Space-Time Codes and MIMO Systems

DISCLAIMER OF WARRANTY

The technical descriptions, procedures, and computer programs in this book have been developed with the greatest of care and they have been useful to the author in a broad range of applications; however, they are provided as is, without warranty of any kind. Artech House, Inc. and the author and editors of the book titled *Space-Time Codes and MIMO Systems* make no warranties, expressed or implied, that the equations, programs, and procedures in this book or its associated software are free of error, or are consistent with any particular standard of merchantability, or will meet your requirements for any particular application. They should not be relied upon for solving a problem whose incorrect solution could result in injury to a person or loss of property. Any use of the programs or procedures in such a manner is at the user's own risk. The editors, author, and publisher disclaim all liability for direct, incidental, or consequent damages resulting from use of the programs or procedures in this book or the associated software.

For a listing of recent titles in the
Artech House Universal Personal Communications Series,
turn to the back of this book.

Space-Time Codes and MIMO

Systems Mohinder Jankiraman

Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data

Jankiraman, Mohinder.
Space-time codes and MIMO systems / Mohinder Jankiraman.
p. cm. — (Artech House universal personal communications series)
Includes bibliographical references and index.
ISBN 1-58053-865-7 (alk. paper)
1. Space time codes. 2. MIMO systems. 3. Wireless communication systems. 4. Antenna arrays. I. Title. II. Series.

TK5103.4877.J36 2004

621.382—dc22 2004050668

British Library Cataloguing in Publication Data
Jankiraman, Mohinder
Space-time codes and MIMO systems. — (Artech House universal personal communications library)
1. Coding theory 2. Wireless communication systems
I. Title
621.3’822

ISBN 1-58053-865-7

Cover design by Yekaterina Ratner

685 Canton Street
Norwood, MA 02062

International Standard Book Number: 1-58053-865-7
Library of Congress Catalog Card Number: 2004050668

10 9 8 7 6 5 4 3 2 1

Dedicated to my late father
Kuppuswamy Janakiraman
and to my mother
Gnanambal Janakiraman
Their extraordinary sacrifices made it possible for me to come this far

Contents

Preface

This book is intended to introduce space-time coding and multiantenna systems. The endeavor is to impart a working knowledge of the subject not just for students and researchers but for the entire wireless community.

The birth of multiantenna systems is the direct result of the long-standing struggle to achieve data rates without compromising the quality of the reception. Indeed this has been the case since the inception of wireless communications. A binding constraint in the evolution of high data rate systems is the stringent limita tion imposed on the available spectrum. This, in turn, has given rise to more efficient signaling techniques. Recent study has shown that multiple antennas yield substantial increases in channel capacity. Toward this end, multiple-input multiple output (MIMO) systems have been constructed comprising multiple antenna arrays at both ends of the wireless link. Space-time coding, as the name suggests, involves coding across space and time and is aimed at approaching the capacity limits of MIMO channels. Today space-time coding and MIMO systems are widely regarded as the most likely candidates for futuristic high data rate systems and are already being designed by many companies for the high data rate market.

This book is intended for postgraduate students, practicing engineers, and researchers. It is assumed that the reader is familiar with basic digital communica tions, linear algebra, and probability theory.

In view of the fact that space-time coding theory has become such a widely discussed and researched topic in recent times, it is not possible to cover all aspects in any detail. Therefore, an effort has been made to impart to the reader a ‘‘flavor’’ of the subject just enough to whet his/her appetite, prompting further detailed study of areas of particular interest. Toward this end, the style of writing has been kept as simple as possible and technical cliche´s and jargon have been kept to a minimum. All effort has been made to explain the basics in a cogent and conversa tional manner.

The reader is also introduced to a new technique of interfacing CDMA to OFDM systems called ‘‘Hybrid OFDM/CDMA.’’ This approach is different from the popular OFDM/CDMA systems proposed by Fazel et al [1], wherein an MC-CDMA system transmits each bit using OFDM modulation. In the approach

suggested by the author a CDMA system is directly interfaced to an OFDM system. This approach yields a CDMA system capable of handling high throughputs, bandwidth permitting, along with the added bonus of user separation based on CDMA codes. It is, to the best of the author's knowledge and belief, the first time such a concept has been discussed anywhere and had formed part of the author's Ph.D. thesis [2]. This concept has been recently proved in the field by NTT DoCoMo

of Japan, wherein the company tested such a system for outdoor use using a bandwidth of 100 MHz to achieve a throughput of 100 Mbps [3] in the downlink. Further details are not known at the time of going to press. The company plans to introduce this, as part of its 4G effort, by 2010.

The teaching effort in this book is aided by a set of accompanying software. This software has been divided into a set of two broad classes—narrowband and wideband. The narrowband software is distributed on the basis of chapters and directly pertains to topics discussed in those chapters. The wideband software is orthogonal frequency division multiplexing (OFDM) based and is included as part of Chapters 7, 8, and 9. The entire coding has been kept simple and sometimes very unprofessional to enable readers to clearly understand the various steps involved in the implementation of the program. If the reader finds this annoying, the error is mine and is deeply regretted! The entire coding has been implemented at baseband and the radio frequency (RF) aspects of coding have been avoided for similar reasons. Because the software is basically intended as a ''skeleton,'' the user is encouraged to modify it in any manner or means by adding to its RF capability and so on to suit one's convenience. This is an excellent method to learn the subject. The software presupposes a sound understanding of MATLAB® and SIMULINK® and has been tested on MATLAB® Version 6.0 (with Signal Processing and Commu nication Toolboxes) and above with the SIMULINK® option with DSP and Commu nication Blocksets. It is important to note that in a technology of this nature the best way to assimilate the subject is by programming. Coding an operation forces the user to look at all aspects of the subject. This is similar to learning mathematics through solving problems.

A consistent set of notations has been used throughout the book and excessive mathematics has been avoided. Emphasis is placed on imparting to the reader a physical understanding of the subject so that the reader has a clear grasp of the processes involved.

There will be errors even though every effort has been made to detect and eliminate them. Any inconvenience to the readers as a result is deeply regretted.

References

[1] Fazel, K., and L. Papke, ''On the Performance of Convolutionally-Coded CDMA/OFDM for Mobile Radio Communication Systems,'' *PIMRC,* 1993.

[2] Jankiraman, M., ''Wideband Multimedia Solution Using Hybrid CDMA/OFDM/SFH Techniques,'' *Ph.D. Thesis,* University of Aalborg, Denmark, September 2000. [3] ''NTT DoCoMo Successsfully Completes 4G Mobile-Communications Experiment Includ ing 100

Mbps Transmission," *nG Japan,* Vol. 1, No. 11, November 25, 2002, InfoCom Research Inc., Tokyo, Japan.

Acknowledgments

It is not possible to generate a work of this nature without invaluable help from many other participants, since no man is an island. It is not possible to list all of them but I gratefully acknowledge those who helped in the preparation of this book. This invaluable help came in two principal areas—the preparation of the text and the coding of the software. I gratefully thank Dr. John Terry of Nokia Research Center, Dallas, Texas, and Michelle Johnson of Pearson Education for permission to reproduce certain figures in this book. I also thank Professor Ted Rappaport of the University of Texas at Austin for his help with obtaining permis sion for figures from his book, as well as Emily McGee of Pearson Education. My thanks are also due to Dr. Branka Vucetic of the University of Sydney, Australia, for her patience with my many questions. I also thank Duncan James of John Wiley & Sons for permission to use certain figures in this book. Special thanks are also due to Steve Thanos of Pentek for permission to reproduce Figure 1.3. I thank Dr. Donald Shaver of Texas Instruments for permission to reproduce Figure 1.2, and Dr. Peter Rysavy of Rysavy Research for permission to reproduce Figure 1.4(b). Thanks are also due to Dr. Richard van Nee of Airgonetworks, Netherlands, for permission to use certain figures from his book. My thanks are due to Professor Arogyaswami Paulraj of Stanford University and Ted Gerney of Cambridge Univer sity Press for permission to use certain figures in this book. Finally, I thank Jacque line Hansson and Claudio Stanziola of IEEE for their kind permission to reproduce certain papers and figures in this book. You all have been very patient with my requests!

In addition, I thank Devendra Prasad and Albena Mihovska of the University of Aalborg, Denmark, for their invaluable help with this book. Without their help, this effort would not have been possible.

The software developed for this book has many participants. Thanks are due to Beza Negash Getu of the University of Aalborg, Denmark, for his help with the software in Chapters 2 and 4. We had many hours of fruitful discussion. Thanks are also due to Dr. Persifoni Kyritsi of Stanford University for her help with the software in Chapter 6. Persa and I thoroughly analyzed the MIMO channel. I also thank Kamil Anis of the Department of Radio Engineering, Faculty of Electrical Engineering, at Czech Technical University in Prague for his invaluable help with the space-time trellis coding in Chapter 5. In fact many of his programming ideas are incorporated in the software of Chapter 5. I also thank Dr. John Terry and Juha Heiskala of Nokia Research Center in Dallas for helping me understand the coding aspects of OFDM-based MIMO. They were very understanding and patient. I also thank Martin Clark of Mathworks, the makers of MATLAB, for his help

in obtaining permission from Mathworks to use his software in Chapter 8. Finally, I thank Efrayim Metin of Versatel Nederland B.V., Amsterdam,

Netherlands, for his invaluable help with my many software problems. I am truly grateful.

A work of this nature also requires salubrious surroundings. I thank my sister Uma and her husband, Dr. M. J. Rangaraj of Monroe, Louisiana, for inviting me to their wonderful home on the Louisiana bayou. There is no better spot than sitting on the banks of the bayou on a sunny day and mulling over technical problems surrounded by cormorants, butterflies, and noisy frogs. I solved many of my knotty problems in Louisiana. I also thank my children, Pavan and Pallavi, for their patience and encouragement.

Finally, I wish to acknowledge the advice and comments from my anonymous reviewers. I thank them all. Thanks are also due to Christine Barnaby Daniele, Tiina Ruonamaa, Julie Lancashire, Judi Stone, and Jill Stoodley of Artech House for their patience with my many corrections and repeated ‘‘final versions.’’ I acknowledge their superb support and efficient handling of this publication project.

Introduction

‘‘If I have seen farther, it is by standing on the shoulders of giants.’’

—Sir Isaac Newton

1.1 The Crowded Spectrum

It is all about spectrum. Marconi pioneered the wireless industry 100 years ago. Today life does not seem possible without wireless in some form or the other. In fact wireless permeates every aspect of our lives. The demands on bandwidth and spectral availability are endless. Currently wireless finds its widest expression in fixed and mobile roles. In the fixed role, wireless is used extensively for data transfer, especially from desktop computers and laptops. In the mobile role, wireless networks provide mobility for use from fast vehicles for both voice and data. Consequently wireless designers face an uphill task of limited availability of radio frequency spectrum and complex time varying problems in the wireless channel, such as fading and multipath, as well as meeting the demand for high data rates. Simultaneously, there is an urgent need for better quality of service (QoS), compared with that obtainable from DSL and cable.

1.2 Need for High Data Rates

The gradual evolution of mobile communication systems follows the quest for high data rates, measured in bits/sec (bps) and with a high spectral efficiency, measured in bps/Hz. The first mobile communications systems were analog and are today referred to as systems of the first generation. In the beginning of the 1990s, the first digital systems emerged, denoted as *second generation* (2G) systems. In Europe, the most popular 2G system introduced was the *global system for mobile communi cations* (GSM), which operated in the 900 MHz or the 1,800-MHz band and supported data rates up to 22.8 kbit/s. In

many parts of the world today, GSM is still in vogue. Basically GSM is a cellular system [i.e., it typically uses a single *base transceiver station* (BTS), which marks the center of a cell and which serves several *mobile stations* (MS), meaning the users]. In the United States, the most popular 2G system is the *TDMA/136*, which is also a digital cellular system. TDMA stands for time-division multiple access.

To accomplish higher data rates, two add-ons were developed for GSM, namely *high-speed circuit switched data* (HSCSD) and the *general packet radio service* (GPRS), providing data rates up to 38.4 kbit/s and 172.2 kbit/s, respectively.

The demand for yet higher data rates forced the development of a new genera tion of wireless systems, the so-called *third generation* (3G). 3G systems are charac terized by a maximum data rate of at least 384 kbit/s for mobile and 2 mbit/s for indoors.

One of the leading technologies for 3G systems is the now well-known *universal mobile telephone system* (UMTS) [also referred to as *wideband code-division multi plex* (WCDMA) or UTRA FDD/TDD]. UMTS represents an evolution in terms of services and data speeds from today's ''second generation'' mobile networks. As a key member of the global family of 3G mobile technologies identified by the International Telecommunication Union (ITU), UMTS is the natural evolutionary choice for operators of GSM networks, currently representing a customer base of more than 850 million end users in 195 countries and representing over 70% of today's digital wireless market. UMTS is already a reality. Japan launched the world's first commercial WCDMA network in 2001, and WCDMA networks are now operating commercially in Austria, Italy, Sweden and the United Kingdom, with more launches anticipated during 2004. Several other pilot and precommercial trials are operational in the Isle of Man, Monaco, and other European territories. UMTS is also a cellular system and operates in the 2-GHz band. Compared with the 2G systems, UMTS is based on a novel technology. To yield the 3G data rates, an alternative approach was made with the *enhanced data rates for GSM evolution* (*EDGE*) *concept.* The EDGE system is based on GSM and operates in the same frequency bands. The significantly enhanced data rates are obtained by means of a new modulation scheme, which is more efficient than the GSM modulation scheme. As for GSM, two add-ons were developed for EDGE, namely *enhanced circuit switched data* (ECSD) and the *enhanced general packet radio service* (EGPRS). The maximum data rate of the EDGE system is 473.6 kbit/s, which is accomplished by means of EGPRS. EDGE was introduced in the United States as a generic air interface to enhance the TDMA/136 system. Some 200 operators worldwide are also giving their customers a taste of faster data services with so called 2.5G systems based on GPRS technology, a natural evolutionary stepping stone toward UMTS.

The new IEEE and High Performance Radio Local Area Network (HIPERLAN) standards specify bit rates of up to 54 mbit/s, although 24 mbit/s will be the typical rate used in most applications. Such high data rates impose large bandwidths, thus pushing carrier frequencies for values higher than the UHF band. HIPERLAN has frequencies allocated in the 5- and 17-GHz bands;

multimedia broadcasting systems (MBS) will occupy the 40- and 60-GHZ bands; and even the infrared band is being considered for broadband wired local area networks (WLANs).

A comparison of several systems, based on two of the key features (mobility and data rate) is shown in Figure 1.1 [1], where it is clear that no competition exists between the different approaches.

The applications and services of the various systems are also different. IEEE 802.11 is mainly intended for communications between computers (thus being an extension of WLANs). Future wireless *broadband applications* are likely to require

Figure 1.1 Comparison of mobility and data rates for several systems. (*From:* [1]. © 2000, Artech House. Reprinted with permission.)

data rates that are hundreds of megabits per second—up to 250 times the maximum data rate promised for UMTS. Such a broadband service could, for example, be wireless high-quality video conferencing (up to 100 mbit/s) or wireless virtual reality (up to 500 mbit/s, when allowing free body movements).

Therefore, the goal of the next generation of wireless systems—the *fourth generation* (4G)—is to provide data rates yet higher than the ones of 3G while granting the same degree of user mobility. 4G is the short term for fourth generation wireless, the stage of broadband mobile communications that will follow the still burgeoning 3G that is expected to reach maturity between 2003–2005. 4G services are expected to be introduced first in Japan, as early as 2006—four years ahead of the previous target date. The major distinction of 4G over 3G communications is increased data transmission rates, just as it is for 3G over 2G and 2.5G (the current state of wireless services, hovering somewhere between 2G and 3G). According to *NTT-DoCoMo,* the leading Japanese wireless company, the current download speed for i-Mode (mobile Internet service) data is, theoretically, 9.6 kbit/s, although in practice the rates tend to be slower. 3G rates are expected to reach speeds 200 times higher, and 4G to yield further increases, reaching 20–40 mbit/s (about 10–20 times the current rates of ADSL service). 4G is expected to deliver more advanced versions of the same improvements promised by 3G, such as enhanced

multimedia, smooth streaming video, universal access, and portability across all types of devices. Industry insiders are reluctant to predict the direction that less-than-immediate future technology might take, but 4G enhancements are expected to include world wide ‘‘roaming’’ capability. As was projected for the ultimate 3G system, 4G might actually connect the entire globe and be operable from any location on—or above— the surface of the earth. This aspect makes it distinctly different from the techno logies developed until now. These technologies were built for or overlaid onto

proprietary networking equipment. In fact the outlook for 3G is uncertain. Network providers in Europe and North America maintain separate standards bodies (3GPP for Europe and Asia; 3GPP2 for North America). These different standards bodies reflect the difference in air interface technologies. In addition to 3G’s technical challenges, there are problems from a financial aspect, such as justifying the large expense of building systems based on less-than-compatible 2G technologies. In contrast, 4G wireless networks that are Internet Protocol (IP)-based have an intrinsic advantage over their predecessors. IP tolerates a variety of radio protocols. It allows you to design a core network that gives you complete flexibility as to what shape the access network will take. One can support diverse technologies like 802.11, WCDMA, Bluetooth, HIPERLAN and so on, as well as some new CDMA protocols. A 4G IP network also has certain financial advantages. Equipment costs are much lower than what they used to be for 2G and 3G systems. Wireless service providers would no longer be bound by single-system vendors of proprietary equipment. An all-IP wireless core network would enable services that are sufficiently varied for consumers. This implies improved data access for mobile Internet devices. Currently wireless communications are heavily biased toward voice. However, studies indicate that growth in wireless data traffic is rising exponentially relative to the demand for voice traffic. This is especially true of 802.11 data transfer protocol, a wireless LAN standard developed by IEEE, as a distinct data access technology that can work in a variety of radio spectrums, including infrared. Chapters 8 and 9 are exclusively devoted to this technology. Because an all-IP core layer is easily accessi ble, it is ideally suited to meet this challenge. It will not be surprising if we leapfrog directly into 4G from the current 2.5G.

WLANs are already characterized by data rates significantly higher than 2 mbit/s. WLANs cover comparably small areas, often hot spots in shopping malls, airports, hotels or office buildings. Typical properties of WLANs are comparably small cells and mobile terminals moving, at most, with pedestrian speed.

The *Bluetooth* system, already commercially available, operates in the 2-GHz band and provides data rates of < 1 mbit/s. Bluetooth is normally employed to interconnect certain electronic devices rather than to establish a complete wireless LAN. A wireless Bluetooth interconnection for example between a laptop com puter and a *personal digital assistant* (PDA), is typically characterized by a short range of about 5–10m. The U.S. standard *IEEE 802.11b,* as implemented in current WLAN products, operates in the same band and provides data rates up to 11 mbit/s over a distance of 50–100m. The succeeding standard *IEEE 802.11a* delivers data rates up to 54 mbit/s using the 5-GHz band. The European standard corresponding to IEEE 802.11a is *HIPERLAN Type 2.* Data rates in WLANs could be significantly enhanced by exploiting yet higher frequency bands. In this context, for example, the 60-GHz

band is the subject of current research activities. In particular, a new WLAN standard for 60 GHz could be developed such that it is compatible with IEEE 802.11a or HIPERLAN/2. This would enable 60-GHz systems to use the 5-GHz band as a fallback option, in case the channel quality at 60 GHz becomes insufficient.

