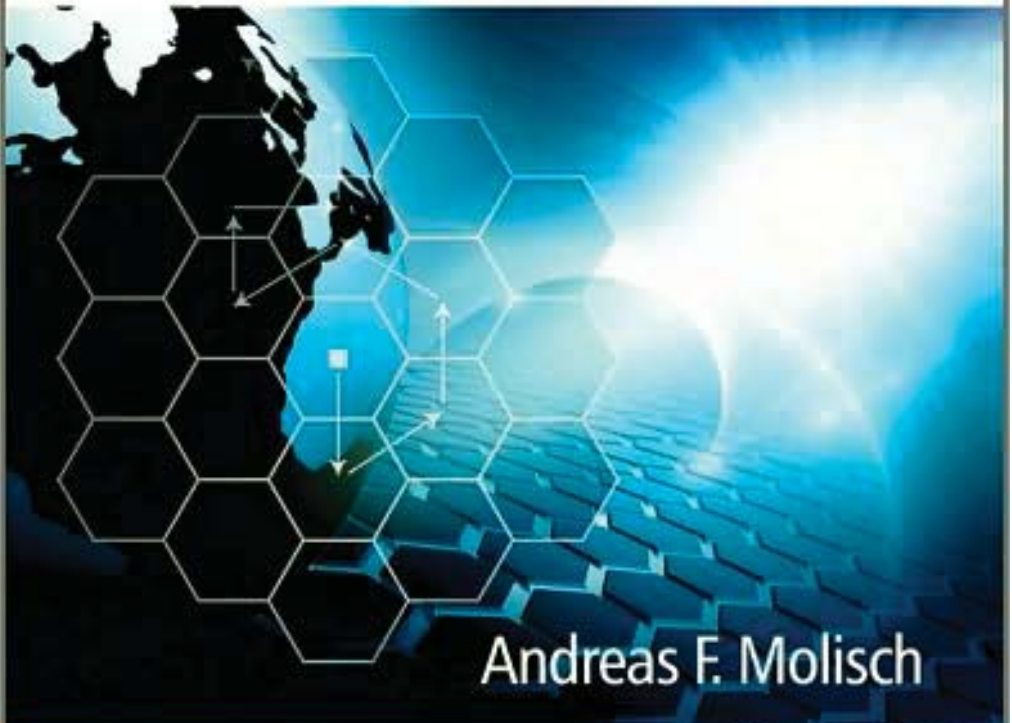


Second Edition

Wireless Communications



Andreas F. Molisch

 WILEY

 IEEE



WIRELESS COMMUNICATIONS

Wireless Communications, Second Edition Andreas F. Molisch
© 2011 John Wiley & Sons Ltd. ISBN: 978-0-470-74187-0

WIRELESS COMMUNICATIONS

Second Edition

Andreas F. Molisch, *Fellow, IEEE*
University of Southern California, USA



A John Wiley and Sons, Ltd., Publication

This edition first published 2011
© 2011 John Wiley & Sons Ltd.

First edition published 2005

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloguing-in-Publication Data

Molisch, Andreas F.

Wireless communications / Andreas F. Molisch. – 2nd ed.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-74187-0 (cloth) – ISBN 978-0-470-74186-3 (pbk.)

1. Wireless communication systems—Textbooks. I. Title.

TK5103.2.M65 2011

621.3845'6—dc22

2010017177

A catalogue record for this book is available from the British Library.

Print ISBN: 9780470741870 (H/B)

Print ISBN: 9780470741863 (P/B)

ePDF ISBN: 9780470666692

Typeset in 9/11 Times by Laserwords Private Limited, Chennai, India.

Contents

Preface and Acknowledgements to the Second Edition	xxiii
Preface to the First Edition	xxv
Acknowledgments to the First Edition	xxix
Abbreviations	xxxi
Symbols	xlvi
 Part I INTRODUCTION	 1
 1 Applications and Requirements of Wireless Services	 3
1.1 History	4
1.1.1 <i>How It All Started</i>	4
1.1.2 <i>The First Systems</i>	4
1.1.3 <i>Analog Cellular Systems</i>	5
1.1.4 <i>GSM and the Worldwide Cellular Revolution</i>	6
1.1.5 <i>New Wireless Systems and the Burst of the Bubble</i>	7
1.1.6 <i>Wireless Revival</i>	8
1.2 Types of Services	8
1.2.1 <i>Broadcast</i>	8
1.2.2 <i>Paging</i>	9
1.2.3 <i>Cellular Telephony</i>	10
1.2.4 <i>Trunking Radio</i>	12
1.2.5 <i>Cordless Telephony</i>	12
1.2.6 <i>Wireless Local Area Networks</i>	14
1.2.7 <i>Personal Area Networks</i>	14
1.2.8 <i>Fixed Wireless Access</i>	14
1.2.9 <i>Ad hoc Networks and Sensor Networks</i>	15
1.2.10 <i>Satellite Cellular Communications</i>	16
1.3 Requirements for the Services	16
1.3.1 <i>Data Rate</i>	16
1.3.2 <i>Range and Number of Users</i>	17
1.3.3 <i>Mobility</i>	18
1.3.4 <i>Energy Consumption</i>	19
1.3.5 <i>Use of Spectrum</i>	20

1.3.6	<i>Direction of Transmission</i>	21
1.3.7	<i>Service Quality</i>	21
1.4	Economic and Social Aspects	22
1.4.1	<i>Economic Requirements for Building Wireless Communications Systems</i>	22
1.4.2	<i>The Market for Wireless Communications</i>	23
1.4.3	<i>Behavioral Impact</i>	24
2	Technical Challenges of Wireless Communications	27
2.1	Multipath Propagation	27
2.1.1	<i>Fading</i>	27
2.1.2	<i>Intersymbol Interference</i>	31
2.2	Spectrum Limitations	32
2.2.1	<i>Assigned Frequencies</i>	32
2.2.2	<i>Frequency Reuse in Regulated Spectrum</i>	34
2.2.3	<i>Frequency Reuse in Unregulated Spectrum</i>	35
2.3	Limited Energy	35
2.4	User Mobility	36
3	Noise- and Interference-Limited Systems	37
3.1	Introduction	37
3.2	Noise-Limited Systems	37
3.2.1	<i>Link Budget</i>	40
3.3	Interference-Limited Systems	43
Part II	WIRELESS PROPAGATION CHANNELS	45
4	Propagation Mechanisms	47
4.1	Free Space Attenuation	47
4.2	Reflection and Transmission	49
4.2.1	<i>Snell's Law</i>	49
4.2.2	<i>Reflection and Transmission for Layered Dielectric Structures</i>	51
4.2.3	<i>The d^{-4} Power Law</i>	53
4.3	Diffraction	54
4.3.1	<i>Diffraction by a Single Screen or Wedge</i>	55
4.3.2	<i>Diffraction by Multiple Screens</i>	59
4.4	Scattering by Rough Surfaces	63
4.4.1	<i>The Kirchhoff Theory</i>	64
4.4.2	<i>Perturbation Theory</i>	65
4.5	Waveguiding	66
4.6	Appendices: please see companion website (www.wiley.com/go/molisch)	67
4.A:	<i>Derivation of the d^{-4} Law</i>	
4.B:	<i>Diffraction Coefficients for Diffraction by a Wedge or Cylinder</i>	
	Further Reading	67
5	Statistical Description of the Wireless Channel	69
5.1	Introduction	69
5.2	The Time-Invariant Two-Path Model	71
5.3	The Time-Variant Two-Path Model	72
5.4	Small-Scale Fading without a Dominant Component	74

5.4.1	<i>A Computer Experiment</i>	75
5.4.2	<i>Mathematical Derivation of the Statistics of Amplitude and Phase</i>	78
5.4.3	<i>Properties of the Rayleigh Distribution</i>	80
5.4.4	<i>Fading Margin for Rayleigh-Distributed Field Strength</i>	82
5.5	Small-Scale Fading with a Dominant Component	83
5.5.1	<i>A Computer Experiment</i>	83
5.5.2	<i>Derivation of the Amplitude and Phase Distribution</i>	83
5.5.3	<i>Nakagami Distribution</i>	87
5.6	Doppler Spectra and Temporal Channel Variations	88
5.6.1	<i>Temporal Variations for Moving MS</i>	88
5.6.2	<i>Temporal Variations in Fixed Wireless Systems</i>	90
5.7	Temporal Dependence of Fading	91
5.7.1	<i>Level Crossing Rate</i>	91
5.7.2	<i>Average Duration of Fades</i>	92
5.7.3	<i>Random Frequency Modulation</i>	94
5.8	Large-Scale Fading	95
5.9	Appendices: please see companion website (www.wiley.com/go/molisch)	99
5.A:	<i>The Lindeberg–Feller Theorem</i>	
5.B:	<i>Derivation of the Rayleigh Distribution</i>	
5.C:	<i>Derivation of the Level Crossing Rate</i>	
	Further Reading	99
6	Wideband and Directional Channel Characterization	101
6.1	Introduction	101
6.2	The Causes of Delay Dispersion	102
6.2.1	<i>The Two-Path Model</i>	102
6.2.2	<i>The General Case</i>	104
6.3	System-Theoretic Description of Wireless Channels	106
6.3.1	<i>Characterization of Deterministic Linear Time Variant Systems</i>	106
6.3.2	<i>Stochastic System Functions</i>	107
6.4	The WSSUS Model	109
6.4.1	<i>Wide-Sense Stationarity</i>	110
6.4.2	<i>Uncorrelated Scatterers</i>	110
6.4.3	<i>WSSUS Assumption</i>	111
6.4.4	<i>Tapped Delay Line Models</i>	111
6.5	Condensed Parameters	112
6.5.1	<i>Integrals of the Correlation Functions</i>	113
6.5.2	<i>Moments of the Power Delay Profile</i>	113
6.5.3	<i>Moments of the Doppler Spectra</i>	114
6.5.4	<i>Coherence Bandwidth and Coherence Time</i>	114
6.5.5	<i>Window Parameters</i>	116
6.6	Ultra Wideband Channels	118
6.6.1	<i>UWB Signals with Large Relative Bandwidth</i>	118
6.6.2	<i>UWB Channels with Large Absolute Bandwidth</i>	120
6.7	Directional Description	120
6.8	Appendices: please see companion website (www.wiley.com/go/molisch)	123
6.A:	<i>Validity of WSSUS in Mobile Radio Channels</i>	
6.B:	<i>Instantaneous Channel Parameters</i>	
	Further Reading	123

7	Channel Models	125
7.1	Introduction	125
7.2	Narrowband Models	126
7.2.1	<i>Modeling of Small-Scale and Large-Scale Fading</i>	126
7.2.2	<i>Path Loss Models</i>	127
7.3	Wideband Models	128
7.3.1	<i>Tapped Delay Line Models</i>	128
7.3.2	<i>Models for the Power Delay Profile</i>	129
7.3.3	<i>Models for the Arrival Times of Rays and Clusters</i>	130
7.3.4	<i>Standardized Channel Model</i>	131
7.4	Directional Models	131
7.4.1	<i>General Model Structure and Factorization</i>	131
7.4.2	<i>Angular Dispersion at the Base Station</i>	132
7.4.3	<i>Angular Dispersion at the Mobile Station</i>	133
7.4.4	<i>Polarization</i>	133
7.4.5	<i>Model Implementations</i>	134
7.4.6	<i>Standardized Directional Models</i>	137
7.4.7	<i>Multiple-Input Multiple-Output Matrix Models</i>	137
7.5	Deterministic Channel-Modeling Methods	138
7.5.1	<i>Ray Launching</i>	139
7.5.2	<i>Ray Tracing</i>	140
7.5.3	<i>Efficiency Considerations</i>	140
7.5.4	<i>Geographical Databases</i>	142
7.6	Appendices: please see companion website (www.wiley.com/go/molisch)	142
7.A:	<i>The Okumura–Hata Model</i>	
7.B:	<i>The COST 231–Walfish–Ikegami Model</i>	
7.C:	<i>The COST 207 GSM Model</i>	
7.D:	<i>The ITU-R Models</i>	
7.E:	<i>The 3GPP Spatial Channel Model</i>	
7.F:	<i>The ITU-Advanced Channel Model</i>	
7.G:	<i>The 802.15.4a UWB Channel Model</i>	
	Further Reading	142
8	Channel Sounding	145
8.1	Introduction	145
8.1.1	<i>Requirements for Channel Sounding</i>	145
8.1.2	<i>Generic Sounder Structure</i>	145
8.1.3	<i>Identifiability of Wireless Channels</i>	147
8.1.4	<i>Influence on Measurement Data</i>	149
8.2	Time-Domain Measurements	150
8.2.1	<i>Impulse Sounder</i>	150
8.2.2	<i>Correlative Sounders</i>	151
8.3	Frequency Domain Analysis	152
8.4	Modified Measurement Methods	153
8.4.1	<i>Swept Time Delay Cross Correlator (STDCC)</i>	153
8.4.2	<i>Inverse Filtering</i>	154
8.4.3	<i>Averaging</i>	154
8.4.4	<i>Synchronization</i>	155
8.4.5	<i>Vector Network Analyzer Measurements</i>	156

8.5	Directionally Resolved Measurements	157
8.5.1	<i>Data Model for Receive Arrays</i>	158
8.5.2	<i>Beamforming</i>	160
8.5.3	<i>High-Resolution Algorithms</i>	160
8.5.4	<i>Multiple Input Multiple Output Measurements</i>	162
8.6	Appendix: please see companion website (www.wiley.com/go/molisch)	164
8.A:	<i>The ESPRIT Algorithm</i>	
	Further Reading	164
9	Antennas	165
9.1	Introduction	165
9.1.1	<i>Integration of Antennas into Systems</i>	165
9.1.2	<i>Characteristic Antenna Quantities</i>	165
9.2	Antennas for Mobile Stations	169
9.2.1	<i>Monopole and Dipole Antennas</i>	169
9.2.2	<i>Helical Antennas</i>	170
9.2.3	<i>Microstrip Antennas</i>	171
9.2.4	<i>Planar Inverted F Antenna</i>	172
9.2.5	<i>Radiation Coupled Dual L Antenna</i>	173
9.2.6	<i>Multiband Antennas</i>	173
9.2.7	<i>Antenna Mounting on the Mobile Station</i>	174
9.3	Antennas for Base Stations	175
9.3.1	<i>Types of Antennas</i>	175
9.3.2	<i>Array Antennas</i>	175
9.3.3	<i>Modifying the Antenna Pattern</i>	176
9.3.4	<i>Impact of the Environment on Antenna Pattern</i>	176
	Further Reading	178
Part III	TRANSCIVERS AND SIGNAL PROCESSING	179
10	Structure of a Wireless Communication Link	181
10.1	Transceiver Block Structure	181
10.2	Simplified Models	186
	Further Reading	186
11	Modulation Formats	187
11.1	Introduction	187
11.2	Basics	188
11.2.1	<i>Pulse Amplitude Modulation</i>	188
11.2.2	<i>Multipulse Modulation and Continuous Phase Modulation</i>	192
11.2.3	<i>Power Spectrum</i>	193
11.2.4	<i>Signal Space Diagram</i>	194
11.3	Important Modulation Formats	196
11.3.1	<i>Binary Phase Shift Keying</i>	196
11.3.2	<i>Quadrature-Phase Shift Keying</i>	199
11.3.3	<i>$\pi/4$-Differential Quadrature-Phase Shift Keying</i>	201
11.3.4	<i>Offset Quadrature-Phase Shift Keying</i>	204
11.3.5	<i>Higher Order Modulation</i>	206

11.3.6	<i>Binary Frequency Shift Keying</i>	208
11.3.7	<i>Minimum Shift Keying</i>	212
11.3.8	<i>Demodulation of Minimum Shift Keying</i>	214
11.3.9	<i>Gaussian Minimum Shift Keying</i>	215
11.3.10	<i>Pulse Position Modulation</i>	215
11.3.11	<i>Summary of Spectral Efficiencies</i>	219
11.4	Appendix: please see companion website (www.wiley.com/go/molisch)	219
11.A:	<i>Interpretation of MSK as OQPSK</i>	
	Further Reading	219
12	Demodulation	221
12.1	Demodulator Structure and Error Probability in Additive White Gaussian Noise Channels	221
12.1.1	<i>Model for Channel and Noise</i>	221
12.1.2	<i>Signal Space Diagram and Optimum Receivers</i>	222
12.1.3	<i>Methods for the Computation of Error Probability</i>	225
12.2	Error Probability in Flat-Fading Channels	232
12.2.1	<i>Average BER – Classical Computation Method</i>	232
12.2.2	<i>Computation of Average Error Probability – Alternative Method</i>	234
12.2.3	<i>Outage Probability versus Average Error Probability</i>	238
12.3	Error Probability in Delay- and Frequency-Dispersive Fading Channels	239
12.3.1	<i>Physical Cause of Error Floors</i>	239
12.3.2	<i>Computation of the Error Floor Using the Group Delay Method</i>	242
12.3.3	<i>General Fading Channels: The Quadratic Form Gaussian Variable Method</i>	245
12.3.4	<i>Bit Error Probability</i>	247
	Further Reading	247
13	Diversity	249
13.1	Introduction	249
13.1.1	<i>Principle of Diversity</i>	249
13.1.2	<i>Definition of the Correlation Coefficient</i>	250
13.2	Microdiversity	251
13.2.1	<i>Spatial Diversity</i>	252
13.2.2	<i>Temporal Diversity</i>	254
13.2.3	<i>Frequency Diversity</i>	254
13.2.4	<i>Angle Diversity</i>	256
13.2.5	<i>Polarization Diversity</i>	258
13.3	Macrodiversity and Simulcast	258
13.4	Combination of Signals	259
13.4.1	<i>Selection Diversity</i>	260
13.4.2	<i>Switched Diversity</i>	262
13.4.3	<i>Combining Diversity</i>	263
13.5	Error Probability in Fading Channels with Diversity Reception	268
13.5.1	<i>Error Probability in Flat-Fading Channels</i>	268
13.5.2	<i>Symbol Error Rate in Frequency-Selective Fading Channels</i>	270
13.6	Transmit Diversity	273
13.6.1	<i>Transmitter Diversity with Channel State Information</i>	273

13.6.2	<i>Transmitter Diversity Without Channel State Information</i>	274
13.7	Appendix: please see companion website (www.wiley.com/go/molisch)	274
13.A:	<i>Correlation Coefficient of Two Signals with Time Separation and Frequency Separation</i>	
	Further Reading	275
14	Channel Coding and Information Theory	277
14.1	Fundamentals of Coding and Information Theory	277
14.1.1	<i>History and Motivation of Coding</i>	277
14.1.2	<i>Fundamental Concepts of Information Theory</i>	278
14.1.3	<i>Channel Capacity</i>	279
14.1.4	<i>Power–Bandwidth Relationship</i>	281
14.1.5	<i>Relationship to Practical Systems</i>	281
14.1.6	<i>Classification of Practical Codes</i>	282
14.2	Block Codes	283
14.2.1	<i>Introduction</i>	283
14.2.2	<i>Encoding</i>	284
14.2.3	<i>Decoding</i>	285
14.2.4	<i>Recognition and Correction of Errors</i>	287
14.2.5	<i>Concatenated Codes</i>	288
14.3	Convolutional Codes	288
14.3.1	<i>Principle of Convolutional Codes</i>	288
14.3.2	<i>Viterbi Decoder – Classical Representation</i>	290
14.3.3	<i>Improvements of the Viterbi Algorithm</i>	293
14.4	Trellis Coded Modulation	294
14.4.1	<i>Basic Principle</i>	294
14.4.2	<i>Set Partitioning</i>	297
14.5	Bit Interleaved Coded Modulation (BICM)	299
14.6	Turbo Codes	300
14.6.1	<i>Introduction</i>	300
14.6.2	<i>Encoder</i>	300
14.6.3	<i>Turbo Decoder</i>	301
14.7	Low Density Parity Check Codes	303
14.7.1	<i>Definition of Low Density Parity Check Codes</i>	304
14.7.2	<i>Encoding of Low Density Parity Check Codes</i>	305
14.7.3	<i>Decoding of Low Density Parity Check Codes</i>	305
14.7.4	<i>Performance Improvements</i>	309
14.8	Coding for the Fading Channel	309
14.8.1	<i>Interleaving</i>	309
14.8.2	<i>Block Codes</i>	312
14.8.3	<i>Convolutional Codes</i>	313
14.8.4	<i>Concatenated Codes</i>	313
14.8.5	<i>Trellis Coded Modulation in Fading Channels</i>	313
14.9	Information-Theoretic Performance Limits of Fading Channels	315
14.9.1	<i>Ergodic Capacity vs. Outage Capacity</i>	315
14.9.2	<i>Capacity for Channel State Information at the Receiver (CSIR) Only</i>	315
14.9.3	<i>Capacity for CSIT and CSIR – Waterfilling</i>	316
14.10	Appendices: please see companion website (www.wiley.com/go/molisch)	317

14.A:	<i>ARQ and HARQ</i>	
	Further Reading	317
15	Speech Coding	319
	Gernot Kubin	
15.1	Introduction	319
15.1.1	<i>Speech Telephony as Conversational Multimedia Service</i>	319
15.1.2	<i>Source-Coding Basics</i>	320
15.1.3	<i>Speech Coder Designs</i>	320
15.2	The Sound of Speech	322
15.2.1	<i>Speech Production</i>	322
15.2.2	<i>Speech Acoustics</i>	323
15.2.3	<i>Speech Perception</i>	325
15.3	Stochastic Models for Speech	326
15.3.1	<i>Short-Time Stationary Modeling</i>	326
15.3.2	<i>Linear Predictive voCoder (LPC)</i>	327
15.3.3	<i>Sinusoidal Modeling</i>	329
15.3.4	<i>Harmonic + Noise Modeling</i>	330
15.3.5	<i>Cyclostationary Modeling</i>	330
15.4	Quantization and Coding	331
15.4.1	<i>Scalar Quantization</i>	331
15.4.2	<i>Vector Quantization</i>	332
15.4.3	<i>Noise Shaping in Predictive Coding</i>	335
15.4.4	<i>Analysis by Synthesis</i>	336
15.4.5	<i>Joint Source Channel Coding</i>	338
15.5	From Speech Transmission to Acoustic Telepresence	339
15.5.1	<i>Voice Activity Detection</i>	339
15.5.2	<i>Receiver End Enhancements</i>	340
15.5.3	<i>Acoustic Echo and Noise</i>	340
15.5.4	<i>Service Augmentation for Telepresence</i>	341
	Further Reading	342
16	Equalizers	343
16.1	Introduction	343
16.1.1	<i>Equalization in the Time Domain and Frequency Domain</i>	343
16.1.2	<i>Modeling of Channel and Equalizer</i>	344
16.1.3	<i>Channel Estimation</i>	345
16.2	Linear Equalizers	347
16.2.1	<i>Zero-Forcing Equalizer</i>	348
16.2.2	<i>The Mean Square Error Criterion</i>	349
16.2.3	<i>Adaptation Algorithms for Mean Square Error Equalizers</i>	350
16.2.4	<i>Further Linear Structures</i>	352
16.3	Decision Feedback Equalizers	353
16.3.1	<i>MMSE Decision Feedback Equalizer</i>	354
16.3.2	<i>Zero-Forcing Decision Feedback Equalizer</i>	355
16.4	Maximum Likelihood Sequence Estimation – Viterbi Detector	355
16.5	Comparison of Equalizer Structures	358
16.6	Fractionally Spaced Equalizers	358
16.7	Blind Equalizers	359
16.7.1	<i>Introduction</i>	359

16.7.2	<i>Constant Modulus Algorithm</i>	359
16.7.3	<i>Blind Maximum Likelihood Estimation</i>	360
16.7.4	<i>Algorithms Using Second- or Higher Order Statistics</i>	360
16.7.5	<i>Assessment</i>	361
16.8	Appendices: please see companion website (www.wiley.com/go/molisch)	361
16.A:	<i>Equivalence of Peak Distortion and Zero-Forcing Criterion</i>	
16.B:	<i>Derivation of the Mean-Square-Error Criterion</i>	
16.C:	<i>The Recursive-Least-Squares Algorithm</i>	
	Further Reading	361
Part IV	MULTIPLE ACCESS AND ADVANCED TRANSCEIVER SCHEMES	363
17	Multiple Access and the Cellular Principle	365
17.1	Introduction	365
17.2	Frequency Division Multiple Access	366
17.2.1	<i>Multiple Access via Frequency Division Multiple Access</i>	366
17.2.2	<i>Trunking Gain</i>	367
17.3	Time Division Multiple Access	371
17.4	Packet Radio	373
17.4.1	<i>ALOHA</i>	373
17.4.2	<i>Carrier Sense Multiple Access</i>	375
17.4.3	<i>Packet Reservation Multiple Access</i>	376
17.4.4	<i>Comparison of the Methods</i>	376
17.4.5	<i>Routing for Packet Radio</i>	377
17.5	Duplexing	378
17.6	Principles of Cellular Networks	379
17.6.1	<i>Reuse Distance</i>	379
17.6.2	<i>Cellshape</i>	380
17.6.3	<i>Cell Planning with Hexagonal Cells</i>	380
17.6.4	<i>Methods for Increasing Capacity</i>	383
17.7	Appendix: please see companion website (www.wiley.com/go/molisch)	385
17.A:	<i>Adjacent Channel Interface</i>	
17.B:	<i>Information Theory of Multi-User Channels</i>	
	Further Reading	385
18	Spread Spectrum Systems	387
18.1	Frequency Hopping Multiple Access (FHMA)	387
18.1.1	<i>Principle Behind Frequency Hopping</i>	387
18.1.2	<i>Frequency Hopping for Multiple Access (FHMA)</i>	388
18.2	Code Division Multiple Access	389
18.2.1	<i>Basic Principle Behind the Direct Sequence-Spread Spectrum</i>	390
18.2.2	<i>Multiple Access</i>	392
18.2.3	<i>Mathematical Representation</i>	393
18.2.4	<i>Effects of Multipath Propagation on Code Division Multiple Access</i>	394
18.2.5	<i>Synchronization</i>	397
18.2.6	<i>Code Families</i>	398
18.3	Cellular Code-Division-Multiple-Access Systems	401
18.3.1	<i>Principle Behind Code Division Multiple Access – Revisited</i>	401
18.3.2	<i>Power Control</i>	403

18.3.3	<i>Methods for Capacity Increases</i>	405
18.3.4	<i>Combination with Other Multiaccess Methods</i>	406
18.4	Multuser Detection	406
18.4.1	<i>Introduction</i>	406
18.4.2	<i>Linear Multuser Detectors</i>	407
18.4.3	<i>Nonlinear Multuser Detectors</i>	408
18.5	Time Hopping Impulse Radio	411
18.5.1	<i>Simple Impulse Radio</i>	411
18.5.2	<i>Time Hopping</i>	412
18.5.3	<i>Impulse Radio in Delay-Dispersive Channels</i>	414
	Further Reading	415
19	Orthogonal Frequency Division Multiplexing (OFDM)	417
19.1	Introduction	417
19.2	Principle of Orthogonal Frequency Division Multiplexing	418
19.3	Implementation of Transceivers	418
19.4	Frequency-Selective Channels	420
19.4.1	<i>Cyclic Prefix</i>	420
19.4.2	<i>Performance in Frequency-Selective Channels</i>	422
19.4.3	<i>Coded Orthogonal Frequency Division Multiplexing</i>	425
19.5	Channel Estimation	425
19.5.1	<i>Pilot-Symbol-Based Methods</i>	426
19.5.2	<i>Methods Based on Scattered Pilots</i>	426
19.5.3	<i>Methods Based in Eigen Decompositions</i>	428
19.6	Peak-to-Average Power Ratio	429
19.6.1	<i>Origin of the Peak-to-Average Ratio Problem</i>	429
19.6.2	<i>Peak-to-Average Ratio Reduction Techniques</i>	431
19.7	Inter Carrier Interference	432
19.8	Adaptive Modulation and Capacity	436
19.8.1	<i>Channel Quality Estimation</i>	436
19.8.2	<i>Parameter Adaptation</i>	436
19.8.3	<i>Signaling of Chosen Parameters</i>	438
19.9	Multiple Access – OFDMA	439
19.10	Multicarrier Code Division Multiple Access	440
19.11	Single-Carrier Modulation with Frequency Domain Equalization	442
	Further Reading	443
20	Multiantenna Systems	445
20.1	Smart Antennas	445
20.1.1	<i>What are Smart Antennas?</i>	445
20.1.2	<i>Purpose</i>	446
20.1.3	<i>Capacity Increase</i>	446
20.1.4	<i>Receiver Structures</i>	449
20.1.5	<i>Algorithms for Adaptation of Antenna Weights</i>	453
20.1.6	<i>Uplink versus Downlink</i>	458
20.1.7	<i>Algorithms for the Adaptation of the Antenna Weights in the Downlink</i>	461
20.1.8	<i>Network Aspects</i>	462
20.1.9	<i>Multuser Diversity and Random Beamforming</i>	462
20.2	Multiple Input Multiple Output Systems	464
20.2.1	<i>Introduction</i>	464

