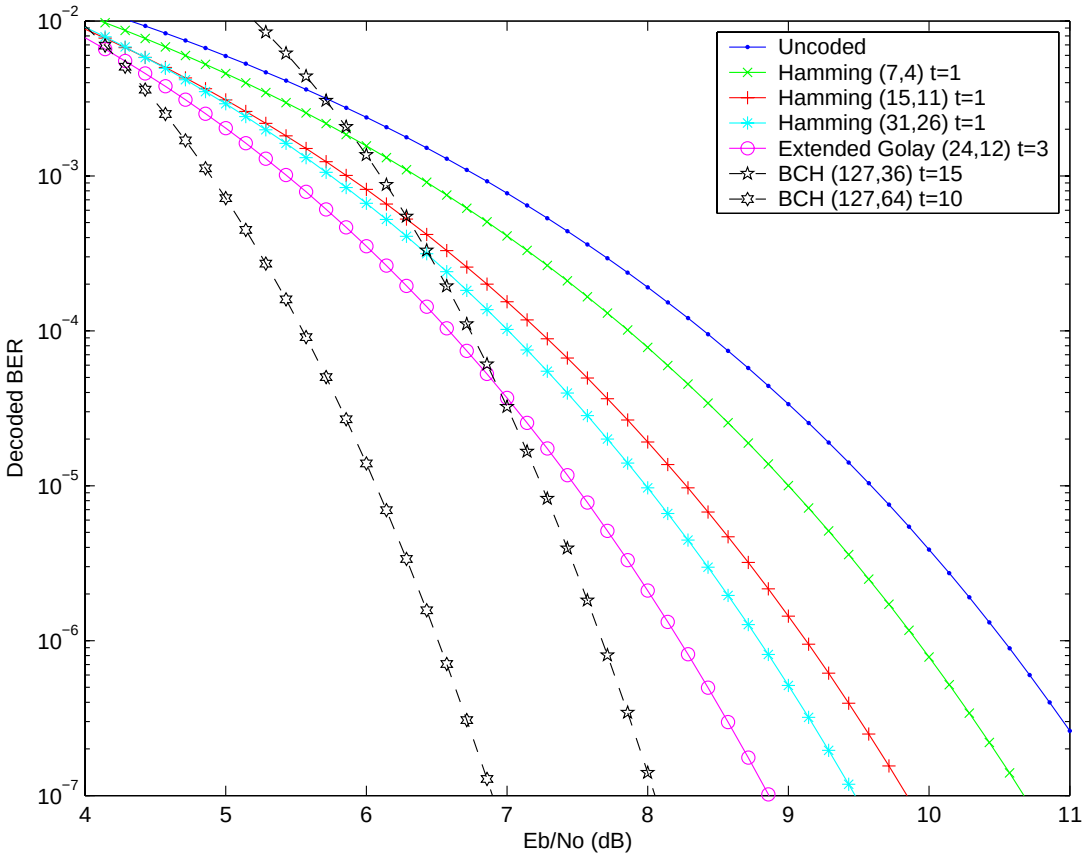


Figure 8.4: P_b for different BCH codes.

point for a constellation point other than the transmitted point) is P_M, then the probability of symbol error associated with the nonbinary code is upper bounded by

$$P_s \leq \sum_{j=t+1}^{N} \binom{N}{j} P_M^j (1-P_M)^{N-j}, \tag{8.51}$$

similar to the form for the binary code (8.36). The probability of bit error is then

$$P_b = \frac{2^{k-1}}{2^k - 1} P_s. \tag{8.52}$$

8.2.10 Block Coding and Interleaving for Fading 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. However, codes designed for AWGN channels cannot typically correct for error bursts longer than t, their error correction capability. In practice slowly fading channels exhibit error bursts much longer than the t associated with codes of reasonable complexity. Therefore, on 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, such that each received codeword only exhibits a few simultaneous symbol errors,

which can be corrected for. The spreading out of burst errors is accomplished by an interleaver and the error correction is accomplished by the code. The size of the interleaver must be large enough so that each symbol in the codeword exhibits independent fading when transmitted over the channel. Slowly fading channels require large interleavers, which in turn can lead to large delays, as we now discuss in more detail.

A block interleaver is an array with d rows and n columns, as shown in Figure 8.5. 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: i.e. 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. 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 have been corrupted by the channel.) The deinterleaver output is read into the decoder by rows, i.e. one codeword at a time.

Figure 8.5 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 2 or fewer errors, the codeword will be decoded in error. However, if the codeword is put through an interleaver then, as shown in Figure 8.5, 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 8.3.7.

To analyze the performance of coding and interleaving, we approximate error bursts caused by Rayleigh fading as independent errors caused by AWGN with codeword symbol error probability determined by the average probability of error in fading. In particular, the probability of error bounds in Sections 8.2.6 are used to determine the probability of error for block codes with interleaving transmitted over fading channels with HDD. The probability of symbol error p in these bounds corresponds to the codeword symbol error probability in fading as determined in Chapter 6. For example, for BPSK modulation in Rayleigh fading,

$$p = \frac{1}{2}\left[1 - \sqrt{\frac{\overline{\gamma}_b}{1 + \overline{\gamma}_b}}\right]. \tag{8.53}$$

Coding and interleaving is a suboptimal coding technique, since the correlation of the fading which affects subsequent bits contains information about the channel which could be used in a true maximum-likelihood decoding scheme. By essentially throwing away this information, the inherent capacity of the channel is decreased [5]. Despite this capacity loss, interleaving codes designed for AWGN is a common coding technique for fading channels, since the complexity required to do maximum-likelihood decoding on correlated coded symbols is prohibitive. More details on code design for fading channels and their performance can be found in [2, Section 14.6].

Example 8.8: Consider a Rayleigh fading channel with a Doppler of $B_d = 80$ Hz. The system uses a (5,2) block code and interleaving to compensate for the fading. If the codeword symbols are sent through

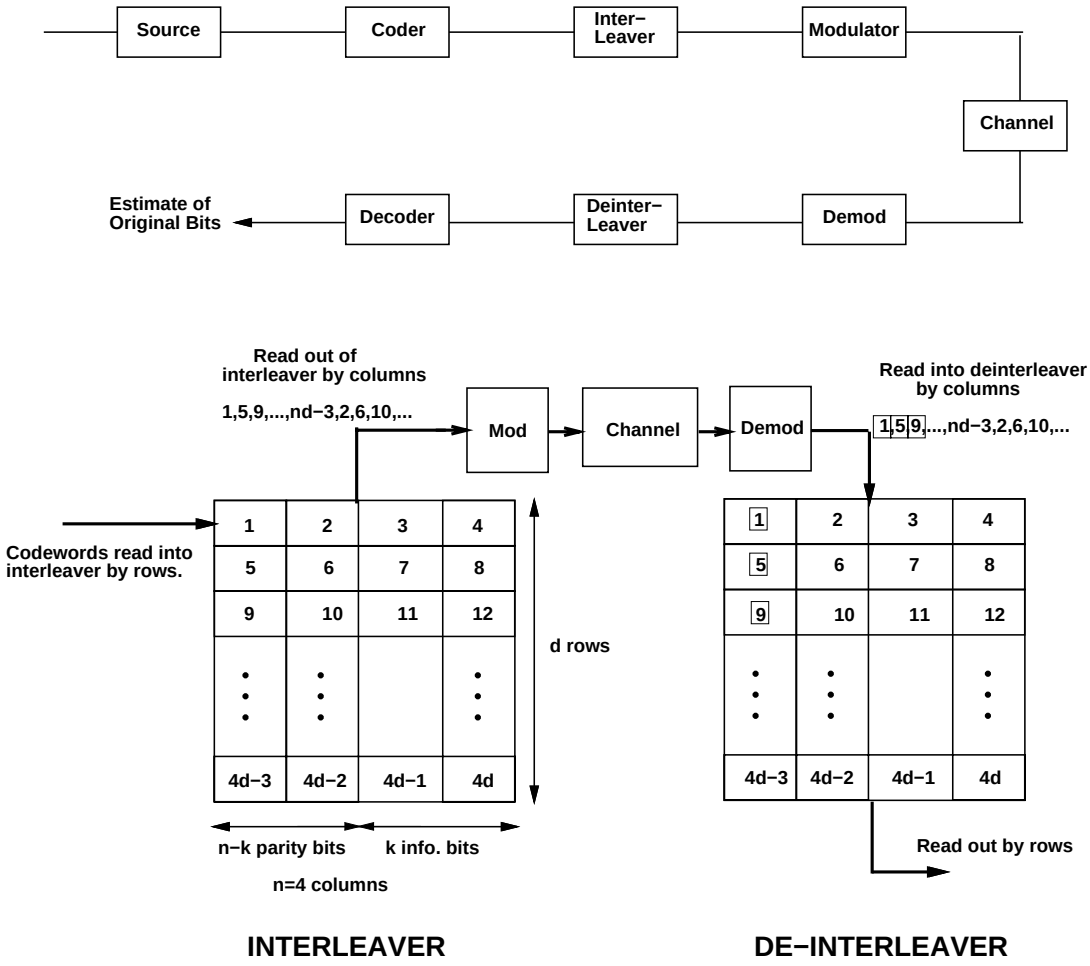


Figure 8.5: The Interleaver/De-interleaver operation.

the channel at 30 Kbps, what is the required interleaver depth 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) code has a minimum distance of 3 so it can correct $t = .5(3-1) = 1$ 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. So the system can tolerate an error burst of 375 symbols. However, all rows of the interleaver must be filled before it can read out by columns, hence the total delay of the interleaver is $ndT_s = 5 \cdot 375 \cdot 3.3 \cdot 10^{-5} = 62.5$ msec. This delay exceeds the delay that can be tolerated in a voice system. We thus see that the price paid for correcting long error bursts through coding and interleaving is significant delay.

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 8.6. 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 number of shift register stages K is called the *constraint length* of the code. It is clear from Figure 8.6 that a length-n codeword depends on kK input bits, in contrast to a block code which only depends on k input bits. 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).

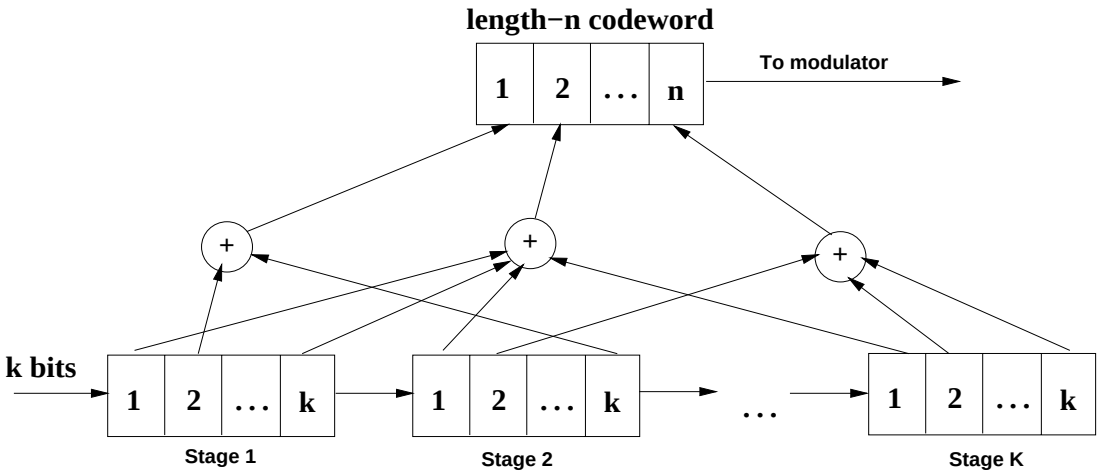


Figure 8.6: Convolutional Encoder.

8.3.1 Code Characterization: Trellis Diagrams

When a length-n codeword is generated by a convolutional encoder, this codeword 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 a convolutional code, we must characterize how the codeword generation depends both on the k input bits and the encoder state, which has 2^{K-1} possible values. There are multiple ways to characterize convolutional codes, including a tree diagram, state diagram, and trellis diagram [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 will focus on the trellis representation of a convolutional code since it is the most common characterization. The details of the trellis diagram representation are best described by an example.

Consider the convolutional encoder shown in Figure 8.7 with $n = 3$, $k = 1$, and $K = 3$. In this encoder, one bit at a time is shifted into Stage 1 of the 3-stage shift register. At a given time t we denote

the bit in Stage i of the shift register as S_i. The 3 stages of the shift register are used to generate a codeword of length 3, $C_1C_2C_3$, where from the figure we see that $C_1 = S_1 + S_2$, $C_2 = S_1 + S_2 + S_3$, and $C_3 = S_3$. 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_3 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 how the new input bit changes the encoder state for the next input bit.

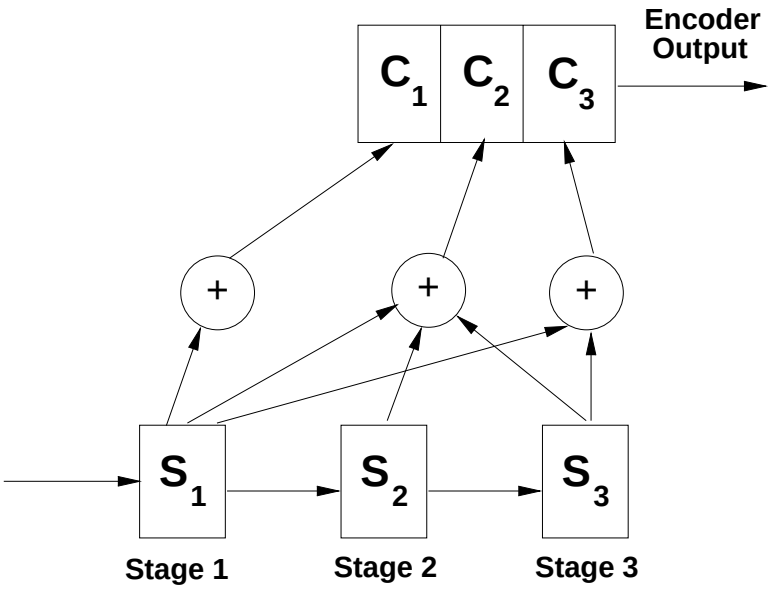


Figure 8.7: Convolutional Encoder Example, ($n = 3$, $k = 1$, $K = 3$).

The trellis diagram for this code is shown in Figure 8.8. The solid lines in Figure 8.8 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 Figure 8.8. 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 $C_1 = S_1 + S_2 = 1$, $C_2 = S_1 + S_2 + S_3 = 1$, and $C_3 = S_3 = 0$. This output 110 is drawn next to the dashed line transitioning from state $S = 00$ to state $S = 10$ in Figure 8.8. 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 8.8 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.

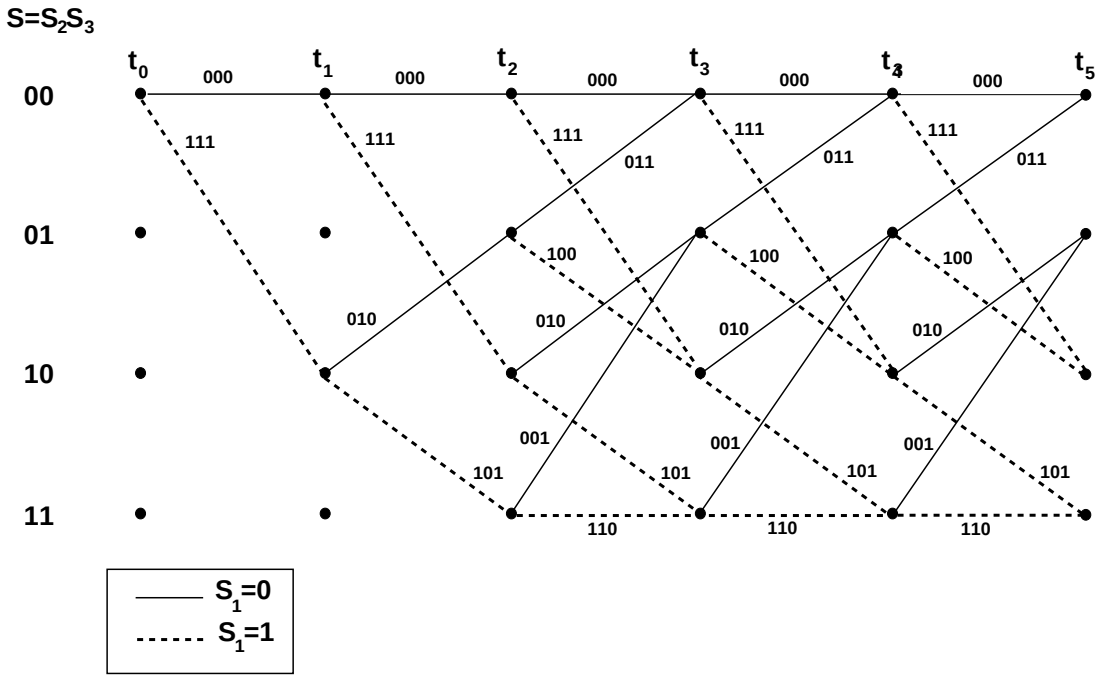


Figure 8.8: Trellis Diagram

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 8.9: Consider the convolution code represented by the trellis in Figure 8.8. For an initial state $S = S_2S_3 = 01$, find the state sequence S and the encoder output C for input bit sequence $\mathbf{U} = 011$.

Solution: The first occurence of $S = 01$ in the trellis is at time t_2. We see at t_2 that if the information bit $S_1 = 0$ 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 particular, for a received sequence $\mathbf{R}$, the decoder decides that coded symbol sequence $\mathbf{C}^*$ was transmitted if

$$p(\mathbf{R}|\mathbf{C}^*) \geq p(\mathbf{R}|\mathbf{C}) \forall \mathbf{C}. \tag{8.54}$$

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 (8.54) as

$$p(\mathbf{R}|\mathbf{C}) = \prod_{i=0}^{\infty} p(R_i|C_i) = \prod_{i=0}^{\infty} \prod_{j=1}^{n} p(R_{ij}|C_{ij}), \tag{8.55}$$

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

$$\log p(\mathbf{R}|\mathbf{C}) = \sum_{i=0}^{\infty} \log p(R_i|C_i) = \sum_{i=0}^{\infty} \sum_{j=1}^{n} \log p(R_{ij}|C_{ij}). \tag{8.56}$$

The expression

$$B_i = \sum_{j=1}^{n} \log p(R_{ij}|C_{ij}) \tag{8.57}$$

is called the **branch metric** since it indicates the component of (8.56) 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 since 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, from (8.56), 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 decision or soft decision 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 L bits long and differ in d places (i.e. their Hamming distance is d), then

$$p(\mathbf{R}|\mathbf{C}) = p^d(1-p)^{L-d}$$

and

$$\log p(\mathbf{R}|\mathbf{C}) = -d \log \frac{1-p}{p} + L \log(1-p). \tag{8.58}$$

Since $p < .5$, (8.58) is minimized 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 then

$$R_{ij} = \sqrt{E_c}(2C_{ij} - 1) + n_{ij}, \tag{8.59}$$

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,

$$p(R_{ij}|C_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left(R_{ij} - \sqrt{E_c}(2C_{ij}-1)\right)^2}{2\sigma^2} \right\}. \tag{8.60}$$

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 (8.56), 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 (8.56) without changing the maximizing sequence. Thus, by neglecting scaling factors and terms in (8.60) that are common to any C_{ij}, we can replace $\sum_{j=1}^{n} \log p(R_{ij}|C_{ij})$ in (8.56) with the **equivalent branch metric**

$$\mu_i = \sum_{j=1}^{n} R_{ij}(2C_{ij} - 1) \tag{8.61}$$

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 8.7 with the trellis diagram in Figure 8.8. For simplicity, we will only consider 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 8.8 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 $\mathbf{C_1} = 110110011$ 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 2. As discussed above, 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

$$M_0 = \sum_{i=0}^{2} \sum_{j=1}^{3} \log P(R_{ij}|C_{ij}) = 6 \log p + 3 \log(1-p), \tag{8.62}$$

while the path metric for the other path is

$$M_1 = \sum_{i=0}^{2} \sum_{j=1}^{3} \log P(R_{ij}|C_{ij}) = 2 \log p + 7 \log(1-p). \tag{8.63}$$

Assuming $p << 1$, which is generally the case, this yields $M_0 \approx 3$ and $M_1 \approx 7$. 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 $\mathbf{Z} = (.8, -.35, -.15, 1.35, 1.22, -.62, .87, 1.08, .91)$. The path metric for the all zero path is

$$M_0 = \sum_{i=0}^{2} \mu_i = \sum_{i=0}^{2} \sum_{j=1}^{3} R_{ij}(2C_{ij} - 1) = \sum_{i=0}^{2} \sum_{j=1}^{3} -R_{ij} = -5.11.$$

The path metric for the second path is

$$M_1 = \sum_{i=0}^{2} \sum_{j=1}^{3} R_{ij}(2C_{ij} - 1) = 6.65.$$

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 (8.56) grows exponentially with the length i of the coded sequence, 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, discovered by Viterbi in 1967 [6] 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 8.9, where Path 1, Path 2, and Path 3 enter Node N (at trellis depth n) with partial path metrics $P^l = \sum_{i=0}^{N} B_i^l, l = 1, 2, 3$ up to this node. Assume P^1 is the largest of these partial path metrics. The complete path with the highest metric, shown in bold, 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, $l = 1, 2$, or 3 up to node N and the maximum likelihood 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 8.9, Path 1 is the survivor at node N and Paths 2 and 3 are discarded from further consideration. Thus, 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, as shown in Figure 8.10. This figure shows the surviving paths at time t_{k+3}. We see in this 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 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, this delay depends on k, K, and the specific code properties. To avoid a random decoding delay, the Viterbi algorithm is typically modified such 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. While this modification does not yield exact maximum likelihood decoding, for n sufficiently large (typically $n \geq 5K$) it is a good approximation.

The Viterbi algorithm must keep track of $2^{k(K-1)}$ surviving paths and their corresponding metrics. At each stage, 2^k metrics must be computed for each node to determine the surviving path, corresponding

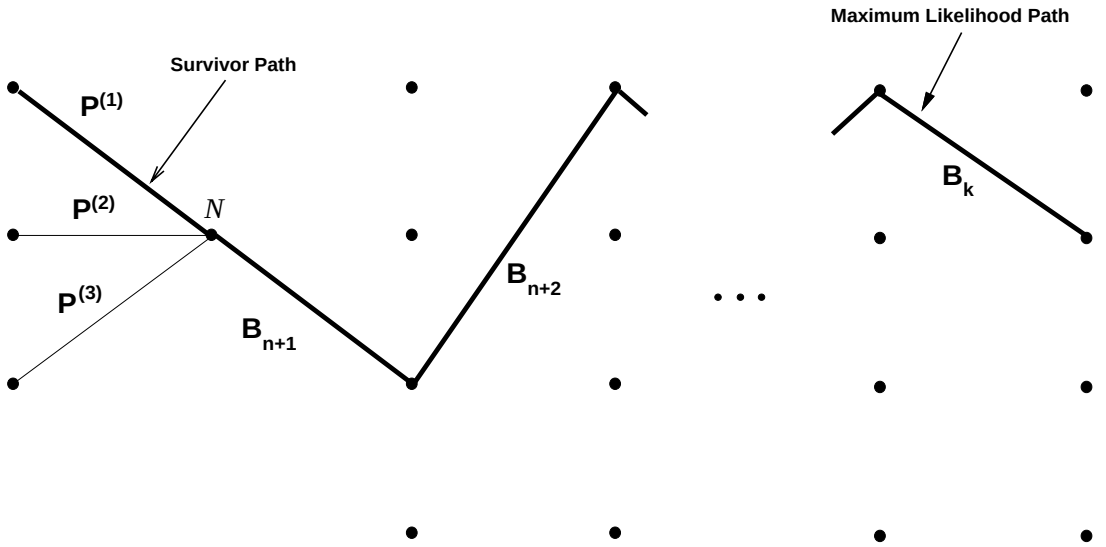


Figure 8.9: Partial Path Metrics on Maximum Likelihood Path

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 for practical implementations 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 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, such 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 8.11 we draw all paths in Figure 8.8 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 therefore 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 same branches as Path 1, 00-10-01 and then later 01-00, plus the branches 01-10-01. Thus we need not consider a longer time interval to find the minimum distance path. For each path in Figure 8.8 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 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.

We define the **minimum free distance** d_{free} of a convolutional code, also called the free distance,

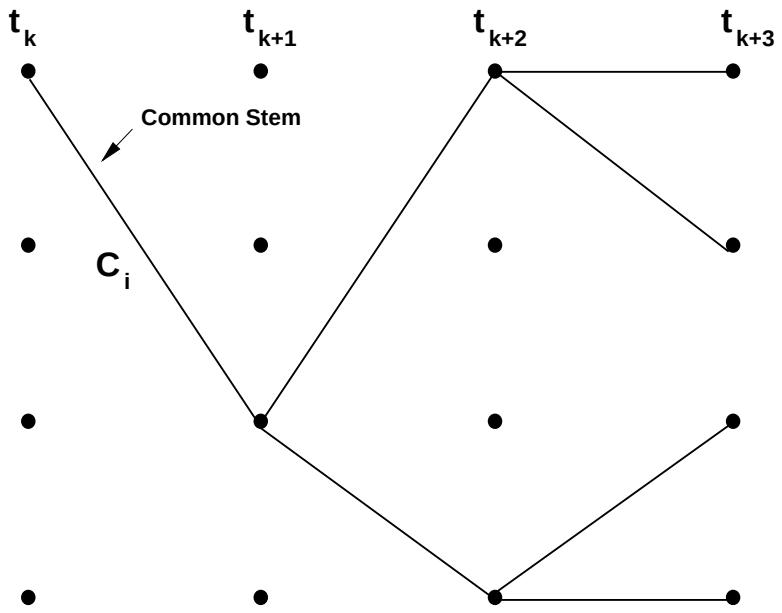


Figure 8.10: Common Stem for All Survivor Paths in the Trellis

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

$$t = \lfloor \frac{d_f - 1}{2} \rfloor.$$

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 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 8.8 is shown in Figure 8.12, 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 $D^0 = 1, D^1$, or D^2, 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 since it does not contribute to the distance properties of the code.

The state diagram can be represented by state equations for each state. For the example of Figure 8.8 we obtain state equations corresponding to the four states:

$$X_c = D^3 X_a + D X_b X_b = D X_c + D X_d X_d = D^2 X_c + D X_d X_e = D^2 X_b, \tag{8.64}$$

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

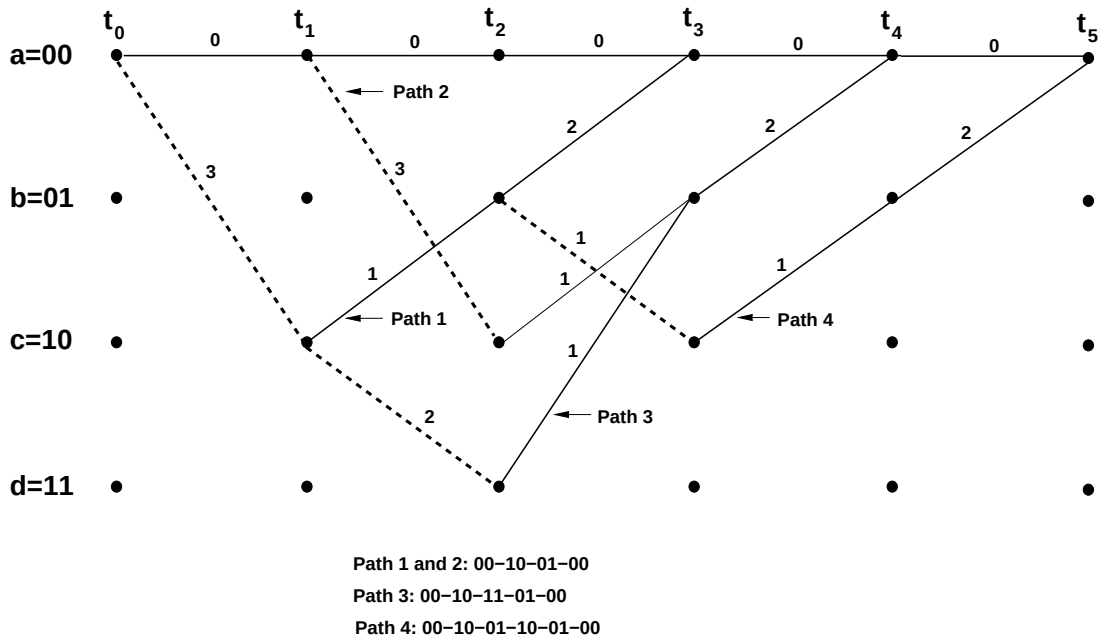


Figure 8.11: Path Distances to the All-Zero Path

equations for the code, which can be done using standard techniques such as Mason's formula, we obtain a transfer function of the form

$$T(D) = \sum_{d=d_f}^{\infty} a_d D^d, \tag{8.65}$$

where a_d is the number of paths with Hamming distance d from the all-zero path. As stated above, 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 8.8, we can solve the state equations given in 8.64 to get the transfer function

$$T(D) = \frac{D^6}{1 - 2D^2} = D^6 + 2D^8 + 4D^{10} + \ldots \tag{8.66}$$

We see from the transfer function that there is one path with minimum distance $d_f = 6$, and 2 paths with Hamming distance 8, which is consistent with Figure 8.11. 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.

While 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 8.12). 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 8.8 is shown in Figure 8.13.

The extended state diagram is also represented by state equations. For the example of Figure 8.13

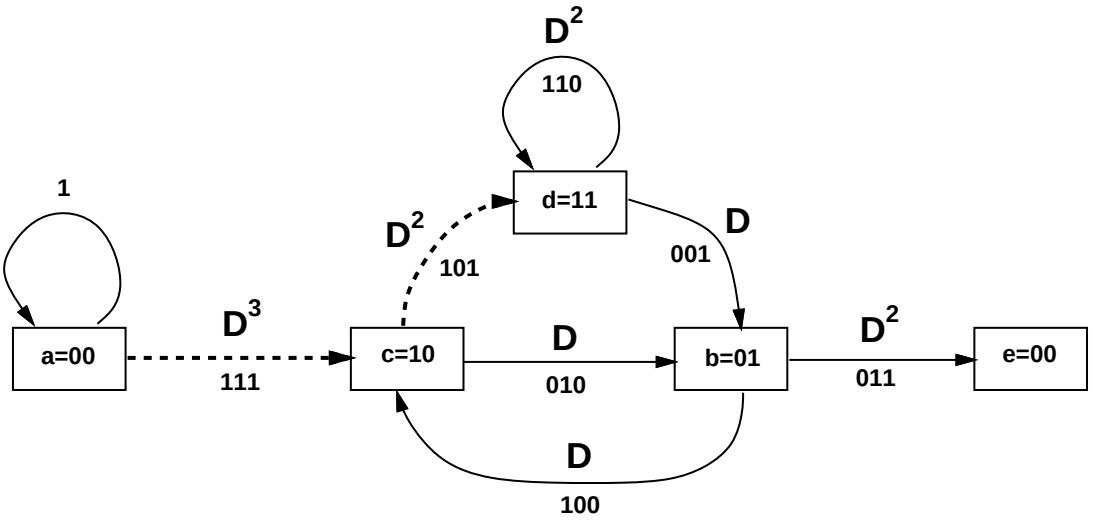


Figure 8.12: State Diagram

these are given by:

$$X_c = JND^3X_a + JNDX_bX_b = JDX_c + JDX_dX_d = JND^2X_c + JND^2X_dX_e = JD^2X_b, \tag{8.67}$$

Similar 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

$$T(D, N, J) = \frac{J^3ND^6}{1 - JND^2(1 + J)} = J^3ND^6 + J^4N^2D^8 + J^5N^2D^8 + J^5N^3D^10 + \ldots. \tag{8.68}$$

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

$$T(D, N) = T(D, N, J = 1). \tag{8.69}$$

The transfer function for the extended state tells us more information about the diverging and remerging paths; namely, the minimum distance path with Hamming distance 6 is of length 3 and results in a single bit error (exponent of N is one), one path of Hamming distance 8 is of length 4 and results in 2 bit errors, and the other path of Hamming distance 8 is of length 5 and results in 2 bit errors, consistent with Figure 8.11. The extended transfer function is a convenient shorthand to represent the Hamming distance, length, and number of bit errors corresponding to each diverging and remerging path of a code from the all zero path. We will see in the next section that this convenient representation is very 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 assuming that the all-zero sequence is transmitted, and 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.

We 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

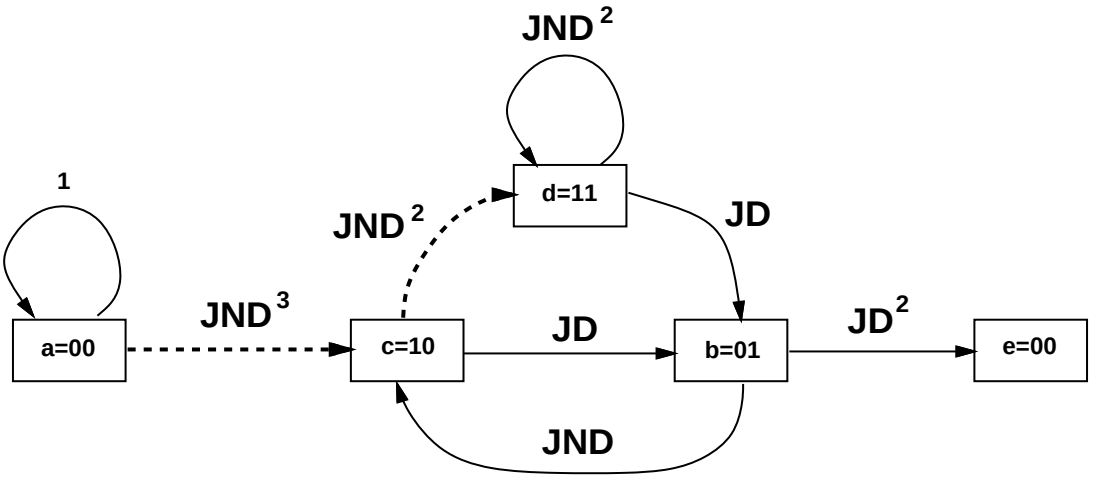


Figure 8.13: Extended State Diagram

encoder are sent over an AWGN channel using coherent BPSK modulation with energy $E_c = R_c E_b$, then it can be shown that if the all-zero sequence is transmitted, the probability of mistaking this sequence with a sequence Hamming distance d away is [2]

$$P_2(d) = Q\left(\sqrt{\frac{2E_c}{N_0}d}\right) = Q\left(\sqrt{2\gamma_b R_c d}\right). \tag{8.70}$$

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

$$P_e \leq \sum_{d_f}^{\infty} a_d Q\left(\sqrt{2\gamma_b R_c d}\right), \tag{8.71}$$

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 an exponential to upper bound the Q function, i.e. we use the fact that

$$Q\left(\sqrt{2\gamma_b R_c d}\right) \leq e^{-\gamma_b R_c d}.$$

We then get the upper bound

$$P_e < T(D)|_{D=e^{-\gamma_b R_c}}. \tag{8.72}$$

While this upper bound tells us the probability of mistaking one sequence for another, it does not yield the probability of bit error, which is more fundamental. 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

$$T(D,N) = \sum_{d=d_{free}}^{\infty} a_d D^d N^{f(d)}, \tag{8.73}$$

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 [2]

$$P_b \leq \sum_{d_f}^{\infty} a_d f(d) Q\left(\sqrt{2\gamma_b R_c d}\right), \tag{8.74}$$

where the only difference with (8.71) is the weighting factor $f(d)$ corresponding to the number of bit errors in each incorrect path. If the Q function is upper bounded by the complex exponential as above we get the upper bound

$$P_b < \left.\frac{dT(D,N)}{dN}\right|_{N=1,D=e^{-\gamma_b R_c}}. \tag{8.75}$$

If $k > 1$ then we divide (8.74) or (8.75) 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

$$P_2(d) = \sum_{k=.5(d+1)}^{d} \binom{d}{k} p^k (1-p)^{(d-k)}, \tag{8.76}$$

where p is the probability or error on the channel. This is because the incorrect path will be selected only if the decoded path is closer to the incorrect path than to the all-zero path, i.e. 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 of the number of errors is exactly $.5d$. We can simplify the pairwise error probability using the Chernoff bound to yield

$$P_2(d) < [4p(1-p)]^{d/2}. \tag{8.77}$$

Following the same approach as in soft decision decoding, we then obtain the error probability bound as

$$P_e < \sum_{d_f}^{\infty} a_d [4p(1-p)]^{d/2} < T(D)|_{D=\sqrt{4p(1-p)}}, \tag{8.78}$$

and

$$P_b < \sum_{d_f}^{\infty} a_d f(d) P_2(d) = \left.\frac{dT(D,N)}{dN}\right|_{N=1,D=\sqrt{4p(1-p)}}. \tag{8.79}$$

8.3.7 Convolutional Coding and Interleaving for Fading Channels

As with block codes, convolutional codes suffer performance degradation in fading channels, since 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 to both spread out burst errors and to work well with the incremental nature of convolutional code generation [7, 8].

An example block diagram for a convolutional interleaver is shown in Figure 8.14. 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 $N(N-1)T_s$ [1], where T_s is the symbol time for transmitting the coded symbols over the channel.

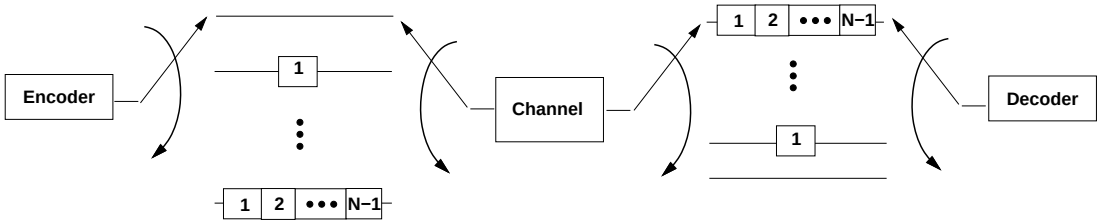


Figure 8.14: Convolutional Coding and Interleaving

Example 8.10: Consider a channel with coherence time $T_c = 12.5$ msec and a coded bit rate of $R_s = 100,000$ 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 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 a very high delay for either voice or data.

8.4 Concatenated Codes

A concatenated code uses two levels of coding: an inner code and an outer code, as show in Figure 8.15. 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 correctly by the inner code). Concatenated codes may have the inner and outer codes separated by an interleaver to break up block errors introduced by the channel. Concatenated codes typically achieve very low error probability with less complexity than a single code with the same error probability performance. A common concatenated code used in CD recordings has a convolutional inner code and a Reed Soloman (block) outer code [4]. The decoding of concatenated codes is typically 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 ML 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 concatanated codes can be obtained based on iterative decoding, and that is the basic premise behind turbo codes, described in the next section.

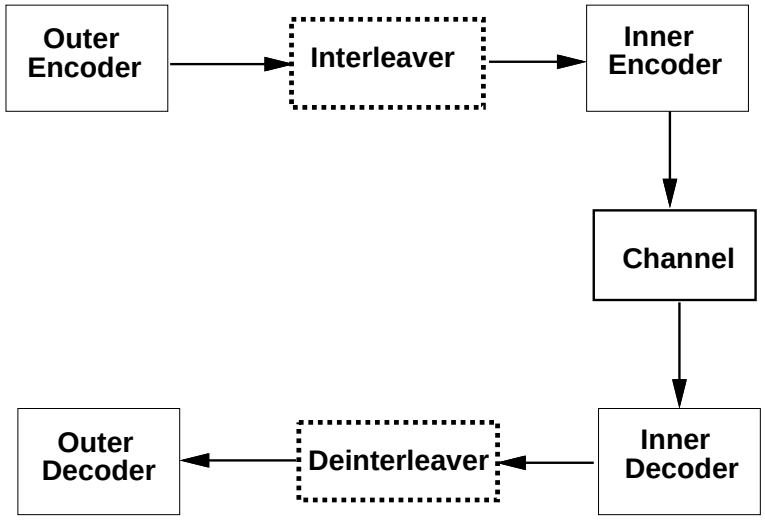


Figure 8.15: Concatenated Coding

8.5 Turbo Codes

Turbo codes, introduced in 1993 in a landmark paper by Berrou, Glavieux, and Thitimajshima [9], are very powerful codes that can come within a fraction of a dB of the Shannon capacity limit on AWGN channels. Turbo codes and the more general family of codes on graphs with iterative decoding algorithms [11, 12] 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], however at the time these coding techniques were thought impractical and were generally not pursued by researchers in the field. The landmark 1993 paper on turbo codes [9] provided more than enough motivation to revisit Gallager's and other's work 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 [9, 13]. A typical parallel concatenated encoder is shown in Figure 8.16. 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 call 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 elimination of error floors [13].

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 8.17 for the example encoder of Figure 8.16. 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 the received codeword (m, X_1). This reliability 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

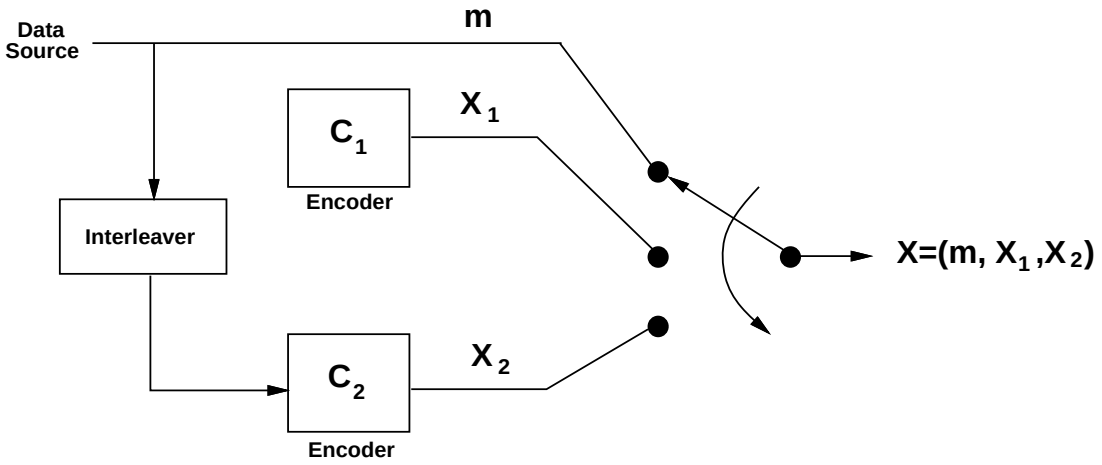


Figure 8.16: Parallel Concatenated (Turbo) Encoder.

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 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; i.e., the decoders cannot agree on the value of m. Several methods have been proposed for detecting convergence (if it occurs), including bit estimate variance [Berr96] and neural net-based techniques [14].

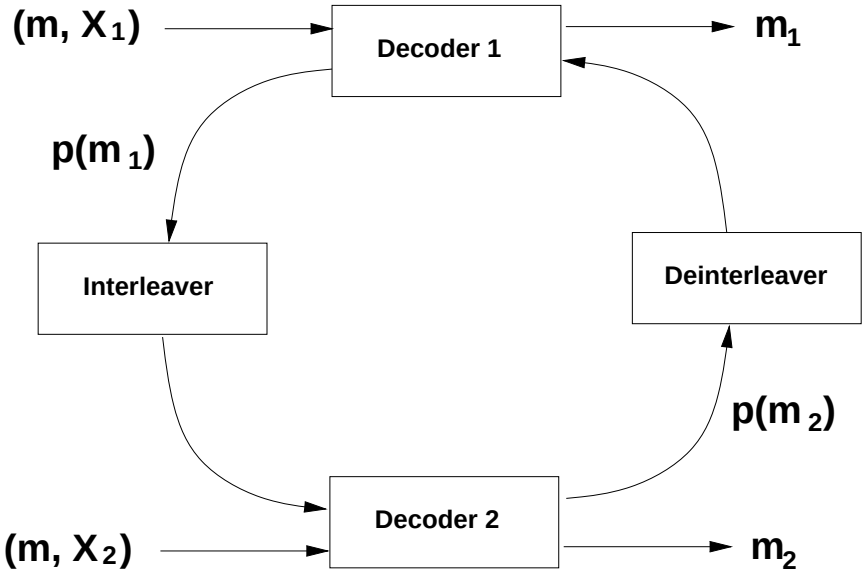


Figure 8.17: Turbo Decoder.

The simulated performance of turbo codes over multiple iterations of of the decoder is shown in Figure 8.18 for a code composed of two rate 1/2 convolutional codes with constraint length $K = 5$ separated by an interleaver of depth $d = 2^{16} = 65,536$. The decoder converges after approximately 18

iterations. This curve indicates several important aspects of turbo codes. First, note their exception 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 [9] 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 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 8.18 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.

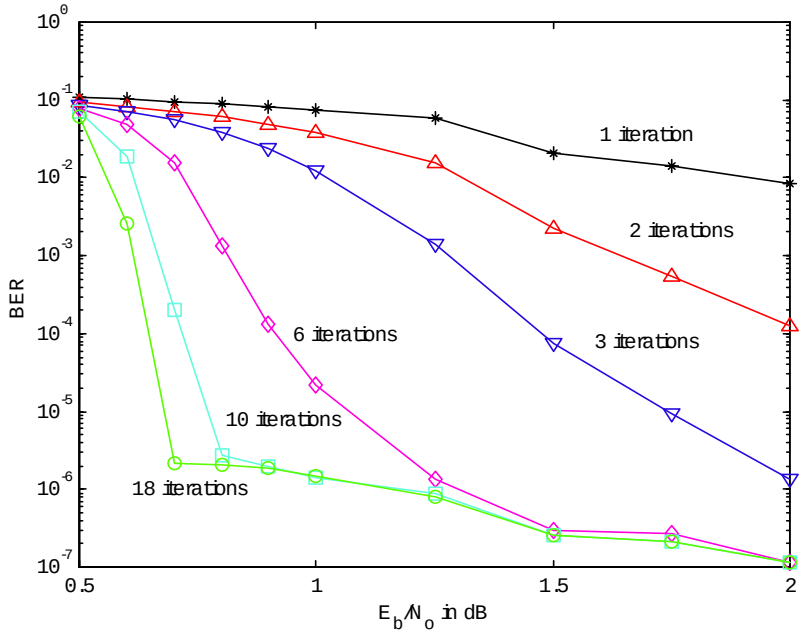


Figure 8.18: Turbo Code Performance (Rate 1/2, $K = 5$ component codes with interleaver depth 2^{16}).

An alternative to parallel concatenated coding is serial concatenated coding [15]. In this coding technique, one component code serves as an outer code, and 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, e.g. [15, 17, 16]. While both codes perform very well under similar delay and complexity conditions, serial concatenated coding in some cases 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 in his 1961 Masters thesis [10]. However, these codes were largely ignored until the introduction of turbo codes, which rekindled some of same ideas. Subsequent to the landmark paper on turbo codes in 1993 [9], LDPC codes were reinvented by Mackay and Neil [18] and by Wiberg [19] in 1996. Shortly thereafter it was recognized that these new code designs were actually reinventions of Gallager's original work, 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.

LDPC 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 ones in each column and d_c ones 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, and hence the name low density parity check codes.

Provided that the codeword length is long, LDPC codes achieve performance close to the Shannon limit, in some cases surpassing the performance of parallel or serially concatenated codes [24]. 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 (quadratic in blocklength) but low decoding complexity. In particular, like turbo codes, LDPC decoding uses iterative techniques, which are related to Pearl's belief propagation commonly used by the artificial intelligence community [25]. However, the belief propagation corresponding to LDPC decoding is simpler than for turbo decoding, thereby making the LDPC iterative decoder much simpler [25, 26]. In addition, the belief propagation decoding is parallelizable and can be closely approximated with very low complexity decoders [20]. Finally, the decoding algorithm for LDPC codes can detect when a correct codeword has been detected, which is not necessarily the case for turbo codes.

Additional work in the area of LDPC codes includes finding capacity limits for these codes [20], determining effective code designs [29] and efficient encoding and decoding algorithms [20, 28], and expanding the code designs to include nonregular [24] and nonbinary LDPC codes [21] as well as coded modulation [22].

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, such as block convolutional, and turbo codes, reduce transmit power at the expense of increased bandwidth or reduced data rate.

The spectrally-efficient coding breakthrough came when Ungerboeck [30] 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, has remained superior over more recent developments in coded modulation (coset and lattice codes), as well as more complex trellis codes [31]. We now outline the general principles of this coding technique. Comprehensive treatments of trellis, lattice, and coset codes can be found in [30, 32, 31].

8.7.1 Coded Modulation for AWGN Channels

The basic scheme for trellis and lattice coding, or more generally, any type of coset coding, is depicted in Figure 8.19. There are five elements required to generate the coded-modulation:

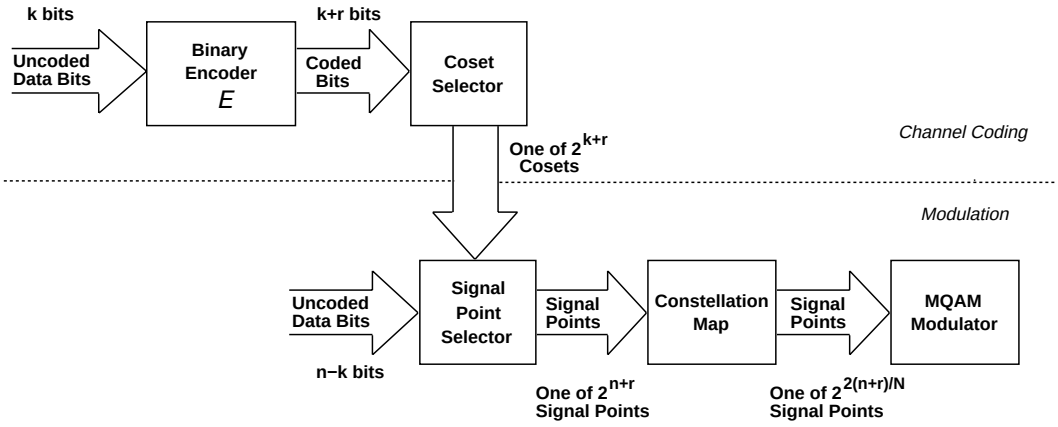


Figure 8.19: General Coding Scheme.

1. A binary encoder E, block or convolutional, that operates on k uncoded data bits to produce $k+r$ coded bits.

2. A subset 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.

5. An MQAM modulator (or other M-ary modulator).

The first two steps described above are the channel coding, and the remaining steps 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 coded-modulation scheme is referred to as a trellis code; for E a block encoder, it is called a lattice (or block) code.

The steps described above essentially decouple the channel coding gain from 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 shape gain (for asymptotically large N), independent of the channel coding scheme[1]. However, the performance improvement from shape 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 doesn't affect the shape or coding gain.

The channel coding gain results from a selection of 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

[1]A square constellation has 0 dB of shape 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 [31].

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_{min}^2. 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 allowable sequence points (the sequence points obtained through coding), multiplied by the density of the allowable sequence points. Specifically, if the minimum distance and density of points in the sequence space are denoted by d_0 and Δ_0, respectively, and if the minimum distance and density of points in the sequence space obtained through coding are denoted by d_{min} and Δ, respectively, then maximum-likelihood sequence detection yields a channel coding gain of

$$G_c = \left(\frac{d_{min}^2}{d_0^2}\right)\left(\frac{\Delta}{\Delta_0}\right). \tag{8.80}$$

The second bracketed term in this expression is also refered to as the **constellation expansion factor**, and equals 2^{-r} (per N dimensions) for a redundancy of r bits in the encoder E [31].

Some of the nominal coding gain in (8.80) is lost due to correct 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 codes in [31]. 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-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 **sub-lattice** is a subset of a lattice that is itself a lattice. The sequence space for *uncoded* M-QAM 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-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, however 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 has been solved for all N for which the decoder complexity is manageable[3]. Once the densest lattice is known, we can form partioning subsets, or **cosets**, of the lattice through translation of any sublattice. The choice of the partitioning sublattice will determine the size of the partition, i.e. the number of subsets that the subset selector in Figure 8.19 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 denote 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

$$G_c = 2^{-2r/N} d_{min}^2, \tag{8.81}$$

where $2^{-2r/N}$ is the constellation expansion factor (in two dimensions) from the r extra bits introduced by the binary channel encoder.

[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.

Returning to Figure 8.19, suppose that we want to send $m = n + r$ bits per dimension, so an N sequence conveys mN bits. If we use the densest lattice in N space that lies within an N sphere, where the radius of the sphere is just large enough to enclose 2^{mN} points, then we achieve a total coding gain which combines the channel gain (resulting from the lattice density and the encoder properties) with the shape gain of the N sphere over the N rectangle. Clearly, the channel coding gain is decoupled from the shape gain. An increase in signal power would allow us to use a larger N 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 [32] 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 [32]. First, the two-dimensional signal constellation is partitioned into four subsets as shown in Figure 8.20, where the subsets are represented by the points A_0, A_1, B_0, and B_1, respectively. From this subset partition, we form an 8-dimensional lattice by taking all sequences of four points in which all points are either A points or B points and moreover, 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 two). Thus, three data bits and one parity check bit are used to determine the lattice subset. The minimum distance resulting from this subset partition is four times the minimum distance 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 from Chapter 5.3.3 costs and additional 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, so for a data rate of m bits/symbol, the four lattice subsets must each have 2^{m-1} points[4].

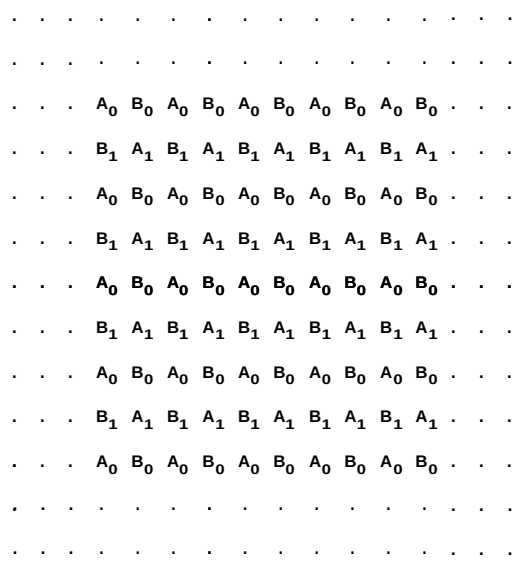


Figure 8.20: Subset Partition for an Eight-Dimensional Lattice.

[4]This yields $m - 1$ bits/symbol, with the additional bit/symbol conveyed by the channel code.

Coded modulation using turbo codes has also been investigated [33, 34, 35]. This work shows that turbo trellis coded modulation can come very close to the Shannon limit for nonbinary signalling.

8.7.2 Coded Modulation with Interleaving for Fading Channels

Coded modulation for fading channels also uses the coding and interleaving approach of block and convolutional codes, however the interleaver is matched to the block or convolutional encoder in coded modulation design [36, 37]. However, the minimum distance error event in a trellis code depends both on the parallel transitions and the minimum distance error event through the trellis. Thus, the dominating error event is not always obvious, which complicates code design. A good overview of trellis code design for fading channels, including the impact of interleaving and channel fade information, can be found in [36] and the references therein. There is no good rule of thumb for these code designs, and in many cases simulations must be used to evaluate performance and choose between different code designs.

8.8 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 packets 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 or bits requiring low P_b will be received correctly with high probability.

Practical implementation of a multilevel code was first studied by Imai and Hirakawa [38]. Binary UEP codes were later considered both for combined speech and channel coding [39], and combined image and channel coding [40]. 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 have been proposed: time-multiplexing of bandwidth-efficient coded modulation [41], and coded-modulation techniques applied to both uniform and nonuniform signal constellations [42, 43]. All of these multilevel codes can be designed for either AWGN or fading channels. We now briefly summarize these UEP techniques; specifically, we describe the principles behind multilevel coding and multistate decoding, and 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

$$s = \sum_{i=1}^{M} s_i \times 2^{i-1}. \tag{8.82}$$

For example, if $M = 3$ and the signal constellation is 8PSK, then the chosen signal point will have phase $2\pi s/8$.

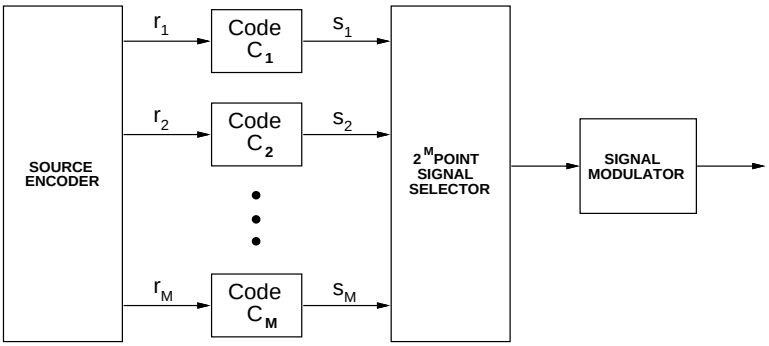


Figure 8.21: Multilevel Encoder

Optimal decoding of the multilevel code uses a maximum-likelihood decoder, which determines the input sequence that maximizes the received sequence probability. The maximum-likelihood 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, $i = 1, \ldots, M$, the number of states in the optimal decoder is $2^{\mu_1 + \cdots + \mu_M}$. Due to the high complexity of optimal decoding, the suboptimal technique of multistage decoding, introduced in [38], is used for most implementations. Multistage decoding is accomplished by decoding the component codes sequentially. First, the most robust 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 correct for code decisions on the less robust code sequences.

The binary encoders of this multilevel code require extra code bits to achieve their coding gain, thus they are not bandwidth-efficient. An alternative approach recently proposed in [42] uses time-multiplexing of the trellis codes described in Chapter 8. In this approach, different conventional coded modulation schemes, such as lattice or trellis codes, 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 different priority bit streams. 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 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 comprising $R_1T_1/(R_1T_1 + R_2T_2)$ 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 2 encoders, where the extension to M encoders is straightforward. Recall that in trellis coding, bits are encoded to select the lattice subset, and 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 used to choose the subset of the partitioned constellation, while the low-priority bits are uncoded or lightly coded and used to select the constellation signal point.

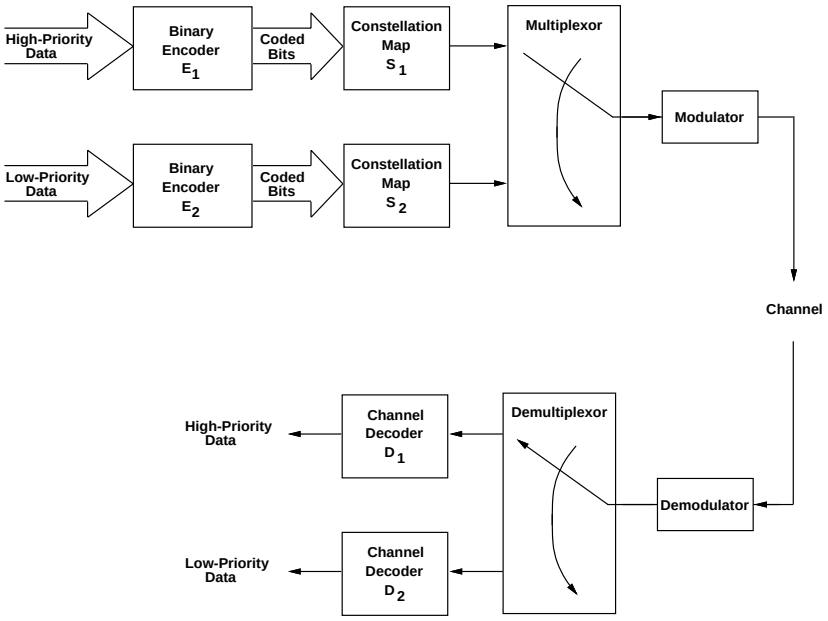


Figure 8.22: Transceiver for Time-Multiplexed Coded Modulation

8.9 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 [45], this result holds only in the limit of infinite source code dimension, infinite channel code block length, 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.