The first 4G systems are likely to be an *integration* of 3G systems and WLAN systems. By this means, considerable data rates can be granted at hot spots. On the other hand, the interworking of WLAN and 3G systems will provide a good

degree of mobility, given that *seamless* handover is accomplished between several heterogeneous systems.

Figure 1.2 reveals the possible candidates for 4G systems. The figure is self explanatory. There are, however, a few interesting points. The ‘‘hottest’’ candidates for 4G appear to be:

- *BWIF:* The Broadband Wireless Internet Forum (BWIF) [3] is the principal organization chartered with creating and developing next generation fixed wireless standards. The broadband wireless specifications are based on *vector orthogonal frequency division multiplexing* (VOFDM) technology and *data over cable service interface specification* (DOCSIS). BWIF was formed to address the needs of the quickly emerging wireless broadband market. Fur ther, through BWIF, members establish product road maps that lower prod uct costs, simplify deployment of advanced services, and ensure the availability of interoperable solutions. BWIF extends the partnership model to all companies offering expanded broadband wireless technology to multi ple markets. At the core of the partnerships, membership includes:
  - *ASIC semiconductor* companies, which develop new ASICs based on VOFDM technology;
  - *Customer premise equipment* (CPE) companies, which use the chips to build new subscriber equipment;
  - *Systems integrators,* which design and deploy the networks based on these products;
  - *Service providers,* which incorporate VOFDM products and technology into their network infrastructure to offer to new services customers;
  - *RF/ODU manufacturers,* which supply subsystems to the total wireless solution offering.

Figure 1.2 Possible candidates for 4G systems. (*From:* [2]. Reprinted with the permission of Donald Shaver, Texas Instruments.)

- Many companies associated with wireless technology see the market poten tial for new applications, products, and services. BWIF members have com mitted to the VOFDM specification for optimized and open broadband fixed wireless access. Founding companies and promoting members do not collect royalties for the intellectual property they contribute to support VOFDM technology. In addition, BWIF is organized as a program of the *IEEE Indus try Standards and Technology Organization* (IEEE-ISTO), which acts as the managing body for this forum.
- *TD-SCDMA:* Time-division synchronous CDMA (TD-SCDMA) is the Chi nese contribution to the ITU's IMT-2000 specification for 3G wireless mobile services. It endeavors to integrate with the existing GSM system. It is designed to manage both symmetric circuit-switched services, such as speech or video, as well as asymmetric packet-switched services, such as mobile Internet data flows. TD-SCDMA combines two leading technologies—an advanced TDMA system with an adaptive CDMA component—to overcome this chal lenge. Further details are beyond the scope of this book. The interested reader is advised to refer to [4].
- *HSDPA: High-speed downlink packet access* (HSDPA) is a packet-based data service in WCDMA downlink with data transmission up to 8–10 mbit/s [and 20 mbit/s for multiple-input multiple-output (MIMO) systems] over a 5-MHz bandwidth in WCDMA downlink. HSDPA implementations include adaptive modulation and coding (AMC), MIMO, hybrid automatic request (HARQ), fast cell search, and advanced receiver design. In *third generation partnership project* (3GPP) standards, Release 4 specifications provide effi cient IP support, enabling provision of services through an all-IP core net work, and Release 5 specifications focus on HSDPA to provide data rates up to approximately 10 mbit/s to support packet-based multimedia services. MIMO systems are the work item in Release 6 specifications, which will support even higher data

transmission rates up to 20 mbit/s. HSDPA is evolved from and backward compatible with Release 99 WCDMA systems. The interested reader is referred to [5].

It can be seen that there are many approaches toward 4G. The final standard has not been defined as yet, but it is expected to be finalized by 2007. The common theme in all these proposals, as can be seen from Figure 1.2, is increasing spectrum efficiency using MIMO techniques. These techniques are the subject of this book.

1.3 Multiple-Input Multiple-Output Systems

This technique is mainly based on the theoretical work developed by Teletar [6] and Foschini [7]. The core of this idea is to use multiple antennas both for transmis sion and reception. This increases the capacity of the wireless channel. Capacity is expressed as the maximum achievable data rate for an arbitrarily low probability of error. Hence, the thrust has been toward the development of codes and schemes that would enable systems to approach their Shannon capacity limit [8]. This technology received a fillip when Tarokh et al. [9] introduced their space-time

trellis coding techniques and Alamouti introduced his space-time block coding techniques to improve link-level performance based on diversity [10]. It received another boost when Bell Laboratories introduced its Bell Laboratories Layered Space-Time (BLAST) coding technique [11], demonstrating spectral efficiencies as high as 42 bit/s/Hz. This represents a tremendous boost in spectral efficiency compared with the current 2–3 bit/s/Hz achieved in cellular mobile and wireless LAN systems. There is, therefore, a need for communication engineers to understand this remarkable technology. This book has been expressly written to fulfill such a need.

We will now discuss a MIMO system pioneered by AT&T Labs-Research in Middletown, New Jersey, [12]. It conducted field tests to characterize the mobile MIMO radio channel. The company measured the capacity of a system with four antennas on a laptop computer and four antennas on a rooftop base station. The field tests showed that close to the theoretical fourfold increase in capacity over a single antenna system can be supported in a 30-KHz channel with dual polarized spatially separated base station and mobile terminal antennas. Figure 1.3 shows the arrangement. Note the mounting of the four antennas on the laptop computer and the rooftop antennas.

Figure 1.3 Upper: Transmitter with four antennas on a laptop and the 1,900-MHz coherent trans mitters. Lower: The four receivers with real-time baseband processing and rooftop antennas. (*From:* [12]. Reprinted with the permission of AT&T Labs-Research and Pentek Inc.)

The base station rooftop antenna array used dual-polarized antennas separated by 11.3 feet, which is approximately 20 wavelengths apart and a multibeam antenna. The laptop-mounted antennas included a vertically polarized array and a dual-polarized array with elements spaced half a wavelength apart. Different signals were transmitted out of each antenna simultaneously in the same bandwidth and then separated at the receiver. With four antennas at the transmitter and receiver, this has the potential to provide four times the data rate of a single antenna system without an increase in transmit power or bandwidth, provided the multipath environment is rich enough. This means that high capacities are theoretically possi ble unless there is a direct line of sight between transmitter and receiver. These and other concepts will be examined in this book.

We discussed earlier the aspect of 4G being IP-based. We now examine briefly as to what this entails, as wireless IP will be a principal beneficiary of MIMO.

1.4 Internet Protocol

The Internet is a collection of thousands of networks interconnected by gateways. It was developed under the U.S. Department of Defense Advanced Research Projects Agency (DARPA) support as an effort to connect the myriad of local area networks at U.S. universities to the ARPANET. The protocol developed to interconnect the various networks was called the IP.

The individual networks comprising the Internet are joined together by gate ways. A gateway appears as an external site to an adjacent network, but appears as a node to the Internet as a whole. In terms of layering, the Internet sublayer is viewed as sitting on top of the various network layers. The Internet sublayer packets are called datagrams and each datagram can pass through a path of gateways and individual networks on the way to its final destination.

1.4.1 Routing Operations

The IP gateway makes routing decisions based on the routing list. If the destination host resides in another network, the IP gateway must decide how to route to the other network. If multiple hops are involved, then each gateway must be traversed and the gateway must make decisions about the routing.

Each gateway maintains a routing table that contains the next gateway on the way to the final destination network. In effect, the table contains an entry for each reachable network. These tables could be static or dynamic. The IP module makes a routing decision on all the datagrams it receives.

1.4.2 The Transmission Control Protocol

IP is not designed to recover from certain problems nor does it guarantee the delivery of traffic. IP is designed to discard datagrams that are outdated or have exceeded the number of permissible transit hops in an Internet. Certain user applica tions require assurance that all datagrams have been delivered safely to the destina tion. Furthermore, the transmitting user may need to know that the traffic has

been delivered at the receiving host. The mechanisms to achieve these important services reside in the Transmission Control Protocol (TCP).

TCP resides in the transport layer of the layered model. It is situated above IP and below the upper layers. Figure 1.4 illustrates that TCP is not loaded into a gateway. It is designed to reside in the host computer or in a machine that is tasked with end-to-end integrity of the transfer of user data.

Figure 1.4 (a) TCP structure. (b) Wireless network IP that is packet-based. (*From:* [13]. Reprinted with the permission of Peter Rysavy, Rysavy Research.)

The figure also shows that TCP is designed to run over the IP. Since IP is a connectionless network, the tasks of reliability, flow control, sequencing, opens, and closes are given to TCP. TCP and IP are tied together so closely that they are used in the same context. Many of the TCP functions (such as flow control, reliability, sequencing and so on) could be handled within an application program. But it makes little sense to code these functions into each application. The preferred approach is to develop generalized software that provides community functions applicable to a wide range of applications and then invoke these programs from the application software. This allows the application programmer to concentrate on solving the application problem and it relieves

the programmer from the details and problems of networks.

TCP provides the following principal services to the upper layers:

- *Connection-oriented data management:* This term refers to the fact that TCP maintains status and state information about each user data stream flowing into and out of the TCP module. Used in this context, the term means TCP is responsible for the end-to-end transfer of data across multiple networks (i.e., the three networks in Figure 1.2) to a receiving user application.
- *Reliable data transfer:* To ensure reliability, TCP uses checksum routines and timers to ensure that the lapse of time is not excessive before remedial measures are taken for either transmission of acknowledgements from the receiving site and/or retransmission of data at the transmitting site.
- *Stream-oriented data transfer:* Stream-oriented transmission is designed to send individual characters and not blocks, frames, datagrams and so on. The bytes are sent on a stream basis, byte by byte. When they arrive at the TCP layer, the bytes are grouped into TCP segments. These segments are then passed to the IP.
- *Flow control:* The receiver's TCP module is able to flow control the sender's data to prevent buffer overruns. This is based on a transmission "window" that allows a specific number of bytes.
- *Full-duplex transmission:* TCP provides full-duplex transmission between two TCP entities. This permits simultaneous two-way transmission without having to wait for a turnaround signal.
- *Multiplexing:* TCP also has a very useful facility for multiplexing multiple user sessions within a single host computer. This is accomplished through a simple naming convention for ports and sockets between TCP and IP modules.
- *Precedence and security:* TCP also provides the user with the capability to specify levels of security and precedence (priority level) for the connection.
- *Graceful close:* TCP provides a graceful close to the logical connection between two users. A graceful close ensures that all traffic has been acknowl edged before the virtual circuit is removed.

The preceding paragraphs were a brief overview of IP. Until now, IP has been implemented as the foundation of the Internet and virtually all multivendor private Internet works. The currently deployed version of IP is actually IP version 4.

Previous versions of IP (1 through 3) were successively defined and replaced to reach IPv4. This protocol is reaching the end of its useful life and a new protocol known as IPv6 (IP version 6) has been defined to ultimately replace IP. The driving motivation for the adoption of a new version of IP was the limitation imposed by the 32-bit address field in IPv4. In addition, IP is a very old protocol and new requirements in the areas of security, routing flexibility, and traffic support have developed. To meet these needs, IPv6 has been defined [14, 15] and includes functional and formatting enhancements over IPv4.

1.5 Wireless Internet Protocol

Having briefly examined IP, we now extend this concept to wireless IP, which is illustrated in Figure 1.4(b).

In an ideal world, a computer connected over a wireless network would work just like a computer on a LAN. But wireless networks currently operate at lower speeds with higher latency and connections can be lost at any moment, especially when mobile. Experiments conducted in the field showed that the basic bottleneck was low data rates, which made downloading of pages inordinately long, as much as 15 seconds for screen updates. Therefore, a need exists to develop a modulation scheme well-suited to this requirement. The chosen scheme should not only support high bit rates, but should have a very fast synchronization scheme suited to packet transmission in fading channels. Toward this end, multiple antenna systems were developed with a view to increasing spectral efficiencies.

In Chapter 2, we examine the capacities of MIMO systems both in deterministic as well as random channels. The material has been drawn substantially from the outstanding pioneering research work conducted over the past six years by Dr. Paulraj and his team at Stanford University and discussed in their books and publications [16]. I have retained in this book the symbols and expressions they used in their papers and publications. I wish to add that I did this after a lot of deliberation, as over the years, these symbols and expressions have *de facto* passed into the *lingua franca* of MIMO technology. Hence, since this purports to be a book for the beginner as well, it is best to familiarize the reader with these symbols and expressions early on in the learning stage itself, so as to enable the reader to follow MIMO technical papers in the future for further understanding of the subject.

This chapter also examines the channel capacities under conditions when the channel is narrowband and is known to the transmitter and when it is unknown to the transmitter. It then examines frequency selective channels for the same two cases. We also discuss the ergodic and outage capacities of random channels.

Chapter 2 examines the MIMO wireless channel and determines its capacity under different conditions.

Chapter 3 provides an overview of the fading channel models and channel propagation.

Chapter 4 introduces the reader to space-time block coding techniques. We also examine the performance of space-time block codes in the presence of imperfect

channel estimation and antenna correlation. We conclude with dominant eigenmode transmission.

Chapter 5 introduces the reader to space-time trellis codes and its design and performance evaluation in fast and slow fading channels.

Chapter 6 discusses layered space-time codes and BLAST algorithms. Chapter 7 introduces the reader to orthogonal frequency division multiplexing (OFDM) techniques, which have become an integral part of MIMO technology. We also examine synchronization and channel estimation of OFDM signals. Chapter 8 introduces the reader to the IEEE 802.11a standard, which, as we

discussed earlier, has a lot of potential as a packet transmission system for MIMO systems.

Chapter 9 discusses the behavior of the space-time algorithms in a broadband wireless channel. We also discuss the capacity of spatially multiplexed signals and their performance. In addition, the reader is also introduced to a new topic called ‘‘hybrid OFDM/CDMA.’’ This is a new hardware-oriented technique intended primarily for field engineers and is totally unlike existing OFDM/CDMA tech niques. The idea here is to extend the capabilities of the current CDMA systems (used in cell phones) using the high throughput capability imparted by OFDM. The CDMA-OFDM system is then coupled with MIMO to yield a powerful solution to the ever-increasing requirement of high throughput cell phones.

Chapter 10 discusses topics for further reading in the MIMO field.

References

[1] van Nee, R., and R. Prasad, *OFDM for Wireless Multimedia Communications,* Norwood, MA: Artech House, 2000.

[2] Donald Shaver, Texas Instruments.

[3] Broadband Wireless Internet Forum, www.bwif.org.

[4] TD-SCDMA Forum, www.td-scdma-forum.org.

[5] UMTS Web site, www.umtsworld.com/technology/hsdpa.htm.

[6] Telatar, E., ‘‘Capacity of Multiantenna Gaussian Channels,’’ *European Transactions on Telecommunications,* Vol. 10, No. 6, November/December 1999, pp. 585–595. [7] Foschini, G. J., and M. J. Gans, ‘‘On Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas,’’ *Wireless Personal Communications,* Vol. 6, 1998, pp. 311–335.

[8] Shannon, C. E., ‘‘A Mathematical Theory of Communication,’’ *Bell Syst. Tech. J.,* Vol. 27, October 1948, pp. 379–423 (Part One), pp. 623–656 (Part Two), Reprinted in Book Form, Urbana, IL: University of Illinois Press, 1949.

[9] Tarokh, V., N. Seshadri, and A. R. Calderbank, ‘‘Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction,’’ *IEEE Trans. Inform. Theory,* Vol. 44, No. 2, March 1998, pp. 744–765.

[10] Alamouti, S. M., ‘‘A Simple Transmit Diversity Technique for Wireless Communications,’’ *IEEE Journal Select. Areas Commun.,* Vol. 16, No. 8, October 1998, pp. 1451–1458. [11] Foschini, G. J., ‘‘Layered Space-Time Architecture for Wireless Communications in a Fading Environment When Using Multiple Antennas,’’ *Bell Labs. Tech. J.,* 1996, Vol. 6, No. 2, pp. 41–59.

[12] ‘‘Smart Antenna Experiments for 3G and 4G Cellular Systems,’’ *The Pentex Pipeline,* Summer 2001, Vol. 10, No. 2, AT&T Labs-Research and Pentek Inc., Upper Saddle River, NJ.

[13] Rysavy, Peter, ‘‘Wireless IP: A Case Study,’’ www.pcsdata.com.

[14] Stallings, W., *Data and Computer Communications,* 5th edition, Upper Saddle River, NJ: Prentice Hall, 1996.

[15] Huitema, C., *IPv6: The New Internet Protocol,* Upper Saddle River, NJ: Prentice Hall, 1996.

[16] Paulraj, Arogyaswami, Rohit Nabar, and Dhananjay Gore, *Introduction to Space-Time Wireless Communications,* Cambridge, UK: Cambridge University Press, 2003.

2.1 Introduction

It is best to begin at the beginning. We will start by examining a few terms that are part of the MIMO antenna systems. These terms will be used throughout this book and, in fact, will become part of the MIMO vocabulary. We shall first discuss the MIMO system model. This will be followed by a study of frequency-flat MIMO channel capacities for cases when the channel is known to the transmitter and when it is unknown to the transmitter. This study will be carried out both for deterministic as well as for random channels. We shall then investigate the effects of the physical parameters of the channel, such as fading correlation, line-of sight problems and cross-polarization discrimination (XPD) problems of MIMO antennas and their impact on channel capacity. This will be followed by a study of the phenomenon of keyhole effect or degenerate channels and the effect of such channels on MIMO capacity. We conclude with a similar study of MIMO capacity on frequency-selective channels.

2.2 Preliminaries

2.2.1 Multiantenna Systems

Figure 2.1 illustrates different antenna configurations used in defining space-time systems. Single-input single-output (SISO) is the well-known wireless configuration, single-input multiple-output (SIMO) uses a single transmitting antenna and multiple (M_R) receive antennas, multiple-input single-output (MISO) has multiple (M_T) transmitting antennas and one receive antenna, MIMO has multiple (M_T) transmit ting antennas and multiple (M_R) receive antennas and, finally, MIMO-multiuser (MIMO-MU), which refers to a configuration that comprises a base station with multiple transmit/receive antennas interacting with multiple users, each with one or more antennas. We now examine the meaning of certain terms.