20.2.2	<i>How Does Spatial Multiplexing Work?</i>	465
20.2.3	<i>System Model</i>	466
20.2.4	<i>Channel State Information</i>	467
20.2.5	<i>Capacity in Nonfading Channels</i>	468
20.2.6	<i>Capacity in Flat-Fading Channels</i>	470
20.2.7	<i>Impact of the Channel</i>	473
20.2.8	<i>Layered Space–Time Structure</i>	478
20.2.9	<i>Diversity</i>	480
20.2.10	<i>Tradeoffs between Diversity, Beamforming Gain, and Spatial Multiplexing</i>	484
20.2.11	<i>Feedback for MIMO</i>	484
20.3	Multuser MIMO	488
20.3.1	<i>Performance Limits</i>	489
20.3.2	<i>Scheduling</i>	490
20.3.3	<i>Linear Precoding – Uplink</i>	491
20.3.4	<i>Linear Precoding – Downlink</i>	492
20.3.5	<i>Closed-Loop Systems and Quantized Feedback</i>	496
20.3.6	<i>Base Station Cooperation</i>	497
	<i>Further Reading</i>	497
Part V	STANDARDIZED WIRELESS SYSTEMS	499
21	Cognitive Radio	501
21.1	<i>Problem Description</i>	501
21.2	<i>Cognitive Transceiver Architecture</i>	504
21.3	<i>Principles of Interweaving</i>	505
21.4	<i>Spectrum Sensing</i>	505
21.4.1	<i>Spectrum Sensing in a Hierarchical System</i>	506
21.4.2	<i>Types of Detectors</i>	507
21.4.3	<i>Multinode Detection</i>	509
21.4.4	<i>Cognitive Pilots</i>	510
21.5	<i>Spectrum Management</i>	510
21.5.1	<i>Spectrum Opportunity Tracking</i>	510
21.6	<i>Spectrum Sharing</i>	511
21.6.1	<i>Introduction</i>	511
21.6.2	<i>Non-Cooperative Games</i>	511
21.6.3	<i>Games with Partial Coordination</i>	511
21.6.4	<i>Centralized Solutions</i>	513
21.7	<i>Overlay</i>	514
21.8	<i>Underlay Hierarchical Access – Ultra Wide Bandwidth System Communications</i>	516
	<i>Further Reading</i>	520
22	Relaying, Multi-Hop, and Cooperative Communications	521
22.1	<i>Introduction and Motivation</i>	521
22.1.1	<i>Principle of Relaying</i>	521
22.2	<i>Fundamentals of Relaying</i>	523
22.2.1	<i>Fundamental Protocols</i>	523
22.2.2	<i>Decode-and-Forward</i>	525
22.2.3	<i>Amplify-and-Forward</i>	527

22.2.4	<i>Compress-and-Forward</i>	528
22.3	Relaying with Multiple, Parallel Relays	529
22.3.1	<i>Relay Selection</i>	530
22.3.2	<i>Distributed Beamforming</i>	531
22.3.3	<i>Transmission on Orthogonal Channels</i>	532
22.3.4	<i>Distributed Space–Time Coding</i>	533
22.3.5	<i>Coded Cooperation</i>	534
22.3.6	<i>Fountain Codes</i>	535
22.4	Routing and Resource Allocation in Multi-Hop Networks	537
22.4.1	<i>Mathematical Preliminaries</i>	538
22.4.2	<i>Goals and Classifications of Routing Protocols</i>	539
22.4.3	<i>Source Routing</i>	539
22.4.4	<i>Link-State Based Routing</i>	541
22.4.5	<i>Distance Vector Routing</i>	542
22.4.6	<i>Geography-Based Routing</i>	543
22.4.7	<i>Hierarchical Routing</i>	544
22.4.8	<i>Impact of Node Mobility</i>	544
22.4.9	<i>Data-Driven Routing</i>	545
22.4.10	<i>Power Allocation Strategies</i>	546
22.4.11	<i>Routing for Multiple Messages – Stochastic Network Optimization</i>	547
22.4.12	<i>Scaling Laws</i>	550
22.5	Routing and Resource Allocation in Collaborative Networks	551
22.5.1	<i>Edge-Disjoint Routing and Anypath Routing</i>	551
22.5.2	<i>Routing with Energy Accumulation</i>	552
22.5.3	<i>Fountain Codes</i>	554
22.5.4	<i>Other Collaborative Routing Problems</i>	554
22.5.5	<i>Scaling Laws</i>	556
22.6	Applications	556
22.6.1	<i>Dedicated Relays</i>	556
22.6.2	<i>Relaying and User Cooperation in Ad hoc Networks</i>	558
22.7	Network Coding	558
22.7.1	<i>Two-Way Relaying</i>	558
22.7.2	<i>Basics of Network Coding</i>	559
22.7.3	<i>Applications to Wireless Systems</i>	561
22.7.4	<i>Interference Alignment</i>	561
	Further Reading	562
23	Video Coding	565
	Anthony Vetro	
23.1	Introduction	565
23.1.1	<i>Digital Video Representation and Formats</i>	565
23.1.2	<i>Video Coding Architecture</i>	566
23.2	Transform and Quantization	568
23.2.1	<i>Discrete Cosine Transform</i>	568
23.2.2	<i>Scalar Quantization</i>	569
23.3	Prediction	571
23.3.1	<i>Intraframe Prediction</i>	571
23.3.2	<i>Interframe Prediction</i>	572
23.4	Entropy Coding	573
23.4.1	<i>Huffman Coding</i>	573

23.4.2	<i>Arithmetic Coding</i>	574
23.5	Video Coding Standards	576
23.6	Layered Video Coding	577
23.6.1	<i>Scalable Video Coding</i>	577
23.6.2	<i>Multiview Video Coding</i>	579
23.7	Error Control	579
23.7.1	<i>Transport Layer Mechanisms</i>	580
23.7.2	<i>Error-Resilient Encoding of Video</i>	580
23.7.3	<i>Error Concealment at the Decoder</i>	583
23.8	Video Streaming	583
	Further Reading	585
24	GSM – Global System for Mobile Communications	587
24.1	Historical Overview	587
24.2	System Overview	589
24.2.1	<i>Base Station Subsystem</i>	589
24.2.2	<i>Network and Switching Subsystem</i>	590
24.2.3	<i>Operating Support System</i>	590
24.3	The Air Interface	591
24.4	Logical and Physical Channels	595
24.4.1	<i>Logical Channels</i>	595
24.4.2	<i>Mapping Between Logical and Physical Channels</i>	597
24.5	Synchronization	600
24.5.1	<i>Frequency Synchronization</i>	600
24.5.2	<i>Time Synchronization</i>	600
24.5.3	<i>Timing Advance</i>	601
24.5.4	<i>Summary of Burst Structures</i>	601
24.6	Coding	602
24.6.1	<i>Voice Encoding</i>	602
24.6.2	<i>Channel Encoding</i>	603
24.6.3	<i>Cryptography</i>	605
24.6.4	<i>Frequency Hopping</i>	606
24.7	Equalizer	606
24.8	Circuit-Switched Data Transmission	607
24.9	Establishing a Connection and Handover	608
24.9.1	<i>Identity Numbers</i>	609
24.9.2	<i>Identification of a Mobile Subscriber</i>	609
24.9.3	<i>Examples for Establishment of a Connection</i>	610
24.9.4	<i>Examples of Different Kinds of Handovers</i>	611
24.10	Services and Billing	614
24.10.1	<i>Available Services</i>	614
24.10.2	<i>Billing</i>	616
24.11	Glossary for GSM	617
24.12	Appendices: please see companion website (www.wiley.com/go/molisch)	620
24.A:	<i>The Original Global System for Mobile Communications Speech Coder</i>	
24.B:	<i>General Packet Radio Service</i>	
	Further Reading	620
25	IS-95 and CDMA 2000	621
25.1	Historical Overview	621

25.2	System Overview	621
25.3	Air Interface	622
25.3.1	<i>Frequency Bands and Duplexing</i>	622
25.3.2	<i>Spreading and Modulation</i>	622
25.3.3	<i>Power Control</i>	623
25.3.4	<i>Pilot Signal</i>	623
25.4	Coding	623
25.4.1	<i>Speech Coders</i>	624
25.4.2	<i>Error Correction Coding</i>	624
25.5	Spreading and Modulation	625
25.5.1	<i>Long and Short Spreading Codes and Walsh Codes</i>	625
25.5.2	<i>Spreading and Modulation in the Uplink</i>	626
25.5.3	<i>Databurst Randomization and Gating for the Uplink</i>	627
25.5.4	<i>Spreading and Modulation in the Downlink</i>	629
25.5.5	<i>Discussion</i>	629
25.6	Logical and Physical Channels	631
25.6.1	<i>Traffic Channels</i>	631
25.6.2	<i>Access Channel</i>	631
25.6.3	<i>Pilot Channels</i>	632
25.6.4	<i>Synchronization Channel</i>	632
25.6.5	<i>Paging Channel</i>	632
25.6.6	<i>Power Control Subchannel</i>	632
25.6.7	<i>Mapping Logical Channels to Physical Channels</i>	633
25.7	Handover	633
25.8	Appendices: please see companion website (www.wiley.com/go/molisch)	633
25.A:	<i>CDMA 2000 –History</i>	
25.B:	<i>CDMA 2000 –1x Mode</i>	
25.C:	<i>CDMA 2000 –3x Mode</i>	
25.D:	<i>CDMA 2000 –1xEV-DO</i>	
	Further Reading	634
26	WCDMA/UMTS	635
26.1	Historical Overview	635
26.2	System Overview	636
26.2.1	<i>Physical-Layer Overview</i>	636
26.2.2	<i>Network Structure</i>	637
26.2.3	<i>Hierarchical Cellular Structure</i>	637
26.2.4	<i>Data Rates and Service Classes</i>	638
26.3	Air Interface	639
26.3.1	<i>Frequency Bands and Duplexing</i>	639
26.3.2	<i>Time Domain Duplexing and Frequency Domain Duplexing Modes</i>	640
26.3.3	<i>Radio-Frequency-Related Aspects</i>	640
26.4	Physical and Logical Channels	641
26.4.1	<i>Logical Channels</i>	641
26.4.2	<i>Physical Channels</i>	643
26.5	Speech Coding, Multiplexing, and Channel Coding	645
26.5.1	<i>Speech Coder</i>	645
26.5.2	<i>Multiplexing and Interleaving</i>	646
26.6	Spreading and Modulation	649

26.6.1	<i>Frame Structure, Spreading Codes, and Walsh–Hadamard Codes</i>	649
26.6.2	<i>Uplink</i>	650
26.6.3	<i>Downlink</i>	655
26.7	Physical-Layer Procedures	657
26.7.1	<i>Cell Search and Synchronization</i>	657
26.7.2	<i>Establishing a Connection</i>	658
26.7.3	<i>Power Control</i>	659
26.7.4	<i>Handover</i>	660
26.7.5	<i>Overload Control</i>	661
26.8	Glossary for WCDMA	661
	Further Reading	663
27	3GPP Long-Term Evolution	665
27.1	Introduction	665
27.1.1	<i>History</i>	665
27.1.2	<i>Goals</i>	666
27.2	System Overview	667
27.2.1	<i>Frequency Bands and Spectrum Flexibility</i>	667
27.2.2	<i>Network Structure</i>	668
27.2.3	<i>Protocol Structure</i>	669
27.2.4	<i>PHY and MAC Layer Overview</i>	669
27.3	Physical Layer	672
27.3.1	<i>Frames, Slots, and Symbols</i>	672
27.3.2	<i>Modulation</i>	674
27.3.3	<i>Mapping to Physical Resources – Downlink</i>	674
27.3.4	<i>Mapping to Physical Resources – Uplink</i>	676
27.3.5	<i>Pilots or Reference Signals</i>	676
27.3.6	<i>Coding</i>	680
27.3.7	<i>Multiple-Antenna Techniques</i>	683
27.4	Logical and Physical Channels	684
27.4.1	<i>Mapping of Data onto (Logical) Subchannels</i>	684
27.4.2	<i>Synchronization Signals</i>	686
27.4.3	<i>Broadcast Channel</i>	687
27.4.4	<i>General Aspects of Control Channels Associated with a DL-SCH</i>	687
27.4.5	<i>Physical Control Format Indicator CHannel</i>	688
27.4.6	<i>Physical HARQ Indicator CHannel</i>	688
27.4.7	<i>Physical Downlink Control CHannel</i>	688
27.4.8	<i>Physical Random Access CHannel</i>	690
27.4.9	<i>General Aspects of Control Signals Associated with PUSCH</i>	691
27.4.10	<i>PUCCH</i>	692
27.4.11	<i>PUSCH</i>	693
27.5	Physical Layer Procedures	693
27.5.1	<i>Establishing a Connection</i>	693
27.5.2	<i>Retransmissions and Reliability</i>	694
27.5.3	<i>Scheduling</i>	696
27.5.4	<i>Power Control</i>	696
27.5.5	<i>Handover</i>	696
27.6	Glossary for LTE	697
	Further Reading	698

28	WiMAX/IEEE 802.16	699
28.1	Introduction	699
28.1.1	<i>History</i>	699
28.1.2	<i>WiMAX versus Existing Cellular Systems</i>	700
28.2	System Overview	701
28.2.1	<i>Physical Layer Overview</i>	701
28.2.2	<i>Frequency Bands</i>	701
28.2.3	<i>MAC Layer Overview</i>	702
28.2.4	<i>Network Structure</i>	702
28.3	Modulation and Coding	703
28.3.1	<i>Modulation</i>	703
28.3.2	<i>Coding</i>	705
28.4	Logical and Physical Channels	707
28.4.1	<i>Frames and Zones</i>	707
28.4.2	<i>Details of Frame Structure</i>	707
28.4.3	<i>Mapping of Data onto (Logical) Subchannels</i>	709
28.4.4	<i>Principles of Preamble and Pilots</i>	710
28.4.5	<i>PUSC</i>	711
28.4.6	<i>TUSC</i>	716
28.4.7	<i>FUSC</i>	717
28.4.8	<i>AMC</i>	718
28.4.9	<i>Channel Sounding</i>	719
28.5	Multiple-Antenna Techniques	720
28.5.1	<i>Space–Time Coding and Spatial Multiplexing</i>	720
28.5.2	<i>MIMO Precoding</i>	723
28.5.3	<i>SDMA and Soft Handover in MIMO</i>	724
28.6	Link Control	724
28.6.1	<i>Establishing a Connection</i>	724
28.6.2	<i>Scheduling and Resource Request</i>	725
28.6.3	<i>QoS</i>	726
28.6.4	<i>Power Control</i>	726
28.6.5	<i>Handover and Mobility Support</i>	727
28.6.6	<i>Power Saving</i>	728
28.7	Glossary for WiMAX	728
	Further Reading	729
29	Wireless Local Area Networks	731
29.1	Introduction	731
29.1.1	<i>History</i>	731
29.1.2	<i>Applications</i>	733
29.1.3	<i>Relationship between the Medium Access Control Layer and the PHY</i>	733
29.2	802.11a/g – Orthogonal Frequency Division Multiplexing-Based Local Area Networks	734
29.2.1	<i>Frequency Bands</i>	735
29.2.2	<i>Modulation and Coding</i>	735
29.2.3	<i>Headers</i>	737
29.2.4	<i>Synchronization and Channel Estimation</i>	738
29.3	IEEE 802.11n	739
29.3.1	<i>Overview</i>	739
29.3.2	<i>Modulation and Coding</i>	740

29.3.3	<i>Multiple-Antenna Techniques</i>	741
29.3.4	<i>20-MHz and 40-MHz Channels</i>	742
29.3.5	<i>Headers and Preambles</i>	743
29.3.6	<i>Channel Estimation</i>	744
29.4	Packet Transmission in 802.11 Wireless Local Area Networks	745
29.4.1	<i>General Medium Access Control Structure</i>	745
29.4.2	<i>Frame Formats</i>	746
29.4.3	<i>Packet Radio Multiple Access</i>	747
29.5	Alternative Wireless Local Area Networks and Future Developments	749
29.6	Glossary for WLAN	749
	Further Reading	750
30	Exercises	751
	Peter Almers, Ove Edfors, Hao Feng, Fredrik Floren, Anders Johanson, Johan Karedal, Buon Kiong Lau, Christian Mehlführer, Andreas F. Molisch, Jan Plasberg, Barbara Resch, Jonas Samuelson, Junyang Shen, Andre Stranne, Fredrik Tufvesson, Anthony Vetro and Shurjeel Wyne	
30.1	Chapter 1: Applications and Requirements of Wireless Services	751
30.2	Chapter 2: Technical Challenges of Wireless Communications	751
30.3	Chapter 3: Noise- and Interference-Limited Systems	752
30.4	Chapter 4: Propagation Mechanisms	752
30.5	Chapter 5: Statistical Description of the Wireless Channel	754
30.6	Chapter 6: Wideband and Directional Channel Characterization	757
30.7	Chapter 7: Channel Models	758
30.8	Chapter 8: Channel Sounding	759
30.9	Chapter 9: Antennas	761
30.10	Chapter 10: Structure of a Wireless Communication Link	762
30.11	Chapter 11: Modulation Formats	762
30.12	Chapter 12: Demodulation	763
30.13	Chapter 13: Diversity	765
30.14	Chapter 14: Channel Coding	768
30.15	Chapter 15: Speech Coding	770
30.16	Chapter 16: Equalizers	773
30.17	Chapter 17: Multiple Access and the Cellular Principle	775
30.18	Chapter 18: Spread Spectrum Systems	777
30.19	Chapter 19: Orthogonal Frequency Division Multiplexing (OFDM)	779
30.20	Chapter 20: Multiantenna Systems	780
30.21	Chapter 21: Cognitive Radio	782
30.22	Chapter 22: Relaying, Multi-Hop, and Cooperative Communications	784
30.23	Chapter 23: Video Coding	786
30.24	Chapter 24: GSM – Global System for Mobile Communications	787
30.25	Chapter 25: IS-95 and CDMA 2000	788
30.26	Chapter 26: WCDMA/UMTS	788
30.27	Chapter 27: 3GPP Long Term Evolution	788
30.28	Chapter 28: WiMAX/IEEE 802.16	790
30.29	Chapter 29: Wireless Local Area Networks	790
	References	793
	Index	817

2

Technical Challenges of Wireless Communications

In the previous chapter, we described the requirements for wireless communications systems stemming from the applications and user demands. In this chapter, we give a high-level description of the technical challenges to wireless communications systems. Most notably, they are as follows:

- multipath propagation: i.e., the fact that a transmit signal can reach the receiver via different paths (e.g., reflections from different houses or mountains);
- spectrum limitations;
- energy limitations;
- user mobility.

This sets the stage for the rest of the book, where these challenges, as well as remedies, are discussed in more detail.

As a first step, it is useful to investigate the differences between wired and wireless communications. Let us first repeat some important properties of wired and wireless systems, as summarized in Table 2.1.

2.1 Multipath Propagation

For wireless communications, the transmission medium is the radio channel between transmitter TX and receiver RX. The signal can get from the TX to the RX via a number of different propagation paths. In some cases, a Line Of Sight (LOS) connection might exist between TX and RX. Furthermore, the signal can get from the TX to the RX by being reflected at or diffracted by different *Interacting Objects* (IOs) in the environment: houses, mountains (for outdoor environments), windows, walls, etc. The number of these possible propagation paths is very large. As shown in Figure 2.1, each of the paths has a distinct amplitude, delay (runtime of the signal), direction of departure from the TX, and direction of arrival; most importantly, the components have different *phase shifts* with respect to each other. In the following, we discuss some implications of the multipath propagation for system design.

2.1.1 Fading

A simple RX cannot distinguish between the different *Multi Path Components* (MPCs); it just adds them up, so that they interfere with each other. The interference between them can be constructive

3

Noise- and Interference-Limited Systems

3.1 Introduction

This chapter explains the principles of link budgets, and the planning of wireless systems with one or multiple users. In Section 3.2, we set up link budgets for noise-limited systems and compute the minimum transmit power (or maximum range) that can be achieved in the absence of interference. Such computations give a first insight into the basic capabilities of wireless systems and also have practical applications. For example, Wireless Local Area Networks (WLANs) and cordless phones often operate in a noise-limited mode, if no other Base Station (BS) is in the vicinity. Even cellular systems sometimes operate in that mode if the user density is low (this happens, e.g., during the build-up phase of a network).

In Section 3.3, we discuss interference-limited systems. As we described in the first two chapters, the unregulated use of spectrum leads to interference that cannot be controlled by the user. When the spectrum is regulated, the network operator can determine the location of BSs, and thus impact the Signal-to-Interference Ratio (SIR). For either case, it is important to set up the link budgets that take the presence of interference into account; Section 3.3 describes these link budgets. In Chapter 17, we then see how these calculations are related to the cellular principle and the reuse of frequencies in different cells.

3.2 Noise-Limited Systems

Wireless systems are required to provide a certain minimum transmission quality (see Section 1.3). This transmission quality in turn requires a minimum *Signal-to-Noise Ratio* (SNR) at the receiver (RX). Consider now a situation where only a single BS transmits, and a Mobile Station (MS) receives; thus, the performance of the system is determined only by the strength of the (useful) signal and the noise. As the MS moves further away from the BS, the received signal power decreases, and at a certain distance, the SNR does not achieve the required threshold for reliable communications. Therefore, the range of the system is noise limited; equivalently, we can call it *signal power limited*. Depending on the interpretation, it is too much noise or too little signal power that leads to bad link quality.

Let us assume for the moment that the received power decreases with d^2 , the square of the distance between BS and MS. More precisely, let the received power P_{RX} be

$$P_{RX} = P_{TX} G_{RX} G_{TX} \left(\frac{\lambda}{4\pi d} \right)^2 \quad (3.1)$$

where G_{RX} and G_{TX} are the gains of the receive and transmit antennas, respectively,¹ λ is the wavelength, and P_{TX} is the transmit power (see Chapter 4 for a derivation of this equation and for more details).

The noise that disturbs the signal can consist of several components, as follows:

1. *Thermal noise*: The power spectral density of thermal noise depends on the environmental temperature T_e that the antenna “sees.” The temperature of the Earth is around 300 K, while the temperature of the (cold) sky is approximately $T_e \approx 4$ K (the temperature in the direction of the Sun is of course much higher). As a first approximation, it is usually assumed that the environmental temperature is isotropically 300 K. Noise power spectral density is then

$$N_0 = k_B T_e \quad (3.2)$$

where k_B is Boltzmann’s constant, $k_B = 1.38 \cdot 10^{-23}$ J/K, and the noise power is

$$P_n = N_0 B \quad (3.3)$$

where B is RX bandwidth (in units of Hz). It is common to write Eq. (3.2) using logarithmic units (power P expressed in units of dBm is $10 \log_{10} (P/1 \text{ mW})$):

$$N_0 = -174 \text{ dBm/Hz} \quad (3.4)$$

This means that the noise power contained in a 1-Hz bandwidth is -174 dBm. The noise power contained in bandwidth B is

$$-174 + 10 \log_{10}(B) \text{ dBm} \quad (3.5)$$

The logarithm of bandwidth B , specifically $10 \log_{10}(B)$, has the units dBHz.

2. *Man-made noise*: We can distinguish two types of man-made noise:

- (a) *Spurious emissions*: Many electrical appliances as well as radio transmitters (TXs) designed for other frequency bands have spurious emissions over a large bandwidth that includes the frequency range in which wireless communications systems operate. For urban outdoor environments, car ignitions and other impulse sources are especially significant sources of noise. In contrast to thermal noise, the noise created by impulse sources decreases with frequency (see Figure 3.1). At 150 MHz, it can be 20 dB stronger than thermal noise; at 900 MHz, it is typically 10 dB stronger. At Universal Mobile Telecommunications System (UMTS) frequencies, Neubauer et al. [2001] measured 5-dB noise enhancement by man-made noise in urban environments and about 1 dB in rural environments. Note that frequency regulators in most countries impose limits on “spurious” or “out-of-band” emissions for *all* electrical devices. Furthermore, for communications operating in licensed bands, such spurious emissions are the only source of man-made noise. It lies in the nature of the license (for which the license holder usually has paid) that no other intentional emitters are allowed to operate in this band. In contrast to thermal noise, man-made noise is not necessarily Gaussian distributed. However, as a matter of convenience, most system-planning tools, as well as theoretical designs, assume *Gaussianity* anyway.

¹ Roughly speaking, “receive antenna gain” is a measure of how much more power we can receive (from a certain direction) by using a specific antenna, compared with the use of an isotropic antenna; the definition for transmit antennas is similar. See Chapter 9 and/or Stutzman and Thiele [1997] for details.

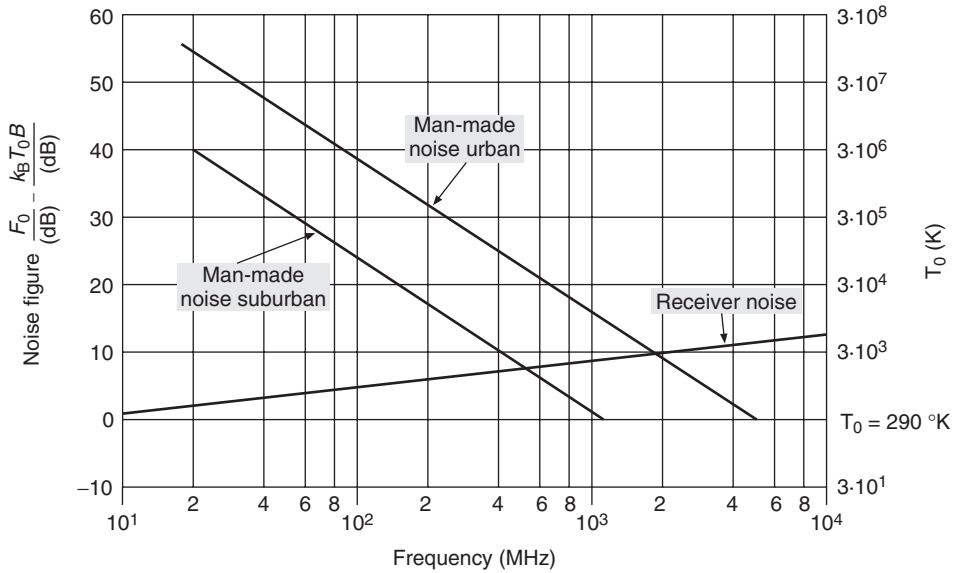


Figure 3.1 Noise as a function of frequency.

Reproduced with permission from Jakes [1974] © IEEE.

- (b) *Other intentional emission sources*: Several wireless communications systems operate in unlicensed bands. In these bands, everybody is allowed to operate (emit electromagnetic radiation) as long as certain restrictions with respect to transmit power, etc. are fulfilled. The most important of these bands is the 2.45-GHz Industrial, Scientific, and Medical (ISM) band. The amount of interference in these bands can be considerable.
3. *Receiver noise*: The amplifiers and mixers in the RX are noisy, and thus increase the total noise power. This effect is described by the noise figure F , which is defined as the SNR at the RX input (typically after downconversion to baseband) divided by the SNR at the RX output. As the amplifiers have gain, noise added in the later stages does not have as much of an impact as noise added in the first stage of the RX. Mathematically, the total noise figure F_{eq} of a cascade of components is

$$F_{eq} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \dots \quad (3.6)$$

where F_i and G_i are noise figures and noise gains of the individual stages in absolute units (not in decibels (dB)). Note that for this equation, passive components, like attenuators with gain $m < 1$, can be interpreted as *either* having a noise figure of $F = 1/m$ and unit gain of $G = 1$, *or* unit noise figure $F = 1$, and gain $G = m$.

For a digital system, the transmission quality is often described in terms of the *Bit Error Rate* (BER) probability. Depending on the modulation scheme, coding, and a range of other factors (discussed in Part III of this book), there is a relationship between SNR and BER for each digital communications systems. A minimum transmission quality can thus be linked to the minimum SNR, SNR_{min} , by this mapping (see Figure 3.2). Thus, the planning methods of all analog and digital links in noise-limited environments are the same; the goal is to determine the minimum

9

Antennas

9.1 Introduction

9.1.1 Integration of Antennas into Systems

Antennas are the interface between the wireless propagation channel and the transmitter (TX) and receiver (RX), and thus have a major impact on the performance of wireless systems. This chapter therefore discusses some important aspects of antennas for wireless systems, both at the Base Station (BS) and at the Mobile Station (MS). We concentrate mainly on the antenna aspects that are specific to practical wireless systems.

Antenna design for wireless communications is influenced by two factors: (i) performance considerations and (ii) size and cost considerations. The latter aspect is especially important for antennas on MSs. These antennas must show not only good electromagnetic performance but must also be small, mechanically robust, and easy to produce. Furthermore, the performance of these antennas is influenced by the casing on which they are mounted, and by the person operating the handset. Antennas for BSs, on the other hand, are more similar to “conventional” antennas. Both size and cost are less restricted, and – at least for outdoor applications – the immediate surroundings of the antennas are clear of obstacles (some exceptions to this rule are also discussed in this chapter).

The remainder of this chapter thus distinguishes between MS antennas and BS antennas. Different types of antennas are discussed, as is the impact of the environment on antenna performance.

9.1.2 Characteristic Antenna Quantities

Directivity

The *directivity* D of an antenna is a measure of how much a transmit antenna concentrates the emitted radiation to a certain direction, or how much a receive antenna emphasizes radiation from a certain direction. More precisely, it is defined as [Vaughan and Andersen 2003]

$$D(\Omega) = \frac{\text{Total power radiated per unit solid angle in a direction } \Omega}{\text{Average power radiated per unit solid angle}} \quad (9.1)$$

Due to the principle of reciprocity, directivity is the same in the transmit and in the receive case. It is related to the far-field antenna power pattern $G(\Omega)$.¹ It is worth keeping in mind that the antenna

¹ Note that $G(\Omega)$ refers to antenna power, while we define $\tilde{G}(\Omega)$ as the complex amplitude gain, so that $G(\Omega) = |\tilde{G}(\Omega)|^2$. Both quantities are defined for the far field.

power pattern is normalized so that

$$\frac{1}{4\pi} \int G(\Omega) d\Omega = 1 \quad (9.2)$$

In many cases, antennas have different patterns for different polarizations. In such cases,

$$D(\phi_0, \theta_0) = \frac{G_\phi(\phi_0, \theta_0) + G_\theta(\phi_0, \theta_0)}{\frac{1}{4\pi} \int \int (G_\phi(\phi, \theta) + G_\theta(\phi, \theta)) \sin(\theta) d\theta d\phi} \quad (9.3)$$

where θ and ϕ are elevation and azimuth, respectively, and G_ϕ and G_θ are the power gains for radiation polarized in ϕ and θ , respectively.

The *gain* of an antenna in a certain direction is related to directivity. However, the gain also has to account for losses – e.g., ohmic losses. Those losses are described by antenna efficiency, which is discussed in the next subsection.

Efficiency

Losses in the antenna can be caused by several different phenomena. First, ohmic losses (i.e., due to the finite conductivity of the antenna material) can occur. Second, polarizations between the receive antenna and the incident radiated field can be misaligned (see below). Finally, losses can occur because of imperfect matching. The efficiency can thus be written as

$$\eta = \frac{R_{\text{rad}}}{R_{\text{rad}} + R_{\text{ohmic}} + R_{\text{match}}} \quad (9.4)$$

where R_{rad} is the radiation resistance; it is defined as the resistance of an equivalent network element so that the radiated power P_{rad} can be written as $0.5|I_0|^2 R_{\text{rad}}$, where I_0 is the excitation current magnitude. For example, the input impedance of a half-wavelength dipole is $Z_0 = 73 + j42 \Omega$, so that the radiation resistance is $R_{\text{rad}} = 73 \Omega$.