Previous 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. UEP 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 [40, 46, 47]. A comprehensive treatment of matching RCPC channel codes or multilevel quadrature amplitude modulation (MQAM) to subband and linear predictive speech coding in both AWGN and Rayleigh fading channels, can be found in [48]. 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 [49], 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.

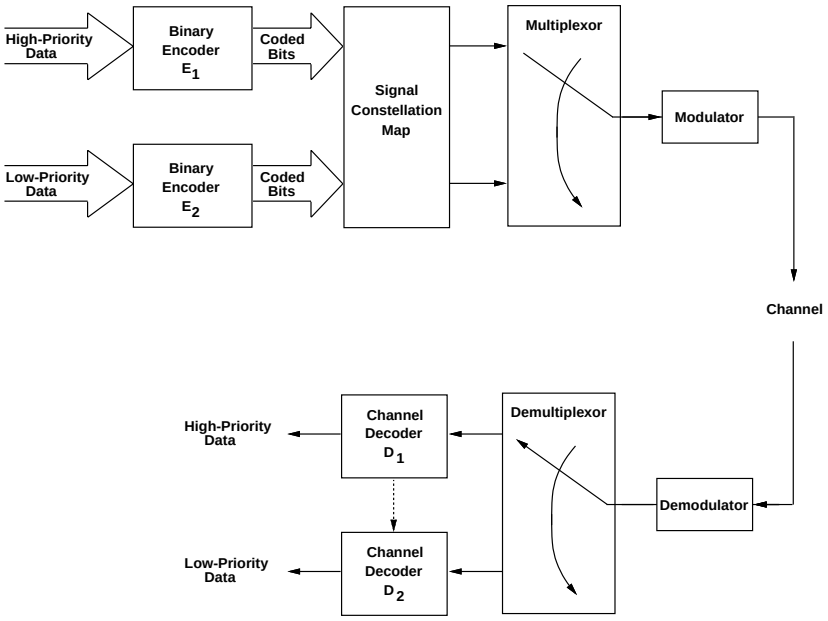


Figure 8.23: Joint Optimization of Signal Constellation

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 independent of the source. Examples of work taking this approach include the channel-optimized vector quantizer (COVQ) and its scalar variation [50, 51]. Source-optimized channel coding and modulation can be combined with channel-optimized source coding using an iterative design. This approach is used for the joint design of a COVQ and multicarrier modulation in [52] and for the joint design of a COVQ and RCPC channel code in [53]. 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 [54, 55]. 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.

Bibliography

[1] B. Sklar, *Digital Communications - Fundamentals and Applications*. Prentice Hall 1988.

[2] J.G. Proakis, *Digital Communications*. 4th Ed. New York: McGraw-Hill, 2001.

[3] D. G. Wilson, *Digital Modulation and Coding*. Prentice Hall 1996.

[4] S. Lin and J.D.J. Costello, *Error Control Coding*, 2nd Ed., Prentice Hall, 2004.

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

[6] A.J. Viterbi, "Error bounds for convolutional codes and asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, pp. 260–269, 1967.

[7] G.D. Forney, "Burst error correcting codes for the classic bursty channel," *IEEE Trans. Commun. Tech.* pp. 772–781, Oct. 1971.

[8] J.L. Ramsey, "Realization of optimum interleavers," *IEEE Trans. Inform. Theory*, pp. 338–345, 1970.

[9] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes," *Proc. of ICC'93*, pp. 54-58.

[10] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inform. Theory,* pp. 21–28, Jan. 1962.

[11] IEEE Trans. on Inform. Theory, Special Issue on Codes and Graphs and Iterative Algorithms, Feb. 2001.

[12] S. B. Wicker and S. Kim, Codes, Graphs, and Iterative Decoding, Boston: Kluwer Academic Press, 2002.

[13] C. Heegard and S. B. Wicker, Turbo Coding, Boston: Kluwer Academic Press, 1999.

[14] 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 - 576, April 2000.

[15] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial concatenation of interleaved codes: performance analysis, design and iterative decoding," *IEEE Trans. Inform. Theory*, pp. 909-926, May 1998.

[16] 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, Jan. 2000.

[17] H. Jin and R.J. McEliece, "Coding theorems for turbo code ensembles," *IEEE Trans. Inform. Theory*, pp. 1451 - 1461, June 2002.

[18] D.J.C. MacKay and R.M. Neal, "Near Shannon limit performance of low density parity check codes," *Elec. Letts.*, pg. 1645, Aug. 1996.

[19] N. Wiberg, N.-A. Loeliger, and R. Kotter, "Codes and iterative decoding on general graphs," *Euro. Trans. Telecommun.*, pp. 513-525, June 1995.

[20] T. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message passing decoding," *IEEE Trans. Inform. Theory*, pp. 599–618, Feb. 2001.

[21] M.C. Davey and D. MacKay, "Low density parity-check codes over GF(q)," *IEEE Commun. Letters,* pp. 165–167, June 1998.

[22] 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. 2141-2155, Sept. 2003.

[23] R.G. Gallager, "Low density parity check codes," *IRE Trans. Inform. Theory*, pp. 21-28, Jan. 1962. See also *Low density parity check codes*, no. 21 in Research Monograph Series, Cambridge, MA: MIT Press, 1963.

[24] T. Richardson, A. Shokrollahi, and R. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Inform. Theory*, pp. 619–637, Feb. 2001.

[25] R. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's "belief propagation" algorithm," *IEEE J. Select Areas Commun.*, pp. 140-152, Feb. 1998.

[26] F.R. Kschischang and D. Frey, "Iterative decoding of compound codes by probability propagation in graphical models," *IEEE J. Select Areas Commun.*, pp. 219–230, Feb. 1998.

[27] D. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, pp. 399–431, March 1999.

[28] M. Fossorier, "Iterative reliability-based decoding of low-density parity check codes, " *IEEE J. Select Areas Commun.*, pp. 908–917, May 2001.

[29] S-Y Chung, G. D. Forney, T. Richardson, and R. Urbanke, "On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit," *IEEE Commun. Letters,* pp. 58–60, Feb. 2001.

[30] G. Ungerboeck, "Channel coding with multi-level/phase signals," *IEEE Trans. Info. Theory*, Vol. IT-28, No. 1, pages 55-67, Jan. 1982.

[31] G.D. Forney, "Coset codes, I: Introduction and geometrical classification, and II: Binary lattices and related codes. *IEEE Trans. Inform. Theory*, pp. 1123 - 1187, Sept. 1988.

[32] 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. Selected Areas Commun.*, Vol. SAC-2, No. 5, pp. 632–647, Sept. 1984.

[33] S. Benedetto, D. Divsalar, G. Montorsi, F Pollara, "Parallel concatenated trellis coded modulation," *Proc. Intl. Comm. Conf. Rec*, pp. 974 - 978, June 1996.

[34] P. Robertson and T. Worz, "Bandwidth-efficient turbo trellis-coded modulation using punctured component codes," *IEEE J. Select. Areas Commun.*, pp. 206–218, Feb. 1998.

[35] C. Fragouli and R.D. Wesel, "Turbo-encoder design for symbol-interleaved parallel concatenated trellis-coded modulation," *IEEE Trans. Commun.*, pp. 425 - 435, March 2001

[36] C.-E.W. Sundberg and N. Seshadri "Coded modulation for fading channels - an overview," *Europ. Trans. on Telecomm. and Related Technol.* Vol. 4, No. 3, pages 309–324, May-June 1993.

[37] L.-F. Wei, "Coded M-DPSK with built-in time diversity for fading channels," *IEEE Trans. on Info. Theory*, Vol. IT-39, No. 6, pages 1820–1839, Nov. 1993.

[38] H. Imai and S. Hirakawa, "A new multilevel coding method using error correcting codes," *IEEE Trans. Inform. Theory*, Vol IT-23, No. 3, pp. 371–377, May 1977.

[39] 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.*, Vol. SP-39, No. 8, pp. 1717–1731, Aug. 1991.

[40] J.W. Modestino and D.G. Daut, "Combined source-channel coding of images," *IEEE Trans. Commun.*, Vol. COM-27, No. 11, pp. 1644–1659, Nov. 1979.

[41] A.R. Calderbank and N. Seshadri, "Multilevel codes for unequal error protection," *IEEE Trans. Inform. Theory*, Vol IT-39, No. 4, pp. 1234–1248, July 1993.

[42] L.-F. Wei, "Coded modulation with unequal error protection," *IEEE Trans. Commun.*, Vol. COM-41, No. 10, pp. 1439–1449, Oct. 1993.

[43] N. Seshadri and C.-E.W. Sundberg, "Multilevel trellis coded modulations for the Rayleigh fading channel," *IEEE Trans. Commun.*, Vol. COM-41, No. 9, pp. 1300–1310, Sept. 1993.

[44] C.-E.W. Sundberg and N. Seshadri, "Coded modulations for fading channels: An overview," *Europ. Trans. Telecomm. and Related Technol.* Vol. 4, No. 3, pp. 309–323, May-June 1993.

[45] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, Part 4, pp. 142-163, 1959.

[46] N. Tanabe and N. Farvardin, "Subband image coding using entropycoded quantization over noisy channels," *IEEE J. Select. Areas Commun.*, pp. 926-943, June 1992.

[47] H. Jafarkhani, P. Ligdas, and N. Farvardin, "Adaptive rate allocation in a joint source/channel coding framework for wireless channels," *Proc. IEEE VTC'96*, pp. 492-496, April 1996.

[48] W. C. Wong, R. Steele, and C.-E. W. Sundberg, Source-Matched Mobile Communications. London, U.K.: Pentech, New York: IEEE Press, 1995.

[49] K.-P. Ho and J. M. Kahn, "Transmission of analog signals using multicarrier modulation: A combined source-channel coding approach," IEEE Trans. Commun., vol. 44, pp. 1432-1443, Nov. 1996.

[50] N. Farvardin and V. Vaishampayan, "On the performance and complexity of channel-optimized vector quantizers," *IEEE Trans. Inform. Theory*, pp. 155-160, Jan. 1991.

[51] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Inform. Theory*, pp. 827-838, Nov. 1987.

[52] K.-P. Ho and J. M. Kahn, "Combined source-channel coding using channel-optimized quantizer and multicarrier modulation," *Proc. IEEE ICC'96*, pp. 1323-1327, June 1996.

[53] A.J. Goldsmith and M. Effros, "Joint Design of Fixed-Rate Source Codes and Multiresolution Channel Codes," *IEEE Trans. Commun.*, pp. 1301-1312, Oct. 1998.

[54] E. Ayanoglu and R. M. Gray, "The design of joint source and channel trellis waveform coders," *IEEE Trans. Inform. Theory*, pp. 855-865, Nov. 1987.

[55] T. R. Fischer and M. W. Marcellin, "Joint trellis coded quantization/ modulation," *IEEE Trans. Commun.*, pp. 172-176, Feb. 1991.

Chapter 8 Problems

1. Consider a (3,1) linear block code where each codeword consists of 3 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 is the error-correction and error-detection capability 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 transmitted codeword $\mathbf{C} = [1011010]$, the received codeword has a syndrome polynomial of zero.

6. The weight distribution of a Hamming code of block length n is given by

$$N(x) = \sum_{i=0}^{n} N_i x^i = \frac{1}{n+1}\left[(1+x)^n + n(1+x)^{.5(n-1)}(1-x)^{.5(n+1)}\right],$$

   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 (8.40) for a Hamming (7,4) code, assuming BPSK modulation of the coded bits with $\gamma = 10$ dB. Compare with the probability of error from the looser bound (8.41) 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 8PSK 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 (5,2) linear block code with $d_{min} = 3$ and $0 \leq E_b/N_0 \leq 20$ dB using the union bound for probability of codeword error. Assume 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. Find the approximate coding gain (8.49) 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 for hard decoding and the approximation (8.49) for soft decoding. What is the difference in coding gain at high SNR for the two decoding techniques?

11. Evalute the upper and lower bounds on codeword error probability, (8.37) and (8.38) 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 8PSK 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 8PSK have $P_M = 10^{-3}$.

13. 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 ($dT_s >> T_c$), 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.

14. Consider a Rayleigh fading channel with BPSK modulation, average SNR of 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 which corrects essentially all of the bit errors, where the interleaver delay is constrained to be less than 5 msec. Your design should include the dimensions of the interleaver, as well as the block code type and the values of n and k.

15. For the trellis of Figure 8.8, determine the state sequence and encoder output assuming an initial state $S = 00$ and information bit sequence $\mathbf{U} = [0110101101]$.

16. Consider the convolutional code generated by the encoder shown in Figure 8.24.

(a) Sketch the trellis diagram of the code.

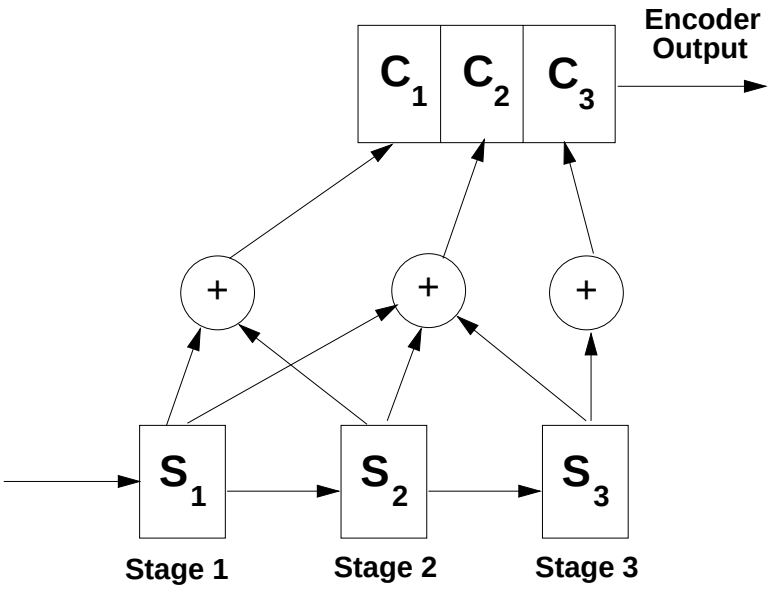


Figure 8.24: Convolutional Encoder for Problems 16 and 17

(b) 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 p as in part (b).

17. 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 (8.75) on probability of bit error for this code assuming SDD and BPSK modulation with $\gamma_b = 10$ dB.

(e) Compute the upper bound (8.79) 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?

18. Consider a channel with coherence time $T_c = 10$ msec and a coded bit rate of $R_s = 50,000$ 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 the system.

19. Suppose you have a 16QAM signal constellation which is trellis encoded using the following scheme: The set partitioning for 16 QAM is shown in Figure 8.20.

(a) Assuming that parallel transitions dominate the error probability, what is the coding gain of this trellis code relative to uncoded 8PSK, given that d_0 for the 16QAM is .632 and d_0 for the 8PSK is .765?

(b) Draw the trellis for this scheme, and assign subsets to the transitions according to the heuristic rules of Ungerboeck.

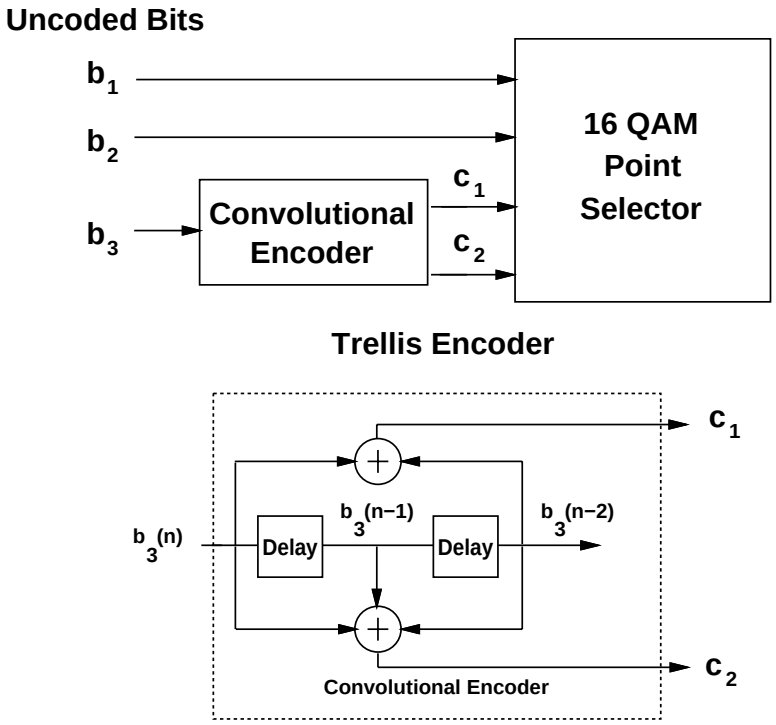


Figure 8.25: 16QAM Trellis Encoder.

(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?

(e) Draw the trellis structure and assign transitions assuming that the convolutional encoder is rate 2/3 (so uncoded bits b_2 and b_3 are input, and 3 coded bits are output).

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 8PSK modulation. 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 8PSK modulation is $T_s = 10$ μsec, what is the data rate for each of the 3 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 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.

Chapter 9

Adaptive Modulation

9.1 Introduction

High-speed wireless data transmission requires robust and spectrally-efficient communication techniques for flat-fading channels. When the channel can be estimated and this estimate sent back to the transmitter, the transmission scheme can be adapted relative to the channel characteristics. Most modulation and coding techniques do not adapt to fading conditions. These nonadaptive methods require a fixed link margin to maintain acceptable performance when the channel quality is poor. Thus, these systems are effectively designed for the worst-case channel conditions, resulting in insufficient utilization of the full channel capacity. Adapting to the signal fading allows the channel to be used more efficiently, since power and rate can be allocated to take advantage of favorable channel conditions. In Chapter 4.3.4, the optimal adaptive transmission scheme that achieves the Shannon capacity of a fading channel was derived. In this chapter we develop practical variable-rate variable-power MQAM modulation techniques for fading channels.

Adaptive transmission, which requires accurate channel estimates at the receiver and a reliable feedback path between the receiver and transmitter, was first proposed in the late sixties [3]. 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. The fact that these issues are less constraining in current systems, coupled with the growing demand for spectrally-efficient communication, has revived interest in adaptive modulation methods. The basic idea behind adaptive transmission is to maintain a constant E_b/N_0 by varying the transmitted power level [3], symbol transmission rate [4], constellation size [5, 6, 7], coding rate/scheme [8], or any combination of these parameters [9, 10, 11]. Thus, without increasing probability of error, also called the Bit Error Rate (BER), these schemes provide high average spectral efficiency by transmitting at high speeds under favorable channel conditions, and reducing throughput as the channel degrades. The performance of these schemes is further improved by combining them with space diversity [12, 13]. Adaptive techniques are also used for high-speed modems [14, 15], satellite links [16, 17, 18], and to minimize distortion or satisfy Quality-of-Service requirements in end-to-end wireless applications. [19, 20]. Our approach is novel relative to all of these adaptive techniques in that we optimize *both* the transmission rate and power to maximize spectral efficiency, while satisfying average power and BER constraints. Although we restrict ourselves to MQAM signal constellations, the same adaptive techniques can be applied to lattice-based constellations [21], which exhibit 1-1.5 dB of shaping gain relative to MQAM.

Cellular systems exploit the power falloff with distance of signal propagation to reuse the same frequency channel at spatially separated locations. While frequency-reuse provides more efficient use

of the limited available spectrum within a given area, it also introduces co-channel interference, which ultimately determines the data rates and corresponding BERs available to each user. Thus, although adaptive modulation techniques increase the spectral efficiency (b/s/Hz) of a single channel, these techniques may also increase co-channel interference levels in a cellular system, thereby requiring a higher reuse distance to mitigate this increased interference power. Adaptive modulation may therefore reduce the *area spectral efficiency*[1] of a cellular system, defined as its average b/s/Hz/km^2. Indeed, while we show in this chapter that channel inversion can significantly reduce the spectral efficiency of a single user relative to optimal adaptation, this inversion is necessary in CDMA cellular systems without multiuser detection to reduce the near-far effect [22, 23]. The area spectral efficiency of FDMA/TDMA cellular systems with the adaptive policies described in this chapter are analyzed in [24], where it is shown that power adaptation typically reduces area spectral efficiency, while rate adaptation improves it. We do not consider the effect of co-channel interference in our analysis below. Thus, our results apply to systems without frequency reuse, or to cellular systems where the co-channel interference is significantly mitigated through cell isolation, sectorization, or adaptive antennas.

There are several practical constraints which determine when adaptive modulation should be used. If the channel is changing faster than it can be estimated and fed back to the transmitter, adaptive techniques will perform poorly, and other means of mitigating the effects of fading should be used. In Chapter 9.6 we find that, for a target BER of 10^{-6}, the BER remains at its target level as long as the total delay of the channel estimator and feedback path is less than $.001\lambda/v$, where v is the vehicle speed and λ the signal wavelength. Thus, at pedestrian speeds of 3.6 Km/Hr the total delay should not exceed 1 ms, and at vehicle speeds of 90 Km/Hr the total delay should not exceed 40 μsec. The former constraint is within the capabilities of existing estimation techniques and feedback channels, while the latter constraint is more challenging. However, a higher BER target loosens the delay constraint: at 10^{-3} BER a total delay constraint of less than $.01\lambda/v$ suffices for good performance. The effects of estimation error are also characterized in Section 7, where we find that the estimation error must be less than 1 dB to maintain the target BER. In Rayleigh fading this bound on estimation error can be achieved using the pilot-symbol assisted estimation technique described in [25] with appropriate choice of parameters.[2] Finally, hardware constraints may dictate how often the transmitter can change its rate and/or power. As part of our analysis we will derive a closed-form expression for how often the transmitter must adapt its signal constellation as a function of the Doppler frequency $f_D = v/\lambda$.

9.2 System Model

Consider a discrete-time channel with stationary and ergodic time-varying gain $\sqrt{g[i]}$ and additive white Gaussian noise $n[i]$. Let $\overline{S}$ denote the average transmit signal power, $N_0/2$ denote the power spectral density of the complex noise $n[i]$, B denote the received signal bandwidth, and $\overline{g}$ denote the average channel gain. The received noise power is thus $2B \times N_0/2 = N_0B$. With appropriate scaling of $\overline{S}$ we can assume that $\overline{g} = 1$. For a constant transmit power $\overline{S}$, the instantaneous received SNR is $\gamma[i] = \overline{S}g[i]/(N_0B)$ and the average received SNR is $\overline{\gamma} = \overline{S}/(N_0B)$. Suppose, however, that we adapt the transmit power at time i based on the channel estimate $\hat{g}[i]$ or, equivalently, on $\hat{\gamma}[i] = \overline{S}\hat{g}[i]/(N_0B)$. We denote the transmit power at time i with this adaptive scheme by $S(\hat{\gamma}[i])$, and the received power at time i is then $\gamma[i]\frac{S(\hat{\gamma}[i])}{\overline{S}}$. Since $g[i]$ is stationary, the distribution of $\gamma[i]$ is independent of i, and we denote this

[1] Unfortunately, the area spectral efficiency is often referred to as just spectral efficiency, which causes some confusion between the two definitions. In this chapter spectral efficiency refers to the b/s/Hz of a single-user channel.

[2] There will be some loss of spectral efficiency for this estimation technique since the pilot symbol rate must be subtracted from the transmitted symbol rate.

distribution by $p(\gamma)$. When the context is clear, we will omit the time reference i relative to γ and $S(\gamma)$.

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 τ_e and that this same estimate is available to the transmitter after a combined estimation and feedback path delay of $\tau = \tau_e + \tau_f$. We also assume ideal coherent phase detection. The channel gain estimation error $\epsilon[i]$ is defined as $\epsilon[i] = \hat{g}[i]/g[i] = \hat{\gamma}[i]/\gamma[i]$. We assume that the feedback path does not introduce any errors, which can be assured by increasing its delay time and using an ARQ transmission protocol. The availability of channel information at the transmitter allows it to adapt its transmission scheme relative to the channel variation. We will initially ignore the effects of estimation error and delay, assuming $\epsilon = 1$ and $\tau = 0$. We then relax these assumptions and determine closed-form expressions for the increase in BER resulting from these effects.

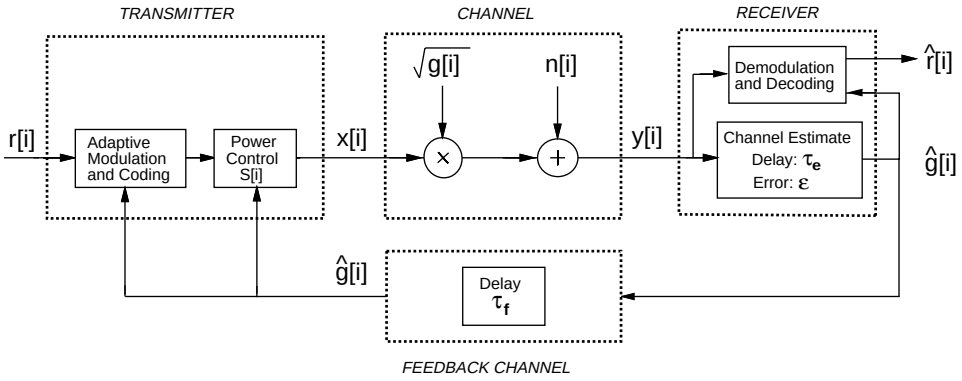


Figure 9.1: System Model.

We will assume $p(\gamma)$ to be either log-normal or exponential (Rayleigh fading) in the numerical calculations below, although our formulas apply for any distribution of γ. The log-normal distribution arises from attenuation of the transmitted signal by surrounding buildings, and the exponential distribution arises from multipath [26]. Although both types of fading will typically be superimposed on the received signal, we consider the two distributions separately for the following reasons. At low speeds the log-normal shadowing is essentially constant, and the Rayleigh fading is sufficiently slow so that it can be estimated and fed back to the transmitter with an estimation error and delay that does not significantly degrade performance. At high speeds these effects may become nonnegligible. In this case, most of the Rayleigh fading can be removed with a sufficient number of diversity branches at the transmitter or receiver, in which case the adaptive modulation need only respond to log-normal channel variations.

The rate of channel variation will dictate how often the transmitter must adapt its rate and/or power, and will also impact the BER increase due to estimation error and delay. For Rayleigh fading we assume the standard Jakes model for the autocorrelation of the channel power gain over time [26]:

$$A_g(\tau) = J_o^2(2\pi v\tau/\lambda), \tag{9.1}$$

where v is the mobile user's velocity and λ is the RF wavelength.

The autocorrelation function for log-normal shadowing is not well-characterized. However, measurements reported in [27] support an autoregressive model:

$$A_g(\tau) = e^{-v|\tau|/X_c}, \tag{9.2}$$

where X_c is the effective autocorrelation distance of the log-normal shadowing. This distance is on the order of 10-100 m, depending on propagation distance [28].

9.3 Variable-Rate Variable-Power MQAM

Shannon capacity places no restriction on the complexity or delay of the multiplexed transmission scheme which achieves capacity. In fact, Shannon theory doesn't tell us anything about how to design this scheme. Therefore, the main emphasis of this chapter is on practical adaptive modulation methods and their spectral efficiency relative to the theoretical capacity results obtained in Chapter 4. Specifically, we consider a variable-rate and variable-power modulation method using MQAM signal constellations. We will see that the same optimization of power and rate that achieves capacity (Chapter 6.3.3) can be applied to our MQAM design. We also obtain a formula for the efficiency difference between our adaptive MQAM technique and the fading channel capacity.

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 and we assume ideal Nyquist data pulses ($\mathrm{sinc}[t/T_s]$) for each constellation[3]. Let $\overline{S}$, B, N_0, $\gamma = \frac{\overline{S}g}{N_0B}$, and $\overline{\gamma} = \frac{\overline{S}}{N_0B}$ be as given in our system model. Since each of our MQAM constellations have Nyquist data pulses ($B = 1/T_s$), the average E_s/N_0 equals the average SNR:

$$\frac{\overline{E_s}}{N_0} = \frac{\overline{S}T_s}{N_0} = \overline{\gamma}. \tag{9.3}$$

The spectral efficiency of our modulation scheme equals its data rate per unit bandwidth (R/B). For fixed M, $R = (\log_2 M)/T_s$. The spectral efficiency for fixed M is therefore $\log_2 M$, the number of bits per symbol. This efficiency is typically parameterized by the average transmit power $\overline{S}$ and the BER of the modulation technique.

In [30] the BER for an AWGN channel with MQAM modulation, ideal coherent phase detection, and SNR γ is bounded by

$$\mathrm{BER} \leq 2e^{-1.5\gamma/(M-1)}. \tag{9.4}$$

A tighter bound good to within 1dB for $M \geq 4$ and $0 \leq \gamma \leq 30$dB is

$$\mathrm{BER} \leq .2e^{-1.5\gamma/(M-1)}. \tag{9.5}$$

Note that these expressions are only bounds, and may differ from BER expressions found in other textbooks. We use these bounds since they are easy to invert, so we can obtain M as a function of a target BER, as we will see shortly.

In a fading channel with nonadaptive transmission (constant transmit power and rate), the received SNR varies with time. The BER in this case is obtained by integrating the BER in AWGN over the fading distribution $p(\gamma)$. For BPSK ($M = 2$) in Rayleigh fading, this integration yields BER $\approx \frac{1}{4\overline{\gamma}}$ at large SNRs (6.58). Without transmitter adaptation, we therefore require $\overline{\gamma} = 24$ dB to obtain a spectral efficiency of 1 at 10^{-3} BER. For $M \geq 4$ we can bound the average BER by integrating over (9.5):

$$\mathrm{BER} \leq \int .2e^{-1.5\gamma/(M-1)}p(\gamma)d\gamma, \quad M \geq 4. \tag{9.6}$$

Setting $M = 4$ in (9.6) yields a required average SNR of $\overline{\gamma} = 26$ dB to obtain a spectral efficiency of 2 at 10^{-3} BER. We will see below that adaptive techniques yield much higher spectral efficiencies at these BER and power specifications.

We now consider adapting the transmit power $S(\gamma)$ relative to γ, subject to the average power constraint $\overline{S}$. The received SNR is then $\gamma S(\gamma)/\overline{S}$, and the BER bound for each value of γ becomes

$$\mathrm{BER}(\gamma) \leq .2\exp\left[\frac{-1.5\gamma}{M-1}\frac{S(\gamma)}{\overline{S}}\right]. \tag{9.7}$$

[3]Practical Nyquist filters with non-zero excess bandwidth will reduce the spectral efficiency.

We can also adjust M and $S(\gamma)$ to maintain a fixed BER. Rearranging (9.7) yields the following maximum constellation size for a given BER:

$$M(\gamma) = 1 + \frac{1.5\gamma}{-\ln(5\text{BER})}\frac{S(\gamma)}{\overline{S}} = 1 + \gamma K\frac{S(\gamma)}{\overline{S}}, \tag{9.8}$$

where

$$K = \frac{-1.5}{\ln(5\text{BER})} < 1. \tag{9.9}$$

We maximize spectral efficiency by maximizing

$$E[\log_2 M(\gamma)] = \int \log_2\left(1 + \frac{K\gamma S(\gamma)}{\overline{S}}\right) p(\gamma)d\gamma, \tag{9.10}$$

subject to the power constraint

$$\int S(\gamma)p(\gamma)d\gamma = \overline{S}. \tag{9.11}$$

The power control policy that maximizes (9.10) has the same form as the optimal power control policy (4.12) that achieves capacity:

$$\frac{S(\gamma)}{\overline{S}} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma K} & \gamma \geq \gamma_0/K \\ 0 & \gamma < \gamma_0/K \end{cases}, \tag{9.12}$$

where γ_0 is the optimized cutoff fade depth. If we define $\gamma_K = \gamma_0/K$ and substitute (9.12) into (9.8) and (9.10) we get that the instantaneous rate is given by $M(\gamma) = \gamma/\gamma_K$ and the maximum spectral efficiency is given by

$$\frac{R}{B} = \int_{\gamma_K}^{\infty} \log_2\left(\frac{\gamma}{\gamma_K}\right) p(\gamma)d\gamma. \tag{9.13}$$

Comparing the power adaptations (4.12) and (9.12) and the spectral efficiencies (4.13) and (9.13) we see that the power adaptation and spectral efficiency for both the optimal transmission scheme and our MQAM technique are the same, with an effective power loss of K in the latter case. In other words, there is a simple relationship between the maximum spectral efficiency of a fading channel and the spectral efficiency of our uncoded adaptive MQAM technique: uncoded MQAM has an effective power loss of K relative to the optimal transmission scheme, *independent of the fading distribution.* Thus, if the capacity of a fading channel is R bps/Hz at SNR $\overline{\gamma}$, uncoded adaptive MQAM requires a received SNR of $\overline{\gamma}/K$ to achieve the same rate. Equivalently, K is the maximum possible coding gain for our adaptive MQAM method. We discuss coding techniques for our adaptive modulation later in the chapter. It is interesting to note that this constant gap between Shannon capacity and the spectral efficiency of MQAM has also been reported for time-invariant channels with ISI and decision-feedback equalization [32, 33].

We compute the efficiency (9.13) at BERs of 10^{-3} and 10^{-6} for both log-normal shadowing (relative to the average dB received power and for a standard deviation $\sigma = 8$dB) and Rayleigh fading in Figures 9.2 and 9.3, respectively. We also plot the capacity (6.10) in these figures for comparison. Note that the gap between (9.13) and (6.10) is the constant K, which is a simple function of the BER (9.9).

We can also apply the suboptimal policies of total and truncated channel inversion to adaptive MQAM. The spectral efficiency with total channel inversion is obtained by substituting $S(\gamma)/S = \sigma/\gamma$ in (9.8) with $\sigma = (\overline{[1/\gamma]})^{-1}$:

$$\frac{R}{B} = \log_2\left(1 + \frac{-1.5}{\ln(5\text{BER})\overline{[1/\gamma]}}\right). \tag{9.14}$$

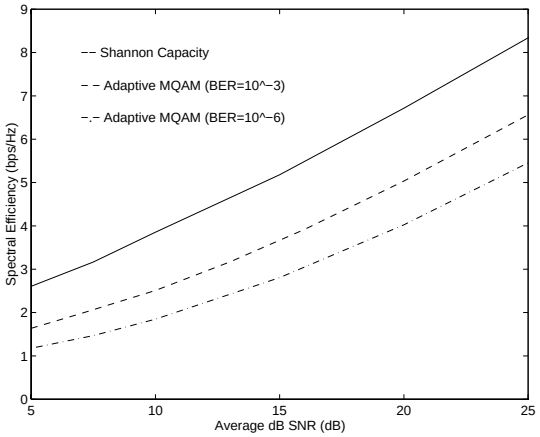


Figure 9.2: Efficiency in Log-Normal Shadowing ($\sigma = 8$dB).

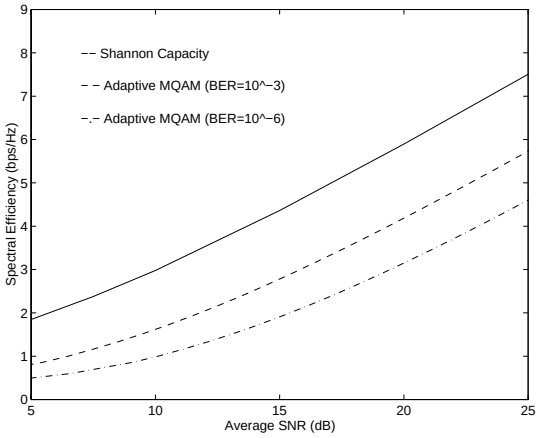


Figure 9.3: Efficiency in Rayleigh Fading.

This spectral efficiency is based on the tight bound (9.5); if $R/B < 4$ the loose bound (9.4) must be used and the spectral efficiency recalculated.

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 (6.13) into (9.8) and multiplying by the probability that $\gamma > \gamma_0$. The maximum value is obtained by optimizing relative to the cutoff level γ_0:

$$\frac{R}{B} = \max_{\gamma_0} \log_2 \left(1 + \frac{-1.5}{\ln(5\text{BER})\overline{[1/\gamma]}_{\gamma_0}} \right) p(\gamma > \gamma_0). \tag{9.15}$$

The spectral efficiency of MQAM with these suboptimal policies, in both log-normal and Rayleigh fading, is evaluated in [34, Figures 3-4].

The spectral efficiencies (9.13), (9.14), and (9.15) place no restrictions on the constellation size; indeed, the size is not even restricted to integer values. While transmission at noninteger rates is possible, the complexity is quite high [40]. 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 impact the spectral efficiency though, surprisingly, not by very much.

9.4 Constellation Restriction

We now restrict ourselves to MQAM constellations of size $M_0 = 0$, $M_1 = 2$, and $M_j = 2^{2(j-1)}, j = 2, ..., N$. We use square constellations for large M due to their inherent spectral efficiency and ease of implementation [31]. We first consider the impact of this restriction on the spectral efficiency of the optimal adaptation policy. We then determine the effect on the channel inversion policies.

9.4.1 Optimal Adaptation

We now consider optimizing the variable-rate variable-power MQAM transmission scheme subject to the constellation restrictions described above. Thus, at each symbol time we transmit a symbol from a constellation in the set $\{M_j : j = 0, 1, \ldots, N\}$: 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 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 the 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 $N + 1$ *fading regions* and associate the constellation M_j with the jth region. The data rate for γ values falling in the jth region is thus $\log_2 M_j$ bits per symbol. If the symbol time $T = 1/B$ then we get a data rate of $(1/T) \log_2 M_j = B \log_2 M_j$ bits per second.

We set the region boundaries as follows. Define

$$M(\gamma) = \frac{\gamma}{\gamma_K^*}, \tag{9.16}$$

where $\gamma_K^* > 0$ is a parameter that will later be optimized to maximize spectral efficiency. Note that substituting (9.12) into (9.8) yields (9.16) with $\gamma_K^* = \gamma_K$. Therefore the appropriate choice of γ_K^* in (9.16) defines the optimal constellation size for each γ when there is no constellation restriction.

Assume now that γ_K^* is fixed and define $M_{N+1} = \infty$. To obtain the constellation size M_j for a fixed γ, we first compute $M(\gamma)$ from (9.16). 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 $\{M_j : j = 0, \ldots, N\}$ that is smaller than $M(\gamma)$. For example, if the fade level γ satisfies $2 \leq \gamma/\gamma_K^* < 4$ we transmit the 2-QAM signal constellation. The region boundaries are located at $\gamma = \gamma_K^* M_j, j = 0, \ldots, N + 1$. Clearly, increasing the number of discrete signal constellations N yields a better approximation to the continuous adaptation (9.8), resulting in a higher spectral efficiency.

Once the regions and associated constellations are fixed we must find a power control policy that satisfies the BER requirement and the power constraint. By (9.8) we can maintain a fixed BER for the constellation $M_j > 0$ using the power control policy

$$\frac{S_j(\gamma)}{\overline{S}} = \begin{cases} (M_j - 1)\frac{1}{\gamma K} & M_j < \frac{\gamma}{\gamma_K^*} \leq M_{j+1} \\ 0 & M_j = 0 \end{cases}. \tag{9.17}$$

A fixed BER implies that the received E_s/N_0 for each constellation M_j is constant:

$$\frac{E_s(j)}{N_0} = \frac{\gamma S_j(\gamma)}{\overline{S}} = \frac{M_j - 1}{K}, \tag{9.18}$$

where $S_j(\gamma)/\overline{S}$ is defined in (9.17). In Table 1 we tabulate the constellation size and power adaptation as a function of γ and γ_K^* for 5 fading regions.

Region(j)	γ Range	M_j	$S_j(\gamma)/\overline{S}$
0	$0 \le \gamma/\gamma_K^* < 2$	0	0
1	$2 \le \gamma/\gamma_K^* < 4$	2	$\frac{1}{K\gamma}$
2	$4 \le \gamma/\gamma_K^* < 16$	4	$\frac{3}{K\gamma}$
3	$16 \le \gamma/\gamma_K^* < 64$	16	$\frac{15}{K\gamma}$
4	$64 \le \gamma/\gamma_K^* < \infty$	64	$\frac{63}{K\gamma}$

Table 9.1: Rate and Power Adaptation for 5 Regions.

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:

$$\frac{R}{B} = \sum_{j=1}^{N} \log_2(M_j) p(M_j \le \gamma/\gamma_K^* < M_{j+1}). \tag{9.19}$$

Since M_j is a function of γ_K^*, we can maximize (9.19) relative to γ_K^*, subject to the power constraint

$$\sum_{j=1}^{N} \int_{\gamma_K^* M_j}^{\gamma_K^* M_{j+1}} \frac{S_j(\gamma)}{\overline{S}} p(\gamma) d\gamma = 1, \tag{9.20}$$

where $S_j(\gamma)/\overline{S}$ is defined in (9.17). There is no closed-form solution for the optimal γ_K^*: in the calculations below it was found using numerical search techniques.

In Figures 9.4 and 9.5 we show the maximum of (9.19) versus the number of regions $(N+1)$ for log-normal shadowing and Rayleigh fading, respectively. We assume a BER of 10^{-3} for both plots. From Figure 9.4 we see that restricting our adaptive policy to just 6 different signal constellations ($M_j = 0, 2, 4, 16, 64, 256$) results in a spectral efficiency that is within 1 dB of the efficiency obtained with continuous-rate adaptation (9.13). A similar result holds for Rayleigh fading using 5 constellations (M_j =0,2,4,16,64).

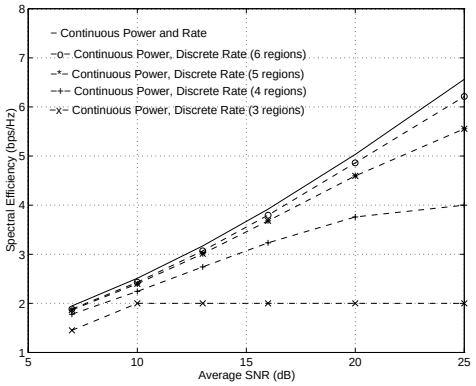


Figure 9.4: Discrete-Rate Efficiency in Log-Normal Shadowing ($\sigma = 8$dB.)

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. We call this the discrete-power discrete-rate policy. Ideally, the fixed transmit power associated with each region should be optimized to maximize spectral efficiency. However, since we do not have a

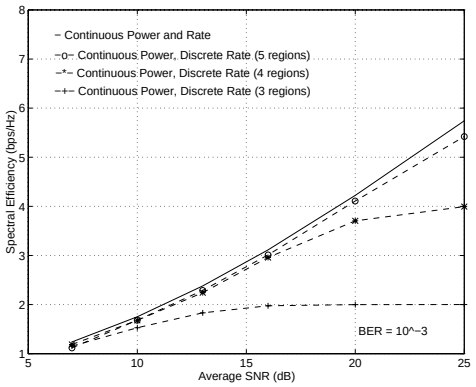


Figure 9.5: Discrete-Rate Efficiency in Rayleigh Fading.

closed-form expression for the spectral efficiency of this policy, we cannot perform the optimization. We will present simulation results for this policy in Chapter 9.5 using suboptimal transmit power values. Even with this suboptimal choice, these simulations demonstrate that keeping the transmit power constant in each region results in less than 2 dBs of power loss relative to the continuous-power discrete-rate policy.

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. For constellation adaptation on a per-symbol basis, the number of regions must be chosen such that the channel gain stays within one region over a symbol time. However, hardware constraints may dictate that the constellation remain 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 and a description of a VLSI prototype can be found in [35]. Hardware advances will eventually make today's constraints obsolete. However, determining how long the channel gain remains within a particular region is of interest, since it determines the tradeoff between the number of regions and the rate of power and constellation adaptation. We now derive this value.

Let $\overline{\tau_j}$ 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 defined above. The jth fading region is then defined as $\{\gamma : A_j \leq \gamma < A_{j+1}\}$. We call $\overline{\tau_j}$ the jth average fade region duration (AFRD). This definition is similar to the average fade duration (AFD) (Chapter 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 τ_j requires a complex derivation based on the joint density $p(\gamma, \dot{\gamma})$, and remains an open problem. However, we can obtain a good approximation using the finite-state Markov model derived in [36]. In that paper, fading is approximated as a discrete-time Markov process with time discretized to the symbol period T_s. The underlying assumption of the model is that γ remains within one region over a symbol period and from a given region the process can only transition to the same region or to adjacent regions. These assumptions are consistent with our model, where γ stays within one region over a symbol time. The transition probabilities between regions under this assumption are given as

$$p_{j,j+1} = \frac{N_{j+1} T_s}{\pi_j}, \quad p_{j,j-1} = \frac{N_j T_s}{\pi_j}, \quad p_{j,j} = 1 - p_{j,j+1} - p_{j,j-1}, \tag{9.21}$$

where N_j is the level-crossing rate at A_j and π_j is the steady-state distribution corresponding to the jth region: $\pi_j = p(A_j \leq \gamma < A_{j+1})$. Since the time in which the Markov process stays in a given state is geometrically distributed [37, 2.66], $\overline{\tau_j}$ is given by

$$\overline{\tau_j} = \frac{T_s}{p_{j,j+1} + p_{j,j-1}} = \frac{\pi_j}{N_{j+1} + N_j}. \tag{9.22}$$

The value of $\overline{\tau_j}$ is thus a simple function of the level crossing rate and the fading distribution. While the level crossing rate is known for Rayleigh fading [26, Section 1.3.4], it cannot be obtained for log-normal shadowing since the joint distribution $p(\gamma, \dot{\gamma})$ for this fading type is unknown.

In Rayleigh fading the level crossing rate is given by

$$N_j = \sqrt{\frac{2\pi A_j}{\overline{\gamma}}} f_D e^{-A_j/\overline{\gamma}}, \tag{9.23}$$

where $f_D = v/\lambda$ is the Doppler frequency. Substituting (9.23) into (9.22) it is easily seen that $\overline{\tau_j}$ is inversely proportional to the Doppler frequency. Moreover, since π_j and A_j do not depend on f_D, if we compute $\overline{\tau_j}$ for a given Doppler frequency f_D, we can compute $\overline{\hat{\tau}_j}$ corresponding to another Doppler frequency $\hat{f}_D$ as

$$\overline{\hat{\tau}_j} = \frac{f_D}{\hat{f}_D} \overline{\tau_j}. \tag{9.24}$$

We tabulate below the $\overline{\tau_j}$ values corresponding to five regions ($M_j = 0, 2, 4, 16, 64$) in Rayleigh fading[4] for $f_D = 100$Hz and two average power levels: $\overline{\gamma} = 10$dB ($\gamma_K^* = 1.22$) and $\overline{\gamma} = 20$dB ($\gamma_K^* = 1.685$). The AFRD for other Doppler frequencies is easily obtained using the table values and (9.24). This table indicates that, even at high velocities, for symbol rates of 100 Kilosymbols/sec the discrete-rate discrete-power policy will maintain the same constellation and transmit power over tens to hundreds of symbols.

Region(j)	$\overline{\gamma} = 10$dB	$\overline{\gamma} = 20$dB
0	2.23ms	.737ms
1	.830ms	.301ms
2	3.00ms	1.06ms
3	2.83ms	2.28ms
4	1.43ms	3.84ms

Table 9.2: Average Fade Region Duration $\overline{\tau_j}$ for $f_D = 100$Hz.

In shadow fading we can obtain a coarse approximation for $\overline{\tau_j}$ based on the autocorrelation function (9.2). Specifically, if $\overline{\tau_j} \approx .1X_c/v$ then the correlation between fade levels separated in time by $\overline{\tau_j}$ is .9. Thus, for a small number of regions it is likely that γ will remain within the same region over this time period.

9.4.2 Suboptimal Policies

A restriction on allowable signal constellations will also affect the total channel inversion and truncated channel inversion policies. Specifically, although the power adaptation policies remain the same, the constellation must be chosen from the signal set $\mathcal{M} = \{0, 2, 4, 16, 64, 256\}$. For total channel inversion the spectral efficiency with this restriction is thus

[4]The validity of the finite-state Markov model for Rayleigh fading channels has been confirmed in [38].

$$\frac{R}{B} = \left\lfloor \log_2\left(1 + \frac{-1.5}{\ln(5\text{BER})\overline{[1/\gamma]}}\right)\right\rfloor_{\mathcal{M}}, \tag{9.25}$$

where $\lfloor x \rfloor_{\mathcal{M}}$ denotes the largest number in the set 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 $\overline{\gamma}$ values where the spectral efficiency without constellation restriction (9.14) equals M. For truncated channel inversion the spectral efficiency is given by

$$\frac{R}{B} = \max_{\gamma_0} \left\lfloor \log_2\left(1 + \frac{-1.5}{\ln(5\text{BER})\overline{[1/\gamma]}_{\gamma_0}}\right)\right\rfloor_{\mathcal{M}} p(\gamma > \gamma_0). \tag{9.26}$$

9.5 Simulation Results

We now present simulation results for adaptive modulation performance. The simulations were done using COSSAP, where fixed-rate transmitter and receiver modules were used as building blocks for the variable-rate transmitter and receiver simulation. The Rayleigh and log-normal shadowing simulation modules in the COSSAP library were used [39], with velocity entered as a parameter. The velocity was chosen so that, over a symbol time T_s, γ stays within one region with high probability. The constellation size transmitted at each symbol time was determined using the discrete-rate adaptive policy outlined in the previous section, assuming perfect instantaneous knowledge of the simulated fade level γ at the transmitter and receiver. The target BER of the adaptive policy was 10^{-3}. We also assumed coherent phase detection at the receiver. Gray coding was used for bit mapping to the MQAM constellations.

We expect our simulated BER to be slightly smaller than the target BER, since (9.7) is an upper bound. An exact BER expression for MQAM with two-dimensional Gray coding is [41]

$$\text{BER}(M) = \alpha_M \operatorname{erfc}\left(\sqrt{\beta_M \frac{E_b}{N_0}}\right) + \text{H.O.T.s}, \tag{9.27}$$

where α_M and β_M are constants which depend on M and the higher order terms (H.O.T.s) are negligible. Moreover, for our continuous-power discrete-rate policy, the E_b/N_0 for the jth signal constellation is approximately

$$\frac{E_b(j)}{N_0} = \frac{E_s(j)}{N_0}\frac{1}{\log_2 M_j} = \frac{M_j - 1}{K \log_2 M_j}. \tag{9.28}$$

We obtain the exact BER for our adaptive policy by averaging over the BER (9.27) for each signal constellation:

$$\text{BER} = \sum_{j=1}^{N} \alpha_{M_j} \operatorname{erfc}\left(\sqrt{\frac{\beta_{M_j}(M_j - 1)}{K \log_2 M_j}}\right) \int_{\gamma_K^* M_j}^{\gamma_K^* M_{j+1}} p(\gamma) d\gamma. \tag{9.29}$$

We plot (9.29) and the simulated BER in Figures 9.6 and 9.7 for log-normal shadowing and Rayleigh fading, respectively. These simulation results are slightly better than the analytical calculation of (9.29), and both are smaller than the target BER of 10^{-3}, for $\overline{\gamma} > 10$ dB. The BER bound of 10^{-3} breaks down at low SNRs, since (9.5) is not applicable to 2-QAM, and we must use the looser bound (9.4). Since our adaptive policy will use the 2-QAM constellation often at low SNRs, the BER will be larger than that predicted from the tight bound (9.5).

The fact that the simulated BER is less than our target at high SNRs implies that the analytical calculations in Figures 9.4 and 9.5 are pessimistic. A slightly higher efficiency could be achieved while still maintaining the target BER of 10^{-3}.

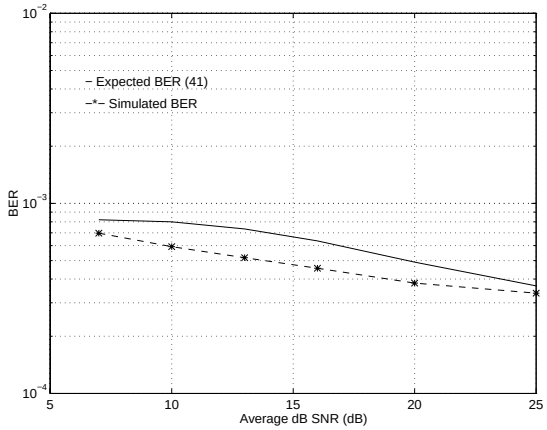


Figure 9.6: BER for Log-Normal Shadowing (6 Regions).

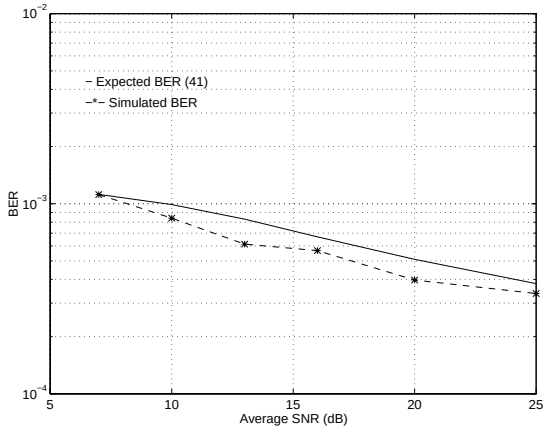


Figure 9.7: BER for Rayleigh Fading (5 Regions).

In Figures 9.8 and 9.9 we show the simulated spectral efficiency corresponding to this simulated BER for the continuous-power discrete-rate policy. These figures also show the simulated efficiency of the discrete-power discrete-rate policy, where the transmit power for each region was chosen to achieve the same simulated BER as the continuous-power discrete-rate policy. We see that even with this suboptimal choice of power assignment, keeping the power constant for each transmit constellation results in a power loss of just 1-2 dB relative to continuous power adaptation. For comparison, we also plot the maximum efficiency (9.13) for continuous power and rate adaptation. Both discrete-rate policies have simulated 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. The average BER in this case is obtained by integrating the probability of error

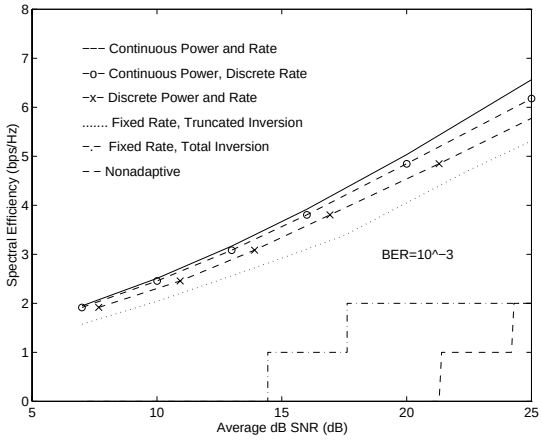


Figure 9.8: Efficiency in Log-Normal Shadowing ($\sigma = 8$dB).

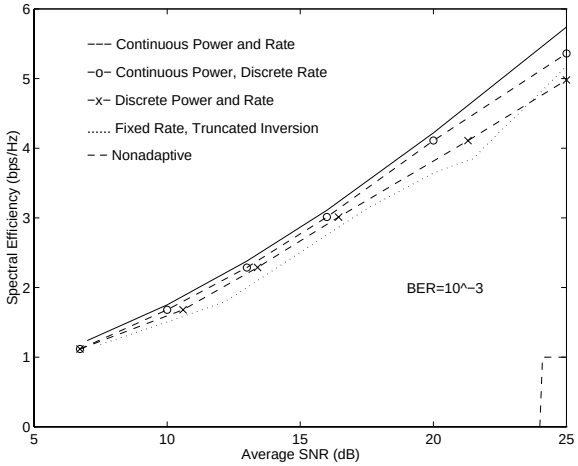


Figure 9.9: Efficiency in Rayleigh Fading.

(9.27) against the fade distribution $p(\gamma)$. The spectral efficiency is obtained by determining the value of M which yields a 10^{-3} BER for the given value of $\overline{\gamma}$. Nonadaptive transmission clearly suffers a large efficiency loss in exchange for its simplicity. However, if the channel varies rapidly and cannot be accurately estimated, nonadaptive transmission may be the best alternative. Similar curves are 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.

9.6 Channel Estimation Error and Delay

We now relax our earlier assumptions about estimation error and delay to consider the case when the estimation error $\epsilon = \hat{\gamma}/\gamma \neq 1$ and the delay $\tau = \tau_f + \tau_\epsilon \neq 0$. We first consider the estimation error. Suppose the transmitter adapts its power and rate relative to a target BER_0 based on the channel estimate $\hat{\gamma}$ instead of the true value γ. From (9.7) the BER is then bounded by

$$\mathrm{BER}(\gamma,\hat{\gamma}) \leq .2\exp\left[\frac{-1.5\gamma}{M(\hat{\gamma})-1}\frac{S(\hat{\gamma})}{\overline{S}}\right] = .2[5\mathrm{BER}_0]^{1/\epsilon}, \tag{9.30}$$

where the second equality is obtained by substituting the optimal rate (9.8) and power (9.12) policies. For $\epsilon = 1$ (9.30) reduces to the target BER_0. For $\epsilon \neq 1$, $\epsilon > 1$ yields an increase in BER, and $\epsilon < 1$ yields a decrease in BER.

The effect of estimation error on BER is given by

$$\overline{\mathrm{BER}} \leq \int_0^\infty \int_{\gamma_0}^\infty .2[5\mathrm{BER}_0]^{\gamma/\hat{\gamma}} p(\gamma,\hat{\gamma}) d\gamma d\hat{\gamma}. \tag{9.31}$$

The joint distribution $p(\gamma,\hat{\gamma})$ will depend on the channel estimation technique. It has been shown recently that when the channel is estimated using pilot symbols, the joint distribution of the signal envelope and its estimate is bi-variate Rayleigh [42]. This joint distribution was then used in [42] 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 the effect of estimation error using (9.30). We plot the BER increase as a function of a constant ϵ in Figure 9.10. This figure shows that for a target BER of 10^{-3} the estimation error should be less than 1dB, and for a target BER of 10^{-6} it should be less than .5dB. 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 [25] with appropriate choice of parameters. When the channel is underestimated ($\epsilon < 1$) the BER decreases but there will also be some loss in spectral efficiency, since the mean of the channel estimate $\overline{\hat{\gamma}}$ will differ from the true mean $\overline{\gamma}$. The effect of this average power estimation error is characterized in [44].

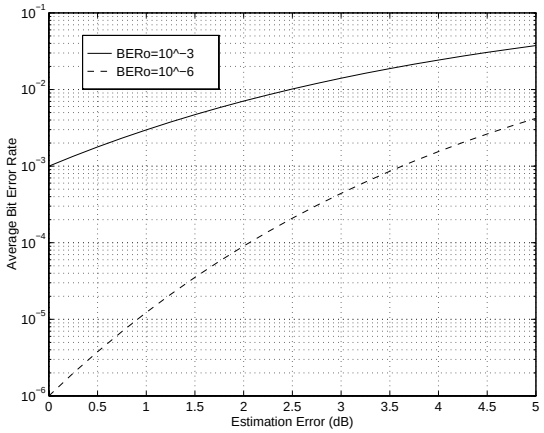


Figure 9.10: Effect of Estimation Error on BER.