2.2.2 Array Gain

Array gain is the average increase in the signal-to-noise ratio (SNR) at the receiver that arises from the coherent combining effect of multiple antennas at the receiver or transmitter or both. If the channel is known to the multiple antenna transmitter, the transmitter will weight the transmission with weights, depending on the channel

Figure 2.1 Different antenna configurations in space-time systems.

coefficients, so that there is coherent combining at the single antenna receiver (MISO case). The array gain in this case is called transmitter array gain. Alternately, if we have only one antenna at the transmitter and no knowledge of the channel and a multiple antenna receiver, which has perfect knowledge of the channel, then the receiver can suitably weight the incoming signals so that they coherently add up at the output (combining), thereby enhancing the signal. This is the SIMO case. This is called receiver array gain. Basically, multiple antenna systems require perfect channel knowledge either at the transmitter or receiver or both to achieve this array gain.

2.2.3 Diversity Gain

Multipath fading is a significant problem in communications. In a fading channel, signals experience fades (i.e., they fluctuate in their strength). When the signal power drops significantly, the channel is said to be in a fade. This gives rise to high bit error rates (BER). We resort to diversity to combat fading. This involves providing replicas of the transmitted signal over time, frequency, or space. There are three types of diversity schemes in wireless communications.

- *Temporal diversity:* In this case replicas of the transmitted signal are provided across time by a combination of channel coding and time interleaving strate gies. The key requirement here for this form of diversity to be effective is that the channel must provide sufficient variations in time. It is applicable in cases where the coherence time of the channel is small compared with the desired interleaving symbol duration. In such an event, we are assured that the interleaved symbol is independent of the previous symbol. This makes it a completely new replica of the original symbol.
- *Frequency diversity:* This type of diversity provides replicas of the original signal in the frequency domain. This is applicable in cases where the

coherence bandwidth of the channel is small compared with the

bandwidth of the signal. This assures us that different parts of the relevant spectrum will suffer independent fades.

- *Spatial diversity:* This is also called antenna diversity and is an effective method for combating multipath fading. In this case, replicas of the same transmitted signal are provided across different antennas of the receiver. This is applicable in cases where the antenna spacing is larger than the coherent distance to ensure independent fades across different antennas. The traditional types of diversity schemes are [1] *selection diversity, maximal ratio diversity,* and *equal gain diversity.* These schemes will be investigated in Chapter 3. Space-time codes exploit diversity across space and time. These will be examined in Chapters 4, 5, and 6.

Basically the effectiveness of any diversity scheme lies in the fact that at the receiver we must provide *independent* samples of the basic signal that was trans mitted. In such an event we are assured that the probability of two or more relevant parts of the signal undergoing deep fades will be very small. The constraints on coherence time, coherence bandwidth, and coherence distance ensure this. The diversity scheme must then optimally combine the received diversified waveforms so as to maximize the resulting signal quality. We can also categorize diversity under the subheading of spatial diversity, based on whether diversity is applied to the transmitter or to the receiver.

- *Receive diversity:* Maximum ratio combining is a frequently applied diversity scheme in receivers to improve signal quality. In cell phones it becomes costly and cumbersome to deploy. This is one of the main reasons transmit diversity became popular, since transmit diversity is easier to implement at the base station.
- *Transmit diversity:* In this case we introduce controlled redundancies at the transmitter, which can be then exploited by appropriate signal processing techniques at the receiver. Generally this technique requires complete channel information at the transmitter to make this possible. But with the advent of space-time coding schemes like Alamouti's scheme [2], discussed in Chapter 4, it became possible to implement transmit diversity *without* knowledge of the channel. This was one of the fundamental reasons why the MIMO industry began to rise. Space-time codes for MIMO exploit both transmit as well as receive diversity schemes, yielding a high quality of reception.

Therefore, in MIMO we talk a lot about receive antenna diversity or transmit antenna diversity. In receive antenna diversity, the receiver that has multiple anten nas receives multiple replicas of the same transmitted signal, assuming that the transmission came from the same source. This holds true for SIMO channels. If the signal path between each antenna pair fades independently, then when one path is in a fade, it is extremely unlikely that all the other paths are also in deep fade. Therefore, the loss of signal power due to fade in one path is countered by the same signal but received through a different path (route). This is like a line of soldiers. When one soldier falls in battle, another is ready to take his place. Hence,

extending this analogy further, the more the soldiers, the stronger the line. The same is the argument in diversity. The more the diversity, the easier we can

combat fades in a channel. Diversity is characterized by the number of independent fading branches, or paths (routes). These paths are also known as diversity order and are equal to the number of receive antennas in SIMO channels. Logically, the higher the diversity order (independent fading paths, or receive antennas), the better we combat fading. If the number of receive antennas tends to infinity, the diversity order tends to infinity and the channel tends to additive white Gaussian noise (AWGN). This is illustrated in Figure 2.2 [3]. In the figure, the sharp drops in power are called ‘‘fading margins.’’ Note that with rising diversity order, the fading margins come down in intensity. This has been measured over a time period of 900 samples.

In the category of spatial diversity there are two more types of diversity that we need to consider. These are:

- *Polarization diversity:* In this type of diversity horizontal and vertical polar ization signals are transmitted by two different polarized antennas and received correspondingly by two different polarized antennas at the receiver. Different polarizations ensure that there is no correlation between the data streams, without having to worry about coherent distance of separation between the antennas.
- *Angle diversity:* This applies at carrier frequencies in excess of 10 GHz. At such frequencies, the transmitted signals are highly scattered in space. In such an event the receiver can have two highly directional antennas facing in totally different directions. This enables the receiver to collect two samples of the same signal, which are totally independent of each other.

2.2.4 Data Pipes

The term *data pipe* is derived from fluid mechanics. Pipes are used to transfer water to a tank/reservoir. The more the number of pipes, the greater the quantum

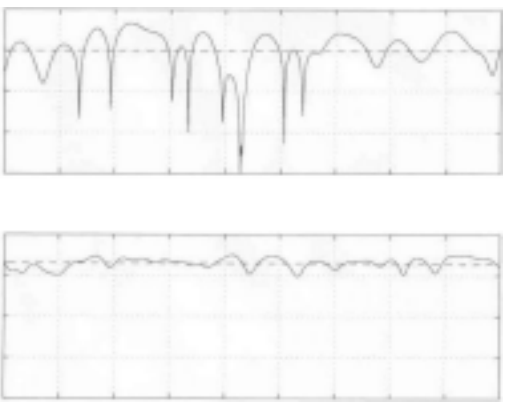

Figure 2.2 Rayleigh fading margins for (a) 1-input 1-output and (b) 2-input 2-output system. (*From:* [3]. Reprinted with permission. CRC Press, Boca Raton, Florida.)

of flow of water into a tank/reservoir. This is similar to data pipes, but the analogy of communications with fluid mechanics ends there. We consider a case of two data pipes between the transmitter and receiver. In this situation there are two cases; either the data in the data pipes are identical to each other or they are independent samples, completely different from each other. In the

former case, effectively the data going through is as if it is going through one data pipe, with the other pipe merely being a replica of the first one. This is a case of full correlation and because of this correlation, we do not get any throughput (bits per second) advantage. However, we do get a diversity advantage of two. The latter case deals with a situation where there is absolutely no correlation between the data carried by the two pipes. The data streams are independent. Hence, there is no diversity, but the throughput (output in bit/s) is definitely higher than in the first case. Therefore, the more the data pipes, the higher the throughput, provided the signals in the data pipes are not replicas of each other or correlated. In such an event the same signal is going through both pipes, so no new information is getting trans ferred. Therefore, correlation is not a good thing and it does reduce capacity, as we shall see. Remember that transmit diversity comes at the cost of throughput and vice versa. If we wish to eat the cake and still have it, then one way out is to sacrifice transmit diversity at the cost of throughput and incorporate diversity in the receiver (receive diversity). This way we at least have receive diversity, rather than no diversity at all in the system. This is what is done in spatial multiplexing.

2.2.5 Spatial Multiplexing

Spatial multiplexing offers a linear (in the number of transmit-receive antenna pairs or min (M_R , M_T) increase in the transmission rate (or capacity) for the same bandwidth and with no additional power expenditure. It is only possible in MIMO channels [4, 5]. Consider the case of two transmit and two receive antennas. This can be extended to more general MIMO channels.

The bit stream is split into two half-rate bit streams, modulated and transmitted simultaneously from both the antennas. The receiver, having complete knowledge of the channel, recovers these individual bit streams and combines them so as to recover the original bit stream. Since the receiver has knowledge of the channel it provides receive diversity, but the system has no transmit diversity since the bit streams are completely different from each other in that they carry totally different data. Thus spatial multiplexing increases the transmission rates proportionally with the number of transmit-receive antenna pairs.

This concept can be extended to MIMO-MU. In such a case, two users transmit their respective information simultaneously to the base station equipped with two antennas. The base station can separate the two signals and can likewise transmit two signals with spatial filtering so that each user can decode his or her own signal correctly. This allows capacity to increase proportionally to the number of antennas at the base station and the number of users.

2.2.6 Additional Terms

Automatic request for repeat (ARQ). This is an error control mechanism in which received packets that cannot be corrected are retransmitted. This is a type of temporal diversity.

Forward error correction (FEC). This is a technique that inserts redundant bits during transmission to help detect and correct bit errors during reception.

Coding gain. The improvement in SNR at the receiver because of FEC is called

coding gain.

Interleaving. A form of data scrambling that spreads bursts of bit errors evenly over the received data allowing efficient forward error correction.

Multiplexing gain. Capacity gain at no additional power or bandwidth consump tion obtained through the use of multiple antennas at both sides of a wireless link.

2.3 MIMO System Model

We consider a MIMO system with a transmit array of M_T antennas and a receive array of M_R antennas. The block diagram of such a system is shown in Figure 2.3. The transmitted matrix is a $M_T \cdot 1$ column matrix **s** where s_i is the ith component, transmitted from antenna i. We consider the channel to be a Gaussian channel such that the elements of **s** are considered to be independent identically distributed (i.i.d.) Gaussian variables. If the channel is unknown at the transmitter, we assume that the signals transmitted from each antenna have equal powers of E_s / M_T. The covariance matrix for this transmitted signal is given by

$$R_{ss} = E_{\underline{s}}$$

$$M_T I_{M^T} \quad (2.1)$$

where E_s is the power across the transmitter irrespective of the number of antennas M_T and I_{M^T} is an $M_T \cdot M_T$ identity matrix. The transmitted signal bandwidth is so narrow that its frequency response can be considered flat (i.e., the channel is

Figure 2.3 Block diagram of a MIMO system.

memoryless). The channel matrix **H** is a $M_R \cdot M_T$ complex matrix. The component $h_{i\ j}$ of the matrix is the fading coefficient from the jth transmit

antenna to the *i*th receive antenna. We assume that the received power for each of the receive antennas is equal to the total transmitted power E_s. This implies we ignore signal attenuation, antenna gains, and so on. Thus we obtain the normalization constraint for the elements of **H**, for a deterministic channel as

$$\sum_{j=1}^{M_T} |h_{ij}|^2 = M_T,\ i = 1, 2, \ldots, M_R \quad (2.2)$$

If the channel elements are not deterministic but random, the normalization will apply to the expected value of (2.2).

We assume that the channel matrix is known at the receiver but unknown at the transmitter. The channel matrix can be estimated at the receiver by transmitting a training sequence. If we require the transmitter to know this channel, then we need to communicate this information to the transmitter via a feedback channel. The elements of **H** can be deterministic or random.

The noise at the receiver is another column matrix of size $M_R \cdot 1$, denoted by **n**. The components of **n** are zero mean circularly symmetrical complex Gaussian (ZMCSCG) variables. The covariance matrix of the receiver noise is

$$\mathbf{R_{nn}} = E\left[\mathbf{n}\mathbf{n}^H\right] \quad (2.3)$$

If there is no correlation between components of **n**, the covariance matrix is obtained as

$$\mathbf{R_{nn}} = N_0 \mathbf{I}_{MR} \quad (2.4)$$

Each of the M_R receive branches has identical noise power of N_0. The receiver operates on the maximum likelihood detection principle over M_R receive antennas. The received signals constitute a $M_R \cdot 1$ column matrix denoted by **r**, where each complex component refers to a receive antenna. Since we assumed that the total received power per antenna is equal to the total transmitted power, the SNR can be written as

$$\gamma = \frac{E_s}{N_0} \quad (2.5)$$

Therefore, the received vector can be expressed as

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n} \quad (2.6)$$

The received signal covariance matrix defined as $E\left[\mathbf{r}\mathbf{r}^H\right]$, is given by [using (2.6)]

$$\mathbf{R_{rr}} = \mathbf{H}\mathbf{R_{ss}}\mathbf{H}^H \quad (2.7)$$

while the total signal power can be expressed as tr($\mathbf{R_{rr}}$).

2.4 MIMO System Capacity

The system capacity is defined as the maximum possible transmission rate such that the probability of error is arbitrarily small [6].

We assume that the channel knowledge is unavailable at the transmitter and known only at the receiver.

The capacity of MIMO channel is defined as [4, 5]

$$C = \max_{f(\mathbf{s})} I(\mathbf{s}; \mathbf{y}) \quad (2.8)$$

where $f(\mathbf{s})$ is the probability distribution of the vector $\mathbf{s}$ and $I(\mathbf{s}; \mathbf{y})$ is the mutual information between vectors $\mathbf{s}$ and $\mathbf{y}$. We note that

$$I(\mathbf{s}; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{s}) \quad (2.9)$$

where $H(\mathbf{y})$ is the differential entropy of the vector $\mathbf{y}$, while $H(\mathbf{y} \mid \mathbf{s})$ is the conditional differential entropy of the vector $\mathbf{y}$, given knowledge of the vector $\mathbf{s}$. Since the vectors $\mathbf{s}$ and $\mathbf{n}$ are independent, $H(\mathbf{y} \mid \mathbf{s}) = H(\mathbf{n})$. From (2.9),

$$I(\mathbf{s}; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{n}) \quad (2.10)$$

If we maximize the mutual information $I(\mathbf{s}; \mathbf{y})$ reduces to maximizing $H(\mathbf{y})$.

The covariance matrix of $\mathbf{y}$, $R_{\mathbf{yy}} = \mathrm{e}\mathbf{H}\mathbf{y}\mathbf{y}^H\mathbf{J}$, satisfies

$$\mathbf{R}_{\mathbf{yy}} = E_{\underline{s}} M_T \mathbf{H}\mathbf{R}_{\mathbf{ss}}\mathbf{H}^H + N_0 \mathbf{I}_{M_R} \quad (2.11)$$

where $\mathbf{R}_{\mathbf{ss}} = \mathrm{e}\mathbf{H}\mathbf{s}\mathbf{s}^H\mathbf{J}$ is the covariance matrix of $\mathbf{s}$. Among all vectors $\mathbf{y}$ with a given covariance matrix $\mathbf{R}_{\mathbf{yy}}$, the differential entropy $H(\mathbf{y})$ is maximized when $\mathbf{y}$ is ZMCSCG. This implies that $\mathbf{s}$ must also be ZMCSCG vector, the distribution of which is completely characterized by $\mathbf{R}_{\mathbf{ss}}$. The differential entropies of the vectors $\mathbf{y}$ and $\mathbf{n}$ are given by

$$H(\mathbf{y}) = \log_2 (\det(\mathrm{p}\mathrm{e}\mathbf{R}_{\mathbf{yy}})) \text{ bps/Hz} \quad (2.12)$$

$$H(\mathbf{n}) = \log_2 \mathrm{X}\det\mathrm{X}\mathrm{p}\mathrm{e}s^2 \mathbf{I}_{M_R}\mathrm{CC} \text{ bps/Hz} \quad (2.13)$$

Therefore, $I(\mathbf{s}; \mathbf{y})$ in (2.10) reduces to [4]

$$M_T N_0 \mathbf{H}\mathbf{R}_{\mathbf{ss}}\mathbf{H}^H\mathrm{D} \text{ bps/Hz} \quad (2.14)$$

$$I(\mathbf{s}; \mathbf{y}) = \log_2 \det\mathrm{S}\mathbf{I}_{M_R} + E_{\underline{s}}$$

and from (2.8), the capacity of the MIMO channel is given by

$$M_T N_0 \mathbf{H}\mathbf{R}_{\mathbf{ss}}\mathbf{H}^H \Big| \text{ bps/Hz} \quad (2.15)$$

$$\text{Tr}(\mathbf{R}_{\mathbf{ss}}) = M_T$$

$$C = \max \log_2 \det\left(\mathbf{I}_{M_R} + \frac{E_s}{N_0}\mathbf{H}\mathbf{R}_{\mathbf{ss}}\mathbf{H}^H\right) \text{ bps/Hz}$$

The capacity C in (2.15) is also called error-free spectral efficiency or data rate per unit bandwidth that can be sustained reliably over the MIMO link. Thus if our bandwidth is W Hz, the maximum achievable data rate over this bandwidth using MIMO techniques is WC bit/s [7].

2.5 Channel Unknown to the Transmitter

If the channel is unknown to the transmitter, then the vector $\mathbf{s}$ is statistically independent (i.e., $\mathbf{R}_{\mathbf{ss}} = \mathbf{I}_{M_T}$). This implies that the signals are independent and the power is equally divided among the transmit antennas. The capacity in such a case is, (from 2.15)

$$M_T N_0 \mathbf{H}\mathbf{H}^H \Big| \quad (2.16)$$

$$C = \log_2 \det\left(\mathbf{I}_{M_R} + \frac{E_s}{N_0}\mathbf{H}\mathbf{H}^H\right)$$

The reader is cautioned that this is not Shannon capacity since it is possible to outperform $\mathbf{R}_{\mathbf{ss}} = \mathbf{I}_{M_T}$, if one has the channel knowledge. Nevertheless we shall refer to (2.16) as capacity. Now $\mathbf{H}\mathbf{H}^H$ is an $M_R \cdot M_R$ positive semidefinite Hermitian matrix. The eigendecomposition of such a matrix is given by $\mathbf{Q}\Lambda\mathbf{Q}^H$ [8], where $\mathbf{Q}$ is an $M_R \cdot M_R$ matrix satisfying $\mathbf{Q}^H\mathbf{Q} = \mathbf{Q}\mathbf{Q}^H = \mathbf{I}_{M_R}$ and $\Lambda = diag\{\lambda_1\ \lambda_2 \ldots \lambda_{M_R}\}$ with $\lambda_i \geq 0$. We assume that the eigenvalues are ordered so that $\lambda_i \geq \lambda_{i+1}$. Then

$$\lambda_i = \begin{cases} s_i^2, \text{ if } i = 1, 2, \ldots r \\ 0 \text{ if } i = r, r+1, \ldots M_R \end{cases} \quad (2.17)$$

where s_i are the singular values obtained as $S = diag\{s_1\ s_2 \ldots s_r\}$ from the singular value decomposition of $\mathbf{H} = \mathbf{U}S\mathbf{V}^H$. Then the capacity of the MIMO channel is given by

$$M_T N_0 \mathbf{Q}\Lambda\mathbf{Q}^H \Big| \quad (2.18)$$

$$C = \log_2 \det\left(\mathbf{I}_{M_R} + \frac{E_s}{N_0}\mathbf{H}\mathbf{H}^H\right)$$

Using $\det(\mathbf{I}_m + \mathbf{A}\mathbf{B}) = \det(\mathbf{I}_n + \mathbf{B}\mathbf{A})$ for matrices $\mathbf{A}(m \cdot n)$ and $\mathbf{B}(n \cdot m)$ and $\mathbf{Q}^H\mathbf{Q} = \mathbf{I}_{M_R}$, (2.18) simplifies to

$$M_T N_0 \left. \right] D \quad (2.19)$$

$$C = \log_2 \det \left(\mathbf{I}_{M_R} + \frac{E_s}{} \right)$$

or

$$C = \sum_{i=1}^{r} \log_2 \left(1 + \frac{E_s}{M_T N_0} \lambda_i \right) \quad (2.20)$$

where r is the rank of the channel and λ_i ($i = 1, 2, \ldots, r$) are the positive eigenvalues of $\mathbf{H}\mathbf{H}^H$. Equation (2.20) expresses the capacity of the MIMO channel as a sum of the capacities of r SISO channels, each having a power gain of λ_i ($i = 1, 2, \ldots, r$) and transmit power E_s / M_T.