For antennas that are to operate on just a single frequency, perfect matching can be achieved. However, there are limits to how well an antenna can be matched over a larger band. The so-called Fano bound for the case that the antenna impedance can be modeled as a resistor and a capacitor in series reads

$$\int \frac{1}{(2\pi f_c)^2} \ln \left(\frac{1}{|\rho(f)|} \right) df < \pi RC \quad (9.5)$$

where ρ is the reflection coefficient, and R and C are resistance and capacitance, respectively.

High efficiency is a key criterion for any wireless antenna. From the point of view of the transmit antenna, high efficiency reduces the required power of the amplifier to achieve a given field strength. From the point of view of the receive antenna, the achievable Signal-to-Noise Ratio (SNR) is directly proportional to antenna efficiency. Antenna efficiency thus enters the battery lifetime of the MS, as well as the transmission quality of a link. It is difficult, however, to define an absolute goal for efficiency. Ideally, $\eta = 1$ should be obtained over the whole bandwidth of interest. However, it must be noted that in recent years the antenna efficiency of MS antennas has *decreased*. It has been sacrificed mainly for cosmetic reasons – namely, to decrease the size of the antennas.

Another factor influencing radiation efficiency is the presence of dielectric and/or conducting material – namely, the user – in the vicinity of an MS. When computing antenna gain that can be used in a link budget, this material has to be taken into account, as it leads to strong distortions of

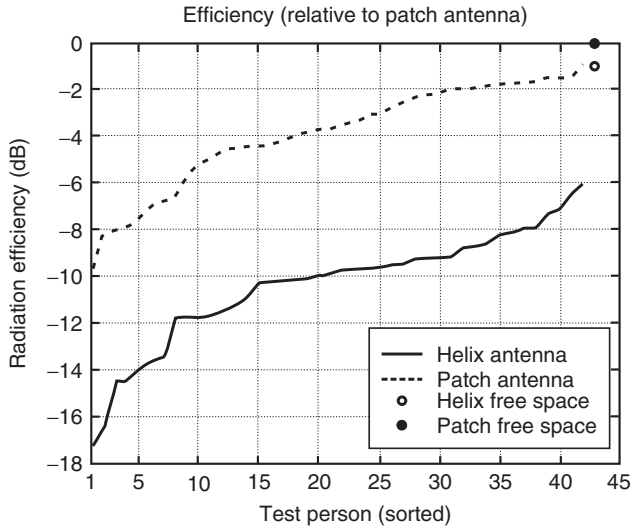


Figure 9.1 Efficiency of mobile station antennas close to a human body, relative to the efficiency of the patch antenna in free space. Different users and positions of the MS were used in the ensemble.

Reproduced with permission from Pedersen et al. [1998] © IEEE.

the antenna pattern and absorption of energy. Therefore, radiation efficiency is decreased when these effects are taken into account. The effective gain should be analyzed for a large number of users in order to eliminate the specific properties of any one user (at which angle with respect to the head is the MS being held?, is the user right-handed or left-handed?, what are the dielectric properties of the hand holding the device?, etc.). Figure 9.1 shows an exemplary cumulative distribution function for a patch antenna and a helical antenna. Note that the difference between effective antenna gain and theoretical values can exceed 10 dB. This has a large impact on network planning.

Q-Factor

A fundamental quantity of antennas is their Q-factor, defined as [Vaughan and Andersen 2003]

$$Q = 2\pi \frac{\text{Energy stored}}{\text{Energy dissipated per cycle}} \quad (9.6)$$

The Q-factor can be related to input impedance as

$$Q = \frac{f_c}{2R} \frac{\partial X}{\partial f} \quad (9.7)$$

where input impedance $Z = R + jX$. For an antenna contained in a sphere with diameter L_a , the Q-factor is given as

$$Q \geq \frac{1}{(k_0 L_a/2)^3} + \frac{1}{k_0 L_a/2} \quad (9.8)$$

where k_0 is again $2\pi/\lambda$.

Mean Effective Gain

The antenna pattern, and thus antenna directivity, is defined for the far field – i.e., assumes the existence of homogeneous plane waves – and is a function of the considered direction. It is measured in an anechoic chamber, where all obstacles and distorting objects are removed from the vicinity of the antenna. The patterns and gains measured this way are very close to the numbers found in any book on antenna theory – e.g., a Hertzian dipole has a gain of 1.5. When the antenna is operated in a random scattering environment, it becomes meaningful to investigate the *Mean Effective Gain* (MEG), which is an average of the gain over different directions when the directions of incident radiation are determined by a random environment [Andersen and Hansen 1977, Taga 1990]. The MEG is defined as the ratio of the average power received at the mobile antenna and the sum of the average power of the vertically and horizontally polarized waves received by isotropic antennas.

Polarization

In some wireless systems (e.g., satellite TV), the polarizations of the radiation and the receive antenna should be carefully aligned. This alignment can be measured by the Poincaré sphere. Each polarization state is associated with a point on the sphere: right-hand circular and left-hand circular polarizations are the north and south poles, respectively, while the different linear polarization states are on the equator; all other points correspond to elliptical polarizations. The angle between two points on the sphere is a measure of their mismatch.

For non-line-of-sight scenarios, requirements for a specific polarization of the antennas are typically not very stringent. Even if the transmit antenna sends mainly with a single polarization, propagation through the wireless channel leads to depolarization (see Chapter 7), so that cross-polarization at the *RX* is rarely higher than 10 dB. This fact is advantageous in many practical situations: as the orientation of MS antennas cannot be predicted (different users hold handsets in different ways), low sensitivity of the antenna to polarization of the incident radiation and/or uniform polarization of the incident radiation is advantageous.

One situation where good cross-polarization discrimination of antennas is required is the case of polarization diversity (see Chapter 13). In that case, the absolute polarization of the antennas is not critical; however, cross-polarization discrimination of two antenna elements plays an important role.

Bandwidth

Antenna bandwidth is defined as the bandwidth over which antenna characteristics (reflection coefficient, antenna gain, etc.) fulfill the specifications. Most wireless systems have a relative bandwidth of approximately 10%. This number already accounts for the fact that most systems use Frequency Domain Duplexing ((FDD), see Chapter 17) – i.e., transmit and receive on different frequencies. For example, in a GSM1800² system (Chapter 24), bandwidth is about 200 MHz, while the carrier frequency is around 1.8 GHz. As it is desirable to have only a single antenna for TX and RX cases, the antenna bandwidth must be large enough to include both the TX and RX bands. As we have mentioned above, a large bandwidth implies that the matching circuits can no longer be perfect. This effect is the stronger, the smaller the physical dimensions of the antenna are.

Another interesting special case involves dual- or multimode devices. For example, most GSM phones have to be able to operate at both the 900- and the 1,800-MHz carrier frequencies. Antennas for such devices can be constructed by requiring good performance across the whole 0.9–2-GHz band. However, using such an antenna with 50% relative bandwidth would be “overkill,” as the frequency range from 1 GHz to 1.7 GHz need not be covered. It is therefore better to just design

² Global System for Mobile communications in the 1,800-MHz band.

antennas that have good performance in the required bands; this goal is easier to fulfill if the different carrier frequencies are integer multiples of each other. Still, the design of antennas that cover two or more bands with high efficiency is a very challenging task. For BS antennas, it is thus often preferable from a technical point of view to use different antennas for different frequency bands. However, multiband antennas are to be preferred for esthetical reasons and reduced visual impact.

9.2 Antennas for Mobile Stations

9.2.1 Monopole and Dipole Antennas

Linear antennas are the “classical” antennas for MSs, and have for long determined the typical “look” of these devices. The most common ones are electric monopoles, located above a conducting plane (the casing), and dipoles. The antenna pattern of a short (Hertzian) dipole oriented along the z -axis is uniform in azimuth, and sine-shaped in polar angle θ (measured from the z -axis):

$$\tilde{G}(\varphi, \theta) \propto \sin(\theta) \quad (9.9)$$

with a maximum gain:

$$G_{\max} = 1.5 \quad (9.10)$$

A $\lambda/2$ dipole has the following properties (see Figure 9.2):

$$\tilde{G}(\varphi, \theta) \propto \frac{\cos\left(\frac{\pi}{2} \cos(\theta)\right)}{\sin(\theta)} \quad (9.11)$$

and a maximum gain:

$$G_{\max} = 1.64 \quad (9.12)$$

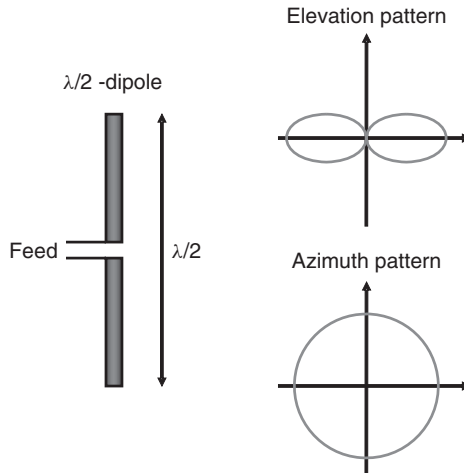


Figure 9.2 Shape and radiation pattern of a $\lambda/2$ dipole antenna.

We can thus see that the patterns of the $\lambda/2$ dipole and the Hertzian dipole do not differ dramatically. However, the radiation resistance can be quite different. For a dipole with *uniform current distribution*, the radiation resistance is

$$R_{\text{rad}}^{\text{uniform}} = 80\pi^2 (L_a/\lambda)^2 \quad (9.13)$$

For dipoles with a tapered current distribution (maximum at the feed, and linear decrease towards the end), the radiation resistance is $0.25R_{\text{rad}}^{\text{uniform}}$.

From the image principle it follows that the radiation pattern of a monopole located above a conducting plane is identical to that of a dipole antenna in the upper half-plane. Since the energy transmitted into the upper half-space $0 \leq \theta \leq \pi/2$ is twice that of the dipole, the maximum gain is twice and the radiation resistance half that of the dipole. Note, however, that the image principle is valid only if the conducting plane extends infinitely – this is not the case for MS casings. We can thus also anticipate considerable radiation in the lower half-space even for monopole antennas.

The reduction in radiation resistance is often undesirable, since it makes matching more difficult, and leads to a reduction in efficiency due to ohmic losses. One way of increasing efficiency without increasing the physical size of the antenna is to use *folded* dipoles. A folded dipole consists of a pair of half-wavelength wires that are joined at the non-feed end [Vaughan and Andersen 2003]; these increase the input impedance.

The biggest plus of monopole and dipole antennas is that they can be produced easily and cheaply. The relative bandwidth is sufficient for most applications in single-antenna systems. The disadvantage is the fact that a relatively long metal stick must be attached to the MS casing. In the 900-MHz band, a $\lambda/4$ monopole is 8 cm long – often longer than the MS itself. Even when realized as a retractable element, it is easily damaged. For this reason, shorter and/or integrated antennas, which are less efficient, are becoming increasingly widespread. This does not pose a significant problem in Europe and Japan, where coverage is usually good, and most systems are interference-limited anyway. However, in the U.S.A. and other countries where coverage is somewhat haphazard in many regions, this can have considerable influence on performance.

9.2.2 Helical Antennas

The geometry of helical antennas is outlined in Figure 9.3. A helical antenna can be seen as a combination of a loop antenna and a linear antenna, and thus has two modes of operation. The dimensions of the antenna determine which mode it is operating in. If the dimensions of the helix are much smaller than a wavelength, then the antenna is operating in normal mode. It behaves similar to a linear antenna, and has a pattern that is shaped mainly in the radial direction. This is the operating condition used in MS antennas. In general, the polarization is elliptical, though it can

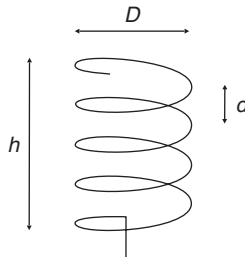


Figure 9.3 Geometry of a helical antenna.

become circular if the ratio $2\lambda d/(\pi D)^2$ becomes unity [Balanis 2005]. The polarization becomes vertical if the helical antenna is arranged over a conducting plane, as the horizontal components of the actual antenna and its image cancel out. When the circumference of the helix is on the order of one wavelength, the antenna pattern has its maximum along the axis of the helix, and polarization is almost circular.

Since the helical antenna in normal mode is similar to a linear antenna, the number of turns the antenna makes does not influence the antenna pattern. However, the bandwidth, efficiency, and radiation resistance increase with increasing h . In general, a helical antenna has lower bandwidth and a smaller input impedance than a monopole antenna; however, a relative bandwidth of 10% can be achieved by appropriate matching circuits. The main advantage of the helical antenna is its smaller size; this has made it (together with linear antennas) the most widely used external antenna for MSs.

9.2.3 Microstrip Antennas

A microstrip antenna (patch antenna) consists of a thin dielectric substrate, which is covered on one side by a thin layer of conducting material (ground plane), while on the other side there is a *patch* of conducting material. The configuration is outlined in Figure 9.4 (see also Fujimoto [2008] and Fujimoto et al. [1987]).

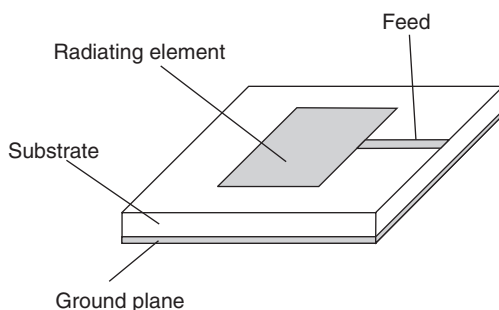


Figure 9.4 Geometry of a microstrip antenna.

The properties of a microstrip antenna are determined by the shape and dimension of the metallic patch, as well as by the dielectric properties of the used substrate. Essentially, the patch is a resonator whose dimensions have to be multiples of the effective dielectric wavelength. Thus, a high dielectric constant of the substrate allows the construction of small antennas. The most commonly used patch shapes are rectangular, circular, and triangular.

The patch is usually fed either by a coaxial cable or a microstrip line. It is also possible to feed the patch via electromagnetic coupling. This latter case uses a substrate where the ground plane is sandwiched between two layers of dielectric material. On the top of one material is the patch, while the feedline is at the bottom of the other dielectric layer. Coupling is effected through a slot (*aperture*) in the ground plane. These antennas are thus called aperture-coupled patch antennas. This design has the advantage that the dielectric properties of the two layers can be chosen differently, depending on the requirements for the patch and the feedline. Furthermore, this design shows a larger bandwidth than conventional patch antennas.

As mentioned above, the size and efficiency of the microstrip antenna are determined by the parameters of the dielectric substrate. A large ϵ_r reduces the size. This follows immediately from

the fact that in a resonator the length of one side must be

$$L = 0.5\lambda_{\text{substrate}} \quad (9.14)$$

where

$$\lambda_{\text{substrate}} = \lambda_0 / \sqrt{\epsilon_r} \quad (9.15)$$

Unfortunately, a reduction in physical size also leads to a smaller bandwidth, which is usually undesirable. For this reason, substrates used in practice usually have a very low ϵ_r – even air is used quite frequently. A further possibility for reducing the size of patch antennas is the use of short-circuited resonators, which reduces the required size of a resonator from $\lambda/2$ to $\lambda/4$ (see Figure 9.5).

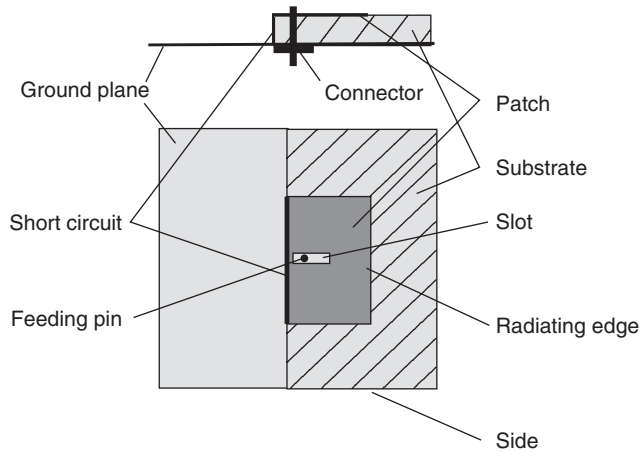


Figure 9.5 Short-circuited $\lambda/4$ patch antenna.

The bandwidth of microstrip antennas can be increased by various measures. The most straightforward one is an increase in antenna volume – i.e., the use of thicker substrates with a lower ϵ_r . Alternatives are the use of matching circuits and the use of parasitic elements.

Microstrip antennas have several important advantages for wireless applications:

- They are small and can be manufactured cheaply.
- The feedlines can be manufactured on the same substrate as the antenna.
- They can be integrated into the MS, without sticking out from the casing.

However, they also have serious weaknesses:

- They have a low bandwidth (usually just a few percent of the carrier frequency).
- They have low efficiency.

9.2.4 Planar Inverted F Antenna

Some of the problems of microstrip antennas can be alleviated by a *Planar Inverted F Antenna* (PIFA). The shape of the PIFA is similar to that of a $\lambda/4$ short-circuited microstrip antenna (see

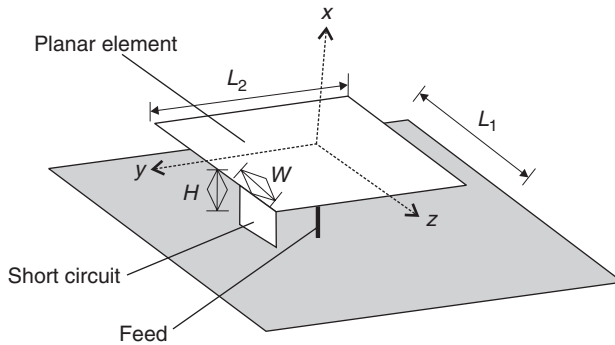


Figure 9.6 Planar inverted-F antenna.

Figure 9.6). A planar, radiating element is located parallel to a ground plane. This element is short-circuited over a distance W . If W is chosen equal to the length of the edge L , then we obtain a short-circuited $\lambda/4$ microstrip antenna. If W is chosen smaller, then the resonance length increases, and the current distribution on the radiating element changes.

9.2.5 Radiation Coupled Dual L Antenna

A further improvement is achieved by the so-called *Radiation Coupled Dual L Antenna* (RCDLA) (see Figure 9.7 and Rasinger et al. [1990]). It consists of two L-shaped angular structures only one of which is fed directly (conductively). The other L-shaped structure is fed by the first L by means of radiation coupling. This increases the bandwidth of the total arrangement. By optimally placing the antenna on the casing, relative bandwidths of up to 10% can be achieved, which is about twice the bandwidth of a PIFA.

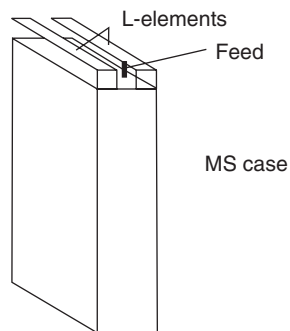


Figure 9.7 Radiation-coupled dual-L antenna.

9.2.6 Multiband Antennas

Modern cellular handsets are anticipated to be able to handle different frequencies for communications. As discussed in Section 9.2.2 a GSM handset, e.g., needs to be able to deal at least with 900 and 1,800 MHz foreseen in the specifications for most countries. As an added difficulty, many

handsets should also be able to cope with the 1,900-MHz mode used in the U.S.A., as well as 2.4 GHz if Bluetooth connections (e.g., to a wireless headset) are required. The situation becomes even more complicated for dual-mode devices that can handle both GSM and Wideband Code Division Multiple Access (WCDMA) (see Chapter 26). The design of internal multiband antennas is very complicated, and few rules for a closed-form design are available. Figure 9.8 shows an example of a microstrip multiband antenna.

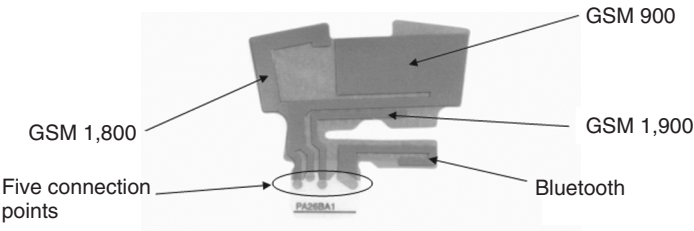


Figure 9.8 Integrated multiband antenna.

Reproduced with permission from Ying and Anderson [2003] © Z. Ying.

9.2.7 *Antenna Mounting on the Mobile Station*

Antennas do not operate in empty space but are placed on top of the casing, which can be considered to be part of the radiator. Furthermore, antenna characteristics are influenced by the hand and the head of the user; this influence also depends on the mounting of the antenna on the MS. It is therefore important to investigate different options for placing the antenna, and to see how this placement influences performance.

Linear and helical antennas are usually placed on the upper, narrow side of the casing – i.e., they stick out from that part of the casing. This has mainly ergonomic reasons – if they were sticking out from the lower side, they would feel uncomfortable to the user, and the hand of the user would often cover the antenna, leading to additional attenuation.

For microstrip antennas, PIFAs, and RCDLAs, there are more options for placements. These antennas are usually used as internal antennas – i.e., integrated into the casing or enclosed within the casing. This greatly reduces the danger of mechanical damage. However, there is an increased probability that users will place their hands over the antenna, which increases absorption of the electromagnetic energy and thus worsens link performance [Erätuuli and Bonek 1997]. Examples for the positioning of RCDLAs can be found in Figure 9.9; other types of microstrip antennas can be placed in a similar way.

The impact of the human body on antenna patterns is discussed in Section 9.3.4.

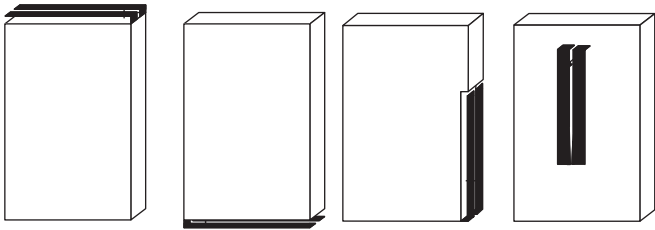


Figure 9.9 Placement of radiation-coupled dual-L antennas on the casing of a mobile station.

9.3 Antennas for Base Stations

9.3.1 Types of Antennas

The design requirements for BS antennas are different from those of MS antennas. Cost has a smaller impact, as BSs are much more expensive in the first place. Also, size restrictions are less stringent: for macrocells, it is only required that (i) the mechanical stress on the antenna mast, especially due to wind forces, must remain reasonable and (ii) the “cosmetic” impact on the surroundings must be small. For micro- and picocells, antennas need to be considerably smaller, as they are mounted on building surfaces, street lanterns, or on office walls. The desired antenna pattern is quite different for BS antennas compared with MS antennas. As the physical placement and orientation of the BS antenna is known, patterns should be shaped in such a way that no energy is wasted (see also Section 9.3.3). Such pattern shaping can be most easily implemented by multielement antennas.

For these reasons, macrocellular antennas typically are antenna arrays or Yagi antennas whose elements are linear antennas. For micro- and picocells, antenna arrays consisting of patch antennas are common. All considerations from Section 9.2 about bandwidth, efficiency, etc., remain valid, except for the size requirements which are relaxed quite a bit.

9.3.2 Array Antennas

Array antennas are often used for BS antennas. They result in an antenna pattern that can be more easily shaped. This shaping can either be done in a predetermined way (see Section 9.3.3) or adaptively (see Chapters 13 and 20). The pattern of an antenna array can be written as the product of the pattern of a single element and the array factor $M(\phi, \theta)$. In the plane of the array antenna, the array factor of a uniform linear array is (compare Section 8.5.1) [Stutzman and Thiele 1997]

$$M(\phi) = \sum_{n=0}^{N_t-1} w_n \exp[-j \cos(\phi) 2\pi d_a n / \lambda] \quad (9.16)$$

where d_a is the distance between the antenna elements,³ and w_n are the complex weights of element excitations. For the case that all $|w_n| = 1$ and $\arg(w_n) = n\Delta$ the array factor becomes

$$|M(\phi)| = \left| \frac{\sin \left[\frac{N_t}{2} \left(\frac{2\pi}{\lambda} d_a \cos \phi - \Delta \right) \right]}{\sin \left[\frac{1}{2} \left(\frac{2\pi}{\lambda} d_a \cos \phi - \Delta \right) \right]} \right| \quad (9.17)$$

The phase shift Δ of the feed currents determines the direction of the antenna main lobe. This principle is well known from the theory of phased array antennas ([Hansen 1998], see also Chapter 8). By imposing $\varphi_n = n\Delta$, the degrees of freedom have reduced to one; while the main lobe can be put into an arbitrary direction, the placement of the sidelobes and the nulls follows uniquely from that direction.

³ Distance is usually chosen as $d_a = \lambda/2$ ($d_a \leq \lambda/2$ is necessary to avoid spatial aliasing – i.e., periodicities in the antenna pattern). In the following, we assume that the elements of the linear array are on the x-axis. Obviously, the same principles are valid when antenna elements are stacked vertically, and the elevation pattern of the array should be shaped.

Example 9.1 Consider a BS antenna mounted at a height of 51 m, providing coverage for a cell with a radius of 1 km. The antenna consists of eight vertically stacked short dipoles, separated by $\lambda/2$. How much should the phase shift Δ be so that the maximum points at the cell edge at an MS height of 1 m?

In order to get constructive interference of the contributions from the different antennas in the direction θ_0 , the angle Δ should fulfill

$$\Delta = \frac{2\pi}{\lambda} d_a \cos \theta_0 \quad (9.18)$$

Obviously, if $\theta_0 = \pi/2$, then $\Delta = 0$ – i.e., no phase shift is necessary. In our example, we are looking for a tilt angle:

$$\theta_0 = \frac{\pi}{2} + \arctan\left(\frac{51 - 1}{1000}\right) = \frac{\pi}{2} + 0.05 \quad (9.19)$$

The phase shift thus has to be

$$\Delta = \frac{2\pi}{\lambda} d_a \cos\left(\frac{\pi}{2} + 0.05\right) = -0.05\pi \quad (9.20)$$

The impact of the element pattern can be neglected.

9.3.3 Modifying the Antenna Pattern

It is undesirable that a BS antenna radiates isotropically. Radiation emitted in a direction that has a large elevation angle (i.e., into the sky) is not only wasted but actually increases interference with other systems. The optimum antenna pattern is achieved when the received power is constant within the whole cell area, and vanishes outside. The problem is then to synthesize such an elevation pattern. This can be achieved approximately by means of array antennas, where the complex weights are chosen in such a way as to minimize the mean-square error of the pattern. An even simpler approach is to use an antenna array with $w_n = 1$, but tilt the main lobe down by about 5° . This downtilt can be achieved either mechanically (by tilting the whole antenna array) or electronically.

The desired azimuthal antenna pattern is either (i) omnidirectional – i.e., uniform in $[0, 2\pi)$ – or (ii) uniform within a sector, and zero outside. The angular range of a sector is usually 60° or 120° . For omnidirectional antennas, which are mostly used in rural macrocells, linear antennas are used. The antennas are arranged in a linear array that extends only along the vertical axis. For sector antennas, either linear antennas with appropriately shaped reflectors or microstrip antennas (which have inherently nonuniform antenna patterns) can be used.

9.3.4 Impact of the Environment on Antenna Pattern

The antenna pattern of BS antennas is usually defined for the case when an antenna is in free space, or above an ideally conducting plane. This is what is measured in an anechoic chamber, and is also the easiest to compute. However, at its location of operation, a BS antenna is surrounded by different objects of finite extent and conductivity. The antenna patterns in such surroundings can deviate significantly from theoretical patterns. The antenna mast, which is usually metallic, and thus highly conductive, can lead to distortions. Similarly, roofs made out of certain building materials, like reinforced concrete, can distort antenna patterns (see Figure 9.10).

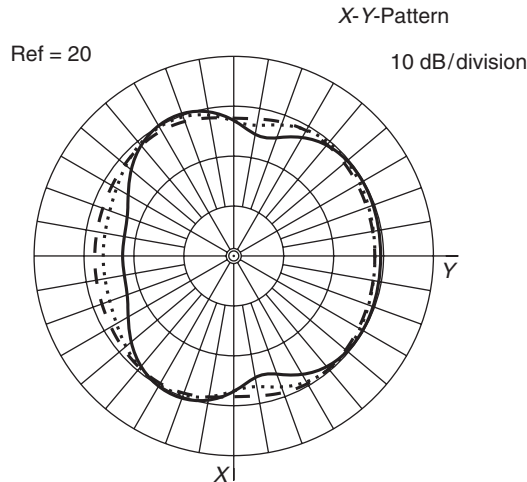


Figure 9.10 Antenna pattern of an omnidirectional antenna close to an antenna mast. Distance from the antenna mast 30 cm. Diameter of the mast: very small (solid), 5 cm (dashed), 10 cm (dotted).

Reproduced with permission from Molisch et al. [1995] © European Microwave Association.

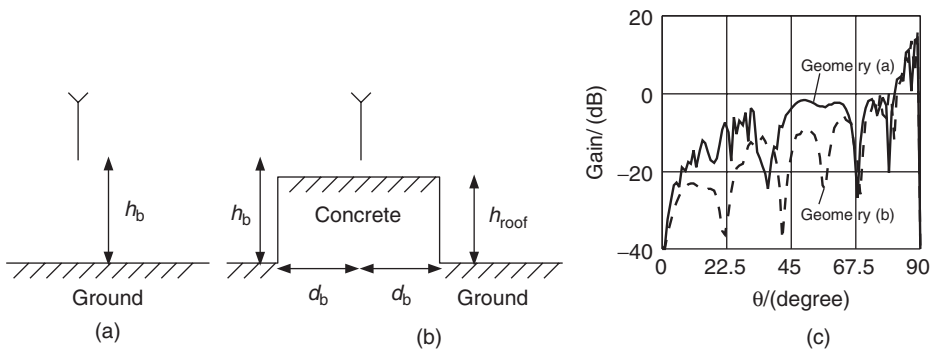


Figure 9.11 Distortion in the vertical antenna pattern. (a) Idealized geometry. (b) Real geometry, including the roof. (c) Distortions in antenna patterns. Properties of all materials: relative dielectric constant $\epsilon_r = 2$; conductivity 0.01 S/m ; $d_b = 20 \text{ m}$.