Suppose now that the channel is estimated perfectly ($\epsilon = 1$) but the delay τ of the estimation and feedback path is nonzero. Thus, at time t the transmitter will use the delayed version of the channel estimate $\hat{\gamma}(t) = \gamma(t-\tau)$ to adjust its power and rate. The resulting increase in BER is obtained in the same manner as (9.30),

$$\mathrm{BER}(\gamma(t),\hat{\gamma}(t)) \leq .2\exp\left[\frac{-1.5\gamma(t)}{M(\hat{\gamma}(t))-1}\frac{S(\hat{\gamma}(t))}{\overline{S}}\right] = .2[5BER_0]^{\gamma(t)/\gamma(t-\tau)}. \tag{9.32}$$

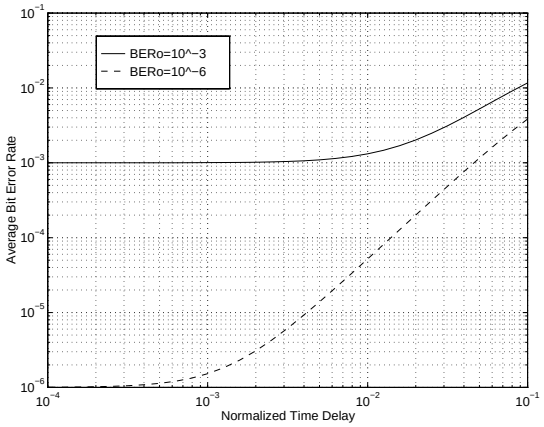


Figure 9.11: Effect of Normalized Delay (τf_D) on BER.

Define $\xi(t,\tau) = \gamma(t)/\gamma(t-\tau)$. Since $\gamma(t)$ is stationary and ergodic, the distribution of $\xi(t,\tau)$ conditioned on $\gamma(t)$ depends only on τ and the value of $\gamma = \gamma(t)$. We denote this distribution by $p_\tau(\xi|\gamma)$. The average BER is obtained by integrating over ξ and γ. Specifically, it is shown in [45] that

$$\text{BER}(\tau) = \int_{\gamma_K}^{\infty} \left[\int_0^{\infty} .2[5BER_0]^{\xi} p_\tau(\xi|\gamma) d_\xi \right] p(\gamma) d\gamma, \tag{9.33}$$

where γ_K is the cutoff level of the optimal policy and $p(\gamma)$ is the fading distribution. The distribution $p_\tau(\xi|\gamma)$ will depend on the autocorrelation of the fading process. A closed-form expression for $p_\tau(\xi|\gamma)$ in Nakagami fading (of which Rayleigh fading is a special case), based on the autocorrelation function (9.1), is derived in [45]. Using this distribution in (9.33) we obtain the average BER in Rayleigh fading as a function of the delay parameter τ. A plot of (9.33) versus the normalized time delay τf_D is shown in Figure 9.11. From this figure we see that the total estimation and feedback path delay must be kept to within $.001/f_D$ to keep the BER near its desired target.

9.7 Coding Issues and Capacity Revisited

A convolutional or block code can be applied to the uncoded bit stream before modulation to reduce the BER. If adaptive modulation is applied to these coded bits, they will not suffer burst errors typically exhibited on fading channels. Since the adaptive modulation keeps the BER constant under all fading conditions by adjusting the transmit power and rate, the probability of error in a deep fade is the same as with little or no fading, thereby eliminating error bursts and the need for an interleaver. Standard decoding algorithms can be applied to the demodulated bits, although some buffering may be required. Unfortunately, block and convolutional codes are not spectrally-efficiency, and would therefore reduce some of the efficiency gains of the variable-rate scheme. A more effective coding scheme is to superimpose a trellis code on top of the adaptive modulation. This superimposed coding technique is investigated in Chapter 9.8, where we find that it is difficult to obtain more than 4 dB of coding gain using a trellis code of reasonable complexity. Thus, the constant gap (9.9) between the spectral efficiency of adaptive modulation and Shannon capacity, exhibited in Figures 9.2 and 9.3, cannot be fully closed. This discrepancy between Shannon capacity and achievable rates arises from the lack of complexity and implementation constraints inherent to Shannon theory. However, the derivation and general form of the optimal power and rate adaptation for our MQAM scheme were identical to that of the Shannon

capacity analysis. Thus, although we cannot reach the Shannon limit, the intuition and general strategy of optimal adaptation in a Shannon sense was a useful guide in our adaptive modulation design.

Bibliography

[1] A.J. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*. Nov. 1997.

[2] A. J. Goldsmith "Design and performance of high-speed communication systems in time-varying radio channels," Ph.D. dissertation, Dept. Elec. Engin. Comput. Science, University of California at Berkeley, 1994.

[3] J. F. Hayes, "Adaptive feedback communications," *IEEE Transactions on Communication Technology*, vol. COM-16, pp. 29–34, February 1968.

[4] J. K. Cavers, "Variable-rate transmission for Rayleigh fading channels," *IEEE Transactions on Communications*, vol. COM-20, pp. 15–22, February 1972.

[5] S. Otsuki, S. Sampei, and N. Morinaga, "Square-QAM adaptive modulation/TDMA/TDD systems using modulation level estimation with Walsh function," *Electronics Letters*, vol. 31, pp. 169–171, February 1995.

[6] W. T. Webb and R. Steele, "Variable rate QAM for mobile radio," *IEEE Transactions on Communications*, vol. COM-43, pp. 2223–2230, July 1995.

[7] 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," in *Proceedings of the IEEE VTC'95*, pp. 221–225, July 1995.

[8] B. Vucetic, "An adaptive coding scheme for time-varying channels," *IEEE Transactions on Communications*, vol. COM-39, pp. 653–663, May 1991.

[9] S. M. Alamouti and S. Kallel, "Adaptive trellis-coded multiple-phased-shift keying for Rayleigh fading channels," *IEEE Transactions on Communications*, vol. COM-42, pp. 2305–2314, June 1994.

[10] T. Ue, S. Sampei, and N. Morinaga, "Symbol rate and modulation level controlled adaptive modulation/TDMA/TDD for personal communication systems," in *Proceedings of the IEEE VTC'95*, pp. 306–310, July 1995.

[11] H. Matsuoka, S. Sampei, N. Morinaga, and Y. Kamio, "Symbol rate and modulation level controlled adaptive modulation/TDMA/TDD for personal communication systems," in *Proceedings of the IEEE VTC'96*, pp. 487–491, April 1996.

[12] S. Sampei, N. Morinaga, and Y. Kamio, "Adaptive modulation/TDMA with a BDDFE for 2 Mbit/s multi-media wireless communication systems," in *Proceedings of the IEEE VTC'95*, pp. 311–315, July 1995.

[13] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques." *IEEE Trans. Vehic. Technol.*, pp. 1165–1181, July 1999.

[14] J.A.C. Bingham, "Multicarrier modulation for data transmission: an idea whose time has come," *IEEE Comm. Mag.*, Vol. 28, No. 5, pp. 5–14, May 1990.

[15] P.S. Chow, J.M. Cioffi, and John A.C. Bingham, "A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels," *IEEE Trans. Commun.*, Vol. 43, No. 2/3/4, Feb.-Apr. 1995.

[16] M.Filip and E. Vilar, "Optimum utilization of the channel capacity of a satellite link in the presence of amplitude scintillations and rain attenuation," *IEEE Trans. Commun.,* Vol. 38, No. 11, pp. 1958–1965, Nov. 1990.

[17] A.M. Monk and L.B. Milstein, "Open-loop power control error in a land mobile satellite system," *IEEE J. Select. Areas. Commun.*, pp. 205–212, Feb. 1995.

[18] J.L. Rose, "Satellite communications in the 30/20 GHz band," *Satellite Communications*, pp. 155–162, On-line Publications, 1985.

[19] R.V. Cox, J. Hagenauer, N. Seshadri, and C.-E.W. Sundberg, "Subband speech coding and matched convolutional channel coding for mobile radio channels," *IEEE Trans. Signal Proces.*, Vol. 39, pp. 1717–1731, Aug. 1991.

[20] L.C. Yun and D.G. Messerschmitt, "Variable Quality of Service in CDMA systems by statistical power control, " *IEEE Intl. Commun. Conf. Rec.*, pp. 713–719, June 1995.

[21] G. David Forney, "Trellis shaping," *IEEE Trans. Inform. Theory*, pp. 281–300, March 1992.

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

[23] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Trans. Vehic. Technol.*, pp. 57–62, Feb. 1992.

[24] M.-S. Alouini and A. J. Goldsmith, "Area spectral efficiency of cellular mobile radio systems," *IEEE Trans. Vehic. Technol.*, pp. 1047–1066, July 1999.

[25] J. K. Cavers, "An analysis of pilot symbol assisted modulation for Rayleigh fading channels," *IEEE Trans. Vehic. Technol.*, pp. 686–693, Nov. 1991.

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

[27] R. Vijayan and J. M. Holtzman, "Foundations for level crossing analysis of handoff algorithms," *Proc. IEEE ICC Conf.*, pp. 935–939, June 1993.

[28] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electron. Lett.*, Vol 27, pp. 2145–2146, Nov. 7, 1991.

[29] W.C.Y. Lee, "Estimate of channel capacity in Rayleigh fading environment," *IEEE Trans. Vehic. Technol.*, Vol VT-39, No. 3, pp. 187–189, Aug. 1990.

[30] G.J.Foschini and J.Salz, "Digital communications over fading radio channels," *Bell Systems Technical Journal*, pp. 429-456, Feb. 1983

[31] J.G. Proakis, *Digital Communications*, 2nd Ed., New York: McGraw-Hill, 1989.

[32] R. Price, "Non-linearly feedback equalized PAM versus capacity for noisy filter channels," *Proc. IEEE ICC Conf.*, June 1972.

[33] M. V. Eyuboglu, "Detection of coded modulation signals on linear, severely distorted channels using decision feedback noise prediction with interleaving," *IEEE Trans. Commun.,* Vol-COM 36, No. 4, pp. 401–409, April 1988.

[34] A. Goldsmith and P. Varaiya, "Increasing spectral efficiency through power control," in *Proc. IEEE ICC Conf.*, pp. 600–604, June 1993.

[35] M. Filip and E. Vilar, "Implementation of adaptive modulation as a fade countermeasure," *Intl. J. Sat. Commun.,* Vol. 12, pp. 181–191, 1994.

[36] H. S. Wang and N. Moayeri, "Finite-state Markov channel - a useful model for radio communication channels," *IEEE Trans. Vehic. Technol.,* Vol VT-44, No. 1, pp. 163–171, Feb. 1995.

[37] L. Kleinrock *Queueing Systems Volume I: Theory*, Wiley: 1975.

[38] H. S. Wang and P.-C. Chang, "On verifying the first-order Markov assumption for a Rayleigh fading channel model," *IEEE Trans. Vehic. Technol.,* Vol VT-45, No. 2, pp. 353–357, May 1996.

[39] *COSSAP Model Libraries*, Volume 1, Version 6.7, Synopsys 1994.

[40] 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. Selected Areas Commun.*, Vol. SAC-2, No. 5, pp. 632–647, Sept. 1984.

[41] M.K.Simon, S.M.Hinedi, W.C.Lindsey, *Digital Communication Techniques*, New Jersey: Prentice Hall, 1995

[42] X. Tang, M.-S. Alouini, and A. Goldsmith. "The effect of channel estimation error on MQAM BER performance in Rayleigh fading channels," *IEEE Transactions on Communications*, Vol 47, No. 12, pp. 1856-1864, Dec. 1999.

[43] M.G. Jansan and R. Prasad, "Capacity, throughput, and delay analysis of a cellular DS CDMA system with imperfect power control and imperfect sectorization," *IEEE Trans. Vehic. Technol.*, Vol. 44, pp. 67–75, Feb. 1995.

[44] A. J. Goldsmith and L. J. Greenstein, "Effect of average power estimation error on adaptive MQAM modulation," *Proc. IEEE ICC'97.*

[45] M.-S. Alouini and A. J. Goldsmith, *Kluwer Journal on Wireless Personal Communications.*, pp. 119-143, May 2000.

[46] S.-G. Chua and A.J. Goldsmith, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, pp. 595-602, May 1998.

Chapter 10

Multiple Antenna Systems

Multiple antennas at the transmitter and/or receiver of a mobile system can increase data rates and performance (multiple input multiple output systems) or reduce ISI and interference from other users (smart antennas). In this chapter we treat both of these techniques and discuss the performance improvement that can be achieved via each technique.

10.1 Multiple Input Multiple Output (MIMO) Systems

MIMO systems are defined as point-to-point communication links with multiple antennas at both the transmitter and receiver. The use of multiple antennas at both transmitter and receiver clearly provide enhanced performance over diversity systems where either the transmitter or receiver, but not both, have multiple antennas. In particular, recent research has shown that MIMO systems can significantly increase the data rates of wireless systems without increasing transmit power or bandwidth. The cost of this increased rate is the added cost of deploying multiple antennas, the space requirements of these extra antennas (especially on small handheld units), and the added complexity required for multi-dimensional signal processing. Recent work in MIMO systems includes capacity of these systems under different assumptions about channel knowledge, optimal coding and decoding for these systems, and transmission strategies for uncoded systems.

10.1.1 The Narrowband Multiple Antenna System Model

A narrowband (flat-fading) point to point communication system employing n transmit and m receive antennas is shown in Figure 10.1

This system can be represented by the following discrete time model:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} N_1 \\ \vdots \\ N_m \end{bmatrix}$$

or simply as $\overline{y} = H\overline{x} + \overline{N}$. Here $\overline{x}$ represents the n-dimensional transmitted symbol, $\overline{N}$ is the m-dimensional additive white Gaussian noise (AWGN) vector, and the channel matrix H consists of zero mean (Rayleigh Fading) complex circular Gaussian random variables h_{ij} representing the channel gain from transmit antenna j to receive antenna i. Without loss of generality we normalize the noise so that the noise covariance matrix is an identity matrix. Note that although the dependence on time is

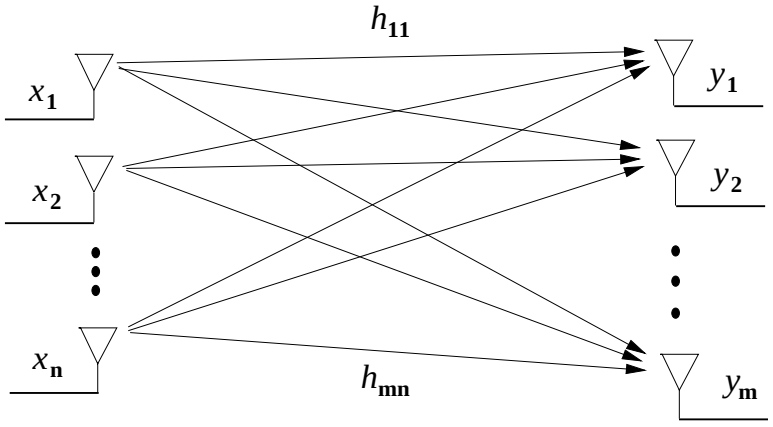


Figure 10.1: MIMO Systems.

suppressed here, $\overline{x}$, $\overline{y}$, $\overline{N}$ and H are all stochastic processes. We assume that the receiver is able to estimate the channel state H perfectly. So at each instant H is known at the receiver.

The transmit power constraint is given as

$$\sum_{i=1}^{n} \mathrm{E}[x_i x_i^*] = P,$$

or, equivalently, as

$$\mathrm{trace}(\mathrm{E}[\overline{x}\overline{x}^\dagger]) \;=\; P.$$

10.1.2 Transmit Precoding and Receiver Shaping

In general an R symbols/s input data stream can be split into r parallel, independent data streams, producing r-tuples $\overline{\tilde{x}}$ at a rate R/r symbols/s. The actual input to the antennas $\overline{x}$ is generated through a linear transformation on $\overline{\tilde{x}}$ as

$$\overline{x} = M\overline{\tilde{x}},$$

where M is an $n \times r$ fixed matrix. This operation is sometimes called transmit precoding. A similar operation, called receiver shaping, can be performed at the receiver by multiplying the channel output with a $r \times n$ matrix F, as shown in Figure 10.2.

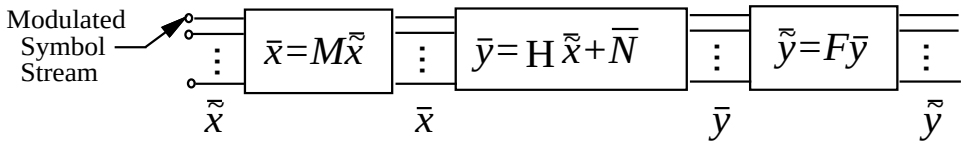


Figure 10.2: Transmit Precoding and Receiver Shaping.

The relevance of these operations will become obvious in the next section. The overall system can be described as follows:

$$\begin{aligned} \overline{\tilde{y}} &= F\overline{y} \\ &= FH\overline{x} + F\overline{N} \\ &= FHM\overline{\tilde{x}} + F\overline{N} \end{aligned}$$

Note that the rank of the input covariance matrix $Q = \mathrm{E}[\overline{x}\overline{x}^\dagger]$ is equal to r, the number of independent streams being simultaneously transmitted. For example if $\overline{x} = \overline{M}x$ where $\overline{M}$ is a constant vector, the input covariance matrix $Q = \mathrm{E}[xx^*]\overline{M}\overline{M}^\dagger$ has unit rank.

Optimal decoding of the received signal requires maximum likelihood demodulation. However, if the modulated symbols are chosen from an alphabet of size $|\mathcal{X}|$, then ML demodulation requires an exhaustive search over $|\mathcal{X}|^r$ possibilities for the input r-tuple. In general (for a non-trivial H), when the transmitter does not know H this complexity cannot be reduced further. So the optimal decoding complexity without the channel state information at the transmitter (CSIT) is exponential in the rank of the input covariance matrix, which is the same as the number of independent streams being transmitted simultaneously. This decoding complexity is typically prohibitive for even small numbers of antennas. However, decoding complexity is significantly reduced if the channel can be measured at the receiver and fed back to the transmitter, as we see in the next section.

10.1.3 Parallel Decomposition of the MIMO Channel

Let us consider the case of perfect Channel State Information at the Transmitter (CSIT). In other words, both the transmitter and the receiver know H at each instant. Further let the instantaneous channel matrix have a singular value decomposition (SVD)

$$H = U\Lambda V, \tag{10.1}$$

where U and V are unitary matrices (i.e. $UU^\dagger = I_n$ and $VV^\dagger = I_m$) and Λ is the diagonal matrix of singular values of H. Now suppose the transmitter chooses $M = V^\dagger$ and the receiver chooses $F = U^\dagger$. The Multiple Input Multiple Output (MIMO) channel is then transformed into r ($\leq \min(m, n)$) parallel non-interfering Single Input Single Output (SISO) channels:

$$\begin{aligned} \overline{\tilde{y}} &= U^\dagger U\Lambda V V^\dagger \overline{\tilde{x}} + U^\dagger \overline{N} \\ &= \Lambda\overline{\tilde{x}} + \overline{\tilde{N}}, \end{aligned}$$

where $\overline{\tilde{N}} = +U^\dagger\overline{N}$. This parallel decomposition is shown in Figure 10.3.

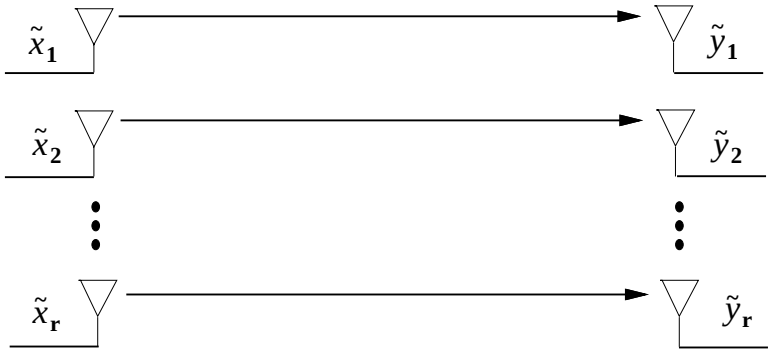


Figure 10.3: Parallel Decomposition of the MIMO Channel.

Since the SISO channels do not interfere the optimal (maximum likelihood) demodulation complexity is now only $r|\mathcal{X}|$ instead of $|\mathcal{X}|^r$. Note that multiplication by a unitary matrix does not change the distribution of white Gaussian noise, i.e. $\overline{N}$ and $\overline{\tilde{N}}$ are identically distributed.

10.1.4 MIMO Channel Capacity

The MIMO decomposition described above allows a simple characterization of the MIMO channel capacity when both transmitter and receiver have perfect knowledge of the channel matrix H. The capacity formula is [**?**]:

$$C = \max_{Q:Tr(Q)\leq P} \log|I + HQH^\dagger|, \tag{10.2}$$

where the maximum is taken over all matrices Q that satisfy the average power constraint.

By substituting the matrix SVD (10.1) into (10.2) and using properties of unitary matrices yields

$$C = \max_{\{P_i\}:\sum_i P_i\leq P} \sum_i B\log\left(1+\frac{\lambda_i^2 P_i}{N_0 B}\right), \tag{10.3}$$

which is similar to the capacity formula in flat fading (4.9) or in frequency-selective fading with constant channel gains (**??**). We therefore get a similar water-filling power allocation for the MIMO channel with the channel gain given by the eigenvalues:

$$\frac{P_i}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_i} & \gamma_i \geq \gamma_0 \\ 0 & \gamma_i < \gamma_0 \end{cases} \tag{10.4}$$

for some cutoff value γ_0, where $\gamma_i = \lambda_i^2 P/(N_0 B)$. The resulting capacity is then

$$C = \sum_{i=1(\gamma_i\geq\gamma_0)} B\log(\gamma_i/\gamma_0). \tag{10.5}$$

10.1.5 Beamforming

In this section we consider the case when the transmitter does not know the instantaneous channel. It is no longer possible to transform the MIMO channel into non-interfering SISO channels. Since the decoding complexity is exponential in r, we can keep the complexity low by keeping r small. Of particular interest is the case where $r = 1$. A transmit strategy where the input covariance matrix has unit rank is called *beamforming*. This corresponds to the precoding matrix being just a column vector $M = \overline{c}$, the beamforming vector, as shown in Figure 10.4

Spatial matched filtering yields a single SISO AWGN channel as follows.

$$\begin{aligned} \tilde{y} &= \frac{\overline{c}^\dagger H^\dagger}{||\overline{c}^\dagger H^\dagger||} y \\ &= \frac{\overline{c}^\dagger H^\dagger}{||\overline{c}^\dagger H^\dagger||} H\overline{c}x + \frac{\overline{c}^\dagger H^\dagger}{||\overline{c}^\dagger H^\dagger||}\overline{N} \\ &= ||H\overline{c}||x + \tilde{N} \end{aligned}$$

where $\tilde{N}$ is zero-mean, unit-variance AWGN.

The optimal demodulation complexity with beamforming is of the order of $|\mathcal{X}|$, the size of the modulation symbol alphabet. Recall that $\overline{c}$ does not change with time. For a given choice of $\overline{c}$ and a given channel matrix H the SNR becomes

$$\text{SNR} = \overline{c}^\dagger H^\dagger H\overline{c}\text{E}[xx^*]$$

Define the optimal choice of $\overline{c}$ as one that maximizes the average SNR (averaged over the distribution of H). Note that optimality can also be defined so that the information theoretic capacity of this fading channel is maximized. However, for now, we are interested in uncoded systems and therefore we choose the average SNR as the optimality criterion.

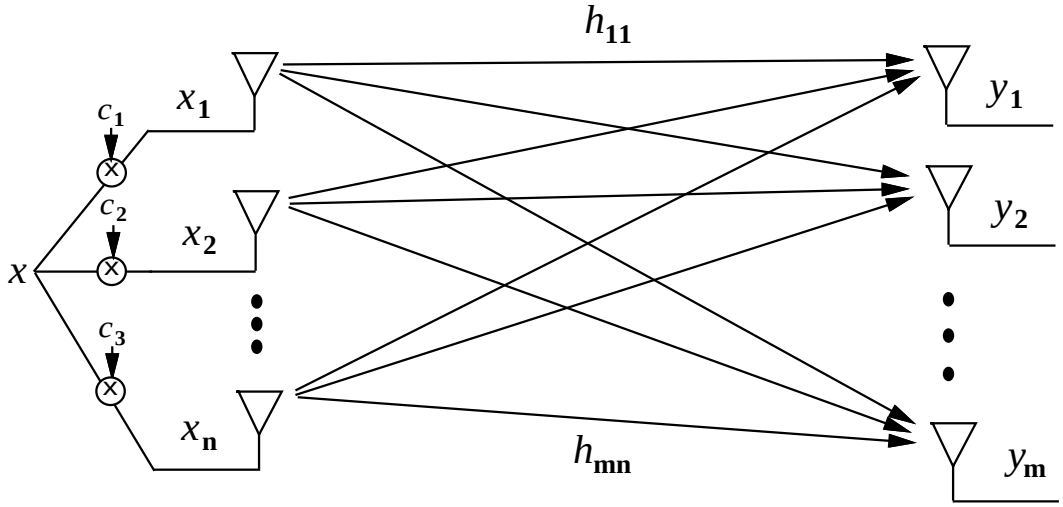


Figure 10.4: MIMO Channel with Beamforming.

Solving for the optimal beamforming vector

We wish to choose the beamforming vector $\overline{c}$ to maximize the average SNR given by

$$\mathrm{E}[\mathrm{SNR}] = P\overline{c}^{\dagger}\mathrm{E}[H^{\dagger}H]\overline{c}$$

subject to

$$\overline{c}^{\dagger}\overline{c} = 1.$$

We need this constraint in order to satisfy the transmit power constraint. But the solution to this optimization problem is simply the unit norm principal eigenvector (the eigenvector corresponding to the maximum eigenvalue) of the positive definite matrix $\mathrm{E}[H^{\dagger}H]$.

I.i.d. Fading

For i.i.d. fades, i.e. when the channel fades between any transmit-receive antenna pair are independent and identically distributed, $\mathrm{E}[H^{\dagger}H]$ is a multiple of the identity matrix. Thus without loss of generality, we could choose $\overline{c} = [1, 0, 0, \cdots, 0]^T$. So for i.i.d. fades there is no gain from using multiple transmit antennas. However the magnitude of the average received SNR is directly proportional to the number of receive antennas. Hence multiple receive antennas improve average received SNR with i.i.d. fading.

Independent Fading

Each row of the channel matrix H is an n-dimensional random vector. Let the covariance matrix for the i^{th} row be denoted by K_i. For independent fades between all transmit-receive antenna pairs, the K_i are all diagonal matrices. $\mathrm{E}[H^{\dagger}H] = \sum_{i=1}^{m} K_i$ is also a diagonal matrix. Again the principal eigenvector is $\overline{c} = [\ 0, 0, \cdots, 0, 1, 0, \cdots, 0\ ]^T$. This again corresponds to using just one transmit antenna alone. It can easily be verified that the transmit antenna is the one that has the highest sum of average channel power gains to all the receive antennas. Again, multiple receive antennas improve the received SNR.

Correlated Fading

In general, for correlated fading, the principal eigenvector of $\mathrm{E}[H^\dagger H]$ may use all transmit antennas. It is easy to verify this by constructing an example. We leave this as an exercise to the reader. So for correlated fading, one does gain from using multiple transmit antennas as well as multiple receive antennas.

10.2 Space-time codes

The key result discussed in the previous subsection motivates the study of channel codes, called space-time codes to pursue the very high throughput predicted by information theory. As we saw earlier, if the transmitter knows the channel it is possible to transform it into several parallel non-interfering SISO channels and the codec technology for SISO channels is well established. However if the transmitter does not know the instantaneous channel, inherently multi-dimensional codes are required. Codewords are now long matrices instead of vectors. The optimal decoding complexity of these codewords is exponential in the number of antennas. Designing these codewords itself is a complex problem and represents a vast area of research in itself. Some of the approaches explored include treating the transmission from each antenna as an independent user using conventional scalar codes in conjunction with multiuser detection techniques at the receiver (layered space time codes). However most of these suboptimal approaches suffer significant performance penalties.

10.3 Smart Antennas

Smart antennas generally consist of an antenna array combined with signal processing in both space and time. The spatial processing introduces a new degree of freedom in the system design with enormous potential to improve performance, including range extension, capacity enhancement, higher data rates, and better BER performance [**?**].

The main impediments to high-performance wireless communications are the interference from other users (cochannel interference) and the intersymbol interference (ISI) and signal fading caused by multipath. The cochannel interference limits the system capacity, defined as the number of users which can be serviced by the system. However, since interference typically arrives at the receiver from different directions, smart antennas can exploit these differences to reduce cochannel interference, thereby increasing system capacity. The reflected multipath components of the transmitted signal also arrive at the receiver from different directions, and spatial processing can be used to attenuate the multipath, thereby reducing ISI and flat-fading. Since data rate and BER are degraded by these multipath effects, reduction in multipath through spatial processing can lead to higher data rates and better BER performance.

The complexity of spatial domain processing along with the required real estate of an antenna array make the use of smart antennas in small, lightweight, low-power handheld devices unlikely in next-generation systems. However the base stations for these systems can use antenna arrays with space-time processing at the transmitter to reduce cochannel interference and multipath, providing similar performance advantages as smart antennas in the receiver. An excellent overview of smart antennas can be found in [**?**].

Appendix 10.A
Derivations of the Alternate Representations of the Gaussian Q-function and its Square

A byproduct of Craig's work on the probability of error for two-dimensional signal constellations [**?**] was the alternate representation of the Gaussian Q-function given in (**??**). An extension of this representation for the square of the Gaussian Q-function (**??**) was obtained by Simon and Divsalar [**?**]. In this appendix we present another simple method of proving the alternate representations of the Gaussian Q-function and its square.

A-1 Proof of Eqn. (??)

The proposed proof is an extension of the classical method to evaluate the Laplace-Gauss integral [**?**, Eqn. (3.321.3)]:

$$J(a) \stackrel{\triangle}{=} \int_0^\infty e^{-a^2\ x^2}\ dx = \frac{\sqrt{\pi}}{2\ a}; \quad a > 0. \tag{10.6}$$

Let us consider the double integral

$$\int_0^\infty \int_x^\infty e^{-\frac{u^2+v^2}{2}}\ du\ dv; \quad x \geq 0. \tag{10.7}$$

Because of separability (10.7) can be rewritten as

$$\underbrace{\int_0^\infty e^{-u^2/2}\ du}_{J\left(1/\sqrt{2}\right)} \underbrace{\int_x^\infty e^{-v^2/2}\ dv}_{\sqrt{2\pi}\ Q(x)} = \pi\ Q(x), \tag{10.8}$$

where we see that each integral in the LHS of (10.8) is a well-defined function. Further, transformation to polar coordinates $u = r\cos\phi$ and $v = r\sin\phi$ ($du\ dv = rdr\ d\phi$) may be carried out in (10.7) giving

$$\begin{aligned} \int_0^\infty \int_x^\infty e^{-\frac{u^2+v^2}{2}}\ du\ dv &= \int_0^{\pi/2} \int_{x/\sin\phi}^\infty e^{-r^2/2}\ rdr\ d\phi \\ &= \int_0^{\pi/2} \exp\left(-\frac{x^2}{2\ \sin^2\phi}\right)\ d\phi. \end{aligned} \tag{10.9}$$

Equating the RHS of (10.8) and (10.9) we obtain an alternate proof of the desired result (**??**). Note that another purely algebraic proof of the result (**??**) which can be implied from the work of Pawula *et al.* [**?**] is given in detail in [**?**, Appendix 4A].

A-2 Proof of Eqn. (??)

The proof presented in Appendix A-1 can be easily extended to arrive at the alternate representation of $Q^2(\cdot)$ given in (**??**). Let us now consider the following double integral

$$\int_x^\infty \int_x^\infty e^{-\frac{u^2+v^2}{2}}\ du\ dv; \quad x \geq 0. \tag{10.10}$$

Again because of separability, (10.10) can be rewritten as

$$\underbrace{\int_x^\infty e^{-u^2/2}\ du}_{\sqrt{2\pi}\ Q(x)}\ \underbrace{\int_x^\infty e^{-v^2/2}\ dv}_{\sqrt{2\pi}\ Q(x)} = 2\ \pi\ Q^2(x), \tag{10.11}$$

where each integral in the LHS of (10.11) is the Gaussian Q-function multiplied by $\sqrt{2\pi}$. The transformation to polar coordinates $u = r\cos\phi$ and $v = r\sin\phi$ ($du\ dv = rdr\ d\phi$) is carried out in (10.10) and by symmetry the rectangular region of integration is divided into two equal triangular parts giving

$$\begin{aligned}\int_x^\infty \int_x^\infty e^{-\frac{u^2+v^2}{2}}\ du\ dv &= 2\int_0^{\pi/4}\int_{x/\sin\phi}^\infty e^{-r^2/2} r dr d\phi \\ &= 2\int_0^{\pi/4} \exp\left(-\frac{x^2}{2\sin^2\phi}\right)\ d\phi. \end{aligned}\tag{10.12}$$

Equating (10.11) and (10.12) we obtain an alternate proof of the Simon-Divsalar result (**??**).

Appendix 10.B
Closed-Form Expressions for $\int_0^{\pi/2}\left(\frac{\sin^2\phi}{\sin^2\phi+c}\right)^m\ d\phi$

The alternate representation of the Gaussian Q-function can also be used to find closed-form expressions for integrals not tabulated in classical table of integrals such as [**?**, **?**]. As an example we evaluate in this appendix the integral $I_m(c)$ defined by

$$I_m(c) \triangleq \int_0^{\pi/2}\left(\frac{\sin^2\phi}{\sin^2\phi + c}\right)^m\ d\phi. \tag{10.13}$$

To do so consider first the integral $J_m(a,b)$ defined by

$$J_m(a,b) \triangleq \frac{a^m}{\Gamma(m)}\int_0^{+\infty} e^{-at} t^{m-1} Q(\sqrt{bt}) dt, \quad m \geq 0. \tag{10.14}$$

This integral (10.14) has a known closed-form expression. When m is a positive real number the integral $J_m(a,b)$ is given by [**?**, Eqn. (A8)]

$$J_m(a,b) \triangleq J_m(c) = \frac{\sqrt{c/\pi}}{2(1+c)^{m+1/2}}\frac{\Gamma(m+1/2)}{\Gamma(m+1)}\ {}_2F_1\left(1, m+1/2; m+1; \frac{1}{1+c}\right), \tag{10.15}$$

where $c = b/(2a)$ and ${}_2F_1(.,.;.;.)$ denotes the the *hypergeometric series* (known also as the *Gauss hypergeometric function*). When m is a positive integer, the integral $J_m(a,b)$ reduces to [**?**, Eqn. (7.4.15)], [**?**, Eqn. (A13)]

$$J_m(a,b) \triangleq J_m(c) = [P(c)]^m \sum_{k=0}^{m-1}\binom{m-1+k}{k}\ [1-P(c)]^k, \tag{10.16}$$

where

$$P(x) = \frac{1}{2}\left(1 - \sqrt{\frac{x}{1+x}}\right); \quad x \geq 0. \tag{10.17}$$

Using the alternate representation of the Gaussian Q-function (**??**) in (10.15), we obtain

$$J_m(a,b) = \frac{a^m}{\Gamma(m)} \int_0^\infty e^{-at} t^{m-1} \left(\frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{b\ t}{2\ \sin^2\phi} \right)\ d\phi \right)\ dt. \tag{10.18}$$

Interchanging the order of integration in (10.18), then using (**??**), gives

$$J_m(a,b) \stackrel{\triangle}{=} J_m(c) = \frac{1}{\pi}\ \int_0^{\pi/2} \left(\frac{\sin^2\phi}{\sin^2\phi + c} \right)^m\ d\phi = \frac{1}{\pi}\ I_m(c), \tag{10.19}$$

which is the desired closed-form expression for $I_m(c)$. A similar equivalence can be made between a result derived by Chennakeshu and Anderson [**?**] and the integrals $\int_0^{(M-1)\ \pi/M} \left(\frac{\sin^2\phi}{\sin^2\phi+c} \right)^m\ d\phi$ and $\int_0^{\pi/M} \left(\frac{\sin^2\phi}{\sin^2\phi+c} \right)^m\ d\phi$. Full details on these equivalences can be found in [**?**, Appendix 5A]. The reason for mentioning these equivalences and the resulting closed-form expressions is that they can be used, for example, to simplify calculations involving the performance BPSK and M-PSK with selection diversity over correlated Nakagami-m fading channels [**?**].

Appendix 10.C
Key Result on Multiple Antenna System Capacity

While the idea of using multiple antennas at either the transmitter or the receiver to achieve diversity gains or directional transmission has been around for a long time, the recent surge of interest in dual-antenna-array systems (systems using multiple antennas at both the transmitter and receiver) is mostly due to the following result by Foschini and Gans. They show that with n transmit and n receive antennas and i.i.d. fades at different antenna elements, if the receiver has a perfect estimate of the channel the mutual information grows linearly with n for a given fixed average transmitter power and bandwidth. We provide some insight into the mathematical basis of this result.

Since the transmitter does not know the channel, we assume that equal power is transmitted from each transmit antenna. The mutual information of the n-transmit, n-receive antenna system with equal power allocation is:

$$I_n = \log\det[I_{n\times n} + \frac{P}{n} HH^\dagger],$$

where the total transmit power is P. If we denote the eigenvalues of $HH^\dagger$ as $\lambda_i, 1 \le i \le n$ we can express this as:

$$I_n = \sum_{i=1}^{n} \log(1 + \frac{P}{n}\lambda_i).$$

Now comes the really interesting result from theory of large random matrices that says that the eigenvalues of the random matrix $HH^\dagger$ grow linearly in n, asymptotically as $n \to \infty$. This is true for any distribution of the entries H_{ij}, as long as the entries are i.i.d. with unit variance. Even more interestingly if λ_{max} is the largest eigenvalue of $HH^\dagger$, then the following statements are true with probability one,

$$\lim_{n\to\infty} \frac{\lambda_{max}}{n} = 4,$$

and the random empirical distribution of the scaled eigenvalues ($\frac{\lambda_i}{n}$) converges to the following deterministic density:

$$g(\lambda) = \frac{1}{\pi}\sqrt{\frac{1}{\lambda} - \frac{1}{4}} \quad \text{for} \quad 0 \leq \lambda \leq 4 \text{ and } 0 \text{ otherwise.}$$

The asymptotic behavior of the mutual information I_n follows directly from this result:

$$\frac{I_n}{n} = \frac{1}{n}\sum_{i=1}^{n} \log(1 + P\frac{\lambda_i}{n}) \rightarrow \int_0^4 \log(1 + P\lambda)g(\lambda)d\lambda.$$

Thus the mutual information scales linearly with n. Beyond its theoretical beauty this result is exciting since the linear growth predicted by the asymptotic analysis is observed even for reasonably small number of antennas. Also it was shown recently that even for correlated fades between antenna elements the capacity growth rate is still linear in n, albeit smaller than under independent fading.

Recent work in [**?**, 12, **?**] indicates that substantial capacity improvements can be made on MIMO systems even with just channel correlation information available at the transmitter (this is not true for SISO systems). Moreover, results in [**?**] indicate that in some scenarios a beamforming transmission strategy achieves close to channel capacity. This is interesting since beamforming corresponds to scalar coding with linear preprocessing at the transmit antenna array. Thus, the complexity involved is only a fraction of the vector coding complexity for typical array sizes. These results are quite new and have not yet been translated to practical transmission strategies for MIMO systems. However, these results suggest that the capacity enhancement promised by MIMO systems can be achieved in real systems with techniques of reasonable complexity. Practical transmission strategies for MIMO channels generally fall into two categories: space-time coding and space-time signal processing. In space-time coding the codes are designed to take advantage of the extra degrees of freedom in the space domain [**?**, **?**]. Space-time processing focuses on estimation, equalization, and filtering techniques to accurately estimate a signal transmitted over a MIMO channel [13, 14].

Bibliography

[1] M. Simon and M.-S. Alouini, *Digital Communication over Fading Channels A Unified Approach to Performance Analysis*. Wiley, 2000.

[2] W. Lee, *Mobile Communications Engineering*. New York: McGraw-Hill, 1982.

[3] J. Winters, "Signal acquisition and tracking with adaptive arrays in the digital mobile radio system is-54 with flat fading," *IEEE Trans. Vehic. Technol.*, vol. 43, pp. 1740–1751, Nov. 1993.

[4] G. L. Stuber, *Principles of Mobile Communications, 2nd Ed.* Kluwer Academic Publishers, 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. 1887–1895, Dec. 1979.

[6] A. Abu-Dayya and N. Beaulieu, "Switched diversity on microcellular ricean channels," *IEEE Trans. Vehic. Technol.*, pp. 970–976, Nov. 1994.

[7] A. Abu-Dayya and N. Beaulieu, "Analysis of switched diversity systems on generalized-fading channels," *IEEE Trans. Commun.*, pp. 2959–2966, Nov. 1994.

[8] M. Yacoub, *Principles of Mobile Radio Engineering*. CRC Press, 1993.

[9] M. K. Simon and M. -S. Alouini, "A unified approach to the performance analysis of digital communications over generalized fading channels," *Proc. IEEE*, vol. 86, pp. 1860–1877, September 1998.

[10] 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.*, vol. COM-46, pp. 1625–1638, December 1998.

[11] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, pp. 1451–1458, Oct. 1998.

[12] E. Visotsky and U. Madhow, "Space-time transmit precoding with imperfect feedback," *Proc. Intl. Symp. Inform. Theory*, pp. 357–366, June 2000.

[13] A. Paulraj, "Space-time modems for wireless personal communications," *IEEE Pers. Commun. Mag.*, vol. 5, pp. 36–48, Feb. 1998.

[14] R. Kohno, "Spatial and temporal communication theory using adaptive antenna array," *IEEE Pers. Commun. Mag.*, vol. 5, pp. 36–48, Feb. 1998.

Chapter 11

Equalization

We have seen in Chapter 5 that delay spread causes intersymbol interference (ISI), which in turn produces an irreducible error floor in most digital modulation techniques. There are several techniques we can use as countermeasures to delay spread. These techniques fall in two broad categories: signal processing and antenna solutions. In a broad sense, equalization defines any signal processing technique used at the receiver to alleviate the ISI problem caused by delay spread. Signal processing can also be used at the transmitter to make the signal less susceptible to delay spread: spread spectrum and multicarrier modulation fall in this category of transmitter signal processing techniques. In this chapter we focus on equalization. Multicarrier modulation and spread spectrum are the topics of Chapters 11 and 12, respectively. Antenna solutions can also be used to change the propagation environment such that delay spread is reduced or eliminated: techniques that fall in this category include distributed antennas, directive antennas, and adaptive antennas.

An irreducible error floor arises when the channel symbol time T_s is not much larger than the average or rms delay spread (μ_{T_m} or σ_{T_m}). For example, cordless phones typically operate indoors, where the delay spread is small. Since voice is also a relative low-rate application, equalization is generally not used in cordless phones. However, in digital cellular systems which operate outdoors, $\sigma_{T_m} \approx T_s$, so equalization is typically used. Clearly higher data rate applications are even more sensitive to delay spread, and generally require high-performance equalizers. In fact, mitigating the impact of delay spread is the most challenging hurdle for high-speed wireless data systems.

The goal of equalization is to mitigate the effects of ISI. However, this goal must be balanced so that in the process of removing ISI, the noise power in the received signal is not enhanced. A simple example, shown in Figure 11.1, illustrates the pitfalls of removing ISI without considering this effect on noise. Consider a signal $s(t)$ that is passed through a channel with frequency response $H(f)$. At the receiver front end white Gaussian noise $n(t)$ is added to the signal, so the signal input to the receiver is $W(f) = S(f)H(f)+N(f)$, where $N(f)$ has power spectral density N_0. If the bandwidth of $s(t)$ is B then the noise power within the signal bandwidth of interest is N_0B. Suppose we wish to equalize the received signal so as to completely remove the ISI introduced by the channel. This is easily done by introducing an analog equalizer in the receiver defined by

$$H_{eq}(f) = 1/H(f). \tag{11.1}$$

The receiver signal $W(f)$ after passing through this equalizer becomes $[S(f)H(f) + N(f)]H_{eq}(f) = S(f) + N'(f)$, where $N'(f)$ is colored Gaussian noise with power spectral density $N_0/|H(f)|^2$. Thus, all ISI has been removed from the transmitted signal $S(f)$.

However, if $H(f)$ has a spectral null ($H(f_0) = 0$ for some f_0) at any frequency within the bandwidth

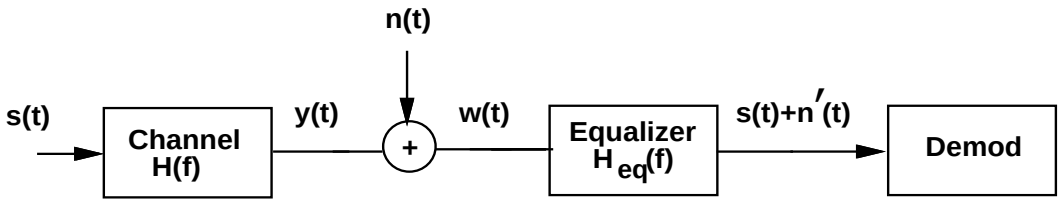


Figure 11.1: Analog Equalizer Illustrating Noise Enhancement.

of $s(t)$, then the power of the noise $N'(f)$ is infinite. Even without a spectral null, if some frequencies in $H(f)$ are greatly attenuated, the equalizer $H_{eq}(f) = 1/H(f)$ will greatly enhance the noise power at those frequencies. In this case even though the ISI effects are removed, the equalized system will perform poorly due to its greatly reduced SNR. Thus, the true goal of equalization is to balance mitigation of the effects of ISI with maximizing the SNR of the post-equalization signal. We will discuss different approaches to achieve this balance in more detail below.

For an equalizer to mitigate the ISI introduced by the channel, it must have an estimate of the channel impulse or frequency response. Since the wireless channel varies over time, the equalizer must learn the frequency response of the channel (training) and then update its estimate of the frequency response as the channel changes (tracking). The process of equalizer training and tracking is often referred to as adaptive equalization, since the equalizer adapts to the changing channel. In general, the training is done by sending a fixed-length known bit sequence over the channel. The equalizer at the receiver uses the known training sequence to adapt its filter coefficients to match the channel frequency response. Specifically, the equalizer filter coefficients are updated to minimize the error between the actual channel output and the channel output resulting from the known training sequence transmitted through the estimate of the channel frequency response. The training process assumes that the channel is relatively constant over the length of the training sequence, otherwise equalization should not be used, since the channel cannot be estimated properly. After training, the equalizer coefficients are matched to the channel, and data can be transmitted. During transmission of user data, an adaptive algorithm is used on the received data to continually update the equalizer coefficients.

If the channel changes relatively slowly, adaptive algorithms are usually sufficient to track the channel changes. However, if the channel is changing quickly, the training sequence may be transmitted periodically to insure that the equalizer coefficients do not drift significantly from their optimal values. It is clearly desirable to avoid periodic retraining, since no useful data is sent during the training interval. When periodic retraining is necessary, the length of the training sequence determines how much bandwidth is wasted on training. The length of the training sequence depends on the equalizer structure and its tap update algorithm, as well as the channel delay spread and coherence time. The convergence rate of several common equalizer algorithms are given below.

An equalizer can be implemented at baseband, RF, or IF. Most equalizers are implemented at baseband using DSP, since such filters are easily tuneable and cheap to implement.

11.1 Equalizer Types

Equalization techniques fall into two broad categories: linear and nonlinear. The linear techniques are generally the simplest to implement and to understand conceptually. However, linear equalization techniques typically suffer from *noise enhancement* on frequency-selective fading channels, and are therefore not used in most wireless applications. Among nonlinear equalization techniques, decision-feedback equal-

ization (DFE) is the most common, since it is fairly simple to implement and does not suffer from noise enhancement. However, on channels with low SNR, the DFE suffers from error propagation when bits are decoded in error, leading to poor performance. The optimal equalization technique to use is maximum likelihood sequence estimation (MLSE). Unfortunately, the complexity of this technique grows exponentially with memory length, and is therefore impractical on most channels of interest. However, the performance of the MLSE is often used as an upper bound on performance for other equalization techniques. Figure 11.2 summarizes the different equalizer types, along with their corresponding structures and tap updating algorithms, which are discussed in more detail in [1].

Equalizers can also be categorized as symbol-by-symbol (SBS) or sequence estimators (SE). SBS equalizers remove ISI from each symbol and then detect each symbol individually. All linear equalizers in Figure 11.2 as well as the DFE are SBS equalizers. SE equalizers detect sequences of symbols, so the effect of ISI is part of the estimation process. Maximum likelihood sequence estimation (MLSE) is the optimal form of sequence detection, but is highly complex.

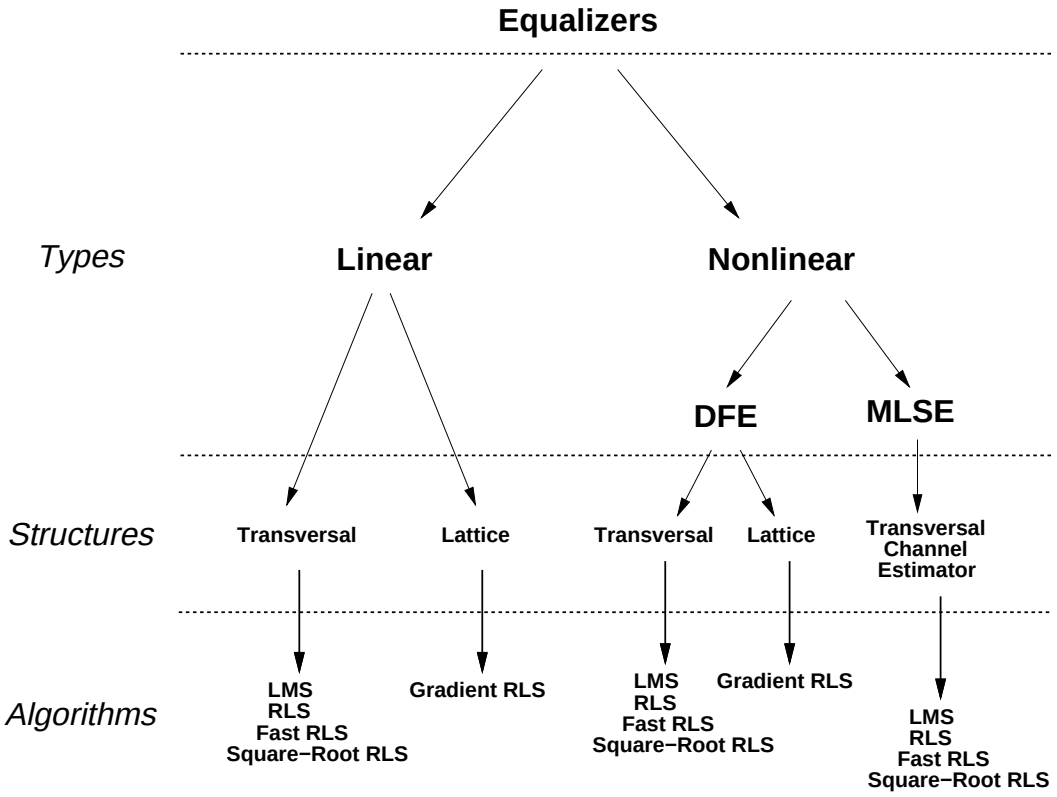


Figure 11.2: Equalizer Types, Structures, and Algorithms.

11.2 Folded Spectrum and ISI-Free Transmission

Figure 11.3 shows a block diagram of an end-to-end system using equalization. The input symbol d_k is passed through a pulse shape filter $p(t)$ which is then transmitted over the ISI channel with impulse response $c(t)$. We define the equivalent channel impulse response $h(t) = p(t) * c(t)$, and the transmitted signal is thus given by $d(t) * p(t) * c(t)$ for $d(t) = \sum_k d_k \delta(t - kT)$ the train of information symbols. The

pulse shape $p(t)$ improves the spectral properties of the transmitted signal, as described in Chapter 5.5. This pulse shape is under the control of the system designer, whereas the channel $c(t)$ is introduced by nature and is outside the designer's control.

At the receiver front end white Gaussian noise $n(t)$ is added to the received signal for a resulting signal $w(t)$. The first operation on the received signal $w(t)$ is to pass the signal through an analog matched filter $g^*(-t)$. The purpose of the matched filter is to maximize the SNR of the signal before sampling and subsequent processing[1]. Recall from Chapter 5.1 that in AWGN the SNR of the received signal is maximized using a matched filter that is matched to the pulse shape. This result also indicates that for the system shown in Figure 11.3, SNR is maximized by passing $w(t)$ through a filter matched to $h(t)$, so ideally we would have $g(t) = h(t)$. However, since the channel impulse response $c(t)$ is time-varying and sometimes unknown, and analog filters are not easily tuneable, it is not possible to have $g^*(t) = h^*(-t)$. Thus, part of the art of equalizer design is to chose $g^*(t)$ to get good performance. The fact that $g^*(t)$ cannot be matched to $h(t)$ can result in significant performance degradation and also makes the receiver extremely sensitive to timing error. These problems are somewhat mitigated by oversampling $w(t)$ at a rate much faster than the symbol rate: this process is called fractionally-spaced equalization [1]

The output of the matched filter is sampled and then passed through a digital equalizer. Digital implementation of the equalizer is highly desirable, since digital filters are cheap, easy to implement, and easily tuneable to adjust for changing channel conditions. Let $f(t)$ denote the combined baseband impulse response of the transmitter, channel, and matched filter:

$$f(t) = p(t) * c(t) * g^*(-t). \tag{11.2}$$

Then the matched filter output is given by

$$y(t) = d(t) * f(t) + n_g(t) = \sum d_k f(t - kT) + n_g(t), \tag{11.3}$$

where $n_g(t) = n(t) * g^*(-t)$ is the equivalent baseband noise at the equalizer input and T is the symbol time. If we sample $y(t)$ every T seconds we obtain $y_n = y(nT)$ as

$$\begin{aligned} y_n &= \sum_{k=-\infty}^{\infty} d_k f(nT - kT) + n_g(nT) \\ &\triangleq \sum_{k=-\infty}^{\infty} d_k f_{n-k} + \nu_n \\ &= d_n f_0 + \sum_{k \neq n} d_k f_{n-k} + \nu_n. \end{aligned} \tag{11.4}$$

where the first term in (11.4) is the desired data bit, the second term is the ISI, and the third term is the sampled baseband noise. We see from (11.4) that we get zero ISI if $f_{n-k} = 0$ for $k \neq n$, i.e. $f_k = \delta_k f_0$. In this case (11.4) reduces to $y_n = d_n f_0 + \nu_n$.

We now show that the condition for ISI-free transmission, $f_k = \delta_k f_0$, is satisfied if and only if

$$F_\Sigma(f) \triangleq \frac{1}{T} \sum_{n=-\infty}^{\infty} F(f + \frac{n}{T}) = f_0. \tag{11.5}$$

The function $F_\Sigma(f)$ is often called the folded spectrum, and $F_\Sigma(f) = f_0$ implies that the folded spectrum is flat.

[1]While the matched filter could be more efficiently implemented digitally, the analog implementation before the sampler allows for a smaller dynamic range in the sampler, which significantly reduces cost.

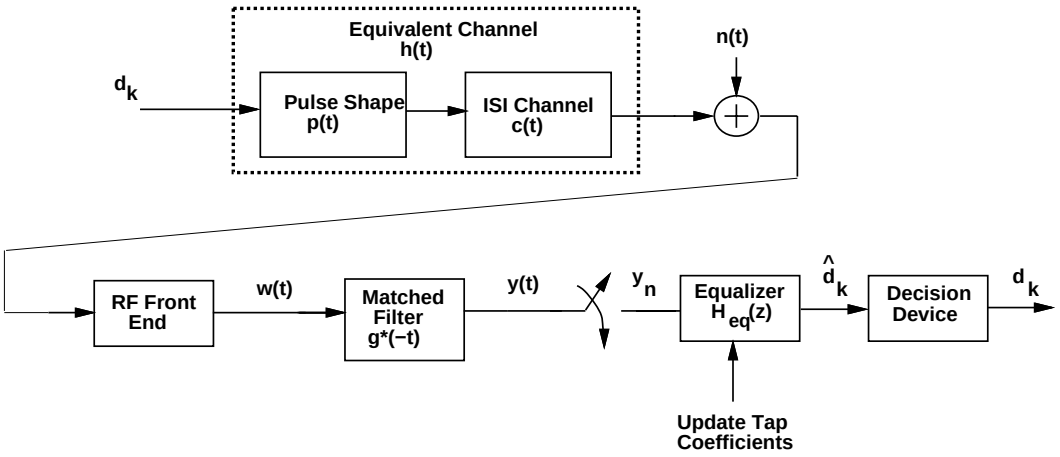


Figure 11.3: End-to-End System.

To show this equivalence, first note that

$$\begin{aligned} f_k &= \int_{-\infty}^{\infty} F(f)e^{j2\pi kfT}df \\ &= \sum_{n=-\infty}^{\infty} \int_{(2n-1)/2T}^{(2n+1)/2T} F(f)e^{j2\pi kfT}df \\ &= \sum_{n=-\infty}^{\infty} \int_{-1/2T}^{1/2T} F\left(f' + \frac{n}{T}\right) e^{j2\pi k(f'+n/T)T}df' \\ &= \int_{-1/2T}^{1/2T} e^{j2\pi kfT}\left[\sum_{n=-\infty}^{\infty} F\left(f + \frac{n}{T}\right)\right] df. \end{aligned} \tag{11.6}$$

We first show that a flat folded spectrum implies that $f_k = \delta_k f_0$. Suppose (11.5) is true. Then by (11.6),

$$f_k = T\int_{-1/2T}^{1/2T} e^{-j2\pi kfT} f_0 T df = \frac{\sin \pi k}{\pi k} f_0 = \delta_k f_0, \tag{11.7}$$

which is the desired result. We now show that $f_k = \delta_k f_0$ implies a flat folded spectrum. If $f_k = \delta_k f_0$ then by (11.6),

$$f_k = T\int_{-1/2T}^{1/2T} F_\Sigma(f) e^{j2\pi kfT} df. \tag{11.8}$$

So f_k is the Fourier transform of $F_\Sigma(f)$. Therefore, if $f_k = \delta_k f_0$, $F_\Sigma(f) = f_0$.

11.3 Linear Equalizers

If $F_\Sigma(f)$ is not flat, we can use the equalizer $H_{eq}(z)$ in Fig. 11.3 to reduce ISI. In this section we assume a linear equalizer:

$$H_{eq}(z) = w_0 + w_1 z^{-1} + \ldots w_N z^{-N}. \tag{11.9}$$

The length of the equalizer N is typically dictated by implementation considerations, since a large N usually entails higher complexity. For a given equalizer size N the only task for the equalizer design is

to determine the equalizer coefficients $\{w_i\}_{i=0}^N$. Recall that our performance metric in wireless systems is probability of error (or outage probability), so the optimal choice of equalizer coefficients would be the coefficients that minimize probability of error. Unfortunately it is extremely difficult to optimize the $\{w_i\}$s with respect to this criterion. Since we cannot directly optimize for our desired performance metric, we must instead use an indirect optimization that balances ISI mitigation with the prevention of noise enhancement, as discussed relative to the simple analog example above. We now describe two linear equalizers: the Zero Forcing (ZF) equalizer and the Minimum Mean Square Error (MMSE) equalizer. The former equalizer cancels all ISI, but can lead to considerable noise enhancement. The latter technique minimizes the expected mean squared error between the transmitted symbol and the symbol detected at the equalizer output, thereby providing a better balance between ISI mitigation and noise enhancement. Because of this more favorable balance, MMSE equalizers tend to have better BER performance than systems using the ZF algorithm.