This means that the technique of multiple antennas at the transmitter and receiver opens up multiple scalar spatial data pipes between the transmitter and receiver. Furthermore, equal transmit energy is allocated to each spatial data pipe. This is for the case when the channel is unknown at the transmitter.

We define the squared Frobenius norm of $\mathbf{H}$, as $\|\mathbf{H}\|_F^2 = \mathrm{Tr}(\mathbf{H}\mathbf{H}^H) = \sum_{i=1}^{M_R} \sum_{j=1}^{M_T} |h_{i,j}|^2$. Frobenius norm is interpreted as the total power gain of the channel.

Also $\|\mathbf{H}\|_F^2 = \sum_{i=1}^{M_R} \lambda_i$ where λ_i ($i = 1, 2, \ldots, M_R$) are the eigenvalues of $\mathbf{H}\mathbf{H}^H$. We fix this total power so that $\|\mathbf{H}\|_F^2 = b$. Then if the channel matrix is of full rank such that $M_T = M_R = M$, the capacity C in (2.20) is maximized when $\lambda_i = \lambda_j = b/M$ ($i, j = 1, 2, \ldots, M$) (remember, the channel is unknown, so equal power distribution). To achieve this, $\mathbf{H}\mathbf{H}^H = \mathbf{H}^H \mathbf{H} = (b/M)\mathbf{I}_M$, (i.e., the channel matrix $\mathbf{H}$ should be orthogonal). This gives

$$N_0 M^2 \left. \right] D \quad (2.21)$$

$$C = M \log_2 \left(1 + \frac{b E_s}{} \right)$$

If the elements of $\mathbf{H}$ have ones along the diagonal, then $\|\mathbf{H}\|_F^2 = M^2$ and N_0

$$\left. \right] D \quad (2.22)$$

$$C = M \log_2 \left(1 + \frac{E_s}{} \right)$$

The capacity of an orthogonal MIMO channel is therefore M times the scalar

channel capacity [7]. This conclusion is once again verified using a different approach for OFDM-based spatial multiplexing, as discussed in Chapter 9.

2.6 Channel Known to the Transmitter

It is possible by various means, which will be discussed in Chapter 4, to learn the channel state information (CSI) at the transmitter. In such an event the capacity can be increased by resorting to the so-called ‘‘water filling principle’’ [4], by assigning various levels of transmitted power to various transmitting antennas. This power is assigned on the basis that the better the channel gets, the more power it gets and vice versa. This is an optimal energy allocation algorithm.

2.6.1 Water-Pouring Principle

Consider a MIMO channel where the channel parameters are known at the trans mitter. The ‘‘water-pouring principle’’ or ‘‘water-filling principle’’ can be derived

by maximizing the MIMO channel capacity under the rule that more power is allocated to the channel that is in good condition and less or none at all to the bad channels.

This analysis is taken from [7]. Consider a ZMCSCG signal vector $\mathbf{s}\tilde{}$ of dimen sion $r \cdot 1$ where r is the rank of the channel $\mathbf{H}$ to be transmitted. We note from Figure 2.4 that the vector is multiplied by a matrix $\mathbf{V}$ prior to transmission (based on the fact that $\mathbf{H} = \mathbf{U}\mathrm{S}\mathbf{V}^H$ through singular value decomposition). At the receiver, the received signal vector $\mathbf{y}$ is multiplied by the matrix $\mathbf{U}^H$.

The input-output relationship for this operation is given by

$$\mathbf{y}\tilde{} = \sqrt{\frac{E_s}{M_T}}\mathbf{U}^H\mathbf{H}\mathbf{V}\mathbf{s}\tilde{} + \mathbf{U}^H\mathbf{n} \quad (2.23)$$

$$= \sqrt{\frac{E_s}{M_T}}\mathrm{S}\mathbf{s}\tilde{} + \mathbf{n}\tilde{}$$

where $\mathbf{y}\tilde{}$ is the transformed received signal vector of size $r \cdot 1$ and $\mathbf{n}\tilde{}$ is the ZMCSCG transformed noise vector of size $r \cdot 1$ with the covariance matrix $\mathrm{e}\{\mathbf{n}\tilde{}\mathbf{n}\tilde{}^H\} = N_0\mathbf{I}_r$. The vector $\mathbf{s}\tilde{}$ satisfies $\mathrm{e}\{\mathbf{s}\tilde{}\mathbf{s}\tilde{}^H\} = M_T$ to constrain the total transmit energy. Equation (2.23) shows us that with channel knowledge at the transmitter, $\mathbf{H}$ can be explicitly decomposed into r parallel SISO channels satisfying

$$y\tilde{}_i = \sqrt{\frac{E_s}{M_T}}\sqrt{\lambda_i}s\tilde{}_i + n\tilde{}_i,\ i = 1, 2, \ldots, r \quad (2.24)$$

The capacity of the MIMO channel is the sum of the individual parallel SISO channel capacities and is given by

$$C = \sum_{i=1}^{r} \log_2 \left(1 + \frac{E_s}{M_T N_0} \gamma_i \lambda_i\right) \quad (2.25)$$

where $\gamma_i = E\{|s_i|^2\}$ ($i = 1, 2, \ldots, r$) is the transmit energy in the ith subchannel such that $\sum_{i=1}^{r} \gamma_i = M_T$.

To maximize mutual information, the transmitter can access the individual subchannels and allocate variable power levels to them. Hence, the mutual informa tion maximization problem becomes,

Figure 2.4 Decomposition of **H** when the channel is known to the transmitter and receiver.

$$C = \max_{\sum_{i=1}^{r} \gamma_i = M_T} \sum_{i=1}^{r} \log_2 \left(1 + \frac{E_s}{M_T N_0} \gamma_i \lambda_i\right) \quad (2.26)$$

Using Lagrangian methods, the optimal energy allocation procedure is $E_s \lambda_i$

$$= \left(\mu - \frac{M_T N_0}{\lambda_i}\right)_+, \; i = 1, 2, \ldots, r \text{ and} \quad (2.27)$$

$$\gamma_i^{opt}$$

$$\sum_{i=1}^{r} \gamma_i^{opt} = M_T$$

where μ is a constant and $(x)_+$ implies

$$(x)_+ = \begin{cases} x \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases} \quad (2.28)$$

We determine this optimal energy allocation iteratively through the ‘‘water pouring algorithm’’ [9]. We now describe the algorithm.

We set the iteration count p to 1 and calculate the constant m in (2.27):

$$\mathrm{m} = \frac{M_T}{E_s}\left[\sum_{i=1}^{r-p+1} \lambda_i^{-1} + \frac{N_0}{(r-p+1)}\right] \qquad (2.29)$$

Using this value of m, the power allocated to the ith subchannel is calculated as

$$g_i = \left[\mathrm{m} - \frac{M_T N_0}{E_s}\lambda_i^{-1}\right]^{+}, \quad i = 1, 2, \ldots, r-p+1 \qquad (2.30)$$

If the power allotted to the channel with the lowest gain is negative (i.e., $\lambda_{r-p+1} < 0$), we discard this channel by setting $g_{r-p+1}^{opt} = 0$ and rerun the algorithm with the iteration count p incremented by 2. The optimal power allocation strategy, therefore, allocates power to those spatial subchannels that are non-negative. Figure 2.5 illustrates the water-pouring algorithm. Obviously, since this algorithm only concentrates on good-quality channels and rejects the bad ones during each channel realization, it is to be expected that this method yields a capacity that is equal or better than the situation when the channel is unknown to the transmitter.

Channel capacity for the case when the channel is unknown to the transmitter and receiver is an area of ongoing research [10, 11].

2.6.2 Capacity When Channel Is Known to the Transmitter

This has been already discussed and is given by (2.26)

$$C = \max_{\sum_{i=1}^{r} g_i = M_T N_0} \sum_{i=1}^{r} \log_2\left(1 + \frac{E_s}{N_0} g_i \lambda_i\right) \qquad (2.31)$$

Figure 2.5 Schematic of the water-pouring algorithm.

2.7 Deterministic Channels

2.7.1 SIMO Channel Capacity

In a SIMO channel, $M_T = 1$ and there are M_R receive antennas. In such a case the channel matrix is a column matrix

$$H = \begin{bmatrix} h_1 & h_2 & \ldots & h_{M_R} \end{bmatrix}^T \quad (2.32)$$

where $(?)^T$ denotes matrix transpose. Since $M_R > M_T$, (2.16) is modified as M_T

$$N_0 \mathbf{H}^H \mathbf{H} \quad (2.33)$$

$$C = \log_2 \det \left(\mathbf{I}_{M_T} + \frac{E_s}{} \right)$$

Now $\mathbf{H}^H \mathbf{H} = \sum_{i=1}^{M_R} |h_i|^2$ and $M_T = 1$. Hence,

$$C = \log_2 \det \left(1 + \sum_{i=1}^{M_R} \right)$$

$$N_0 \quad (2.34) \quad |h_i|^2 \frac{E_s}{}$$

If the channel matrix elements are equal and normalized as

$$|h_1|^2 = |h_2|^2 = \ldots |h_{M_R}|^2 = 1$$

then capacity when the channel is unknown at the transmitter, is N_0 (2.35)

$$C = \log_2 \det\left(I + M_R E_s\right)$$

The system achieves a diversity gain of M_R relative to the SISO case. For $M_R = 4$ and SNR = 10 dB, the SIMO capacity is 5.258 bit/s/Hz. The addition

of receive antennas yields a logarithmic increase in capacity in SIMO channels. Knowledge of the channel at the transmitter in this case provides no additional benefit.

2.7.2 MISO Channel Capacity

In MISO channels, $M_R = 1$ and there are M_T transmit antennas. In this case, since $M_T > M_R$, we use (2.16) as it is. The channel is represented by the row matrix

$$H = \left[h_1\ h_2 \ldots h_{M_T}\right]$$

As $\mathbf{H}\mathbf{H}^H = \left(\sum_{j=1}^{M_T} |h_j|^2\right)$, from (2.16) we obtain

$$C = \log_2\left(1 + \sum_{j=1}^{M_T} \frac{|h_j|^2 E_s}{M_T N_0}\right) \quad (2.36)$$

If the channel coefficients are equal and normalized as $\left(\sum_{j=1}^{M_T} |h_j|^2 = M_T\right)$, then the capacity for the MISO case becomes

$$C = \log_2\left(1 + \frac{E_s}{N_0}\right) \quad (2.37)$$

We note that (2.37) is the same as for a SISO case (i.e., the capacity did not increase with the number of antennas). This is for the case when the channel is unknown at the transmitter. The reason for this result is that there is no array gain at the transmitter because the transmitter has no knowledge of the channel parameters. Array gain is the average increase in the SNR at the receiver that arises from the coherent combining effect of multiple antennas at the receiver or transmitter or both. If the channel is known to the transmitter, the transmitter will weight the transmission with weights depending on the channel coefficients, so that there is coherent combining at the receiver (MISO case). If we take the case when the channel is known at the transmitter, we apply (2.31). Since the channel matrix has rank 1, there is only one term in the sum in (2.31)

and only one nonzero eigenvalue given by

$$\lambda = \sum_{j=1}^{M_T} |h_j|^2$$

Hence, capacity is

$$C = \log_2\left(1 + \sum_{j=1}^{M_T} |h_j|^2 \frac{E_s}{N_0}\right) \quad (2.38)$$

If the channel coefficients are equal and normalized as $\sum_{j=1}^{M_T} |h_j|^2 = M_T$, the capacity becomes

$$C = \log_2\left(1 + M_T \frac{E_s}{N_0}\right) \quad (2.39)$$

For $M_T = 4$ and SNR = 10 dB, the MISO capacity is 5.258 bit/s/Hz. *This is with channel knowledge at the transmitter.* In both cases of SIMO and MISO there is only one spatial data pipe (i.e., the rank of the channel matrix is one). Basically, the channel matrix is a $M_R \cdot M_T$ matrix. In a MISO case, $M_R = 1$ and in a SIMO case, $M_T = 1$. In either case, the channel matrix has only one eigenvalue and its rank is 1. Physically, this means that there is only one route from transmitter to receiver for the signals to pass through. Hence, we have one data pipe. If we had $M_T = M_R = 2$, then we would have a MIMO case with a channel matrix of rank 2 and having two eigenvalues, hence, two routes from transmitter to receiver (i.e., we have two data pipes and so on).

2.8 Random Channels

We have until now discussed MIMO capacity when the channel is a deterministic channel. We now consider the case when **H** is chosen randomly according to a Rayleigh distribution in a quasi-static channel. This is a real-life situation encoun tered, for example, in wireless LANs with high data rates and low fade rates. We assume that the receiver has perfect knowledge of the channel and the transmitter has no knowledge of the channel. Since the channel is random, the information rate associated with it is also random. The cumulative distribution function (CDF) of the information rate of a flat fading MIMO channel is shown in Figure 2.6 for a $2 \cdot 2$ system. The SNR is 10 dB and the channel is unknown to the transmitter.

Figure 2.6 CDF of information rate for i.i.d. channel matrix with a 2 · 2 system and SNR = 10 dB.

2.8.1 Ergodic Capacity

The ergodic capacity of a MIMO channel is the ensemble average of the information rate over the distribution of the elements of the channel matrix **H** [7]. It is the capacity of the channel when every channel matrix **H** is an independent realization [i.e., it has no relationship to the previous matrix but is typically representative of it class (ergodic)]. This implies that it is a result of infinitely long measurements. Since the process model is ergodic, this implies that the coding is performed over an infinitely long interval. Hence, it is the Shannon capacity of the channel. Based on (2.20) the ergodic capacity is expressed as

$$C = \mathrm{e}\,5\,\sum_{i=1}^{r} \log_2 \mathrm{S}_{1+\mathrm{r}\,M_T\,\mathrm{l}}\,i\,\mathrm{D6} \quad (2.40)$$

where $\mathrm{r} = E_s / N_0$. The expectation operator applies in this case because the channel is random. Since **H** is random, the information rate associated with it is also random. The CDF of the information rate is depicted in Figure 2.6.

The ergodic capacity is the median of the CDF curve. In this case it is 7.0467 bit/s/Hz. Figure 2.7 shows the ergodic capacity over different system configurations as a function of r. We note that ergodic capacity increases with increasing r and with increasing M_T and M_R .

Ergodic capacity when the channel is known to the transmitter is based on the water-filling algorithm and is given from (2.25)

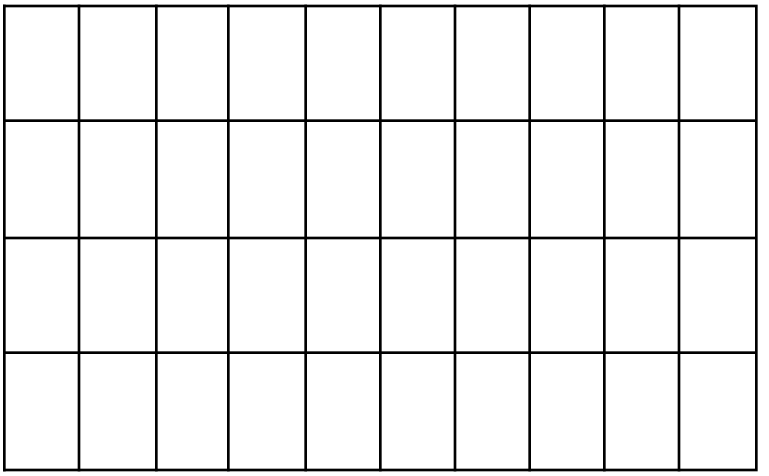

Figure 2.7 Ergodic capacity for different antenna configurations with $M_T = M_R = M$.

$$C = \mathrm{e}\,5\ \sum_{i=1}^{r} \log^2 \mathrm{S}\,1 + \frac{E_s}{M_T N_0}\,\mathrm{g}_i \,\mathrm{l}_i\,\mathrm{D6} \quad (2.41)$$

Equation (2.41) is the ensemble average of the capacity achieved when the water-filling optimization is performed for each realization of **H**. Figure 2.8 shows the performance comparison of ergodic capacity of a MIMO channel with $M_T = M_R = 4$ when the channel is unknown to the transmitter and also when known and the channel is Rayleigh i.i.d.

The ergodic capacity when the channel is known to the transmitter is always higher then when it is unknown. This advantage reduces at high SNRs. This is because at high SNRs (2.41) tends to (2.40). Another way of looking at this situation is to appreciate the fact that at high SNRs, all eigenchannels perform equally well (i.e., there is no difference in quality between them). Hence, all the channels will perform to their capacities, making both cases nearly identical.

2.8.2 Outage Capacity

In reality, the block lengths are finite. The common example is speech transmission. In such cases, we speak of outage capacity. Outage capacity is the capacity that is guaranteed with a certain level of reliability. We define p% outage capacity as the information rate that is guaranteed for (100-p)% of the channel realizations, that is, $P(C \leq C_{\text{out}}) = p\%$ [12]. We show 10% outage capacity in Figure 2.9.

Figure 2.9 shows the 10% outage capacity for several MIMO cases, when the channel is i.i.d. and unknown at the transmitter. We note that as the SNR increases, the capacity increases and as the number of antennas increases, so does the capacity. From (2.16) for the case when $M_T = M_R = M$ and the channel is i.i.d.,

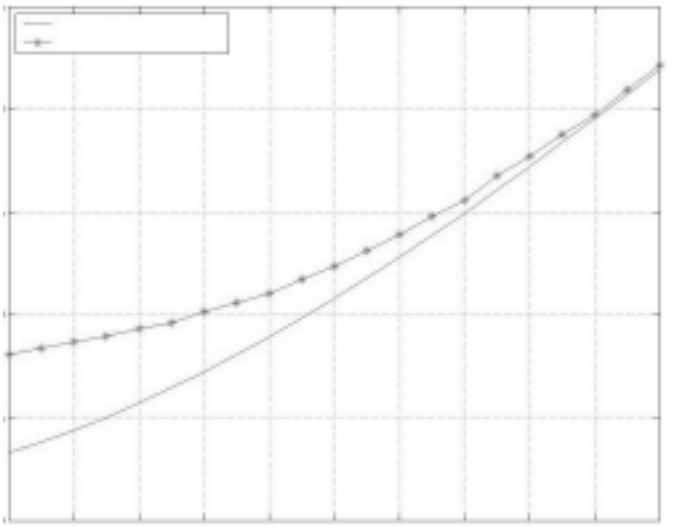

Figure 2.8 Ergodic capacity of an $M = 4$ channel with and without channel knowledge at the transmitter. The difference in ergodic capacities decreases with SNR.