Reproduced with permission from Molisch et al. [1995] © European Microwave Association.

Distortions in the vertical pattern are plotted in Figure 9.11. They arise from two effects: the fact that the (finite extent) roof is much closer to the antenna than the ground and that the dielectric properties of the roof are different from those of the ground.

The presence of a human body also distorts antenna patterns. One way of taking this into account is to consider the antenna and the human body as a “superantenna,” whose characteristics (efficiency, radiation pattern) can be measured and characterized in the same manner as “regular” antennas. Figure 9.12 shows example measurements of antenna patterns by a human head and body.

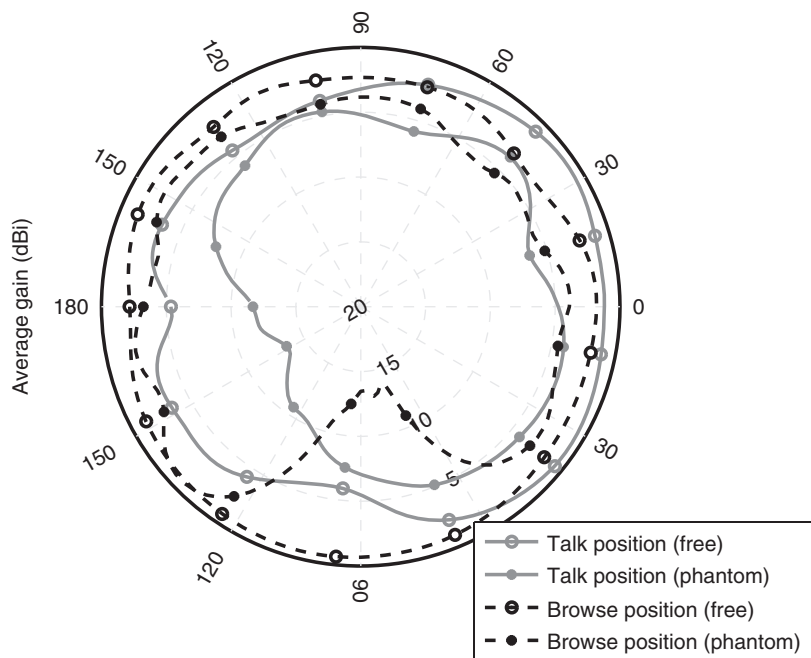


Figure 9.12 Pattern of antenna distortion by human body in talk and data-browsing position.

Reproduced with permission from Harryson et al. [2010]. Copyright IEEE.

Further Reading

For a general introduction to antenna theory, we just refer to the many excellent books on antenna theory: Balanis [2005], Kraus and Marhefka [2002], Ramo et al. [1967], and Stutzman and Thiele [1997], as well as the somewhat more advanced text of Collin [1985]; the latter also devotes a separate chapter to the properties of receiving antennas. For more details on antenna specifically for wireless communications, see Godara [2001] and Vaughan and Andersen [2003]. Antennas for the MS are discussed in Fujimoto [2008] and Hirasawa and Haneishi [1991]. Antenna design for BSs is surveyed in the monograph of Chen and Luk [2009] and in the conference paper by Beckman and Lindmark [2007]. Finally, phased array antennas are treated in detail in Hansen [1998] and Mailloux [1994]. Discussions on beam tilting can be found in Manholm et al. [2003]. The impact of the human head and body on antenna characteristics is discussed, e.g., in Ogawa and Matsuyoshi [2001], Kivekaes et al. [2004] and [Harryson et al. 2010].

For updates and errata for this chapter, see wides.usc.edu/teaching/textbook

13

Diversity

13.1 Introduction

13.1.1 Principle of Diversity

In the previous chapter, we treated conventional transceivers that transmit an uncoded bitstream over fading channels. For Additive White Gaussian Noise (AWGN) channels, such an approach can be quite reasonable: the Bit Error Rate (BER) decreases exponentially as the Signal-to-Noise Ratio (SNR) increases, and a 10-dB SNR leads to BERs on the order of 10^{-4} . However, in Rayleigh fading the BER decreases only linearly with the SNR. We thus would need an SNR on the order of 40 dB in order to achieve a 10^{-4} BER, which is clearly impractical. The reason for this different performance is the fading of the channel: the BER is mostly determined by the probability of channel attenuation being large, and thus of the instantaneous SNR being low. A way to improve the BER is thus to change the effective channel statistics – i.e., to make sure that the SNR has a smaller probability of being low. Diversity is a way to achieve this.

The principle of diversity is to ensure that the same information reaches the receiver (RX) on statistically independent channels. Consider the simple case of an RX with two antennas. The antennas are assumed to be far enough from each other that small-scale fading is independent at the two antennas. The RX always chooses the antenna that has instantaneously larger receive power.¹ As the signals are statistically independent, the probability that both antennas are in a fading dip *simultaneously* is low – certainly lower than the probability that one antenna is in a fading dip. The diversity thus changes the SNR statistics at the detector input.

Example 13.1 Diversity reception in a two-state fading channel.

To quantify this effect, let us consider a simple numerical example: the noise power within the RX filter bandwidth is 50 pW, the average received signal power is 1 nW, the SNR is thus 13 dB. In an AWGN channel, the resulting BER is 10^{-9} , assuming that the modulation is differentially detected Frequency Shift Keying (FSK). Now consider a fading channel where during 90% of the time the received power is 1.11 nW, and the SNR is thus 13.5 dB, while for the remainder, it is zero. This means that during 90% of the time, the BER is 10^{-10} ; the remainder of the time,

¹ We will see later on that this scheme is only one of many different possible diversity schemes.

it is 0.5; the average BER is thus

$$0.9 \cdot 10^{-10} + 0.1 \cdot 0.5 = 0.05 \quad (13.1)$$

For the case of two-antenna diversity, the probability that the received signal power is 0 at both antennas simultaneously is $0.1 \cdot 0.1 = 0.01$. The probability that the received power is 1.11 nW at both antennas simultaneously is $0.9 \cdot 0.9 = 0.81$; the probability that it is 1.11 nW at one antenna and 0 at the other is 0.18. Assuming selection diversity, in both the latter cases, the SNR at the detector is 13.5 dB. The total BER is thus

$$0.01 \cdot 0.5 + 0.99 \cdot 10^{-10} = 0.005 \quad (13.2)$$

This is approximately the square of the BER for a single-antenna system. If we have three antennas, then the probability that the signal power is 0 at all three antennas simultaneously is 0.1^3 ; the total BER is then $0.5 \cdot 0.001 + 0.999 \cdot 10^{-10} = 0.0005$; this is approximately the third power of the BER for a single-antenna system.

Later sections will give exact equations for the BER with diversity in Rayleigh-fading channels. The general concepts, however, are the same as in the simple example described above. With N_r diversity antennas, we obtain a bit error probability $\propto BER_{oc}^{N_r}$, where BER_{oc} is the BER with just one receive channel.²

In the following, we first describe the characterization of correlation coefficients between different transmission channels. We then give an overview about how transmission over independent channels can be realized – spatial antenna diversity described above is one, but certainly not the only approach. Next, we describe how signals from different channels can best be combined, and what performance can be achieved with the different combining schemes.

13.1.2 Definition of the Correlation Coefficient

Diversity is most efficient when the different transmission channels (also called diversity branches) carry independently fading copies of the same signal. This means that the joint probability density function (pdf) of field strength (or power) $pdf_{r_1, r_2, \dots}(r_1, r_2, \dots)$ is equal to the product of the marginal pdfs for the channels, $pdf_{r_1}(r_1), pdf_{r_2}(r_2), \dots$. Any correlation between the fading of the channels decreases the effectiveness of diversity.

The *correlation coefficient* characterizes the correlation between signals on different diversity branches. A number of different definitions are being used for this important quantity: complex correlation coefficients, correlation coefficient of the phase, etc. The most important one is the correlation coefficient of signal envelopes x and y :

$$\rho_{xy} = \frac{E\{x \cdot y\} - E\{x\} \cdot E\{y\}}{\sqrt{(E\{x^2\} - E\{x\}^2) \cdot (E\{y^2\} - E\{y\}^2)}} \quad (13.3)$$

For two statistically independent signals, the relationship $E\{xy\} = E\{x\}E\{y\}$ holds; therefore, the correlation coefficient becomes zero. Signals are often said to be “effectively” decorrelated if ρ is below a certain threshold (typically 0.5 or 0.7).

² Since in a Rayleigh-fading channel, $BER_{oc} \propto SNR^{-1}$, we find that for a diversity system with N_r independently fading channels, $BER \propto SNR^{-N_r}$. Quite generally in fading channels with diversity, $BER \propto SNR^{-d_{div}}$, where d_{div} is known as *diversity order*.

13.2 Microdiversity

As mentioned in the introduction, the basic principle of diversity is that the RX has multiple copies of the transmit signal, where each of the copies goes through a statistically independent channel. This section describes different ways of obtaining these statistically independent copies.

We concentrate on methods that can be used to combat small-scale fading, which are therefore called “microdiversity.” The five most common methods are as follows:

1. *Spatial diversity*: several antenna elements separated in space.
2. *Temporal diversity*: transmission of the transmit signal at different times.
3. *Frequency diversity*: transmission of the signal on different frequencies.
4. *Angular diversity*: multiple antennas (with or without spatial separation) with different antenna patterns.
5. *Polarization diversity*: multiple antennas with different polarizations (e.g., vertical and horizontal).

When we speak of antenna diversity, we imply that there are multiple antennas at the *receiver*. Only in Section 13.6 (and later in Chapter 20) will we discuss how multiple *transmit* antennas can be exploited to improve performance.

The following important equation will come in handy: Consider the correlation coefficient of two signals that have a temporal separation τ and a frequency separation $f_1 - f_2$. As shown in Appendix 13.A (see www.wiley.com/go/molisch), the correlation coefficient is

$$\rho_{xy} = \frac{J_0^2(k_0 v \tau)}{1 + (2\pi)^2 S_\tau^2 (f_2 - f_1)^2} \quad (13.4)$$

Note that for moving Mobile Stations (MSs), temporal separation can be easily converted into spatial separation, so that temporal and spatial diversity become mathematically equivalent. Equation (13.4) is thus quite general in the sense that it can be applied to spatial, temporal, and frequency diversity. However, a number of assumptions were made in the derivation of this equation: (i) validity of the Wide Sense Stationary Uncorrelated Scatterer (WSSUS) model, (ii) no existence of Line Of Sight (LOS), (iii) exponential shape of the Power Delay Profile (PDP), (iv) isotropic distribution of incident power, and (v) use of omnidirectional antennas.

Example 13.2 Compute the correlation coefficient of two frequencies with separation (i) 30 kHz, (ii) 200 kHz, (iii) 5 MHz, in the “typical urban” environment, as defined in COST 207³ channel models.

For zero temporal separation, the Bessel function in Eq. (13.4) is unity, and the correlation coefficient is thus only dependent on rms delay spread and frequency separation. Rms delay spread is calculated according to Eq. (6.39), where the PDP of the COST 207 typical urban channel model is given in Section 7.6.3. We obtain

$$S_\tau = \sqrt{\frac{\int_0^{7 \cdot 10^{-6}} e^{-\tau/10^{-6}} \tau^2 d\tau}{\int_0^{7 \cdot 10^{-6}} e^{-\tau/10^{-6}} d\tau}} - \left(\frac{\int_0^{7 \cdot 10^{-6}} e^{-\tau/10^{-6}} \tau d\tau}{\int_0^{7 \cdot 10^{-6}} e^{-\tau/10^{-6}} d\tau} \right)^2$$

$$= 0.977 \mu\text{s} \quad (13.5)$$

³ European COoperation in the field of Scientific and Technical research.

The correlation coefficient thus becomes

$$\begin{aligned}\rho_{xy} &= \frac{1}{1 + (2\pi)^2(0.977 \cdot 10^{-6})^2(f_1 - f_2)^2} \\ &= \begin{cases} 0.97, & f_1 - f_2 = 30 \text{ kHz} \\ 0.4, & f_1 - f_2 = 200 \text{ kHz} \\ 1 \cdot 10^{-3}, & f_1 - f_2 = 5 \text{ MHz} \end{cases} \quad (13.6)\end{aligned}$$

From this, we can see that the correlation between two neighboring 30-kHz bands – as used, e.g., in the old IS-136⁴ Time Division Multiple Access (TDMA) cellular system – is very high; correlation over 200-kHz bands (two neighboring channels in Global System for Mobile communication (GSM) are 200 kHz apart, see Chapter 24) is also appreciable, but two neighboring channels in Wideband Code Division Multiple Access (WCDMA) (5 MHz apart) are uncorrelated in this environment. Furthermore, carriers that are, e.g., 45 MHz apart – used for Base Station (BS)–MS and MS–BS communication in GSM (see Chapter 24 and also Chapter 17) are completely uncorrelated.

13.2.1 Spatial Diversity

Spatial diversity is the oldest and simplest form of diversity. Despite (or because) of this, it is also the most widely used. The transmit signal is received at several antenna elements, and the signals from these antennas are then further processed according to the principles that will be described in Section 13.4. But, irrespective of the processing method, performance is influenced by correlation of the signals between the antenna elements. A large correlation between signals at antenna elements is undesirable, as it decreases the effectiveness of diversity. A first important step in designing diversity antennas is thus to establish a relationship between antenna spacing and the correlation coefficient. This relationship is different for BS antennas and MS antennas, and thus will be treated separately.

1. *MS in cellular and cordless systems*: it is a standard assumption that waves are incident from all directions at the MS (see also Chapter 5). Thus, points of constructive and destructive interference of Multi Path Components (MPCs) – i.e., points where we have high and low received power, respectively – are spaced approximately $\lambda/4$ apart. This is therefore the distance that is required for decorrelation of received signals. This intuitive insight agrees very well with the results from the exact mathematical derivation (Eq. (13.4), with $f_2 - f_1 = 0$), given in Figure 13.1: decorrelation, defined as $\rho = 0.5$, occurs at an antenna separation of $\lambda/4$. Compare also Example 5.3.

The above considerations imply that the minimum distance for antenna elements in GSM (at 900 MHz) is about 8 cm, and for various cordless and cellular systems at the 1,800-MHz band it is about 4 cm. For Wireless Local Area Networks (WLANs) (at 2.4 and 5 GHz), the distances are even smaller. It is thus clearly possible to place two antennas on an MS of a cellular system.

2. *BS in cordless systems and WLANs*: in a first approximation, the angular distribution of incident radiation at indoor BSs is also uniform – i.e., radiation is incident with equal strength from all directions. Therefore, the same rules apply as for MSs.
3. *BSs in cellular systems*: for a cellular BS, the assumption of uniform directions of incidence is no longer valid. Interacting Objects (IOs) are typically concentrated around the MS (Figure 13.2, see also Chapter 7). Since all waves are incident essentially from one direction, the correlation

⁴ IS-136 is a (now defunct) second-generation cellular system using TDMA.

coefficient (for a given distance between antenna elements d_a) is much higher. Expressed differently, the antenna spacing required to obtain sufficient decorrelation increases.

To get an intuitive insight, we start with the simple case when there are only two MPCs whose wave vectors are at an angle α with respect to each other (Figure 13.3, see also Chapter 5). It is obvious that the distance between the maxima and minima of the interference pattern is larger the smaller α is. For very small α , the connection line between antenna elements lies on a “ridge” of the interference pattern and antenna elements are completely correlated.

Numerical evaluations of the correlation coefficient as a function of antenna spacing are shown in Figure 13.4. The first column shows the results for rectangular angular power spectra; the results for Gaussian distributions are shown in the second column. We can see that antenna spacing has to be on the order of 2–20 wavelengths for angular spreads between 1° and 5° in order to achieve decorrelation. We also find that it is mostly rms angular spread that determines the required antenna spacing, while the shape of the angular power spectrum has only a minor influence.

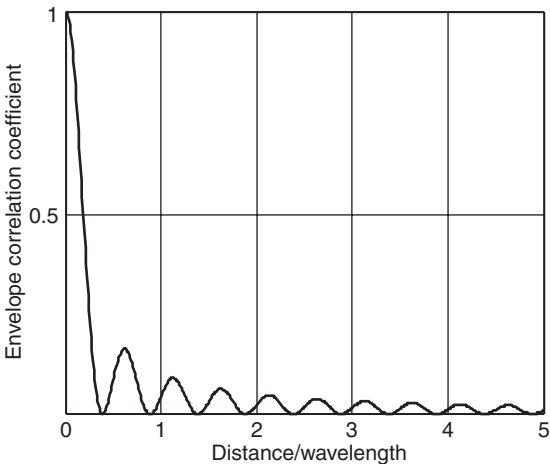


Figure 13.1 Envelope correlation coefficient as a function of antenna separation.

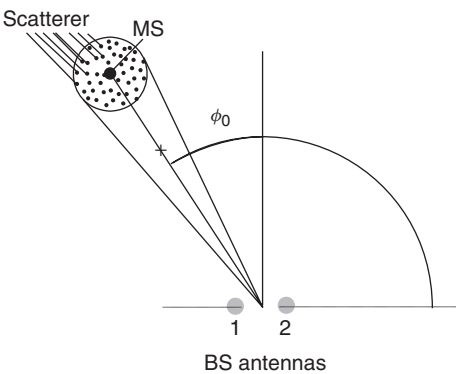


Figure 13.2 Scatterers concentrated around the mobile station.

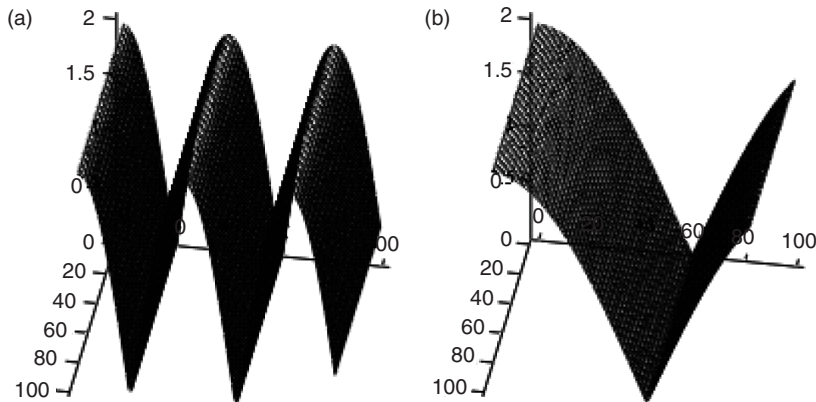


Figure 13.3 Interference pattern of two waves with 45° (a) and 15° (b) angular separation.

13.2.2 Temporal Diversity

As the wireless propagation channel is time variant, signals that are received at different times are uncorrelated. For “sufficient” decorrelation, the temporal distance must be at least $1/(2\nu_{\max})$, where $\nu_{\max}$ is the maximum Doppler frequency. In a static channel, where neither transmitter (TX), RX, nor the IOs are moving, the channel state is the same at all times. Such a situation can occur, e.g., for WLANs. In such a case, the correlation coefficient is $\rho = 1$ for all time intervals, and temporal diversity is useless.

Temporal diversity can be realized in different ways:

1. *Repetition coding*: this is the simplest form. The signal is repeated several times, where the repetition intervals are long enough to achieve decorrelation. This obviously achieves diversity, but is also highly bandwidth inefficient. Spectral efficiency decreases by a factor that is equal to the number of repetitions.
2. *Automatic Repeat reQuest (ARQ)*: here, the RX sends a message to the TX to indicate whether it received the data with sufficient quality (see Appendix 14.A at www.wiley.com/go/molisch). If this is not the case, then the transmission is repeated (after a wait period that achieves decorrelation). The spectral efficiency of ARQ is better than that of repetition coding, since it requires multiple transmissions only when the first transmission occurs in a bad fading state, while for repetition coding, retransmissions occur always. On the downside, ARQ requires a feedback channel.
3. *Combination of interleaving and coding*: a more advanced version of repetition coding is forward error correction coding with interleaving. The different symbols of a codeword are transmitted at different times, which increases the probability that at least some of them arrive with a good SNR. The transmitted codeword can then be reconstructed. For more details, see Chapter 14.

13.2.3 Frequency Diversity

In frequency diversity, the same signal is transmitted at two (or more) different frequencies. If these frequencies are spaced apart by more than the coherence bandwidth of the channel, then their fading is approximately independent, and the probability is low that the signal is in a deep fade at both frequencies simultaneously. For an exponential PDP, the correlation between two

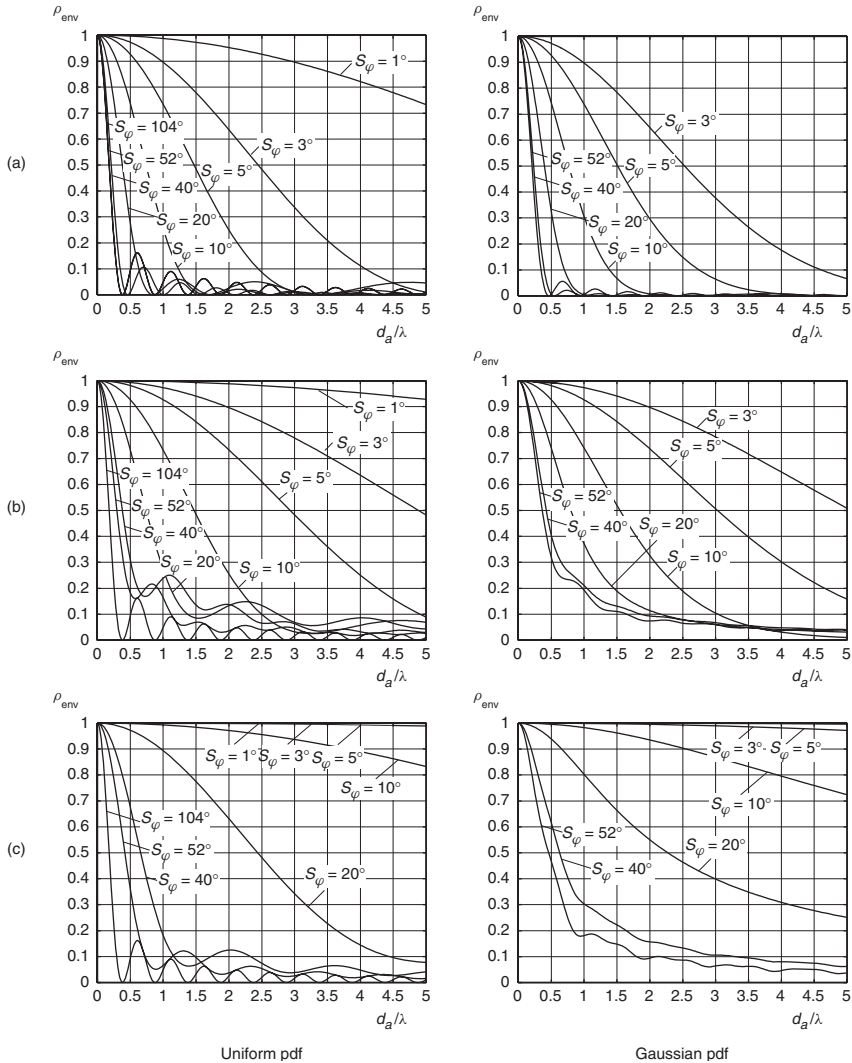


Figure 13.4 Envelope correlation coefficient at the BS for uniform and Gaussian probability density function (pdf) of the directions of arrival (a) $\phi_0 = 90^\circ$, (b) $\phi_0 = 45^\circ$, (c) $\phi_0 = 10^\circ$, and different values of angular spread S_φ . Reproduced with permission from Fuhl et al. [1998] © IEEE.

frequencies can be obtained from Eq. (13.4) by setting the numerator to unity as the signals at the two frequencies occur at the same time. Thus

$$\rho = \frac{1}{1 + (2\pi)^2 S_t^2 (f_2 - f_1)^2} \quad (13.7)$$

This again confirms that the two signals have to be at least one coherence bandwidth apart from each other. Figure 13.5 shows ρ as a function of the spacing between the two frequencies. For a more general discussion of frequency correlation, see Chapter 6.

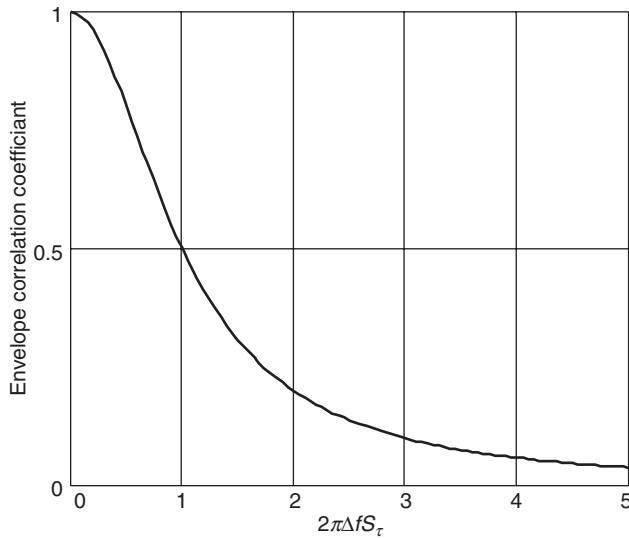


Figure 13.5 Correlation coefficient of the envelope as a function of normalized frequency spacing.

It is not common to actually repeat the same information at two different frequencies, as this would greatly decrease spectral efficiency. Rather, information is spread over a large bandwidth, so that small parts of the information are conveyed by different frequency components. The RX can then sum over the different frequencies to recover the original information.

This spreading can be done by different methods:

- *Compressing the information in time*: – i.e., sending short bursts that each occupy a large bandwidth – TDMA (see Chapter 17).
- *Code Division Multiple Access (CDMA)*: (Section 18.2).
- Multicarrier CDMA (Section 19.9) and coded orthogonal frequency division multiplexing (Section 19.4.3).
- *Frequency hopping in conjunction with coding*: different parts of a codeword are transmitted on different carrier frequencies (Section 18.1).

These methods allow the transmission of information without wasting bandwidth and will be described in greater detail in Chapters 17–19. For the moment, we just stress that the use of frequency diversity requires the channel to be frequency selective. In other words, frequency diversity (delay dispersion) can be exploited by the system to make it more robust, and decrease the effects of fading. This seems to be a contradiction to the results of Chapter 12, where we had shown that frequency selectivity leads to an *increase* of the BER, and even an error floor. The reason for this discrepancy is that Chapter 12 considered a very simple RX that takes no measures to combat (or exploit) the effects of frequency selectivity.

13.2.4 Angle Diversity

A fading dip is created when MPCs, which usually come from different directions, interfere destructively. If some of these waves are attenuated or eliminated, then the location of fading dips changes.

In other words, two colocated antennas with different patterns “see” differently weighted MPCs, so that the MPCs interfere differently for the two antennas. This is the principle of *angle diversity* (also known as *pattern diversity*).

Angular diversity is usually used in conjunction with spatial diversity; it enhances the decorrelation of signals at closely spaced antennas. Different antenna patterns can be achieved very easily. Of course, different types of antennas have different patterns. But even identical antennas can have different patterns when mounted close to each other (see Figure 13.6). This effect is due to *mutual coupling*: antenna B acts as a reflector for antenna A, whose pattern is therefore skewed to the left.⁵ Analogously, the pattern of antenna B is skewed to the right due to reflections from antenna A. Thus, the two patterns are different.

The different patterns are even more pronounced when the antennas are located on different parts of the casing. While dipole antennas are usually restricted to the top of the casing, patch antennas and inverted-F antennas (see Chapter 9) can be placed on all parts of the casing (see Figure 13.7). In all of these cases, decorrelation is good even if the antennas are placed very closely to each other.

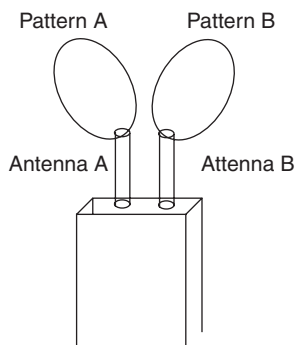


Figure 13.6 Angle diversity for closely spaced antennas.

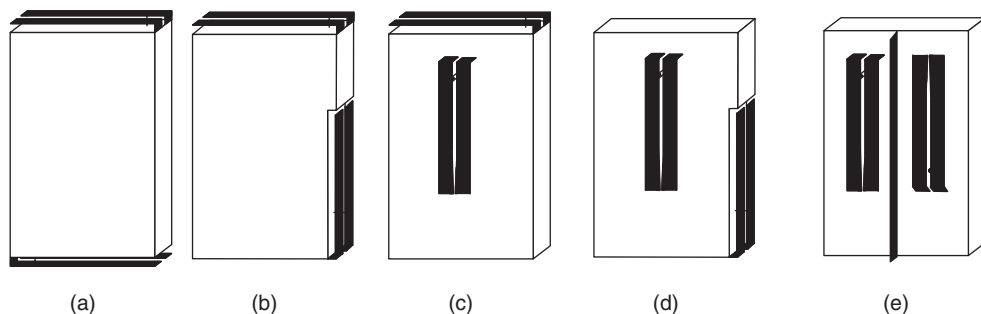


Figure 13.7 Configurations of diversity antennas at a mobile station.

Reproduced with permission from Erätuuli and Bonek [1997] © IEEE.

⁵ This arrangement can also be considered as a Yagi antenna. It depends on the spacing of the two elements whether antenna B acts as director or reflector, and thus which direction the pattern is skewed into.

13.2.5 Polarization Diversity

Horizontally and vertically polarized MPCs propagate differently in a wireless channel,⁶ as the reflection and diffraction processes depend on polarization (see Chapters 4 and 7). Even if the transmit antenna only sends signals with a single polarization, the propagation effects in the channel lead to depolarization so that both polarizations arrive at the RX. The fading of signals with different polarizations is statistically independent. Thus, receiving both polarizations using a dual-polarized antenna, and processing the signals separately, offers diversity. This diversity can be obtained without any requirement for a minimum distance between antenna elements.

Let us now consider more closely the situation where the transmit signal is vertically polarized, while the signal is received in both vertical and horizontal polarization. In that case, fading of the two received signals is independent, but the average received signal strength in the two diversity branches is *not* identical. Depending on the environment, the horizontal (i.e., cross-polarized) component is some 3–20 dB weaker than the vertical (co-polarized) component. As we will see later on, this has an important impact on the effectiveness of the diversity scheme. Various antenna arrangements have been proposed in order to mitigate this problem.