11.3.1 Zero Forcing (ZF) Equalizers

The samples $\{y_n\}$ input to the equalizer can be represented based on the discretized combined system response $f(t) = h(t) * g^*(-t)$ as

$$Y(z) = D(z)F(z) + N(z), \tag{11.10}$$

where $N(z)$ is the power spectrum of the white noise after passing through the matched filter $G^*(1/z^*)$ and the equalizer $H_{eq}(z)$ and

$$F(z) = H(z)G^*(1/z^*) = \sum_n f(nT)z^{-n}. \tag{11.11}$$

The zero-forcing equalizer removes all ISI introduced in the combined response $f(t)$. From (11.10) we see that the equalizer to accomplish this is given by

$$H_{ZF}(z) = \frac{1}{F(z)}. \tag{11.12}$$

This is the discrete-time equivalent to the analog equalizer (11.1) described above, and it suffers from the same noise enhancement properties. Specifically, the power spectrum $N(z)$ is given by

$$N(z) = N_g(z)|H_{ZF}(z)|^2 = \frac{N_0|G^*(1/z^*)|^2}{|F(z)|^2} = \frac{N_0|G^*(1/z^*)|^2}{|H(z)|^2|G^*(1/z^*)|^2} = \frac{N_0}{|H(z)|^2}. \tag{11.13}$$

We see from (11.13) that if the channel $H(z)$ is sharply attenuated at any frequency within the bandwidth of interest, as is common on frequency-selective fading channels, the noise power will be significantly increased. This motivates an equalizer design that better optimizes between ISI mitigation and noise enhancement. One such equalizer is the MMSE equalizer, which we describe in the next section.

The ZF equalizer defined by $H_{ZF}(z) = 1/F(z)$ may not be implementable as a finite impulse response (FIR) filter. Specifically, it may not be possible to find a finite set of coefficients $w_0, \ldots, w_N$ such that

$$w_0 + w_1 z^{-1} + \ldots + w_N z^{-N} = \frac{1}{F(z)}. \tag{11.14}$$

In this case we find the set of coefficients $\{w_i\}$ that best approximates the zero-forcing equalizer. Note that this is not straightforward since the approximation must be valid for all values of z. There are many ways we can make this approximation. One technique is to represent $H_{ZF}(z)$ as an infinite impulse response (IIR) filter, $1/F(z) = \sum_{i=-\infty}^{\infty} c_i z^{-i}$ and then set $w_i = c_i$. Another technique is to take $w_i = c_i$ where $\{c_i\}$ is the inverse z-transform of $1/F(z)$ (it can be shown that this minimizes the L_2 norm of $\frac{1}{F(z)} - (w_0 + w_1 z^{-1} + \ldots + w_N z^{-N})$ at $z = e^{j\omega}$). Other approximations are also used in practice.

11.3.2 Minimum Mean Square Error (MMSE) Equalizer

In MMSE equalization the goal of the equalizer design is to minimize the average mean square error (MSE) between the transmitted symbol d_k and its estimate $\hat{d}_k$ at the output of the equalizer, i.e we want to find the $\{w_i\}$s to minimize $\mathrm{E}[d_k - \hat{d}_k]^2$. Since we are dealing with linear equalizers, the equalizer output $\hat{d}_k$ is a linear combination of the input samples y_k:

$$\hat{d}_k = \sum_{i=0}^{N} w_i y_{k-i}. \tag{11.15}$$

As such, finding the optimal filter coefficients $\{w_i\}$ becomes a standard problem in linear estimation. In fact, if the noise input to the equalizer is white, this is a standard Weiner filtering problem. The problem is that because of the matched filter $g^*(-t)$ at the receiver front end, the noise input to the equalizer is not white but colored with power spectrum $N_0|G^*(1/z^*)|^2$. Therefore, in order to apply known techniques for optimal linear estimation, we expand the filter $H_{eq}(z)$ into two components, a noise whitening component $1/G^*(1/z^*)$ and an ISI removal component $\hat{H}_{eq}(z)$, as shown in Figure 11.4.

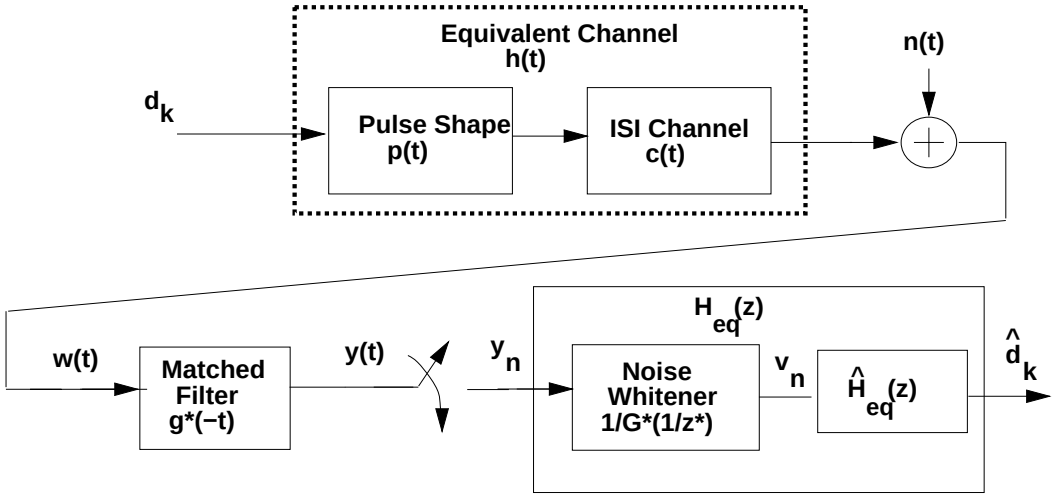


Figure 11.4: MMSE Equalizer with Noise Whitening Filter.

The purpose of the noise whitening filter, as indicated by the name, is to whiten the noise such that the noise component output from this filter has a constant power spectrum. Since the noise input to this receiver has power spectrum $N_0|G^*(1/z^*)|^2$, the appropriate noise whitening filter is $1/G^*(1/z^*)$ so that the noise power spectrum at the output of the noise whitening filter is $N_0|G^*(1/z^*)|^2/|G^*(1/z^*)|^2 = N_0$. Note that the filter $1/G^*(1/z^*)$ is not the only filter that will whiten the noise, and another noise whitening filter with more desirable properties (like stability) may be chosen. It might seem odd at first to introduce the matched filter $g^*(-t)$ at the receiver front end only to cancel its effect in the equalizer. Recall, however, that the matched filter is meant to maximize the SNR at the A/D input. By removing the effect of this matched filter through noise whitening we merely simplify the design of $\hat{H}_{eq}(z)$ to minimize MSE. In fact if the noise whitening filter does not yield optimal performance then its effect would be cancelled by the $\hat{H}_{eq}(z)$ filter design, as we will see below in the case of IIR MMSE equalizers.

We assume the filter $\hat{H}_{eq}(z)$, with input v_n, is a linear filter with N taps:

$$\hat{H}_{eq}(z) = w_0 + w_1 z^{-1} + \ldots + w_N z^{-N}. \tag{11.16}$$

Our goal is to design the filter coefficients $\{w_i\}$ so as to minimize $\mathrm{E}[d_k - \hat{d}_k]^2$. This is the same goal as for the total filter $H_{eq}(z)$, we've just added the noise whitening filter to make solving for these coefficients simpler. We define the following column vectors $\mathbf{v} = (\mathbf{v_k}, \mathbf{v_{k-1}}, \ldots, \mathbf{v_{k-N}})$ and $\mathbf{w} = (\mathbf{w_0}, \ldots, \mathbf{w_N})$. Then the output of the filter $\hat{H}_{eq}(z)$ is

$$\hat{d}_k = \mathbf{w^T v} = \mathbf{v^T w}. \tag{11.17}$$

Thus, we want to minimize the mean square error

$$J = \mathrm{E}[d_n - \hat{d}_n]^2 = \mathrm{E}\left[\mathbf{w^T v v^H w^*} - \mathbf{2}\Re\{\mathbf{v^H w^* d_n}\} + |\mathbf{d_n}|^\mathbf{2}\right], \tag{11.18}$$

where for a vector $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x_N}]$, $\mathbf{x^*} = [\mathbf{x_1^*}, \ldots, \mathbf{x_N^*}]$, and $\mathbf{x^H} = [\mathbf{x_1^*}, \ldots, \mathbf{x_N^*}]^\mathbf{T}$. Define $M_v = \mathrm{E}[\mathbf{v v^H}]$ and $\mathbf{v_d} = \mathbf{E}[(\mathbf{v^H d_n}]$. The matrix M_v is an $N \times N$ Hermitian matrix and $\mathbf{v_d}$ is a length N row vector. Assume $\mathrm{E}|d_k|^2 = 1$. Then the MSE J is

$$J = \mathbf{w^T M_v w^*} - \mathbf{2}\Re\{\mathbf{v_d w^*}\} + \mathbf{1}. \tag{11.19}$$

We obtain the optimal tap vector $\mathbf{w}$ by setting the gradient $\bigtriangledown_\mathbf{w} J = 0$ and solving for $\mathbf{w}$. From (11.19) it can be shown that [5, 3]

$$\bigtriangledown_\mathbf{w} J = \left(\frac{\partial J}{\partial w_0}, \ldots, \frac{\partial J}{\partial w_N}\right) = 2\mathbf{w^T M_v} - \mathbf{2 v_d}. \tag{11.20}$$

Setting this to zero yields $\mathbf{w^T M_v} = \mathbf{v_d}$ or, equivalently, that the optimal tap weights are given by

$$\mathbf{w_{opt}} = \left(\mathbf{M_v^T}\right)^{-\mathbf{1}} \mathbf{v_d^H}. \tag{11.21}$$

Note that solving for $\mathbf{w_{opt}}$ requires a matrix inversion with respect to the filter inputs. Thus, the complexity of this computation is quite high, typically on the order of N^2 to N^3 operations. Substituting in these optimal tap weights we obtain the minimum mean square error as

$$J_{min} = 1 - \mathbf{v_d M_v^{-1} v_d^H}. \tag{11.22}$$

For an infinite length equalizer, $\mathbf{v} = (\mathbf{v_{n+\infty}}, \ldots, \mathbf{v_n}, \mathbf{v_{n-\infty}})$ and $\mathbf{w} = (\mathbf{w_{-\infty}}, \ldots, \mathbf{w_0}, \ldots, \mathbf{w_\infty})$. Then $\mathbf{w^T M_v} = \mathbf{v_d}$ can be written as

$$\sum_{i=-\infty}^{\infty} w_i (f_{j-i} + N_0)\delta_{ij} = g_{-j}^*, \quad -\infty \le j \le \infty. \tag{11.23}$$

Taking z transforms and noting that $\hat{H}_{eq}(z)$ is the z transform of the filter coefficients $\mathbf{w}$ yields

$$\hat{H}_{eq}(z)(F(z) + N_0) = G^*(1/z^*). \tag{11.24}$$

Solving for $\hat{H}_{eq}(z)$ yields

$$\hat{H}_{eq}(z) = \frac{G^*(1/z^*)}{F(z) + N_0}. \tag{11.25}$$

Since the MMSE equalizer consists of the noise whitening filter $1/G^*(1/z^*)$ plus the ISI removal component $\hat{H}_{eq}(z)$, we get that the full MMSE equalizer, when it is not restricted to be finite length, becomes

$$H_{eq}(z) = \frac{\hat{H}_{eq}(z)}{G^*(1/z^*)} = \frac{1}{F(z) + N_0}. \tag{11.26}$$

There are three interesting things to notice about this result. First of all, the ideal infinite length MMSE equalizer cancels out the noise whitening filter. Second, this infinite length equalizer is identical to the ZF filter except for the noise term N_0, so in the absence of noise the two equalizers are equivalent. Finally, this ideal equalizer design clearly shows a balance between inverting the channel and noise enhancement: if $F(z)$ is highly attenuated at some frequency the noise term N_0 in the denominator prevents the noise from being significantly enhanced by the equalizer. Yet at frequencies where the noise power spectral density N_0 is small compared to the composite channel $F(z)$, the equalizer effectively inverts $F(z)$.

For the equalizer (11.26) it can be shown [3] that the minimum MSE (11.22) can be expressed in terms of the folded spectrum $F_\Sigma(f)$ as

$$J_{min} = T \int_{-.5/T}^{.5/T} \frac{N_0}{F_\Sigma(f) + N_0} df. \tag{11.27}$$

This expression for MMSE has several interesting properties. First it is readily seen, as expected, that $0 \leq J_{min} = \mathrm{E}[d_k - \hat{d}_k]^2 \leq 1$. In addition, $J_{min} = 0$ in the absence of noise ($N_0 = 0$) as long as $F_\Sigma(f) \neq 0$ within the signal bandwidth of interest. Also, as expected, $J_{min} = 1$ if $N_0 = \infty$.

11.4 Maximum Likelihood Sequence Estimation

Maximum-likelihood sequence estimation (MLSE) avoids the problem of noise enhancement since it doesn't use an equalizing filter: instead it estimates the sequence of transmitted symbols [4]. The structure of the MLSE is the same as in Figure 11.3 except that the equalizer $H_{\mathrm{eq}}(z)$ and decision device are replaced by the MLSE algorithm. Given the channel response $h(t)$, the MLSE algorithm chooses the input sequence $\{d_k\}$ that maximizes the likelihood of the received signal $w(t)$. We now investigate this algorithm in more detail.

Using a Gram-Schmidt orthonormalization procedure we can express $w(t)$ on a time interval $[0, LT]$ as

$$w(t) = \sum_{n=1}^{N} w_n \phi_n(t), \tag{11.28}$$

where $\{\phi_n(t)\}$ form a complete set of orthonormal basis functions. The number N of functions in this set is a function of the channel memory, since $w(t)$ on $[0, LT]$ depends on $d_0, \ldots, d_L$. With this expansion we have

$$w_n = \sum_{k=-\infty}^{\infty} d_k h_{nk} + n_n = \sum_{k=0}^{L} d_k h_{nk} + n_n, \tag{11.29}$$

where

$$h_{nk} = \int_0^T h(t - kT) \phi_n^*(t) dt \tag{11.30}$$

and

$$n_n = \int_0^T n(t) \phi_n^*(t) dt. \tag{11.31}$$

The n_n are complex Gaussian random variables with mean zero and covariance $.5E[n_n^* n_m] \times N_0 \delta(n - m)$. Thus, $W^N = (w_1, \ldots, w_N)$ has a multivariate Gaussian distribution

$$p(W^N | d^L, h(t)) = \Pi_{n=1}^{N} \frac{1}{\pi N_o} \exp\left[-\frac{1}{N_0} \left| w_n - \sum_{k=0}^{L} d_k h_{nk} \right|^2 \right]. \tag{11.32}$$

Given a received signal $w(t)$ or, equivalently, W^N, the MLSE decodes this as the symbol sequence d^L that maximizes the likelihood function $p(W^N|d^L, h(t))$ (or the log of this function) that W^N is received given that d^L was transmitted. That is, the MLSE outputs the sequence

$$\begin{aligned}
\hat{d}^L &= \arg\max \left[\log p(W^N|d^L, h(t))\right] \\
&= \arg\max \left[-\sum_{n=1}^{N} |w_n - \sum_k d_k h_{nk}|^2\right] \\
&= \arg\max \left[-\sum_{n=1}^{N} |w_n|^2 + \sum_{n=1}^{N}\left(w_n^* \sum_k d_k h_{nk} + w_n \sum_k d_k^* h_{nk}^*\right) - \sum_{n=1}^{N}\left(\sum_k d_k h_{nk}\right)\left(\sum_m d_m^* h_{nm}^*\right)\right] \\
&= \arg\max \left[2\Re\left\{\sum_k d_k^* \sum_{n=1}^{N} w_n h_{nk}^*\right\} - \sum_k \sum_m d_k d_m^* \sum_{n=1}^{N} h_{nk} h_{nm}^*\right]. \qquad (11.33)
\end{aligned}$$

Note that

$$\sum_{n=1}^{N} w_n h_{nk}^* = \int_{-\infty}^{\infty} w(\tau) h^*(\tau - nT) d\tau = y_n, \qquad (11.34)$$

and

$$\sum_{n=1}^{N} h_{nk} h_{nm}^* = \int_{-\infty}^{\infty} h(\tau - kT) h^*(\tau - mT) d\tau = f_{km}. \qquad (11.35)$$

Combining (11.33), (11.34), and (11.35) we have that

$$\hat{d}^L = \arg\max \left[2\Re\left\{\sum_k d_k^* y_k\right\} - \sum_k \sum_m d_k d_m^* f_{km}\right]. \qquad (11.36)$$

We see from this equation that the MLSE output depends only on the sampler output $\{y_k\}$ and the channel parameters $f_{nk} = f(nT - kT)$ where $f(t) = h(t) * h^*(-t)$. Since the derivation of the MLSE is based on the channel output only, our derivation implies that the receiver matched filter in Figure 11.1 is optimal for MLSE detection (typically the matched filter is optimal for detecting signals in AWGN, but this derivation shows that it is also optimal for detecting signals in the presence of ISI if MLSE is used).

The MLSE algorithm is also used in ML decoding, and the Viterbi algorithm can be used for MLSE. However, the complexity of this equalization technique grows exponentially with the channel delay spread. A nonlinear technique with significantly less complexity is the decision-feedback decoder, or DFE.

11.5 Decision-Feedback Equalization

The DFE consists of a feedforward filter with the received sequence as input (similar to the linear equalizer) followed by a feedback filter with the previously detected sequence as input. The DFE structure is shown in Figure 11.5. Effectively, the DFE determines the ISI contribution from the detected symbols $\{d_n\}$ by passing them through the feedback filter. The resulting ISI is then subtracted from the incoming symbols. Since the feedback filter $D(z)$ in Figure 11.5 sits in a feedback loop, it must be strictly causal, or else the system is unstable. The feedback filter of the DFE does not suffer from noise enhancement because it estimates the channel frequency response rather than its inverse. For channels with deep spectral nulls, DFEs generally perform much better than linear equalizers.

DFEs exhibit feedback errors if $\hat{\hat{d}}_n \neq d_n$, since the ISI subtracted by the feedback path is not the true ISI corresponding to d_n. This error therefore propagates to later bit decisions. Moreover, this error

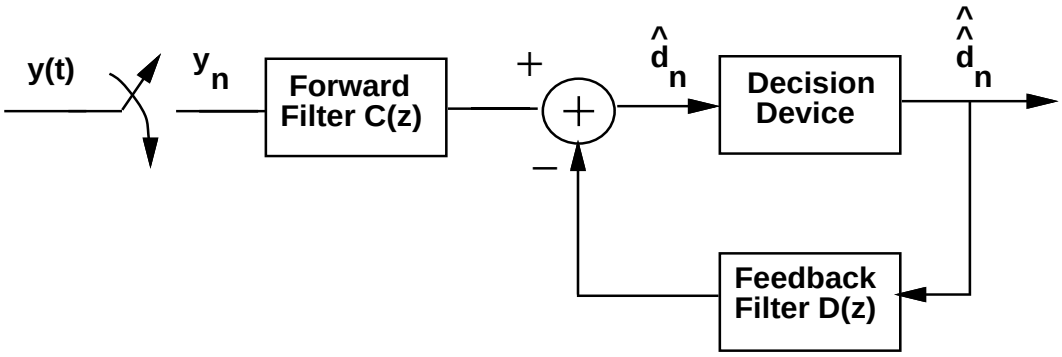


Figure 11.5: Decision-Feedback Equalizer Structure.

propagation cannot be improved through channel coding, since the feedback path operates on coded channel symbols before decoding. That is because the ISI must be subtracted immediately, which doesn't allow for any decoding delay. The error propagation therefore seriously degrades performance on channels with low SNR. DFEs use versions of the LMS and RLS algorithms to update the filter coefficients of $C(z)$ and $D(z)$. Details of these algorithms and their relative performance for DFEs can be found in [1].

11.6 Equalizer Training and Tracking

All of the equalizers described so far are designed based on a known value of the composite channel response $h(t) = p(t) * c(t)$. Since the channel $c(t)$ in generally not known when the receiver is designed, the equalizer must be tunable so it can adjust to different values of $c(t)$. Moreover, since in wireless channels $c(t) = c(\tau, t)$ will change over time, the system must periodically estimate the channel $c(t)$ and update the equalizer coefficients accordingly. This process is called equalizer training.

During training the N coefficients of the equalizer are updated at time k based on a known training sequence $[d_{k-M}, \ldots, d_k]$ that has been sent over the channel. The length M of the training sequence depends on the number of equalizer coefficients that must be determined and the convergence speed of the training algorithm. Note that the equalizer must be retrained when the channel decorrelates, i.e. at least every T_c seconds where T_c is the channel coherence time. Thus, if the training algorithm is slow relative to the channel coherence time then the channel may change before the equalizer can learn the channel. Specifically, if $MT > T_c$ then the channel will decorrelate before the equalizer has finished training. In this case equalization is not an effective countermeasure for ISI, and some other technique (e.g. multicarrier modulation or CDMA) is needed.

Let $\{\hat{d}_k\}$ denote the bit decisions output from the equalizer given a transmitted training sequence $\{d_k\}$. Our goal is to update the N equalizer coefficients at time $k+1$ based on the training sequence we have received up to time k. We will denote these updated coefficients as $\{w_0(k+1), \ldots, w_N(k+1)\}$. We will use the MMSE as our criterion to update these coefficients, i.e. we will chose $\{w_0(k+1), \ldots, w_N(k+1)\}$ as the coefficients that minimize the MSE between d_k and $\hat{d}_k$. Recall that $\hat{d}_k = w_0(k)y_k + w_1(k)y_{k-1} + \ldots + w_N(k)y_{k-N}$, where y_k is the output of the sampler in Figure 11.1 at time k with the known training sequence as input. The $\{w_0(k+1), \ldots, w_N(k+1)\}$ that minimize MSE are obtained via a Weiner filter [5, 3]. Specifically,

$$\mathbf{w(k+1)} = \{\mathbf{w_0(k+1)}, \ldots, \mathbf{w_N(k+1)}\} = \mathbf{R^{-1}p}, \tag{11.37}$$

where $\mathbf{p} = \mathbf{d_k}[\mathbf{y_k y_{k-1}} \cdots \mathbf{y_{k-N}}]^{\mathbf{T}}$ and

$$R = \begin{bmatrix} |y_k|^2 & y_k y_{k-1}^* & \cdots & y_k y_{k-N}^* \\ y_{k-1} y_k^* & |y_{k-1}|^2 & \cdots & y_{k-1} y_{k-N}^* \\ \vdots & \ddots & \ddots & \vdots \\ y_{k-N} y_k^* & \cdots & \cdots & |y_{k-N}|^2 \end{bmatrix}. \tag{11.38}$$

Note that the optimal tap updates in this case requires a matrix inversion, which requires $N^2 - N^3$ multiply operations on each iteration (each bit time T). However, the convergence of this algorithm is very fast: it often converges in around N bit times for N the number of equalizer tap weights.

If complexity is an issue then the large number of multiply operations needed to do MMSE training can be prohibitive. A simpler technique is the least mean square (LMS) algorithm [5]. In this algorithm the tap weight vector $\mathbf{w(k+1)}$ is updated linearly as

$$\mathbf{w(k+1)} = \mathbf{w(k)} + \mathbf{\Delta} \epsilon_{\mathbf{k}} [\mathbf{y_k^* \cdots y_{k-N}^*}], \tag{11.39}$$

where $\epsilon_k = d_k - \hat{d}_k$ is the error between the bit decisions and the training sequence and Δ is the step size of the algorithm, which is a parameter that can be chosen. The choice of Δ dictates the convergence speed and stability of the algorithm. For small values of Δ convergence is very slow, at it takes many more than N bits for the algorithm to converge to the proper equalizer coefficients. However, if Δ is chosen to be large then the algorithm can go unstable, basically skipping over the desired tap weights at every iteration. Thus, for good performance of the LMS algorithm Δ is typically small and convergence is typically slow. However, the LMS algorithm exhibits significantly reduced complexity compared to the MMSE algorithm since the tap updates only require approximately $2N + 1$ multiply operations per iteration. Thus, the complexity is linear in the number of tap weights. Other algorithms, such as the root-least-squares (RLS), Square-root-least-squares, and Fast Kalman provide various tradeoffs in terms of complexity and performance that lie between the two extremes of the LMS algorithm (slow convergence but low complexity) and the MMSE algorithm (fast convergence but very high complexity). A description of these other algorithms is given in [1]. The table below summarizes the specific number of multiply operations and the relative convergence rate of all these algorithms.

Note that the bit decisions $\hat{d}_k$ output from the equalizer are typically passed through a threshold detector to round the decision to the nearest bit value[2]. The resulting roundoff error can be used to adjust the equalizer coefficients during data transmission. This is called equalizer tracking. Tracking is based on the premise that if the roundoff error is nonzero then the equalizer is not perfectly trained, and the roundoff error can be used to adjust the channel estimate inherent in the equalizer. The procedure works as follows. The equalizer output bits $\hat{d}_k$ and threshold detector output bits $\hat{\hat{d}}_k$ are used to adjust an estimate of the baseband equivalent composite channel $H(z)$. In particular, the coefficients of $H(z)$ are adjusted to minimize the MSE between $\hat{d}_k$ and $\hat{\hat{d}}_k$, using the same MMSE procedures described earlier in this chapter. The updated version of $H(z)$ is then taken to equal the composite channel and used to update the equalizer coefficients accordingly. More details can be found in [5, 3].

A summary of the training and tracking characteristics for the different algorithms as a function of the number of taps N is given in the following table.

[2] A bit value of zero or one corresponds to binary decisions. For higher level modulations the threshold detector rounds to the nearest constellation point.

Algorithm	# of multiply operations	Complexity	Convergence	Tracking
LMS	$2N+1$	Low	Slow ($>> NT_s$)	Poor
MMSE	N^2-N^3	Very High	Fast ($\approx NT_s$)	Good
RLS	$2.5N^2+4.5N$	High	Fast	Good
Fast Kalman	$20N+5$	Fairly Low	Fast	Good
Square Root RLS	$1.5N^2+6.5N$	High	Fast	Good

Note that the Fast Kalman and Square Root RLS may be unstable in their convergence and tracking, which is the price paid for their fast convergence with relatively low complexity.

Bibliography

[1] J.G. Proakis, "Adaptive equalization for TDMA digital mobile radio," *IEEE Trans. Vehic. Technol.* Vol. 40, No. 2, pp. 333–341, May 1991.

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

[3] G.L. Stüber, *Principles of Mobile Communications*. Kluwer Academic Publishers, 1996.

[4] G.D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory*, Vol. IT-18, pp. 363–378, May 1972.

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

[6] S.U. Qureshi, "Adaptive equalization," *Proc. IEEE*, Vol. 73, pp. 1349–1387, Sept. 1985.

[7] C. A. Belfiore and J. H. Park, Jr., "Decision-feedback equalization," *Proc. IEEE*, Vol. 67, No. 8, pp. 1143–1156, Aug. 1979.

Chapter 12

Multicarrier Modulation

The basic idea of multicarrier modulation is to divide the transmitted bitstream into many different substreams and send these over many different subchannels. Typically the subchannels are orthogonal under ideal propagation conditions, in which case multicarrier modulation is often referred to as orthogonal frequency division multiplexing (OFDM). The data rate on each of the subchannels is much less than the total data rate, and the corresponding subchannel bandwidth is much less than the total system bandwidth. The number of substreams is chosen to insure that each subchannel has a bandwidth less than the coherence bandwidth of the channel, so the subchannels experience relatively flat fading. Thus, the ISI on each subchannel is small. Moreover, in the discrete implementation of OFDM, often called discrete multitone (DMT), the ISI can be completely eliminated through the use of a cyclic prefix. The subchannels in OFDM need not be contiguous, so a large continuous block of spectrum is not needed for high rate multicarrier communications.

Over the past few years, there has been increasing interest in multicarrier modulation for a variety of applications. However, multicarrier modulation is not a new technique: it was first used for military HF radios in the late 1950's and early 1960's. For the last ten years, multicarrier modulation has been used in many applications [1], including Digital Audio Broadcasting in Europe [2], high-speed digital subscriber lines (HDSL) [3], and the most recent generation of wireless LANs (IEEE 802.11a). The multicarrier technique can be implemented in multiple ways and are sometimes called by different names, including frequency division multiplexing (FDM) [4] and vector coding [5], as well as DMT [3] and OFDM [6].

There is some debate as to whether multicarrier modulation is better for ISI channels than single carrier transmission with equalization. It is claimed in [2] that for mobile radio applications, single carrier with equalization has roughly the same performance as multicarrier modulation with channel coding, frequency-domain interleaving, and weighted maximum-likelihood decoding. Adaptive loading was not taken into account in [2], which has the potential to significantly improve multicarrier performance [7]. But there are other problems with multicarrier modulation which impair its performance, most significantly frequency offset and timing jitter, which impair the orthogonality of the subchannels. In addition, the peak-to-average power ratio of multicarrier is significantly higher than that of single carrier systems, which is a serious problem when nonlinear amplifiers are used. The relative performance of multicarrier systems versus single carrier, along with techniques to overcome the performance impairments of multicarrier, are current topics of intense research.

12.1 Orthogonal Frequency Division Multiplexing (OFDM)

The simplest form of multicarrier modulation divides the data stream into multiple substreams to be transmitted over different orthogonal subchannels centered at different subcarrier frequencies. The number of substreams is chosen to make the symbol time on each substream much greater than the delay spread of the channel or, equivalently, to make the substream bandwidth less then the channel coherence bandwidth. This insures that the substreams will not experience significant ISI.

Consider a system with baseband bandwidth B (passband bandwidth $2B$) and a desired data rate R. The coherence bandwidth for the channel is assumed to be $B_c \geq B$. We set N sufficiently large so that the baseband subchannel bandwidth $B_N = B/N << B_c$, which insures relatively flat-fading on each subchannel. The bit stream is divided into N substreams that are linearly-modulated (typically via MQAM or MPSK) relative to the subcarrier frequencies f_n and then transmitted in parallel over the N subchannels. For nonoverlapping channels we set $f_n = f_c + n(2B_N), n = 0, \ldots, N-1$. The multicarrier system with nonoverlapping channels is shown in Figure 12.1. The data rate for each substream is $R_N = R/N$, and the symbol time for each substream is T_N. The transmitted signal over one symbol time T_N is given by

$$s(t) = \mathcal{R}\left\{\sum_{n=0}^{N-1} s_n g(t) e^{j2\pi f_n t}\right\}., \tag{12.1}$$

where s_n is the complex symbol associated with the nth subcarrier. If we assume raised cosine pulses for $g(t)$ we get $T_N = .5(1+\beta)/B_N$, where β is the rolloff factor (e.g. for rectangular pulses, $T_N = .5/B_N$). The substreams are sent in their respective orthogonal subchannels with passband bandwidth $2B_N$, yielding a total passband bandwidth $N2B_N = 2B$ and data rate $NR_N = R$. Thus, this form of multicarrier modulation does not change the data rate or signal bandwidth relative to single-carrier systems, but it almost completely eliminates ISI since the subchannels have $B_N << B_c$ and therefore they experience relatively little frequency-selective fading.

OFDM with nonoverlapping subchannels is basically a form of frequency-division (see Chapter 14.2.1), a technique that allows multiple users to share the same system bandwidth. The advantage of using nonoverlapping subchannels is that small frequency offsets and timing jitter do not have much impact on the orthogonality of the subchannels. However, the frequency division approach is not spectrally efficient. We can improve on the spectral efficiency of OFDM by overlapping the subcarriers. The subcarriers must still be orthogonal so that they can be separated out by the demodulator in the receiver. Note that the baseband subcarriers $\{\cos(2\pi j/T_N), j = 1, 2 \ldots\}$ form a set of orthonormal basis functions on the interval $[0, T_n]$. Moreover, it is easily shown that no set of subcarriers with a smaller frequency separation forms an orthonormal set on $[0, T_n]$. This implies that the minimum frequency separation required for subcarriers to remain orthogonal over the symbol interval $[0, T_N]$ is $1/T_N$. So if we use raised cosine pulses with $\beta = 1$, we would have $T_N = 1/B_N$, and a carrier separation of B_N. Since the passband bandwidth of each subchannel is $2B_N$, the passband subchannels in this system would overlap, as illustrated in Figure 12.2.

Clearly, in order to separate out overlapping subcarriers, a different receiver structure is needed than the one shown in Figure 12.1. In particular, overlapping subchannels are demodulated with the receiver structure shown in Figure 12.3. We now show that this structure demodulates the appropriate symbol without interference from overlapping subchannels. For simplicity, our analysis will assume rectangular pulse shapes and in-phase signaling only, so that s_n in (12.1) is real and modulated with a cosine carrier. The same structure in Figure 12.3 using sine carriers would be used to demodulated the quadrature signal component, and the analysis for this component would be basically the same as the in-phase analysis.

Consider the received symbol on the ith branch, $\hat{s}_i$, in the absence of noise ($n(t) = 0$). The passband subcarriers with separation $1/T_N$ are given by $f_j = f_c + j/T_N, j = 0, 1, \ldots, N-1$. For transmitted/received

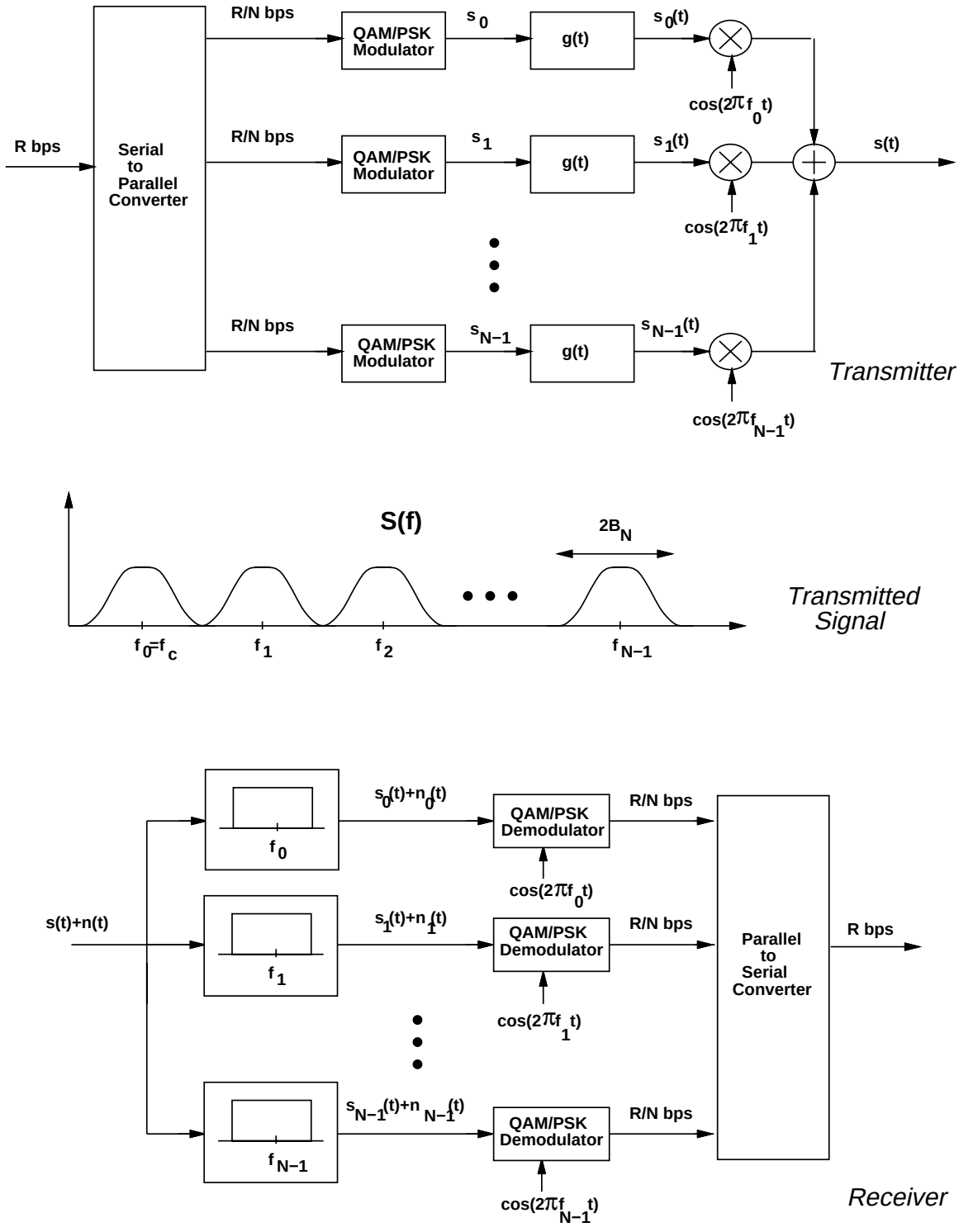


Figure 12.1: Multicarrier Transmitter and Receiver.

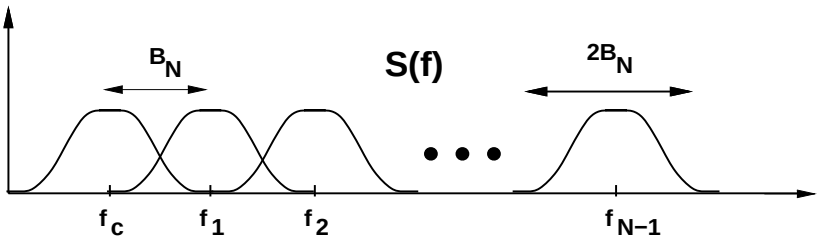


Figure 12.2: OFDM Signal with Overlapping Subcarriers.

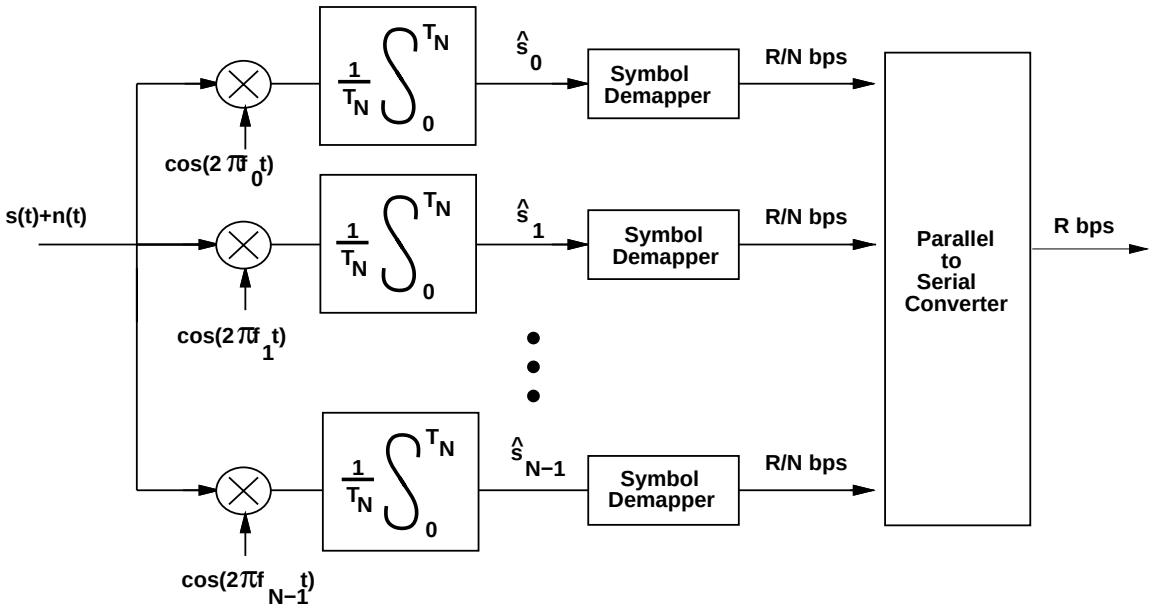


Figure 12.3: OFDM Receiver for Overlapping Subcarriers.

signal $s(t)$ given by (12.1) we have

$$\begin{aligned}
\hat{s}_i &= \frac{1}{T_N}\int_0^{T_N}\left(\sum_{j=0}^{N-1} s_j\cos(2\pi f_j t)\right)\cos(2\pi f_i t)dt \\
&= \frac{1}{T_N}\sum_{j=0}^{N-1} s_j\int_0^{T_N}\cos(2\pi(f_c+j/T_N)t)\cos(2\pi(f_c+i/T_N)t)dt \\
&= \frac{1}{2T_N}\sum_{j=0}^{N-1} s_j\left[\int_0^{T_N}\cos(2\pi(j-i)t/T_N)dt+\int_0^{T_N}\cos(2\pi(2f_c+j+i)/T_N)t)dt\right] \qquad (12.2)\\
&\approx \frac{1}{2}\sum_{j=0}^{N-1} s_j\delta(j-i) \\
&= \frac{1}{2}s_i, \qquad (12.3)
\end{aligned}$$

where the approximation holds for $f_c >> 1/T_N$, in which case the second integral in (12.2) is approximately zero. This multicarrier system has twice the bandwidth efficiency of the system depicted in

Figure 12.1 with nonoverlapping subcarriers. Note, however, that since the subcarriers overlap, their orthogonality is compromised by timing jitter, frequency offset, and fading. These effects, even when relatively small, can significantly degrade performance, as they cause subchannels to interfere with each other [9].

12.2 Discrete Implementation of OFDM (Discrete Multitone)

Although OFDM was invented in the 1950s, its requirement for separate modulators and demodulators on each subchannel was far too complex for most system implementations at the time. However, the development of the FFT and IFFT twenty years later, combined with the realization that OFDM can be implemented very simply and cheaply with these algorithms, ignited its widespread use. In this section we illustrate the discrete implementation of OFDM using FFT and IFFT hardware. This discrete implementation is sometimes refered to as discrete multitone modulation (DMT).

The DMT implementation of OFDM is shown in Figure 12.4 The input data stream is modulated by a QAM modulator, resulting in a complex symbol X. This symbol stream is passed through a serial-to-parallel converter, whose output is a set of N parallel QAM symbols corresponding to the symbols transmitted over each of the subcarriers. Thus, the N symbols output from the serial-to-parallel converter are the discrete frequency components of OFDM output signal $s(t)$. In order to generate $s(t)$, we therefore need to convert these frequency components into time samples. We therefore perform an inverse DFT on these N symbols, which is efficiently implemented using the IFFT algorithm. The IFFT yields a set of parallel outputs $\{x_0, \ldots, x_{N-1}$, where

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, \;\; n = 0, 1, \ldots, N-1. \tag{12.4}$$

If we ignore for the moment the addition of a cyclic prefix in the next block (discussed in detail below), the time samples $\{x_0, \ldots, x_{N-1}$ are ordered by the parallel-to-serial converter and passed through a D/A converter, resulting in the baseband OFDM signal $x(t)$. Specifically,

$$x(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, \;\; 0 \le t \le T_N \tag{12.5}$$

where, as defined earlier, T_N is the duration of the OFDM symbols. The subcarrier frequencies with this implementation are given by $f_i = i/T_N, \;\; i = 0, \ldots, N-1$, and the samples $\{x_0, \ldots, x_{N-1}$ represent samples of $x(t)$ every T_N/N seconds. The baseband OFDM signal $x(t)$ is upconverted to the carrier frequency, resulting in the transmitted signal $s(t)$. The receiver performs the reverse operation of the transmitter, so that in the absence of noise or channel distortion the original data sequence is perfectly recovered. Note that each subchannel has a symbol duration of T_N, which is chosen sufficiently large ($B/N << B_c$) to remove most ISI that might be introduced by the channel.

While the removal of most ISI generally yields good performance, it is possible to remove all ISI when the maximum delay spread of the channel is known. This can be done by either adding a guard time between symbol transmissions equal to the channel delay spread, in which case symbols do not interfere with subsequent symbols, or by adding a cyclic prefix after the IFFT (as shown in Figure 12.4), which is subsequently removed in the receiver. Let us now consider the cyclic prefix in more detail. Assume that the channel delay spread has a maximum value of $\mu T_N/N$, recalling that T_N/N is the sampling rate of the continuous time signal $x(t)$. Thus, the channel delay spread lasts a maximum of μ additional samples. The cyclic prefix $\{x_{N-\mu}, \ldots, x_{N-1}$ consists of the last μ values of the $\{x_n\}$ sequence. These μ

samples are appended to the beginning of each block of samples, yielding the new input to the D/A of $\{x_{N-\mu}, \ldots, x_{N-1}, x_0, x_1, \ldots, X_{N-1}\}$. Note that the cyclic prefix increases the number of samples in the $\{x_n\}$ sequence to $N + \mu$.

Let us now consider the channel impulse response $c(t)$. The received signal in the absence of noise will be $r(t) = x(t) * c(t)$. Consider sampling $c(t)$ every T_N/N seconds, which yields the sample set $\{c_0, \ldots, c_\mu\}$. If we convolve $\{x_{N-\mu}, \ldots, x_{N-1}$ with $\{c_0, \ldots, c_\mu\}$ we obtain the received sequence $\{r_n\}$, which has duration $N + \mu$. In the receiver we remove the cyclic prefix by removing the last μ samples in $\{r_n\}$. Note then that the input to the FFT is the circular convolution of $\{x_n\}$ and $\{c_n\}$, which does not necessarily correspond to samples of $r(t) = x(t) * c(t)$. However, adding the cyclic prefix at the transmitter effectively converts the circular convolution associated with the FFT to a linear convolution. Thus, the FFT output in the absence of noise is $\hat{X}_k = C_k X_k$, where C_k is the FFT of $\{c_0, \ldots, c_\mu\}$. This indicates that the effects of the channel $c(t)$ can be completely removed by frequency equalization, i.e. multiplying each $\hat{X}_k$ by $1/C_k$ to completely remove the effects of the channel. However, when noise is present this leads to noise enhancement and no net gain in the subchannel SNR, as discussed in more detail below.

12.3 Fading across Subcarriers

The advantage of multicarrier modulation is that each subchannel is relatively narrowband, which mitigates the effect of delay spread. However, each subchannel experiences flat-fading, which can cause large BERs on some of the subchannels. In particular, if the transmit power on subcarrier i is P_i, and the fading on that subcarrier is α_i, then the received SNR is $\gamma_i = P_i\alpha_i^2/(N_o B)$, where B is the bandwidth of each subchannel. If α_i is small then the received SNR on the ith subchannel is quite low, which can lead to a high BER on that subchannel. Moreover, in wireless channels the α_i will vary over time according to a given fading distribution, so we have the same flat-fading problem as discussed in the context of narrowband channels. We know from Chapter 6 that flat fading can seriously degrade performance, so it is important to somehow compensate for subchannels with a low SNR. There are several techniques for doing this, including frequency equalization, precoding, coding across subchannels, and adaptive loading. We now describe each of these techniques in more detail.

12.3.1 Frequency Equalization

In frequency equalization the fading α_i is basically inverted in the receiver, so the received signal is multiplied by $1/\alpha_i$ which gives a resultant signal power $P_i\alpha_i^2/\alpha_i^2 = P_i$. While this removes the impact of fading on the desired signal, it enhances the noise. Specifically, the incoming noise signal is also multiplied by $1/\alpha_i$, so the noise power becomes $N_0 B/\alpha_i^2$. Thus, the resultant SNR after frequency equalization, SNR_{eq}, is the same as before equalization. Therefore, frequency equalization does not really change the impact of the different subcarrier fading. More details on frequency equalization can be found in [2].

12.3.2 Precoding

Precoding uses the same idea as frequency equalization, except that the fading is inverted at the transmitter instead of the receiver. This technique requires that the transmitter have knowledge of the subchannel fading α_i. In this case, if the desired received signal power in the ith subchannel is P_i, and the channel introduces fading of α_i in that subchannel, then the transmitter sends the signal in the ith subchannel with power P_i/α_i^2. This signal is multiplied by the channel gain α_i, so the received signal power is $P_i\alpha_i^2/\alpha_i^2 = P_i$, as desired. Note that the channel inversion takes place at the transmitter instead of

the receiver, so the noise power remains as N_0B. Precoding is quite common on wireline multicarrier systems like HDSL. There are two main problems with precoding in a wireless setting. First, precoding is basically channel inversion, and we know from Section 6.3.5 that inversion is not power-efficient in fading channels. In fact, an infinite amount of power is needed to do channel inversion on a Rayleigh channel. The other problem with precoding is the need for accurate channel estimates at the transmitter. This same problem is encountered in adaptive modulation.

12.3.3 Adaptive Loading

Adaptive loading is based on the adaptive modulation techniques discussed in Chapter 9. The basic idea is to vary the data rate and power assigned to each subchannel relative to that subchannel gain. As in the case of adaptive modulation, this requires knowledge of the subchannel fading $\{\alpha_i, i = 1, \ldots, N\}$ at the transmitter. We can optimize the power and rate associated with each subchannel to maximize capacity or to maximize the rate of a variable-rate variable-power modulation scheme like MQAM. Let's first consider the capacity maximization. The capacity of the multicarrier system with N subchannels of baseband bandwidth B_N and corresponding subchannel gains $\{\alpha_i, i = 1, \ldots, N\}$ is given by[1]:

$$C = \sum_{i=1}^{N} B_N \log\left(1 + \frac{\alpha_i^2 P_i}{N_o B_N}\right), \tag{12.6}$$

where we assume a power constraint across subchannels as $\sum_i P_i = \overline{P}$. We would like to find the power P_i to allocate to each subchannel that maximizes this expression. But this is the same optimization problem as in Chapter 4.4. The optimal power allocation is thus the same water-filling given by Equation (4.24):

$$\frac{P_i}{\overline{P}} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_i} & \gamma_i \geq \gamma_0 \\ 0 & \gamma_i < \gamma_0 \end{cases} \tag{12.7}$$

for some cutoff value γ_0, where $\gamma_i = \alpha_i^2 \overline{P}/(N_0 B_N)$. The cutoff value is obtained by substituting the power adaptation formula into the power constraint. The capacity then becomes

$$C = \sum_{i=1(\gamma_i \geq \gamma_0)}^{N} B_N \log(\gamma_i/\gamma_0). \tag{12.8}$$

Suppose we now apply the variable-rate variable-power MQAM modulation scheme described in Chapter 9 to the subchannels. Then our total data rate will be given by

$$R = B_N \sum_{i=1}^{N} \log(1 + K\gamma_i P_i/\overline{P}), \tag{12.9}$$

where $K = -1.5/\ln(5\text{BER})$. Optimizing this expression relative to the P_is yields the optimal power allocation

$$\frac{P_i}{\overline{P}} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{K\gamma_i} & \gamma_i \geq \gamma_0/K \\ 0 & \gamma_i < \gamma_0/K \end{cases} \tag{12.10}$$

[1]This summation is the exact capacity when the α_is are independent. However, in order for the α_is to be independent, the subchannels must be separated by the coherence bandwidth of the channel, which would imply that the subchannels are no longer flat fading. Since the subchannels are designed to be flat-fading, the subchannel gains $\{\alpha_i, i = 1, \ldots, N\}$ will be correlated, in which case the capacity obtained by summing over the capacity in each subchannel is an upper bound on the true capacity. We will take this bound to be the actual capacity, since in practice the bound is quite tight.

and corresponding data rate

$$R = B_N \sum_{i=1,\gamma_i \geq \gamma_0/K}^{N} \log(K\gamma_i/\gamma_0). \tag{12.11}$$

12.3.4 Coding across Subchannels

The basic idea in coding across subchannels is to encode incoming bits into a length-N codeword, where N is the number of subchannels in the multicarrier system. Then if most of the subchannels have a high SNR, the codeword will have most coded bits received correctly, and the errors associated with the few bad subchannels can be corrected. This technique only works well for channels with a large delay spread, where the coherence bandwidth of the channel is on the order of the bandwidth of each subchannel. In that case each subchannel has independent fading, and coding across subchannels provides frequency diversity. However, if the coherence bandwidth of the channel is large, then the fading across subchannels will be highly correlated, which will significantly reduce the effect of coding. Note that coding across subchannels is the only technique discussed in this section that takes advantage of the fact that the data on all the subcarriers can be processed simultaneously. The other techniques discussed in this section are all basically flat-fading compensation techniques, which apply equally to multicarrier systems as well as narrowband flat-fading channels.

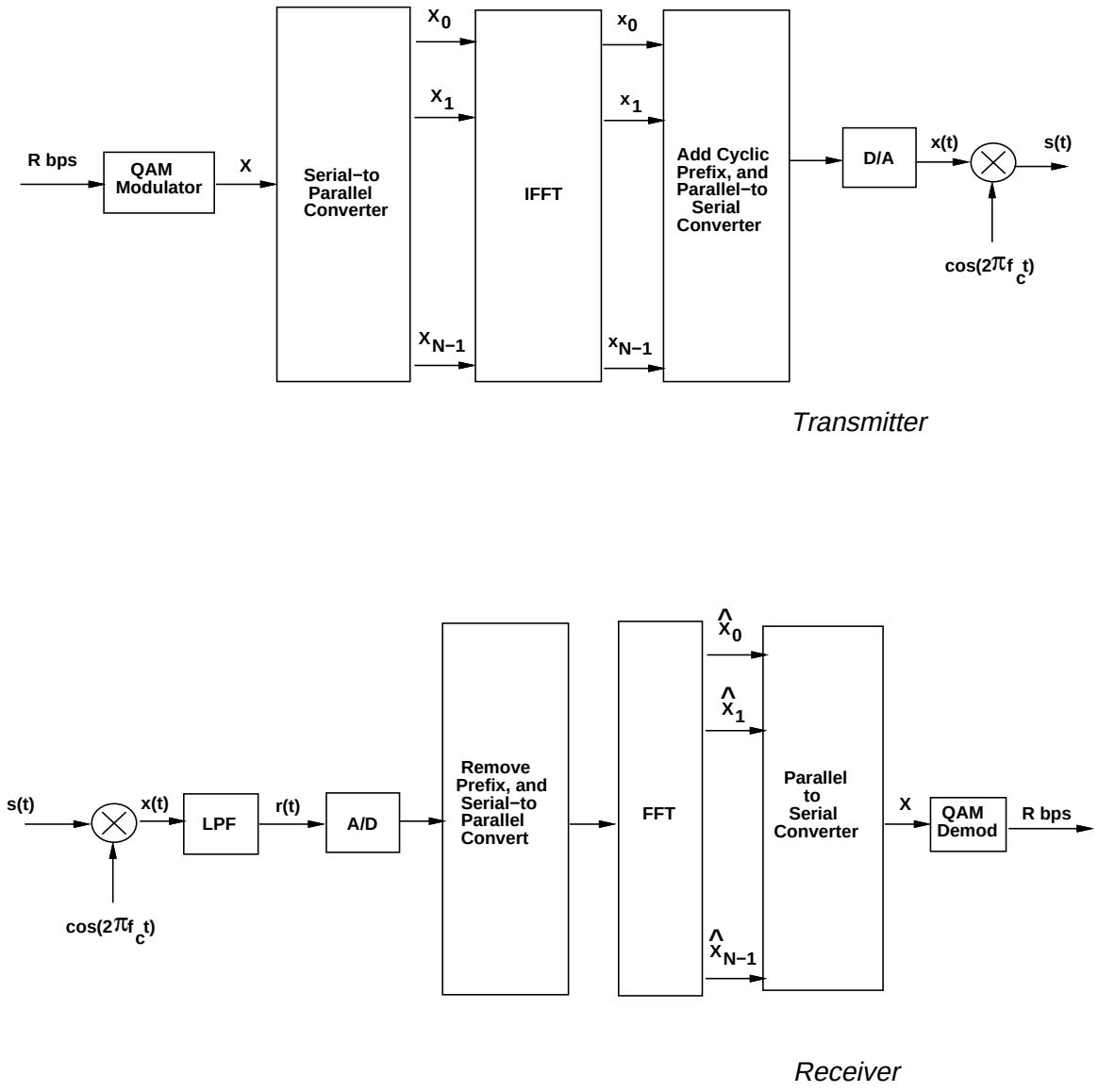


Figure 12.4: OFDM with IFFT/FFT Implementation.

Bibliography

[1] J. Bingham, "Multicarrier modulation for data transmission: an idea whose time has come," *IEEE Commun. Mag.* Vol. 28, No. 5, pp. 5-14, May 1990.

[2] H. Sari, G. Karam, and I. Jeanclaude, "Transmission techniques for digital terrestrial TV broadcasting," *IEEE Commun. Mag.* Vol. 33, No. 2, pp. 100–109.

[3] J.S. Chow, J.C. Tu, and J.M. Cioffi, "A discrete multitone transceiver system for HDSL applications," *IEEE J. Select. Areas. Commun.*, Vol. 9, No. 6, pp. 895–908, Aug. 1991.

[4] I. Kalet and N. Zervos, "Optimized decision feedback equalization versus optimized orthogonal frequency division multiplexing for high-speed data transmission over the local cable network," *Proc. of ICC'89*, pp. 1080–1085, Sept. 1989.

[5] S. Kasturia, J.T. Aslanis, and J.M. Cioffi, "Vector coding for partial response channels," *IEEE Trans. Inform. Theory,* Vol. 36, No. 4, pp. 741–762, July 1990.

[6] L.J. Cimini, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing," *IEEE Trans. Inform. Theory,* Vol. 33, No. 7, pp. 665–675, July 1985.

[7] P.S. Chow, J.M. Cioffi, and John A.C. Bingham, "A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels," *IEEE Trans. Commun.*, Vol. 43, No. 2/3/4, Feb.-Apr. 1995.

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

[9] A. R. S. Bahai and B. R. Saltzberg, *Multi-Carrier Digital Communications - Theory and Applications of OFDM*, Kluwer Academic Publisher: Plenum Press, 1999.

Chapter 13

Spread Spectrum and RAKE Receivers

Although bandwidth is a valuable commodity in wireless systems, *increasing* the transmit signal bandwidth (e.g. with coding) can sometimes improve performance. Spread spectrum is a technique which increases signal bandwidth to reduce ISI and narrowband interference. In conjunction with a RAKE receiver, spread spectrum also provides a form of diversity, called code diversity. Spread spectrum first achieved widespread use in military applications due to its inherent property of "hiding" the signal below the noise floor during transmission, and its resistance to narrowband jamming. Since both of these properties are desirable in wireless systems as well, it has become increasingly pervasive in wireless system designs, and is now one of the two standards for digital cellular in the U.S.

13.1 Spread Spectrum Modulation

Spread spectrum is a modulation technique which increases the transmit signal bandwidth. There are several benefits obtained in exchange for this increased bandwidth. First, spread spectrum modulation mitigates the effect of intersymbol interference (ISI) and narrowband interference. The narrowband interference rejection also applies to hostile jamming signals, and for that reason spread spectrum is often used in military systems. In addition, spread spectrum also "hides" the signal beneath the noise floor. This property means that the transmitted signal is unlikely to be detected (low probability of interceptions, or LPI), another desirable property for military communication systems. Finally, spread spectrum modulation can be used as a multiple access technique, which is the basis of the digital cellular standard IS-95.

There are two common forms of spread spectrum: *direct sequence* and *frequency hopping*. Since direct sequence is more commonly used, we will focus on this technique. Frequency-hopping is briefly described in §13.6. In direct sequence spread spectrum modulation, the data signal $s_b(t)$ is multiplied by a pseudorandom sequence $s_c(t) = \pm 1$, where the bit duration T_b is some multiple K of the spreading code bit duration T_c. The spreading code bits are usually referred to as *chips*. The multiplication of spreading code and BPSK data sequence is illustrated in Figure 13.1.