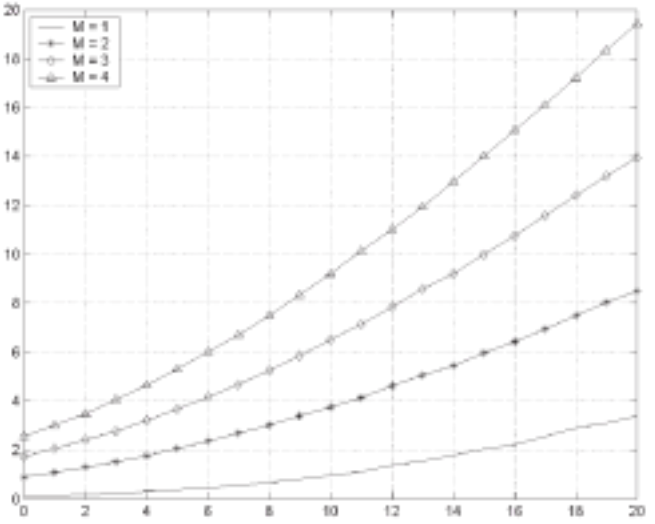

Figure 2.9 10% outage capacity for various antenna configurations. Outage capacity improves with rising $M_T = M_R = M$.

$$\frac{1}{M}\mathbf{H}_v\mathbf{H}^H \qquad _v \rightarrow \mathbf{I}_M \text{ as } M \rightarrow \infty$$

Therefore,

$$C \rightarrow M \log_2 (1 + r) \text{ where r is the SNR} \quad (2.42)$$

Asymptotically in *M*, the capacity in spatially white MIMO channel becomes deterministic and increases linearly with *M* for a fixed SNR. Also for every 3-dB increase in SNR, we get *M* bit/s/Hz increase in capacity for a MIMO channel, compared with 1 bit/s/Hz in a SISO channel. The outage capacity curves substantiate this conclusion. If the channel is known at the transmitter, Figure 2.10 shows that water-filling is a superior solution.

The same arguments for convergence of the curves at high SNRs apply to Figure 2.10 as for Figure 2.8 but in the context of outage capacities.

2.9 Influence of Fading Correlation on MIMO Capacity

In reality, the channel is not ideally Rayleigh i.i.d. There are various factors that cause it to deviate from this and, as a result, the performance of MIMO systems deteriorate. One of these is correlation. Correlation problems arise because of the separation distance between *antenna elements* in a base station. Usually this separation distance is in the order of a few centimeters, whereas the separation between the mobile and the base station is in the order of a few kilometers! Hence,

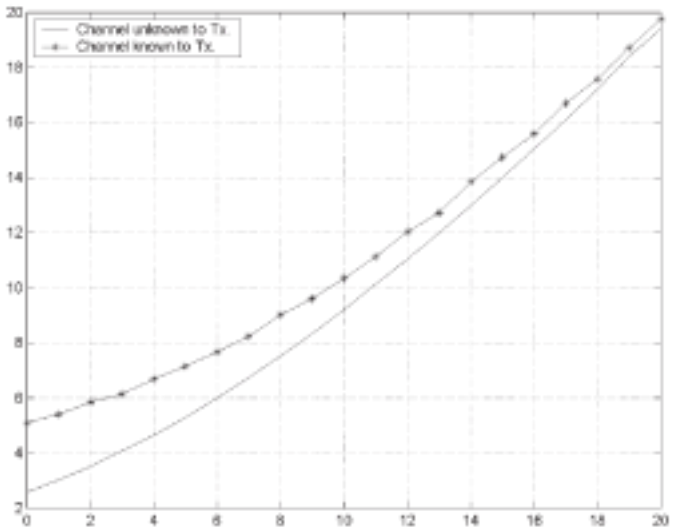

Figure 2.10 10% outage capacity of an $M = 4$ channel with and without channel knowledge at the transmitter. The difference in outage capacities decreases with SNR.

the signals arriving at the base station from a receiver will necessarily be very close together, giving rise to correlation between them. This occurs because all the antenna elements receive the same signal, due to the geometry of the phenomenon. The degree of ‘‘sameness’’ determines the correlation coefficient with 1 as maximum correlation and 0 as no correlation. This is overcome in a base station by:

- Using independent dipole antennas separated by a distance *D* that

exceeds the coherent distance for that channel.

- Using two separate antenna arrays separated by a distance D that exceeds the coherent distance for that channel.

These cases are illustrated in Figure 2.11.

In Figure 2.11, both cases (a) and (b) are feasible. The separation distance D is usually of the order of 10 to 16 wavelengths for a base station, because it is on a high vantage point and far from the mobile receivers. The problem is not so severe for mobile phones because they are invariably located in a high scattering environment. In such cases the separation distance is usually 2 to 3 wavelengths. In the event of correlation, the elements of the channel matrix are correlated and may be modeled as [7]

$$\text{vec}(\mathbf{H}) = \mathbf{R}^{1/2}\,\text{vec}(\mathbf{H}_v) \quad (2.43)$$

where $\mathbf{H}_v$ is a Rayleigh i.i.d. spatially white MIMO channel matrix of size $M_R \cdot M_T$ and $\mathbf{R}$ is a $M_T M_R \cdot M_T M_R$ covariance matrix defined as

Figure 2.11 The correlation problem. We can deploy (a) separate dipole antennas well separated from each other or (b) antenna arrays well separated from each other.

$$\mathbf{R} = e\left\{vec\,(\mathbf{H})\;vec\,(\mathbf{H})^H\right\} \quad (2.44)$$

$\mathbf{R}$ is a positive semidefinite Hermitian matrix. If $\mathbf{R}$ is full rank (i.e., $\mathbf{R} = \mathbf{I}_{M_T M_R}$), then in such a case $\mathbf{H} = \mathbf{H}_v$. The idea of such a model is to efficiently portray the correlation effects in the channel. This approach is elaborated by using a more generalized model given by

$$\mathbf{H} = \mathbf{R}_r^{1/2}\,\mathbf{H}_v\mathbf{R}_t^{1/2} \quad (2.45)$$

where $\mathbf{R}_t$ is the $M_T \cdot M_T$ transmit covariance matrix and $\mathbf{R}_r$ is the $M_R \cdot M_R$ receive

covariance matrix. Both $\mathbf{R}_t$ and $\mathbf{R}_r$ are positive semidefinite Hermitian matrixes. Equation (2.45) is explained as follows:

- The transmitted signal, when it reaches the receiver, is correlated by virtue of the geometry at the receiver ($\mathbf{R}_r$). The channel *per se* has been portrayed as Rayleigh i.i.d. ($\mathbf{H}_v$).
- The transmitted signal is correlated at the transmitter itself due to the geome try at the transmitter ($\mathbf{R}_t$) or due to a low angle of spread.
- $\mathbf{R}$, $\mathbf{R}_t$ and $\mathbf{R}_r$ are related by $\mathbf{R} = \mathbf{R}^T_t \otimes \mathbf{R}_r$ where $\otimes$ denotes Kronecker product.

We note that $\mathbf{H}_v$ is full rank *per se,* but the effective rank of $\mathbf{H}$ gets reduced due to correlation at the transmitter or at the receiver or both and this effective rank is expressed as min ($r(\mathbf{R}_r)$, $r(\mathbf{R}_t)$) where $r(\mathbf{A})$ denotes rank of A. If we assume that both the matrixes $\mathbf{R}_r$ and $\mathbf{R}_t$ are normalized so that they have unity values along their diagonals, this yields $e\mathbf{H}|h_{i,j}|^2\mathbf{J} = 1$. The capacity of the MIMO channel in the presence of spatial fading correlation without channel knowledge at the transmitter follows from (2.16) as

$$C = \log_2 \det\mathbf{S}\mathbf{I}_{M_R} + \frac{M_T}{M}\mathbf{R}^{1/2}_r\mathbf{H}_v\mathbf{R}^{H/2}_t\mathbf{D}\mathbf{R}^{1/2}_r\mathbf{H}_v\mathbf{R}_t\mathbf{H}^H\mathbf{D} \quad (2.46)$$

Assume $M_R = M_T = M$ and that the receive and transmit correlation matrixes are full rank. Then, at high SNR, the capacity can be approximated as

$$C \approx \log_2 \det\mathbf{S}\frac{1}{M}\mathbf{H}_v\mathbf{H}^H\mathbf{D} + \log_2 \det(\mathbf{R}_r) + \log_2 \det(\mathbf{R}_t) \quad (2.47)$$

We note from (2.47) that both correlation matrixes have the same impact on the channel capacity. We now examine the conditions on $\mathbf{R}_r$ that maximize capacity. The same arguments apply to $\mathbf{R}_t$.

$$\det(\mathbf{R}_r) = \mathbf{P}_{i=1}^{M} \lambda_i(\mathbf{R}_r) \le 1 \quad (2.48)$$

Remember that there is a power constraint in that $\mathbf{S}_{i=1}^{M} \lambda_i(\mathbf{R}_r) = M$. This means that $\log_2 \det(\mathbf{R}_r) \le 0$. It can only equal zero if all eigenvalues of $\mathbf{R}_r$ are equal (i.e., $\mathbf{R}_r = \mathbf{I}_M$). Therefore, fading signal correlation does reduce the

number of eigenvalues and thereby reduces the MIMO channel capacity. This loss in ergodic or outage capacity is given by ($\log_2 \det(\mathbf{R}_r) + \log_2 \det(\mathbf{R}_t)$) bit/s/Hz.

If we assume an orthogonal channel where $M_T = M_R = 2$ and further assume that there is correlation only at the receiver, then we choose a receive correlation matrix as

$$\mathbf{R}_r = \begin{bmatrix} 1 & s \\ s & 1 \end{bmatrix} \quad (2.49)$$

We take a correlation coefficient of 0.8.

We note from Figure 2.12 that there is a loss of 2.47 bit/s/Hz at high SNR compared with the case with no correlation. This is the loss expected from the $\log_2 \det(\mathbf{R}_r)$ component. If the correlation coefficient of either or both of $\mathbf{R}_r$ and $\mathbf{R}_t$ is unity, then the $\mathbf{H}$ matrix will also become rank 1 (i.e., it becomes an SISO channel). Hence, correlation is not a good thing!

2.10 Influence of LOS on MIMO Capacity

We now examine another aspect, which makes a channel deviate from Rayleigh i.i.d. Until now we have only considered a Rayleigh i.i.d. channel. This is far removed from reality. It is better to depict the real-world channel as

$$\mathbf{H} = \mathbf{H}_{Ric} + \mathbf{R}_r^{1/2} \mathbf{H}_v \mathbf{R}_t^{1/2} \quad (2.50)$$

where $\mathbf{H}_{Ric}$ is the Rician or line-of-sight (LOS) component. The other terms were discussed earlier. The LOS is a component that exists by virtue of a direct path

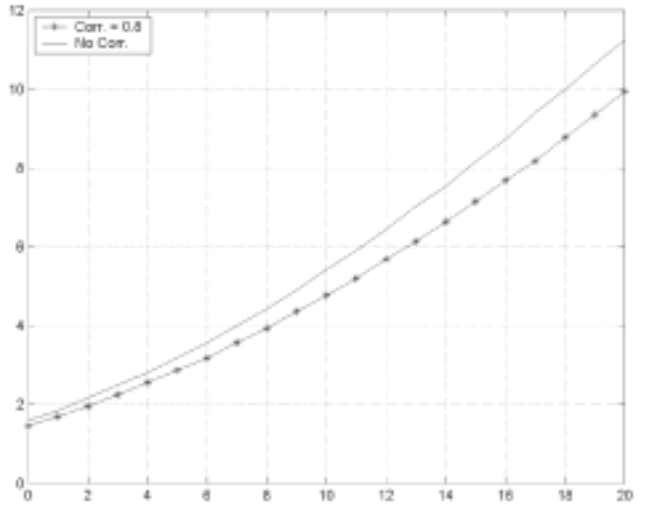

Figure 2.12 Ergodic capacity with high and low correlation. The loss in ergodic capacity is about 2.47 bit/s/Hz when s = 0.8.

between the transmitter and the receiver, which are so located as to be within line of sight of each other. The LOS in (2.50) can also be shown as a sum of a fixed component and a scattered component as follows [13]

$$\mathbf{H} = \sqrt{\frac{K}{1+K}}\mathbf{H} + \sqrt{\frac{1}{1+K}}\mathbf{H}_v \quad (2.51)$$

where $\sqrt{K/(1+K)}\mathbf{H} = e(\mathbf{H})$ is the LOS component of the channel and $\sqrt{1/(1+K)}\mathbf{H}_v$ is the fading component that assumes uncorrelated fading. The elements of **H** are assumed to have unit power. K in (2.51) is the Rician K-factor of the system and is essentially the ratio of the power in the LOS component of the channel to the power in the fading component. $K = 0$ corresponds to a pure Rayleigh i.i.d. channel, whereas $K = \infty$ corresponds to a nonfading channel. The LOS component manifests itself in the following two cases:

- The separation distance between antennas as previously discussed.
- LOS component created due to a poor scattering environment. This is shown in Figure 2.13.

In Figure 2.13, we discuss two indoor wireless environment cases, like a WLAN environment. We have a laptop with two receiving antennas. In a poor scattering environment we are likely to encounter a situation as shown in the left half of the figure. Due to colocated antennas, we have a LOS component. If the scattering is rich enough, the antennas do not appear colocated, as shown in the right half of the figure. This sort of environment is close to Rayleigh i.i.d.

and is desirable. The

Figure 2.13 Colocation problem in a fixed WLAN environment.

former gives rise to a LOS component. Hence, we note that the LOS phenomenon can occur both in indoor as well as outdoor environments.

In either case, the end result is the same regarding correlation. We take a **H** matrix of

$$\mathbf{H} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \quad (2.52)$$

Equation (2.52) pertains to a correlation coefficient of 0.8, similar to the correlation effect in the example in Section 2.9. In Figure 2.14, we have plotted ergodic capacity using this channel matrix with varying *K*-factor.

We note from Figure 2.14 that rising *K*-factor is detrimental to capacity. Hence, we must be careful to minimize the LOS component. This is one of the major engineering hurdles in MIMO technology.

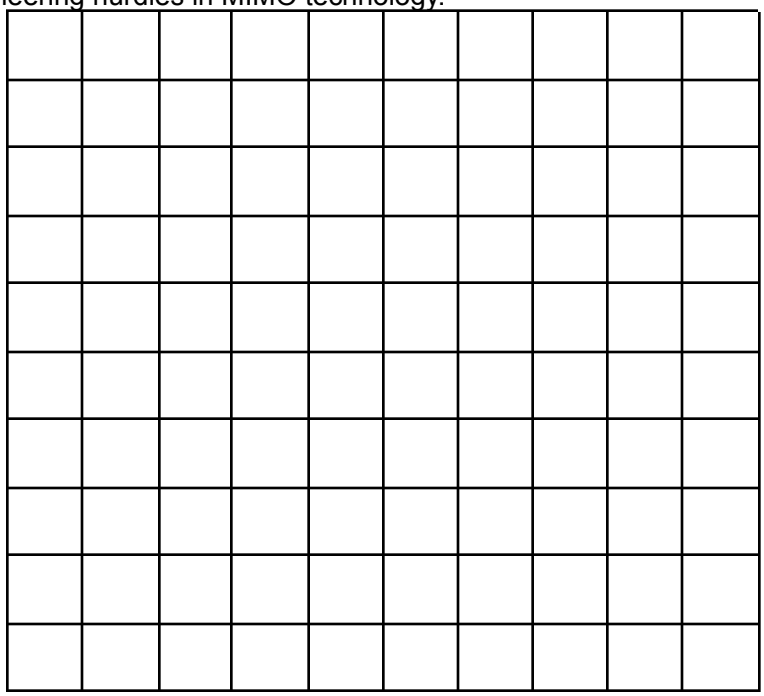

Figure 2.14 Ergodic capacity versus *K*-factor for a MIMO channel. Capacity declines with rising

2.11 Influence of XPD on MIMO Capacity

The channel models discussed so far assume that the antennas at the base station and at the receivers have identical polarizations. The use of antennas with orthogonal polarizations at the transmitter and receiver leads to a gain (or power) and correla tion imbalance between the elements of **H** [14]. These polarizations are usually ±45° or horizontal/vertical (0°/90°). Basically they require being orthogonal to each other. This ideally ensures zero coupling between the antennas. Therefore, signals with vertical polarizations, for example, are transmitted by one set of antennas and received by another set of vertical polarized antennas at the receiver. The same is the case with the horizontal polarized antennas. In view of the fact that these polarizations are orthogonal to each other, the signals do not ‘‘see’’ each other (i.e., they are independent). This is the ideal case. The reality is quite different. A certain amount of each signal ‘‘leaks’’ into the other signal and vice versa. We introduce the terms XPD and cross-polarization coupling (XPC). The former tells us as to how well one antenna discriminates its polarization from the other antenna. The latter term refers to the coupling between these polarizations during their propagation through the channel and is caused due to the rich scattering nature of the environment [15]. These phenomena are collectively defined by a constant a ($0 \le a \le 1$), where *in the absence of XPC,* 0 implies that we have good XPD (i.e., the antennas discriminate between each other’s polarizations extremely well (no interference) and 1 implies no XPD, meaning that the antennas cannot discrimi nate at all between each other’s signals). It was found that, typically, at distances of 2.6 Kms and above, a = 1, due to the rich scattering nature of the environment [16]. If we assume the power in the individual channel elements to be

$$E\{|h_{1,1}|^2\} = E\{|h_{2,2}|^2\} = 1 \quad (2.53)$$

$$E\{|h_{1,2}|^2\} = E\{|h_{2,1}|^2\} = a \quad (2.54)$$

Assuming a Rayleigh i.i.d. channel, the channel **H** with cross-polarized antennas may be modeled approximately as

$$\mathbf{H} = b\,(\,X\mathbf{R}^{1/2} \qquad {}_{t}C \quad (2.55)$$

where

${}_{r}\mathbf{H}_{v}\mathbf{R}^{1/2}$

$$b = \begin{bmatrix} 1 & \sqrt{a} \end{bmatrix} \qquad \sqrt{a\ 1}\,G \quad (2.56)$$

and (stands for the Hadamard product (if $\mathbf{A} = \mathbf{B}$ ($\mathbf{C}$ then $[\mathbf{A}]_{i,j} = [\mathbf{B}]_{i,j}[\mathbf{C}]_{i,j}$). The covariance matrixes $\mathbf{R}_r$ and $\mathbf{R}_t$ are already well known to us as portraying the correlations extant at the receiver and the transmitter, respectively, and also include XPD, XPC, and antenna spacing as factors influencing their structure.