It has also been claimed that the diversity order that can be achieved with polarization diversity is up to 6: three possible components of the E-field and three components of the H-field can all be exploited [Andrews et al. 2001].⁷ However, propagation characteristics as well as practical considerations prevent a full exploitation of that diversity order especially for outdoor situations. This is usually not a serious restriction for diversity systems, as we will see later on that going from diversity order 1 (i.e., no diversity) to diversity order 2 gives larger benefits than increasing the diversity order from 2 to higher values. However, it is an important issue for Multiple Input Multiple Output (MIMO) systems (see Section 20.2).

13.3 Macrodiversity and Simulcast

The previous section described diversity methods that combat small-scale fading – i.e., the fading created by interference of MPCs. However, not all of these diversity methods are suitable for combating large-scale fading, which is created by shadowing effects. Shadowing is almost independent of transmit frequency and polarization, so that frequency diversity or polarization diversity are not effective. Spatial diversity (or equivalently, temporal diversity with moving TX/RX) can be used, but we have to keep in mind that the correlation distances for large-scale fading are on the order of tens or hundreds of meters. In other words, if there is a hill between the TX and RX, adding antennas on either the BS or the MS does not help to eliminate the shadowing caused by this hill. Rather, we should use a separate base station (BS2) that is placed in such a way that the hill is not in the connection line between the MS and BS2. This in turn implies a large distance between BS1 and BS2, which gives rise to the word *macrodiversity*.

The simplest method for macrodiversity is the use of *on-frequency repeaters* that receive the signal and retransmit an amplified version of it. *Simulcast* is very similar to this approach; the same signal is transmitted simultaneously from different BSs. In cellular applications the two BSs should be synchronized, and transmit the signals intended for a specific user in such a way that the two waves arrive at the RX almost simultaneously (timing advance).⁸ Note that synchronization

⁶ For simplicity, we henceforth speak of horizontal and vertical polarization. However, the considerations are valid for any two orthogonal polarizations.

⁷ Note that this result is surrounded by some controversy, and (at the time of this writing) there is no general agreement in the scientific community whether diversity order 6 can really be achieved from polarization only.

⁸ If the RX cannot easily deal with delay dispersion, it is desirable that the two signals arrive exactly at the same time. For advanced RXs (Chapters 16–19) it can be desirable to have a small amount of delay dispersion; if the timing of the BSs and the runtimes of the signals to the RX are exactly known, this can also be achieved.

can only be obtained if the runtimes from the two BSs to the MS are known. Generally speaking, it is desirable that the synchronization error is no larger than the delay dispersion that the RX can handle. Especially critical are RXs in regions where the strengths of the signals from the two BSs are approximately equal.

Simulcast is also widely used for broadcast applications, especially digital TV. In this case, the exact synchronization of all possible RXs is not possible – each RX would require a different timing advance from the TXs.

A disadvantage of simulcast is the large amount of signaling information that has to be carried on landlines. Synchronization information as well as transmit data have to be transported on landlines (or microwave links) to the BSs. This used to be a serious problem in the early days of digital mobile telephony, but the current wide availability of fiber-optic links has made this less of an issue.

The use of on-frequency repeaters is simpler than that of simulcast, as no synchronization is required. On the other hand, delay dispersion is larger, because (i) the runtime from BS to repeater, and repeater to MS is larger (compared with the runtime from a second BS), and (ii) the repeater itself introduces additional delays due to the group delays of electronic components, filters, etc.

13.4 Combination of Signals

Now we turn our attention to the question of how to use diversity signals in a way that improves the total quality of the signal that is to be detected. To simplify the notation, we speak here only about the combination of signals from different antenna signals at the RX. However, the mathematical methods remain valid for other types of diversity signals as well. In general, we can distinguish two ways of exploiting signals from the multiple diversity branches:

1. *Selection diversity*, where the “best” signal copy is selected and processed (demodulated and decoded), while all other copies are discarded. There are different criteria for what constitutes the “best” signal.
2. *Combining diversity*, where all copies of the signal are combined (before or after the demodulator), and the combined signal is decoded. Again, there are different algorithms for combination of the signals.

Combining diversity leads to better performance, as all available information is exploited. On the downside, it requires a more complex RX than selection diversity. In most RXs, all processing is done in the baseband. Thus, an RX with combining diversity needs to downconvert all available signals, and combine them appropriately in the baseband. Thus, it requires N_r antenna elements as well as N_r complete Radio Frequency (RF) (downconversion) chains. An RX with selection diversity requires only *one* RF chain, as it processes only a single received signal at a time.

In the following, we give a more detailed description of selection (combination) criteria and algorithms. We assume that different signal copies undergo statistically independent fading – this greatly simplifies the discussion of both the intuitive explanations and the mathematics of the signal combination. Discussions on the impact of a finite correlation coefficient are relegated to Section 13.5.

In these considerations, we also have to keep in mind that the gain of multiple antennas is due to two effects: *diversity gain* and *beamforming gain*. Diversity gain reflects the fact that it is improbable that several antenna elements are in a fading dip simultaneously; the probability for very low signal levels is thus decreased by the use of multiple antenna elements. Beamforming gain reflects the fact that (for combining diversity) the combiner performs an averaging over the noise at different antennas. Thus, even if the signal levels at all antenna elements are identical, the combiner output SNR is larger than the SNR at a single-antenna element.

13.4.1 Selection Diversity

Received-Signal-Strength-Indication-Driven Diversity

In this method, the RX selects the signal with the largest instantaneous power (or *Received Signal Strength Indication* – *RSSI*), and processes it further. This method requires N_r antenna elements, N_r RSSI sensors, and a N_r -to-1 multiplexer (switch), but only one RF chain (see Figure 13.8). The method allows simple tracking of the selection criterion even in fast-fading channels. Thus, we can switch to a better antenna as soon as the RSSI becomes higher there.

- 1. If the BER is determined by noise, then RSSI-driven diversity is the best of all the selection diversity methods, as maximization of the RSSI also maximizes the SNR.
- 2. If the BER is determined by co-channel interference, then RSSI is no longer a good selection criterion. High receive power can be caused by a high level of interference, such that the RSSI criterion makes the system select branches with a low signal-to-interference ratio. This is especially critical when interference is caused mainly by one dominant interferer – a situation that is typical for Frequency Division Multiple Access (FDMA) or TDMA systems.
- 3. Similarly, RSSI-driven diversity is suboptimum if the errors are caused by the frequency selectivity of the channel. RSSI-driven diversity can still be a reasonable approximation, because

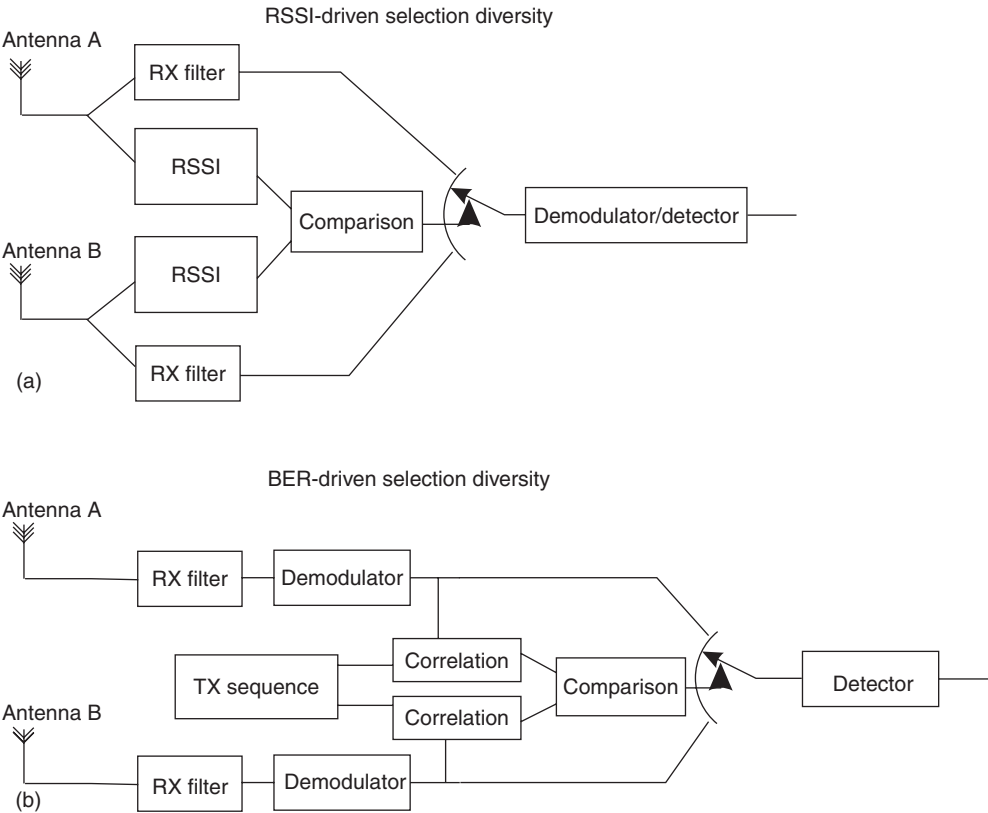


Figure 13.8 Selection diversity principle: (a) Received-signal-strength-indication-controlled diversity. (b) Bit-error-rate-controlled diversity.

we have shown in Chapter 12 that errors caused by signal distortion occur mainly in the fading dips of the channel. However, this is only an approximation, and it can be shown that (uncoded, unequalized) systems with RSSI-driven selection diversity have a BER that is higher by a constant factor compared with optimum (BER-driven) diversity.

For an exact performance assessment (Section 13.5), it is important to obtain the SNR distribution of the output of the selector. Assume that the instantaneous signal amplitude is Rayleigh distributed, such that the SNR of the n th diversity branch, γ_n , is (see Eq. 12.48)

$$\text{pdf}_{\gamma_n}(\gamma_n) = \frac{1}{\bar{\gamma}} \exp\left(-\frac{\gamma_n}{\bar{\gamma}}\right) \quad (13.8)$$

where $\bar{\gamma}$ is the mean branch SNR (assumed to be identical for all diversity branches). The cumulative distribution function (cdf) is then

$$\text{cdf}_{\gamma_n}(\gamma_n) = 1 - \exp\left(-\frac{\gamma_n}{\bar{\gamma}}\right) \quad (13.9)$$

The cdf is, by definition, the probability that the instantaneous SNR lies below a given level. As the RX selects the branch with the largest SNR, the probability that the chosen signal lies below the threshold is the product of the probabilities that the SNR at each branch is below the threshold. In other words, the cdf of the selected signal is the product of the cdfs of each branch:

$$\text{cdf}_{\gamma}(\gamma) = \left[1 - \exp\left(-\frac{\gamma}{\bar{\gamma}}\right)\right]^{N_r} \quad (13.10)$$

Example 13.3 Compute the probability that the output power of a selection diversity system is 5 dB lower than the mean power of each branch, when using $N_r = 1, 2, 4$ antennas.

The threshold is $\gamma_{\text{dB}} = \bar{\gamma}_{\text{dB}} - 5 \text{ dB}$, which in linear scale is $\gamma = \bar{\gamma} \cdot 10^{-0.5}$. Using Eq. (13.10), the probability that the output power is less than $\bar{\gamma} \cdot 10^{-0.5}$ becomes:

$$\begin{aligned} \text{cdf}_{\gamma}(\bar{\gamma} \cdot 10^{-0.5}) &= [1 - \exp(-10^{-0.5})]^{N_r} \\ &= \begin{cases} 0.27, & N_r = 1 \\ 7.4 \cdot 10^{-2}, & N_r = 2 \\ 5.4 \cdot 10^{-3}, & N_r = 4 \end{cases} \end{aligned} \quad (13.11)$$

Example 13.4 Consider now the case that $N_r = 2$, and that the mean powers in the branches are $1.5\bar{\gamma}$ and $0.5\bar{\gamma}$, respectively. How does the result change?

In this case, the probability is

$$\text{cdf}_{\gamma}(\bar{\gamma} \cdot 10^{-0.5}) = \left[1 - \exp\left(-\frac{1}{1.5} \cdot 10^{-0.5}\right)\right] \left[1 - \exp\left(-\frac{1}{0.5} \cdot 10^{-0.5}\right)\right] \quad (13.12)$$

$$= 8.9 \cdot 10^{-2} \quad (13.13)$$

This demonstrates that diversity is less efficient when the average branch powers are different.

Bit-Error-Rate-Driven Diversity

For BER-driven diversity, we first transmit a *training sequence* – i.e., a bit sequence that is known at the RX. The RX then demodulates the signal from each receive antenna element and compares it with the transmit signal. The antenna whose associated signal results in the smallest BER is judged to be the “best,” and used for the subsequent reception of data signals. A similar approach is the use of the mean square error of the “soft-decision” demodulated signal, or the correlation between transmit and receive signal.

If the channel is time variant, the training sequence has to be repeated at regular intervals and selection of the best antenna has to be done anew. The necessary repetition rate depends on the coherence time of the channel.

BER-driven diversity has several drawbacks:

1. The RX needs either N_r RF chains and demodulators (which makes the RX more complex), or the training sequence has to be repeated N_r times (which decreases spectral efficiency), so that the signal at all antenna elements can be evaluated.
2. If the RX has only one demodulator, then it is not possible to continuously monitor the selection criterion (i.e., the BER) of all diversity branches. This is especially critical if the channel changes quickly.
3. Since the duration of the training sequence is finite, the selection criterion – i.e., bit error probability – cannot be determined exactly. The variance of the BER around its true mean decreases as the duration of the training sequence increases. There is thus a tradeoff between performance loss due to erroneous determination of the selection criterion, and spectral efficiency loss due to longer training sequences.

13.4.2 Switched Diversity

The main drawback of selection diversity is that the selection criteria (power, BER, etc.) of *all* diversity branches have to be monitored in order to know when to select a different antenna. As we have shown above, this leads to either increased hardware effort or reduced spectral efficiency. An alternative solution, which avoids these drawbacks, is *switched diversity*. In this method, the selection criterion of just the active diversity branch is monitored. If it falls below a certain threshold, then the RX switches to a different antenna.⁹ Switching only depends on the quality of the active diversity branch; it does not matter whether the other branch actually provides a better signal quality or not.

Switched diversity runs into problems when both branches have signal quality below the threshold: in that case, the RX just switches back and forth between the branches. This problem can be avoided by introducing a hysteresis or hold time, so that the new diversity branch is used for a certain amount of time, independent of the actual signal quality. We thus have two free parameters: switching threshold and hysteresis time. These parameters have to be selected very carefully: if the threshold is chosen too low, then a diversity branch is used even when the other antenna might offer better quality; if it is chosen too high, then it becomes probable that the branch the RX switches to actually offers lower signal quality than the currently active one. If hysteresis time is chosen too long, then a “bad” diversity branch can be used for a long time; if it is chosen too short, then the RX spends all the time switching between two antennas.

Summarizing, the performance of switched diversity is worse than that of selection diversity; we will therefore not consider it further.

⁹ The method is mostly applied when two diversity branches are available.

13.4.3 Combining Diversity

Basic Principle

Selection diversity wastes signal energy by discarding $(N_r - 1)$ copies of the received signal. This drawback is avoided by combining diversity, which exploits *all* available signal copies. Each signal copy is multiplied by a (complex) weight and then added up. Each complex weight w_n^* can be thought of as consisting of a phase correction,¹⁰ plus a (real) weight for the amplitude:

- Phase correction causes the *signal amplitudes* to add up, while, on the other hand, noise is added incoherently, so that *noise powers* add up.
- For amplitude weighting, two methods are widely used: *Maximum Ratio Combining* (MRC) weighs all signal copies by their amplitude. It can be shown that (using some assumptions) this is an optimum combination strategy. An alternative is *Equal Gain Combining* (EGC), where all amplitude weights are the same (in other words, there is no weighting, but just a phase correction). The two methods are outlined in Figure 13.9.

Maximum Ratio Combining

MRC compensates for the phases, and weights the signals from the different antenna branches according to their SNR. This is the optimum way of combining different diversity branches – if several assumptions are fulfilled. Let us assume a propagation channel that is slow fading and flat fading. The only disturbance is AWGN. Under these assumptions, each channel realization can be written as a time-invariant filter with impulse response:

$$h_n(\tau) = \alpha_n \delta(\tau) \quad (13.14)$$

where α_n is the (instantaneous) gain of diversity branch n . This signals at the different branches are multiplied with weights w_n^* and added up, so that the SNR becomes

$$\frac{\left| \sum_{n=1}^N w_n^* \alpha_n \right|^2}{P_n \sum_{n=1}^N |w_n|^2} \quad (13.15)$$

where P_n is the noise power per branch (assumed to be the same in each branch). According to the Cauchy–Schwartz inequality, $\left| \sum_{n=1}^N w_n^* \alpha_n \right|^2 \leq \sum_{n=1}^N |w_n^*|^2 \sum_{n=1}^N |\alpha_n|^2$, where equality holds if and only if $w_n = \alpha_n$. Thus, the SNR is maximized by choosing the weights as

$$w_{\text{MRC}} = \alpha_n \quad (13.16)$$

i.e., the signals are phase-corrected (remember that the received signals are multiplied with w^*) and weighted by the amplitude. We can then easily see that in that case the output SNR of the diversity combiner is the *sum* of the branch SNRs:

$$\gamma_{\text{MRC}} = \sum_{n=1}^{N_r} \gamma_n \quad (13.17)$$

¹⁰ Note that in our notation the signals are weighted with the *complex conjugate* of w ; this simplifies notation later on.

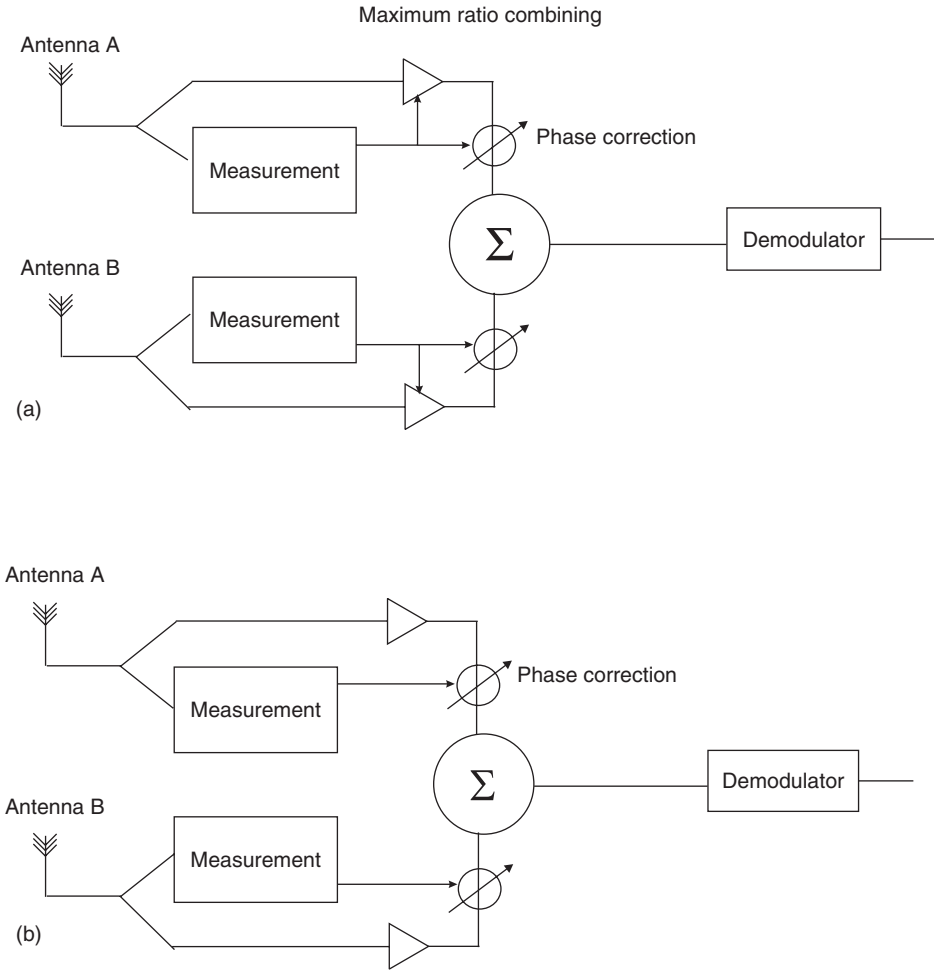


Figure 13.9 Combining diversity principle: (a) maximum ratio combining, (b) equal gain combining.

If the branches are statistically independent, then the moment-generating function of the total SNR can be computed as the product of the characteristic functions of the branch SNRs. If, furthermore, the SNR distribution in each branch is exponential (corresponding to Rayleigh fading), and all branches have the same mean SNR $\overline{\gamma}_n = \overline{\gamma}$, we find after some manipulations that

$$pdf_{\gamma}(\gamma) = \frac{1}{(N_r - 1)!} \frac{\gamma^{N_r - 1}}{\overline{\gamma}^{N_r}} \exp\left(-\frac{\gamma}{\overline{\gamma}}\right) \quad (13.18)$$

and the mean SNR of the combiner output is just the mean branch SNR, multiplied by the number of diversity branches:

$$\overline{\gamma}_{\text{MRC}} = N_r \overline{\gamma} \quad (13.19)$$

Figure 13.10 compares the statistics of the SNR for RSSI-driven selection diversity and MRC. Naturally, there is no difference between the diversity types for $N_r = 1$, since there is no diversity.

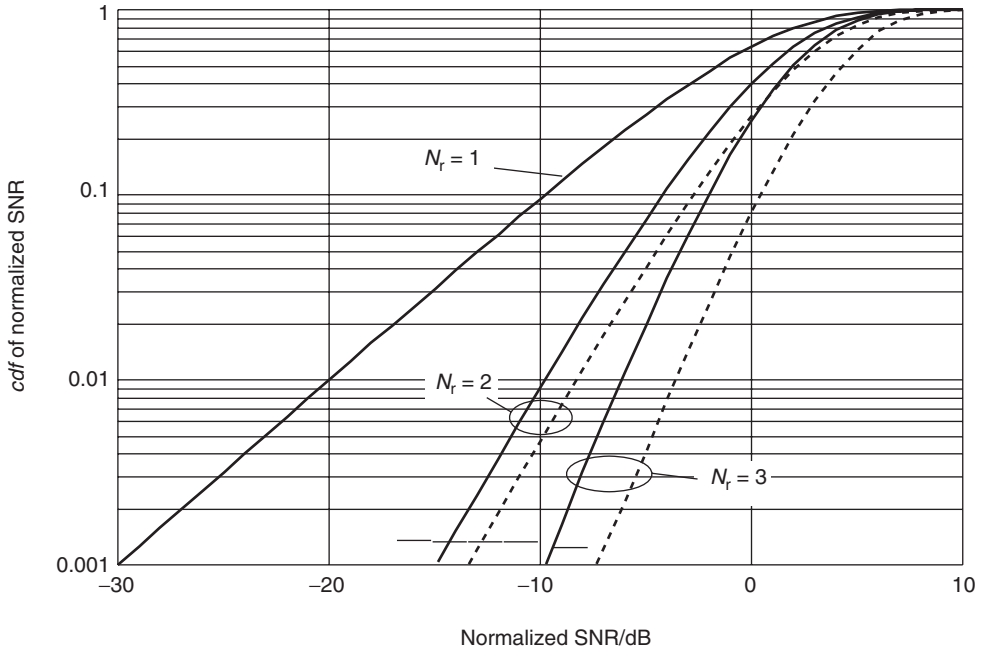


Figure 13.10 Cumulative distribution function of the normalized instantaneous signal-to-noise ratio $\gamma/\bar{\gamma}$ for received-signal-strength-indication-driven selection diversity (solid), and maximum ratio combining (dashed) for $N_r = 1, 2, 3$. Note that for $N_r = 1$, there is no difference between diversity types.

We furthermore see that the slope of the distribution is the same for MRC and selection diversity, but that the difference in the mean values increases with increasing N_r . This is intuitively clear, as selection diversity discards $N_r - 1$ signal copies – something that increases with N_r . For $N_r = 3$, the difference between the two types of diversity is only about 2 dB.

Equal Gain Combining

For EGC, we find that the SNR of the combiner output is

$$\gamma_{\text{EGC}} = \frac{\left(\sum_{n=1}^{N_r} \sqrt{\gamma_n} \right)^2}{N_r} \quad (13.20)$$

where we have assumed that noise levels are the same on all diversity branches. The mean SNR of the combiner output can be found to be

$$\bar{\gamma}_{\text{EGC}} = \bar{\gamma} \left(1 + (N_r - 1) \frac{\pi}{4} \right) \quad (13.21)$$

if all branches suffer from Rayleigh fading with the same mean SNR $\bar{\gamma}$. Remember that we only assume here that the *mean* SNR is the same in all branches, while instantaneous branch SNRs (representing different channel realizations) can be different. It is quite remarkable that EGC performs worse than MRC by only a factor $\pi/4$ (in terms of mean SNR). The performance difference between EGC and MRC becomes bigger when mean branch SNRs are also different.

Equations for the pdf of the SNR can be computed from Eq. (13.20), but become unwieldy very quickly (results can be found, e.g., in Lee [1982]).

Optimum Combining

One of the assumptions in the derivation of MRC was that only AWGN disturbs the signal. If it is interference that determines signal quality, then MRC is no longer the best solution. In order to maximize the Signal-to-Interference-and-Noise Ratio (SINR), the weights should then be determined according to a strategy called *optimum combining*, first derived in the groundbreaking paper of Winters [1984]. The first step is the determination of the correlation matrix of noise and interference at the different antenna elements:

$$\mathbf{R} = \sigma_n^2 \mathbf{I} + \sum_{k=1}^K E\{\mathbf{r}_k \mathbf{r}_k^\dagger\} \quad (13.22)$$

where expectation is over a time period when the channel remains constant, and $\mathbf{r}_k$ is the receive signal vector of the k th interferer. We furthermore need the complex transfer function (complex gain, since we assume flat-fading channels) of the N_r diversity branches; these are written into the vector $\mathbf{h}_d$. The vector containing the optimum receive weights is then

$$\mathbf{w}_{\text{opt}} = \mathbf{R}^{-1} \mathbf{h}_d \quad (13.23)$$

These weights have to be adjusted as the channel changes. It is easy to see that for a noise-limited system the correlation matrix becomes a (scaled) identity matrix, and optimum combining reduces to MRC.

A further interesting observation is that minimizing the Mean Square Error (MSE) is equivalent to maximizing the SINR, which can be written as

$$\text{SINR} = ((\mathbf{w}^\dagger \mathbf{h}_d \mathbf{h}_d^\dagger \mathbf{w}) / (\mathbf{w}^\dagger \mathbf{R} \mathbf{w})) \quad (13.24)$$

This SINR is a generalized Rayleigh quotient, and is maximized by the generalized eigenvalue corresponding to the maximum generalized eigenvector of

$$\mathbf{h}_d \mathbf{h}_d^\dagger \mathbf{w} = \lambda \mathbf{R} \mathbf{w} \quad (13.25)$$

The weight vector obtained from this generalized eigenvalue problem is the same as $\mathbf{w}_{\text{opt}}$ in Eq. (13.22).

Optimum combining of signals from N_r diversity branches gives N_r degrees of freedom. This allows interference from $N_r - 1$ interferers to be eliminated. Alternatively, $N_s \leq N_r - 1$ interferers can be eliminated, while the remaining $N_r - N_s$ antennas behave like “normal” diversity antennas that can be used for noise reduction. This seemingly simple statement has great impact on wireless system design! As we discussed in Chapter 3, many wireless systems are interference limited. The possibility of eliminating at least some of the interferers by appropriate diversity combining opens up the possibility of drastically improving the link quality of such systems, or of increasing their capacity (more details can be found in Section 20.1).

Example 13.5 Consider a desired signal with Binary Phase Shift Keying (BPSK) propagating through a frequency-flat channel with $\mathbf{h}_d = [1, 0.5 + 0.7j]^T$ and a synchronous, interfering BPSK signal with $\mathbf{h}_{\text{int}} = [0.3, -0.2 + 1.7j]^T$. The noise variance is $\sigma_n^2 = 0.01$. Show that weighting the received signal with the weights as computed from Eq. (13.23) leads to mitigation of interference.

Let the desired transmitted signal be $s_d(t)$, the interfering transmitted signal be $s_{\text{int}}(t)$, and $n_1(t)$ and $n_2(t)$ be independent zero-mean noise processes. It is assumed that desired and interfering signals are uncorrelated, and that $E\{s_d s_d^*\} = 1$, and also $E\{s_{\text{int}} s_{\text{int}}^*\} = 1$.

The noise-plus-interference correlation matrix is computed according to Eq. (13.22):

$$\mathbf{R} = \sigma_n^2 \mathbf{I} + \sum_{k=1}^K E\{\mathbf{r}_k \mathbf{r}_k^*\} \quad (13.26)$$

$$= 0.01 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.3 \\ -0.2 + 1.7j \end{pmatrix} \begin{pmatrix} 0.3 & -0.2 - 1.7j \end{pmatrix} \quad (13.27)$$

$$= \begin{pmatrix} 0.1 & -0.06 - 0.51j \\ -0.06 + 0.51j & 2.94 \end{pmatrix} \quad (13.28)$$

Using Eq. (13.23) and normalizing the weights, we obtain:

$$\mathbf{w} = \frac{\mathbf{R}^{-1} \mathbf{h}_d}{|\mathbf{R}^{-1} \mathbf{h}_d|} \quad (13.29)$$

$$= \begin{pmatrix} 0.979 + 0.111j \\ 0.041 - 0.165j \end{pmatrix} \quad (13.30)$$

Inserting this value for $\mathbf{w}$ into Eq. (13.24), we obtain an SINR of 78. This can be compared to the SNR = 100, which could be obtained if the receiver would not have to suppress an interferer, and the SINR of 0.6 that is available at a single antenna.

Hybrid Selection – Maximum Ratio Combining

A compromise between selection diversity and full signal combining is the so-called hybrid selection scheme, where the best L out of N_r antenna signals are chosen, downconverted, and processed. This reduces the number of required RF chains from N_r to L , and thus leads to significant savings. The savings come at the price of a (usually small) performance loss compared with the full-complexity system. The approach is called *Hybrid Selection/Maximum Ratio Combining* (H-S/MRC), or sometimes also *Generalized Selection Combining* (GSC).