The pseudorandom sequence is also called the chip sequence, and $1/T_c$ is called the chip rate. Because the chip duration is a fraction K of the bit duration, its 3 dB bandwidth B_c is approximately K times bigger than the original signal bandwidth B_b[1]. Thus, multiplying the data sequence by the spreading code results in the convolution of these two signals in the frequency domain. Thus, the transmitted

[1]For square pulses, both the data sequence and the chip sequence have infinite bandwidth. The 3 dB signal bandwidth is defined as the bandwidth where the signal power first reaches half of its maximum value.

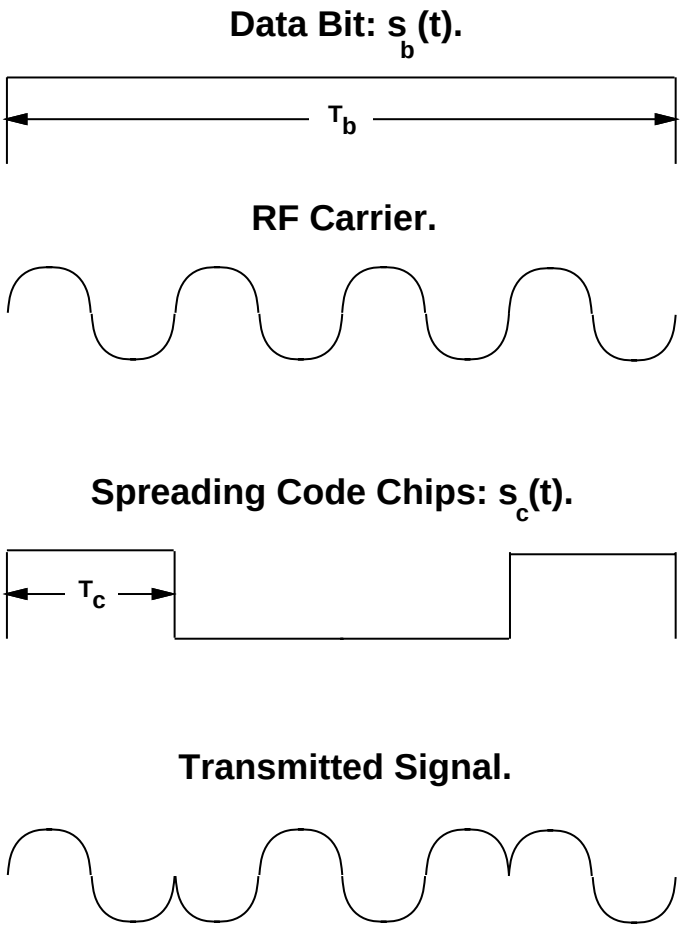


Figure 13.1: Spreading Code Multiplication

signal $s_b(t)s_c(t)$ has frequency response $S(f) = S_b(f) * S_c(f)$, which has a 3 dB bandwidth B of roughly $(K+1)B_b$.

The spreading factor J is defined as the ratio of the signal bandwidth after spreading B to the original data signal bandwidth B_b. The spread factor is generally approximately equal to the number of chips per bit: $J = B/B_b \approx K$. This factor determines how much interference and multipath rejection occurs at the receiver, as will be discussed in more detail below. The value of J depends both on the signal modulation used and the bandwidth of the spreading code. When the code is modulating a pure RF carrier, the transmitted signal is a sequence of pulses, which has the form of a sinc function in the frequency domain. The approximate transmitted bandwidth for this signal equals the 3 dB bandwidth of the main lobe ($.88/T_c$). Even when the data signal is not a pure carrier, this value is usually used to approximate the bandwidth of the transmitted signal.

13.2 Pseudorandom (PN) Sequences (Spreading Codes)

The PN sequences are deterministic, with approximately the same number of +1s and -1s, low correlation between shifted versions of the same sequence, and a low cross correlation between different sequences. Thus, although they are deterministic, they have many of the same characteristics as a random binary

sequence of ± 1. For this reason they are called pseudorandom sequences.

These sequences are typically generated as a binary sequence of 1s and 0s using a shift register with feedback logic, and then the sequence of 1s and 0s is converted to a pulse train of ± 1. The shift register, consisting of n stages, is illustrated in Figure 13.2. The binary sequence output from the shift register is cyclical with a maximum period of $2^n - 1$ cycles.

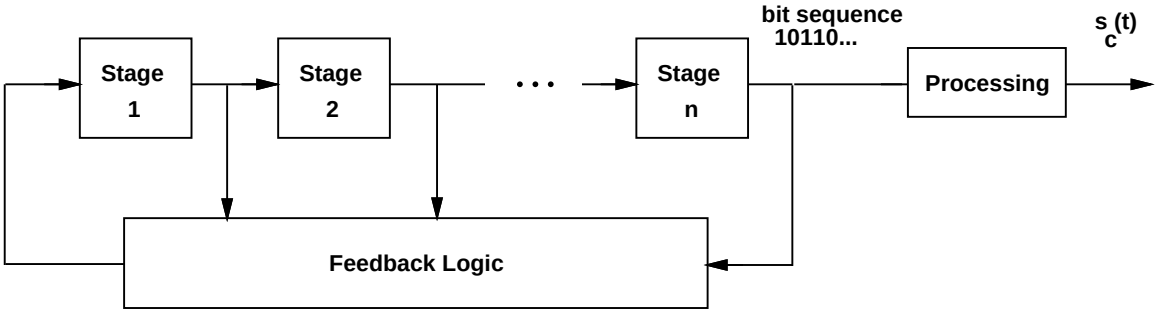


Figure 13.2: Generation of PN Sequences

The autocorrelation of the spreading code $s_c(t)$ over a bit time determines the multipath rejection properties of the spread spectrum signaling. This autocorrelation is defined as

$$\rho_c(\tau) = \frac{1}{T_b}\int_0^{T_b=KT_c} s_c(t)s_c(t-\tau)dt = \frac{1}{K}\sum_{k=1}^{K} s_c(kT_c)s_c(kT_c-\tau). \tag{13.1}$$

As will be evident in the analysis below, the best autocorrelation function for multipath rejection is a delta function: $\rho_c(\tau) = \delta(\tau)$. Unfortunately, we can't design codes with the autocorrelation equal to a delta function for finite values of n. Much work in spread spectrum in the 60s and 70s was focussed on designing codes with autocorrelation close to a delta function.

Among all linear codes, maximal linear codes were found to have the best autocorrelation properties for ISI rejection. Maximal codes are the longest codes that can be generated by a shift register of a given length. Specifically, the codes have length $N = 2^n - 1$ for a binary shift register of length n, so the codes repeat every NT_c seconds. For maximal codes, the number of negative ones in a sequence $(2^{n-1} - 1)$ is approximately equal to the number of ones (2^{n-1}), and the statistical distribution of these two chip values doesn't change over the course of the sequence. This property has some desirable features relative to implementation [1], and also insures maximal spreading (if a sequence of 1s is very long, on the order of N, then the data bit roughly just gets multiplied by 1 and there is no spreading). The autocorrelation function of maximal codes, assuming the number of chips per bit $K = N$, is given by

$$\rho_c(\tau) = \begin{cases} 1 - \frac{|\tau|(1+1/N)}{T_c} & |\tau| \le T_c \\ -1/N & |\tau| > T_c \end{cases} \tag{13.2}$$

Thus, for all delays bigger than a chip time, the correlation value is $-N = -1/(2^n - 1)$, which decreases exponentially with the size of the spreading sequence n. For delays between $-T_c$ and T_c, the correlation varies linearly from $-1/N$ to 1, with the maximum corresponding to zero delay, as illustrated in Figure 13.3. We will see in the next section why this autocorrelation property has good ISI rejection, assuming that $K = N$. When $K < N$ the autocorrelation function is not as steep, and thus maximal length codes are not as effective at removing ISI. More details about code properties and design can be found in [1, 2, 3].

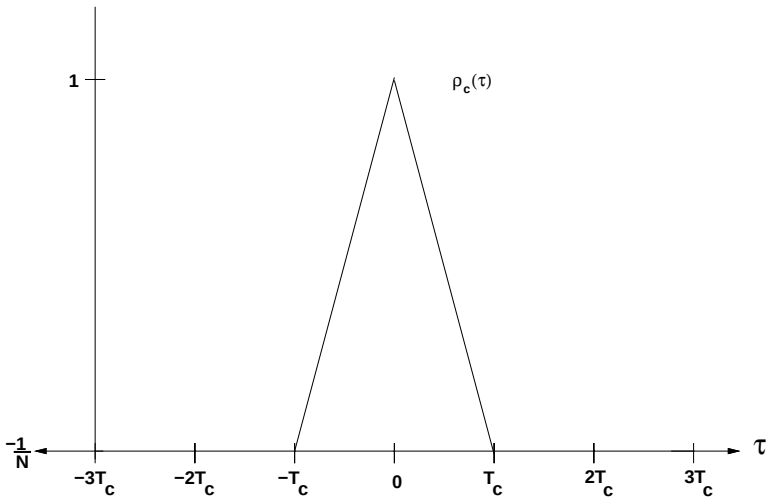


Figure 13.3: Autocorrelation of Maximal Code ($K = N$)

13.3 Direct Sequence Spread Spectrum

An end-to-end direct sequence spread spectrum system is illustrated in Figure 13.4. The data bits are first modulated to form the baseband data signal $x(t) = \sum_k d_k g(t - kT)$, where $g(t)$ is the modulator shaping pulse. This signal is then multiplied by the spreading code $s_c(t)$ with chip time $T_c = T_b/K$, and then upconverted through multiplication by the carrier $\cos 2\pi f_c t$. The spread signal passes through the channel $h(t)$ which also introduces additive noise $n(t)$ and narrowband interference $I(t)$.

The receiver input $r(t)$ is used to synchronize the PN generator to the spreading code delay τ introduced in transmission. This synchronization procedure can be quite complex, especially for ISI channels, and synchronization circuitry makes up a large part of any spread spectrum receiver. Details on synchronization can be found in [1, 3].

Assuming perfect synchronization, the received signal $r(t)$ is despread by multiplying it with a synchronized version $s_c(t - \tau)$ of the original spreading code $s_c(t)$. The value of τ will depend on the channel and the synchronizer. If $h(t) = \delta(t)$, i.e. no multipath is introduced by the channel, then $\tau = 0$. However, if the channel introduces multipath, then the synchronizer will generally synchronize either to the strongest multipath component or the first multipath component above a given threshold. Specifically, if $h(t) = \sum_i \alpha_i \delta(t - \tau_i)$ the synchronizer locks to the strongest multipath component by setting $\tau = \tau_i$ for $\alpha_i = \max_j \alpha_j$. It locks to the first component above a given threshold α_{thres} by setting $\tau = \tau_i$ for $\alpha_i = \min_j : \alpha_j \geq \alpha_{thres}$. The synchronizer also helps with carrier recovery, and therefore has input to the demodulator also. After despreading, the signal $\hat{x}(t)$ passes through a conventional demodulator and decision device. This baseband recovery of the data signal is identical to the optimal detectors defined in Chapter 5.3 for different linear modulation techniques. Nonlinear (constant envelope) modulation is typically used in conjunction with frequency-hopping spread spectrum. Thus, there are two stages in the receiver demodulation for direct sequence spread spectrum: despreading and narrowband demodulation. We now examine these two stages in more detail.

We assume for simplicity that modulator is simple BPSK with rectangular pulse shapes ($g(t) = 1, 0 \leq t \leq T_b$). The matched filter then simply multiplies $\hat{x}(t)$ by the carrier $\cos 2\pi f_c t$, then multiplies by $1/T_b$ and integrates from zero to T_b. The multipath and interference rejection occurs in this demodulation

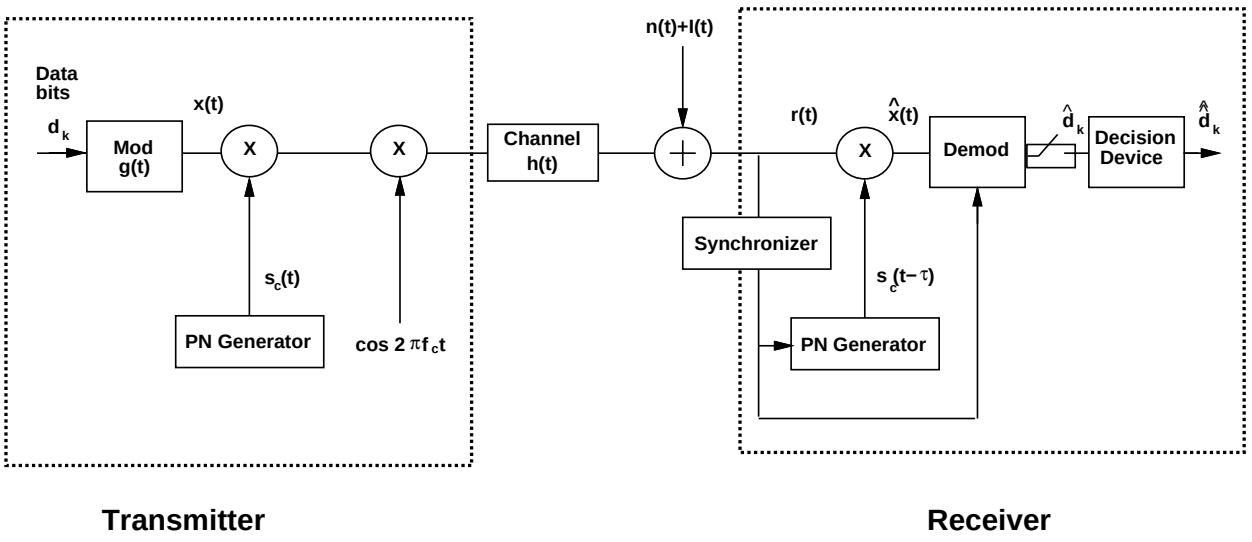


Figure 13.4: Spread Spectrum System

process. Specifically, the input to the demodulator is given by

$$\hat{x}(t) = [x(t)s_c(t)\cos(2\pi f_c t) * h(t)]s_c(t-\tau) + n(t)s_c(t-\tau) + I(t)s_c(t-\tau). \tag{13.3}$$

Note that in the absence of multipath $h(t) = \delta(t)$, the spreading/despreading process has no impact on the baseband signal $x(t)$. Specifically, the PN code consists of ± 1, so multiplying $s_c(t)$ by an synchronized copy of itself yields $s_c^2(t) = 1$ for all t. For $h(t) = \delta(t)$, the baseband signal component of $\hat{x}(t)$ is $[x(t)s_c(t)]s_c(t) = x(t)$, since $s_c^2(t) = 1$. We assume that the statistics of the noise do not change when it is multiplied by the wideband PN sequence[2]. Therefore, without interference $I(t) = 0$ the spreading/despreading process has no impact on performance. We now discuss its interference and multipath rejection properties. We will treat interference and multipath rejection separately: when both are present we can combine the analyses to determine the combined effect.

For narrowband interference rejection, assume $h(t) = \delta(t)$. Then the synchronizer in Figure 13.4 will have $\tau = 0$, yielding

$$\hat{x}(t) = x(t)s_c^2(t) + n(t)s_c(t) + I(t)s_c(t) = x(t) + n'(t) + I(t)s_c(t). \tag{13.4}$$

Then the demodulator output is given by

$$\begin{aligned} \hat{d}_k &= \frac{1}{T_b}\int_0^{T_b} d_k s_c^2(t)\cos^2(2\pi f_c t)dt + \frac{1}{T_b}\int_0^{T_b} n(t)s_c(t)\cos(2\pi f_c t)dt + \frac{1}{T_b}\int_0^{T_b} I(t)s_c(t)\cos(2\pi f_c t)dt \\ &= .5d_k + n_k + I_k. \end{aligned} \tag{13.5}$$

Since multiplying an AWGN process by a carrier does not change its statistics, n_k is an AWGN sample. The narrowband interference rejection is apparent in the last term of (13.5). Specifically, we assume that $I(t)$ is narrowband (with a bandwidth on the order of T_b^{-1}), so it will be approximately constant over a bit time T_b. Recall that the spreading sequence $s_c(t)$ is a sequence of ± 1 that changes every chip time $T_c = T_b/K$. Thus, integrating the product $I(t)s_c(t)$ over a bit time T_b will yield approximately zero.

[2]This is not true in general, but is reasonably accurate for analysis purposes, and greatly simplifies the analysis.

Also, the cosine term is changing rapidly relative to $I(t)$ and $s_c(t)$ since $f_c >> T_c^{-1} >> T_b^{-1}$. Therefore, $I_k \approx 0$. The narrowband interference rejection can be seen more precisely in the frequency domain, as shown in Figure 13.5. We see in this figure that when the narrowband interference $I(f)$ is multiplied by the spreading code in the receiver, its bandwidth is increased by the spreading factor $J \approx K$. When this spread interference signal goes through the demodulator, it passes through an integrator or equivalently, a narrowband filter the size of the data signal bandwidth. Thus, all of the spread interference outside the narrowband filter is removed, as shown in Figure 13.5. Thus, the interference power in the demodulator is reduced by approximately a factor of K. The noise is not impacted by the spreading process in the receiver.

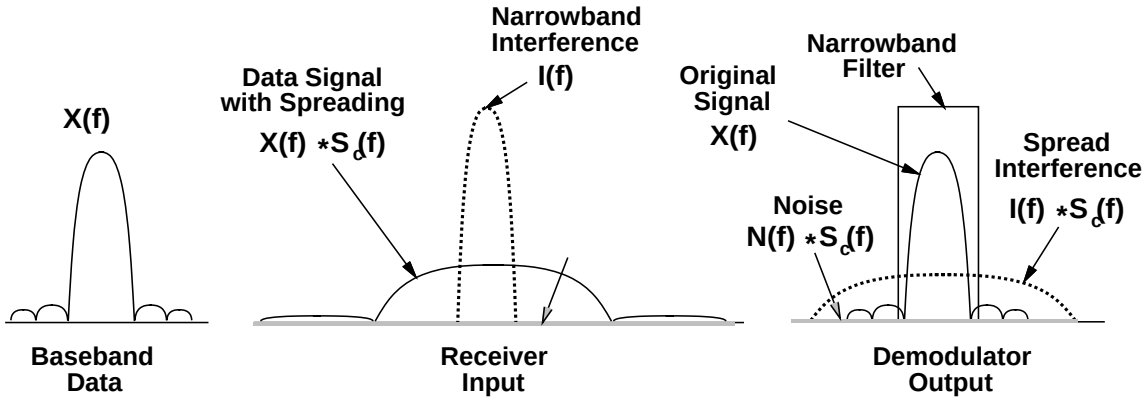


Figure 13.5: Signal and Interference after Despreading

The ISI rejection is even stronger when the spreading code has good autocorrelation properties over a bit time. Let us assume a multipath channel with one delayed component: $h(t) = \alpha_1\delta(t) + \alpha_2\delta(t - \tau_0)$. Suppose that the first multipath component is stronger than the second: $\alpha_1 > \alpha_2$, so that the receiver synchronizes to the first component ($\tau = 0$ in Figure 13.4). Then, in the absence of narrowband interference ($I(t) = 0$), we have

$$\hat{x}(t) = \alpha_1 x(t) s_c^2(t)\cos(2\pi f_c t) + \alpha_2 x(t-\tau_0)s_c(t-\tau_0)s_c(t)\cos(2\pi f_c(t-\tau_0)) + n(t)s_c(t). \tag{13.6}$$

The demodulator output is then given by

$$\begin{aligned}\hat{d}_k &= \frac{1}{T_b}\int_0^{T_b} \alpha_1 d_k s_c^2(t)\cos^2(2\pi f_c t)dt + \frac{1}{T_b}\int_0^{T_b} \alpha_2 d_{k-k_0} s_c(t)s_c(t-\tau_0)\cos(2\pi f_c t)\cos(2\pi f_c(t-\tau_0))dt \\ &+ \frac{1}{T_b}\int_0^{T_b} n(t)s_c(t)\cos(2\pi f_c t)dt \\ &\approx .5\alpha_1 d_k + .5\alpha_2 d_{k0} + n_k, \end{aligned} \tag{13.7}$$

where d_{k-k0} is the symbol corresponding to time $t - \tau_0$ and the approximation assumes $f_c >> 1/T_b$. The noise n_k is an AWGN sample as described above. Let us consider the term d_{k0}:

$$\begin{aligned} d_{k0} &= d_{k-k_0}\frac{1}{T_b}\int_0^{T_b} s_c(t)s_c(t-\tau_0)\cos(2\pi f_c t)\cos(2\pi f_c(t-\tau_0))dt \\ &= d_{k-k_0}\frac{1}{T_b}\int_0^{T_b} s_c(t)s_c(t-\tau_0)(\cos(2\pi f_c\tau_0) + \cos(4\pi f_c t - 2\pi f_c\tau_0))dt \\ &\stackrel{a}{=} d_{k-k_0}\cos(2\pi f_c\tau_0)\frac{1}{T_b}\int_0^{T_b} s_c(t)s_c(t-\tau_0)dt \end{aligned}$$

$$\stackrel{b}{=} \quad d_{k-k_0}\cos(2\pi f_c\tau_0)\rho_c(\tau_0), \tag{13.8}$$

where a follows from the fact that $f_c >> T_c^{-1}$ and b follows from the definition of $\rho_c(t)$. We therefore see that the multipath rejection is a direct function of the spreading code autocorrelation. If we assume a maximal length code with $K = N$ and $\tau_0 > T_c$ then the multipath term $d_{k-k_0}\cos(2\pi f_c\tau_0)\rho_c(\tau_0) = -d_{k-k_0}\cos(2\pi f_c\tau_0)/K$, i.e. the power of all multipath components at delays greater than a chip time is reduced by roughly the spreading gain. Since the spreading gain is generally quite large, this effectively removes most of the ISI. There is some constructive and destructive interference from multipath delayed by less than a chip time, which gives rise to Rician fading statistics of the LOS path.

The problem with the above receiver design is that the exact delay of the LOS component is not known. Moreover, the LOS path may not be the strongest path, or it may be blocked entirely. In practice, a spread spectrum receiver has a header with known data which is used to *acquire* the delay of the spreading code. The acquisition loop steps through the code until it finds the delay which is highly correlated with the incoming signal. It then further refines the delay until it is perfectly synchronized with the incoming signal. The synchronization is continually adjusted throughout data demodulation since, if the receiver code becomes delayed by more than a small fraction of T_c relative to the transmitted signal code, performance is significantly compromised. The code acquisition and tracking process is the hardest part of implementation in spread spectrum systems. See [1, 3] for detailed design and analysis of this part of the receiver design.

A more complicated receiver can have several branches, with each branch synchronized to a different multipath component (so the time delay of the PN code between branches is T_c). The receiver determines which branch has the strongest signal, and this signal is passed to the demodulator. This turns out to be the simplest implementation of a RAKE receiver, shown in Figure 13.6 below, where choosing the branch with the strongest signal corresponds to selection diversity. We now describe the general RAKE structure, which can use any of the diversity combining techniques discussed in Chapter 7.

13.4 RAKE receivers

A RAKE receiver uses the autocorrelation properties of the code to coherently combine all multipath components. The RAKE receiver structure is shown in Figure 13.6. The RAKE is essentially another form of diversity combining, since the spreading code induces a time diversity on the transmitted signal so that independent multipath components separated by more than a chip time can be resolved. Any of the combining techniques discussed in Chapter 7 may be used, although equal gain combining is the most common, since it doesn't require knowledge of the multipath amplitudes. A more detailed description of the RAKE receiver for the discrete-time multipath channel model with unknown delays and amplitudes can be found in [2]. If we ignore the effects of interference and synchronization errors, and we also assume a steep autocorrelation function which equals zero for codes delayed by more than a chip time and one for codes within a chip time, then performance of the RAKE receiver with M branches is identical to any other M-branch diversity technique. Since these assumptions usually don't hold in practice and spread spectrum is more difficult to implement than other diversity techniques, spread spectrum is not usually used for diversity alone. However, if spread spectrum signaling is chosen for its other benefits, such as its multiuser or interference rejection capabilities, then M branch diversity comes almost for free.

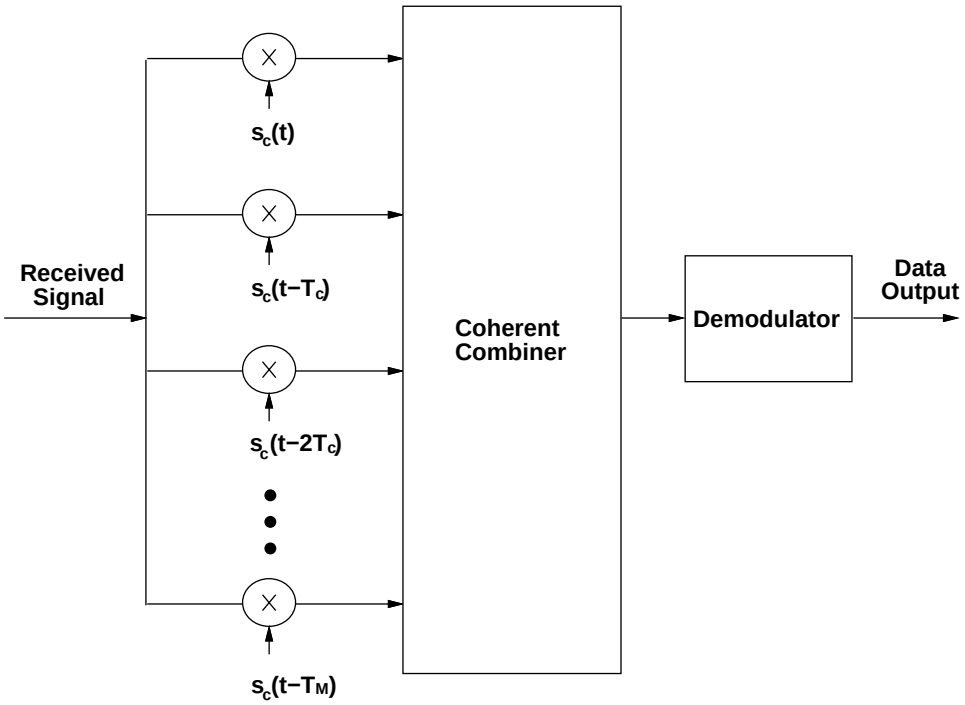


Figure 13.6: RAKE receiver

13.5 Spread Spectrum Multiple Access

In spread spectrum multiple access each user has a unique spreading code assigned to him which is used to modulate his transmitted signal. The users, modulated by their unique spreading codes, all occupy the same bandwidth, so they are superimposed in time and in frequency. However, the receiver can use the properties of the spreading codes to separate out different users. The spreading codes, which can be orthogonal or semi-orthogonal, consist of chip sequences at a much higher rate than the data rate, so the spreading code has a larger bandwidth than the data signal. Hence, modulating the data signal with these spreading codes results in a larger transmit signal bandwidth. However, since signals modulated by different spreading codes can occupy the same bandwidth and still be separated out at the receiver, this technique is bandwidth efficient. In fact, under ideal conditions spread spectrum multiple access with orthogonal codes can accommodate the same number of users as time-division and frequency-division, since all of these techniques provide orthogonal channels between users. Spread spectrum multiple access with semi-orthogonal codes can support more users than these orthogonal techniques, since there is often an infinite number of semi-orthogonal codes that can be assigned to different users. However, with semi-orthogonal codes users interfere with each other, so that although there is no hard limit on the total number of users that can share the channel, if too many users access the channel simultaneously then all users will have poor performance. We thus say that systems with semi-orthogonal codes are "interference-limited."

13.5.1 Spreading Codes for Multiple Access

Multiple access using direct sequence spread spectrum is accomplished by assigning each user a unique spreading code sequence $s_{c_i}(t)$. As we saw earlier, the autocorrelation function of this code determines

its multipath rejection properties. The autocorrelation, which is typically defined over one symbol time, is given by

$$\rho(\tau) = \frac{1}{T_s}\int_0^{T_s} s_{c_i}(t)s_{c_i}(t-\tau)dt. \tag{13.9}$$

The cross correlation properties of the code set determine the amount of interference between users. The cross correlation between the codes assigned to user i and user j over one symbol time is given by

$$\rho_{ij}(\tau) = \frac{1}{T_s}\int_0^{T_s} s_{c_i}(t)s_{c_j}(t-\tau)dt. \tag{13.10}$$

Ideally we would like $\rho(\tau) = \delta(\tau)$ to eliminate multipath and $\rho_{ij}(\tau) = 0 \ \ \forall\tau$ to eliminate multiple access (MAC) interference. However, $\rho_{ij}(\tau) = 0 \ \forall\tau$ only if we have orthogonal codes. Specifically, if we have K chips per bit we can obtain, using Walsh-Hadamard codes for example [1, 4], exactly K orthogonal codes ($\rho_{ij}(\tau) = 0 \ \forall\tau$). Suppose our information signal has bandwidth B. Since the spreading code with K chips per bit has a bandwidth expansion of roughly K, each user's spread signal requires a bandwidth of roughly KB. Since these users share the same spectrum, with orthogonal coding we require a bandwidth of K times the original signal bandwidth to support K users. This is the same requirement as frequency-division, so under ideal conditions these two techniques are equivalent. In practical scenarios spread spectrum with orthogonal codes is more susceptible to multipath (since it is a wideband signal) and it is more complex to implement. In addition, multipath in the channel typically compromises the orthogonality of the codes, leading to MAC interference. For these reasons multiple access is not typically implemented using orthogonal spread spectrum coding.

We can accommodate more than K users in a total bandwidth of KB using semi-orthogonal codes, but then the signals modulated by these codes cannot be completed separated in the receiver. Thus semi-orthogonal codes exhibit nonzero MAC interference. However, we can design spreading codes to make this interference as small as possible (i.e. $\rho_{ij}(\tau) \approx 0 \, \forall\tau$). In spreading code design there is usually a tradeoff between good multipath rejection properties ($\rho(\tau) \approx \delta(\tau)$) and good MAC interference rejection properties ($\rho_{ij}(\tau) \approx 0$). It is difficult to design codes which are good at both multipath and MAC interference rejection.

We can typically get a very large number of semi-orthogonal codes with cross correlation

$$\rho_{ij}(\tau) \stackrel{\triangle}{=} \frac{1}{\sqrt{G}} \approx \frac{1}{\sqrt{J}}, \tag{13.11}$$

where J is the bandwidth expansion factor of the codes (the ratio of the spread signal bandwidth to the original signal bandwidth) of the codes. Gold codes are an example of codes with this property [1, 4].

13.5.2 Broadcast Channels

Let us first consider a broadcast channel with semi-orthogonal codes. The transmitter for this system is shown in Figure 13.7 and the channel and receiver in Figure 13.8. In a broadcast channel the signals of all users are sent simultaneously by the transmitter (base station), and each receiver must demodulate its individual signal. Specifically, for a K-user system the transmitter has K modulated signals $s_1(t), \ldots, s_K(t)$ to send to the K users. Assuming linear modulation these signals are given by

$$s_i(t) = \sum_{l=1}^{\infty} d_{il}g(t - lT_s), \tag{13.12}$$

where d_{il} is the ith user's symbol over symbol time $[(l-1)T_s, lT_s]$, $g(t)$ is the pulse shape and T_s the symbol time. For simplicity in our analysis we assume a baseband system with binary modulation ($T_s = T_b$) and rectangular pulse shapes. The analysis easily extends to more general linear modulations at passband.

The transmitter consists of K branches, where the ith branch multiplies the ith signal $s_i(t)$ with a semi-orthogonal spreading code $s_{c_i}(t)$. The branches are summed together, resulting in the signal

$$x(t) = \sum_{i=1}^{K} s_i(t) s_{c_i}(t) \tag{13.13}$$

which is transmitted over the channel.

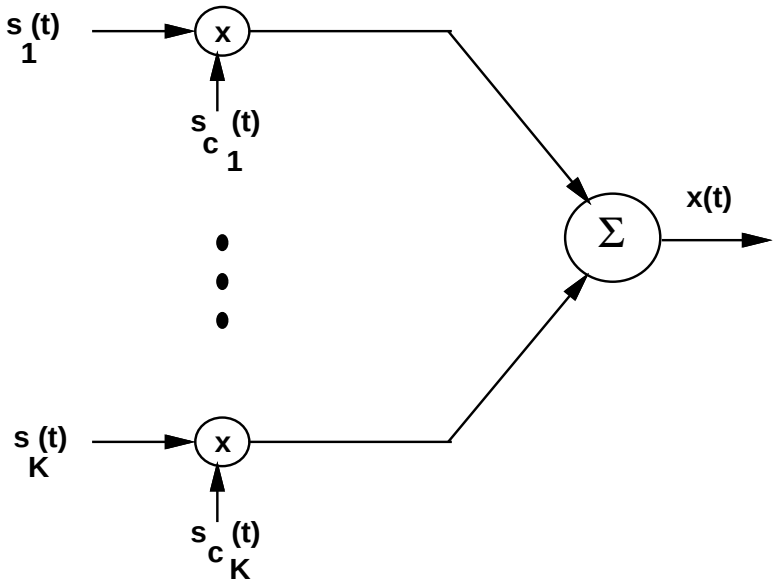


Figure 13.7: Broadcast Transmitter.

The signal received by user i first passes through the ith user's channel, which has impulse response $h_i(t)$ and AWGN. Thus the received signal at the ith user's receiver is $x(t) * h_i(t) + n(t)$. This signal is first multiplied by the ith user's spreading code $s_{c_i}(t)$, which is perfectly synchronized to the corresponding code in the transmitted signal[3]. The signal is then integrated over a bit time (symbol time for nonbinary modulation). The output of the integrator is

$$\hat{s}_i(t) + I_i(t) = \hat{d}_{il} + I_{il}, \tag{13.14}$$

where $\hat{d}_{il}$ is the demodulated bit from the ith user at time l and I_{il} is the interference from other users over this bit time.

Let us first assume that each user has a channel with no multipath, so $h_i(t) = \alpha_i \delta(t)$. First consider the desired signal component $\hat{s}_i(t) = \hat{d}_{il}$. We have

$$\hat{s}_i(t) = \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} [s_i(t) s_{c_i}(t) * h_i(t) + n(t)] s_{c_i}(t) dt = \alpha_i d_{il} + n_{il} = \hat{d}_{il}, \tag{13.15}$$

[3]This synchronization is even more difficult than in the single-user case, since it must be done in the presence of multiple spread signals. In fact some spreading code sets are obtained by shifting a single spreading code by some time period. For these systems there must be some control channel to inform the receiver which time shift corresponds to its desired signal. More details on the synchronization for these systems can be found in [1].

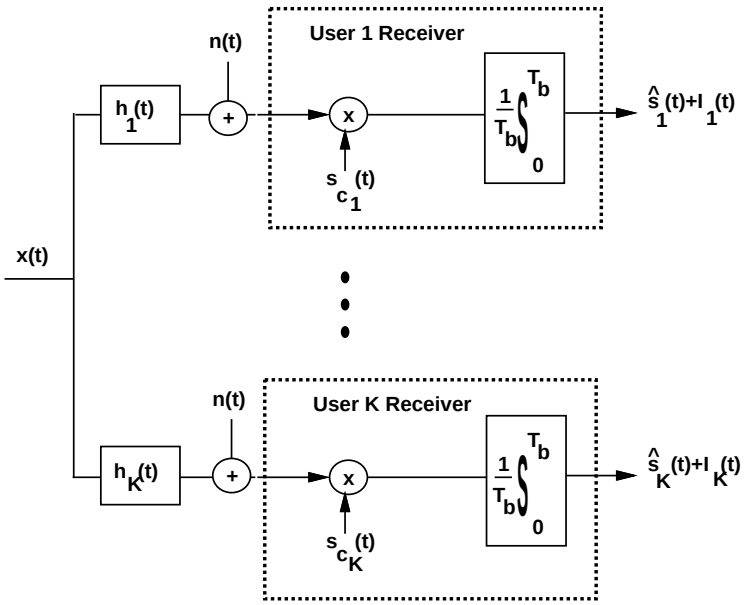


Figure 13.8: Broadcast Channel and Receiver.

where $n_{il} = \int_{(l-1)T_b}^{lT_b} n(t)s_{c_i}(t)dt$ is the noise sample at the output of the integrator at time l. Thus, in the absence of other users, the demodulated bit differs from the original bit due to signal attenuation and noise, as in a single-user channel.

Now consider the interference signal $I_i(t)$. We have

$$I_i(t) = \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} [s_j(t)s_{c_j}(t) * h_i(t)]s_{c_i}(t)dt = \alpha_i \sum_{\substack{j=1 \\ j\neq i}}^{K} d_{jl} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_j}(t)s_{c_i}(t)dt = \alpha_i \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{d_{jl}}{\sqrt{G}} = I_{il}, \tag{13.16}$$

where $\sqrt{G}$ is defined by the cross correlation (13.11). We see that both the interference $I_i(t)$ and signal $\hat{s}_i(t)$ are attenuated by the path gain α_i, and therefore this path gain has no impact of the receiver signal-to-interference power ratio (SIR)[4]. In particular, for a transmitted signal power of S the received signal power at the ith receiver is $\alpha_i^2 S$ and the interference power is $I = \alpha_i^2 S(K-1)/G$, so SIR $= G/(K-1)$. Typically semi-orthogonal code systems are designed for a large number of users, so that $K-1 >> N_0B/S$. Thus we can typically neglect noise in our performance analysis, since the MAC interference power is much bigger than the noise power. In this case we say that these systems are interference-limited.

Now consider a more general channel $h_i(t) = \sum_{n=1}^{N} \alpha_{in}\delta(t-\tau_{in})$. First consider the desired signal component:

$$\begin{aligned} \hat{s}_i(t) &= \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} [s_i(t)s_{c_i}(t) * h_i(t) + n(t)]s_{c_i}(t)dt \\ &= \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} \left(\sum_{n=1}^{N} \alpha_{in}s_i(t-\tau_{in})s_{c_i}(t-\tau_{in})s_{c_i}(t) + n(t)s_{c_i}(t) \right) dt \end{aligned}$$

[4]The path gain does impact SNR, but noise is typically neglected in analysis of spread spectrum MAC systems, since they tend to be interference-limited.

$$
\begin{aligned}
&= \sum_{n=1}^{N} \alpha_{in} d_{il} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_i}(t-\tau_{in}) s_{c_i}(t) dt + n_{il} \\
&= \sum_{n=1}^{N} \alpha_{in} d_{il} \rho(\tau_{in}) + n_{il} \qquad (13.17)
\end{aligned}
$$

where n_{il} is the same as before and we assume $\tau_{in} < T_b$ so that $s_i(t-\tau_{in}) = d_{il}$ over the lth bit interval (For $\tau_{in} > T_b$, $s_i(t-\tau_{in}) \neq d_{il}$ so we get ISI between between bits). Thus, in the absence of other users, the demodulated bit has fading due to the multipath and the code autocorrelation, as in the single-user case.

Now consider the interference signal $I_i(t)$ for this more general channel. We have

$$
\begin{aligned}
I_i(t) &= \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} [s_j(t) s_{c_j}(t) * h_i(t)] s_{c_i}(t) dt \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_i}(t) \left(\sum_{n=1}^{N} \alpha_{in} s_{c_j}(t-\tau_{in}) s_j(t-\tau_{in}) \right) dt \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \sum_{n=1}^{N} \alpha_{in} s_j(t-\tau_{in}) \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_i}(t) s_{c_j}(t-\tau_{in}) dt \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \sum_{n=1}^{N} \alpha_{in} d_{jl} \rho_{ij}(\tau_{in}) \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \sum_{n=1}^{N} \frac{\alpha_{in} d_{jl}}{\sqrt{G}} \\
&= I_{il}, \qquad (13.18)
\end{aligned}
$$

where we again assume $\tau_{in} < T_b$, so $s_j(t-\tau_{in}) = d_{jl}$ over the lth bit interval. We see that the interference also experiences fading due to multipath, and the interference power on each path is reduced by the code cross correlation.

13.5.3 Multiple Access Channels

We now consider a multiple access channel (MAC) with semi-orthogonal codes. In a MAC channel each user's signal is generated by the individual user. These signals pass through the user's individual channel and are then summed together at the receiver, along with AWGN. Since each signal goes through a different channel, the received interference power can be much larger than the received signal power, as we now show in more detail.

The transmitter and channel for each individual user in a K-user MAC is shown in Figure 13.9. We assume that the users are synchronized, so that the lth bit interval is the same for all users (the asynchronous MAC is more complex to analyze, and generally has worse performance than the synchronous MAC). We see from Figure 13.9 that the ith user generates an information signal $s_i(t)$. As in the broadcast model above we assume binary linear modulation, a baseband system, and rectangular pulses for $s_i(t)$, but these assumptions can be generalized without significantly changing the analsys. The ith user

multiplies his information signal by his spreading code $s_{c_i}(t)$ before sending it over his channel, which has impulse response $h_i(t)$. All the user's signals are summed at the receiver front end and corrupted by AWGN $n(t)$.

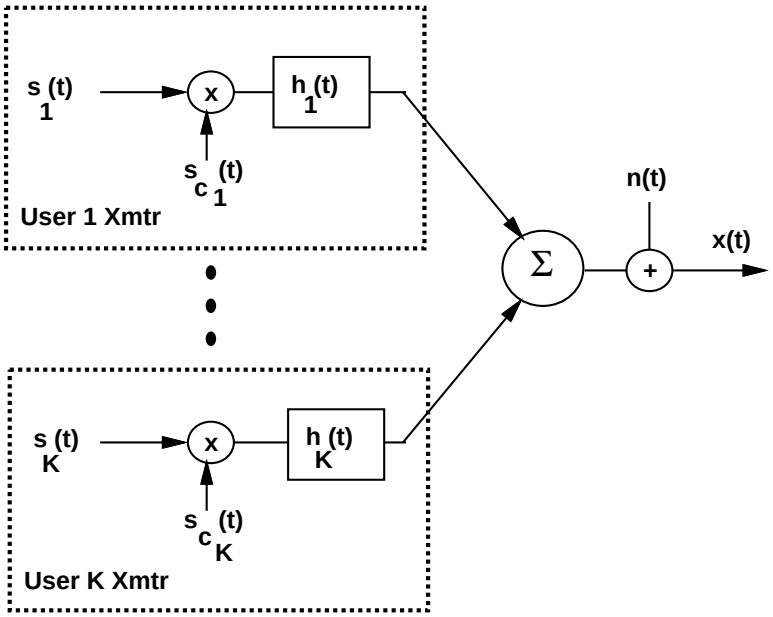


Figure 13.9: MAC Transmitter and Channel.

The signal entering the MAC receiver is given by

$$x(t) = \left[\sum_{i=1}^{K} s_i(t)s_{c_i}(t) * h_i(t)\right] + n(t). \tag{13.19}$$

The receiver consists of K branches corresponding to the K received signals, as shown in Figure 13.10. The ith branch multiplies the received signal by the ith user's spreading code $s_{c_i}(t)$, which is perfectly synchronized to the corresponding code in the ith user's transmitter. The signal is then integrated over a bit time (symbol time for nonbinary modulation). The output of the integrator is

$$\hat{s}_i(t) + I_i(t) = \hat{d}_{il} + I_{il}, \tag{13.20}$$

where $\hat{d}_{il}$ is the demodulated bit from the ith user at time l and I_{il} is the interference from other users over this bit time.

Assume that each user has a channel with no multipath, so $h_i(t) = \alpha_i\delta(t)$. The desired signal component $\hat{s}_i(t)$ is

$$\hat{s}_i(t) = \frac{1}{T_b}\int_{(l-1)T_b}^{lT_b} [s_i(t)s_{c_i}(t) * h_i(t) + n(t)]s_{c_i}(t)dt = \alpha_i d_{il} + n_{il} = \hat{d}_{il}, \tag{13.21}$$

where n_{il} is the noise sample at the output of the integrator at time l. Thus, in the absence of other users, the demodulated bit differs from the original bit due to signal attenuation α_i and noise, as in a single-user and broadcast channels.

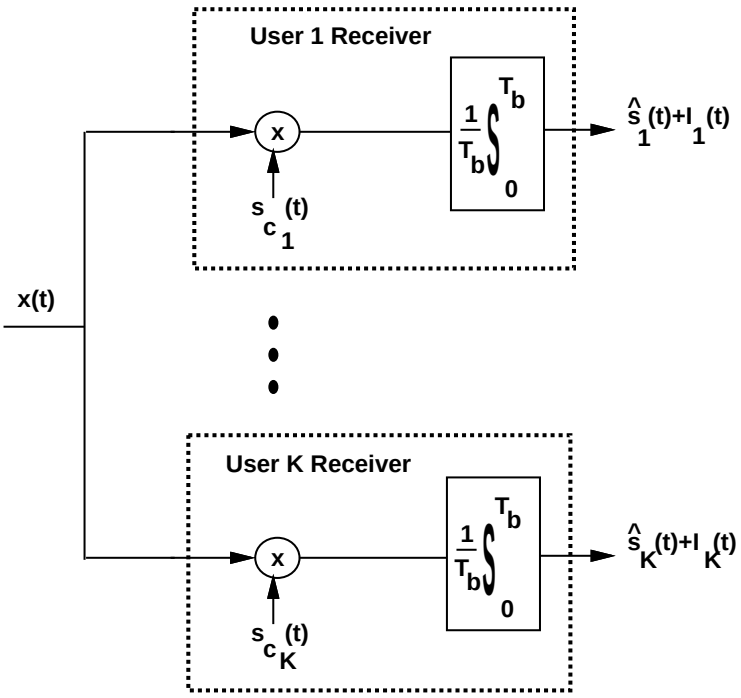


Figure 13.10: MAC Receiver.

Now consider the interference signal $I_i(t)$. We have

$$I_i(t) = \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} [s_j(t)s_{c_j}(t) * h_j(t)]s_{c_i}(t)dt = \sum_{\substack{j=1 \\ j\neq i}}^{K} \alpha_j d_{jl} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_j}(t)s_{c_i}(t)dt = \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{\alpha_j d_{jl}}{\sqrt{G}}, \quad (13.22)$$

where α_j is the path gain of the jth user's channel. We see that, in contrast to the broadcast channel, the interference $I_i(t)$ and signal $\hat{s}_i(t)$ are attenuated by different path gains. Therefore, if $\alpha_j >> \alpha_i \forall j \neq i$, the MAC interference can be quite large. In particular, if $\alpha_j = \alpha >> \alpha_i$ for $j \neq i$ we get that for transmitted signal power S the received signal power on the ith branch is $S_i = \alpha_i^2 S$ and the received interference power on this branch is $I_i = \alpha^2(K-1)/G$ leading to an SIR of

$$\frac{S_i}{I_i} = \frac{G\alpha_i^2}{(K-1)\alpha^2} << \frac{G}{K-1}. \quad (13.23)$$

Since these systems are designed such that $\frac{G}{K-1}$ equals the required SIR for acceptable performance, we see that MAC channel performance can be significantly degraded by path loss.

The solution to this problem is to use power control based on channel inversion, where each user transmits signal power S/α_i^2 so that his received signal power is S, regardless of his path loss. This will lead to an SIR of $G/(K-1)$ for each user. The disadvantage of this form of power control is that channel inversion leads to poor channel capacity and can also cause significant interference to adjacent cells in a cellular system. Despite these problems, channel inversion is used on the mobile-to-base connection in the IS-95 cellular system standard.

For more general channels $h_i(t) = \sum_{n=1}^{N} \alpha_{in}\delta(t-\tau_{in})$ the desired signal component, assuming $\tau_{in} < T_b$,

is the same as in the case of the broadcast channel:

$$\hat{s}_i(t) = \sum_{n=1}^{N} \alpha_{in} d_{il} \rho(\tau_{in}) + n_{il} \tag{13.24}$$

Thus, in the absence of other users, the demodulated bit has fading due to the multipath and the code autocorrelation, as in the single-user and broadcast cases.

Now consider the interference signal $I_i(t)$ for this more general channel, assuming $\tau_{jn} < T_b$. We have

$$\begin{aligned}
I_i(t) &= \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} [s_j(t) s_{c_j}(t) * h_j(t)] s_{c_i}(t) dt \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_i}(t) \left(\sum_{n=1}^{N} \alpha_{jn} s_{c_j}(t-\tau_{jn}) s_j(t-\tau_{jn}) \right) dt \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \sum_{n=1}^{N} \alpha_{jn} s_j(t-\tau_{jn}) \frac{1}{T_b} \int_{(l-1)T_b}^{lT_b} s_{c_i}(t) s_{c_j}(t-\tau_{jn}) dt \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \sum_{n=1}^{N} \alpha_{jn} d_{jl} \rho_{ij}(\tau_{jn}) \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{K} \sum_{n=1}^{N} \frac{\alpha_{jn} d_{jl}}{\sqrt{G}}.
\end{aligned} \tag{13.25}$$

We see that the interference also experiences fading due to multipath, with path gains that are different from those on the desired signal component. The interference power on each path is reduced by the code cross correlation.

13.5.4 Multiuser Detection

Interference signals in spread spectrum multiple access need not be treated as noise. If the spreading code of the interference signal is known, then that signal can be detected and subtracted out. Alternatively, the received signal can be projected onto a subspace that is orthogonal to the direction of the interfering signal. Multiuser detection is an active field of research in the spread spectrum area. The optimal multiuser detector was obtained by Verdu based on the Viterbi algorithm. Unfortunately, this algorithm has exponential complexity in the number of users, and also requires channel knowledge. Simpler multiuser detection algorithms include the decorrelating detector and the successive interference canceller. A good overview of multiuser detection can be found in [5] and a more systematic treatment in [6]. Recent work in this area for ISI channels is described in [7]

13.6 Frequency-Hopping

Frequency-hopping uses the same idea of bandwidth spreading and diversity as in direct sequence, however the diversity is in frequency rather than in time. Specifically, if the data signal has bandwidth B, and bandwidth B_t is available for signal transmission, then B_t is divided into equally spaced channels of bandwidth B with corresponding center frequency f_i. The data signal is then transmitted, or

hopped, over the different carrier frequencies, with the carrier frequency changing every T_c seconds. The sequence of carrier frequencies is determined by a pseudorandom sequence. The receiver must lock to the pseudorandom sequence, then use a frequency synthesizer to demodulate the narrowband signal at the appropriate carrier. Frequency-hopping has the same interference rejection capability as direct-sequence, since the signal only occupies the interference bandwidth for a short period of time. It is also resistant to frequency-selective fading, since the signal hops over many frequency bands, thus the channels with deep nulls are averaged out. A more detailed description of frequency-hopping and its fading and interference immunity can be found in [3, 8]. Slow frequency-hopping is also used in cellular systems to average out interference from other cells. Frequency-hopping has some benefits over direct-sequence in multiuser systems, and it can also be easily combined with narrowband signaling techniques, as in the GSM system.

Bibliography

[1] R.C. Dixon, *Spread Spectrum Systems with Commercial Applications*. 3rd Ed. New York: Wiley, 1994.

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

[3] M. K. Simon, J.K. Omura, R.A. Scholtz, B.K. Levitt, *Spread Spectrum Communications Handbook*. New York: McGraw Hill, 1994.

[4] A.J. Viterbi, *CDMA Principles of Spread Spectrum Communications*. Addison-Wesley 1995.

[5] A. Duel-Hallen, J. Holtzman, and Z. Zvonar, "Multiuser detection for CDMA systems," *IEEE Personal Communications Magazine,* April 1995.

[6] S. Verdu, *Multiuser Detection*, Cambridge University Press, 1998.

[7] X. Wang and V. Poor, "Blind equalization and multiuser detection in dispersive CDMA channels," *IEEE Transactions on Communications*, Jan. 1998

[8] A.A.A. Saleh, A.J. Rustako, L.J. Cimini, G.J. Owens, and R.S. Roman, "An experimental TDMA indoor radio communications system using slow frequency hopping and coding," *IEEE Trans. Commun.*, Vol. 39, No. 1, pp. 152–162, Jan. 1991.

Chapter 14

Multiuser Systems

The topics covered up until now deal with communication techniques for a single user. In multiuser systems the system resources (power, bandwidth, etc.) must be divided among the multiple users. In addition, since bandwidth is a precious resource, systems can take advantage of the signal power falloff with distance to reuse bandwidth at spatially-separate locations. The concept of frequency reuse is the fundamental technology underlying cellular system design, which are dealt with in the next chapter. In this chapter we discuss channelization methods for multiuser systems (time-division, frequency-division, and code-division). We then develop the fundamental capacity limits of multiuser broadcast and multiple access channels under these different channelization methods. We also discuss random multiple access techniques (random access) and scheduling.

14.1 Multiuser Channels: Broadcast and Multiple Access

A multiuser channel refers to any channel which must be shared among multiple users. There are two different types of multiuser channels: the broadcast channel and the multiple access channel, which are illustrated in Figure 14.1. A broadcast channel has one transmitter sending to many receivers, and thus the bandwidth and power of the transmitter must be divided accordingly. Examples of broadcast channels include all radio and television transmissions, the downlink (satellite-to-earth station) of a satellite system, and the base station-to-mobile transmission of a cellular system. A multiple access channel has many transmitters sending signals to one receiver. The transmit power for each of the transmitters may vary, but the receiver bandwidth must be divided among the different users. Examples of multiple access channels include an Ethernet connected to many computers, standard telephone lines (which are time multiplexed between many voice signals), and the mobile-to-base station transmission of cellular systems. The goal of multiuser communications is to utilize the limited system resources (power and bandwidth) in an efficient manner while creating minimal (no) interference between users.

The most important system resource to be divided is the signal bandwidth, since the bandwidth is assigned by the FCC, and is usually scarce (or expensive). The bandwidth is typically divided into channels using a channelization method based on time, frequency, or code division, discussed in more detail below. When dedicated channels are allocated to users it is often called *multiple access*: since voice signals require a dedicated channel, multiple access is the common model for channel allocation in telephony systems. Bandwidth sharing for users with bursty transmissions generally use some form of random channel allocation which does not guarantee channel access. Bandwidth sharing using random channel allocation is called random multiple access or simply random access. In general, the choice of whether to use multiple access or random access, and which channelizaton technique to use for each access

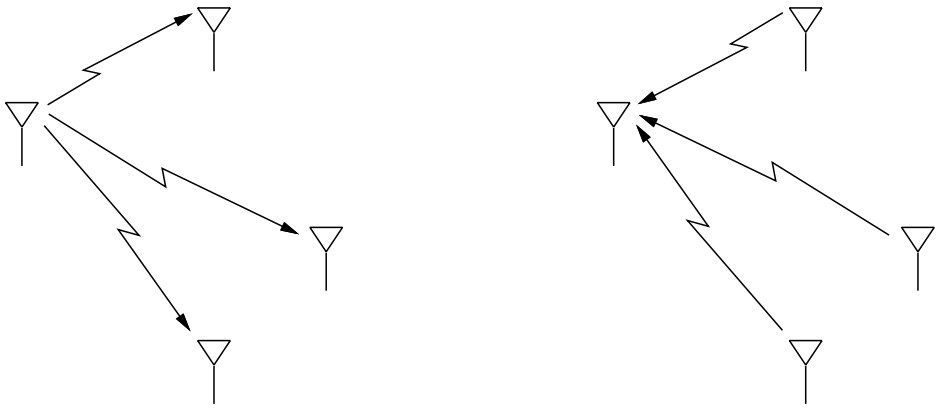


Figure 14.1: Broadcast and Multiple Access Channels.

type, will depend on the traffic characteristics of the system, the state of current access technology, and compatibility with other systems.

14.2 Multiple Access

Multiple access techniques use a channelization method to divide up the system resources and then assign dedicated channels to users that access the system. The channels are created by an orthogonal or semi-orthogonal division of the system resources. Channelization methods for multiple access can be applied to either broadcast or multiple access channels, although the relative performance of the different bandwidth division techniques are different for these two channel types. Methods to divide the spectrum include frequency-division, time-division, code-division, and hybrid combinations of these methods. Code division can use either orthogonal or semiorthogonal coding techniques. Theoretically, time-division frequency-division, and orthogonal code division are equivalent for homogeneous users (equal power and bandwidth requirements), since they all divide up the signal space orthogonally [1]. Thus, as we will see in Sections 14.3 and 14.4, the capacity region for all of these techniques is the same. These sections also demonstrates that code division using semiorthogonal codes achieves higher user rates if multiuser detection is used. However, if multiuser detection is not used, the achievable rates are lower than that of the other orthogonal techniques. We now describe each of these division techniques in more detail. More thorough treatments of channelization methods can be found in [2, 3] and the references therein.

14.2.1 Frequency Division

In frequency division, the bandwidth is divided into nonoverlapping channels. Each user is then assigned a different channel for transmission and reception. In time division, time is divided into orthogonal time slots, which are then allocated to the different users. The channels typically have a guard band between then to compensate for imperfect filters, adjacent channel interference, and spectral spreading due to Doppler. In frequency division the user channels are typically narrowband, so they experience flat fading and do not require compensation techniques for ISI. Frequency-division is generally the simplest division technique to implement, however it is rather inflexible. In particular, it is difficult to allocate multiple

channels on demand to a single user, since it requires simultaneous demodulation of multiple narrowband channels. Frequency division is used in most analog cellular systems, and is part of the standard. In some sense, all systems can be considered to use frequency division, since they are only allocated a finite amount of bandwidth.

14.2.2 Time-Division

In time-division, time is divided into orthogonal time slots, and each user is assigned a cyclically-repeating timeslot for transmission, and another one for reception. Time-division is often combined with frequency division, as in the GSM standard. Since each user occupies a cyclically-repeating timeslot, transmission is not continuous. Therefore, digital transmission techniques which allow for buffering are required. The fact that transmission is not continuous makes handoff simpler, since the handoff can be done during the timeslots occupied by other users. In addition, the channel can be sensed by the transmitter during idle times, which allows the mobile to assist in determining which base station it should be handed off to. One difficulty of using time-division in the multiple access channel is synchronization. Since different users in the multiple access channel have different time delays, the timeslots must be synchronized so that they remain orthogonal *after* these respective delays. The user channels associated with a time-division system can be wideband or narrowband depending on the total system bandwidth and whether or not frequency-division is also used. If the channels are wideband, then typically some form of ISI compensation is required. For example, in the IS-54 standard, the channels are roughly 30KHz, and no equalization is used. Conversely, in GSM, the channels are roughly 200KHz, and equalization is required for acceptable performance.

14.2.3 Code-Division

In code-division, time and bandwidth are used simultaneously by different users, modulated by orthogonal or semi-orthogonal codes. The receiver then uses the code structure to separate out the different users. One of the big advantages of spread spectrum is that little dynamic coordination of users in time or frequency is required, since the users can be separated by the code properties alone. In addition, since time and frequency division carve up time and bandwidth in N orthogonal pieces, there is a hard limit of N on how many users can simultaneously occupy the system. This is also true for code-division using orthogonal codes. However, if semi-orthogonal codes are used, the number of users is *interference limited*. Specifically, there is no hard limit on how many users can simultaneously share the channel. However, because semi-orthogonal codes can cause mutual interference to other users sharing the same bandwidth, the more users that are packed into the same channel, the higher the level of interference, which degrades the performance of all the users. Moreover, on multiple access channels, a semiorthogonal code-division system requires *power control* to compensate for the near-far problem. The near-far problem arises because users modulating their signal with different spreading codes interfere with each other. Suppose that one user is very close to his base station, and another user very far away. If both users transmit at the same power level, then the interference from the close user will swamp the signal from the far user. Thus, power control is used on all users such that their received signal powers are roughly the same. This form of power control, which essentially inverts any attenuation and/or fading on the channel, causes each interferer to contribute an equal amount of power, thereby eliminating the near-far problem.

Code-division is the most complex bandwidth division technique due to code synchronization requirements and power control. Code division for direct-sequence spread spectrum was discussed in the previous chapter. Although code-division with multiuser detection does not require power control, receiver complexity is significantly increased when simultaneous detection of all users is required. Moreover,

the detection scheme must have a low probability of bit error, since bits that are incorrectly detected are subtracted from the signals of other users, which may cause them to be decoded in error as well. A similar feedback error problem is found in decision-feedback equalizers.