The XPC phenomenon occurs in a scattering environment. If the environment through which

the signal propagates is nonscattering, then **H** = b (i.e., the right half of (2.55) vanishes).

If we assume the environment as nonscattering (i.e., deterministic), then the capacity for a 2 · 2 system is given by [from (2.21)]

$$C_{a=0} = 2\log^2 S_{1+r_2} D \quad (2.57)$$

and

$$C_{a=1} = \log_2 (1 + 2r) \quad (2.58)$$

where r is SNR. For (2.58), the **H** matrix is all ones, yielding eigenvalues of 0 and 4.

At very low SNR (r << 1), using $\log_2 (1 + x) \approx x \log_2 e$ for $x << 1$, $C_{a=0} \approx r$

$$\log_2 e \quad (2.59)$$

and

$$C_{a=1} \approx 2r \log_2 e \quad (2.60)$$

Hence, good XPD is detrimental to capacity at low SNR.

In high SNR conditions (r >> 1),

$$C_{a=0} \approx 2\log^2 S$$

$$_2D \quad (2.61)\ r$$

$$C_{a=1} \approx \log_2 (2r) \quad (2.62)$$

Hence, in high SNR conditions, good XPD (a = 0) performs better than poor XPD, which is exactly the reverse of the case at low SNRs. Figure 2.15 confirms this performance for a 2 · 2 channel.

2.12 Keyhole Effect: Degenerate Channels

We consider a system with two transmit and two receive antennas surrounded by scatterers. These antennas are uncorrelated. If the channel were a Rayleigh i.i.d. channel, this would have yielded a channel matrix of full rank and size as $M_R \cdot M_T$. Now suppose, as shown in Figure 2.16, a screen separates these two sets of antennas with a small hole in it. This gives rise to a propagation condition called ‘‘keyhole.’’ The only way for the transmitted signals to propagate is to pass through the keyhole. The transmitted signal vector is given

by

$$\mathbf{S} = [s_1\ s_2]^T \quad (2.63)$$

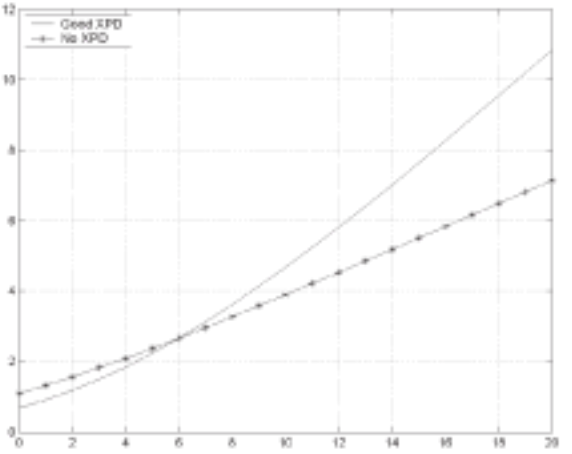

Figure 2.15 Ergodic capacity of a MIMO channel with good XPD (a = 0) and no XPD (a = 1).

Figure 2.16 The keyhole effect.

where s_1 and s_2 are the signals transmitted from the first and second antennas, respectively. The signal incident at the keyhole is given by

$$\mathbf{r}_1 = \mathbf{H}_t\,\mathbf{S} \quad (2.64)$$

where $\mathbf{H}_t$ is given by

$$\mathbf{H}_t = [h_1\ h_2] \quad (2.65)$$

where h_1 and h_2 are the channel coefficients corresponding to transmitted signals s_1 and s_2 respectively. h_1 and h_2 are independent complex Gaussian variables. The signal across the keyhole is given by

$$\mathbf{r}_2 = q\mathbf{r}_1 \quad (2.66)$$

where q is the keyhole attenuation.

The signal vector at the receive antenna across the keyhole, denoted by $\mathbf{r}_3$, is given by

$$\mathbf{r}_3 = \mathbf{H}_r \mathbf{r}_2 \quad (2.67)$$

where $\mathbf{H}_r$ is the channel matrix describing the propagation on the right-hand side of the keyhole and is given by

$$\mathbf{H}_r = [h_3 \; h_4]^T \quad (2.68)$$

where h_3 and h_4 are the channel coefficients corresponding to the first and second receive antennas, respectively.

The effective channel $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{H}_r \mathbf{H}^T_t \quad (2.69)$$

The received signal vector at the right-hand side of the keyhole is then given by

$$\mathbf{r}_3 = q\mathbf{HS} \quad (2.70)$$

The channel matrix from (2.69) is

$$\mathbf{H} = \begin{matrix} h_1 h_3 & h_2 h_3 \\ h_1 h_4 & h_2 h_4 \end{matrix} \quad (2.71)$$

Since $\mathbf{H}$ is constructed from the product of two vectors, every realization of the channel $\mathbf{H}$ is rank-deficient with a rank of 2. The distribution of $\mathbf{H}$ is double Rayleigh [17] and is given by

$$f(x) = \int_0^\infty \frac{x}{\sigma_r^4} e^{-\frac{x^2}{2\sigma_r^2}} \cdot \frac{2\sigma_r^2}{\sigma_r^4 + \sigma_r^4} d\sigma_r, \; x \geq 0 \quad (2.72)$$

where the amplitude distribution is the product of two independent Rayleigh distri butions, each with the power of $2\sigma_r^2$. There is only one channel or data-pipe between the transmitter and the receiver. The corresponding channel capacity is given by [from (2.20)]

$$C = \log_2 \begin{matrix} 1 & 0 \\ 0 & 1 + \rho_2 \end{matrix} \quad (2.73)$$

where r is SNR and l is the solitary eigenvalue. The capacity with increasing SNR increases logarithmically, although the underlying channel is a MIMO channel. Figure 2.17 shows the performance of a degenerate channel. The drop in capacity compared with a regular channel is evident.

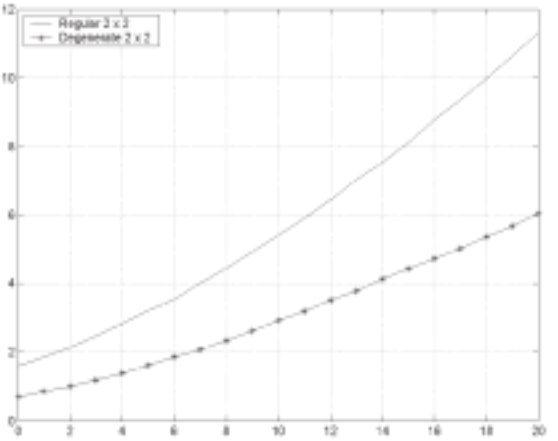

Figure 2.17 Performance of a degenerate channel for a 2 · 2 system.

Keyhole effects occur sometimes when the transmitted wavefront arrives with no angle spread (e.g., signals penetrating into buildings through small windows and also in narrow streets).

2.13 Capacity of Frequency Selective MIMO Channels

We now consider a real-life situation wherein the channel is not narrowband but frequency selective. Intuitively, subdividing the wideband channel into *N* nar rowband ones, and then summing the capacities of these *N* frequency flat channels can achieve this. The bandwidth of each of these subchannels will be *B*/*N* Hz where *B* is the overall channel bandwidth. This is provided the coherent bandwidth of the channel permits this (i.e., it is more than or equal to *B*/*N* Hz), as otherwise the subchannels will not be frequency flat.

We take the *i*th subchannel. The input-output relationship is defined as [from (2.6)],

$$\mathbf{r}_i = \mathbf{H}_i\mathbf{s}_i + \mathbf{n}_i \quad (2.74)$$

where $\mathbf{r}_i$ is the $M_R \cdot 1$ received signal vector, $\mathbf{s}_i$ is the $M_T \cdot 1$ transmitted signal vector and $\mathbf{n}_i$ is the $M_R \cdot 1$ noise vector for the *i*th subchannel.

Hence, for the overall wideband channel we deal with block matrixes as

$$R = HS + N \quad (2.75)$$

where $R = \mathbf{F}\mathbf{r}^T_1\ \mathbf{r}^T_2 \ldots \mathbf{r}^T_N\ \mathbf{G}^T$ is $M_R N \cdot 1$, $S = \mathbf{F}\mathbf{s}^T_1\ \mathbf{s}^T_2 \ldots \mathbf{s}^T_N\ \mathbf{G}^T$ is $M_T N \cdot 1$,

$N = \mathbf{F}\mathbf{n}^T{}_1\,\mathbf{n}^T{}_2\,\ldots\mathbf{n}^T{}_N\,\mathbf{G}^T$ is $M_R N \cdot 1$ and H is an $M_R N \cdot M_T N$ block diagonal matrix with $\mathbf{H}_i$ as block diagonal elements. $\mathbf{R}_{ss} = \mathrm{e}\mathbf{H}SS^H\mathbf{J}$ is the covariance matrix

of S, constrained so that $\mathrm{Tr}(\mathbf{R}_{ss}) = NM_T$. This constrains the total average transmit power to E_s. From (2.17), the capacity of such a channel is given by

$$M_T N_0 H\,\mathbf{R}_{ss} H^H\mathbf{D} \quad \text{bps/Hz} \tag{2.76}$$

$$\mathrm{Tr}(\mathbf{R}_{SS}) = NM_T \log_2 \det\mathbf{S}_{\mathbf{I}_{M_R N} + E_{\underline{s}}}$$

$$C_{FS} = {}^{B}_{N} \max$$

We now examine the two usual cases of when the channel is unknown to the transmitter and when it is known to the transmitter.

2.13.1 Channel Unknown to the Transmitter

In this case, we should choose $\mathbf{R}_{ss} = \mathbf{I}_{M_T N}$, which implies that the covariance matrix is of full rank (no correlation) and this in turn means that transmit power is allocated evenly across space (transmit antennas) and frequency (subchannels). This yields a deterministic capacity of [from (2.16)]

$$M_T N_0 \mathbf{H}_i \mathbf{H}^H{}_i \mathbf{D} \quad \text{bps/Hz} \tag{2.77}$$

$$C_{FS} \approx {}^{B}_{N} \sum^{N}{}_{i=1} + E_{\underline{s}}$$

$$\log_2 \det\mathbf{S}_{\mathbf{I}_{M_R}}$$

If the frequency response of the channel is flat (we are talking about the *entire* channel being narrowband), [i.e., $\mathbf{H}_i = \mathbf{H}$ (i = 1, 2, . . . , N)], then $M_T N_0 \mathbf{H}\mathbf{H}^H\mathbf{D}$

(2.78)

$$C_{FS} = \log_2 \det\mathbf{S}_{\mathbf{I}_{M_R} + E_{\underline{s}}}$$

which is the same as (2.16), the capacity of a frequency flat MIMO channel.

Further if all $\mathbf{H}_i$ are i.i.d. (i.e., the coherence bandwidth is B/N Hz), then as

$$N \to \infty,\ C_{FS} \to C^{\infty}{}_{FS} \tag{2.79}$$

(i.e., the capacity of such a frequency selective channel approaches a fixed quantity). If the channel is random, we then have the usual two cases of ergodic

and outage capacities. The ergodic capacity is given by

$$M_T N_0 \mathbf{H}_i \mathbf{H}^{H_i} \mathrm{D6} \quad \text{bps/Hz} \quad (2.80)$$

$$C_{FS} \approx \mathrm{e}\,5 B_N \quad + E_{\underline{s}}$$
$$\sum^{N}{}_{i=1}$$
$$\log_2 \det \mathrm{S}_{\mathbf{I}_{MR}}$$

The outage capacity is similarly defined. However, this outage capacity will be much better (higher) than for the earlier examined cases of frequency flat channels (at low outage rates). This is due to the high amount of frequency diversity present in the frequency selective channel. This is manifest in Figure 2.18.

In Figure 2.18, as the number of narrowband channels increases, with increasing frequency selectivity, the outage capacity also rises proportionately because of rising frequency diversity. Hence, the more the frequency selectivity, the higher the outage capacity. Note also the tendency of the curve to flatten with rising frequency

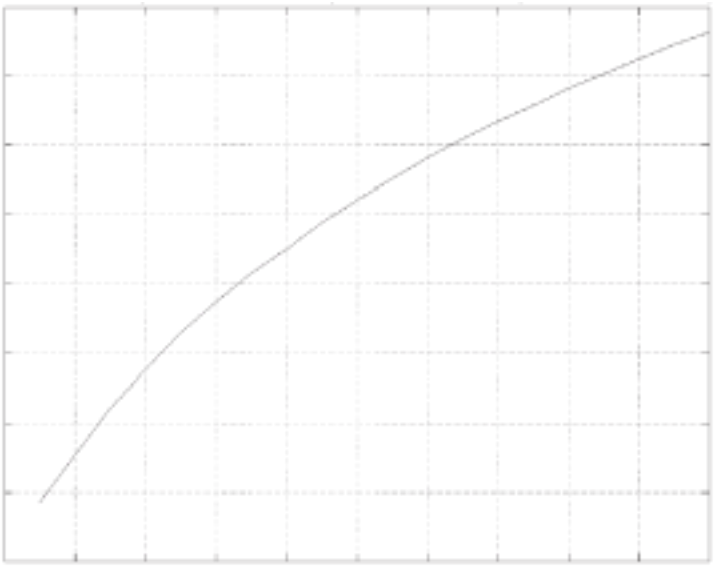

Figure 2.18 Performance of frequency selectivity versus 10% outage capacity.

selectivity and rising N. This bears out the statement in (2.79) that as $N \rightarrow \infty$, the capacity tends to a fixed value. This means that asymptotically (in N), the outage capacity of a sample realization of the frequency selective MIMO channel equals its ergodic capacity (because $N \rightarrow \infty$). The influence of multiple physical parameters such as delay spread, cluster angle spread and total angle spread on ergodic and outage capacity of frequency selective MIMO channels will be studied in Chapter 9.

2.13.2 Channel Known to the Transmitter

The treatment regarding this case is similar as was done earlier for frequency flat channels. In this case, we need to distribute the energy or power across space (antennas) and frequency (subchannels) so as to maximize spectral efficiency. This is called space-frequency water-filling [18]. Since water-filling is applicable only to purely orthogonal channels, it becomes necessary to achieve orthogonal channels by using OFDM techniques (see Chapter 7) to convert a frequency select channel into a set of parallel frequency flat channels, which are orthogonal to each other.

In such an event, if the composite channel H is known to the transmitter, the channel may be decomposed into $r(H)$ space-frequency modes, where $r(\mathbf{A})$ of matrix $\mathbf{A}$ stands for rank. The capacity is then given by [using (2.31)]

$$C_{FS} = \frac{B}{N} \max_{S^{r(H)}} \sum_{\substack{i=1 \\ r(H)}}^{\sum_{i=1} g_i = NM_T} \log_2 \left(1 + \frac{E_s}{M_T N_0} g_i \lambda_i(HH^H)\right), \quad \text{bps/Hz} \quad (2.81)$$

where $\lambda_i(HH^H)$ ($i = 1, 2, \ldots, r(H)$) are the positive eigenvalues of HH^H and g_i is the energy allocated to the ith space-frequency subchannel. We can define ergodic

and outage capacities of such channels, as was done earlier for frequency flat channels.

References

[1] Rappaport, T. S., *Wireless Communications: Principles and Practice,* Upper Saddle River, NJ: Prentice Hall, 1996.

[2] Alamouti, S. M., ‘‘A Simple Transmit Diversity Technique for Wireless Communications,’’ *IEEE Journal Select. Areas Commun.,* Vol. 16, No. 8, October 1998, pp. 1451–1458. [3] Bo¨ lcskei, H., and A. J. Paulraj, ‘‘Multiple-Input Multiple-Output (MIMO) Wireless Sys tems,’’ Chapter in *The Communications Handbook,* 2nd edition, J. Gibson (ed.), CRC Press, 2002, pp. 90.1–90.14.

[4] Telatar, E., ‘‘Capacity of Multi-Antenna Gaussian Channels,’’ *European Transactions on Telecommunications,* Vol. 10, No. 6, November/December 1999, pp. 585–595. [5] Foschini, G. J., and M. J. Gans, ‘‘On Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas,’’ *Wireless Personal Communications,* Vol. 6, 1998, pp. 311–335.

[6] Vucetic, Branka, and Jinhong Yuan, *Space-Time Coding,* Chichester, UK: John Wiley & Sons Ltd., 2003.

[7] Paulraj, Arogyaswami, Rohit Nabar, and Dhananjay Gore, *Introduction to Space-Time Wireless Communications,* Cambridge, UK: Cambridge University Press, 2003. [8] Golub, G., and C. Van Loan, *Matrix Computations,* Baltimore, MD: John Hopkins University Press, 2nd edition, 1989.

[9] Cover, T., and J. Thomas, *Elements of Information Theory,* New York: John Wiley & Sons, 1992.

[10] Marzetta, T., and B. Hochwald, ‘‘Capacity of a Mobile Multiple-Antenna Communication Link in Rayleigh Flat Fading,’’*IEEE Trans. Inf. Theory,* 45(1), January 1999, pp. 139–157. [11] Hassibi, B., and T. Marzetta, ‘‘Multiple Antennas and Isotropically Random Unitary Inputs: The Received Signal Density in Closed Form,’’ *IEEE Trans. Inf. Theory,* 48(6), June 2002,

pp. 1473–1484.

[12] Biglieri, E., J. Proakis, and S. Shamai, ‘‘Fading Channels: Information Theoretic and Communications Aspects,’’ *IEEE Trans. Inf. Theory,* 44(6), October 1998, pp. 2619–2692.

[13] Rashid-Farrokhi, F., et al., ‘‘Spectral Efficiency of Wireless Systems with Multiple-Transmit and Receive Antennas,’’ *PIMRC,* London, UK, September 2000, pp. 373–377. [14] Nabar, R., et al., ‘‘Performance of Multi-Antenna Signaling Techniques in the Presence of Polarization Diversity,’’ *IEEE Trans. Sig. Proc.,* 50(10), October 2002, pp. 2553–2562. [15] Vaughan, R., ‘‘Polarization Diversity in Mobile Communications,’’ *IEEE Trans. Veh. Tech.,* 39(3), August 1990, pp. 177–186.

[16] Baum, D. S., et al., ‘‘Measurement and Characterization Broadband MIMO Fixed Wireless Channels at 2.5 GHz,’’ *Proc. IEEE Int. Conf. Pers. Wireless Comm.,* Hyderabad, India, December 2000, pp. 203–206.

[17] Erceg, V., et al., ‘‘Comparison of the WISE Propagation Tool Prediction With Experimental Data Collected in Urban Microcellular Environments,’’ *IEEE J. Sel. Areas Comm.,* 15(4), May 1997, pp. 677–684.

[18] Raleigh, G., and J. Cioffi, ‘‘Spatio-Temporal Coding for Wireless Communication,’’ *IEEE Trans. Comm.* 46(3), March 1998, pp. 357–366.