It is well known that the output SNR of MRC is just the sum of the SNRs at the different receive antenna elements. For H-S/MRC, the instantaneous output SNR of H-S/MRC looks deceptively similar to MRC – namely:

$$\gamma_{\text{H-S/MRC}} = \sum_{n=1}^L \gamma_{(n)} \quad (13.31)$$

The major difference from MRC is that the $\gamma_{(n)}$ are *ordered* SNRs – i.e., $\gamma_{(1)} > \gamma_{(2)} > \dots > \gamma_{(N_r)}$. This leads to different performance, and poses new mathematical challenges for performance analysis. Specifically, we have to introduce the concept of *order statistics*. Note that selection diversity (where just one of N_r antennas is selected) and MRC are limiting cases of H-S/MRC with $L = 1$ and $L = N_r$, respectively.

H-S/MRC schemes provide good diversity gain, as they select the best antenna branches for combining. Actually, it can be shown that the diversity *order* obtained with such schemes is proportional to N_r , not to L . However, they do not provide full beamforming gain. If the signals at

all antenna elements are completely correlated, then the SNR gain of H-S/MRC is only L , compared with N_r for an MRC scheme.

The analysis of H-S/MRC based on a chosen ordering of the branches at first appears to be complicated, since the SNR statistics of the ordered branches are *not* independent. However, we can alleviate this problem by transforming ordered branch variables into a new set of random variables. It is possible to find a transformation that leads to *independently distributed* random variables (termed *virtual branch variables*). The fact that the combiner output SNR can be expressed in terms of independent identically distributed (iid) virtual branch variables enormously simplifies performance analysis of the system. For example, the derivation of Symbol Error Probability (SEP) for uncoded H-S/MRC systems, which normally would require evaluation of nested N -fold integrals, essentially reduces to evaluation of a *single* integral with finite limits.

The mean and variance of the output SNR for H-S/MRC is thus

$$\bar{\gamma}_{\text{H-S/MRC}} = L \left(1 + \sum_{n=L+1}^{N_r} \frac{1}{n} \right) \bar{\gamma} \quad (13.32)$$

and

$$\sigma_{\text{H-S/MRC}}^2 = L \left(1 + L \sum_{n=L+1}^{N_r} \frac{1}{n^2} \right) \bar{\gamma}^2 \quad (13.33)$$

13.5 Error Probability in Fading Channels with Diversity Reception

In this section we determine the Symbol Error Rate (SER) in fading channels when diversity is used at the RX. We start with the case of flat-fading channels, computing the statistics of the received power and the BER. We then proceed to dispersive channels, where we analyze how diversity can mitigate the detrimental effects of dispersive channels on simple RXs.

13.5.1 Error Probability in Flat-Fading Channels

Classical Computation Method

Analogous to Chapter 12, we can compute the error probability of diversity systems by averaging the conditional error probability (conditioned on a certain SNR) over the distribution of the SNR:

$$\overline{\text{SER}} = \int_0^\infty \text{pdf}_\gamma(\gamma) \text{SER}(\gamma) d\gamma \quad (13.34)$$

As an example, let us compute the performance of BPSK with N_r diversity branches with MRC. The SER of BPSK in AWGN is (see Chapter 12)

$$\text{SER}(\gamma) = Q(\sqrt{2\gamma}) \quad (13.35)$$

Let us apply this principle to the case of MRC. When inserting Eqs. (13.18) and (13.35) into Eq. (13.34), we obtain an equation that can be evaluated analytically:

$$\overline{\text{SER}} = \left(\frac{1-b}{2} \right)^{N_r} \sum_{n=0}^{N_r-1} \binom{N_r-1+n}{n} \left(\frac{1+b}{2} \right)^n \quad (13.36)$$

where b is defined as

$$b = \sqrt{\frac{\bar{\gamma}}{1 + \bar{\gamma}}} \quad (13.37)$$

For large values of $\bar{\gamma}$, this can be approximated as

$$\overline{SER} = \left(\frac{1}{4\bar{\gamma}}\right)^{N_r} \binom{2N_r - 1}{N_r} \quad (13.38)$$

From this, we can see that (with N_r diversity antennas) the BER decreases with the N_r -th power of the SNR.

Computation via the Moment-Generating Function

In the previous section, we averaged the BER over the distribution of SNRs, using the “classical” representation of the Q-function. As we have already seen (Chapter 12), there is an alternative definition of the Q-function, which can easily be combined with the moment-generating function $M_\gamma(s)$ of the SNR. Let us start by writing the SER conditioned on a given SNR in the form (see Chapter 12):

$$SER(\gamma) = \int_{\theta_1}^{\theta_2} f_1(\theta) \exp(-\gamma_{MRC} f_2(\theta)) d\theta \quad (13.39)$$

Since

$$\gamma_{MRC} = \sum_{n=1}^{N_r} \gamma_n \quad (13.40)$$

this can be rewritten as

$$SER(\gamma) = \int_{\theta_1}^{\theta_2} f_1(\theta) \prod_{n=1}^{N_r} \exp(-\gamma_n f_2(\theta)) d\theta \quad (13.41)$$

Averaging over the SNRs in different branches then becomes

$$\overline{SER} = \int d\gamma_1 pdf_{\gamma_1}(\gamma_1) \int d\gamma_2 pdf_{\gamma_2}(\gamma_2) \cdots \int d\gamma_{N_r} pdf_{\gamma_{N_r}}(\gamma_{N_r}) \int_{\theta_1}^{\theta_2} d\theta f_1(\theta) \prod_{n=1}^{N_r} \exp(-\gamma_n f_2(\theta)) \quad (13.42)$$

$$= \int_{\theta_1}^{\theta_2} d\theta f_1(\theta) \prod_{n=1}^{N_r} \int d\gamma_n pdf_{\gamma_n}(\gamma_n) \exp(-\gamma_n f_2(\theta)) \quad (13.43)$$

$$= \int_{\theta_1}^{\theta_2} d\theta f_1(\theta) \prod_{n=1}^{N_r} M_\gamma(-f_2(\theta)) \quad (13.44)$$

$$= \int_{\theta_1}^{\theta_2} d\theta f_1(\theta) [M_\gamma(-f_2(\theta))]^{N_r} \quad (13.45)$$

With that, we can write the error probability for BPSK in Rayleigh fading as

$$\overline{SER} = \frac{1}{\pi} \int_0^{\pi/2} \left[\frac{\sin^2(\theta)}{\sin^2(\theta) + \bar{\gamma}} \right]^{N_r} d\theta \quad (13.46)$$

Example 13.6 Compare the symbol error probability in Rayleigh fading of 8-PSK (Phase Shift Keying) and four available antennas with 10-dB SNRs when using $L = 1, 2, 4$ receive chains.

The SER for M-ary Phase Shift Keying (MPSK) with H-S/MRC can be shown to be

$$\overline{SER}_{e,H-S/MRC}^{\text{MPSK}} = \frac{1}{\pi} \int_0^{\pi(M-1)/M} \left[\frac{\sin^2 \theta}{\sin^2(\pi/M) \overline{\gamma} + \sin^2 \theta} \right]^L \prod_{n=L+1}^{N_r} \left[\frac{\sin^2 \theta}{\sin^2(\pi/M) \overline{\gamma}_n^L + \sin^2 \theta} \right] d\theta \quad (13.47)$$

Evaluating Eq. (13.47) with $M = 8$, $\overline{\gamma} = 10$ dB, $N_r = 4$, and $L = 1, 2, 4$ yields:

L	$\overline{SER}_{e,H-S/MRC}^{\text{MPSK}}$
1	0.0442
2	0.0168
4	0.0090

13.5.2 Symbol Error Rate in Frequency-Selective Fading Channels

We now determine the SER in channels that suffer from time dispersion and frequency dispersion. We assume here FSK with differential phase detection. The analysis uses the correlation coefficient ρ_{XY} between signals at two sampling times that was discussed in Chapter 12.

For binary FSK with selection diversity:

$$\overline{SER} = \frac{1}{2} - \frac{1}{2} \sum_{n=1}^{N_r} \binom{N_r}{n} (-1)^{n+1} \frac{b_0 \text{Im}\{\rho_{XY}\}}{\sqrt{(\text{Im}\{\rho_{XY}\})^2 + n(1 - |\rho_{XY}|^2)}} \quad (13.48)$$

where b_0 is the transmitted bit. This can be approximated as

$$\overline{SER} = \frac{(2N_r - 1)!!}{2} \left(\frac{1 - |\rho_{XY}|^2}{2(\text{Im}\{\rho_{XY}\})^2} \right)^{N_r} \quad (13.49)$$

where $(2N_r - 1)!! = 1 \cdot 3 \cdot 5 \dots (2N_r - 1)$.

For binary FSK with MRC:

$$\overline{SER} = \frac{1}{2} - \frac{1}{2} \frac{b_0 \text{Im}\{\rho_{XY}\}}{\sqrt{1 - (\text{Re}\{\rho_{XY}\})^2}} \sum_{n=0}^{N_r-1} \frac{(2n-1)!!}{(2n)!!} \left(1 - \frac{(\text{Im}\{\rho_{XY}\})^2}{1 - (\text{Re}\{\rho_{XY}\})^2} \right)^n \quad (13.50)$$

which can be approximated as

$$\overline{SER} = \frac{(2N_r - 1)!!}{2(N_r)!} \left(\frac{1 - |\rho_{XY}|^2}{2(\text{Im}\{\rho_{XY}\})^2} \right)^{N_r} \quad (13.51)$$

This formulation shows that MRC improves the SER by a factor $N_r!$ compared with selection diversity. A further important consequence is that the errors due to delay dispersion and random Frequency Modulation (FM) are decreased in the same way as errors due to noise. This is shown by the expressions in parentheses that are taken to the N_r -th power. These terms subsume the errors due to all different effects. The SER with diversity is approximately the N_r -th power of the SER without diversity (see Figures 13.11–13.13).

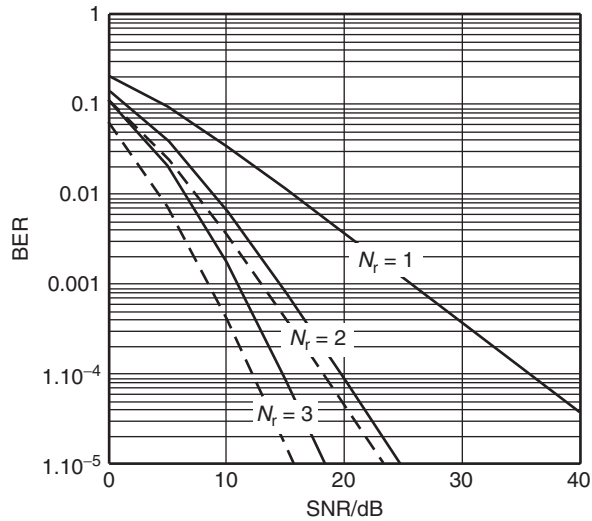


Figure 13.11 Bit error rate of minimum shift keying (MSK) with received-signal-strength-indication-driven selection diversity (solid) and maximum ratio combining (dashed) as a function of the signal-to-noise ratio with N_r diversity antennas.

Reproduced with permission from Molisch [2000] © Prentice Hall.

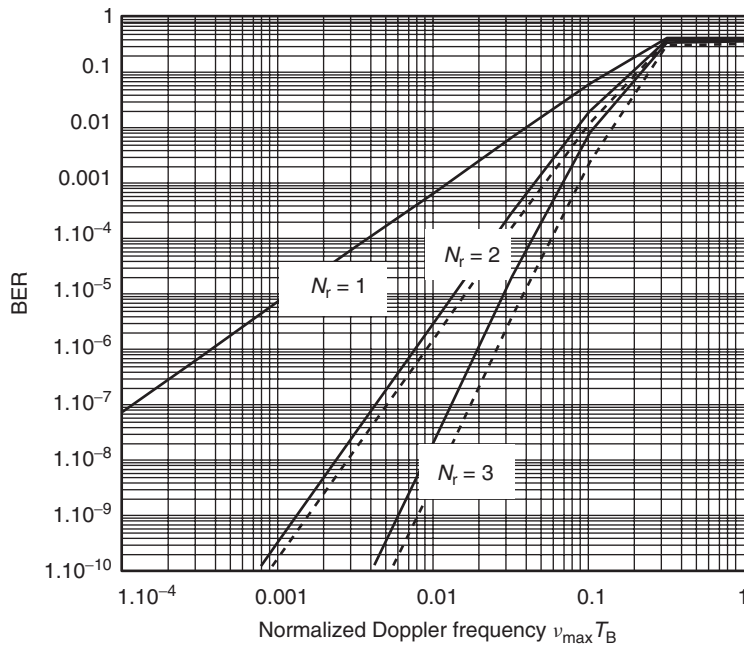


Figure 13.12 Bit error rate of MSK with received-signal-strength-indication-driven selection diversity (solid) and maximum ratio combining (dashed) as a function of the normalized Doppler frequency with N_r diversity antennas.

Reproduced with permission from Molisch [2000] © Prentice Hall.

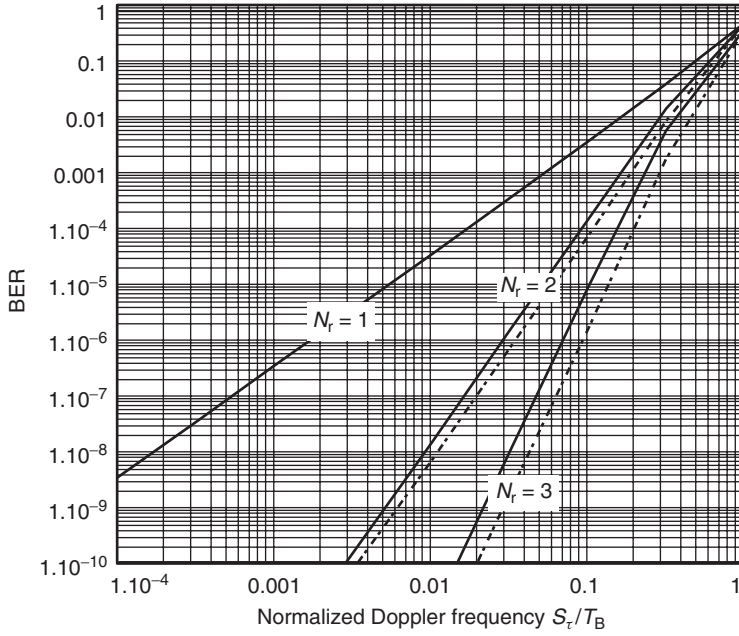


Figure 13.13 Bit error rate of MSK with received-signal-strength-indication-driven selection diversity (solid) and maximum ratio combining (dashed) as a function of the normalized rms delay spread with N_r diversity antennas. Reproduced with permission from Molisch [2000] © Prentice Hall.

For Differential Quadrature-Phase Shift Keying (DQPSK) with selection diversity, the average BER is

$$\overline{BER} = \frac{1}{2} - \frac{1}{4} \sum_{n=1}^{N_r} \binom{N_r}{n} (-1)^{n+1} \left[\frac{b_0 \text{Re}\{\rho_{XY}\}}{\sqrt{(\text{Re}\{\rho_{XY}\})^2 + n(1 - |\rho_{XY}|^2)}} + \frac{b'_0 \text{Im}\{\rho_{XY}\}}{\sqrt{(\text{Im}\{\rho_{XY}\})^2 + n(1 - |\rho_{XY}|^2)}} \right] \quad (13.52)$$

where (b_0, b'_0) is the transmitted symbol.

For DQPSK with MRC:

$$\overline{BER} = \frac{1}{2} - \frac{1}{4} \sum_{n=0}^{N_r-1} \frac{(2n-1)!!}{(2n)!!} \left[\frac{b_0 \text{Re}\{\rho_{XY}\}}{\sqrt{1 - (\text{Im}\{\rho_{XY}\})^2}} \left(\frac{1 - |\rho_{XY}|^2}{1 - (\text{Im}\{\rho_{XY}\})^2} \right)^n + \frac{b'_0 \text{Im}\{\rho_{XY}\}}{\sqrt{1 - (\text{Re}\{\rho_{XY}\})^2}} \left(\frac{1 - |\rho_{XY}|^2}{1 - (\text{Re}\{\rho_{XY}\})^2} \right)^n \right] \quad (13.53)$$

More general cases can be treated with the Quadratic Form Gaussian Variable (QFGV) method (see Section 12.3.3), where Eq. (12.103) is replaced by

$$\begin{aligned}
 P(D < 0) = & Q_M(p_1, p_2) - I_0(p_1 p_2) \exp \left[-\frac{1}{2}(p_1^2 + p_2^2) \right] \\
 & + \frac{I_0(p_1 p_2) \exp \left[-\frac{1}{2}(p_1^2 + p_2^2) \right]}{(1 + v_2/v_1)^{2N_r-1}} \sum_{n=0}^{N_r-1} \binom{2N_r-1}{n} \left(\frac{v_2}{v_1} \right)^n \\
 & + \frac{\exp \left[-\frac{1}{2}(p_1^2 + p_2^2) \right]}{(1 + v_2/v_1)^{2N_r-1}} \cdot \sum_{n=1}^{N_r-1} I_n(p_1 p_2) \sum_{k=0}^{N_r-1-n} \binom{2N_r-1}{k} \\
 & \times \left[\left(\frac{p_2}{p_1} \right)^n \left(\frac{v_2}{v_1} \right)^k - \left(\frac{p_1}{p_2} \right)^n \left(\frac{v_2}{v_1} \right)^{2N_r-1-k} \right] \quad (13.54)
 \end{aligned}$$

RSSI-driven diversity is not the best selection strategy when errors are mostly caused by frequency selectivity and time selectivity. It puts emphasis on signals that have a large amplitude, and not on those with the smallest distortion.¹¹ In these cases, BER-driven selection diversity is preferable. For $N_r = 2$, the BER of Minimum Shift Keying (MSK) with differential detection becomes

$$\overline{BER} \approx \left(\frac{\pi}{4} \right)^4 \left(\frac{S_r}{T_B} \right)^4 \quad (13.55)$$

compared with the RSSI-driven result:

$$\overline{BER} \approx 3 \left(\frac{\pi}{4} \right)^4 \left(\frac{S_r}{T_B} \right)^4 \quad (13.56)$$

13.6 Transmit Diversity

For many situations, multiple antennas can be installed at just one link end (usually the BS). For the uplink transmission from the MS to BS, multiple antennas can act as receive diversity branches. For the downlink, any possible diversity originates at the *transmitter*. In this section, we will thus discuss ways of transmitting signals from several TX antennas and achieve a diversity effect with it. Note that this section specifically refers to *antenna* diversity (which encompasses spatial diversity, pattern diversity, and polarization diversity). Time diversity and frequency diversity inherently involve the TX, and thus need not be discussed again here.

The question of how transmit diversity can be combined with multiple antenna elements at the RX is discussed in Chapter 20.

13.6.1 Transmitter Diversity with Channel State Information

The first situation we analyze is the case where the TX knows the channel perfectly. This knowledge might be obtained from feedback from the RX, or from reciprocity principles; a more detailed discussion of this issue can be found in Section 20.1.6. In this case, we find that (at least for

¹¹ For a similar reason, MRC is not the best combining strategy.

the noise-limited case) there is a complete equivalence between transmit diversity and receive diversity. In other words, the optimum transmission scheme linearly weights signals transmitted from different antenna elements with the complex conjugates of the channel transfer functions from the transmit antenna elements to the single receive antenna. This approach is known as *maximum ratio transmission*.

13.6.2 Transmitter Diversity Without Channel State Information

In many cases, Channel State Information (CSI) is not available at the TX. We then cannot simply transmit weighted copies of the same signal from different transmit antennas, because we cannot know how they would add up at the RX. It is equally likely for the addition of different components to be constructive or destructive; in other words, we would just be adding up MPCs with random phases, which results in Rayleigh fading. We thus cannot gain any diversity (or beamforming).

In order to give benefits, transmission of the signals from different antenna elements has to be done in such a way that it allows the RX to distinguish different transmitted signal components. One way is *delay diversity*. In this scheme, signals transmitted from different antenna elements are delayed copies of the same signal. This makes sure that the effective impulse response is delay dispersive, even if the channel itself is flat fading. So, in a flat-fading channel, we transmit data streams with a delay of 1 symbol duration (relative to preceding antennas) from each of the transmit antennas. The effective impulse response of the channel then becomes

$$h(\tau) = \frac{1}{\sqrt{N_t}} \sum_{n=1}^{N_t} h_n \delta(\tau - nT_s) \quad (13.57)$$

where the h_n are gains from the n th transmit antenna to the receive antenna, and the impulse response has been normalized so that total transmit power is independent of the number of antenna elements. The signals from different transmit antennas to the RX act effectively as delayed MPCs. If antenna elements are spaced sufficiently far apart, these coefficients fade independently. With an appropriate RX for delay-dispersive channels – e.g., an equalizer as described in Chapter 16, or a Rake RX as described in Chapter 18 – we get a diversity order that is equal to the number of antenna elements.

If the channel from a single transmit antenna to the RX is already delay dispersive, then the scheme still works, but care has to be taken in the choice of delays for different antenna elements. The delay between signals transmitted from different antenna elements should be at least as large as the maximum excess delay of the channel.

An alternative method is *phase-sweeping diversity*. In this method, which is especially useful if there are only two antenna elements, the same signal is transmitted from both antenna elements. However, one of the antenna signals undergoes a time-varying phase shift. This means that at the RX the received signals add up in a time-varying way; in other words, we are artificially introducing temporal variations into the channel. The reason for this is that – even if the TX, RX, and the IOs are stationary – the signal does not remain stuck in a fading dip. If this scheme is combined with appropriate coding and/or interleaving, it improves performance.

Yet another possibility for achieving transmit diversity is *space–time coding*. This method is discussed in Chapter 20.

13.7 Appendix

Please go to www.wiley.com/go/molisch

Further Reading

Spatial (antenna) diversity is the oldest form of diversity, and is discussed in textbooks on wireless propagation channels (see, e.g., Vaughan and Anderson [2003]). Evaluations of the antenna correlation coefficient for different angular spectra can be found in Fuhl et al. [1998] and Roy and Fortier [2004]. For antennas on a handset, results are given in Ogawa et al. [2001]. Taga [1993] also shows the effect of mutual coupling on pattern diversity and thus the correlation coefficient, as do a number of recent papers written in the context of MIMO systems [Waldschmidt et al. 2004 and Wallace and Jensen 2004]. Polarization diversity is discussed in more detail, e.g., in Narayanan et al. [2004] and Shafi et al. [2006]; joint spatial, polarization, and pattern (angle) diversity is discussed in Dietrich et al. [2001]. Different combination strategies for antenna signals, and the resulting channel statistics, are discussed in Proakis [2005], and in many papers in the primary literature. The situation is similar to that mentioned in Chapter 12: each type of fading statistics and each combination strategy merits at least one paper. Application of these fading statistics to computation of BERs gives rise to an even greater variety of papers. A nice, unified treatment can be found in Simon and Alouini [2004], who include different fading statistics and antenna combination strategies for a variety of modulation formats. For Rayleigh or Rice fading, the classical QFGV method of Proakis [1968] is applicable; Adachi and Ohno [1991] and Adachi and Parsons [1989] compute the performance for DQPSK and FSK, respectively. The virtual path method for H-S/MRC was introduced in Win and Winters [1999]; Simon and Alouini [2004] also investigate this case. General mathematical aspects of order statistics can be found in David and Nagaraja [2003]. Discussions on the impact of diversity on systems in frequency-selective channels can be found in Molisch [2000, ch. 13]. Maximum ratio transmission was first suggested in Lo [1999]. Delay diversity was suggested by Winters [1994] and Wittneben [1993]. An overview of various transmit diversity techniques can be found in Hottinen et al. [2003].

For updates and errata for this chapter, see wides.usc.edu/teaching/textbook

Multiantenna Systems

Since the 1990s, there has been enormous interest in multiantenna systems. As spectrum became a more and more precious resource, researchers investigated ways of improving the capacity of wireless systems without actually increasing the required spectrum. Multiantenna systems offer such a possibility.

When discussing multiantenna systems, we distinguish between *smart antenna systems* (systems with multiantenna elements at one link end only), and *Multiple Input Multiple Output* (MIMO) systems, which have multiantenna elements at both link ends. Both of these systems are discussed in this chapter.

20.1 Smart Antennas

20.1.1 What are Smart Antennas?

Let us start out with a general definition: (receiver (RX)) smart antennas are “antennas with multiple elements, where signals from different elements are combined by an adaptive (intelligent) algorithm”; for the transmit case, the signals at the antenna elements are *created* by the algorithm.¹

Intelligence (smartness) is not in the antenna, but rather in signal processing. In the simplest case, combination of antenna signals is a linear combination using a weight vector $\mathbf{w}$. The ways of determining $\mathbf{w}$ essentially differentiates smart antenna systems. We see immediately that there is a strong relationship between multiantenna systems and diversity systems – as a matter of fact, an RX with antenna diversity *is* a smart antenna. This chapter thus reuses many results from Chapter 13, while concentrating more on system aspects and the impact smart antennas have on multiple access.

The definition of smart antennas as a combiner of different antenna signals stresses the fact that we exploit signals from different *spatial locations*. Alternatively, we can also say that smart antennas exploit the *directional properties of the channel*. Remember from Chapter 8 and 13 that an RX with multiantenna elements can distinguish Multi Path Components (MPCs) with different Directions Of Arrival (DOAs). Thus, one way to interpret smart antennas is as a spatial Rake² RX that distinguishes between MPCs with different DOAs and processes them separately. This allows the RX to coherently add up the different MPCs, and thus reduce fading; MPCs from interferers can also be suppressed.

¹ In most practical situations, smart antennas are located at the Base Station (BS), and this is also the case we assume in this section, unless otherwise specified.

² A device that enables Code Division Multiple Access (CDMA) systems to deal with a channel’s MPCs (see Section 18.2).

Yet another interpretation is that the smart antenna adaptively forms an antenna array pattern that enhances (coherently combines) the desired MPCs, e.g., by pointing a beam with strong gain in the direction of desired MPCs. Furthermore, the antenna arrays can also form notches in their patterns. The latter is important for suppressing co-channel interference. If there are only few interferers, the ability to suppress interferers makes a bigger contribution to the improvement of the Signal-to-Interference Ratio (SIR) than the enhancement of the desired signal.

20.1.2 Purpose

Smart antennas can be used for various purposes:

1. *Increase of coverage*: assume that the smart antenna is at the RX. Now, if the spatial (angular) position of the transmitter (TX) is known, the RX can form an array pattern in the direction of the TX (beamforming). This results in higher receive power – e.g., an antenna array with eight elements can increase the Signal-to-Noise Ratio (SNR) by 9 dB compared with a single antenna. In a noise-limited cellular system, an improvement in the SNR increases the area that can be covered by one BS. Conversely, the coverage range can be kept constant while the transmit power is decreased.
2. *Increase of capacity*: smart antennas can increase the SIR, e.g., through the use of optimum combining (see Chapter 13); this in turn allows the number of users in the system to be increased. This is the practically most important advantage of smart antennas, and are discussed in more detail below.
3. *Improvement of link quality*: by increasing signal power and/or decreasing interference power, we can also increase the transmission quality on each single link.
4. *Decrease of delay dispersion*: by suppressing MPCs with large delays, delay dispersion can be reduced. This feature can be especially useful in systems with a very high data rate.
5. *Improvement of user position estimation*: knowledge of the DOAs, especially for the (quasi-) Line Of Sight (LOS) component, improves geolocation. This is of value both for *location-based services* and for the ability to locate users in case of emergency.

It is *not* possible to have all these advantages simultaneously to their fullest extent. For example, we can use the capability of smart antennas to reduce interference in order to *either* improve the quality of a single link *or* have more users in the system or to *slightly* increase both the quality per link and the number of users. For system design, the engineer has to decide which aspect is the most important.

20.1.3 Capacity Increase

As mentioned above, increasing the capacity is the most important application of smart antennas. Depending on whether the considered system uses Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), or CDMA, there are different approaches for achieving such capacity gains (Figure 20.1):

1. *Spatial Filtering for Interference Reduction (SFIR)*: this is used in TDMA/FDMA systems to reduce the reuse distance. A conventional TDMA/FDMA system cannot reuse the same frequency in each neighboring cell since the interference from the adjacent cells would be too strong (Section 17.6). Rather, there is a “cluster” of cells where each cell uses a different frequency, and then a neighboring cluster uses the same set of frequencies. Cluster size is a measure for how spectrally efficient a cellular network is: for Global System for Mobile communications

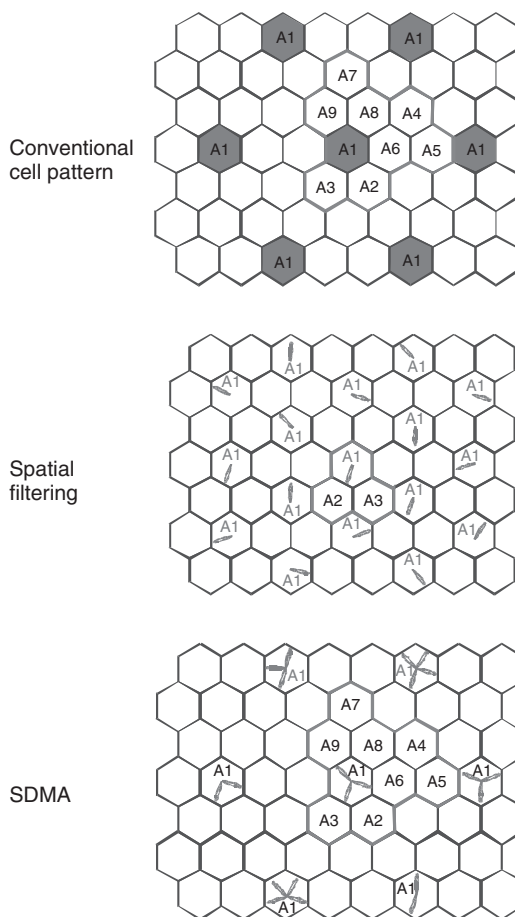


Figure 20.1 Principle behind spatial filtering for interference reduction and space division multiple access.