14.2.4 Standards Debate

Commercially, the primary competing standards in the U.S. for cellular and PCS multiple access are frequency-division (IS-54), spread spectrum code-division (IS-95), or a combination of time- division and slow frequency hopping (GSM). The spread spectrum systems do not currently use multiuser detection, and therefore they require stringent power control to maintain a constant received signal power at the receiver for each of the users. If this constant power is not maintained, co-channel interference from strong signals will degrade the quality of other signals (the near-far problem). Stringent power control is difficult to maintain in a fading environment, and is one of the major challenges of spread spectrum multiple access.

The debate among cellular and personal communication standards committees and equipment providers over which approach to use has led to countless analytical studies claiming superiority of one technique over the other [4, 5, 6]. In many cases the a priori assumptions used in these analyses bias the results in favor of one technique over the other alternatives; usually the technique that is of some economic interest to the authors of the study.

This debate about multiple access was primarily for voice systems. Data and mixed media systems pose different challenges, and analysis based on multiple access for voice systems is not necessarily valid for data and multimedia systems. The best multiple or random access technique in this case will depend on the traffic statistics.

14.3 Broadcast Channel Capacity Region

When several users share the same channel, the channel capacity can no longer be characterized by a single number. At the extreme, if only one user occupies the channel then the single-user capacity results of the previous section apply. However, since there is an infinite number of ways to "divide" the channel between many users, the multiuser channel capacity is characterized by a *rate region*, where each point in the region is a vector of achievable rates that can be maintained by all the users simultaneously. The union of all achievable rate vectors is called the *capacity region* of the multiuser system.

In this section we analyze the capacity region of a broadcast channel with AWGN and with fading. We begin by first reviewing results from [**?**] for the AWGN broadcast channel capacity region using superposition code-division with successive interference cancellation, time-division, and frequency-division. We then extend the code-division analysis to direct sequence spread spectrum for both orthogonal and nonorthogonal codes, and obtain the corresponding capacity regions both with and without interference cancellation. We then extend these results to fading broadcast channels.

We will see that the maximum capacity region is achieved using superposition code-division with interference cancellation. In addition, spread spectrum code division with successive interference cancellation has a capacity penalty relative to superposition coding which increases with spreading gain. Finally, spread spectrum with orthogonal code division can achieve a subset of the time-division and frequency-division capacity regions, but spread spectrum with nonorthogonal coding and no interference cancellation is inferior to all the other spectrum-sharing techniques.

14.3.1 The AWGN Broadcast Channel Model

The broadcast channel consists of one transmitter sending *independent* information to different receivers over a common channel. Thus, it does not model a typical FM radio or TV broadcast channel, where the same signal is received by all users. The capacity region of the broadcast channel characterizes the rates at which information can be conveyed to the different receivers simultaneously. We only consider capacity regions for the two-user broadcast channel, since the general properties and the relative performance of the different spectrum-sharing techniques are the same when the number of users is increased [**?**]. That is because the transmit distribution for each user which achieves the multiuser capacity region is Gaussian [**?**], so interference from other users is accurately modeled as Gaussian noise even for a small number of interferers.

We will use the following notation. The two-user broadcast channel has one transmitter and two distant receivers receiving data at rate R_i, $i = 1, 2$. Each receiver has front-end AWGN of noise density n_i, $i = 1, 2$, and we arbitrarily assume $n_1 \leq n_2$. We denote the transmitter's total average power and bandwidth by S and B, respectively.

If the transmitter allocates all the power and bandwidth to one of the users, then clearly the other user will have a rate of zero. Therefore, the set of simultaneously achievable rates (R_1, R_2) includes the pairs $(C_1, 0)$ and $(0, C_2)$, where

$$C_i = B \log \left[1 + \frac{S}{n_i B} \right]. \tag{14.1}$$

These two points bound the broadcast capacity region. We now consider rate pairs in the interior of the region, which are achieved using more equitable methods of dividing the system resources.

14.3.2 Capacity Region in AWGN under TD, FD, and CD

In time-division, the transmit power S and bandwidth B are allocated to user 1 for a fraction τ of the total transmission time, and then to user 2 for the remainder of the transmission. This time-division scheme achieves a straight line between the points C_1 and C_2, corresponding to the rate pairs

$$\left\{ \bigcup (R_1 = \tau C_1, R_2 = (1 - \tau) C_2) ; \quad 0 \leq \tau \leq 1 \right\}. \tag{14.2}$$

This equal-power time-division capacity region is illustrated in Figures 14.3 and 14.4. In these figures, $n_1 B$ and $n_2 B$ differ by 3dB and 20dB, respectively. This dB difference is a crucial parameter in comparing the relative capacities of the different spectrum-sharing techniques, as we discuss in more detail below.

If we also vary the average transmit power of each user then we can achieve a larger capacity region. Let S_1 and S_2 denote the average power allocated to users 1 and 2 over their respective time slots. The average power constraint then becomes $\tau S_1 + (1 - \tau) S_2 = S$. The capacity region with this power allocation is then

$$\left\{ \bigcup \left(R_1 = \tau B \log \left[1 + \frac{S_1}{n_1 B} \right], R_2 = (1 - \tau) B \log \left[1 + \frac{S_2}{n_2 B} \right] \right) ; \quad \tau S_1 + (1 - \tau) S_2 = S, \quad 0 \leq \tau \leq 1 \right\}. \tag{14.3}$$

We will see in the following section that the rate region defined by (14.3) is the same as the frequency-division capacity region.

In frequency-division the transmitter allocates S_i of its total power S and B_i of its total bandwidth B to user i. The power and bandwidth constraints require that $S_1 + S_2 = S$ and $B_1 + B_2 = B$. The set of achievable rates for a fixed frequency division (B_1, B_2) is thus

$$\left\{ \bigcup \left(R_1 = B_1 \log \left[1 + \frac{S_1}{n_1 B_1} \right], R_2 = B_2 \log \left[1 + \frac{S_2}{n_2 B_2} \right] \right) ; \quad S_1 + S_2 = S \right\}. \tag{14.4}$$

It was shown by Bergmans [?] that, for n_1 strictly less than n_2 and any fixed frequency division (B_1, B_2), there exists a range of power allocations $\{S_1, S_2 : S_1 + S_2 = S\}$ whose corresponding rate pairs exceed a segment of the equal-power time-division line (14.2).

The frequency-division rate region is defined as the union of fixed frequency-division rate regions (14.4) over all bandwidth divisions:

$$\left\{\bigcup\left(R_1 = B_1 \log\left[1 + \frac{S_1}{n_1 B_1}\right], R_2 = B_2 \log\left[1 + \frac{S_2}{n_2 B_2}\right]\right); \quad S_1 + S_2 = S, \quad B_1 + B_2 = B\right\}. \tag{14.5}$$

It was shown in [?] that this capacity region exceeds the equal-power time-division rate region (14.2). This superiority is indicated by interpolating the fixed frequency-division regions in Figures 14.3 and 14.4, although it is difficult to see in Figure 14.3, where the users have a similar received SNR. In fact, when $n_1 = n_2$, (14.5) reduces to (14.2) [?]. Thus, optimal power and/or frequency allocation is more beneficial when the users have very disparate channel quality.

Note that the rate region for time-division with unequal power allocation given by (14.3) is the same as the frequency-division rate region (14.5). This is seen by letting $B_i = \tau_i B$ and $\sigma_i = \tau_i S_i$ in (14.3), where $\tau_1 = \tau$ and $\tau_2 = 1 - \tau$. The power constraint then becomes $\sigma_1 + \sigma_2 = S$. Making these substitutions in (14.3) yields

$$\left\{\bigcup\left(R_1 = B_1 \log\left[1 + \frac{\sigma_1}{n_1 B_1}\right], R_2 = B_2 \log\left[1 + \frac{\sigma_2}{n_2 B_2}\right]\right); \quad \sigma_1 + \sigma_2 = S\right\}. \tag{14.6}$$

Comparing this with (14.4) we see that with appropriate choice of S_i and τ_i, any point in the frequency-division rate region can also be achieved through time-division with unequal power allocation.

Superposition coding with successive interference cancellation, described in more detail in [?], is a multiresolution coding technique whereby the user with the more favorable channel can distinguish the fine resolution of the received signal constellation, while the user with the worse channel can only distinguish the constellation's coarse resolution. An example of a two-level superposition code constellation taken from [?] is 32-QAM with embedded 4-PSK, as shown in Figure 14.2. In this example, the transmitted constellation point is one of the 32-QAM signal points chosen as follows. The user with the worse SNR provides 2 bits to select one of the 4-PSK superpoints. The user with the better SNR provides 3 bits to select one of the 8 constellation points surrounding the selected superpoint. After transmission through the channel, the user with the better SNR can easily distinguish the quadrant in which the constellation point lies. Thus, the 4-PSK superpoint is effectively subtracted out by this user. However, the user with the worse channel cannot distinguish between the 32-QAM points around its 4-PSK superpoints. Thus, the 32-QAM modulation superimposed on the 4-PSK modulation appears as noise to this user. These ideas can be easily extended to multiple users using more complex signal constellations. Since superposition coding achieves multiple rates by expanding its signal constellation, it does not typically require bandwidth expansion.

The two-user capacity region using superposition coding and successive interference cancellation was derived in [?] to be the set of rate pairs

$$\left\{\bigcup\left(R_1 = B \log\left[1 + \frac{S_1}{n_1 B}\right], R_2 = B \log\left[1 + \frac{S_2}{n_2 B + S_1}\right]\right); \quad S_1 + S_2 = S\right\}. \tag{14.7}$$

The intuitive explanation for (14.7) is the same as for the example discussed above. Since $n_1 < n_2$, user 1 correctly receives all the data transmitted to user 2. Therefore, user 1 can decode and subtract out user 2's message, then decode its own message. User 2 cannot decode the message intended for user 1, since it has a less-favorable channel; thus, user 1's message, with power S_1, contributes an additional noise

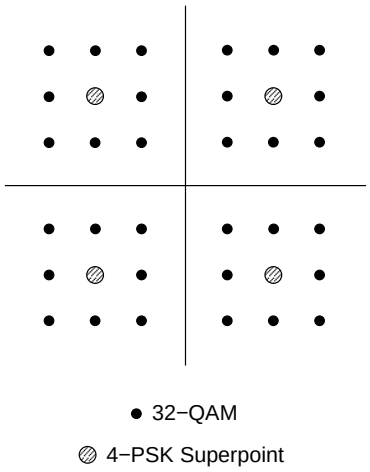


Figure 14.2: 32-QAM with embedded 4-PSK

term to user 2's received message. This same process is used by the successive interference canceller in spread spectrum systems [**?**]. However, it is important to mention that although successive interference cancellation achieves the capacity region (14.7), it is not the best method to use in practice. The capacity analysis assumes perfect signal decoding, whereas real systems exhibit some decoding error. This error leads to decision-feedback errors in the successive interference cancellation scheme. Thus, cancellation methods which mitigate the effect of decision errors work better in practice than successive cancellation.

The rate region defined by (14.7) was shown in [**?**] to exceed the regions achievable through either time- or frequency-division, when $n_1 < n_2$. Moreover, it was also shown in [**?**] that this is the maximum achievable set of rate pairs for any type of coding and spectrum sharing, and thus (14.7) defines the capacity region. However, if the users all have the same SNR, then this capacity region collapses to the equal-power time-division line (14.2). Thus, when $n_1 = n_2$, all the spectrum-sharing methods have the same rate region.

Code-division can also be implemented using direct-sequence spread spectrum, as discussed in Chapter 12.5. As we discussed there, spread spectrum multiplies the modulated data signal by a spreading code, which increases the transmit signal bandwidth by a factor G called the spreading gain. For orthogonal spreading codes, the cross correlation between the respective codes is zero, and these codes require a spreading gain of N to produce N orthogonal codes. For a total bandwidth constraint B, the information bandwidth of each user's signal with these spreading codes is thus limited to B/N. The two-user rate rate region with these spreading codes is then

$$\left\{ \bigcup \left(R_1 = \frac{B}{2} \log \left[1 + \frac{S_1}{n_1 B/2} \right], R_2 = \frac{B}{2} \log \left[1 + \frac{S_2}{n_1 B/2} \right] \right); \quad S_1 + S_2 = S \right\}. \tag{14.8}$$

Comparing (14.8) with (14.4) we see that code-division with orthogonal coding is the same as fixed frequency-division with the bandwidth equally divided ($B_1 = B_2 = B/2$). From (14.6), time-division with unequal power allocation can also achieve all points in this capacity region. Thus, orthogonal code-division with Hadamard-Walsh functions achieves a subset of the time-division and frequency-division capacity regions. More general orthogonal codes are needed to achieve the same region as these other techniques.

We now consider spread spectrum with nonorthogonal spreading codes. As discussed in the previous chapter, these codes are commonly generated using maximal length shift registers, which yield a code

cross correlation of approximately $1/G$. Thus, interference between users is attenuated by a factor of G. Since the signal bandwidth is also increased by this factor, the two-user rate region achievable through spread-spectrum using maximal length spreading codes and successive interference cancellation is given by

$$\left\{\bigcup\left(R_1 = \frac{B}{G}\log\left[1+\frac{S_1}{n_1 B/G}\right], R_2 = \frac{B}{G}\log\left[1+\frac{S_2}{n_2 B/G + S_1/G}\right]\right);\ \ S_1+S_2=S\right\}. \tag{14.9}$$

By the convexity of the log function, the rate region defined by (14.9) for $G > 1$ is smaller than the rate region (14.7) obtained using superposition coding, and the degradation increases with increasing values of G. This implies that for nonorthogonal coding, the spreading gain should be minimized in order to maximize capacity.

With maximal length spreading coding and no interference cancellation, the receiver treats all signals intended for other users as noise, resulting in the rate region

$$\left\{\bigcup\left(R_1 = \frac{B}{G}\log\left[1+\frac{S_1}{n_1 B/G + S_2/G}\right], R_2 = \frac{B}{G}\log\left[1+\frac{S_2}{n_2 B/G + S_1/G}\right]\right);\ \ S_1+S_2=S\right\}. \tag{14.10}$$

Again using the log function convexity, $G = 1$ maximizes this rate region, and the rate region decreases as G increases. Moreover, by taking the second partial derivatives in (14.10), we get that for any $G \geq 1$,

$$\frac{\partial^2 R_2}{\partial^2 R_1} = \frac{\partial R_1}{\partial \alpha_1}\frac{\partial^2 R_2}{\partial^2 \alpha_1} - \frac{\partial R_2}{\partial \alpha_1}\frac{\partial^2 R_1}{\partial^2 \alpha_1} \geq 0. \tag{14.11}$$

Thus, the rate region for nonorthogonal coding without interference cancellation (14.18) is bounded by a convex function with end points C_1 and C_2, as shown in Figures 14.3 and 14.4. Therefore, the capacity region for nonorthogonal code-division without interference cancellation will lie beneath the regions for time-division and frequency-division, which are bounded by concave functions with the same endpoints.

While the orthogonality of time-division and frequency-division is relatively robust against small multipath delays introduced by the channel, multipath delays bigger than a chip time can compromise the orthogonality of orthogonal spread spectrum codes. This loss of orthogonality causes interference noise between users, so the rate region becomes

$$\left\{\bigcup\left(R_1 = \frac{B}{G}\log\left[1+\frac{S_1}{n_1 B/G + S_2/G'}\right], R_2 = \frac{B}{G}\log\left[1+\frac{S_2}{n_2 B/G + S_2/G'}\right]\right);\ \ S_1+S_2=S\right\}, \tag{14.12}$$

where $1/G'$ equals the code cross correlation with multipath. If $G' \approx G$ then the rate region defined by (14.12) is approximately the same as (14.10). As the multipath effect diminishes, $G' \to \infty$ and the region converges to (14.8). A deeper discussion of multipath impact on spread spectrum coding and techniques to improve orthogonality in the presence of multipath can be found in Section 12.5

The rate regions for equal-power time-division (14.2), frequency-division (14.4), orthogonal code-division (14.8), and nonorthogonal code-division with (14.7) and without (14.10) interference cancellation are illustrated in Figures 14.3 and 14.4, where the SNR between the users differs by 3dB and 20dB, respectively. For the calculation of (14.10) we assume code-division through superposition coding with $G = 1$: spread spectrum code-division with larger values of the spreading gain will result in a smaller rate region.

14.3.3 Fading Broadcast Channel Capacity

We now combine the capacity analysis in §?? for the single-user fading channel with the capacity region analysis in §?? to obtain the capacity region of the fading broadcast channel. The two-user broadcast

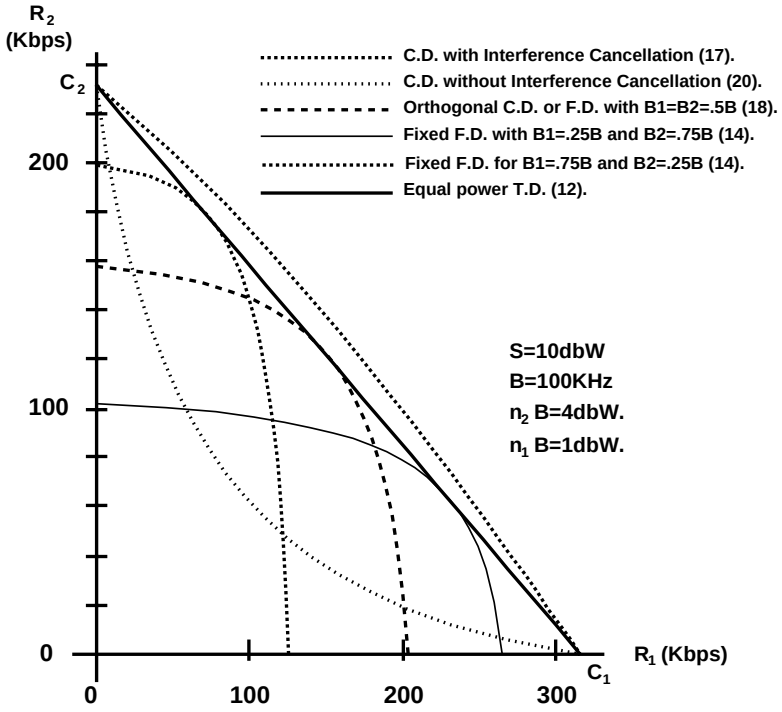


Figure 14.3: Two-User Capacity Region: 3dB SNR Difference.

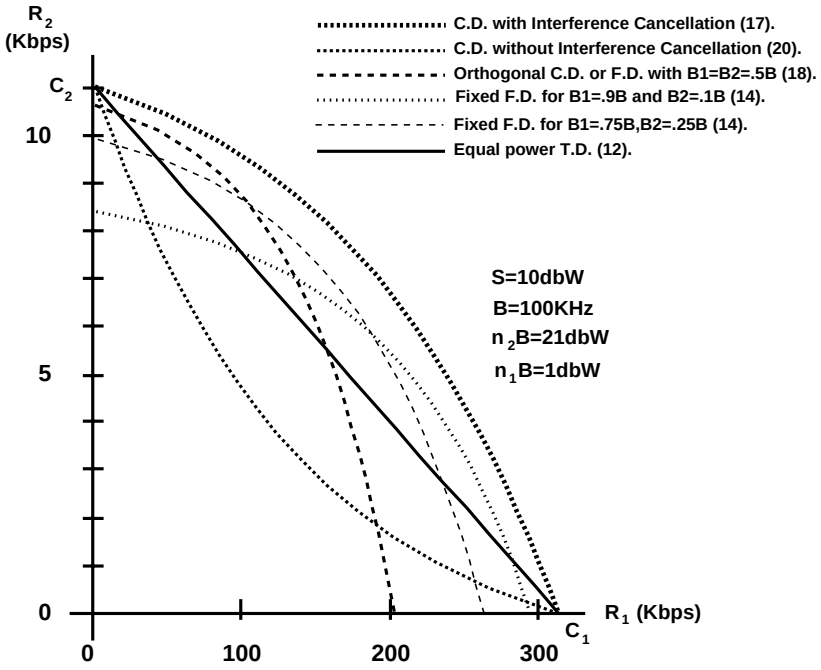


Figure 14.4: Two-User Capacity Region: 20dB SNR Difference.

channel with fading and AWGN has one transmitter with average power S and bandwidth B and two receivers with noise density N_j and time-varying received SNR $\gamma_j[i] = Sg_j[i]/(N_jB), j = 1, 2$. Let $n_j[i] = N_j/g_j[i]$, so $\gamma_j[i] = S/(n_j[i]B)$. We assume that $n_j[i]$ is known to the jth receiver at time i and that both $n_1[i]$ and $n_2[i]$ are known to the transmitter at time i. Thus, the transmitter can vary its power $S[i]$ relative to $n_1[i]$ and $n_2[i]$, subject only to the average power constraint S. For frequency-division, it can also vary the bandwidth $B_j[i]$ allocated to each user, subject to the constraint $B_1[i] + B_2[i] = B$ for all i.

ince time-division allocates orthogonal time slots to each user, the two-user channel with time-division reduces to two orthogonal time-varying single-user channels. Thus, we can apply the single-user capacity results in §**??** to each of the two channels. This yields the rate region

$$\left\{\bigcup (R_1 = \tau C_1[S, B], R_2 = (1-\tau)C_2[S, B]) ; \quad 0 \le \tau \le 1\right\}, \tag{14.13}$$

where $C_i[S, B], i = 1, 2$, is given by (**??**), (**??**), or (**??**), depending on the power adaptation strategy. Clearly, the capacity region is achieved using (**??**) for $C_i[S, B]$ with the corresponding power adaptation (**??**). If the average power allocated to each user is different, the capacity region becomes

$$\left\{\bigcup (R_1 = \tau C_1[S_1, B], R_2 = (1-\tau)C_2[S_2, B]) ; \quad \tau S_1 + (1-\tau)S_2 = S, \quad 0 \le \tau \le 1\right\}. \tag{14.14}$$

As for the AWGN channel, the unequal-power time-division rate region (14.14) is equivalent to the fixed frequency-division rate region (14.15) obtained below.

Fixed frequency division divides the total channel bandwidth B into nonoverlapping segments of width B_1 and B_2, which also reduces the two-user channel to independent single-user channels. As in the time-division case, we can thus apply the results of §**??** to each channel independently, yielding the fixed frequency-division rate region

$$\left\{\bigcup (R_1 = C[S_1, B_1], R_2 = C[S_2, B_2]) ; \quad S_1 + S_2 = S\right\}. \tag{14.15}$$

Again, $C[S_i, B_i]$ is given by (**??**), (**??**), or (**??**), with (**??**) achieving the maximum capacity region. Setting $B_1 = \tau B$ and $S_1 = \tau S$ in (14.15) and comparing with (14.14) shows the equivalence of unequal-power time-division and fixed frequency-division on the fading channel. It is clear that the equal-power time-division capacity region (14.13) will exceed the fixed frequency-division rate region over some range of power allocations $\{S_1, S_2 : S_1 + S_2 = S\}$, in particular when all of the power is allocated to one of the frequency bands. Suppose, however, that both the power and the bandwidth partition vary at each transmission based on the instantaneous noise densities $n_1[i]$ and $n_2[i]$. Clearly the resulting rate region will exceed both fixed frequency-division and time-division, which fixes the allocation of these resources over all time. The rate region for this variable power and bandwidth allocation scheme is

$$\left\{\bigcup \left(R_1 = \int_k C_{1,k}[S_{1,k}, B_{1,k}]\pi_k, R_2 = \int_k C_{2,k}[S_{2,k}, B_{2,k}]\pi_k\right) ; B_{1,k} + B_{2,k} = B, \int_k (S_{1,k} + S_{2,k})\pi_k = S\right\}, \tag{14.16}$$

where π_k denotes the joint noise density distribution $\pi_k = p(n_1[i] = n_{1,k}, n_2[i] = n_{2,k})$, $S_{j,k}$ and $B_{j,k}$ are the bandwidth and power allocated to user j when $n_j[i] = n_{j,k}$, and $C_{j,k}[S_{j,k}, B_{j,k}] = B_{j,k}\log(1 + S_{j,k}/(n_{j,k}B_{j,k}))$. To determine the boundary region of (14.16), both the power and bandwidth allocations must be optimized jointly over time, so the two users are no longer independent. Finding this boundary region requires an exhaustive search or a multidimensional optimization over time subject to the bandwidth and power constraints. We do not evaluate this region in the numerical results presented below. However, this capacity region is bounded above by the capacity region for superposition coding with

successive decoding and bounded below by the union of all fixed frequency-division regions, which are evaluated in Figure 14.5.

The idea of reallocating bandwidth and power as the channel varies is closely related to dynamic channel allocation, where channel allocation is based on the noise (and interference) levels in a particular frequency band [**?**, **?**]. The frequency allocation of (14.16) suggests that instead of using a threshold level to determine which user should occupy the channel, the channel should be allocated to the user which gets the most capacity from it. Similar ideas are currently being investigated for admission control [**?**].

We now consider code division techniques. We first study superposition coding with successive interference cancellation where, at each transmission, the signal constellation is optimized relative to the instantaneous noise densities $n_1[i]$ and $n_2[i]$. In particular, the user with the lower noise density $n_j[i]$ at time i will subtract the interference caused by the other user. The rate region is thus the average of the rate regions in AWGN weighted by the joint probability of the noise densities:

$$\left\{ \bigcup \left(R_1 = \int_k B \log \left[1 + \frac{S_{1,k}}{n_{1,k}B + S_{2,k}\mathbf{1}[n_{1,k} \geq n_{2,k}]} \right] \pi_k, R_2 = \int_k B \log \left[1 + \frac{S_{2,k}}{n_{2,k}B + S_{1,k}\mathbf{1}[n_{2,k} > n_{1,k}]} \right] \pi_k \right) ; \int_k (S_{1,k} + S_{2,k})\pi_k = S \right\}, \tag{14.17}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function ($\mathbf{1}[x] = 1$ if x is true and zero otherwise). Since superposition coding with interference cancellation has a larger rate region than time- and frequency-division on the AWGN channel, we expect this to be true for the fading channel as well. Indeed, consider any rate point in the frequency-division capacity region (14.16). Associated with that point will be a set of frequency divisions $(B_{1,k}, B_{2,k})$ and a set of transmit power values $(S_{1,k}, S_{2,k})$ corresponding to each noise pair $(n_{1,k}, n_{2,k})$. Let $S_k = S_{1,k} + S_{2,k}$. From §**??**, for the broadcast channel with noise density values $(n_{1,k}, n_{2,k})$ there exists a superposition code with total power S_k that has a larger capacity region than frequency-division. Since we can find such a dominating code for all pairs of noise density values, the weighted integral of the superposition rates over all joint noise density pairs will exceed the frequency-division capacity region of (14.16).

The rate region for superposition coding without successive decoding is given by

$$\left\{ \bigcup \left(R_1 = \int_k B \log \left[1 + \frac{S_{1,k}}{n_{1,k}B + S_{2,k}} \right] \pi_k, R_2 = \int_k B \log \left[1 + \frac{S_{2,k}}{n_{2,k}B + S_{1,k}} \right] \pi_k \right) ; \int_k (S_{1,k} + S_{2,k})\pi_k = S \right\}. \tag{14.18}$$

Since the capacity region corresponding to each k term in the integral (14.18) is bounded by a convex function, the resulting rate region will also be bounded by a convex function. Thus, both the equal-power time-division rate region (14.13) and the frequency-division rate region (14.16), which are bounded by concave functions with the same endpoints, will have larger rate regions than that of (14.18).

Obtaining the code division capacity region boundaries either with or without interference cancellation requires either an exhaustive search or a two-dimensional optimization of the power over all time. However, we can obtain a simple lower bound for these boundaries by keeping the transmit power constant. This yields a point in the capacity region which is clearly beneath rate vectors obtained with optimal power adaptation. The resulting capacity region lower bound for Rayleigh fading is shown in Figure 14.5, along with the time-division and fixed frequency-division rate regions, given by (14.13) and (14.15) respectively. From this figure we see that keeping the transmit power constant is clearly suboptimal, since the equal-power time-division rate region exceeds the region obtained by superposition

code-division with interference cancellation near the region end points. In light of this observation, it is interesting to recall our remark in §?? that keeping the transmit power constant has a negligible impact on the capacity of a single-user fading channel. We see now that the effect of power adaptation is much more pronounced in the multiuser case, where power adaptation impacts the interference on other users.

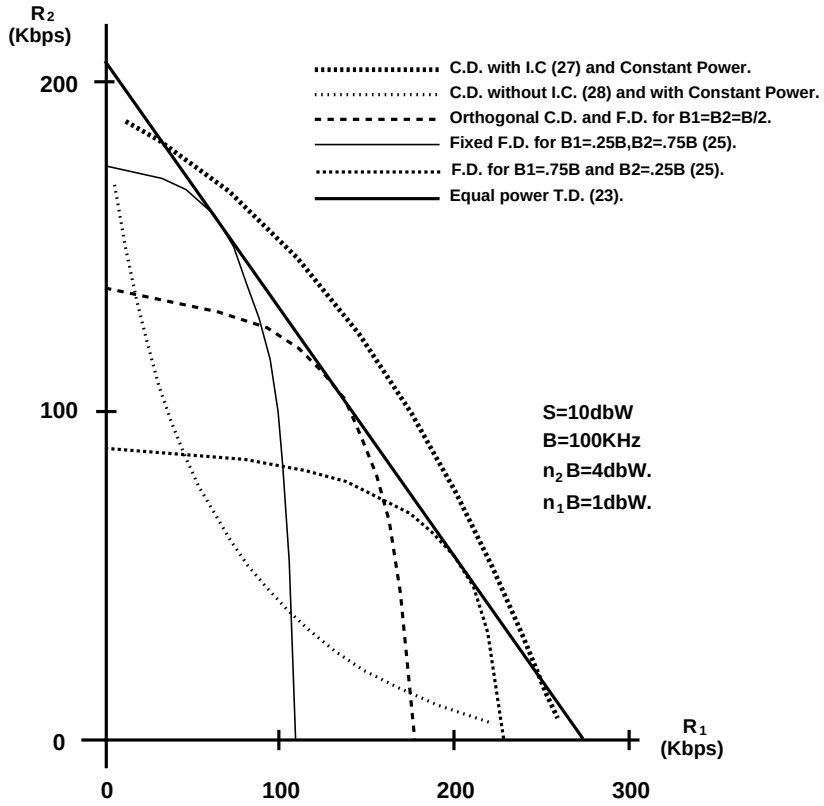


Figure 14.5: Two-User Capacity Region in Rayleigh Fading.

If we compare Figures 14.3 and 14.5 we see that fading decreases the capacity region, even with optimal power and bandwidth, timeslot, or code adaptation relative to the fading. The fact that fading reduces the capacity region is not surprising, since the single-user fading channel capacity evaluated in Chapter 4 is less than the capacity of an AWGN channel with the same average SNR.

To summarize, the time-varying capacity region is obtained by taking a weighted average of time-invariant capacity regions associated with the different noise density pairs, with the weights determined by the joint probability distribution of these pairs. Numerical evaluation of the capacity regions defined by (14.13) and (14.15) is straightforward using the methods defined in §??. These regions have the same general shape as in Figures 14.3 and 14.4, although they are smaller for the fading channel than for the AWGN channel. Evaluation of (14.16), (14.17), and (14.18) requires an exhaustive search or a difficult multidimensional optimization over time. A lower bound for (14.16), the frequency-division rate region with optimal power and bandwidth adaptation, is obtained by maximizing over all fixed frequency-division rate regions (14.15). A lower bound for the code-division rate region with optimal power and transmit constellation adaptation is obtained by keeping the transmit power $S_{1,k} = S_{2,k} = S$ constant in (14.17) and (14.18).

14.4 Multiple Access Channel Capacity Region

14.4.1 The AWGN Multiple Access Channel

The multiaccess channel consists of several transmitters, each with power P_i, sending to a receiver which is corrupted by AWGN of power n. If we denote the ith transmitted signal by X_i, then the received signal is given by $Y = \sum_{i=1}^{K} X_i + N$, where N is an AWGN sample of power n. The two-user multiaccess capacity region was determined by Cover to be the closed convex hull of all vectors (R_1, R_2) satisfying [?]

$$\begin{aligned} R_i &\leq B \log \left[1 + \frac{P_i}{nB}\right], \\ R_1 + R_2 &\leq B \log \left[1 + \frac{P_1 + P_2}{nB}\right]. \end{aligned} \tag{14.19}$$

This region is shown in Figure 14.6, where C_i and C_i^* are given by

$$C_i = B \log \left[1 + \frac{P_i}{nB}\right], \; i = 1, 2, \tag{14.20}$$

$$C_1^* = B \log \left[1 + \frac{P_1}{nB + P_2}\right], \tag{14.21}$$

and

$$C_2^* = B \log \left[1 + \frac{P_2}{nB + P_1}\right]. \tag{14.22}$$

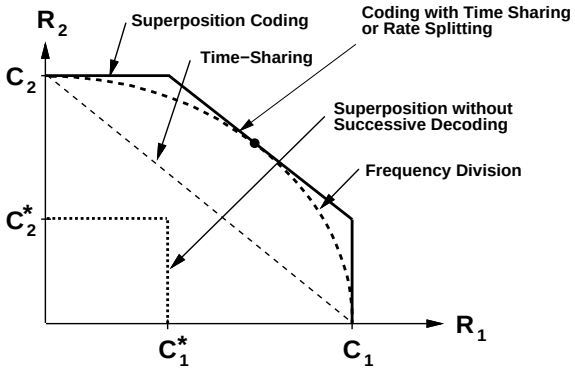


Figure 14.6: Multiaccess Channel Rate Region.

The point $(C_1, 0)$ is the achievable rate vector when transmitter 1 is sending at its maximum rate and transmitter 2 is silent, and the opposite scenario achieves the rate vector $(0, C_2)$. The corner points (C_1, C_2^*) and (C_1^*, C_2) are achieved using the successive decoding technique described above for superposition codes. Specifically, let the first user operate at the maximum data rate C_1. Then its signal will appear as noise to user 2; thus, user 2 can send data at rate C_2^* which can be decoded at the receiver with arbitrarily small error probability. If the receiver then subtracts out user 2's message from its received signal, the remaining message component is just users 1's message corrupted by noise, so rate C_1 can be achieved with arbitrarily small error probability. Hence, (C_1, C_2^*) is an achievable rate vector. A similar argument with the user roles reversed yields the rate point (C_1^*, C_2).

Time division between the two transmitters operating at their maximum rates, given by (14.20), yields any rate vector on the straight line connecting C_1 and C_2. With frequency division, the rates depend on the fraction of the total bandwidth that is allocated to each transmitter. Letting B_1 and B_2 denote the bandwidth allocated to each of the two users, we get the rate region (R_1, R_1) with

$$R_i \leq B_i \log \left[1 + \frac{P_i}{nB_i}\right]. \tag{14.23}$$

Clearly this region dominates time division, since setting $B_1 = \tau B$ and $B_2 = (1-\tau)B$ in (14.23) yields a higher rate region (R_1, R_1) than $(\tau C_1, (1-\tau)C_2)$. Varying the values of B_1 and B_2 subject to the constraint $B_1 + B_2 = B$ yields the frequency division curve shown in Figure 14.6. It can be shown [**?**] that this curve touches the rate region boundary at one point, and this point corresponds to the rate vector which maximizes the sum $R_1 + R_2$. To achieve this point, the bandwidths B_1 and B_2 must be proportional to their corresponding powers P_1 and P_2.

As with the broadcast multiuser channel, we can achieve the same rate region with time division as with frequency division by efficient use of the transmit power. If we take the constraints P_1 and P_2 to be average power constraints, then since user i only uses the channel τ_i percent of the time, its average power over that time fraction can be increased to P_i/τ_i. The rate region achievable through time division is then given by (R_1, R_2) with

$$R_i \leq \tau_i B \log \left[1 + \frac{P_i}{n\tau_i B}\right],\ i = 1, 2, \tag{14.24}$$

and substituting $B_i \stackrel{\triangle}{=} \tau_i B$ in (14.24) yields the same rate region as in (14.23).

Superposition codes without successive decoding can also be used. With this approach, each transmitter's message acts as noise to the others. Thus, the maximum achievable rate in this case cannot exceed (C_1^*, C_2^*), which is clearly dominated by frequency division for some bandwidth allocations, in particular the allocation that intersects the rate region boundary. More work is needed to determine when, if ever, this suboptimal technique achieves better rates than time or frequency division. Clearly, however, $C_1^* \to C_1$ as $R_2 \to 0$ or, equivalently $P_2 \to 0$. Similarly, $C_2^* \to C_2$ as $R_1 \to 0$. Based on this observation it is clear that the suboptimality of superposition codes without successive decoding is most pronounced when both users transmit their full power.

14.4.2 Fading Multiaccess Channels

The two-user fading multiaccess channel has two transmitters with average power P_1 and P_2, respectively, and one receiver with bandwidth B and AWGN of time-varying power $n(t)$. Let $\pi_k = p(n(t) = k)$ We also assume that transmitter i tracks $n_i(t)$, and the receiver tracks both $n_1(t)$ and $n_2(t)$. The transmitters may vary their instantaneous transmit power $P_i(t)$ relative to $n(t)$, subject only to the average power constraint $\overline{P_i(t)} = P_i$ for $i = 1, 2$.

We first consider spectrum sharing through time division. With this technique we can achieve any point $(R_1, R_2) = \int_k \pi_k \left(\tau C_k(\Phi_{k_1}), (1-\tau)C_k(\Phi_{k_2})\right)$, where

$$C_k(\Phi_{k_i}) \stackrel{\triangle}{=} B \log \left[1 + \frac{\Phi_{k_i}}{nB}\right], \tag{14.25}$$

and Φ_{k_i}, the power allocated to the ith user when $n(t) = n$, is subject to the average power constraint $\int_k \pi_k \Phi_{k_i} = P_i$. The Φ_{k_i}s can be optimized independent of each other, since under time division the two users are orthogonal. Optimizing these power allocations subject to the power constraint therefore defines

a straight line connecting the points $C_1(P_1)$ and $C_2(P_2)$, where

$$C_i(P_i) = \max_{\{\Phi_{k_i}:\int_k \pi_k \Phi_{k_i} = P_i\}} \int_k \pi_k C_k(\Phi_{k_i}). \tag{14.26}$$

Fixed frequency division partitions the total bandwidth B into nonoverlapping segments B_1 and B_2, which are then allocated to the respective transmitters. Since the bandwidths are separate, the users are independent, and they can allocate their time-varying power independently, subject only to the total power constraint P_i. The fixed frequency division rate region (R_1, R_2) thus satisfies

$$R_i \leq \max_{\Phi_{k_i}} \int \pi_k C_k(\Phi_{k_i}, B_i), \tag{14.27}$$

where

$$C_k(\Phi_{k_i}, B_i) = B_i \log \left[1 + \frac{\Phi_{k_i}}{nB_i}\right], \tag{14.28}$$

and the Φ_{n_i}s satisfy the power constraint $\int_k \pi_k \Phi_{k_i} = P_i$.

It can be shown [**?**] that fixed frequency division dominates time division. and superposition coding dominates both. Thus, as for the broadcast channel, the relative performance of the different spectrum sharing techniques is the same in AWGN and in fading, although the shape of the capacity region is different.

14.5 Random Access

Given a channelization scheme, each user can be assigned a different channel for some period of time. However, most data users do not require continuous transmission, so dedicated channel assignment can be extremely inefficient. Moreover, most systems have many more total users (active plus idle users) than channels, so at any given time channels can only be allocated to users that need them. Random access strategies are used is such systems to assign channels to the active users.

Random access techniques were pioneered by Abramson with the Aloha protocol [7]. In the ALOHA random access protocol, packets are buffered at each terminal and transmitted over a common channel to a common hub or base station. In unslotted, or "pure" Aloha, no control is imposed on the channel to synchronize transmission from the various users, and therefore the start times of packets from different users in the network can be modeled as a Poisson point process. Should two users "collide," they both wait a random amount of time before retransmitting. The goal, of course, is to prevent the users from colliding once again when they retransmit. Under these circumstances packets from different users will be transmitted with a high probability of success if there is a light to moderate amount of traffic on the network. As the traffic on the network increases the probability of a collision between packets from different users increases.

In slotted Aloha, the users are further constrained by a requirement that they only begin transmitting at the start of a time slot. The use of such time slots increases the maximum possible throughput of the channel [8], but also introduces the need for synchronization of all nodes in the network, which can entail significant overhead. Even in a slotted system, collisions occur whenever two or more users attempt transmission in the same slot. Error control coding can result in correct detection of a packet even after a collision, but if the error correction is insufficient then the packet must be retransmitted, resulting in a complete waste of the energy consumed in the original transmission. A study on design optimization between error correction and retransmission is described in [9].

The pessimistic assumption that a collision results in the loss of two packets is usually made in the analysis of an ALOHA channel. Using this assumption it is possible to show the maximum value of the throughput in an ALOHA channel is about 18% of the peak data rate. In practice such channels are usually sized to operate at about 10% of the peak data rate. Slotted ALOHA has roughly double this peak data rate due the fact that a collision only causes the loss of a single packet.

Collisions can be reduced by Carrier Sense Multiple Access (CSMA), where users sense the channel and delay transmission if they detect that another user is currently transmitting [8]. CSMA only works when all users can hear each other's transmissions, which is typically not the case in wireless systems due to the nature of wireless propagation. This gives rise to the hidden terminal problem, illustrated in Figure 14.7, where each node can hear its immediate neighbor but no other nodes in the network. In this figure both node 3 and node 5 wish to transmit to node 4. Suppose node 5 starts his transmission. Since node 3 is too far away to detect this transmission, he assumes that the channel is idle and begins his transmission, thereby causing a collision with node 5's transmission. Node 3 is said to be hidden from node 5 since it cannot detect node 5's transmission. Aloha with CSMA also creates inefficiencies in channel utilization from the exposed terminal problem, also illustrated in Figure 14.7. Suppose the exposed terminal in this figure - node 2 - wishes to send a packet to node 1 at the same time node 3 is sending to node 4. When node 2 senses the channel it will detect node 3's transmission and assume the channel is busy, even though node 3 does not interfere with the reception of node 2's transmission by node 1. Thus node 2 will not transmit to node 1 even though no collision would have occured.

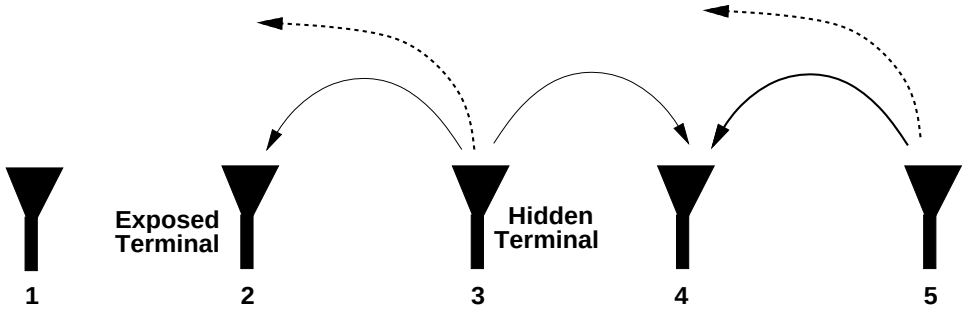


Figure 14.7: Hidden and Exposed Terminals.

The collisions introduced by hidden terminals and inefficiencies introduced by exposed terminals are often addressed by a four-way handshake prior to transmission, as in the 802.11 wireless LAN protocol [10, 11]. However, this handshake protocol is based on single hop routing, and thus its performance in multihop networks is suboptimal [12, 13]. Another technique to avoid hidden and exposed terminals is busy tone transmission. In this strategy users first check to see whether the transmit channel is busy by listening for a "busy tone" on a separate control channel [8]. There is typically not an actual busy tone but instead a bit is set in a predetermined field on the control channel. This scheme works well in preventing collisions when a centralized controller can be "heard" by users throughout the network. In a flat network without centralized control, more complicated measured are used to ensure that any potential interferer on the first channel can hear the busy tone on the second [14, 15]. Hybrid techniques using handshakes, busy tone transmission, and power control are investigated in [15]. Note that while the four-way handshake and busy tone transmission both reduce collisions due to the hidden terminal problem, they tend to aggravate the exposed terminal problem, leading to less efficient utilization of the available channels in the network. A solution to this problem is to have both transmitter and receiver send busy tones [14].

The throughput of a channel is not necessarily the most appropriate figure of merit. The throughput of a channel is simply the fraction of time during which the channel can be used to transmit data. In some cases, such as average power limited satellite channels or battery operated transmitters, the average data rate of the channel for a fixed average transmitter power and a fixed bandwidth is a more appropriate figure of merit. We can define such a figure of merit for multiple access channels, called the efficiency of the channel, which takes into account the system resources of average power and bandwidth. The efficiency of an ALOHA multiple access channel is the ratio of the ALOHA channel capacity to the capacity of the continuous channel using the same average power and the same total bandwidth. When these channel resources are taken into account the picture of ALOHA efficiency that emerges is much different from that of ALOHA throughput. Specifically, the efficiency of an ALOHA channel approaches one for the important case of small values of throughput and small values of the signal to noise power ratio. In other words, under these conditions it is not possible to find a multiple access protocol which has a higher capacity for a given value of average power and a given bandwidth.

By the end of 1996 ALOHA channels have been employed in a wide variety of connection free wireless applications. Various forms of ALOHA channels are used as the signaling channel in all three major digital cellular standards (IS-54, IS-95 and GSM). They are used in the ARDIS and RAM Mobitex packet radio networks, in the Japanese Teleterminal network and in a variety of commercial campus networks, such as the Multipoint mpNET and the ARIA System III. They are used in the request channel of the INMARSAT maritime satellite network to allow tens of thousands of ship stations to request voice and telex capacity and in more than 100,000 very small aperture earth stations (VSAT's) now in operation.

All of these products are narrowband applications typically operating at about 10 Kbs. Conventional first generation ALOHA channels cannot easily provide the much higher bandwidths required for a broadband wireless data network to service a large number of users and the larger markets of interest today. A conventional ALOHA channel cannot be easily implemented when the channel bandwidth is much higher than this because of the demands this puts on the burst power output of the remote terminals. Newer developments of second generation wideband versions of ALOHA, such as Spread ALOHA Multiple Access (SAMA), are expected to change this situation in the future [16].

14.6 Scheduling

Random access protocols work well with bursty traffic where there are many more users than available channels, yet these users rarely transmit. If users have long strings of packets or continuous stream data, then random access works poorly as most transmissions result in collisions. Thus channels must be assigned to users in a more systematic fashion by transmission scheduling. In scheduled access the available bandwidth is channelized into multiple time, frequency, or code division channels. Each node schedules its transmission on different channels in such a way as to avoid conflicts with neighboring nodes while making the most efficient use of the available time and frequency resources. While there has been much work on transmission scheduling, or channel assignment, in cellular systems [17], the centralized control in these systems greatly simplifies the problem. Distributed scheduled access in ad hoc wireless networks in general is an NP- hard problem [18]. Selman et al. have recently discovered that NP-hard problems exhibit a rapid change in complexity as the size of the problem grows [19, 20]. The identification of this "phase transition" provides an opportunity for bounding the complexity of problems like scheduled access by staying on the good side of the phase transition.

Even with a scheduling access protocol, some form of ALOHA will still be needed since a predefined mechanism for scheduling will be, by definition, unavailable at startup. ALOHA provides a means for initial contact and the establishment of some form of scheduled access for the transmission of relatively

large amounts of data. A systematic approach to this initialization that also combines the benefits of random access for bursty data with scheduling for continuous data is packet reservation multiple access (PRMA) [Goodman89]. PRMA assumes a slotted system with both continuous and bursty users (e.g. voice and data users). Multiple users vie for a given time slot under a random access strategy. A successful transmission by one user in a given timeslot reserves that timeslot for all subsequent transmissions by the same user. If the user has a continuous or long transmission then after successfully capturing the channel he has a dedicated channel for the remainder of his transmission (assuming subsequent transmissions are not corrupted by the channel: this corruption causes users to lose their slots and they must then recontend for an unreserved slot, which can entail significant delay). When this user has no more packets to transmit, the slot is returned to the pool of available slots that users attempt to capture via random access. Thus, data users with short transmissions benefit from the random access protocol assigned to unused slots, and users with continuous transmissions get scheduled periodic transmissions after successfully capturing an initial slot. A similar technique using a combined reservation and ALOHA policy is described in [11].

14.7 Power Control

Access protocols can be made more efficient and distributed by taking advantage of power control. Work in this area has mainly focused on maintaining the SINR of each user sharing the channel above a given threshold, which may be different for different users. Necessary and sufficient conditions to ensure that a feasible set of transmit powers for all users exists under which these users can meet their threshold SINR levels given the link gains between them are determined in [28]. Battery power for each user is minimized by finding the minimum power vector within the feasible set. This algorithm can also be performed in a distributed manner, which eliminates the need for centralized power control. Access to the system can be based on whether the new user causes other users to fall below their SINR targets. Specifically, when a new user requests access to the system, a centralized controller can determine if a set of transmit powers exists such that he can be admitted without degrading existing users below their desired SINR threshold. This admission can also be done using the distributed algorithm, where the new user gradually ramps up his power, which causes interference to other existing users in the system. If the new user can be accommodated in the system without violating the SINR requirements of existing users, then the power control algorithms of the new and existing users eventually converge to the feasible power vector under which all users (new and existing) meet their SINR targets. If the new user cannot be accommodated then as he ramps up his power the other users will increase their powers to maintain their SINRs such that the new user remains far from his SINR target. After some number of iterations without reaching his target, the new user will either back off from the channel and try again later or adjust his SINR target to a lower value and try again.

A power control strategy for multiple access that takes into account delay constraints is proposed and analyzed in [28]. This strategy optimizes the transmit power relative to both channel conditions and the delay constraint via dynamic programming. The optimal strategy exhibits three modes: very low power transmission when the channel is poor and the tolerable delay large, higher power when the channel and delay are average, and very high power transmission when the delay constraint is tight. This strategy exhibits significant power savings over constant power transmission while meeting the delay constraints of the traffic.

Bibliography

[1] P. Jung, P.W. Baier, and A. Steil, "Advantages of CDMA and spread spectrum techniques over FDMA and TDMA in cellular mobile radio applications," *IEEE Trans. Vehic. Technol.*, pp. 357–364, Aug. 1993.

[2] T.S. Rappaport, *Wireless Communications - Principles and Practice*, IEEE Press, 1996.

[3] M.D. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press, 1993.

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

[5] B. Gundmundson, J. Sköld, and J.K. Ugland, "A comparison of CDMA and TDMA systems," *IEEE Vehic. Technol. Conf. Rec.*, pp. 732–735, May 1992.

[6] P. Jung, P.W. Baier, and A. Steil, "Advantages of CDMA and spread spectrum techniques over FDMA and TDMA in cellular mobile radio applications," *IEEE Trans. Vehic. Technol.*, pp. 357–364, Aug. 1993.

[7] N. Abramson, "The Aloha system - another alternative for computer communications," Proc. Fall Joint Comput. Conf., AFIPS Conf,. p. 37, 1970.

[8] D. Bertsekas and R. Gallager, Data Networks, 2nd Edition, Prentice Hall 1992.

[9] A. Chockalingam and M. Zorzi, "Energy consumption performance of a class of access protocols for mobile data networks," Proc. IEEE Vehic. Technol. Conf. pp. 820-824, May 1998.

[10] P. Karn, "MACA: A new channel access method for packet radio," Proc. Comp. Net. Conf., pp. 134-140, Sept. 1990.

[11] V. Bharghavan, A. Demers, S. Shenkar, and L. Zhang, "MACAW: A Media Access Protocol for Wireless LAN, Proc. ACM SIGCOMM '94, pp. 212-225, Aug. 1994.

[12] C.-K. Toh, V. Vassiliou, G. Guichal, and C.-H. Shih, "MARCH: A medium access control protocol for multihop wireless ad hoc networks," Proc. IEEE Milt. Commun. Conf. (MILCOM),, 2000, pp. 512-516.

[13] D.A. Maltz, J. Broch, and D.B. Johnson, "Lessons from a full-scale multihop wireless ad hoc network testbed," IEEE Pers. Commun. Mag., pp. 8-15, Feb. 2001.

[14] Z.J. Haas, J. Deng, and S. Tabrizi, "Collision-free medium access control scheme for ad hoc networks, Proc. Milt. Commun. Conf. (MILCOM), pp. 276-280, 1999.

[15]] S.-L. Wu, Y.-C. Tseng and J.-P. Sheu, "Intelligent Medium Access for Mobile Ad Hoc Networks with Busy Tones and Power Control", IEEE J. Select. Areas Commun., pp. 1647- 1657, Sept. 2000.

[16] N. Abramson, "Wide-band random-access for the last mile," *IEEE Pers. Commun. Mag.*, Vol. 3, No. 6, pp. 29–33, Dec. 1996.

[17] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems - a comprehensive survey," IEEE Pers. Commun. Mag., pp. 10-31, June 1996.

[18] K.K. Parhi R. Ramaswami, "Distributed scheduling of broadcasts in a radio network," Proc. IEEE INFOCOM, pages 497-504, March 1989.

[19] Selman, B., "Stochastic Search and Phase Transitions: AI Meets Physics." Proc. Intl. Joint Conf. Artfl. Intell. (IJCAI-95), 1995. (invited paper)

[20] C.P. Gomes, S.B. Wicker, X. Xie, and B. Selman, "Connection between phase transitions in complexity and good decoding," Int. Symp. Inform. Theory Appl., Honolulu, Hawaii, November 5-8, 2000.

[21] P. Agrawal, "Energy efficient protocols for wireless systems," Proc. IEEE Intl. Symp. Personal, Indoor, Mobile Radio Commun., pp. 564-569, Sept. 1998.

[22] S. Kandukuri and N. Bambos, "Power controlled multiple access (PCMA) in wireless communication networks," Proc. IEEE Infocom, pp. 386-395, March 2000.

Bibliography

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

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

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

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

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

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

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

[8] I.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.

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

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

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

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

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

[14] S. Kasturia, J.T. Aslanis, and J.M. Cioffi, "Vector coding for partial response channels," *IEEE Trans. Inform. Theory*, Vol. IT-36, No. 4, pp. 741–762, July 1990.

[15] S.-G. Chua and A.J. Goldsmith, "Variable-rate variable-power MQAM for fading channels," *VTC'96 Conf. Rec.* June 1996. Also submitted to *IEEE Trans. Commun.*

[16] S.-G. Chua and A.J. Goldsmith, "Adaptive coded modulation," *ICC'97 Conf. Rec.* June 1997. Also submitted to *IEEE Trans. Commun.*

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

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

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

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

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

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

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

[24] P. Billingsley. *Probability and Measure*. 2nd Ed. New York: Wiley, 1986.

[25] A. Goldsmith and M Medard, "Capacity of time-varying channels with channel side information," *IEEE Intl. Symp. Inform. Theory*, pg. 372, Oct. 1996. Also submitted to the *IEEE Trans. Inform. Theory*.

[26] M.-S. Alouini and A. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," IEEE Transactions on Vehicular Technology, pp. 1165 -1181, July 1999.

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

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

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

[30] E. Visotsky and U. Madhow, "Space-time transmit precoding with imperfect feedback," *Proc. Intl. Symp. Inform. Theory*, June 2000.

[31] A. Narula, M. Lopez, M. Trott, G. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE JSAC*, Oct. 1998.

[32] A. Narula, M. Trott, G. Wornell, "Performance limits of coded diversity methods for transmitter antenna arrays," *IEEE Trans. Inform. Theory*, Nov. 1999.

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

Chapter 15

Cellular Systems and Infrastructure-Based Wireless Networks

One of the biggest challenges in providing multimedia wireless services is to maximize efficient use of the limited available bandwidth. Cellular systems exploit the power falloff with distance of signal propagation to reuse the same frequency channel at spatially- separated locations. Specifically, in cellular systems a given spatial area (like a city) is divided into nonoverlapping cells, as shown in Figure 15.1. Different frequencies, timeslots, or codes are assigned to different cells. For time and frequency division, cells operating on the same frequency or timeslot are spaced just far enough apart so that their mutual interference is tolerable. In code-division the codes are reused every cell.

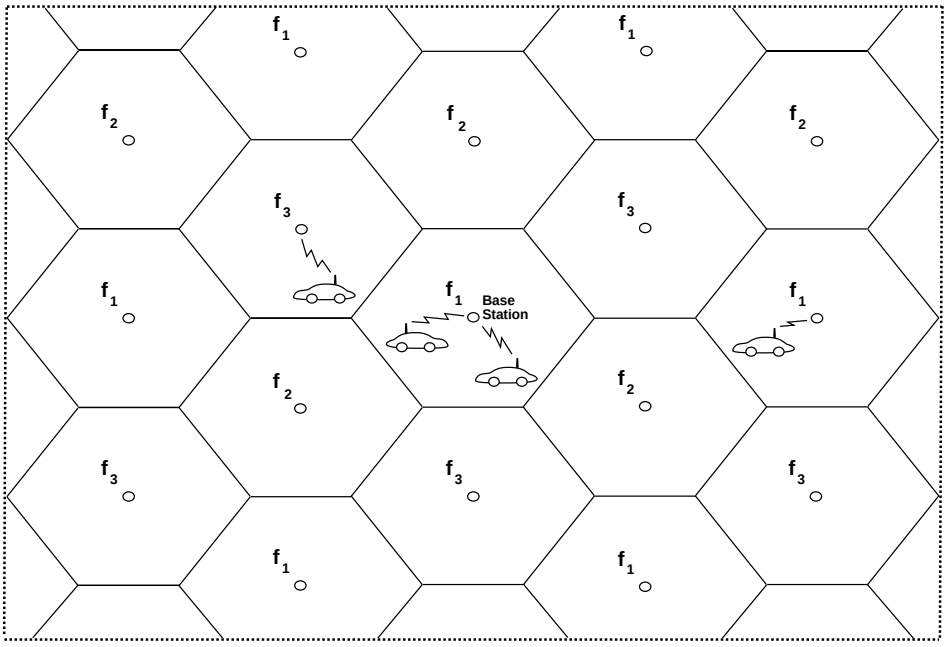


Figure 15.1: Cellular System.

In this chapter we first describe the basic design principles of cellular systems. We then describe a capacity measure for cellular systems, the *area spectral efficiency*, and compute this efficiency for simple cellular models.

15.1 Cellular System Design

For the cellular system shown in Figure 15.1 the central transmitter in each cell is connected to a base station and switching office which act as a central controller. Allocation of channels is performed by this centralized control function, as is the power control in CDMA systems. This controller also coordinates handoff to a neighboring cell when the mobile terminal traverses a cell boundary. The handoff procedure occurs when the base station in the originating cells detects the signal power of the mobile decreasing as it moves towards a cell boundary. This causes the base station in the originating cell to query neighboring cells in order to detect the destination base station. If the destination base station does not have any available channels, the handoff call will be dropped. A call will also be dropped if the originating base station detects a drop in received signal power due to multipath fading or shadowing, and initiates a handoff as a result even though the mobile terminal might be nowhere near a cell boundary. The spectral efficiency per unit area is increased by shrinking the size of a cell, since more users can be accommodated in a given area. However, decreasing the cell size increases the rate at which handoffs occur, which impacts higher level protocols. In general, if the rate of handoff increases the rate of call dropping will also increase proportionally. Routing is also more difficult with small cells, since routes need to be re-established whenever a handoff occurs.