Channel Propagation, Fading, and Link Budget Analysis

3.1 Introduction

In Chapter 2 we examined the mutual information capacity of wireless communica tion based on MIMO channels. We found that this capacity grows linearly with the number of antennas in flat fading channels, due to the increase in the number of spatial data pipes. All this is accomplished *without* increasing the bandwidth or power.

In this chapter, we examine channel fading and propagation issues. We will also discuss a few channel propagation models and carry out link budget analysis. Finally we examine certain diversity combining techniques like selection diversity, maximal ratio combining, and equal gain combining.

3.2 Radio Wave Propagation

The mobile radio channel experiences a lot of limitations on the performance of wireless systems. The transmission path can vary from line-of-sight to one severely obstructed by buildings and foliage. Unlike wired channels, radio channels are extremely random and do not offer easy analysis. The speed of motion, for example, impacts on how the signal level fades as the mobile terminal moves in space. This modeling is therefore based more on statistics and requires specific measurements for an intended communication system.

Broadly the mechanics of electromagnetic wave propagation are confined to reflection, diffraction, and scattering.

Reflection, diffraction, and scattering are the three basic propagation mecha nisms for radio waves. Received power (or its reciprocal, path loss) is generally

the most important parameter predicted by large-scale propagation models and is based on these three phenomena. This is also applicable to small-scale fading and multipath propagation, which will be discussed later in this chapter.

3.2.1 Reflection

This occurs when electromagnetic waves bounce off objects whose dimensions are large compared with the wavelength of the propagating wave. They usually occur

from the surface of the earth and off buildings and walls. A radio wave, when propagating in one medium, impinges on another medium with different electrical properties and is partially transmitted and reflected. The plane wave is incident on a perfect dielectric, part of the energy is transmitted into the second medium and part is reflected back into the first medium without any loss of energy in absorption. If the second medium is a perfect conductor then all of the incident energy is reflected back into the first medium without loss of energy. The electric field intensity of the reflected and transmitted waves may be related to the incident wave in the medium of origin through the *Fresnel reflection coefficient.* This reflec tion coefficient is a function of the material properties and generally depends on the wave polarization, angle of incidence, and frequency of the propagating wave.

3.2.2 Diffraction

Diffraction allows radio signals to propagate around the curved surface of the earth, beyond the horizon, and behind obstructions. The received field strength decreases rapidly as a receiver moves deeper into the obstructed (shadowed) region and is usually of enough intensity so as to produce a discernable signal. The phenomenon of diffraction can be explained by the Huygens principle, which states that all points on a wavefront can be considered as point sources for the production of secondary wavelets and that these secondary wavelets combine to produce a new wavefront in the direction of propagation. Diffraction is caused by the propaga tion of secondary wavelets into a shadowed region. The field strength of a diffracted wave in the shadowed region is the vector sum of the electric field components of all the secondary wavelets in the space around the obstacle.

3.2.3 Scattering

The actual received signal in a mobile radio environment is often stronger than what is predicted by reflection and diffraction models alone. This occurs because when a radio wave impinges on a rough surface, the reflected energy is spread out (diffused) in all directions due to scattering. Objects such as lampposts and trees tend to scatter energy in all directions, thereby providing additional radio energy at the receiver.

Sometimes reflection, diffraction and scattering are collectively referred to as scattering. Further discussions on these phenomena are beyond the scope of this book and the interested reader is referred to [1–4].

Cellular systems usually operate in urban areas, where there is no direct line of-sight path between the transmitter and receiver and where high-rise buildings cause severe diffraction loss. Multiple reflections from various objects cause the electromagnetic waves to travel along different paths of varying lengths. The interac tion between these waves causes multipath fading at a given location, because their phases are such that sometimes they add and sometimes they subtract (fade). The strengths of these waves slowly reduce with distance from the transmitter.

Propagation models based on average-received signal strength at a given dis tance from the transmitter are useful to estimate a radio coverage area and are called *large-scale propagation models* or *macroscopic fading models.* They are

characterized by a large separation—usually a few kilometers—between the trans mitter and receiver. On the other hand, propagation models that characterize the rapid fluctuations of the received signal strength over very short distances (a few wavelengths) or short time durations (on the order of seconds) are called *small scale fading models* or *microscopic fading models.* The latter gives rise to rapid fluctuations as the mobile moves over short distances and the received power sometimes varies as much as 30 to 40 dB when the receiver moves only a fraction of a wavelength. Large-scale fading is manifest when the mobile moves over larger distances, causing the local average signal level to gradually decrease. It is this local average signal level that is predicted by large-scale propagation models. We shall examine a few of these models in this chapter.

Typically, the local average-received power is measured by averaging signal measurements over a measurement track of 5l to 40l. For cellular frequencies in the 1-to 2-GHz band, this works out to movements of 1 to 10m [1]. Figure 3.1 shows an example of small-scale fading and large-scale fading. Small-scale fading movements are rapid fluctuations, whereas large-scale fading movements are much slower average changes in signal strength [1]. The statistical distribution of this mean is influenced by parameters like frequency, antenna heights, environments and so on. However, it has been observed that the received power averaged over microscopic fading approaches a normal distribution when plotted on a logarithmic scale (i.e., in decibels) and is called log-normal distribution [5]. It is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.1)$$

In (3.1) x is in decibels and is a random variable representing the long-term signal power level fluctuation; m and s are, respectively, the mean and standard

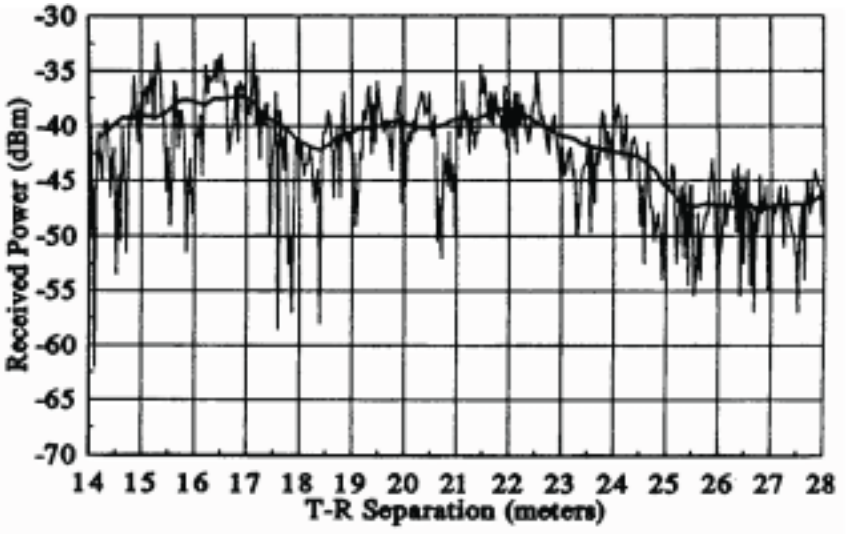

Figure 3.1 Small-scale and large-scale fading. (*From:* [1]. © 2002. Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.)

deviation of *x* expressed in decibels. m is the path loss described earlier. A typical value for s is 8 dB [6].

3.3 Large-Scale Fading or Macroscopic Fading

3.3.1 Free-Space Propagation Model

If there is a clear unobstructed line-of-sight path between the transmitter and receiver, then we resort to the free-space propagation model. Satellite communica tion systems and microwave line-of-sight radio links undergo free-space propaga tion. In this model, the power is presumed to decay with distance from the transmitter according to some power law, usually as square of the distance from the transmitter. The free-space power received by an antenna at a distance *d* from the transmitter is given by [1],

$$P_r(d) = \frac{P_t G_1 G_2 \lambda^2}{(4\pi)^2 d^2 L} \quad (3.2)$$

where P_t is the transmitted power, $P_r(d)$ is the received power as a function of the separation distance d in meters, G_1 is the transmit antenna gain, G_2 is the receive antenna gain, L is the system loss not related to propagation ($L \geq 1$) and λ is the wavelength in meters. The gain of an antenna is related to its effective aperture by

$$G = \frac{4\pi A_e}{\lambda^2} \quad (3.3)$$

l is related to the carrier frequency by

$$\lambda = \frac{c}{f} \quad (3.4)$$

where f is the carrier frequency in Hz and c is the speed of light in meters/sec ($3 \cdot 10^8$ m/sec). The values of P_t and P_r must be expressed in identical units and G_t and G_r are dimensionless quantities. The miscellaneous losses ($L \geq 1$) are usually due to transmission line attenuation (plumbing losses), filter losses, and antenna losses in the communication system. $L = 1$ indicates no losses in the system hardware.

Equation (3.2) shows that the received power falls off as the square of the separation distance d. This implies that the received power decays with distance at a rate of 20 dB/decade.

We define an *isotropic* radiator as an ideal antenna that radiates power with unit gain uniformly in all directions and is often used as a reference antenna gain in wireless systems. The *effective isotropic radiated power* (EIRP) is defined as

$$EIRP = P_t G_t \quad (3.5)$$

and represents the maximum radiated power available from a transmitter in the direction of maximum antenna gain compared with an isotropic radiator. In prac tice, antenna gains are given in units of dBi (dB gain with respect to an isotropic antenna).

The path loss, which is the amount of attenuation suffered by the signal in dBs, is defined as the difference (in dB) between the transmitted power and the received power and is given by

$$P_r = -10 \log \left[G_t G_r \lambda^2 \right]$$

$$PL \text{ (dB)} = 10 \log \frac{P_t}{\left[G \right]_{(4\pi)^2 d^2}} \quad (3.6)$$

It is important to note that the free-space model is only applicable in the so called far-field region of the transmitting antenna or in the *Fraunhofer* region and is defined as

$$d_f = \frac{2D^2}{\lambda} \quad (3.7)$$

where D is the largest physical linear dimension of the antenna (e.g., the length of a rectangular array antenna).

From (3.2), we note that the equation does not hold for $d = 0$. Hence, large scale propagation models use a close-in distance, d_0 , as a known received

power reference point. The received power at any distance $d > d_0$ may then be related to P_r and d_0 . The value $P_r(d_0)$ may be predicted from (3.2) by extrapolation or may be measured in the radio environment by taking the average received power at many points located at a close-in radial distance d_0 from the transmitter. The reference distance must be so chosen that it lies in the far-field (i.e., $d_0 \geq d_f$) and d_0 is chosen to be smaller than any practical distance used in the mobile communication system. Thus from (3.2), the received power in free space at a distance greater than d_0 is given by

$$P_r(d) = P_r(d_0)\left(\frac{d_0}{d}\right)^2, \quad d \geq d_0 \geq d_f \tag{3.8}$$

In mobile radio systems, P_r changes by many orders of magnitude over a typical coverage area of several square kilometers. In view of this very large dynamic range of received power levels, dBm or dBW units are used to express received power levels. dBm is the power in dBs referred to one milliwatt. dBW is the power in dBs referred to one watt. For example,

$$0.001\text{W}\left[\frac{G}{}\right] + 20\log\left(\frac{d_0}{d}\right), \quad d \geq d_0 \geq d_f \tag{3.9}$$

$$P_r(d)\ \text{dBm} = 10\log\left[\frac{P_r(d_0)}{}\right]$$

where $P_r(d_0)$ is in watts.

The reference distance d_0 for practical systems using low-gain antennas in the 1–2 GHz region is typically 1m in indoor environments and 100m or 1 Km in

outdoor environments, so that the numerator in (3.8) and (3.9) is a multiple of 10. This makes loss computations easy in dB units [1].

We now illustrate what we have learned through examples [1].

Example 1

Find the far-field distance for an antenna with maximum dimension of 2m and operating frequency of 900 MHz.

Solution

Given:

Largest dimension of antenna, D = 2m.

Operating frequency f_c = 900 MHz, $\lambda = c/f = \frac{3 \cdot 10^8 \text{ m/s}}{900 \cdot 10^6 \text{ Hz}} = 0.33\text{m}$

Using (3.7),

$$d_f = \frac{2D^2}{\lambda} = \frac{2(2)^2}{}$$

$$0.33 = 24.24\text{m}$$

Example 2

If a transmitter produces 50W of power, express the transmit power in units of (a) dBm and (b) dBW. If 50W is applied to an antenna of gain 1, find the received power in dBm at a free-space distance of 100m from the antenna. What is P_r (10 km)? Assume a gain of 2 for the receiver antenna and no system losses.

Solution

Given:

Transmitter power, $P_t = 50\text{W}$

Carrier frequency, $f_c = 900$ MHz

Using (3.9)

(a) Transmitter power,

$$P_t(\text{dBm}) = 10 \log [P_t(\text{mW})/(1\text{ mW})]$$
$$= 10 \log [50 \cdot 10^3] = 47 \text{ dBm}$$

(b) Transmitter power,

$$P_t(\text{dBW}) = 10 \log [P_t(\text{W})/(1\text{W})]$$
$$= 10 \log [50] = 17 \text{ dBW}$$

The received power can be determined using (3.2)

$$P_r = P_t G_t G_r \lambda^2$$
$$(4\pi)^2 d^2 L = 50(1)(2)(0.33)^2$$
$$(4\pi)^2 (100)^2 (1) = 6.9 \cdot 10^{-3} \text{ mW}$$

$$P_r(\text{dBm}) = 10 \log P_r(\text{mW}) = 10 \log (6.9 \cdot 10^{-3} \text{ mW}) = -21.6 \text{ dBm}$$

The received power at 10 Km can be expressed in terms of dBm using (3.9) where $d_0 = 100$m and $d = 10$ Km

$$10{,}000 \frac{G}{F} = -21.6 - 40 \text{ dB}$$

$$P_r(10 \text{ Km}) = P_r(100) + 20 \log \frac{100}{}$$

$$= -61.6 \text{ dBm}$$

3.3.2 Outdoor Propagation Models

Free-space propagation is rarely encountered in real-life situations. In reality, we need to take into account the terrain profile in a particular area for estimating

path loss. The terrain may vary from a simple curved earth profile to a highly mountainous profile. The presence of trees, buildings, and other obstacles must be taken into account. A number of propagation models are available to predict path loss over irregular terrain. These models differ in their ability to predict signal strength at a particular receiving point or in a specific local area (called a sector) because their approach is different and their results vary in terms of accuracy and complexity. These models are based on iterative experiments conducted over a period of time by measuring data in a specific area. We discuss two such well known models.

3.3.2.1 Okumura Model

This is a widely used model for signal prediction in an urban area. It is applicable for frequencies in the range of 150 to 1,920 MHz and can be extrapolated up to 3 GHz and distances of 1 to 100 Km. It can be used for base station antenna heights ranging from 30 to 1,000m.

Okumura [7] developed a set of curves giving the median attenuation relative to free space (A_{mu}) in an urban area over a quasi-smooth terrain with a base station effective antenna height (h_{te}) of 200m and a mobile antenna height (h_{re}) of 3m. These curves were developed from extensive measurements using vertical omni-directional antennas at both base and mobile and are plotted as a function of frequency in the range 100 to 1,920 MHz and as a function of distance from the base station in the range of 1 to 100 Km. To use these curves, we first deter mine the free-space path loss between the points of interest and then the value of $A_{mu}(f, d)$ (as read from the curves) is added to it along with correction factors to account for the type of terrain. The model is expressed as

$$L_{50}(\text{dB}) = L_F + A_{mu}(f, d) - G(h_{te}) - G(h_{re}) - G_{\text{AREA}} \quad (3.10)$$

where L_{50} is the 50th percentile (i.e., median) value of propagation path loss, L_F is the free-space propagation loss, A_{mu} is the median attenuation relative to free space, $G(h_{te})$ is the base station antenna height gain factor, $G(h_{re})$ is the mobile antenna height gain factor, and G_{AREA} is the gain due to the type of environment. The antenna height gains are strictly a function of height and have nothing to do with the antenna patterns.

Plots of $A_{mu}(f, d)$ and G_{AREA} for a wide range of frequencies are shown in Figures 3.2 and 3.3 [1]. Moreover, Okumura determined that $G(h_{te})$ varies at a rate of 20 dB/decade and $G(h_{re})$ varies at a rate of 10 dB/decade for heights of less than 3m.

$$G(h_{te}) = 20 \log \left(\frac{h_{te}}{200} \right), \quad 1{,}000\text{m} > h_{te} > 30\text{m} \quad (3.11)$$

$$\left(\frac{h_{re}}{3} \right), \quad h_{re} \leq 3\text{m} \quad (3.12)$$

$$G(h_{re}) = 10 \log \left(\frac{h_{re}}{3}\right)^{D}, \quad 10\text{m} > h_{re} > 3\text{m} \quad (3.13)$$

$$G(h_{re}) = 20 \log \left(\frac{h_{re}}{3}\right)^{D}$$

Other corrections may also be applied to Okumura's model. Some of these are terrain undulation height (Dh), isolated ridge height, average slope of the terrain, and the mixed land-sea parameter. Once the terrain-related parameters are calcu lated, the necessary correction factors can be added or subtracted as required. All these correction factors are also available as Okumura curves [7].

Okumura's model is completely based on measured data and there is no analysis to justify it. All extrapolations to these curves for other conditions are highly

Figure 3.2 Median attenuation relative to free space (A_{mu} (f, d)) over a quasi-smooth terrain. (*From:* [7]. © 1968, IEEE.)

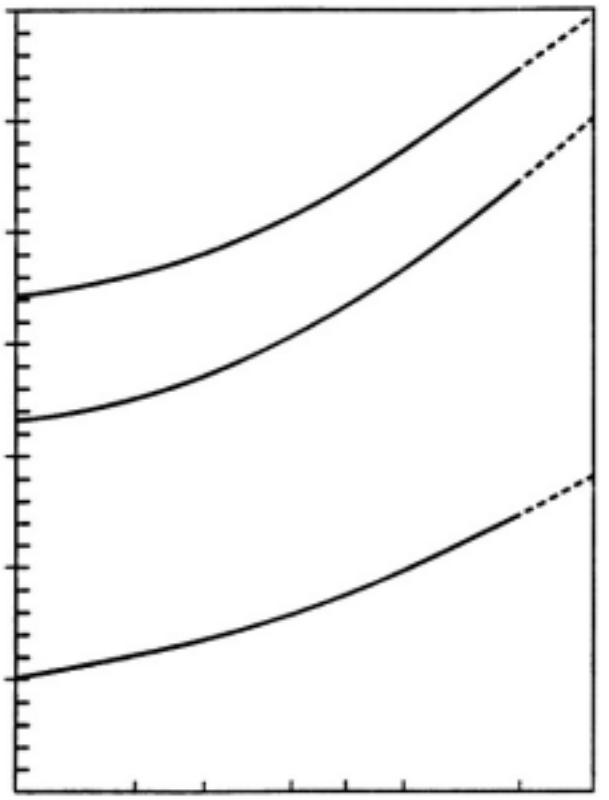

Figure 3.3 Correction factor G_{AREA} for different types of terrain. (*From:* [7]. © 1968, IEEE.)

subjective. Yet it is considered the simplest and best in terms of accuracy in path loss prediction for cellular systems in a cluttered environment. It has become a standard in Japan. The major disadvantage is its slow response to rapid changes in terrain. Hence, it is not so good in rural areas. Common standard deviations between predicted and measured path loss values are around 10 dB to 14 dB.