(GSM)-like systems, it is usually between 3 and 7. Smart antennas reduce interference. For this reason we can put cells with the same frequency closer together – in other words, reduce the cluster size. It is immediately obvious that this leads to an improvement in area spectral efficiency: when frequencies can be used in more cells, then the number of users per area is increased proportionately. Simulations have shown that eight antennas increase capacity by a factor of 3 [Kuchar et al. 1997], while field tests with prototypes [Dam et al. 1999] have shown a somewhat smaller gain of approximately 2.

2. *Space Division Multiple Access (SDMA)*: this is an alternative way of increasing the capacity of TDMA/FDMA systems. In this method, cluster size (frequency reuse) remains unchanged, while the number of users within a given cell is increased. Multiple users can be served on the *same* time/frequency slot, because the BS distinguishes them by means of their different spatial signatures. To understand this mechanism, imagine a situation where there are multiple directional horn antennas at the BS, each of which can be pointed mechanically at a different user, independently of each other. An actual system realizes these directional antennas as the different beams of an antenna array (remember that an antenna array can form multiple beams

simultaneously). The capacity gain that can be achieved with SDMA is larger than with SFIR; however, the required modifications within a system, especially at the BS and the BS controller software, are considerably larger.

3. *Capacity increase in CDMA systems:* smart antennas can also be used to increase capacity in CDMA systems. The way this capacity increase is achieved is subtly different from TDMA/FDMA systems, in that it is based on the enhancement of the desired signal (and not the suppression of a few interferers). In the following, we develop an approximate mathematical model for the capacity enhancement. In a CDMA system, all Mobile Stations (MSs) use the same carrier frequency, and are distinguished only by the spreading codes they employ. The suppression of users with other codes is not perfect, and a cell is judged to be “full” when the residual interference from all the other K users becomes comparable to the admissible SIR (refer to Chapter 18 for more details and a discussion of the assumptions):

$$SIR_{\text{threshold}} = \frac{P_{\text{desired}}}{\sum P_{\text{interfere}}} \quad (20.1)$$

where, for the case of perfect power control, $\sum P_{\text{interfere}} = K \cdot P_{\text{interfere}}$. For typical (voice) CDMA systems, the number of users is on the order of 30 per cell. Thus, there are a large number of possible interferers, each of which contributes only a small part of the interference. It is thus *not* possible to completely suppress interference (remember that the maximum number of interferers that can be suppressed is smaller than the number of antenna elements). Rather, the goal of the smart antenna is to enhance the received signal power. An N_r -element antenna can increase the received signal power by a factor of N_r , so that approximately:

$$P_{\text{desired}} = M_c \cdot N_r \cdot P_{\text{interfere}} \quad (20.2)$$

where M_c is the spreading gain. In that case, the number of admissible users in the cell becomes

$$K = \frac{M_c \cdot N_r}{SIR_{\text{threshold}}} \quad (20.3)$$

and thus increases *linearly* with the number of antenna elements. In practice, such an enormous increase in capacity does not occur; typically, capacity can be doubled with an eight-element array. This is comparable with the gains that can be achieved in a TDMA system. Note that this derivation is based on numerous simplifying assumptions, especially with respect to the channel model. It furthermore assumes that there are a large number of users in a CDMA system, and that all of them contribute to interference in a similar way. This assumption holds for second-generation CDMA systems – e.g., IS-95³ – which are mainly tuned to voice users.

4. *Capacity increase in third-generation CDMA systems:* in third-generation networks, high-rate data transmission is considered an important application. Such a high-rate user creates considerably more interference due to the lower spreading factor, such that suppression of such a user by placing a null in its direction becomes desirable. In other words, high-rate users lead to scenarios and results that are more similar to the SFIR approach than the traditional CDMA approach.

Example 20.1 Consider a CDMA system with spreading factor 128 for voice users, and 4 for data users, and an SIR threshold of 9 dB. There is one data user. How many voice users can be accommodated without smart antennas? How many can be accommodated with a two-element smart antenna?

³ See Chapter 25.

The SIR threshold (on a linear scale) is 8. Without smart antennas, the capacity of the system (in units of voice users) can be obtained from

$$K = \frac{128}{8} = 16 \quad (20.4)$$

Since the spreading factor for data users is smaller by a factor of 32 compared with voice users, each data user contributes (approximately) 32 times the interference. The data user thus requires a capacity of $K_{\text{data}} = 32$, so that $16 - 32 = 0$ voice users can be accommodated.

Consider now the case with smart antennas. When smart antennas are used solely for enhancement of the desired signal power, the capacity of the cell (in units of voice users) becomes

$$K = \frac{128.2}{8} = 32 \quad (20.5)$$

so that now $32 - 32 = 0$ voice users can be serviced.

If, however, smart antennas are used to suppress data users, higher capacity can be achieved: as we saw in Chapter 13, we can suppress 1 interferer completely using two antennas, while retaining a diversity order of 1 for the desired signal. Thus, in our example, we can suppress the data user while retaining $N_r = 1$ for voice users. Thus, the capacity for voice users becomes

$$\frac{128}{8} = 16 \quad (20.6)$$

20.1.4 Receiver Structures

In this section, we describe RX structures that can be used for separation and processing of MPCs. Since our emphasis is on smart antennas at the BS, this implies that we consider the uplink case (see Section 20.1.6 for the downlink). We distinguish between *switched-beam antennas*, *adaptive spatial processing*, *space–time processing*, and *space–time detection*.

Switched-Beam Antennas

A switched-beam antenna is an antenna array that can form just a small set of patterns – i.e., beams pointing in certain discrete directions. A switch then selects one of the possible beams⁴ for downconversion and further processing; it selects the beam that is *best* in the sense that it gives the highest SNR or the highest Signal-to-Interference-and-Noise Ratio (SINR). This approach greatly simplifies RX design.

There are many different ways of realizing switchable beams. For example, the antenna can contain multiple directional elements, oriented in different directions, and the switch selects from them. Another, more popular, variant is a linear array followed by a spatial Fourier transformation. The output of such a spatial Fourier transformation is a number of beams that are all orthogonal to each other, and point in different directions. The spatial Fourier transformation can be realized as a so-called *Butler matrix*. It has a structure that is essentially the butterfly structure that is well known from software implementations of the Fast Fourier Transform (FFT). The elements of each stage are simple phaseshifters that can be realized in the Radio Frequency (RF) domain.

The main advantage of the switched-beam approach is its simplicity: all processing (spatial FFT and selection) is done in the RF domain, so that only a single signal has to be downconverted to the baseband and processed there. Since downconversion circuits are among the most expensive

⁴ More precisely, it selects the signal associated with one of the possible beams.

components in today's wireless systems, this is a significant advantage. The drawback of this scheme is its limited flexibility. The main beam can only be pointed in certain fixed directions, such that gain in the actual direction of the MS might not be the maximum achievable value. Even more importantly, nulls cannot be pointed in arbitrary directions, so that nulling of interferers is very ineffective. For these reasons, switched-beam antennas seem to be more suitable for CDMA applications, where signal enhancement is critical, and not for SDMA or SFIR applications, where interference suppression is essential.

Example 20.2 For a switched-beam antenna, compute the maximum gain and the maximum relative loss due to mismatch of DOA and beam direction for a uniform linear array with Butler matrix and number of antenna elements (spaced $\lambda/2$ apart) equal to $N_r = 2, 4, 8$.

Let us first write down the FFT matrix for N_r elements:

$$W_{\text{FFT},k,n} = w_n(k) = \exp\left(-j\frac{2\pi}{N_r}\right)^{kn} \quad (20.7)$$

The array factor for a uniform linear array was given in Eq. (9.16):

$$M_k(\phi) = \sum_{n=0}^{N_r-1} w_n(k) \exp[-j \cos(\phi) 2\pi d_a n / \lambda] \quad (20.8)$$

With $d_a = \lambda/2$, this becomes

$$M_k(\phi) = \sum_{n=0}^{N_r-1} \exp\left[-j2\pi n \left(\frac{\cos(\phi)}{2} + \frac{k}{N_r}\right)\right] \quad (20.9)$$

and the magnitude of the array factor is (compare Eq. 9.17)

$$|M_k(\phi)| = \left| \frac{\sin\left[\frac{N_r}{2} \left(\pi \cos \phi - 2\pi \frac{k}{N_r}\right)\right]}{\sin\left[\frac{1}{2} \left(\pi \cos \phi - 2\pi \frac{k}{N_r}\right)\right]} \right| \quad (20.10)$$

The crossover points between patterns at FFT output k and $k+1$ are at:

$$\left| \frac{\sin\left[\frac{N_r}{2} \left(\pi \cos \phi - 2\pi \frac{k}{N_r}\right)\right]}{\sin\left[\frac{1}{2} \left(\pi \cos \phi - 2\pi \frac{k}{N_r}\right)\right]} \right| = \left| \frac{\sin\left[\frac{N_r}{2} \left(\pi \cos \phi - 2\pi \frac{k+1}{N_r}\right)\right]}{\sin\left[\frac{1}{2} \left(\pi \cos \phi - 2\pi \frac{k+1}{N_r}\right)\right]} \right| \quad (20.11)$$

Figure 20.2 shows the array factor for different values of N_r . We see that in all cases the value of the array factor of the crossover point is approximately 0.7. The exact values can be found in Table 20.1:

Table 20.1

N_r	Max gain (dB)	Max relative loss due to direction-of-arrival mismatch (dB)
2	3.01	1.51
4	6.02	1.84
8	9.03	1.93

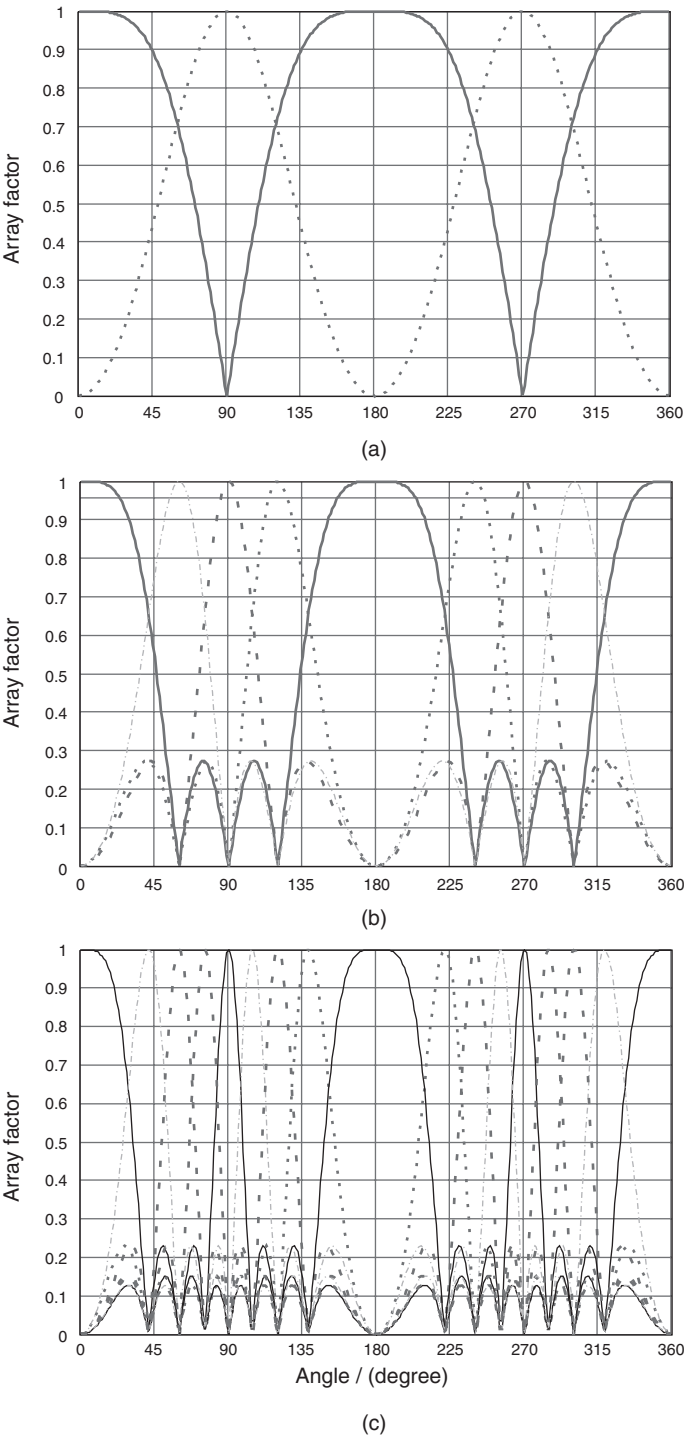


Figure 20.2 Normalized array factor of switched-beam antenna with 2 (a), 4 (b), and 8 (c) elements.

Adaptive Spatial Processing

The adaptive spatial processing approach uses a linear combination of signals (see Figure 20.3), where antenna weighting and summing is done in the baseband. For this reason, there are no restrictions with respect to weights, and they can also be adapted according to the current channel state. In other words, the RX can point maxima and nulls of the array pattern into (almost) arbitrary directions. On the downside, this approach requires in general N_r complete downconversion chains for N_r antenna elements, and is thus considerably more expensive, and consumes more power as well.

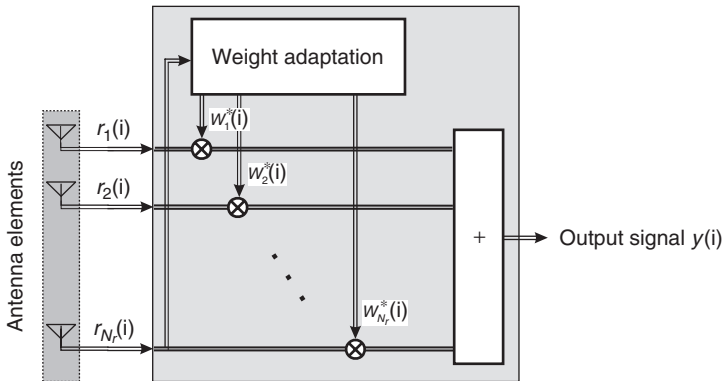


Figure 20.3 Linear combination of antenna signals.

If weights are adapted based on a training sequence, and change whenever the channel realization changes, then adaptive spatial processing is identical to diversity combining. It is thus possible to suppress $K = N_r - 1$ interferers, or alternatively to achieve high gain in the SNR (up to a factor of N_r) with such an approach (for more details, see Section 13.4.3).

Alternatively, antenna weights can be chosen based on angular spectrum (Angular Delay Power Spectrum (ADPS), see Section 6.7), but not adapted to the instantaneous amplitude of MPCs arriving from different directions. In this case, interference suppression capabilities are usually smaller. This aspect is discussed in more detail in Section 20.1.5.

Adaptive Space–Time Processing

The adaptive spatial processing described in the previous section first performs spatial processing and then puts out a single signal for further temporal processing. In other words, Rake reception or equalization is done only on the *single* signal that is put out from the spatial processor. This approach works well only if the temporal properties of the signal (delay dispersion) are the same at all antenna elements, or, conversely, if spatial properties are independent of delay. While this situation can occur in some cases (see also Chapter 7 for channel models that correspond to such an assumption), it is certainly not universally valid. If delay and angular properties cannot be decomposed multiplicatively, then spatial processing followed by temporal processing is no longer optimum.

In order to fully exploit the possibilities of different MPCs, adaptive *space–time processing* has to be used. The optimum linear RX is a two-dimensional Rake RX – i.e., a linear combiner that

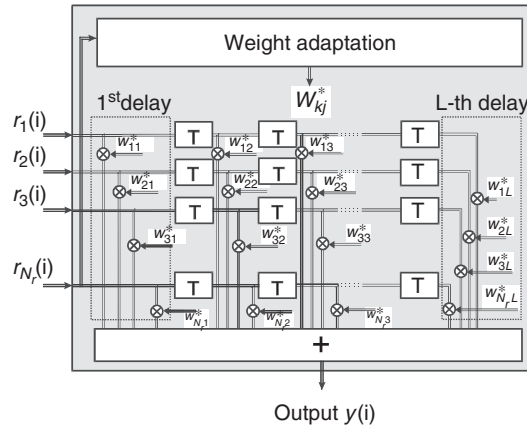


Figure 20.4 Space-time filter.

weights all *resolvable* (in the space-time domain) MPCs and adds them up. Figure 20.4 shows such a space-time processor. If the Rake RX (or equalizer) has L taps, then the total processor has $N_r L$ weights. The output from that processor is sent on to a decoder.

Space-Time Detection

For optimum reception, space-time processing and decoding/detection have to be done jointly. The special properties of the received signal, including finite-alphabet properties, can be taken into account for detection. The optimum detector is a generalized Maximum Likelihood Sequence Estimation (MLSE) RX. However, these structures are not always used in practice, because they are too complex, and the performance gain compared with linear processing does not always justify the additional effort.

20.1.5 Algorithms for Adaptation of Antenna Weights

Algorithms for adaptation of antenna weights can be broadly classified into Spatial Reference (SR) and temporal reference (TR) algorithms.⁵ The distinction between these algorithms is subtle, but important. Let us first discuss their common features: both algorithms are based on *linear* weighting and addition of signals from different antenna elements. Both algorithms effectively form a beam pattern – the weights for any linear combining scheme can be associated with a “beam pattern” even when it sometimes looks somewhat strange, with many maxima and minima. The key difference is which information is used for the choice of the linear weights. In SR algorithms weights are chosen according to the spatial structure (DOAs) of the arriving signal combined with information about the array structure. TR algorithms optimize the SINR after the combiner with the help of a training sequence.

⁵ In the literature, there is often a distinction between diversity and beamforming algorithms. However, this distinction is problematic as diversity (slope of the effective SNR distribution) and beamforming (change in mean SNR) are effects that, among others, depend on antenna arrangement and channel constellation.

Spatial Reference Algorithms

In the *spatial reference (SR)* case, the antenna tries to form a beam pattern that puts maxima in the direction of the main DOAs of the desired user, and nulls in the direction of MPCs coming from interferers. An SR algorithm thus proceeds in the following three steps (see Figure 20.5):

1. Determination of the DOAs, ϕ_n , of the MPCs: as we will see below, a main advantage of using these DOAs is that they show only small variations over time and frequency. The directly observable quantities are the signals at the antenna elements; it is thus necessary to extract the DOAs from these signals. This can be done by the same methods as those for channel sounding (see Chapter 8): spatial Fourier transformations (not recommended) or high-resolution algorithms (Minimum Variance Method MVM, Estimation of Signal Parameters by Rotational Invariance Technique ESPRIT, Space Alternating Generalized Expectation – maximization SAGE) can be applied. There are some important practical differences from the evaluation of channel sounder data: the SNR for the reception of data can be considerably worse than that used for channel sounding – this makes DOA determination more difficult. On the other hand, a lot of data are transmitted within a relatively short time. This allows the use of either averaging or tracking algorithms that observe the change in DOAs [Kuchar et al. 2002]. Note that some DOA estimation algorithms (e.g., ESPRIT) do not require a training sequence, since they are based exclusively on the correlation matrix of the received signal, and can thus estimate the DOA from user data that are unknown. Other algorithms, like SAGE, can eschew the training sequence only at the price of greatly complicating the algorithm.⁶
2. Association of DOAs with specific users: in contrast to channel sounding, where we have just a single possible source for the arriving signal, MPCs during communications can stem from different users – the desired user as well as interferers. It is thus necessary to separate desired and interfering users. For this identification problem, it is necessary to have a training sequence or some other property that is unique for each user.
3. Forming of the beam pattern that maximizes the SINR: this beam pattern is used for actual reception of data.

We note that many DOA estimation algorithms require the validity of a certain channel model (e.g., that the received signal can be written as a finite sum of plane waves). If this model is not valid, the performance of the algorithms, and of the ensuing beamforming, deteriorates.

Temporal Reference Algorithms

Temporal reference algorithms are based on the use of a training signal, which serves as a “temporal reference” to which we try to match the output from the smart antenna – thus the name. The antenna

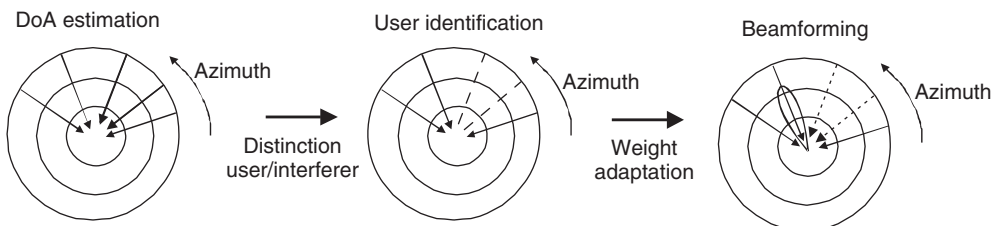


Figure 20.5 Using smart antennas with spatial reference algorithms.

⁶ However, when SAGE is used in conjunction with a training sequence, it usually outperforms ESPRIT.

weights $\mathbf{w}$ are adapted directly in such a way that the deviation of the combiner output from the (known) training sequence is minimized. As a criterion for the deviation we can use the SIR, the Minimum Mean Square Error (MMSE), the Bit Error Rate (BER) during the training sequence, or any other suitable criterion. These criteria and their advantages and disadvantages have already been discussed in Chapter 13 (we stress again that smart antennas with diversity combining at the RX is the “standard” diversity that we have discussed in Chapter 13). In particular, the SINR is maximized by optimum combining. The linear weights lead to the creation of an effective antenna array pattern, but this pattern need not have an intuitive interpretation. There are also no assumptions about the existence of discrete DOAs.

A major difficulty in TR algorithms is the fact that the RX must be synchronized to the incoming signal *before* the determination of the antenna weights can be done. This is required because the sampling instants for the determination of the training sequence have to be known before the weights can be adapted. However, the SINR for the (spatially unfiltered) receive signal can be very poor and can make synchronization to the desired training sequence difficult.

In summary, the TR algorithms proceed according to the following steps (see Figure 20.6):

1. During a training phase, the smart antenna receives the signal at all antenna elements, and adjusts the weights in such a way that the deviation from the known signal is minimized.
2. During the transmission of the user data, the antenna weights stay fixed and are used to weight the incoming signals before they are combined and decoded/demodulated.

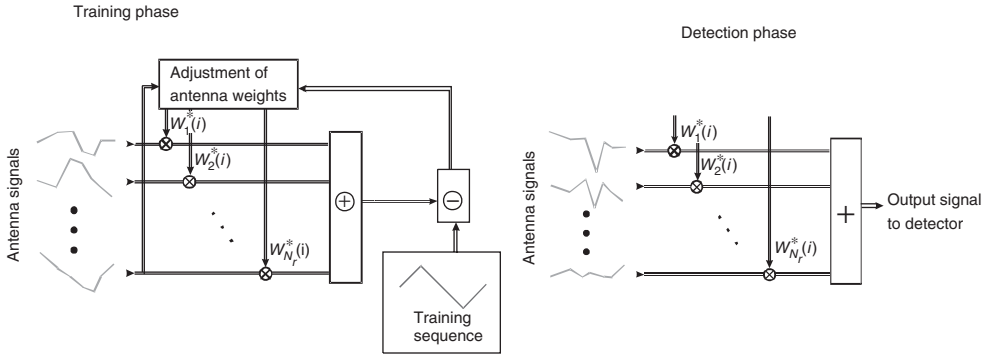


Figure 20.6 Principle behind temporal reference algorithms.

We now look at more details of step 1. Assume that we have determined the channel coefficient, and we want to determine the optimum antenna weights. Paraphrasing the main insights from Section 13.4.3, those optimum antenna weights maximize the SINR. Define the power-weighted channel and interference correlation matrices as

$$\tilde{\mathbf{R}}^{(k)} = P_k \mathbf{h}^{(k)} (\mathbf{h}^{(k)})^\dagger, \quad \tilde{\mathbf{R}}_{\text{ni},k} = \sigma_n^2 \mathbf{I} + \sum_{l \neq k} P_l \mathbf{h}^{(l)} (\mathbf{h}^{(l)})^\dagger \quad (20.12)$$

where P_k is the power of the k -th user. Then the optimum antenna weights $\mathbf{w}^{(k)}$ for the k -th user are obtained by maximization of the SINR, i.e.,

$$\mathbf{w}_{\text{opt}}^{(k)} = \arg \max_{|\mathbf{w}|^2=1} \text{SINR}_k = \arg \max_{|\mathbf{w}|^2=1} \frac{\mathbf{w}^\dagger \tilde{\mathbf{R}}^{(k)} \mathbf{w}}{\mathbf{w}^\dagger \tilde{\mathbf{R}}_{\text{ni}}^{(k)} \mathbf{w}} \quad (20.13)$$

which is the solution of a generalized eigenvalue problem. An even simpler RX, called “decorrelating RX,” can be achieved by setting $\sigma_n = 0$. This RX completely suppresses the interference, but considerably enhances the noise (compare also the discussion of zero-forcing and MMSE equalizers in Chapter 16).

The above derivation assumed that the interference for a specific user is fixed. We can get better results for the average throughput if we can influence the transmit powers of the different users – this requires feedback commands from the BS to the MS (to instruct the MSs about which power level to use), as well as power control capabilities of the MSs; most second- and third-generation cellular MSs have such capabilities. Finding the optimum MS transmission powers together with the optimum receive antenna weights is complicated because adjusting the power for one MS affects the SINR for all users. There is therefore no closed-form solution; rather, we iterate the following two steps until convergence:

1. Fix the transmit powers of the MSs and obtain the BS antenna weights for the reception of the different users by Eq. (20.13).
2. Then, leave the set of antenna weights $w^{(k)}$ unchanged and compute the optimum power allocations P_k . This computation itself does not have a closed-form solution but can be obtained through various numerical optimization methods that depend on the target of the optimization. For example, if we want to optimize the sum throughput, the RX iterates the following equation to convergence (where n is the iteration counter) [Chrisanthopoulos and Tsoukatos 2007]:

$$P_k^{(n+1)} = \min \left[P_{\max}, P_k^{(n)} + \alpha \left(\frac{1}{P_k^{(n)}} - \sum_{l \neq k} \frac{(\mathbf{w}^{(l)})^\dagger \tilde{\mathbf{R}}^{(k)} \mathbf{w}^{(l)}}{\sigma_n^2 + \sum_{m \neq l} (\mathbf{w}^{(l)})^\dagger \tilde{\mathbf{R}}^{(m)} \mathbf{w}^{(l)}} \right) \right] \quad (20.14)$$

where $P_{\max}$ is the maximum power an MS can transmit, and α is a parameter determining the convergence rate; if $P_k^{n+1} < 0$, we set it to 0.

Further performance improvements can be achieved when the RX is not restricted to linear combining of the antenna signals. In particular, the MMSE RX can be combined with a successive-interference canceller (see Section 18.4.3): the BS first decodes a particular user, and subtracts the contribution of this signal from the overall received signal. When now decoding the signal for the next user, a better SINR is available. This decoding strategy is theoretically optimum, but is (like all serial interference cancellation strategies) sensitive to error propagation. It is also noteworthy that the resulting RX has strong similarity to the Horizontal Bell labs LAYERed Space Time (H-BLAST) RX that is discussed in Section 20.2.8.

Blind Algorithms

TR algorithms rely on the existence of a training sequence, while SR algorithms rely on the existence of a certain spatial structure of the arriving signal. Yet another group of algorithms, dubbed *blind* algorithms, makes neither of those assumptions. Rather, they exploit the statistical properties of the transmit signal.

Let us start out by describing the received signal by the equation

$$\mathbf{r}_i = \mathbf{h}_d s_i \quad (20.15)$$

where $\mathbf{r}_i$ is the receive signal vector created by the transmit signal s_i related to the i th bit, and $\mathbf{h}_d$ is the desired channel vector – we omit the noise contribution for convenience. A more general

representation, which also includes possible intersymbol interference, is obtained by stacking signals corresponding to a large number of bits into matrices:

$$\mathbf{Y} = \mathbf{H}_{\text{stack}} \mathbf{X} \quad (20.16)$$

where the matrix $\mathbf{H}_{\text{stack}}$ is related to channel impulse responses at different antenna elements. The channel description is thus similar to the TR approach in the sense that there is no assumption about spatial structure, and the channel is characterized by impulse response (or transfer function) alone. However, a blind algorithm does *not* learn the values of $\mathbf{H}_{\text{stack}}$ from a training sequence. Rather, it tries to compute a factorization of the matrix $\mathbf{Y}$ according to Eq. (20.16), such that signal matrix $\mathbf{X}$ fulfills certain properties – namely, *known* properties of the transmit signal.

Example 20.3 Consider a system with one transmit antenna and three receive antennas. Let the impulse response for each of the antenna elements last for two symbol durations, denoted as $h_{i,l}$, $l = 0, 1$, $i = 1, 2, 3$. For a five-symbol-long transmit signal, $x(n)$, $n = 1, \dots, 5$, write the explicit form of $\mathbf{Y} = \mathbf{H}_{\text{stack}} \mathbf{X}$, assuming that impulse responses stay constant over the duration of the transmit signal.

Writing the n th received symbol at the i th antenna element as $y_i(n)$, we can easily see that

$$y_i(n) = \sum_{l=0}^{L-1} h_{i,l} x(n-l) \quad (20.17)$$

where $h_{i,l}$ is the channel impulse response to the signal at the i th antenna element in the l th channel delay tap, L is the duration of the channel impulse response, and N is the number of transmitted symbols. Let us then define a matrix $\mathbf{X}$ as [Laurila 2000]

$$\mathbf{X} = \begin{bmatrix} x(1) & x(2) & \cdots & x(N) \\ x(0) & x(1) & \cdots & x(N-1) \\ \vdots & \vdots & \vdots & \vdots \\ x(-L+2) & x(-L+3) & \cdots & x(N-L+1) \end{bmatrix} \quad (20.18)$$

which in our case becomes

$$\mathbf{X} = \begin{bmatrix} x(1) & x(2) & x(3) & x(4) & x(5) \\ x(0) & x(1) & x(2) & x(3) & x(4) \end{bmatrix} \quad (20.19)$$

The channel matrix gets the block form:

$$\mathbf{H}_{\text{stack}} = \begin{bmatrix} h_{1,0} & h_{1,1} & \cdots & h_{1,L-1} \\ h_{2,0} & h_{2,1} & \cdots & h_{2,L-1} \\ \vdots & \vdots & \vdots & \vdots \\ h_{N_r,0} & h_{N_r,1} & \cdots & h_{N_r,L-1} \end{bmatrix} \quad (20.20)$$

which in our case becomes

$$\mathbf{H}_{\text{stack}} = \begin{bmatrix} h_{1,0} & h_{1,1} \\ h_{2,0} & h_{2,1} \\ h_{3,0} & h_{3,1} \end{bmatrix} \quad (20.21)$$

Inserting Eqs. (20.19) and (20.21) into Eq. (20.16), we get:

$$\mathbf{Y} = \begin{bmatrix} y_1(1) & y_1(2) & y_1(3) & y_1(4) & y_1(5) \\ y_2(1) & y_2(2) & y_2(3) & y_2(4) & y_2(5) \\ y_3(1) & y_3(2) & y_3(3) & y_3(4) & y_3(5) \end{bmatrix} \quad (20.22)$$

$$= \begin{bmatrix} h_{1,0}x(1) + h_{1,1}x(0) & \cdots & h_{1,0}x(5) + h_{1,1}x(4) \\ h_{2,0}x(1) + h_{2,1}x(0) & \cdots & h_{2,0}x(5) + h_{2,1}x(4) \\ h_{3,0}x(1) + h_{3,1}x(0) & \cdots & h_{3,0}x(5) + h_{3,1}x(4) \end{bmatrix} \quad (20.23)$$

which can easily be seen to be in agreement with Eq. (20.17).