While frequency reuse increases spectral efficiency, it also introduces co-channel interference, which affects the achievable data rate and bit-error-probability of each user. The interference which results from reusing frequencies is small if the users operating at the same frequency have enough distance between them. However, spectral efficiency is maximized by packing the users as close together as possible. Thus, the best cellular system design places users which share the same channel at a separation distance where the co-channel interference is just below the maximum tolerable level for the required data rate and error probability. Equivalently, good cellular system designs are interference-limited, such that the interference power is much larger than the noise power, and thus noise is generally neglected in the study of these systems.

Since co-channel interference is subject to shadowing and multipath fading, a static cellular system design must assume worst-case propagation conditions in determining this separation distance. A better design uses dynamic resource allocation in the cellular system, where power and bandwidth are allocated based on propagation conditions, user demands, and system traffic. Dynamic resource allocation can significantly increase both spectral and power efficiency, but the system complexity also increases dramatically. We discuss dynamic resource allocation is more detail below.

15.2 Frequency Reuse in Cellular Systems

15.2.1 Frequency Reuse in Code-Division Systems

The channels for code-division are semi-orthogonal due to the spreading code properties: these codes allow channel reuse in every cell, but also introduce interference from all users within the same cell (intracell interference) as well as from users in other cells (intercell interference). To compensate for the near-far problem of the intracell interferers, most code-division multiple access systems use power control. Unfortunately, using power control to invert signal attenuation dramatically increases the interference

from neighboring cells: since mobiles close to a cell boundary generally have weak received signal power, power control boosts up the transmit power of these boundary mobiles, which increases their interference to neighboring cells. Both intracell and intercell interference are attenuated by the processing gain of the code [1]. Due to the large number of interferers, the performance analysis of a code-division cellular system is fairly complex, and depends very heavily on the propagation model, cell size, mobility models, and other system parameters [1].

15.2.2 Frequency Reuse in Time and Frequency Division Systems

The channels in frequency-division (FDMA) or time-division (TDMA) are orthogonal, so there is no intracell interference in these systems. However, frequency reuse introduces intercell (co-channel) interference in all cells using the same channel. Thus, the received SNR for each user is determined by the amount of interference at its receiver. If the system is not interference-limited then spectral efficiency could be further increased by allowing more users onto the system or reusing the frequencies at smaller distances.

Consider the cell diagram in Figure 15.2 below. Let R be the distance from the cell center to a vertex. We denote the location of each cell by the pair (i, j) where, assuming cell A to be centered at the origin $(0, 0)$, the location relative to cell A is obtained by moving i cells along the u axis, then turning 60 degrees counterclockwise and moving j cells along the v axis. For example, cell G is located at $(0, 1)$, cell S is located at $(1, 1)$, cell P is located at $(-2, 2)$, and cell M is located at $(-1, -1)$. It is straightforward to show that the distance between cell centers of adjacent cells is $\sqrt{3}R$, and that the distance between the cell center of a cell located at the point (i, j) and the cell center of cell A (located at $(0, 0)$) is given by

$$D = \sqrt{3}R\sqrt{i^2 + j^2 + ij}. \tag{15.1}$$

The formula (15.1) for D suggests a method for assigning frequency A to cells such that the cell separation between cells operating at frequency A is $D = \sqrt{3}R\sqrt{i^2 + j^2 + ij}$. Starting at the origin cell A, move i cells along any chain of hexagons, turn counterclockwise by 60 degress, move j cells along the hexagon chain of this new heading, and assign frequency A to the jth cell. This process is shown in Figure 15.3 below. To assign frequency A throughout the region, this process is repeated starting with any of the new A cells as origin.

Using this process to assign all frequencies results in hexagonal cell clusters, which are repeated at the distance D, as shown in Figure 15.4. Given that the area of a hexagonal cell is $A_{cell} = 3\sqrt{3}R^2/2$ and the area of a hexagonal cluster is $A_{cluster} = \sqrt{3}D^2/2$, the number of cells per cluster is $N = D^2/(3R^2) = i^2 + j^2 + ij$. N is also called the reuse factor, and a small value of N indicates efficient frequency reuse (frequencies reused more often within a given area).

15.3 Dynamic Resource Allocation in Cellular Systems

Initial cellular systems were based on a fixed frequency reuse pattern designed for worst-case signal propagation and interference assumptions. Any system using fixed frequency reuse and base station assignment must be designed relative to worst-case interference assumptions. Dynamic resource allocation is a more efficient strategy, where frequencies, base stations, data rates, and power levels are dynamically assigned relative to to the current interference, propagation, and traffic conditions. Simple dynamic channel allocation techniques have been shown to improve channel efficiency by a factor of two or more, even for relatively simple algorithms [7]. However, this analysis was based on fairly simplistic system

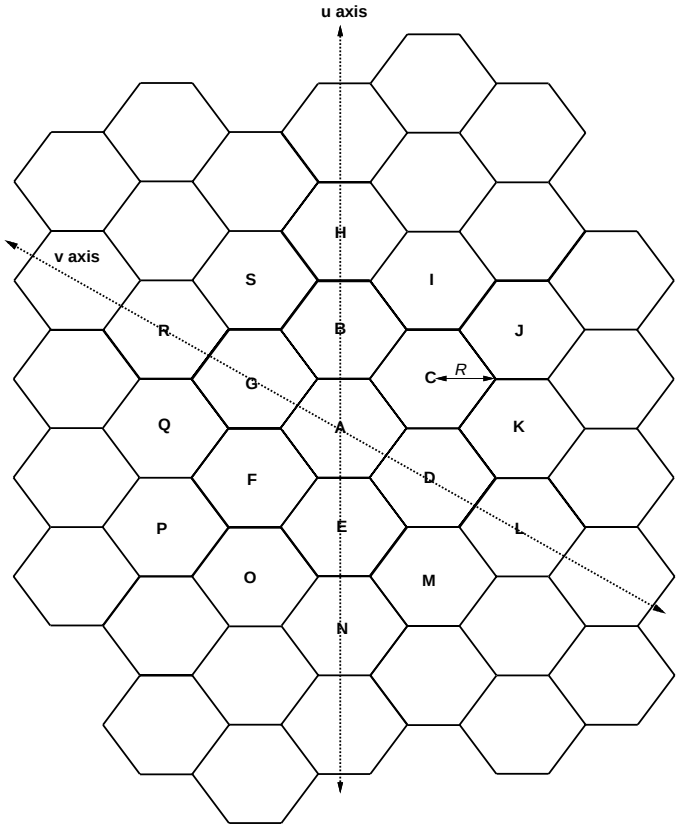


Figure 15.2: Cell Locations.

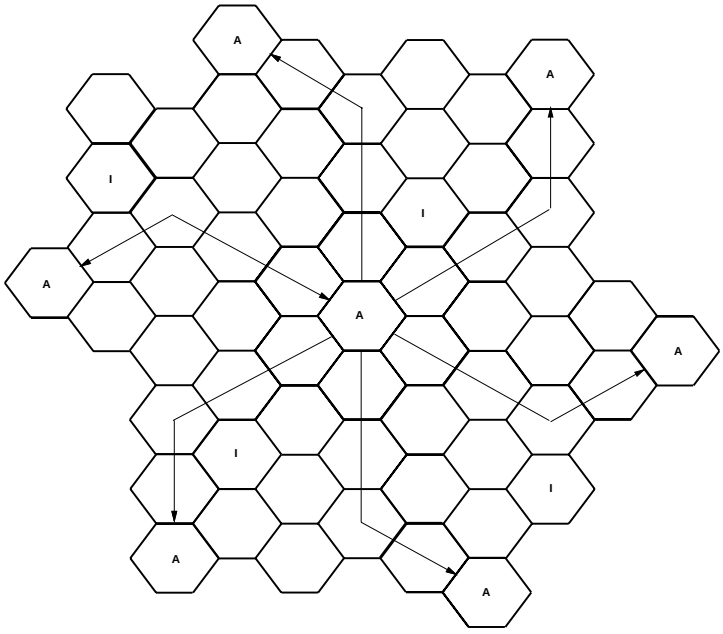


Figure 15.3: Frequency Assignment.

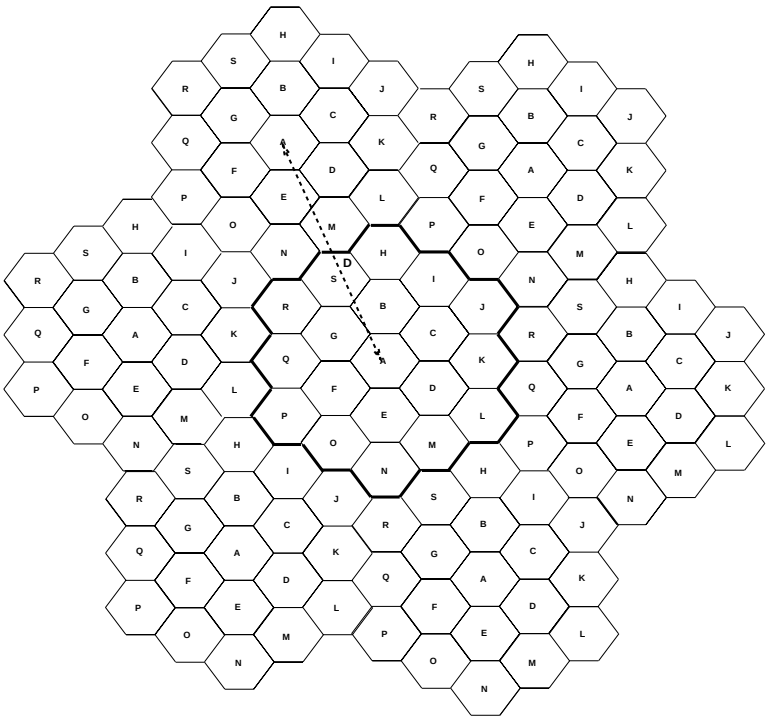


Figure 15.4: Cell Clusters.

assumptions. Performance improvement of dynamic resource allocation under realistic system conditions remains an open and challenging research problem.

Most previous investigations on dynamic channel allocation were based on assumptions of fixed traffic intensity, heterogeneous user demands, fixed reuse constraints, and static channels and users. Even under these simplistic assumptions, optimizing channel allocation is highly complex. An excellent survey on current research in dynamic resource allocation can be found in [7]

Reduced complexity has been obtained by applying neural networks and simulated annealing to the problem; however, these approaches can suffer from lack of convergence, or yield suboptimal results. Little work has been done on optimal or suboptimal resource allocation strategies which consider the simultaneous stochastic variation in traffic, propagation, and user mobility. In addition, current allocation procedures an not easily generalized to incorporate power control, traffic classes (e.g. multimedia), cell handoff, or user priorities. The superior efficiency of dynamic resource allocation is most pronounced under light loading conditions. As traffic becomes heavier, dynamic allocation strategies can suffer from suboptimal allocations which are difficult to reallocate under heavy loading conditions. Thus, the optimal dynamic resource allocation strategy is also dependent on traffic conditions, as was the optimal choice of multiple or random access technique. Finally, the complexity of dynamic resource allocation, particularly in systems with small cells and rapidly-changing propagation conditions and user demands, may be impossible to overcome, at least for the next five to ten years.

15.4 Area Spectral Efficiency

The multiuser capacity results of Chapter 15 assume multiple users sharing the same frequency band through either an orthogonal (FDMA/TDMA) or semi-orthogonal (CDMA) partition of the spectrum.

The spectral efficiency over a large geographical area for any of these partition techniques can generally be increased by reusing the same frequency, time slot, or code at spatially separated cells, where the power falloff with distance reduces the effect of the intercell interference. The magnitude of the intercell interference depends on both the distance between interfering cells, also called the reuse distance R_D, as well as the propagation laws governing the interferers' transmissions and the power adaptation policy. Ideally, we would like to optimize the reuse distance R_D to maximize the multiuser capacity per unit area of the cellular system. We would also like to optimize the power adaptation policy, but this is a very difficult optimization problem, as we discuss below.

In the following sections, we first describe the interference model used for the capacity calculations. We then define the multicell capacity and the area spectral efficiency, which are both functions of reuse distance. We also give a qualitative discussion of the effects of power control on intracell and intercell interference. We conclude by outlining some methods of interference mitigation. These methods include antenna sectorization, voice activity monitoring, and interference cancellation. Since multicell systems are interference limited, any technique to reduce interference will increase the system capacity.

15.5 Interference Model

Most cellular systems are *interference limited*, meaning that the receiver noise power is generally much less than the interference power, and can hence be neglected. The interference distribution for multicell systems is generally assumed to be Gaussian. This is a reasonable assumption for CDMA systems, where there are many intracell and intercell interferers, so the Gaussian distribution follows from the law of large numbers. With FDMA or TDMA, however, there is usually only a few dominant interferers from the first tier of interfering cells. Thus, the Gaussian assumption is usually invalid. In particular, on the forward link, one or two mobiles which are close to the cell boundaries will generally dominate the interference. On the reverse link, there are at most six interfering base stations for hexagonal cells. However, for capacity calculations, the capacity-achieving distribution for all users (i.e. signal and interference) is Gaussian. Thus, modeling the interference as Gaussian noise in capacity calculations is justified for any of the partitioning techniques we've discussed.

15.5.1 Reuse Distance, Multicell Capacity, and Area Efficiency

Define the *reuse distance* R_D to be the minimum distance between any two base stations that use the same code, frequency, or time slot. Since these resources are reused at the distance R_D, the area covered by each resource is roughly the area of a circle with radius $.5R_D$: $\pi(.5R_D)^2$. The larger the reuse distance, the less efficienctly the network resources are used. However, reducing R_D increases the level of interference between cells, thereby reducing the capacity region of each cell. The multicell capacity characterizes this tradeoff between efficient resource use and the capacity region per cell.

Consider a cellular system with N users per cell, a reuse distance R_D, and a total bandwidth allocation B. The multicell system capacity is defined as the multiuser rate region per Hertz divided by the coverage area reserved for the cell resources:

$$C_{\text{multicell}} = \frac{(R_1, R_2, \ldots, R_N)/B}{\pi(.5R_D)^2}, \tag{15.2}$$

where $(R_1, R_2, \ldots, R_N)$ is the set of maximum rates that can be maintained by all users in the cell simultaneously. Clearly, this set of rates will monotonically decrease as the interference from other cells increases. Typically, these interference levels are inversely proportional to R_D. Since the denominator of

(15.2) increases with R_D, there should be an optimal reuse distance which maximizes (15.2). However, deriving this optimal value for the entire rate region is quite complicated, and therefore we instead consider optimizing the reuse distance for the area efficiency, which we now describe.

The area spectral efficiency of a cell is defined as the total bit rate/Hz/unit area that is supported by a cell's resources. Given the multicell system capacity described above, the area efficiency is just

$$A_e = \frac{\sum_{i=1}^{N} R_i/B}{\pi(.5R_D)^2}. \tag{15.3}$$

The rate R_i is just the capacity of the ith user in the cell, which depends on $\gamma_i = S_i/I_i$, the received signal-to-interference power of that user, and B_i, the bandwidth allocated to that user. We could also define R_i to be the maximum possible rate for the ith user under a given set of system parameters (e.g. QPSK modulation with trellis coding, three branch diversity, and a required BER of 10^{-6}. If γ_i is constant, then $R_i = C_i = B_i \log(1 + S_i/I_i)$. Typically, γ_i is not constant, since both the interference and signal power of the ith user will vary with propagation conditions and mobile locations. When γ_i varies with time, R_i equals the time-varying channel capacity of the ith user:

$$R_i = B_i \int \log(1 + \gamma_i) p(\gamma_i) d\gamma_i. \tag{15.4}$$

It can also be defined as the maximum possible rate for the ith user under the given system parameters and time-varying channel conditions.

In general, it is extremely difficult to obtain the distribution $p(\gamma_i)$ in a multicell system, since this distribution depends on the power control policy and channel variations of both the signal and the interferers. The power control policy that maximizes a single user's data rate will not always maximize the area efficiency, since increasing the signal power of one user increases that user's interference to everyone else. Determining the power control policy that maximizes area efficiency is a complex optimization problem which will depend on the spectrum partitioning technique, propagation characteristics, system layout, and the number of users. This optimization is too complex for analysis if all the system characteristics are taken into account. Thus, optimal power control for multicell systems remains an open problem.

If we fix the power control policy, and assume a particular set of system parameters, then the distribution of γ_i can be determined either analytically or via simulation. The distribution of γ_i for CDMA systems (i.e., with both intracell and intercell interference), assuming Gaussian interference and the channel inversion power control policy, has been determined analytically in [1, 2, 3], and via simulation in [4, 5]. The distribution of γ_i for CDMA under other power control policies, and for FDMA and TDMA under any form of power control, has not yet been determined. With these distributions, a comprehensive comparison of area efficiency under different power control policies and spectrum partitioning methods could be done using the methods described above.

15.5.2 Efficiency Calculations

We now give some examples of the area efficiency calculation for the cell uplink with different power control policies. In order to get analytical formulas for the efficiency, we must make very simple assumptions about the system. In particular, we ignore the effects of noise, fading, and shadowing. We will also ignore the effects of user mobility, and calculate the efficiency based on a fixed location for the mobile of interest and the interferers.

Consider first frequency-division, where all users in the cell are assigned the same bandwidth $B_i = B/N$ and transmit power S. We assume the pessimistic model that all the users in the cell are located at the cell boundary, and all the interferers are located at their cell boundaries closest to our cell of interest.

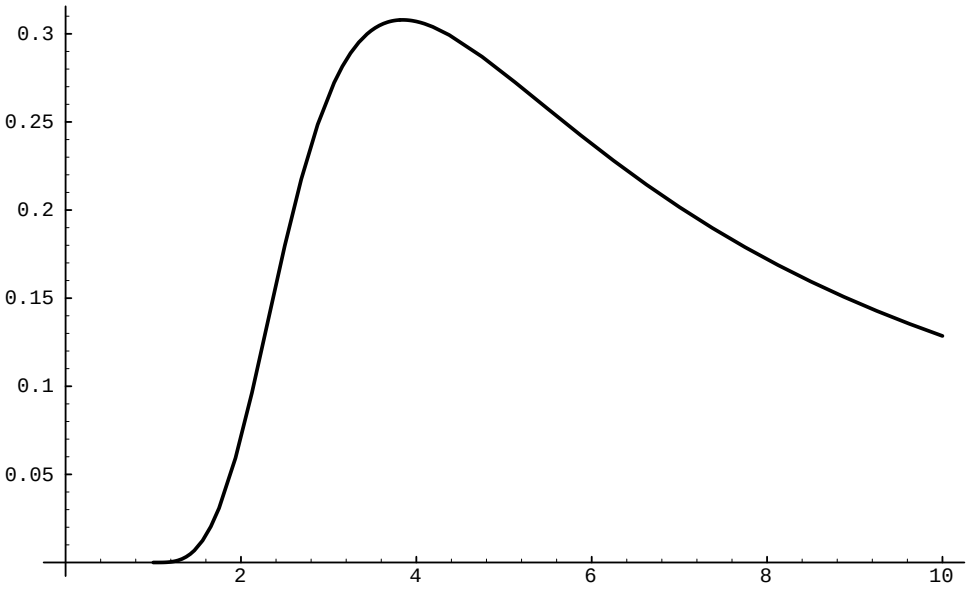


Figure 15.5: Area Efficiency for Frequency Division ($\gamma = 4$)

We assume a propagation model of Kd^{-2} within a cell, and $Kd^{-\gamma}$ outside the cell, where $2 \leq \gamma \leq 4$. With no power control (constant transmit power), the received signal power of the ith user is then $S_i = SKR^{-2}$, and the interference power is $I_i = 6SK(R_D - R)^{-\gamma}$. The capacity of the ith user in the cell is thus

$$C_i = \frac{B}{N} \log \left(1 + \frac{(R_D - R)^\gamma}{6R^2} \right), \tag{15.5}$$

and the area efficiency is

$$A_e = \frac{\sum_{i=1}^{N} C_i/B}{\pi(.5R_D)^2} = \frac{\log \left(1 + \frac{(R_D - R)^\gamma}{6R^2} \right)}{\pi(.5R_D)^2}. \tag{15.6}$$

Plots of A_e versus R_D for $\gamma = 4$ and $\gamma = 2$, are shown in figures 15.5 and 15.6 below. In this plot and all subsequent plots, we normalize the cell radius to $R = 1$. Comparing these figures, we see that, as expected, if the interference propagation loss falls off more slowly, the area efficiency is decreased. However, it is somewhat surprising that the optimal reuse distance is also decreased.

Suppose now that the interferers are not on the cell boundaries. If all interferers are at a distance $R_D - R/2$ from their base stations, then the area efficiency becomes

$$A_e = \frac{\sum_{i=1}^{N} C_i/B}{\pi(.5R_D)^2} = \frac{\log \left(1 + \frac{(R_D - R/2)^\gamma}{6(R)^2} \right)}{\pi(.5R_D)^2}. \tag{15.7}$$

The area efficiency in this case is plotted in the figure below for $\gamma = 4$. As expected, the area efficiency in this case is larger then in Figure 15.5 and the optimal reuse distance is smaller.

Returning to the pessimistic geometry of all users on the cell boundaries, suppose we now use the power control policy which inverts the channel. The received signal power in this case is $S_i = S\rho$, where ρ is a normalizing constant that insures the transmit power satisfies the average power constraint[1]. The

[1]The transmit power of the mobile at distance D from its base is $S\rho D^2/K$ to compensate for the path loss KD^{-2}. For our static channel, with the mobile at distance $D = R$, $\rho = KR^{-2}$ insures an average transmit power of S. In general, ρ will

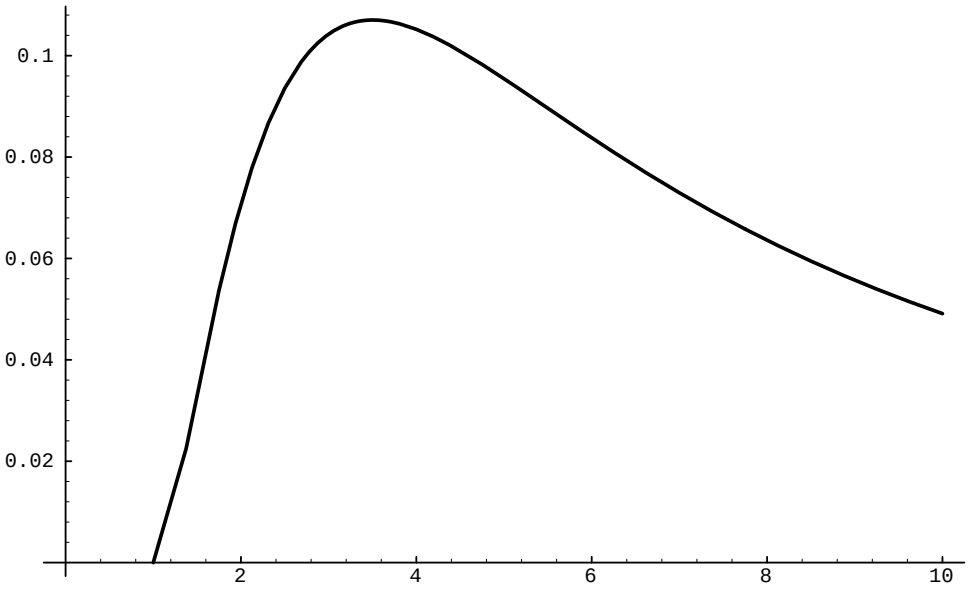


Figure 15.6: Area Efficiency for Frequency Division ($\gamma = 2$)

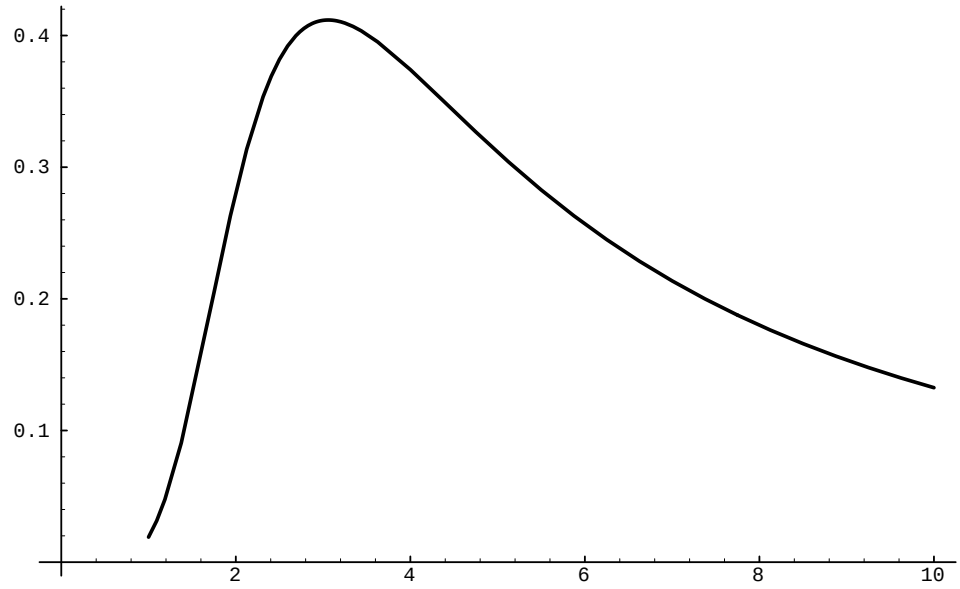


Figure 15.7: Interferer Distance of $R_D - R/2$

interference power is $I_i = 6S\rho[R^2/K][K(R_D - R)^{-\gamma}]$, where the first bracketed term is the power control of the interferer, and the second bracketed term is the propagation loss of that interferer. The received signal-to-interference power in this case is thus

$$\frac{S_i}{I_i} = \frac{(R_D - R)^\gamma}{6R^2}, \tag{15.8}$$

the same as in the case of no power control. So the area efficiency in this case is the same as that shown in Figures 15.5 and 15.6. The reason that power control does not affect the efficiency in this case is symmetry: because the mobile and interferers have the exact same propagation loss and interference conditions, the mobiles and interferers apply the same power control to their transmit signals. Thus, whatever the power control policy is, its effect will be cancelled out. However, when the interferers are closer to their base stations, the symmetry no longer applies.

Assume now that we use channel inversion, and that the interferers are at a distance $R/2$ from their base stations. The received signal power of the ith user is still $S_i = S\rho$. The received interference power is then

$$I_i = 6S\rho[(R/2)^2/K][K(R_D - R/2)^{-\gamma}], \tag{15.9}$$

and the resulting area efficiency is

$$A_e = \frac{\log\left(1 + \frac{(R_D - R/2)^\gamma}{6(R/2)^2}\right)}{\pi(.5R_D)^2}. \tag{15.10}$$

This efficiency is plotted as a function of R_D in the figure below for $\gamma = 4$. Comparing Figures 15.6

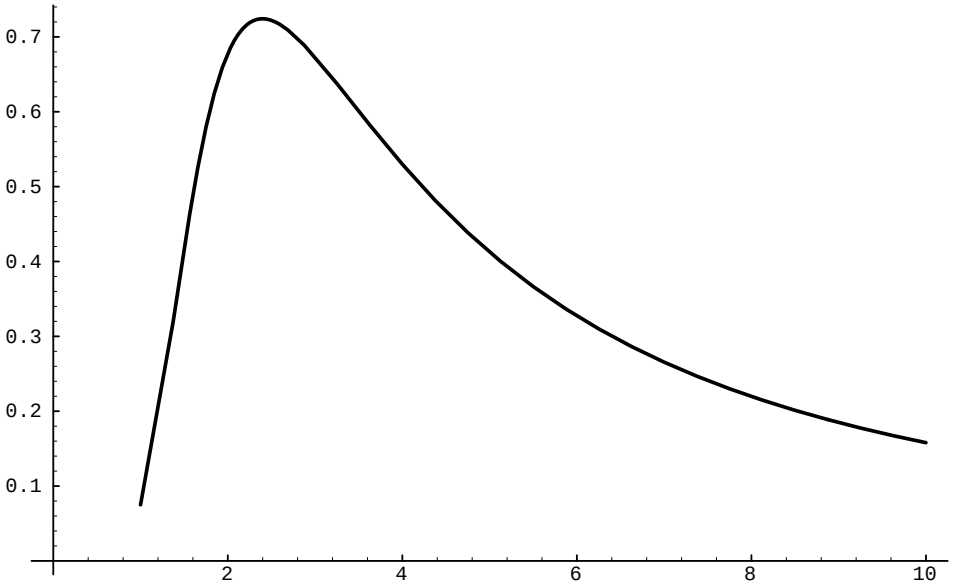


Figure 15.8: Channel Inversion Power Control ($\gamma = 4$)

and 15.8 we see that for this system geometry, channel inversion has a higher efficiency than constant transmit power.

depend on the distribution of the channel variation, e.g. if the received $S/I = \gamma$ is Rayleigh distributed then $\rho = \mathbf{E}1/\gamma = 0$. Since both the mobiles and interferers typically have the same S/I distribution, we assume ρ is the same for both.

Water-filling will give the same efficiency for all mobiles on the cell boundary as the other power control policies. However, as the interferers move closer to their base stations, they will increase their power. Thus, it is not clear what the worst-case interference scenario is for the water-filling power control.

The last example we consider is spread spectrum with channel inversion. Here, the received signal power is $S\rho$. The interference power is the sum of in-cell and out-of-cell interference. We consider only the first tier of interfering cells, and assume that those 6 cells contribute the same interference power. Picking one of the interfering cells at random, the interference contribution from that cell is thus

$$I_{\text{interfering-cell}} = \frac{S}{G}\sum_{i=1} N\frac{(R_D - R_i)^{-\gamma}}{R_i^2}, \tag{15.11}$$

where G is the spreading gain of the code (the cross-correlation inverse) and R_i is the distance of the ith interferer from its base station. The total interference power is thus

$$I_i = \frac{S(N-1)}{G} + \frac{6S}{G}\sum_{i=1} N\frac{(R_D - R_i)^{-\gamma}}{R_i^2}. \tag{15.12}$$

Computing this interference power is fairly complicated, unless we assume that all users in the cell are located at the cell boundary, which is unlikely. However, we can lower bound the area efficiency by considering only the in-cell interference. Since we are ignoring the out-of-cell interference, the optimal reuse distance will be $R_D = 2R$, i.e. codes are reused in every cell. Then for N large,

$$\frac{S_i}{I_i} = \frac{N-1}{G} \approx \frac{N-1}{N} \approx 1, \tag{15.13}$$

where we make the approximation that the spreading gain G is roughly equal to the number of codes N. Plugging this in, we get an area efficiency of

$$A_e = \frac{\log(1+1)}{\pi(.5(2))^2]} = \frac{1}{\pi} = .318. \tag{15.14}$$

Thus, if we completely ignore out-of-cell interference, we get roughly the same capacity as the worst-case interference scenario of frequency-division without power control and interference power falloff with distance of Kd^{-4} (Figure 15.5). If we use the empirical observation that the out-of-cell interference power is roughly the same as the in-cell interference power, we get $A_e = \log 1.5\pi = .186$.

15.6 Power Control Impact on Interference

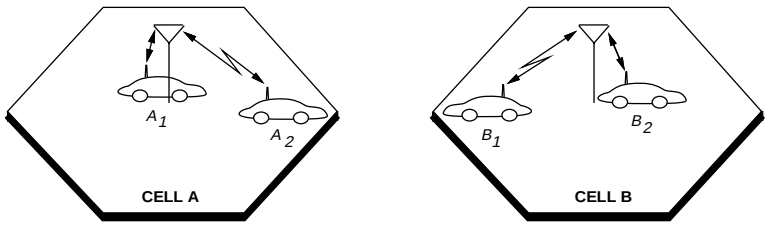


Figure 15.9: Interference Effects.

In this section we describe the qualitative impact of the power control policies we discussed for single-user systems on intracell and intercell interference. Consider first the case of intracell interference

on the forward link (mobile to base station), where two users A_1 and A_2 are transmitting to the same base station, as shown in Figure 15.9. Recall that intracell interference only occurs in CDMA systems, since with FDMA or TDMA only one user is assigned to each frequency or time slot in the cell. If both A_1 and A_2 transmit at the same power level, then the signal received by the base station from A_1 will generally be stronger than the signal received by A_2. Therefore, the interference caused by A_1 to A_2 will be strong even after despreading. This difference in received signal strength is called the near-far effect. To compensate for this effect, power control which equalizes the receive power of all users within a cell is used. With this type of power control, the received power of users A_1 and A_2 at the base station is the same, regardless of their individual path losses, so the signal-to-interference power after receiver processing equals the spreading gain. The "water-filling" power control policy, which increases power when the channel is good, has the opposite effect: since A_1 has a good signal path it will increase its transmit power, while A_2 has a bad signal path, so it will decrease its signal power. Moreover, this policy has a recursive effect: A_1 increasing its power causes A_2 to have an even worse channel, so A_2 will lower its power. This decreases the interference to A_1, so A_1 increases its power further, and so on. Roughly speaking, the constant water-filling tends to *remove* all users from the cell except the one with the most favorable channel. Therefore, if we consider only intracell interference effects, the water-filling policy is unacceptable when all the users within a cell require a guaranteed rate at all times. However, it may have a higher throughput in a system where the users within a cell can tolerate long periods of no transmission with an occasional burst of very high-rate data, as in packet radio systems. This assumes that all the users within a cell will eventually have the best signal path to the base station.

The effect of these two power control policies on intercell interference is quite different. Again referring to Figure 15.9, suppose we have intercell interferers B_1 and B_2 from cell B coupling into cell A. Without power control, the interference power from B_1 will be strong, since it is close to the boundary of cell A, while the interference from B_2 has much farther to travel to the base station of cell A, and will therefore be weaker. With the constant power policy, B_1 will transmit at a high power since it is far from its base station, and this will cause a higher level of interference in cell A. Since B_2 reduces power with this policy, and it is far from cell A's base station, the constant power policy has the effect of magnifying the power of interferers near cell B's boundary while reducing the power of interferers close to cell B's base station. Conversely, the water-filling power control will cause B_1 to lower its power and B_2 to increase its power, so that the intercell interferers in cell B have approximately the same amount of power coupling into cell A's base station, regardless of their location in cell B. Since the dominant intercell interferers are generally near the cell boundaries, water-filling will significantly reduce intercell interference on the forward link.

For the reverse link, the intracell interference and signal are both transmitted from the base station, so their path loss at any point within cell A is the same. Therefore, no power control is required to equalize the received signal strength of the signal and interference (equivalently, the constant power policy for the reverse link is achieved with no power control). Water-filling power control has the same recursive effect as in the forward link: since A_1 has a good path, the base station transmits to A_1 at a high power, which will cause interference to A_2, so transmit power to A_2 is reduced, and so on. Hence, the effect of these two power controls policies on intracell interference is roughly the same for both the forward link and the reverse link.

For intercell interferers, if the base station is sending to B_1 and B_2 at the same power level, then the location of B_1 and B_2 will not affect the amount of power coupling in to cell A. With water-filling, the base station will send at a higher power to B_2 and a lower power to B_1, but these interference signals have the same path loss to the mobiles in cell A. Therefore, it is difficult to say which power control policy will cause worse intercell interference on the reverse link.

15.7 Interference Mitigation

The rate regions for any of the three spectrum-sharing techniques will be increased if interference can be reduced while maintaining the same number of users per cell and the same reuse distance. Several techniques have been proposed to accomplish this, including speech gating, sectorization of the base station antennas, and interference cancellation. We now describe each of these techniques in somewhat more detail.

Speech gating takes advantage of the fact that in duplex voice transmission, each speaker is only active approximately 40% of the time [6]. If voice activity is monitored, and transmission suppressed when no voice is present, then overall interference caused by the voice transmission is reduced. If we denote the average percentage of time that voice is active by ρ, then through speech gating the average power of both intracell and intercell interference is reduced by ρ. Antenna sectorization refers to the use of directional transmit and receive antennas at the base station. For example, if the 360^o omni base station antenna is divided into three sectors to be covered by three directional antennas of 120^o beamwidths, then the interferers seen by each directional antenna is one third the number that would be seen by the omni. If N_S denotes the number of directional antennas used to cover the 360^o beamwidth then, on average, antenna sectorization reduces the total interference power by a factor of N_S.

Another method of mitigating interference in CDMA systems is multiuser detection. The received CDMA signal is a superposition of each user's signal, where user i modulates its data sequence with a unique spreading code. The multiuser detector for such a received signal jointly detects the data sequences of all users: if the data sequences of the interference is known, then it can be subtracted out from the desired signal, as in the superposition coding techniques described above. The optimal receiver for CDMA joint detection was derived by Verdú in [8]; it uses a bank of matched filters and the Viterbi algorithm to determine either the maximum-likelihood set of received signal sequences or the set of signal sequences with minimum error probability. However, the complexity of this optimal receiver structure is exponential in the number of interfering users, making the receiver impractical for systems with many interferers. The detection algorithm also requires knowledge of the signal energies, which is not always available.

Several suboptimal multidetection schemes which are more practical to implement have also been developed. A multiuser decorrelator for joint detection which does not require knowledge of the user energies and with complexity that is only linear in the number of users was proposed in [9] and [10] for synchronous and asynchronous users, respectively. Multistage detectors [11, 12] decode the users' signals sequentially in decreasing order of their received power. Specifically, the highest-power signal is detected using a conventional CDMA receiver (i.e., all interference signals are treated as noise). This signal is then subtracted from the total received signal, and then the highest-power remaining signal is detected. This successive interference cancellation is done until all signals have been estimated. The decision-feedback detector, proposed in [13], uses both forward and feedback filters to remove multiuser interference. As with decision-feedback equalization, this approach suffers from error propagation. The multistage detectors generally yield better performance than the decorrelator and decision-feedback detectors at a cost of increased complexity (although still linear in the number of users). These detectors were designed for AWGN channels, while more recent studies have looked at multiuser detection in fading channels [14, 15].

Bibliography

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

[2] P. Jung, P.W. Baier, and A. Steil, "Advantages of CDMA and spread spectrum techniques over FDMA and TDMA in cellular mobile radio applications," *IEEE Trans. Vehic. Technol.*, Vol. VT-42, No. 3, pp. 357–364, Aug. 1993.

[3] J.-P. Linnartz, *Narrowband Land-Mobile Radio Networks.* Norwood, MA: Artech House, 1993.

[4] T.S. Rappaport and L.B. Milstein, "Effects of radio propagation path loss on DS-CDMA cellular frequency reuse efficiency for the reverse channel," *IEEE Trans. Vehic. Technol.*, Vol. VT-41, No. 3, pp. 231–242, Aug. 1992.

[5] B. Gundmundson, J. Sköld, and J.K. Ugland, "A comparison of CDMA and TDMA systems," *IEEE Vehic. Technol. Conf. Rec.*, pp. 732–735, May 1992.

[6] P.T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell System Tech. J.*, Vol 47, pp. 73–91, Jan. 1968.

[7] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems - a comprehensive survey," *IEEE Pers. Commun. Mag.*, Vol. 3, No. 3, pp. 10–31, June 1996.

[8] S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. Inform. Theory*, Vol IT-32, No. 1, pp. 85–96, Jan. 1986.

[9] R. Lupas and S. Verdú, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, Vol. IT-35, No. 1, pp. 123–136, Jan. 1989.

[10] R. Lupas and S. Verdú, "Near-far resistance of multiuser detectors in asynchronous channels," *IEEE Trans. Commun.*, Vol. COM-38, No. 4, pp. 496–508, April 1990.

[11] M.K. Varanasi and B. Aazhang, "Multistage detection in asynchronous code-division multiple-access communications," *IEEE Trans. Commun.*, Vol. COM-38, No. 4, pp. 509–519, April 1990.

[12] M.K. Varanasi and B. Aazhang, "Near-optimum detection in synchronous code-division multiple-access systems," *IEEE Trans. Commun.*, Vol. COM-39, No. 5, pp. 725–736, May 1991.

[13] A. Duel-Hallen, "Decorrelating decision-feedback multiuser detector for synchronous code-division multiple-access channel," *IEEE Trans. Commun.*, Vol. COM-41, No. 2, pp. 285–290, Feb. 1993.

[14] S. Vasudevan and M.K. Varanasi, "Optimum diversity combiner based multiuser detection for time-dispersive Rician fading CDMA channels," *IEEE J. Selected Areas Commun.*, Vol. SAC-12, No. 4, pp. 580–592, May 1994.

[15] Z. Zvonar and D. Brady, "Multiuser detection in single-path fading channels," *IEEE Trans. Commun.*, Vol. COM-42, No. 2-4, pp. 1729–1739, Feb.-April 1994.

Chapter 16

Ad-Hoc Wireless Networks

An ad hoc wireless network is a collection of wireless mobile nodes that self-configure to form a network without the aid of any established infrastructure, as shown in Figure 16.1. Without an inherent infrastructure, the mobiles handle the necessary control and networking tasks by themselves, generally through the use of distributed control algorithms. Multihop connections, whereby intermediate nodes send the packets towards their final destination, are supported to allow for efficient wireless communication between parties that are relatively far apart. Ad hoc wireless networks are highly appealing for many reasons. They can be rapidly deployed and reconfigured. They can be tailored to specific applications, which fits with the Oxford English Dictionary's definition of ad hoc:"For this purpose, to this end; for the particular purpose in hand or in view." They are also highly robust due to their distributed nature, node redundancy, and the lack of single points-of-failure. These characteristics are especially important for military applications, and much of the groundbreaking research in ad hoc wireless networking was supported by the (Defense) Advanced Research Projects agency (DARPA) [1, 2, 3]. Despite much research activity over the last several decades on wireless communications in general, and ad hoc wireless networks in particular, there remain many significant technical challenges in the design of these networks. In this chapter we describe the basic design principles of ad hoc networks and some of the remaining technical challenges that are still unsolved.

The lack of infrastructure inherent to ad hoc wireless networks is best illustrated by contrast with the most prevalent wireless networks today: cellular systems and wireless local area networks (WLANs). As described in Chapter 15, cellular telephone networks divide the geographic area of interest into regions called cells. A mobile terminal located in a given cell communicates directly with a base station located at or near the center of each cell. Thus, there is no peer-to-peer communication between mobiles. All communication is via the base station through single hop routing. The base stations and backbone network perform all networking functions, including authentication, call routing, and handoff. Most wireless LANs have a similar, centralized, single hop architecture: mobile nodes communicate directly with a centralized access point that is connected to the backbone Internet, and the access point performs all networking and control functions for the mobile nodes. In contrast, an ad hoc wireless network has peer-to-peer communication, distributed networking and control functions among all nodes, and multihop routing.

This discussion should not be taken to mean that ad hoc wireless networks are completely flat; i.e., cannot have any infrastructure or pre-established node hierarchy. Indeed, many ad hoc wireless networks form a backbone infrastructure from a subset of nodes in the network to improve network reliability and capacity [4]. Similarly, some nodes may be chosen to perform as base stations for neighboring nodes [5]. The distinguishing emphasis in the ad hoc approach lies in the design requirements. Ad hoc wireless

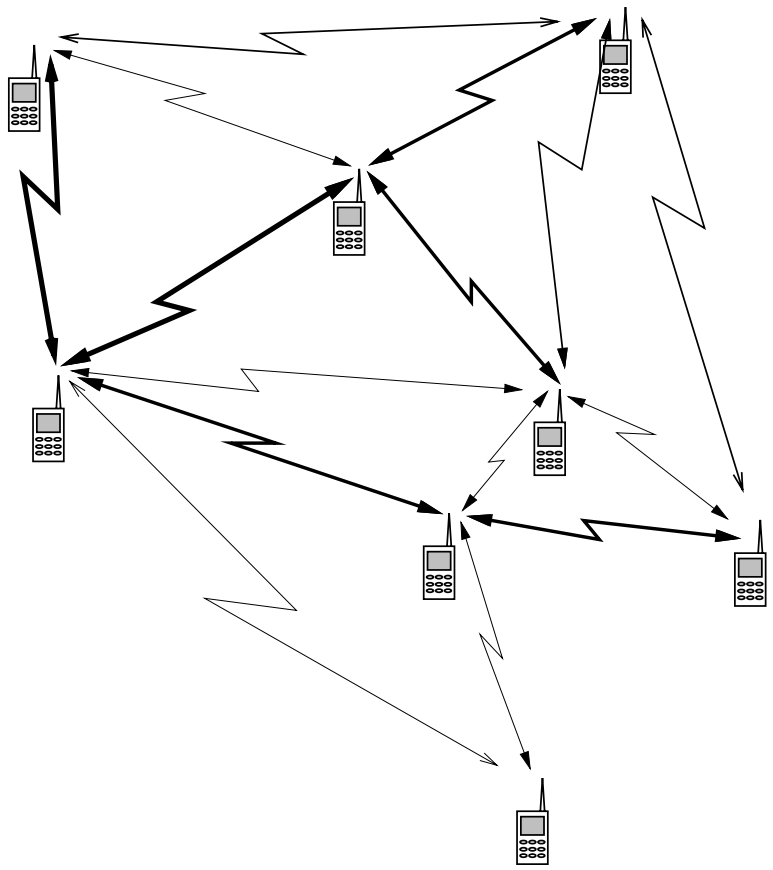


Figure 16.1: Ad Hoc Network.

networks may exploit infrastructure to improve network performance. However, while the infrastructure provides the side benefit of enhanced performance, it is not a fundamental design principle of the network.

Ad hoc networks are quite common in the wired world. Indeed, most LANs, metropolitan area networks (MANs), and wide area networks (WANs), including the Internet, have an ad hoc structure. However, the broadcast nature of the radio channel introduces characteristics in ad hoc wireless networks that are not present in their wired counterparts. In particular, a radio channel allows a node to transmit a signal directly to any other node. The link signal-to-interference-plus-noise power ratio (SINR) between two communicating nodes will typically decrease as the distance between the nodes increases, and will also depend on the signal propagation and interference environment. Moreover, this link SINR varies randomly over time due to the mobility of the nodes which typically changes the transmission distance, propagation environment, and interference characteristics. Link SINR determines the communication performance of the link: the data rate and associated probability of packet error or bit error (bit- error-rate or BER) that can be supported on the link. Links with very low SINRs are not typically used due to their extremely poor performance, leading to partial connectivity among all nodes in the network, as shown in Figure 16.1. However, link connectivity is not a binary decision, as nodes can back off on their transmission rate or increase their transmit power as link SINR degrades and still maintain connectivity [6, 7]. This is illustrated by the different line widths corresponding to different link qualities in Figure 1. Link connectivity also changes as nodes enter and leave the network, and this connectivity can be controlled by adapting the transmit power of existing network nodes to the presence of a new node [8].

The flexibility in link connectivity that results from varying link parameters such as power and data rate has major implications for routing. Nodes can send packets directly to their final destination via single hop routing as long as the link SINR is above some minimal threshold. However, single hop routing can cause excessive interference to surrounding nodes. Routing over a single hop may also require a relatively low rate or have a high probability of bit or packet error if the link SINR is low, thereby introducing excessive delays. Alternatively packets can be forwarded from source to destination by intermediate nodes at a link rate commensurate with the forwarding link SINR. Routing via forwarding by intermediate nodes is called multihop routing. Several recent research results indicate that ideal multihop routing significantly increases the capacity of ad hoc wireless networks [57, 34], but achieving these gains through a decentralized routing strategy remains elusive. The channel and network dynamics of ad hoc wireless systems coupled with multihop routing make it difficult to support multimedia requirements of high speed and low delay. However, flexibility in the link, access, network and application protocols can be exploited to compensate and even take advantage of these dynamics.

Energy constraints are not inherent to all ad hoc wireless networks. Devices in an ad hoc wireless network may be stationary and attached to a large energy source. Mobile devices may be part of a large vehicle, such as a car or tank, that can generate significant amounts of power over the long term. However, many ad hoc wireless network nodes will be powered by batteries with a limited lifetime. Some of the most exciting applications for ad hoc wireless networks follow this paradigm. Thus, it is important to consider the impact of energy-constrained nodes in the design of ad-hoc wireless networks. Devices with rechargeable batteries must conserve energy to maximize time between recharging. Of particular interest are devices that cannot be recharged, i.e. sensors that are imbedded in walls or dropped into a remote region. Energy constraints impact both the hardware operation and the signal transmission associated with node operation. It is often assumed that the transmit power associated with packet transmission dominates power consumption. However, signal processing associated with packet transmission and reception, and even hardware operation in a standby mode, consume nonnegligible power as well [9, 10, 11]. This entails interesting energy tradeoffs across protocol layers. At the link layer many communications techniques that reduce transmit power require a significant amount of signal processing. It is widely assumed that the energy required for this processing is small and continues to decrease with ongoing improvements in hardware technology [10, 12]. However, the results in [9, 11] suggest that these energy costs are still significant. This would indicate that energy-constrained systems must develop energy-efficient processing techniques that minimize power requirements across all levels of the protocol stack and also minimize message passing for network control, as these entail significant transmitter and receiver energy costs. Sleep modes for nodes must be similarly optimized, since these modes conserve standby energy but may entail energy costs at other protocol layers due to associated complications in access and routing. The hardware and operating system design in the node can also be optimized to conserve energy: techniques for this optimization are described in [11, 13].

Another important characteristic of ad hoc wireless networks is mobility in the network nodes. Mobility impacts all layers of the network protocol stack. At the link layer it determines how fast the link characteristics change and whether or not the link connectivity is stable over time. At the medium access control layer it affects how long measurements regarding channel and interference conditions remain in affect and how scheduling algorithms perform. At the network layer mobility has major implications for the performance of different routing protocols. The impact of mobility on network performance ultimately dictates which applications can be supported on a highly mobile network. The impact of mobility on ad hoc wireless network design will be discussed in more detail throughout the paper.

The remainder of this chapter is organized as follows. We first discuss applications for ad hoc wireless networks, including data networks, home networks, device networks, sensor networks, and distributed

control. Next we consider cross layer design in ad hoc wireless networks: what it is, why it is needed, and how it can be done. Link layer design issues are discussed next, followed by consideration of the medium access control (MAC) layer design issues, including the tradeoffs inherent to frequency/time/code channelization and the assignment of users to these channels via random access or scheduling. This section also describes the role power control can play in multiple access. Networking issues such as neighbor discovery, network connectivity, scalability, routing, and network capacity are outlined next. Last we describe techniques for the network to adapt to the application requirements and the application to adapt to network capabilities.

16.0.1 Applications

Since the ad hoc wireless network paradigm tailors the network design to the intended application, it will be useful to consider potential applications in some detail. In what follows we will consider both military and commercial applications. We will see that several design requirements are common to both types of systems, especially the need for energy efficiency. Military applications often require the self-configuring nature and lack of infrastructure inherent to ad hoc wireless networks, even if it results in a significant cost or performance penalty. The lack of infrastructure is also highly appealing for commercial systems, since it precludes a large investment to get the network up and running, and deployment costs may then scale with network success. Other commercial advantages include ease of network reconfiguration and reduced maintenance costs. However, these advantages must be balanced against any performance penalty resulting from the need for distributed network control.

In this section we consider the following applications: data networks, home networks, device networks, sensor networks, and distributed control systems. Note that this list is by no means comprehensive, and in fact the success of ad hoc wireless networks hinges on making them sufficiently flexible so that there can be accidental successes. Therein lies the design dilemma for ad hoc wireless networks. If the network is designed for maximum flexibility to support many applications (a one-size-fits-all network) then it will be difficult to tailor the network to different application requirements. This will likely result in poor performance for some applications, especially those with high rate requirements or stringent delay constraints. On the other hand, if the network is tailored to a few specific applications then designers must predict in advance what these "killer applications" will be - a risky proposition. Ideally an ad hoc wireless network must be sufficiently flexible to support many different applications while adapting its performance to the given set of applications in operation at any given time. The cross layer design discussed in below provides this flexibility along with the ability to tailor protocol design to the energy constraints in the nodes.

Data Networks

Ad hoc wireless networks can support data exchange between laptops, palmtops, personal digital assistants (PDAs), and other information devices. We focus on two types of wireless data networks: LANs with coverage over a relatively small area (a room, floor, or building) and MANs with coverage over several square miles (a metropolitan area or battlefield). The goal of wireless LANs is to provide peak data rates on the order of 10-100 Mbps, similar to what is available on a wired LAN, for low- mobility and stationary users. Commercial wireless LAN standards such as 802.11a and 802.11b provide data rates on this order, however the individual user rates are much less if there are many users accessing the system. Moreover, these commercial LANs are not really based on an ad hoc structure. The normal 802.11 network configuration is a star topology with one wireless access point and single hop routing from the mobile units to the access point. While the 802.11 standard does support a peer-to-peer architecture

in the form of the Independent Base Service Set (IBSS) configuration option, it is not widely used and its performance is somewhat poor [Saadawi01].

Wireless MANs typically require multihop routing since they cover a large area. The challenge in these networks is to support high data rates, in a cost-effective manner, over multiple hops, where the link quality of each hop is different and changes with time. The lack of centralized network control and potential for high-mobility users further complicates this objective. Military programs such as DARPA's GLOMO (Global mobile information systems) have invested much time and money in building high-speed wireless MANs that support multimedia, with limited success [14, 15]. Wireless MANs have also permeated the commercial sector, with Metricom the best example [16]. While Metricom did deliver fairly high data rates throughout several major metropolitan areas, the deployment cost was quite large and significant demand never materialized. Metricom filed for protection under Chapter XI of the Federal Bankruptcy Code in the fall of 2000.

Note that energy efficiency is a major issue in the design of wireless data networks. The canonical example of an ad hoc wireless data network is a distributed collection of laptop computers. Laptops are highly limited in battery power, so power must be conserved as much as possible. In addition, a laptop acting as a router for other laptops could drain its battery forwarding packets for other users. This would leave no power for the laptop user and would initiate a change in network topology. Thus, these networks must conserve battery power in all communication functions, and devise routing strategies that use residual power at each node of the network in a fair and efficient manner.

Home Networks

Home networks are envisioned to support communication between PCs, laptops, PDAs, cordless phones, smart appliances, security and monitoring systems, consumer electronics, and entertainment systems anywhere in and around the home. The applications for such networks are limited only by the imagination. For example, using a PDA in the bedroom one could scan stored music titles on a PC and direct the bedroom stereo to play a favorite piece, check the temperature in the living room and increase it by a few degrees, check the daily TV programming from the Internet and direct the VCR to record a show that night, access voice messages and display them using a voice-to-text conversion software, check stocks on the Internet and send selling instructions to a broker, and start the coffee maker and toaster, all without getting up from the bed. Other applications include smart rooms that sense people and movement and adjust light and heating accordingly, "aware homes" that network sensors and computers for assisted living of seniors and those with disabilities, video or sensor monitoring systems with the intelligence to coordinate and interpret data and alert the home owner and the appropriate police or fire department of unusual patterns, intelligent appliances that coordinate with each other and with the Internet for remote control, software upgrades, and to schedule maintenance, and entertainment systems that allow access to a VCR, Tivo box, or PC from any television or stereo system in the home [17, 18, 19, 20].

There are several design challenges for such networks. One of the biggest is the need to support the varied quality-of-service (QoS) requirements for different home networking applications. QoS in this context refers to the requirements of a particular application, typically data rates and delay constraints, which can be quite stringent for home entertainment systems. Other big challenges include cost and the need for standardization, since all of the devices being supported on this type of home network must follow the same networking standard. Note that the different devices accessing a home network have very different power constraints: some will have a fixed power source and be effectively unconstrained, while others will have very limited battery power and may not be rechargeable. Thus, one of the biggest challenges in home network design is to leverage power in unconstrained devices to take on the heaviest communication and networking burden, such that the networking requirements for all nodes in the

network, regardless of their power constraints, can be met.

One approach for home networking is to use an existing wireless LAN standard such as 802.11 [21]. But 802.11 has several limitations for this type of application. First, it most commonly supports a star architecture with a single access point and all devices talking directly to this access node. This star architecture eliminates the benefits of multihop routing, and while multihop routing is possible in 802.11, as noted above, its performance is poor. In addition, 802.11 uses a statistical multiple access protocol, which makes it difficult to support the quality required in home entertainment systems. 802.11b is also somewhat limited in data rate (1-10Mbps), and while the 802.11a standard supports much higher rates (10-70 Mbps), it is mainly designed for packet data applications and not media streaming. While protocols to support media streaming on top of 802.11a are being developed (802.11e), this type of overlay will likely be insufficient to provide high- quality wireless home entertainment.

A natural choice for home networking is a peer-to-peer ad hoc wireless network. Much of the communication in home networks will take place between peer devices, so peer-to-peer communication eliminates the overhead of going through a centralized node. In addition, many of the devices in a home network will be low power or battery-limited. In an ad hoc wireless network these devices need only communicate with their nearest neighbors (typically a short distance away) to maintain connectivity with (all) other devices in the home. Thus, multihop routing will be very beneficial to such devices in terms of energy savings. Most home networking applications involve stationary or low-mobility nodes, so the protocols need not support high mobility. Ad hoc wireless networks will be challenged to provide high-quality media streaming for home entertainment, and this is an open area of active research.

Home networking is being pushed strongly by the HomeRF working group, which has developed an open industry standard for such networks that combines a centralized and peer-to-peer structure [19]. The working group for HomeRF was initiated by Intel, HP, Microsoft, Compaq, and IBM. The main component of the HomeRF protocol is its Shared Wireless Access Protocol (SWAP). The SWAP protocol is designed to carry both voice and data traffic and to interoperate with the PSTN and the Internet. SWAP is a combination of a managed network that provides isochronous services (such as real-time voice and video) via a centralized network controller (the main home PC) along with an ad hoc peer-to-peer network for data devices. The centralized network controller is not required but it greatly facilitates providing dedicated bandwidth to isochronous applications. Bandwidth sharing is enabled by frequency hopped spread spectrum at 50 hops/sec. HomeRF also supports a time division service for delivery of interactive voice and other time- critical services, and a random access protocol for high speed packet data. The transmit power for HomeRF is specified at 100 mW which provides a data rate of 1-2 Mbps. However, in August 2000 the FCC authorized a five-fold increase in the HomeRF bandwidth, effectively increasing data rates to 10 Mbps. The range of HomeRF covers a typical home and backyard. HomeRF products operating in the 2.4 GHz band are currently on the market in the $100-200$ price range. Details on these products can be found at http://www.homerf.org.

Device Networks

Device networks support short-range wireless connections between devices. Such networks are primarily intended to replace inconvenient cabled connections with wireless connections. Thus, the need for cables and the corresponding connectors between cell phones, modems, headsets, PDAs, computers, printers, projectors, network access points, and other such devices is eliminated. Clearly many of these devices have limited battery life, but are generally rechargeable. Thus, device networks require energy efficiency.

The main technology driver for such networks is Bluetooth [5, 22]. The Bluetooth standard is based on a tiny microchip incorporating a radio transceiver that is built into digital devices. The transceiver takes the place of a connecting cable for electronic devices. Up to eight Bluetooth devices can form

a star-topology network (a piconet) with one node acting as a master and the other nodes acting as slaves. The master node is responsible for synchronization and scheduling transmissions of the slave nodes. Piconets can also be interconnected, leading to a multihop topology. Bluetooth is mainly for short-range communications, e.g. from a laptop to a nearby printer or from a cell phone to a wireless headset. Its normal range of operation is 10 m (at 1 mW transmit power), and this range can be increased to 100 m by increasing the transmit power to 100 mW. The system operates in the unregulated 2.4 GHz frequency band, hence it can be used worldwide without any licensing issues. The Bluetooth standard provides 1 data channel at 721 Kbps and up to three voice channels at 56 Kbps for an aggregate bit rate on the order of 1 Mbps. Networking is done via a packet switching protocol based on frequency hopping at 1600 hops per second. Energy constraints played a large role in the design of Bluetooth, with a goal of using as little energy from the host device as possible. Bluetooth uses a range of techniques in its hardware, communication, and networking protocols to preserve energy, including power-efficient modulation, a limited transmission range, smart packet detection, and intelligent sleep scheduling [22].