We now work out an example by way of illustration [1].

Example 3

Find the median path loss using Okumura's model for d = 50 Km, h_{te} = 100m, h_{re} = 10m in a suburban environment. If the base station transmitter radiates an EIRP of 1 kW at a carrier frequency of 900 MHz, find the power at the receiver (assume a gain of 2 at the receiving antenna).

Solution

The free-space path loss can be calculated using (3.2) as

$$\frac{(4\pi)^2 d^2}{G} = 10 \log \frac{F}{0.33_2}$$

$$L_F = 10 \log \frac{F}{I_2} \qquad \frac{(4\pi)^2 \cdot (50 \cdot 10^3)}{G} = 125.5 \text{ dB}$$

From the Okumura curves

$$A_{mu}\,(900\text{ MHz (50 km)}) = 43\text{ dB}$$

and

$$G_{\text{AREA}} = 9\text{ dB}$$

Using (3.11) and (3.13)

$$_{200}D = 20\log S_{100}$$

$$G(h_{te}) = 20\log S_{\underline{h_{te}}}\ {}_{200}D = -6\text{ dB}$$

Using

$$_{3}D = 20\log S_{10}\ {}_{3}D = 10.46\text{ dB}$$

$$G(h_{re}) = 20\log S_{\underline{h_{re}}}$$

$$L_{50}\,(\text{dB}) = L_F + A_{mu}(f, d) - G(h_{te}) - G(h_{re}) - G_{AREA} = 125.5\text{ dB} + 43\text{ dB} - (-6\text{ dB}) - 10.46\text{ dB} - 9\text{ dB} = 155.04\text{ dB}$$

Therefore, the median received power is

$$P_r(d) = EIRP\,(\text{dBm}) - L_{50}\,(\text{dB}) + G_r\,(\text{dB})$$
$$= 60\text{ dBm} - 155.04\text{ dB} + 3\text{ dB} = -92.04\text{ dBm}$$

3.3.2.2 Hata Model

The Hata model is an empirical formulation of the graphical path loss data provided by Okumura and is valid from 150 to 1,500 MHz. Hata presented the loss as a standard formula and supplied correction equations for application to other situa tions. The standard formula for median path loss in urban areas is given by [1]

$$L_{50}\,(urban)\,(\text{dB}) = 69.55 + 26.16\log f_c - 13.82\log h_{te} - a(h_{re})\quad (3.14) + (44.9 - 6.55\log h_{te})\log d$$

where f_c is the frequency in MHz from 150 to 1,500 MHz, h_{te} is the effective transmitter (base station) antenna height (in meters) ranging from 30 to 200m, h_{re} is the effective receiver (mobile) antenna height (in meters) ranging from 1 to 10m, d is the T-R separation distance (in Km), and $a(h_{re})$ is the correction factor for effective mobile antenna height, which is a function of the size of the coverage area. For a small to medium-sized city, the correction factor is given

by

$$a(h_{re}) = (1.1 \log f_c - 0.7)h_{re} - (1.56 \log f_c - 0.8) \text{ dB} \quad (3.15)$$

and for a large city,

$$a(h_{re}) = 8.29(\log 1.54h_{re})^2 - 1.1 \text{ dB for } f_c \leq 300 \text{ MHz} \quad (3.16)$$

$$a(h_{re}) = 3.2(\log 11.75h_{re})^2 - 4.97 \text{ dB for } f_c \geq 300 \text{ MHz} \quad (3.17)$$

To obtain the path loss in a suburban area, the standard Hata formula in (3.13) is modified as

$$L_{50}(\text{dB}) = L_{50}(urban) - 2[\log (f_c/28)]^2 - 5.4 \quad (3.18)$$

and for path loss in open rural areas, the formula is modified as

$$L_{50}(\text{dB}) = L_{50}(urban) - 4.78(\log f_c)^2 + 18.33 \log f_c - 40.94 \quad (3.19)$$

The predictions of Hata's model compare very closely with the original Oku mura model, if d exceeds 1 Km. This model is well-suited to large cell mobile systems.

This concludes our discussions on outdoor propagation models. The interested reader is referred to [1] and the references listed therein.

3.4 Small-Scale Fading

Small-scale fading or simply *fading* is used to describe the rapid fluctuations of the amplitude, phases, or multipath delays of a radio signal over a short period of time or travel distance, *so that large-scale path loss effects may be ignored.* Fading is caused by a number of signals (two or more) arriving at the reception point through different paths, giving rise to constructive (strengthening) vectorial sum ming of the signal or destructive (weakening) vectorial subtraction of the signals, depending on their phase and amplitude values. These different signals other than the main signal are called *multipath* waves.

Multipath in a radio channel creates small-scale fading effects. These effects are commonly characterized as causing:

- Rapid changes in signal strength over a small travel distance or time interval.
- Random frequency modulation due to varying Doppler shifts on different multipaths.
- Time dispersion (echoes) caused by multipath propagation delays.

Even when a line-of-sight exists, multipath still occurs due to reflections from the ground or surrounding structures. Assume that there is no moving object in the channel. In such a case, fading is purely a spatial phenomenon. The signals add or subtract, creating standing waves in the area where the mobile is located. In such a case, as the mobile moves, it encounters temporal fading as it moves

through the multipath field. In a more serious case, the mobile may stop at a particular point at which the received signal is in deep fade. Maintaining good communication in that case becomes very difficult. It can only be countered using diversity techniques, as discussed in Chapter 2. Figure 2.2 is one such example.

3.4.1 Microscopic Fading

Microscopic fading refers to the rapid fluctuations of the received signal in space, time, and frequency and is caused by the signal scattering off objects between the transmitter and receiver. Since this fading is a superposition of a large number of independent scattered components, then by the central limit theorem, the compo nents of the received signal can be assumed to be independent zero mean Gaussian processes. The envelope of the received signal is consequently Rayleigh distributed and is given by

$$f(x) = \frac{2x}{V} e^{-\frac{x^2}{V}} u(x) \quad (3.20)$$

where V is the average received power and $u(x)$ is the unit step function defined as

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3.21)$$

If there is a direct LOS path between the transmitter and receiver, the signal envelope is no longer Rayleigh and the distribution of the signal is Ricean. The Ricean distribution is often defined in terms of the Ricean factor, *K,* which is the ratio of the power in the mean component of the channel to the power in the scattered component. The Ricean PDF of the envelope is given by

$$f(x) = \frac{2(K+1)x}{V} e^{-K - \frac{(K+1)x^2}{V}} I_0\left(2x\sqrt{\frac{K(K+1)}{V}}\right) u(x) \quad (3.22)$$

where I_0 is the zero-order modified Bessel function of the first kind defined as

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{-x\cos u}\, du \quad (3.23)$$

In the absence of a direct path, $K = 0$ and the Ricean PDF reduces to Rayleigh PDF, since $I_0(0) = 1$. Figure 3.4 shows the combined effects of path loss and macroscopic and microscopic fading on received power in a wireless channel [6].

We note from Figure 3.4 that the mean propagation loss increases monotoni cally with range. Local deviations from this mean occur due to macroscopic and microscopic fading.

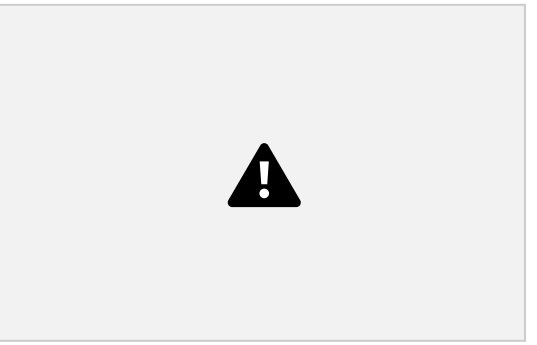

Figure 3.4 Signal power fluctuation versus range in wireless channels. (*From:* [6]. Reprinted with the permission of Cambridge University Press.)

There are three types of microscopic fading [6]:

- Doppler spread-time selective fading;
- Delay spread-frequency selective fading;
- Angle spread-space selective fading.

3.4.1.1 Doppler Spread-Time Selective Fading

Time varying fading due to the motion of a scatterer or the motion of a transmitter or receiver or both results in Doppler spread. The term spread is used to denote the fact that a pure tone frequency f_c in hertz spreads across a finite bandwidth ($f_c \pm f_{\max}$). There is a direct relationship between the autocorrelation function of a signal and its spectrum and is defined by the Wiener-Khinchin equations [8]. The Fourier transform of the time autocorrelation of the channel response to a continu ous wave (CW) tone is defined as Doppler power spectrum $c_{Do}(f)$ with $f_c - f_{\max} \le f \le f + f_{\max}$. Figure 3.5(a) shows the Doppler spectrum [6].

The Doppler power spectrum has a classical U-shaped form and is approxi mated by Jakes model [5]. The Doppler shift of the received signal denoted by f_d is given by

$$f_d = vf\frac{C}{c}\cos u \quad (3.24)$$

where v is the velocity of the moving object (or vehicle speed, if we are talking about static scatterers and a moving vehicle), u is the relative angle between the moving object and the point of reception of the Doppler signal, and c is the speed of light. Obviously, the maximum Doppler will be received at a relative angle of 0° (i.e., when the moving object is ahead or astern).

The root mean square (RMS) bandwidth of $c_{Do}(f)$ is called the Doppler spread and is given by

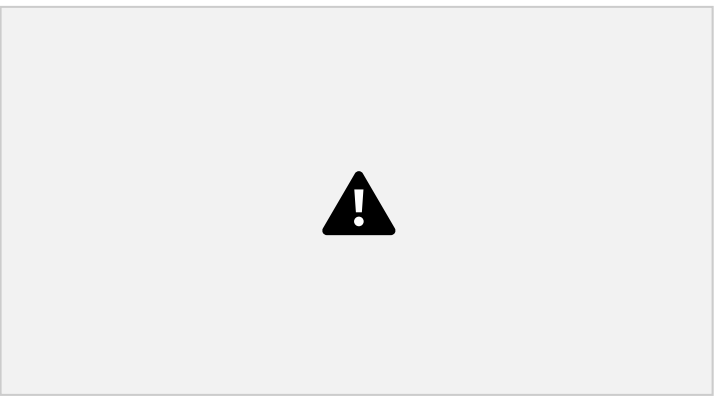

Figure 3.5 (a) Typical Doppler power spectrum. (*From:*[6]. Reprinted with permission of Cambridge University Press.) (b) Fixed wireless Doppler spectra. (2000, IEEE.)

$$f_{RMS} = \sqrt{E_{(f-f)^2} c_{Do}(f)\, df}$$

$$E_{c_{Do}(f)\, df} \tag{3.25}$$

where f is the average frequency of the Doppler spectrum and is given

by $\int_{f_c} C_{Do}(f)\,df$

$$\int_{-\infty}^{\infty} C_{Do}(f)\,df \quad (3.26)$$

$f =$

In LOS cases the spectrum is modified by an additional discrete frequency component given by f_d. We define coherence time of the channel as

$$T_c \approx 1$$

$$f_{RMS} \quad (3.27)$$

where T_c is defined as the time lag for which the signal autocorrelation coefficient reduces to 0.7. It serves as a measure of how fast the channel changes in time, implying that the larger the coherence time, the slower the channel fluctuation. The Doppler spectrum shown in Figure 3.5(a) pertains to a mobile receiver moving at constant speed (e.g., in a car). However, in a fixed wireless channel, the receiver is static but there is movement in the environment (e.g., trees and foliage moving in a random manner due to wind). In such cases, the Doppler spectrum is as shown in Figure 3.5(b) [9].

In Figure 3.5(b), the left-hand graph pertains to a low Doppler spread and the right-hand one pertains to a higher Doppler spread. The curves level off at high Doppler shifts due to the prevailing noise levels. If there is moving traffic around the mobile, Doppler components can occur at much higher frequencies, but the shape of the Doppler spectra will be similar.

3.4.1.2 Delay Spread-Frequency Selective Fading

The small-scale variations of a mobile radio signal can be directly related to the impulse response of the mobile radio channel. This stems from the fact that a mobile radio channel may be modeled as a linear filter with a time varying impulse response, where the time variation is due to receiver motion in space. The filtering nature of the channel is caused by the summation of amplitudes and delays of the multiple arriving waves at any instant of time. Therefore, the impulse response is a useful characterization of the channel because it can be used to predict and compare the performance of many different mobile communication systems and transmission bandwidths for a particular channel condition.

To compare different multipath channels and develop some general design guidelines for wireless systems, certain parameters were decided on as benchmarks to quantify the multipath channel. These parameters are the mean excess delay, RMS delay spread, and excess delay spread and they can be determined from the power delay profile. These are shown in Figure 3.6 [1].

We define [1]:

- *Mean excess delay:* The mean excess delay is the first moment of the power delay profile and is defined as

$$\bar{t} = \frac{\sum_k P(t_k) t_k}{\sum_k P(t_k)} \quad (3.28)$$

- *RMS delay spread:* The RMS delay spread is the square root of the second central moment of the power delay profile and is defined as

$$\sigma_t = \sqrt{\overline{t^2} - (\bar{t})^2} \quad (3.29)$$

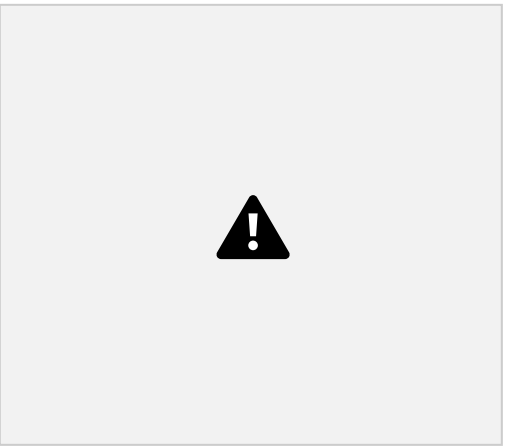

Figure 3.6 Example of an indoor power delay profile. (*From:* [1]. © 2002. Reprinted by permission of Pearson Education Inc., Upper Saddle River, NJ.)

where

$$\overline{t^2} = \frac{\sum_k P(t_k) t_k^2}{\sum_k P(t_k)} \quad (3.30)$$

These delays are measured relative to the first detectable signal arriving at the receiver at $t_0 = 0$. Equations (3.28) to (3.30) do not rely on the absolute power level of $P(t)$, but only on the relative amplitudes of the multipath components within $P(t)$. Typical values of RMS delay spread are on the order of microseconds in outdoor mobile radio channels and on the order of nanoseconds in indoor radio channels.

- *Maximum excess delay* (X dB): This is defined to be the time delay during which multipath energy falls to X dB below the maximum. This implies that maximum excess delay is defined as $t_x - t_0$, where t_0 is the first arriving signal and t_k is the maximum delay at which a multipath component is within X dB of the strongest arriving multipath signal (which does not necessarily arrive at t_0). Figure 3.6 illustrates the computation

of maximum excess delay for multipath components within 10 dB of the maximum. The maximum excess delay tells us how long a multipath exists above a given threshold. This value t_k must be specified with a threshold that relates the multipath noise floor to the maximum received multipath component. This is also sometimes called *excess delay spread.*

In practice, the values for $\bar{t}$, $\overline{t^2}$ and s_l depend on the choice of noise threshold used to process $P(t)$. The noise threshold is used to differentiate between received multipath components and thermal noise. If the noise threshold is too low, then

noise will be processed as multipath, thus giving rise to values of $\bar{t}$, $\overline{t^2}$ and s which are artificially high [1].

Delay spread causes frequency selective fading as the channel acts like a tapped delay line filter. Frequency selective fading can be characterized in terms of coher ence bandwidth, B_c , which is the frequency lag for which the channel's autocorrela tion function reduces to 0.7 (remember that the autocorrelation function and the spectrum are connected by the Wiener-Khinchin equations [8]). We define coherence bandwidth as

$$B_c \approx 1$$

s_t(3.31)

When the coherence bandwidth is comparable with or less than the signal bandwidth, the channel is said to be frequency selective. Otherwise it is frequency flat or non-selective. A ''flat'' channel passes all spectral components with approxi mately equal gain and linear phase. It is not possible to provide an exact relationship between coherence bandwidth and RMS delay spread, as it is a function of specific channel impulse response and applied signals. We need to resort to spectral analysis to determine the exact impact that time varying multipath has on a transmitted signal. Hence, accurate multipath modeling is essential.

3.4.1.3 Rician K-Factor Measurement

There are many techniques to measure Rician K-factor from the power profile. The moment-method estimation of K-factor has found popular appeal [10]. The details are beyond the scope of this book.

We now solve an example to firm up our ideas!

Example 4

Calculate the mean excess delay, RMS delay spread, and maximum excess delay (10 dB) for the multipath profile given in Figure 3.7. Estimate the coherence bandwidth of the channel.

Solution

Using definition of maximum excess delay (10 dB), we determine that $t_{10\ dB} = 3$ msec.

Figure 3.7 Multipath profile for Example 4.

The mean excess delay is [using (3.28)]

$$\bar{t} = (0.01)(0) + (0.01)(1) + (1)(2) + (0.1)(3) + (0.01)(4)$$

$$[0.01 + 0.01 + 1 + 0.1 + 0.01] = 2.35$$

$$1.13 = 2.08 \text{ msec}$$

The second moment for the given power delay profile is (using 3.30)

$$\overline{t^2} = (0.01)(0)^2 + (0.01)(1)^2 + (1)(2)^2 + (0.1)(3)^2 + (0.01)(4)^2$$

$$[0.01 + 0.01 + 1 + 0.1 + 0.01] = 5.07$$

$$1.13 = 4.49 \text{ msec}$$

Therefore the RMS delay spread is (using 3.29)

$$\sigma_t = \sqrt{4.49 - (2.08)^2} = 0.4 \text{ msec}$$

The coherence bandwidth is (using 3.31)

$$B_c = \frac{1}{\sigma_t} = \frac{1}{0.4 \text{ msec}} = 2.5 \text{ MHz}$$

Hence, communications systems operating within a bandwidth of 2.5 MHz need not use equalizers.

3.4.1.4 Angle Spread-Space Selective Fading

Angle spread at the receiver refers to the angle of arrival (AOA) of the multipath components at the receive antenna. Similarly, the angle of departure (AOD) from the transmitter of the multipath that reaches the receiver is called the angle

spread at the transmitter.

We denote AOA by u and the rest of the analysis is as was done for delay spread, the only difference being that instead of t we substitute u. The terminology now becomes RMS angle spread and is given by

$$s_u = \sqrt{u^2 - (u)^2} \quad (3.32)$$

where

$$u^2 = \sum_k P(u_k) u^2_k \quad (3.33)$$

A typical angle (power) spectrum is shown in Figure 3.8.

The RMS angle spread is measured similar to the RMS delay spread. These angles are measured relative to the first detectable signal arriving at the receiver at $u_0 = 0$.