The Bluetooth standard was developed jointly by 3Com, Ericsson, Intel, IBM, Lucent, Microsoft, Motorola, Nokia, and Toshiba. Over 1300 manufacturers have now adopted the standard, and products compatible with Bluetooth are starting to appear on the market now. Some of the products currently available include a wireless headset for cell phones (Ericsson), a wireless USB or RS232 connector (RTX Telecom, Adayma), wireless PCMCIA cards (IBM), and wireless settop boxes (Eagle Wireless). The prognosis for Bluetooth has been varied, progressing from the early euphoria of the late1990s to pessimism and claims of premature death in the year 2000 to the current outlook of guarded optimism.

Sensor Networks

Sensor networks have enormous potential for both consumer and military applications. For the military, it is now painfully clear that the wars of the 21st century will differ significantly from those of the 20th. Enemy targets will be small, mobile, and generally found in extremely hostile terrain. If the war in Afghanistan is any indication, the targets in future combats will be small and difficult to detect from great distances. Future military missions will therefore require that sensors and other intelligence gathering mechanisms be placed close to their intended targets. The potential threat to these mechanisms is therefore quite high, so it follows that the technology used must be highly redundant and require as little support as possible from friendly forces. An apparent solution to these constraints lies in large arrays of passive electromagnetic, optical, chemical, and biological sensors. These can be used to identify and track targets, and can also serve as a first line of detection for various types of attacks. A third function lies in the support of the movement of unmanned, robotic vehicles. For example, optical sensor networks can provide networked navigation, routing vehicles around obstacles while guiding them into position for defense or attack. The design considerations for some industrial applications are quite similar to those for military applications. In particular, sensor arrays can be deployed and used for remote sensing in nuclear power plants, mines, and other industrial venues.

Examples of sensor networks for the home environment include electricity, gas, and water meters that can be read remotely through wireless connections. The broad use of simple metering devices within the home can help consumers identify and regulate devices like air conditioners and hot water heaters that are significant consumers of power and gas. Simple attachments to power plugs can serve as the metering and communication devices for individual appliances. One can imagine a user tracking various types of information on home energy consumption from a single terminal the home computer. Remote control of television usage and content could be monitored in similar ways. Another important home application is smoke detectors that could not only monitor different parts of the house but also communicate to track the spread of the fire. Such information could be conveyed to local firefighters before they arrived on

the scene along with house blueprints. A similar type of array could be used to detect the presence and spread of gas leaks or other toxic fumes.

Sensor arrays also have great potential for use at the sites of large accidents. One may wish to consider, for example, the use of remote sensing in the rescue operations following the collapse of a building. Sensor arrays could be rapidly deployed at the site of an accident and used to track heat, natural gas, and toxic substances. Acoustic sensors and triangulation techniques could be used to detect and locate trapped survivors. It may even be possible to avert such tragedies altogether through the use of sensor arrays. The collapse of walkways and balconies, for example, can be predicted and tragedy averted by building stress and motion sensors into the structures from the outset. One can imagine large numbers of low-cost low-power sensors being directly inserted into the concrete before it is poured. Material fatigue can be detected and tracked over time throughout the structure. Such sensors must be robust and self-configuring, and would require a very long lifetime, commensurate with the lifetime of the structure.

Most sensors will be deployed with non-rechargeable batteries. The problem of battery lifetime in such sensors may be averted through the use of ultra-small energy-harvesting radios. Research on such radios, coined the PicoRadio, promise radios smaller than one cubic centimeter, weighing less than 100 grams, with a power dissipation level below 100 microwatts [23]. This low level of power dissipation enables nodes to extract sufficient power from the environment - energy harvesting - to maintain operation indefinitely. Such picoradios open up new applications for sensor deployment in buildings, homes, and even the human body.

In short, important applications of the future are enabled by large numbers of very small, lightweight, battery-powered sensors. These sensors must be easily and rapidly deployed in large numbers and, once deployed, they must form a suitable network with a minimum of human intervention. All of these requirements must be met with a minimum of power consumption due to battery limitations and, for many applications, the inability to recharge these batteries

Distributed Control Systems

Ad hoc wireless networks enable distributed control, with remote plants, sensors and actuators linked together via wireless communication channels. Such networks are imperative for coordinating unmanned mobile units, and greatly reduce maintenance and reconfiguration costs over distributed control systems with wired communication links. Ad hoc wireless networks are currently under investigation for supporting coordinated control of multiple vehicles in an automated highway system (AHS), remote control of manufacturing and other industrial processes, and coordination of unmanned airborne vehicles (UAVs) for military applications.

Current distributed control designs provide excellent performance as well as robustness to uncertainty in model parameters. However, these designs are based on closed-loop performance that assumes a centralized architecture, synchronous clocked systems, and fixed topology. Consequently, these systems require that the sensor and actuator signals be delivered to the controller with a small, fixed delay. Ad hoc wireless networks cannot provide any performance guarantee in terms of data rate, delay or loss characteristics: delays are typically random and packets may be lost. Unfortunately, most distributed controllers are not robust to these types of communication errors, and effects of small random delays can be catastrophic [24, 25]. Thus, distributed controllers must be redesigned for robustness to the random delays and packet losses inherent to wireless networks. Ideally, the ad hoc wireless network can also be tailored to the requirements of the controller. This is a relatively new area of research: recent results in this area can be found in [25] and the references therein. Energy constraints in distributed control systems will be highly application-dependent: cars in an automated highway will have a large renewable

energy source, whereas sensors in most manufacturing applications will have nonrechargeable batteries.

16.0.2 Cross Layer Design

The different applications for ad-hoc networks have a wide range of network requirements as well as different energy constraints for different network nodes. The network requirements must be met despite variations in the link characteristics on each hop, the network topology, and the node traffic. It is very difficult to ensure performance of the network or the support of real-time or mission critical data in the face of these random variations. There has been significant research directed toward energy constraints, application requirements, and network variability at different levels of the network protocol stack. Examples include diversity, coding, power control, and adaptive techniques at the link layer, power control and scheduling at the MAC layer, energy-constrained and delay-constrained routing at the network layer, and application adaptation at the application layer. However, this work has mainly targeted isolated components of the overall network design, thereby ignoring important interdependencies. Specifically, current ad hoc wireless network protocol design is largely based on a layered approach, as shown in Figure 16.2. In this model each layer in the protocol stack is designed and operated independently, with interfaces between layers that are static and independent of the individual network constraints and applications. This paradigm has greatly simplified network design and led to the robust, scalable protocols in the Internet. However, the inflexibility and suboptimality of this paradigm results in poor performance for ad hoc wireless networks in general, especially when energy is a constraint or the application has high bandwidth needs and/or stringent delay constraints. To meet these requirements a cross layer protocol design that supports adaptivity and optimization across multiple layers of the protocol stack is needed.

In an adaptive cross layer protocol stack, the link layer can adapt rate, power, and coding to meet the requirements of the application given current channel and network conditions. The MAC layer can adapt based on underlying link and interference conditions as well as delay constraints and bit priorities. Adaptive routing protocols can be developed based on current link, network, and traffic conditions. Finally, the application layer can utilize a notion of soft quality-of-service (QoS) that adapts to the underlying network conditions to deliver the highest possible application quality. It is important that the protocols at each layer not be developed in isolation, but rather within an integrated and hierarchical framework to take advantage of the interdependencies between them. These interdependencies revolve around adaptivity at each layer of the protocol stack, general system constraints, such as energy and mobility, and the application(s) the network is supporting.

Adaptivity at each layer of the protocol stack should compensate for variations at that layer based on the time scale of these variations. Specifically, variations in link SINR are very fast, on the order of microseconds for vehicle-based users. Network topology changes more slowly, on the order of seconds, while variations of user traffic may change over tens to hundreds of seconds. The different time scales of the network variations suggest that each layer should attempt to compensate for variation at that layer first. If adapting locally is unsuccessful then information should be exchanged with other layers for a more general response. For example, suppose the link connectivity (link SINR) in the wireless link of an end-to-end network connection is weak. By the time this connectivity information is relayed to a higher level of the protocol stack (i.e. the network layer for rerouting or the application layer for reduced-rate compression), the link SINR will most likely have changed. Therefore, it makes sense for each protocol layer to adapt to variations that are local to that layer. If this local adaptation is insufficient to compensate for the local performance degradation then the performance metrics at the next layer of the protocol stack will degrade as a result. Adaptation at this next layer may then correct or at least mitigate the problem that could not be fixed through local adaptation. For example, consider again the weak link scenario. Link connectivity can be measured quite accurately and quickly at the link level.

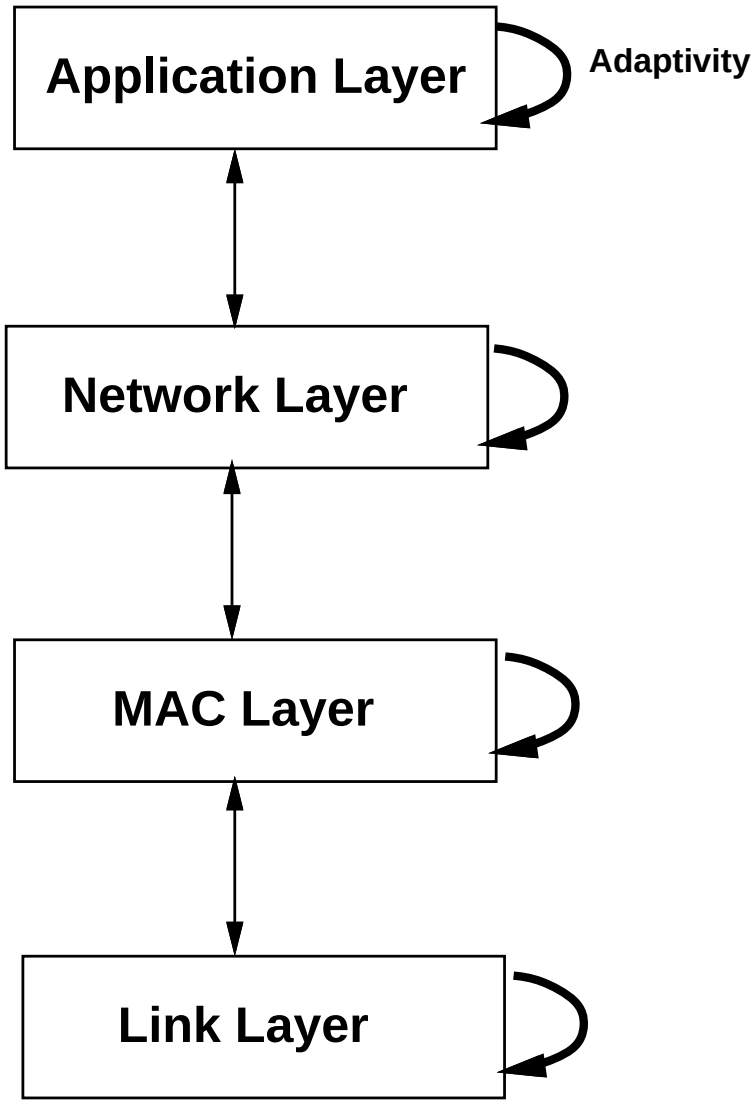


Figure 16.2: The OSI Layered Model for Protocol Design and Operation.

The link protocol can therefore respond to weak connectivity by increasing its transmit power or its error correction coding. This will correct for variations in connectivity due to, for example, multipath flat-fading. However, if the weak link is caused by something difficult to correct for at the link layer, e.g. the mobile unit is inside a tunnel, then it is better for a higher layer of the network protocol stack to respond by, for example, delaying packet transmissions until the mobile leaves the tunnel. Similarly, if nodes in the network are highly mobile then link characteristics and network topology will change rapidly. Informing the network layer of highly-mobile nodes might change the routing strategy from unicast to broadcast in the general direction of the intended user. It is this integrated approach to adaptive networking - how each layer of the protocol stack should respond to local variations given adaptation at higher layers - that forms the biggest challenge in adaptive protocol design.

Energy conservation also requires a cross layer design. For example, Shannon theory indicates that the energy required to communicate one bit of information decreases as the bit time increases [26].

Thus, energy can be conserved by transmitting a bit over a longer period of time. However, this will clearly impact the MAC protocol and also the application. Routing is also an interesting example. The most energy efficient routing protocol in a sensor network may use a centrally-located sensor to forward packets from other sensors. However, the battery of this sensor will be quickly exhausted, which might be undesirable from an application standpoint. Thus, the need for energy efficiency must be balanced against the lifetime of each individual node and the overall life of the network.

The above discussion indicates that in order to support an adaptive cross layer design, the design and operation of the protocol stack must evolve to that shown in Figure 16.3. This figure indicates that information must be exchanged across all layers in the protocol stack. This information exchange allows the protocols to adapt in a global manner to the application requirements and underlying network conditions. In addition, all protocol layers must be jointly optimized with respect to global system constraints and characteristics such as energy and high-mobility nodes. In order to design a protocol stack based on Figure 16.3, two fundamental questions must be answered:

1. What information should be exchanged across protocol layers and how should that information be adapted to?

2. How should global system constraints and characteristics be factored into the protocol designs at each layer.

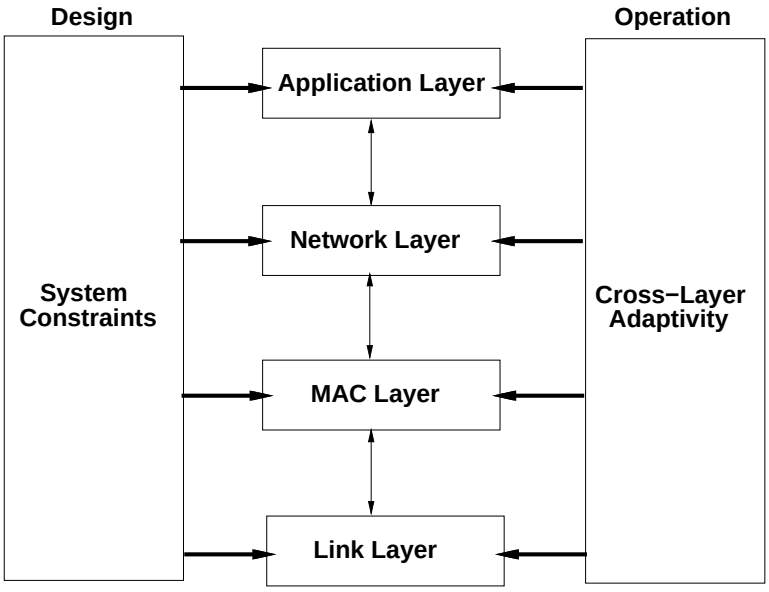


Figure 16.3: Adaptive Cross Layer Design and Application.

In the next several sections we will discuss the design of the different layers in the protocol stack, and then revisit cross layer design relative to these two questions. Cross layer design is an active theme in ad hoc wireless network design today. However, there remains many open questions in the understanding and implementation of this design philosophy.

16.1 Link Design Issues

Many of the design issues for link layer design were covered in previous chapters. We will now briefly review these ideas, and also discuss some new design issues that arise due to energy constraints.

16.1.1 Fundamental Capacity Limits

The fundamental capacity of wireless channels was discussed in Chapter 4. This capacity dictates the maximum data rate that can be transmitted over the channel with arbitrarily small probability of error. In Chapter 4 we analyzed the capacity of AWGN channels and fading channels, and the capacity of multiple antenna channels was given in Chapter 10.1.4. Capacity results for fading channels with perfect transmitter and receiver knowledge indicate that the transmitter should increase power and rate in good channel conditions and decrease them in poor channel conditions. The multiple antenna results indicate that the capacity of wireless channels increases linearly with the number of antennas at the transmitter/receiver, however this requires perfect channel estimates. Degradation in these estimates can significantly degrade the capacity gains resulting from multiple antennas. In general the capacity-achieving codes for wireless channels have asymptotically large block lengths. The long codes and complex decoding in this optimal scheme drive the probability of error to zero for any data rate below capacity, but the complexity of these schemes makes them hard to approximate with practical implementations.

Channel capacity under a hard transmit energy constraint, as opposed to a peak or average power constraint, is a relatively new design problem. With finite energy it is not possible to transmit any number of bits with asymptotically small error probability. This is easy to see intuitively by considering the transmission of a single bit. The only way to ensure that two different values in signal space, representing the two possible bit values, can be decoded with arbitrarily small error is to make their separation arbitrarily large, which requires arbitrarily large energy. Since arbitrarily small error probability is not possible under a hard energy constraint, a different notion of reliable communication is needed. Pioneering work by Gallager in this area defines reliable communication under a finite energy constraint in terms of the capacity per unit energy. This capacity per unit energy is defined as the maximum number of bits per unit energy that can be transmitted such that the maximum likelihood random coding error exponent is positive. This definition ensures that for all rates below the capacity per unit energy error probability decreases exponentially with the total energy, although it will not be asymptotically small for finite-energy channels. Gallager also shows that the capacity per unit energy is achieved using an unlimited number of degrees of freedom per transmitted bit. This translates to either very wideband communication or using many symbols per bit, the opposite of high-rate transmission schemes under a power constraint (e.g. MQAM, with M bits/symbol for M large).

Capacity per unit energy is also explored in [26], and these results can be used to obtain the capacity of finite-energy channels in terms of bits [27]. Capacity in bits dictates the maximum number of bits that can be transmitted over a channel using finite energy given some nonzero probability of bit error (recall that this error probability cannot be driven to zero with finite energy). The capacity of a finite-energy channel in bits is an important concept, since it indicates that ad hoc wireless networks with finite energy nodes only have a finite number of bits that a given node can transmit before exhausting its energy. Allocating those bits to the different requirements of the network: information transmission, exchange of routing information, forwarding bits for other nodes, channel estimation, etc., becomes an interesting and challenging optimization problem that clearly requires cross layer design.

16.1.2 Coding

Channel coding can significantly reduce the power required to achieve a given BER and is therefore a common feature in link layer design. Code designs for both AWGN and fading channels were discussed in Chapter 8. Most wireless systems use some form of error control coding to reduce power consumption. Conventional error control codes use block or convolutional code designs: the error correction capability of these codes is obtained at the expense of an increased signal bandwidth or a lower data rate. Trellis codes use a joint design of the channel code and modulation to provide good error correction without any bandwidth or rate penalty. Turbo codes and the more general family of codes on graphs minimize transmit power required for AWGN channels, but the associated processing complexity may compromise these power gains. All of these codes can also be designed for fading channels to limit required energy.

16.1.3 Multiple Antennas

Multiple antennas at the transmitter and/or receiver play a powerful role in improving the performance and reducing the required transmit power for wireless link layer designs, as described in more detail in Chapter 7. Multiple antenna systems typically use either diversity, beamsteering, or multiple input multiple output (MIMO) techniques. Diversity combining is a common technique to mitigate flat fading by coherently combining multiple independently fading copies of the signal. By significantly reducing the impact of flat- fading, diversity combining can lead to significant power savings.

Beamsteering creates an effective antenna pattern at the receiver with high gain in the direction of the desired signal and low gain in all other directions. Beamsteering is accomplished by combining arrays of antennas with signal processing in both space and time. The signal processing typically adjusts the phase shifts at each antenna to "steer" the beam in the desired direction. A simpler technique uses sectorized antennas with switching between the sectors. Beamsteering significantly improves energy efficiency since transmitter power is focused in the direction of its intended receiver. Beamsteering also reduces interference power along with fading and intersymbol interference due to multipath, since the interference and multipath signals are highly attenuated when they arrive from directions other than that of the line-of-sight (or dominant) signal. Results indicate that beamsteering can significantly improve the transmission range, data rates, and BER of wireless links. Highly mobile nodes can diminish these gains, as the beamsteering direction will be shifting and difficult to determine accurately.

Multiple input multiple output (MIMO) systems, where both transmitter and receiver use multiple antennas, can significantly increase the data rates possible on a given channel. As we saw in Chapter 7, in MIMO systems, if both the transmitter and the receiver have channel estimates, then with N antennas at the transmitter and receiver the MIMO system can be transformed into N separate channels that do not interference with each other, providing a roughly N-fold capacity increase over a system with a single antenna at both the transmitter and receiver. When the transmitter does not know the channel then the optimal transmission strategy is a space-time code, where bits are encoded over both space and time. These codes are highly complex, so in practice suboptimal schemes like layered space-time codes are used and tend to perform very well.

While multiple antenna techniques save transmission power, they are often highly complex and therefore require significant power for signal processing. Given a total energy constraint this tradeoff must be examined relative to each system to determine if multiple antenna techniques result in a net savings in energy.

16.1.4 Power control

Power control is a potent mechanism for improving wireless ad-hoc network performance. At the link layer power control can be used to compensate for random channel variations due to multipath fading, reduce the transmit power required to obtain a given data rate and error probability, minimize the probability of link outage, and reduce interference to neighboring nodes. It can also be used to meet hard delay constraints and prevent buffer overflow.

Power control strategies at the link layer typically either maintain SINR on the link above a required threshold by increasing power relative to fading and interference or use a "water-filling" approach where power and rate are increased for good channel conditions, decreased for poor channel conditions, and set to zero when the channel quality falls below a given cutoff threshold, as described in Chapter 9. The constant SINR strategy works well for continuous stream traffic with a delay constraint, where data is typically sent at a fixed rate regardless of channel conditions. However, this power control strategy is not power efficient, since much power must be used to maintain the constant SINR in deep fading conditions. Optimal variation of transmission rate and power maximizes average throughput and channel capacity, but the associated variable-rate transmission and channel-dependent delay may not be acceptable for some applications. Power control has also been used to meet delay constraints for wireless data links. In this approach power for transmission of a packet increases as the packet approaches its delay constraint, thereby increasing the probability of successful transmission [28]. A more complex approach uses dynamic programming to minimize the transmit power required to meet a hard delay constraint [29], and the resulting power consumption is much improved over power control that maintains a constant SINR.

Before closing this section, we want to emphasize that power control has a significant impact on protocols above the link layer. The level of transmitter power defines the "local neighborhood" - the collection of nodes that can be reached in a single hop - and thus in turn defines the context in which access, routing, and other higher layer protocols operate. Power control will therefore play a key role in the development of efficient cross layer networking protocols. We will discuss integration of power control with multiple access and routing protocols in later sections.

16.1.5 Adaptive Resource Allocation

Adaptive resource allocation in link layer design provides robust link performance with high throughput while meeting application-specific constraints. The basic premise is to adapt the link transmission scheme to the underlying channel, interference, and data characteristics through variation of the transmitted power level, symbol transmission rate, constellation size, coding rate/scheme, or any combination of these parameters. Moreover, adaptive modulation can compensate for SINR variations due to interference as well as multipath fading and can be used to meet different QOS requirements of multimedia [30] by prioritizing delay-constrained bits and adjusting transmit power to meet BER requirements.

Recent work in adaptive resource allocation has investigated combinations of power, rate, code, and BER adaptation ([31] and the references therein). These schemes typically assume some finite number of power levels, modulation schemes, and codes, and the optimal combination is chosen based on system conditions and constraints. Only a small number of power levels, rates, and/or codes are needed to achieve near-optimal performance, since there is a critical number of degrees of freedom needed for good performance of adaptive resource allocation, and beyond this critical number additional degrees of freedom provide minimal performance gain [31]. In particular, power control in addition to variable-rate transmission provides negligible capacity increase in fading channels [32], cellular systems [33, 40], and ad hoc wireless networks [34]. CDMA systems, in addition to varying power, data rate, and channel coding, can also adjust their spreading gain or the number of spreading codes assigned to a given user [35, 36].

The benefits of assigning multiple spreading codes per user are greatest when some form of multiuser detection is used, since otherwise self-interference is introduced [37]. Note also that in adaptive CDMA systems all transmitters sending to a given receiver must coordinate since they interfere with each other.

Other adaptive techniques include variation of the link layer retransmission strategy as well as its frame size. The frame is the basic information block transmitted over the link and includes overhead in the form of header and error control bits. Shorter frames entail a higher overhead, but are less likely to be corrupted by sporadic interference and require less time for retransmission. Recent results have shown that optimizing frame length can significantly improve throughput as well as energy efficiency [38].

Data communications require corrupted packets to be retransmitted so that all bits are correctly received. Current protocols typically discard the corrupted packet and start over again on the retransmission. However, recent work has shown that diversity combining of retransmitted packets or retransmitting additional redundant code bits instead of the entire packet can substantially increase throughput ([39] and the references therein). A performance comparison of incremental redundancy against that of adaptive modulation is given in [40].

16.2 Medium Access Control Design Issues

The medium access control protocol dictates how different users share the available spectrum. There are two components to this spectrum allocation: how to divide the spectrum into different channels, and then how to assign these different channels to different users. The different methods that can be used to divide the spectrum into different channels include frequency-division, time-division, code-division, and hybrid methods. Details on these techniques are given in Chapter 14. When users have very bursty traffic the most efficient mechanism to assign channels is random access, where users contend for a channel whenever they have data to transmit. This contention is inefficient when users have continuous stream data or long packet bursts. In this case some form of scheduling helps to prevent collisions and ensure continuous connections. The design and tradeoff analysis for different channel assignment strategies was given in Chapter 14.5.

Random access protocols can be more energy efficient by limiting the amount of time that a given node spends transmitting and receiving. The paging industry developed a solution to this problem several decades ago by scheduling "sleep" periods for pagers. The basic idea is that each pager need only listen for transmissions during certain short periods of time. This is a simple solution to implement when a central controller is available. It is less obvious how to implement such strategies within the framework of a distributed control algorithm. Access protocols that utilize node sleep times to minimize energy consumption are investigated in [10].

Random access schemes can be made more flexible in general, and more energy aware in particular, by adopting a dynamic programming approach to decisions about transmissions. Under dynamic programming, decision making is based on utility (cost) functions - an agent will act or not, depending on utility of the action as indicated by a utility function computed over some time period. A given protocol can be made energy aware by introducing the cost of a transmission into the utility function. Consider the case of ALOHA. In work conducted by MacKenzie at Cornell, a game-theoretic version of ALOHA was developed that initially focused on a simple "collision game" [41]. In this model the delay and energy cost of transmission are parameters of the cost function associated with transmission. The resulting system is both stable (in the language of game theory, there is a Nash Equilibrium) and distributed. It allows for individual nodes to make autonomous decisions on retransmission strategies. This simple version of the game assumes that the users know the number of backlogged users within the local neighborhood, but it is possible to develop utility functions that reflect less ideal situations. In general, the decision-theoretic

approach provides a convenient way to embed the cost of transmission decisions into random access protocols. Random access protocols work well with bursty traffic where there are many more users than available channels, yet these users rarely transmit. If users have long strings of packets or continuous stream data, then random access works poorly as most transmissions result in collisions. Thus channels must be assigned to users in a more systematic fashion by transmission scheduling, described in more detail in Chapter 14.6. Scheduling still requires some mechanism at startup to establish the schedule.

Scheduling under an energy constraint further complicates the problem. Channel capacity under a finite energy constraint is maximized by transmitting each bit over a very long period of time. However, when multiple users wish to access the channel, the transmission time allocated to each user must be limited. Recent work has investigated optimal scheduling algorithms to minimize transmit energy for multiple users sharing a channel [42]. In this work scheduling was optimized to minimize the transmission energy required by each user subject to a deadline or delay constraint. The energy minimization was based on judiciously varying packet transmission time (and corresponding energy consumption) to meet the delay constraints of the data. This scheme was shown to be significantly more energy efficient than a deterministic schedule with the same deadline constraint.

Power control improves the efficiency of random access and can often be done in a distributed fashion, as described in Chapter 14.7. Specifically, distributed power control algorithms exist that insure all users meet their threshold SINR levels as long as these SINRs are feasible. These algorithms can also modified to prevent user access when this user cannot be accommodated without compromising the target SINRs of existing users. Power control for multiple access can also help users meet delay constraints in a random access environment. Power control has been extensively studied for cellular systems ([43] and the references therein). However, there are few results outside of [8] on the design and performance of power control schemes in ad hoc wireless networks, and this remains an active area of research.

16.3 Network Design Issues

16.3.1 Neighbor Discovery and Network Connectivity

"Neighbor discovery" is one of the first steps in the initialization of a network of randomly distributes nodes. From the perspective of the individual node, this is the process of determining the number and identity of network nodes with which direct communication can be established given some maximum power level and minimum link performance requirements (typically in terms of data rate and associated BER). Clearly the higher the allowed transmit power, the greater the number of nodes in a given neighborhood.

Neighbor discovery begins with a probe of neighboring nodes using an initial power constraint. If the number of nodes thus contacted is insufficient to ensure some minimal connectivity requirements then the power constraint is relaxed and probing repeated. The minimal connectivity requirements will depend on the application, but most ad hoc wireless network applications assume a fully-connected network whereby each node can reach every other node, often through multiple hops. The exact number of neighbors that each node requires to obtain a fully-connected network depends on the exact network configuration but is generally on the order of six to eight for randomly distributed immobile nodes [3, 8]. An analysis of the minimum transmit power required at each node to maintain full connectivity is done in [59]. Clearly the ability of the network to stay connected will decrease with node mobility, and so maintaining full connectivity under high mobility will require larger neighborhoods and an associated increase in transmit power at each node. It is interesting to note that, given a random distribution of nodes, the likelihood of complete connectivity changes abruptly from zero to one as the transmission range of each node is increased [44]. Moreover, the transmission range required for the network to be fully connected increases as the node density decreases, reflecting the increased probability of deep holes, to borrow a term from

the theory of lattices. Connectivity is also heavily influenced by the ability to adapt various parameters at the link layer such as rate, power, and coding, since communication is possible even on links with low SINR if these parameters are adapted [15].

From the standpoint of power efficiency and operational lifetime, it is also very important that nodes be able to decide whether or not to take a nap. These sleep decisions must take into account network connectivity, so it follows that these decisions are local, but not autonomous. Mechanisms that support such decisions can be based on neighbor discovery coupled with some means for ordering decisions within the neighborhood. In a given area, the opportunity to sleep should be circulated among the nodes, ensuring that connectivity is not lost through the coincidence of several, identical decisions to go to sleep.

16.4 Routing

The multihop routing protocol in an ad hoc wireless network is a significant design challenge, especially under energy constraints where the exchange of routing data consumes precious energy resources. Most work in multihop routing protocols falls into three main categories: flooding, proactive routing (centralized or distributed), and reactive routing ([45, 46, 47] and the references therein).

In flooding a packet is broadcast to all nodes within receiving range. These nodes also broadcast the packet, and the forwarding continues until the packet reaches its ultimate destination. Flooding has the advantage that it is highly robust to changing network topologies and requires little routing overhead. In fact, in highly mobile networks flooding may be the only feasible routing strategy. The obvious disadvantage is that multiple copies of the same packet traverse through the network, wasting bandwidth and battery power of the transmitting nodes. This disadvantage makes flooding impractical for all but the smallest of networks.

The opposite philosophy to flooding is centralized route computation. In this approach information about channel conditions and network topology are determined by each node and forwarded to a centralized location that computes the routing tables for all nodes in the network. The criterion used to compute the "optimal" route depends on the optimization criterion: common criteria include minimum average delay, minimum number of hops, and recently, minimum network congestion. While centralized route computation provides the most efficient routing according to the optimality condition, it cannot adapt to fast changes in the channel conditions or network topology, and also requires much overhead for collecting local node information and then disseminating the routing information. Centralized route computation, like flooding, it typically only used in very small networks.

Distributed route computation is the most common routing procedure used in ad hoc wireless networks. In this protocol nodes send their connectivity information to neighboring nodes and then routes are computed from this local information. In particular, nodes determine the next hop in the route of a packet based on this local information. There are several advantages of distributed route computation. First, the overhead of exchanging routing information with local nodes is minimal. In addition, this strategy adapts quickly to link and connectivity changes. The disadvantages of this strategy are that global routes based on local information are typically suboptimal, and routing loops are often common in the distributed route computation.

Both centralized and distributed routing require fixed routing tables that must be updated at regular intervals. An alternate approach is reactive (on-demand) routing, where routes are created only at the initiation of a source node that has traffic to send to a given destination. This eliminates the overhead of maintaining routing tables for routes not currently in use. In this strategy a source node initiates a route-discovery process when it has data to send. This process will determine if one or more routes are available to the destination. The route or routes are maintained until the source has no more data for that

particular destination. The advantage of reactive routing is that globally-efficient routes can be obtained with relatively little overhead, since these routes need not be maintained at all times. The disadvantage is that reactive routing can entail significant delay, since the route discovery process is initiated when there is data to send, but this data cannot be transmitted until the route discovery process has concluded. Recently a combination of reactive and proactive routing has been proposed to reduce the delay associated with reactive routing as well as the overhead associated with proactive routing [46].

Mobility has a huge impact on routing protocols as it can cause established routes to no longer exist. High mobility especially degrades the performance of proactive routing, since routing tables quickly become outdated, requiring an enormous amount of overhead to keep them up to date. Flooding is effective in maintaining routes under high mobility, but has a huge price in terms of network efficiency. A modification of flooding called multipath routing has been recently proposed, whereby a packet is duplicated on only a few paths with a high likelihood of reaching its final destination [47]. This technique has been shown to perform well under dynamically changing topologies.

Energy constraints in the routing protocol significantly change the problem. First of all, the exchange of routing information between nodes entails an energy cost: this cost must be traded against the energy savings that result from using this information to make routes more efficient. In addition, even with perfect information about the links and network topology, the route computation must change to take energy constraints into account. Specifically, a route utilizing a small number of hops (low delay) may use significantly more energy (per node and/or total energy) than a route consisting of a larger number of hops. Moreover, if one node is often used for forwarding packets the battery of that node will die out quickly, making that node unavailable for transmitting its own data or forwarding packets for others. Thus the routing protocol under energy constraints must somehow balance delay constraints, battery lifetime, and routing efficiency.

There has been much recent work on evaluating routing protocols under energy constraints. In [48] simulations were used to compare the energy consumption of different well-known routing protocols. Their results indicate that reactive routing is more energy-efficient. This is not surprising since proactive routing must maintain routing tables via continuous exchange of routing information, which entails a significant energy cost. This work was extended in [49] to more accurately model the energy consumption of radios in a "listening" mode. The energy consumption for this mode, ignored in [48], was significant and based on this more accurate model it was concluded that the proactive and reactive routing schemes analyzed in [48] have roughly the same energy consumption. The paper goes on to propose a sleep mode for nodes that reduces energy consumption by up to 40control and adaptive coding to minimize the energy cost of routes Power control to optimize energy-efficiency in routing is also studied in [50].

16.4.1 Scalability and Distributed Protocols

Scalability arises naturally in the design of self-configuring ad hoc wireless networks. The key to self-configuration lies in the use of distributed network control algorithms: algorithms that adjust local performance to account for local conditions. To the extent that these algorithms forgo the use of centralized information and control resources, the resulting network will be scalable. Work on scalability in ad hoc wireless networks has mainly focused on self-organization [10, 51], distributed routing [52], mobility management [4], QoS support, and security [54]. Note that distributed protocols often consume a fair amount of energy in local processing and message exchange: this is analyzed in detail for security protocols in [55]. Thus interesting tradeoffs arise as to how much local processing should be done versus transmitting information to a centralized location for processing. Most work on scalability in ad hoc wireless networks has focused on relatively small networks, less than 100 nodes. Many ad-hoc network applications, especially sensor networks, could have hundreds to thousands of nodes or even more. The

ability of existing network protocols to scale to such large network sizes remains an open question.

16.4.2 Network Capacity

The fundamental capacity limit of an ad hoc wireless network - the set of maximum data rates possible between all nodes - is a highly challenging problem in information theory. In fact, the capacity for simple channel configurations within an ad hoc wireless network such as the general relay and interference channel remain unsolved [56]. In a recent landmark paper an upper bound on the performance of an asymptotically large ad hoc wireless network in terms of the uniformly achievable maximum data rate was determined [57]. Surprisingly this result indicates that even with optimal routing and scheduling, the per-node rate in a large ad hoc wireless network goes to zero. To a large extent this pessimistic result indicates that in a large network all nodes should not communicate with all other nodes: there should be distributed processing of information within local neighborhoods. This work was extended in [58] to show that node mobility actually increases the per-node rate to a constant, i.e. mobility increases network capacity. This result follows from the fact that mobility introduces variation in the network that can be exploited to improve per-user rates. Other recent work in this area has determined achievable rate regions for ad hoc wireless networks using adaptive transmission strategies [34] and an information theoretic analysis on achievable rates between nodes [59].

16.5 Application Design Issues

In true cross layer protocol design, the highest layer - the application - can play a significant role in network efficiency. In this section we consider network adaptation to the application requirements and application adaptation to the underlying network capabilities.

16.5.1 Adaptive QoS

The Internet today, even with high-speed high-quality fixed communication links, is unable to deliver guaranteed QoS to the application in terms of guaranteed end-to-end rates or delays. For ad hoc wireless networks, with low-capacity error-prone time- varying links, mobile users, and a dynamic topology, the notion of being able to guarantee these forms of QoS is simply unrealistic. Therefore, ad hoc wireless network applications must adapt to time-varying QoS parameters offered by the network. While adaptivity at the link and network level as described in previous sections will provide the best possible QoS to the application, this QoS will vary with time as channel conditions, network topology, and user demands change. Applications must therefore adapt to the QoS that is offered. There can also be a negotiation for QoS such that users with a higher priority can obtain a better QoS by lowering the QoS of less important users.

As a simple example, the network may offer the application a rate- delay tradeoff curve which is derived from the capabilities of the lower layer protocols [25]. The application layer must then decide at which point on this curve to operate. Some applications may be able to tolerate a higher delay but not a lower overall rate. Examples include data applications in which the overall data rate must be high but latency might be tolerable. Other applications might be extremely sensitive to delay (e.g. a distributed-control application) but might be able to tolerate a lower rate (e.g. via a coarser quantization of sensor data). Energy constraints introduce another set of tradeoffs related to network performance versus longevity. Thus, these tradeoff curves will typically be multidimensional to incorporate rate, delay, bit-error-rate, longevity, etc. These tradeoff curves will also change with time as the number of users on the network and the network environment change.

16.5.2 Application Adaptation and Cross Layer Design Revisited

In addition to adaptive QoS, the application itself can adapt to the QoS offered. For example, for applications like video with a hard delay constraint, the video compression algorithm might change its compression rate such that the source rate adjusts to the rate the network can deliver under the delay constraint. Thus, under poor network conditions compression would be higher (lower transmission rate) and the end quality would be poorer. There has been much recent work on application adaptation for wireless networks ([60, 61, 53] and the references therein). This work indicates that even demanding applications like video can deliver good overall performance under poor network conditions if the application is given the flexibility to adapt.

The concept of application adaptation returns us to the cross layer design issue discussed earlier. While the application can adapt to a rate-delay-performance tradeoff curve offered by the network and underlying links, by making the lower layer protocols aware of the tradeoffs inherent to the application adaptation, that tradeoff curve might be adjusted to improve end-to-end performance without using up more resources in the network. In other words, if the application is aware of the lower layer protocol tradeoffs and these protocols aware of the application tradeoffs, these tradeoffs curves can be merged to operate at the best point relative to end-to-end performance. While implementing this philosophy remains a wide open research problem, it holds significant promise for the performance of ad hoc wireless networks.

Bibliography

[1] F. A. Tobagi, "Modeling and performance analysis of multihop packet radio networks," Proc. of the IEEE, pp. 135–155, January 1987.

[2] M.B. Pursley, "The role of spread spectrum in packet radio networks", IEEE Proc., Jan. 1987.

[3] L. Kleinrock and J. Silvester. "Optimum Transmission Radii for Packet Radio Networks or Why Six is a Magic Number." Proc. IEEE Natl. Telecomm. Conf., pages 4.3.1- 4.3.5, Dec. 1978.

[4] S. Basagni, D. Turgut, and S.K. Das, "Mobility- adaptive protocols for managing large ad hoc networks," Proc. IEEE Int. Commun. Conf (ICC), pp. 1539-1543, June 2001.

[5] J. Haartsen, "The Bluetooth radio system," IEEE Pers. Commun. Mag., pp. 28-36, Feb. 2000.

[6] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," IEEE Trans. Vehic. Technol., pp. 57-62, Feb. 1992.

[7] A.J. Goldsmith and S.G. Chua, "Variable-rate variable-power MQAM for fading channels," IEEE Trans. Commun. 1218-1230, Oct. 1997.

[8] N. Bambos, "Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design aspects," IEEE Pers. Commun. Mag., pp. 50-59, June 1998.

[9] W. R. Heinzelman, A. Sinha, and A. P. Chandrakasan, "Energy-scalable algorithms and protocols for wireless microsensor networks," *Proc. IEEE Intl. Conf. Acous., Speech, Signal Process.,* pp. 3722–3725, June 2000.

[10] K. Sohrabi, J. Gao, V. Ailawadhi, and G. Pottie, "Protocols for self-organization of a wireless sensor network," IEEE Pers. Commun. Mag., pp. 16-27, Oct. 2000.

[11] P. Agrawal, "Energy efficient protocols for wireless systems," Proc. IEEE Intl. Symp. Personal, Indoor, Mobile Radio Commun., pp. 564-569, Sept. 1998.

[12] J.M. Kahn, R.H. Katz, and K.S. Pister, "Emerging challenges: mobile networking for "Smart Dust", J. Commun. Networks, pp. 188-196, Aug. 2000.

[13] A. Chandrakasan and R.W. Brodersen, "Low Power Digital CMOS Design. Kluwer Academic Publishers, Norwell, MA 1995.

[14] B. Leiner, R. Ruther, A. Sastry, "Goals and challenges of the DARPA Glomo program (global mobile information systems)", IEEE Pers. Commun. Magazine, pp. 34-43, Dec. 1996.

[15] R. Ramanathan, R. Rosales-Hain, "Topology control of multihop wireless networks using transmit power adjustment,". Proc. IEEE INFOCOM, pp. 404–413, March 2000.

[16] M. Ritter, R. Friday, M. Cunningham, "The architecture of metricom's microcellular data network and details of its implementation as the 2nd and 3rd generation ricochettm wide- area mobile data service", IEEE Emerging Technologies Symposium on Broadband Communications for the Internet Era, pp. 143-152, 2001.

[17] M.N. Huhns, "Networking embedded agents," IEEE Internet Computing, pp. 91-93, Jan/Feb. 1999.

[18] W.W. Gibbs "As we may live" by W. Wayt Gibbs, Scientific America, Nov. 2000.

[19] K. Negus, R. Stephens, . Lansford, "HomeRF: wireless networking for the connected home" IEEE Pers. Commun. Mag, pp. 20-27, Feb. 2000

[20] A. Schmidt, "How to build smart appliances," IEEE Pers. Commun. Mag. pp. 66-71, Aug. 2001.

[21] B. P. Crow, I. Widjaja, J. G. Kim, and P. T. Sakai, "IEEE 802.11 Wireless Local Area Networks", IEEE Commun. Mag., pp. 116-126, Sept. 1997.

[22] J. Haartsen and S. Mattisson, "Bluetooth: a new low-power radio interface providing short-range connectivity," IEEE Proc., pp. 1651-1661, Oct. 2000.

[23] J. Rabaey, M.. Ammer, J. L. da Silva, Jr., D. Roundy, "PicoRadio supports ad hoc untra-low power wireless networking," IEEE Computer, pp. 42-48, July 2000.

[24] J. Nilsson, B. Bernhardsson, and B. Wittenmark, "Stochastic analysis and control of real-time systems with random time delays," Automatica, pp. 57-64, 1998.

[25] X. Liu, S.S. Mahal, A. Goldsmith, and J.K. Hedrick, "Effects of communication delay on string stability in vehicle platoons," IEEE Intl. Conf. Intell. Transp. Sys., Aug. 2001.

[26] Sergio Verdu, "On channel capacity per unit cost," IEEE Trans. Inform. Theory, pp. 1019-1030, Sept. 1990.

[27] H. Mandyam and A.J. Goldsmith, "Capacity of finite energy channels," Proc. Allerton Conf. Commun. Contl. Comp., Oct. 2001.

[28] S. Kandukuri and N. Bambos, "Power controlled multiple access (PCMA) in wireless communication networks," Proc. IEEE Infocom, pp. 386-395, March 2000.

[29] T. Holliday and A. Goldsmith, "Wireless link adaptation policies: QoS for deadline constrained traffic with imperfect channel estimates," To appear: Proc. IEEE Intl. Conf. Commun. (ICC), April 2002.

[30] M.-S. Alouini, X. Tang, and A.J. Goldsmith, "An adaptive modulation scheme for simultaneous voice and data transmission over fading channels," IEEE J. Select. Areas. Commun., pp. 837-850, May 1999.

[31] S.-T. Chung and A. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," IEEE Trans. Commun. pp. 1561-1571, Sept. 2001.

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

[33] M.-S. Alouini and A.J. Goldsmith, "Area spectral efficiency of cellular mobile radio systems," IEEE Trans. Vehic. Technol., pp. 1047-1066, July 1999.

[34] S. Toumpis and A. Goldsmith, "Capacity regions for ad hoc networks", ICC 2002, April 2002. Also to appear *IEEE Trans. Wireless Commun.*

[35] S.A. Jafar, and A.J. Goldsmith, "Adaptive multicode CDMA for uplink throughput maximization," Proc. IEEE Vehic. Technol. Conf., pp546-550, May 2001. Also submitted to IEEE J. Select Areas Commun.

[36] S. Kandukuri and S. Boyd, "Simultaneous rate and power control in multirate multimedia CDMA systems," IEEE Intl. Symp. Spread Spec. Tech. Appl., pp. 570-574, Sept. 2000.

[37] X. Tang and A. Goldsmith, "Admission control and adaptive CDMA for integrated voice and data systems," Proc. IEEE Vehic. Technol. Conf, May 2001.

[38] C. Chien, M. Srivastava, R. Jain, P. Lettieri, V. Aggarwal, and R. Sternowski, "Adaptive radio for multimedia wireless link," IEEE J. Select. Areas Commun. pp. 793-819, May 1999.

[39] S. Kallel, "Analysis of memory and incremental redundancy ARQ schemes over a nonstationary channel," IEEE Trans. Commun., pp. 1474-1480, Sept. 1992.

[40] X. Qiu, J. Chuang, K. Chawla, and J. Whitehead, "Performance comparison of link adaptation and incremental redundancy," Proc., IEEE. Wireless Commun. Net. Conf., pp. 771- 775, Sept. 1999.

[41] A. B. MacKenzie and S. B. Wicker, "Selfish Users in Aloha: A Game-Theoretic Approach," Proc. IEEE Vehic. Technol. Conf., pp. 1354-1357, Oct. 2001.

[42] E. Uysal-Biyikoglu, B.Prabhakar and A. El Gamal , "Energy-Efficient Transmission of Packets in a Wireless Network", to appear in IEEE Trans. Net. Also to appear in Proc. IEEE Infocom, March 2002.

[43] R.D. Yates, "A framework for uplink power control in cellular radio systems," IEEE J. Select. Areas Commun., pp. 3141- 1347, Sept. 1995.

[44] B. Krishnamachari, S. B. Wicker, R. Bejar, and M. Pearlman, "Critical Density Thresholds in Distributed Wireless Networks," to appear in the Festschrift for Ian Blake, 2002.

[45] E. Royer and C.-K. Toh, "A review of current routing protocols for ad hoc mobile wireless networks", , IEEE Pers. Commun. Mag., pp. 46-55, April 1999.

[46] M.R. Pearlman, Z.J. Haas, and S.I.Mir, "Using routing zones to support route maintenance in ad hoc networks," Proc. IEEE Wireless Commun. Net. Conf., pp. 1280-1284, Sept. 2000.

[47] A. Tsirigos and Z.J. Haas, "Multipath routing in the presence of frequency topological changes," IEEE Commun. Mag., pp. 132-138, Nov.2001.

[48] J.-C. Cano and P. Manzoni, "Evaluating the energy- consumption reduction in a MANET by dynamically switching-off network interfaces," Proc. IEEE Symp. Comp. Commun., pp. 186- 191, 2001.

[49] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," To appear, Proc. IEEE Infocom, March 2002.

[50] A. Michail and A. Ephremides, "Energy efficient routing for connection oriented traffic in ad-hoc wireless networks," Proc. IEEE Pers. Indr. Mob. Radio Commun. Conf., pp. 762-766, Sept. 2000.

[51] Subramanian, L.; Katz, R.H., "An architecture for building self-configurable systems", Mobile and Ad Hoc Networking and Computing, 2000.

[52] Jain, R.; Puri, A.; Sengupta, R. "Geographical routing using partial information for wireless ad hoc networks", IEEE Pers. Commun. Mag., pp. 48-57, Feb. 2001.

[53] R. Ramanathan and R. Hain, "An ad hoc wireless testbed for scalable, adaptive QoS support," IEEE WCNC, pp. 998-1002, Nov. 2000.

[54] L. Zhou and Z.J. Haas, "Securing ad hoc networks," IEEE Network, pp. 24-30, Nov/Dec. 1999.

[55] R. Karri and P. Mishra, "Energy management of secure wireless sessions," Preprint.

[56] T. Cover and J.A. Thomas, Elements of Information Theory, Wiley Interscience, New York, 1991.

[57] P. Gupta and P.R. Kumar, "The capacity of wireless networks," IEEE Trans. Inform. Theory, pp. 388-404, March 2000.

[58] M. Grossglauer and D.N. Tse, "Mobility increases the capacity of ad-hoc wireless networks," Proc. IEEE Infocom, pp. 1360-1369, March 2001

[59] P. Gupta and P.R. Kumar, "Towards an information theory of large networks: an achievable rate region," Proc. IEEE Intl. Symp. Inform. Theory, p. 159, June 2001.

[60] A. Alwan, R. Bagrodia, N. Bambos, M. Gerla, L. Kleinrock, J. Short, and J. Villasenor, "Adaptive mobile multimedia networks," IEEE Pers. Commun. Mag., pp. 34-51, April 1996.

[61] M. Mirhakkak, N. Schult, and D. Thomson, "Dynamic bandwidth management and adaptive applications for a variable bandwidth wireless environment," IEEE J. Select. Areas Commun., pp. 1985-1997, Oct. 2001.

Appendix A

Representation of Bandpass Signals and Systems

Many signals in communication systems are bandpass signals occupying a narrow bandwidth B centered around a carrier frequency f_c with $B << f_c$, as illustrated in Figure 16.4. Such signals result from modulation of a baseband signal by a carrier. The bandwidth B of a bandpass signal is roughly equal to the range of frequencies around f_c where the signal has nonnegligible amplitude, and bandpass signals have $B << f_c$.

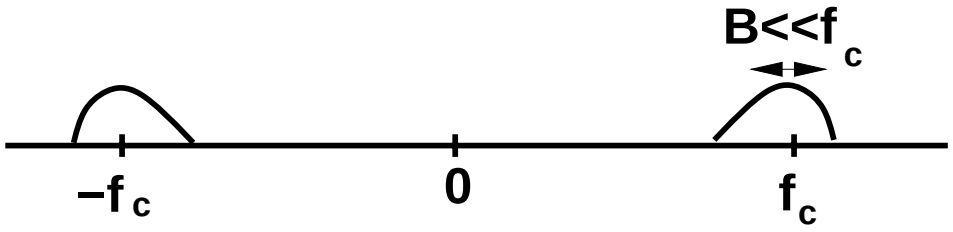


Figure 16.4: Bandpass Signal.

In practice only real signals can be processed at the transmitter and receiver, so both transmitted and received signals must be modeled as real. However, channel models are often based on a real impulse response in the time domain, which in general has a complex Fourier transform. Thus, we require signal models that capture the effect of a complex channel frequency response operating on a real transmitted signal.

Let us begin by representing a bandpass signal $s(t)$ at carrier frequency f_c in the following form:

$$s(t) = x(t)\cos(2\pi f_c t) - y(t)\sin(2\pi f_c t), \tag{16.1}$$

where $x(t)$ and $y(t)$ are real baseband signals of bandwidth $B << f_c$ and we neglect the initial phase of the carrier since it does not affect our development. This is a common representation for bandpass signals and noise. In fact, modulations such as MPSK and MQAM are based on this representation. We call $x(t) = \Re\{u(t)\}$ the **in-phase component** of $s(t)$ and $y(t) = \Im\{u(t)\}$ the **quadrature component** of $s(t)$. Let us define the complex signal $u(t) = x(t) + jy(t)$. Then $u(t)$ is a complex baseband signal of bandwidth B. With this definition we see that

$$s(t) = \Re\{u(t)\}\cos(2\pi f_c t) - \Im\{u(t)\}\sin(2\pi f_c t) = \Re\left\{u(t)e^{j(2\pi f_c t+\phi_0)}\right\}. \tag{16.2}$$

The representation on the right hand side of this equation is called the **complex baseband representation** of the bandpass signal $s(t)$, and the baseband signal $u(t)$ is called the **equivalent baseband signal** for $s(t)$ or its **complex envelope**.

Using properties of the Fourier transform we can show that

$$S(f) = .5[U(f - f_c) + U^*(-f - f_c)]. \tag{16.3}$$

Since $s(t)$ is real, $S(f)$ is symmetric about $f = 0$. However, the lowpass signals $U(f)$ and $U^*(f)$ are not necessarily symmetric about $f = 0$, which leads to an asymmetry of $S(f)$ about the carrier frequency f_c

as shown in Figure 16.4. In fact, $S(f)$ is only symmetric about the carrier frequency if $u(t) = x(t)$, i.e. if there is no quadrature component in $u(t)$. We will see shortly that this asymmetry affects the response of bandpass systems to bandpass signals.

An alternate representation of the equivalent baseband signal is

$$u(t) = a(t)e^{j\theta(t)}, \tag{16.4}$$

with envelope

$$a(t) = \sqrt{x^2(t) + y^2(t)}, \tag{16.5}$$

and phase

$$\theta(t) = \tan^{-1}\left\{\frac{y(t)}{x(t)}\right\}. \tag{16.6}$$

With this representation

$$s(t) = \Re\left\{a(t)e^{j\theta(t)}e^{j(2\pi f_c t)}\right\} = a(t)\cos(2\pi f_c t + \theta(t)). \tag{16.7}$$

Let us now consider a real channel impulse response $h(t)$ with Fourier transform $H(f)$. If $h(t)$ is real then $H^*(-f) = H(f)$. In communication systems we are mainly interested in the channel frequency response around the center frequency f_c and within the bandwidth B of the transmitted signal, since after filtering in the receiver only these frequency components of $H(f)$ affect the received signal. A **bandpass channel** is similar to a bandpass signal, with frequency response $H(f)$ centered at f_c with a bandwidth of $B << f_c$. To capture the frequency response of $H(f)$ around f_c, we develop an **equivalent baseband system** model similar to the equivalent baseband signal model as follows. Since $H(f)$ is a bandpass filter, using the previous development the frequency response $h(t)$ of $H(f)$ can be written in its complex baseband representation

$$h(t) = 2\Re\left\{h_l(t)e^{j2\pi f_c t}\right\}, \tag{16.8}$$

where the extra factor of 2 will be explained shortly. We call $h_l(t)$ the **baseband equivalent filter** for $H(f)$. Using (16.3) this implies that

$$H(f) = H_l(f - f_c) + H_l^*(-f - f_c), \tag{16.9}$$

so $H(f)$ consists of two components: $H_l(f)$ shifted up by f_c, and $H_l^*(f)$ shifted down by f_c. Note that if $H(f)$ is symmetric about the carrier frequency f_c then $h_l(t)$ will be real and its frequency response $H_l(f)$ symmetric about zero. However, in many systems and channels, e.g. frequency-selective fading channels, $H(f)$ is not symmetric about f_c, in which case $h_l(t)$ is complex with in-phase component $h_x(t) = \Re\{h_l(t)\}$ and quadrature component $h_y(t) = \Im\{h_l(t)\}$. Note that if $h_l(t)$ is complex then $H_l(f)$ is not symmetric about zero and thus $H(f)$ is not symmetrics about f_c.

We now use these canonical representations to study the output of a channel to a bandpass signal input. Let $s(t)$ denote the input signal with equivalent lowpass signal $u(t)$. Let $h(t)$ denote the channel impulse response with equivalent lowpass response $h_l(t)$. The channel output $r(t) = s(t) * h(t)$ or, equivalently, $R(f) = H(f)S(f)$. Since $S(f)$ is a bandpass signal, $R(f)$ will also be a bandpass signal, so we can represent it as

$$r(t) = \Re\left\{v(t)e^{j(2\pi f_c t + \phi_0)}\right\}. \tag{16.10}$$

We now consider the relationship between the lowpass equivalent signals corresponding to the channel input $s(t)$, channel impulse response $h(t)$, and channel output $r(t)$. We can express the frequency response of the channel output as

$$R(f) = H(f)S(f) = .5[U(f - f_c) + U^*(-f - f_c)][H_l(f - f_c) + H_l^*(-f - f_c)]. \tag{16.11}$$

For bandpass signals and channels where the bandwidth of $u(t)$ is much less than the carrier frequency, we have

$$U(f - f_c)H_l^*(-f - f_c) = 0$$

and

$$U^*(-f - f_c)H_l(f - f_c) = 0$$

. Thus,

$$R(f) = .5[U(f - f_c)H_l(f - f_c) + U^*(-f - f_c)H_l^*(-f - f_c)]. \tag{16.12}$$

From (16.10) we also have that $R(f) = .5[V(f - f_c) + V^*(-f - f_c)]$. Combining this with (16.12) yields that

$$V(f - f_c) = U(f - f_c)H_l(f - f_c)$$

and

$$V^*(-f - f_c) = U^*(-f - f_c)H_l^*(-f - f_c)$$

or, equivalently, that

$$V(f) = H_l(f)U(f) \tag{16.13}$$

and therefore $v(t) = h_l(t) * u(t)$. In other words, we can obtain the lowpass equivalent signal $v(t)$ for $r(t)$ by taking the convolution of $h_l(t)$ and $u(t)$. The received signal is therefore given by

$$r(t) = \Re\left\{u(t) * h_l(t)e^{j(2\pi f_c t)}\right\}. \tag{16.14}$$

Note that $V(f) = H_l(f)U(f)$ is symmetric about $f = 0$ only if both $U(f)$ and $H_l(f)$ are. In other words, the lowpass equivalent received signal will have both in-phase and quadrature components if either $u(t)$ or $h_l(t)$ are complex. Moreover, if $u(t) = x(t)$ is real (no quadrature component) but the channel impulse resonse $h_l(t) = h_x(t) + jh_y(t)$ is complex (e.g. in frequency-selective fading) then

$$v(t) = x(t) * h_x(t) + jx(t) * h_y(t)$$

is complex, so the received signal will have both an in-phase and quadrature component. More generally, if $u(t) = x(t) + jy(t)$ and $h_l(t) = h_x(t) + jh_y(t)$ then

$$v(t) = [x(t) + jy(t)] * [h_x(t) + jh_y(t)] = [x(t) * h_x(t) - y(t) * h_y(t)] + j[x(t) * h_y(t) + y(t) * h_x(t)]$$

. So the in-phase component of $v(t)$ depends on *both* the in-phase and quadrature components of $u(t)$, and similarly for the quadrature component of $u(t)$. As we discuss in Chapter 6, this creates problems in detection of modulated signals, since it causes the in-phase and quadrature parts of the modulated signal to interfere with each other.

The main point of this appendix is to show that in studying bandpass signals and systems, we can either do the analysis at the bandpass frequency or analyze just the lowpass equivalent models for the transmitted signal, channel impulse response, and received signal. These lowpass equivalent models are often used in analyzing communication systems to remove the dependency on a fixed carrier frequency f_c and to simplify the analysis.