Depending on which signal properties are exploited, we get different algorithms for the determination of $\mathbf{X}$. If there is oversampling in the temporal and/or spatial domain, then cyclostationarity of the sampled signal can be used. Another group of algorithms exploits higher order statistics. Also the finite-alphabet property can be exploited: e.g., we know that for a Binary Phase Shift Keying (BPSK) signal, the elements of $\mathbf{X}$ have to be ± 1 .

Blind algorithms have a number of advantages:

- No assumptions about the channel are required. The method works even for channels that do not exhibit discrete DOAs or separable MPCs.
- Blind algorithms do not require any calibration of the antenna array, and make no assumptions about the specific structure of the array.
- The detection process can be done in one step, yielding directly the desired user data.
- Blind algorithms eliminate the need for a training sequence, thus increasing the spectral efficiency of the system.

However, these advantages are bought at a price:

- Most of the blind algorithms assume that the channel impulse response does not change over the time it takes to establish the statistics of $\mathbf{Y}$. For a good estimate of these statistics, we have to collect samples over a long time, and the assumption of a time-invariant impulse response might be violated. The number of samples that forms the basis of the statistics is thus a compromise between the need to have a large-enough basis for obtaining good statistics and the need to use only samples that correspond to the same channel state.
- Initialization of the estimate for the channel is critical. For this reason, so-called *semi-blind* algorithms have been proposed, which use a very short training sequence at the beginning of transmission. This sequence is chosen to be short enough not to significantly influence the spectral efficiency of the system, but helps to avoid possible convergence problems of the blind algorithms.

These drawbacks are essentially the same as for blind equalization (see Chapter 16), and have prevented widespread use of this method.

20.1.6 Uplink versus Downlink

Up to now, we have considered smart antennas at the RX; since we consider cellular systems in which smart antennas are at the BS, this means that we have been looking at the uplink. The next question is how a smart antenna system behaves in the downlink. The answer to this question

depends critically on whether we are considering a Time Domain Duplex (TDD) or a Frequency Domain Duplexing (FDD) system.

Time Domain Duplexing

In a TDD system in a static environment – i.e., no movement of either the TX or the RX – the channel impulse response is the same for uplink and downlink. Let us make this statement more precise. The transfer function from the MS antenna to the m th BS antenna, $(\mathbf{h}_d)_m$, is the same as the transfer function from the m th BS antenna to the MS antenna. Thus, if we have chosen antenna weights in such a way that they constructively add signals from the antenna elements during the uplink, the *same* antenna weights, when used in the downlink, will ensure that the signals from different BS antenna elements add up constructively at the MS antenna. Furthermore, if BS antenna weights are chosen to suppress interference from other MSs during the uplink, the BS will also cause little or no interference to these other MSs in the downlink transmission phase.

However, interference is not completely reciprocal for the uplink and downlink: the interference that the MS sees stems from other BSs, not from other MSs (see Figure 20.7). However, antenna weights $\mathbf{w}$ that are determined during the uplink cannot take this into account – the desired BS does not even “see” these other BSs. Thus, interference suppression for the downlink cannot be as effective as suppression in the uplink.⁷

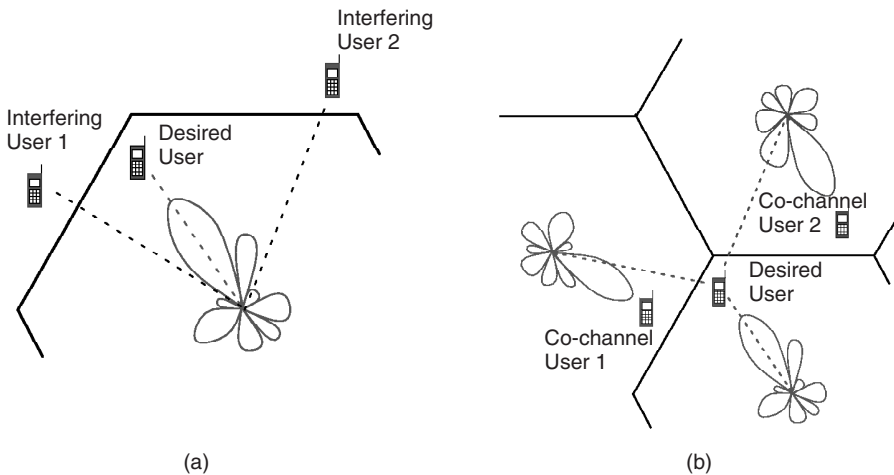


Figure 20.7 Interference situation for the uplink (a) and downlink (b) for a base station with smart antennas.

All we have stated up to now is that the *channel* is reciprocal. However, since antenna weights are determined in the baseband, we also have to require that RF elements and frequency conversion chains – i.e., everything from antenna connectors to the baseband – are reciprocal. This is not self-evident and, in fact, will not be the case for many transceivers. The problem can be eliminated or, at least, mitigated by a calibration procedure: during the setup of the call, the BS estimates the channel

⁷ An exception would be if an MS found itself in a position to notify all the BSs in its surroundings about the details of the interference it sees from them, and the BSs were in a position to cooperate in order to provide as little interference as possible to all users. Such schemes have been proposed, but are rather complex.

impulse response both from the reciprocity principle and by means of feedback from the MS. From the difference of these two measurement methods, the impact of RF circuits can be determined.

Finally, we have to take into account that the channel need not be completely static in the time between uplink and downlink transmission (duplexing time). If the duplexing time is too long (becomes comparable with the coherence time of the channel), then the impulse response of the channel changes due to small-scale fading. The impact of such a decorrelation between the uplink and downlink channel is discussed in the next subsection.

Frequency Domain Duplexing

In an FDD system, the uplink and downlink happen at different frequencies; the frequency separation (duplexing distance) for cellular systems is typically on the order of 100 MHz. This is much larger than the coherence bandwidth of the channel (typically on the order of 1 MHz); as a consequence, small-scale fading of the channel impulse response in the uplink and in the downlink are completely decorrelated. In other words, small-scale fading is created by superposition of MPCs with different phases. The relative phase shifts depend, among other things, on the carrier frequency. For a large-enough duplexing distance, phase shifts are sufficiently different in uplink and downlink that superposition of MPCs occurs in different ways.

For this reason, we find that small-scale fading and, thus, the instantaneous impulse response of the channel, are different for uplink and downlink (a similar situation occurs in a TDD system when the duplexing time is much larger than the coherence time of the channel). What remains constant in uplink and downlink is the *average* channel state: in other words, the correlation matrix (averaged over small-scale fading), as well as the DOAs, delays, and mean powers of the MPCs.

It is still possible to obtain performance enhancements in the downlink for FDD. Antenna weights are now chosen based on the average channel state. For example, the beam pattern is chosen so that it points in the direction of the MPC that is strongest *on average*. Now, it is possible that this MPC is not the strongest *instantaneously*, and that it would therefore be better to choose a different beam pattern. However, since we do not know the instantaneous channel, optimizing on average is the best that we can do with the available information.

These considerations also show the usefulness of SR algorithms. Since these algorithms extract average *Channel State Information* (CSI), they can be easily used in FDD systems.

Feedback

Up to now, we have relied on reciprocity (either for the channel impulse response or for DOAs) to provide information about the channel. An alternative approach is to use feedback. Consider an FDD system. Let the BS transmit a training sequence from the first antenna element, then from the second antenna element, and so on. This enables the MS to learn all the separate impulse responses $(\mathbf{h}_d)_m$. It then feeds the information back to the BS on a separate feedback channel. In that case, the BS knows exactly how to adjust weights such that signals from antenna elements add up in the right way at the MS antenna.

This approach is very effective, but has two drawbacks: first, we have to make sure that feedback occurs within a time that is less than the coherence time of the channel. Second, the feedback channel decreases the spectral efficiency of the system. The information carried on that channel is not user data. System designers thus have to consider carefully how often to feed CSI back, and with what accuracy. Naturally, the performance of the smart antenna gets better the more accurate the feedback information is, but the performance penalty because of wasted resources can become prohibitive. A more detailed discussion can be found in Section 20.2.11.

20.1.7 Algorithms for the Adaptation of the Antenna Weights in the Downlink

Determination of the optimum transmission strategy for the downlink can be considerably more complicated than for the uplink, because changing the transmit strategy (e.g., the antenna weights) for one user influences the SINR for every other user.⁸

We will in the following only consider linear (beamforming) transmission schemes. The BS has a constraint on the *total* (summed over the signals for all users) transmit power $\sum_k P_k \leq P_{\max}$,

which is reasonable, because the total power is limited both by the characteristics of the power amplifier, and by the intercell interference that we are willing to allow. We assume that the BS has complete CSI, i.e., it knows $\mathbf{h}^{(k)}$ for all the k users. For the following mathematical treatment, it is useful to create an *overall channel matrix* $\mathbf{H} = [(\mathbf{h}^{(1)})^T, (\mathbf{h}^{(2)})^T, \dots, (\mathbf{h}^{(K)})^T]^T$. We also define $\tilde{\mathbf{s}}$ as the signal vector transmitted to the BS antenna array, which in the case of beamforming (linear precoding) is

$$\tilde{\mathbf{s}} = \mathbf{T}\mathbf{s}. \quad (20.24)$$

Let us first establish how many users can communicate simultaneously with the BS. We know that in the uplink, the number of users is limited by the number of antenna elements in the BS, N_{BS} . This follows from the principle of “optimum combining” that every additional antenna elements can suppress one interfering signal; thus, one set of BS antenna weights can deal with 1 desired user and $N_{\text{BS}} - 1$ interferers simultaneously. The key insight is now that this admissible number of users is the same for the downlink, i.e., a BS with N_{BS} antenna elements can transmit to N_{BS} different MSs at the same time. Intuitively, this result follows from reciprocity: if – by adapting its weights – the BS in the uplink forms an antenna pattern that nulls out an interferer, then the same BS will not create any interference at this user when the BS transmits.⁹

Let us now formulate these insights mathematically. If the k -th MS should see no interference intended for other users, the transmit signal for the k -th MS must lie in the nullspace of the channel vectors for all other users. The precoding vector for the k -th MS is then the k -th column of the pseudoinverse of $\mathbf{H}$, so that the transmit signal is

$$\tilde{\mathbf{s}} = \mathbf{H}^\dagger (\mathbf{H}\mathbf{H}^\dagger)^{-1} \Xi \mathbf{s} \quad (20.25)$$

where Ξ is a diagonal matrix that contains the power allocations for the different users. Thus, the beamforming can be combined with a power allocation strategy that either maximizes the sum throughput (for the given shape of the antenna patterns), optimize some “sum MSE,” or tries to guarantee a certain minimum quality of service (SINR) for each user.

Clearly, zero interference can only be achieved if no more than N_t users are served at one given time, so we assume henceforth $K \leq N_t$. The key disadvantage of zero-forcing TXs is that in the case of ill-conditioned channel matrices, the transmit power along a particular dimension can become very large, and since the total transmit power is limited, this means that transmit power is used very inefficiently. In other words, if two users are “effectively” very close together (in the sense that

⁸ There are some useful dualities between the uplink and downlink case, which can be exploited to derive good transmission schemes. For example, it can be shown that the same SINRs can be achieved for all users with the same overall transmit power (though the powers associated with each user in the uplink and downlink are different). However, most algorithms developed in the literature for the downlink antenna weights do not make explicit use of this duality.

⁹ Note, however, that the optimum power assignments are different for uplink and downlink, because noise and interference levels are not necessarily reciprocal.

their channel vectors are almost the same – they do not have to be close together geographically), then it is very difficult to transmit to one user while simultaneously canceling out transmissions to the other. This situation is dual to the effect of “noise enhancement” by a decorrelation RX.

The performance can be improved by so-called “regularization” that trades off noise enhancement and interference. The transmit signal becomes

$$\tilde{\mathbf{s}} = \mathbf{H}^{\dagger} (\mathbf{H}\mathbf{H}^{\dagger} + \zeta \mathbf{I})^{-1} \mathbf{\Xi} \mathbf{s} \quad (20.26)$$

where ζ is the so-called “regularization parameter.” This is obviously a dual of an MMSE (optimum combining) RX.

A joint beamforming/power allocation strategy that leads to a maximization of the sum throughput is given – for the case $N_t = K$ in [Stojnic et al. 2006]. It is noteworthy that if we want to maximize the sum rate, fairness is no criterion – in other words, the data streams for some MSs have very high data rates, while other MSs are not serviced at all. Alternative beamforming/power allocation criteria thus try to guarantee a minimum rate for each user. Furthermore, the random beamforming discussed later in this section can be used to provide greater fairness.

There are many similar, but subtly different, optimization problems. For example, the “power control problem” aims to minimize the total transmitted power while guaranteeing a certain minimum SINR (and thus a certain rate) for every user. A number of such resource allocation problems can be tackled by “convex optimization” techniques [Boyd and Vandenberghe 2004].

20.1.8 Network Aspects

In order to really increase capacity by means of smart antennas, it is usually not sufficient to just insert an applique – i.e., replace the single-antenna RF front end and baseband demodulator with their multiantenna counterparts. Rather, the network has to provide additional functionalities. Consider a TDMA/FDMA system that uses SDMA. It can put two different users on the same time/frequency slot and separate them by their DOAs. However, when the two users get too close, such a separation is no longer possible. In such a case, the network has to swap one of these two users with one that is on a different time/frequency slot and sufficiently far away in the angular domain. In other words, the network has to provide intracell handover capabilities. Thus, the control software of the BS has to be designed specifically for the use of SDMA.

Another major problem is the call setup phase. In this phase, the position and thus DOAs of the various users are not known yet, and the call setup information has to be transmitted in all possible directions simultaneously. Compared with actual data transmission, network planning, link budgets, etc., can thus be different for this phase.

20.1.9 Multiuser Diversity and Random Beamforming

Throughout this book, we have assumed that the BS and the MS transmit at predetermined times, irrespective of the channel state. However, it has recently been established that the performance of cellular systems can be improved significantly by scheduling transmission depending on the channel state. It can also be shown that the use of multiple antennas can enhance this idea even further. In the following, we will explain this idea in more detail.

Scheduling and Multiuser Diversity

Let us first consider the downlink in a system where the BS as well as the MSs have only a single antenna, and multiple MSs communicate with one BS. All links undergo flat Rayleigh fading;

we assume for the moment that the mean power is the same for all users. Let us further assume a conventional TDMA system, so that each user is assigned a fixed timeslot and communicates during that timeslot. The probability for user k to “see” a certain SNR γ is then given by the exponential distribution:

$$\Pr(r_k \leq \gamma) = 1 - \exp\left(-\frac{\gamma}{\bar{\gamma}}\right) \quad (20.27)$$

This probability is independent of user k , and is also independent of the *number* of users K . The scheme, also called *round-robin*, is the best we can do if the TX (BS) does not know anything about the channel. Obviously, it treats users fairly in the sense that each user gets the same amount of time to communicate.

However, we can do better if the BS knows the CSI for the downlink of all users. At any given point in time, chances are that there is a very good link to some users, while there are bad links to other users. The optimum strategy for the BS is now to communicate always with the user who has the best link – i.e., the highest instantaneous path gain – and thus the highest instantaneous SNR γ . As the channel changes, different users k become “instantaneously best.” The cumulative distribution function for the active user – i.e., the instantaneously best user – is the same as for selection diversity (see Eq. 13.10). Multiple users thus act like diversity branches – however, we do not need any additional antennas to realize that diversity. This scheme is thus called *multiuser diversity*; it is optimum in the sense that capacity is maximized.

In a cellular system, users that are very close to the BS have a much better SNR than users that are far away. In that case, the danger of selecting the user with the best SNR is that the MS closest to the BS gets to use the channel most often, while MSs that are far away are almost never “instantaneously best,” and thus are chosen much less frequently. As a consequence, the data rate for users at the cell boundary is much lower than for users near the BS. This problem can be solved by *proportionally fair* scheduling. The BS communicates not with the absolutely best user but rather with the user whose ratio of instantaneous to average SNR $\gamma/\bar{\gamma}$ is the highest. In such a case, each user is assigned the same amount of time if the *normalized* fading statistics of the different users are identical – e.g., all users are Rayleigh fading.

A key requirement for scheduling is that the BS has knowledge of the instantaneous downlink channel. As we have discussed in Section 20.1.6, in an FDD system this requires that the MSs feed back their instantaneous SNRs to the BS. The rate of this feedback depends on the coherence time of the channel. Furthermore, it also involves a system design tradeoff. If feedback is done too rarely, then the BS picks a suboptimum user. If feedback is done too often, then the overhead for this feedback becomes prohibitive. It is also noteworthy that *all* MSs have to feed back their SNRs, even though only one will ultimately be picked for communication.¹⁰

All the above considerations were for the downlink; however, the scheme works for the uplink as well. In this case, the BS determines the quality of the uplink channel, and then broadcasts which MS is allowed to transmit.

Delay Considerations and Random Beamforming

Another key problem of multiuser diversity is the inherent delay. If the BS always uses proportionally fair scheduling, it retains communication with the chosen user for approximately one coherence time of the channel. If user mobility is low, the coherence time can be quite large. Remember that

¹⁰ Some tricks can be used to alleviate this problem: e.g., an MS that sees an SNR below a certain threshold does not need to feed its information back since the chance that it will be picked is very low. Again, the choice of this threshold involves an engineering tradeoff: if the threshold is too low, it does not lead to significant savings; if it is too high, then there are times when no MS provides feedback, and the BS thus cannot pick the best user.

each user has to wait until its $\gamma/\bar{\gamma}$ becomes instantaneously largest and it can communicate; this waiting time (latency) is roughly KT_{coh} . In a system with many users and large coherence times, this latency can become prohibitively large.

An ingenious solution to this problem, called *random beamforming*, uses multiple antennas at the BS [Viswanath et al. 2002]. Signals transmitted from the different antenna elements of the BS are multiplied by time-varying complex coefficients. This can be interpreted in two (equivalent) ways:

1. Each vector of transmit weight coefficients can be associated with a beam pattern. No matter what beam is formed, it will enhance the channel to *some* MS. Scheduling (as discussed above) then makes sure that the enhanced MS is chosen for communication. Every time the coefficients are changed, the beam pattern changes, and a different MS gets enhanced.
2. The combination of multiple antennas (with weighting coefficients) and physical channel can be viewed as an “equivalent” channel. By varying the coefficients for antenna elements, we are enforcing time variations on the channel, and thus reducing the coherence time T_{coh} . The effective channel exhibits Rayleigh fading with a coherence time that is approximately the smaller of the coherence time of the physical channel and the time over which antenna weights change.

The main difference from transmit diversity or beamforming (as discussed at the beginning of this section) is the following:

- For *conventional transmit diversity*, the BS needs to know the amplitude and phase of the transfer function from all MSs. It then chooses weights for the signals from different antenna elements in such a way that the different transmit signals add up in an optimum way at the destined MS. Coefficients are changed only if the transfer function between the BS and the chosen MS changes, or if the BS decides to communicate with another MS.
- For *random beamforming*, the BS chooses the coefficients completely at random and changes them according to system parameters (related to latency), but independently of the channel.

The time over which weighting coefficients are kept constant constitutes a compromise between the resulting latency and system overhead for feeding back channel quality.

Impact on System Design

Multiuser diversity is mainly useful for data communications. For voice, stringent delay requirements (total delay less than about 100 ms) preclude the use of “conventional” multiuser diversity, especially when users are stationary or slowly moving (pedestrian speeds), though random beamforming can alleviate this problem somewhat. On the other hand, the scheme seems very well suited to data communications, where larger delays are acceptable.

From a scientific point of view, it is also noteworthy that multiuser diversity leads to a paradigm change in physical-layer design. Many of the transceiver structures and signal-processing methods we have encountered in Chapters 10–20 aim at combating fading by *reducing* variations in SNR for a specific link. For multiuser diversity, however, we exploit and possibly enhance the variations in SNR.

20.2 Multiple Input Multiple Output Systems

20.2.1 Introduction

MIMO systems are systems with *Multiple Element Antennas* (MEAs) at *both* link ends. Originally suggested in Winters [1987], they attracted great attention through theoretical investigations in the

1990s ([Foschini and Gans 1998] and [Telatar 1999]). Since that time, research in these systems has exploded and practical systems based on MIMO have been developed.

The MEAs of a MIMO system can be used for four different purposes: (i) beamforming, (ii) diversity, (iii) interference suppression, and (iv) spatial multiplexing (transmission of several data streams in parallel). The first three concepts are the same as for smart antennas. Having multiple antennas at both link ends leads to some interesting new technical possibilities, but does not change the fundamental effects of this approach. Spatial multiplexing, on the other hand, is a new concept, and has thus drawn the greatest attention. It allows direct improvement of capacity by simultaneous transmission of multiple data streams. We will show below that the (information-theoretic) capacity for a single link increases linearly with the number of antenna elements.

In the early years of MIMO research, the main emphasis was on information-theoretic limits, and this section will also mostly concentrate on these aspects. After 2000, emphasis shifted more to the question of how the theoretical gains of MIMO can be realized in practice. Advances in practical implementation of MIMO systems have also greatly helped in their adoption by international standards organizations. MIMO was included in fourth-generation cellular systems (see Chapters 27 and 28) as well as high-throughput wireless Local Area Networks (LANs) (IEEE 802.11n, see Chapter 29).

20.2.2 How Does Spatial Multiplexing Work?

Spatial multiplexing uses MEAs at the TX for transmission of parallel data streams (see Figure 20.8). An original high-rate data stream is multiplexed into several parallel streams, each of which is sent from one transmit antenna element. The channel “mixes up” these data streams, so that each of the receive antenna elements sees a combination of them. If the channel is well behaved, the received signals represent *linearly independent* combinations. In this case, appropriate signal processing at the RX can separate the data streams. A basic condition is that the number of receive antenna elements is at least as large as the number of transmit data streams. It is clear that this approach allows the data rate to be drastically increased – namely, by a factor of $\min(N_t, N_r)$.

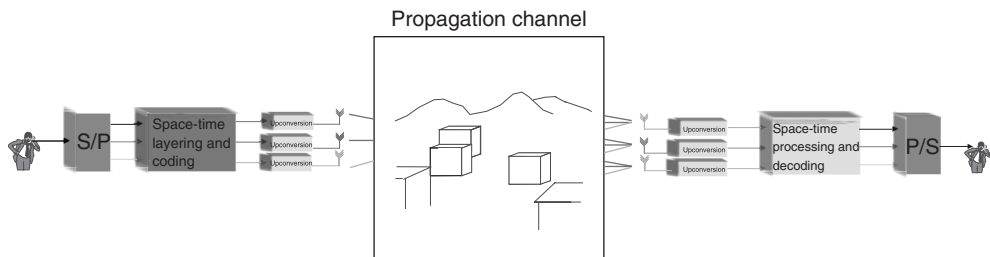


Figure 20.8 Principle behind spatial multiplexing.

For the case when the TX knows the channel, we can also develop another intuition (see Figure 20.9). With N_t transmit antennas, we can form N_t different beams. We point all these beams at different Interacting Objects (IOs), and transmit different data streams over them. At the RX, we can use N_r antenna elements to form N_r beams, and also point them at different IOs. If all the beams can be kept orthogonal to each other, there is no interference between the data streams; in other words, we have established parallel channels. The IOs (in combination with the beams pointing in their direction) play the same role as wires in the transmission of multiple data streams on multiple wires.

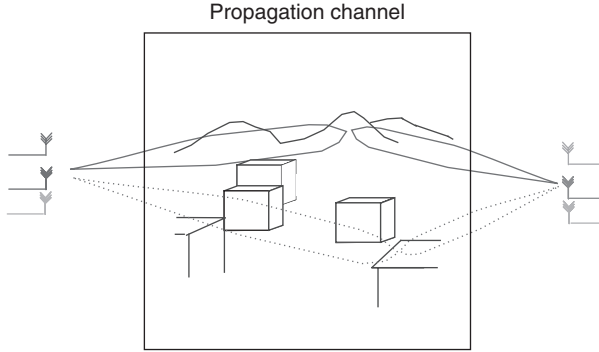


Figure 20.9 Transmission of different data streams via different interacting objects.

From this description, we can also immediately derive some important principles: the number of possible data streams is limited by $\min(N_t, N_r, N_s)$, where N_s is the number of (significant) IOs. We have already seen above that the number of data streams cannot be larger than the number of transmit antenna elements, and that we need a sufficient number of receive antenna elements (at least as many as data streams) to form the receive beams and, thus, be able to separate the data streams. But it is also very important to notice that the number of IOs poses an upper limit: if two data streams are transmitted to the same IO, then the RX has no possibility of sorting them out by forming different beams.

The above intuitive pictures are somewhat simplified. A more exact mathematical treatment follows in subsequent sections.

20.2.3 System Model

Before going into further details, let us first establish the generic system that will be considered for capacity computations. Figure 20.10 exhibits a block diagram. At the TX, the data stream enters an encoder, whose outputs are forwarded to N_t transmit antennas. From the antennas, the signal is sent through the wireless propagation channel, which is assumed to be quasi-static and frequency-flat if not stated otherwise. By quasi-static we mean that the coherence time of the channel is so long that “a large number” of bits can be transmitted within this time.

We denote the $N_r \times N_t$ matrix of the channel as

$$\mathbf{H} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1N_t} \\ h_{21} & h_{22} & \cdots & h_{2N_t} \\ \vdots & \vdots & \cdots & \vdots \\ h_{N_r1} & h_{N_r2} & \cdots & h_{N_rN_t} \end{pmatrix} \quad (20.28)$$

whose entries h_{ij} are complex channel gains (transfer functions) from the j th transmit to the i th receive antenna.

The received signal vector

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n} = \mathbf{x} + \mathbf{n} \quad (20.29)$$

contains the signals received by N_r antenna elements, where $\mathbf{s}$ is the transmit signal vector and $\mathbf{n}$ is the noise vector.

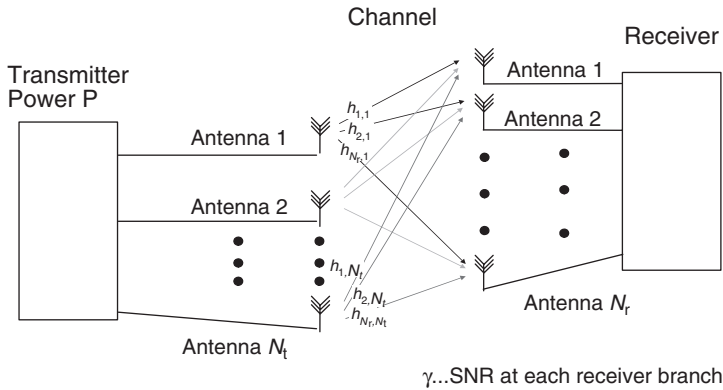


Figure 20.10 Block diagram of a multiple-input multiple-output system.

20.2.4 Channel State Information

Algorithms for MIMO transmission can be categorized by the amount of CSI that they require. We distinguish the following cases:

1. *Full CSI at the TX (CSIT) and full CSI at the RX (CSIR)*: in this ideal case, both the TX and the RX have full and perfect knowledge of the channel. This case obviously results in the highest possible capacity. However, it is difficult to obtain the full CSIT (as we have discussed in Section 20.1.6).
2. *Average CSIT and full CSIR*: in this case, the RX has full information of the instantaneous channel state, but the TX knows only the average CSI – e.g., the correlation matrix of $\mathbf{H}$ or the angular power spectrum. As we have discussed in Section 20.1.6, this is easier to achieve and does not require reciprocity or fast feedback; however, it does require calibration (to eliminate the nonreciprocity of transmit and receive chains) or slow feedback.
3. *No CSIT and full CSIR*: this is the case that can be achieved most easily, without any feedback or calibration. The TX simply does not use any CSI, while the RX learns the instantaneous channel state from a training sequence or using blind estimation.
4. *Noisy CSI*: when we assume “full CSI” at the RX, this implies that the RX has learned the channel state perfectly. However, any received training sequence will be affected by additive noise as well as quantization noise. It is thus more realistic to assume a “mismatched RX,” where the RX processes the signal based on the *observed* channel $\mathbf{H}_{\text{obs}}$, while in reality the signals pass through channel $\mathbf{H}_{\text{true}}$

$$\mathbf{H}_{\text{true}} = \mathbf{H}_{\text{obs}} + \mathbf{\Delta} \quad (20.30)$$

Some papers have taken this into account by ad hoc modification of noise variance (replacing σ_n^2 by $\sigma_n^2 + \sigma_e^2$, where σ_e^2 is the variance of the entries of $\mathbf{\Delta}$).

5. *No CSIT and no CSIR*: it is remarkable that channel capacity is also high when neither the TX nor the RX have CSI. We can, e.g., use a generalization of differential modulation. For high SNR, capacity no longer increases linearly with $m = \min(N_t, N_r)$, but rather increases as $\tilde{m}(1 - \tilde{m}/T_{\text{coh}})$, where $\tilde{m} = \min(N_t, N_r \lfloor T_{\text{coh}}/2 \rfloor)$, and T_{coh} is the coherence time of the channel in units of symbol duration.

20.2.5 Capacity in Nonfading Channels

The first key step in understanding MIMO systems is the derivation of the capacity equation for MIMO systems in nonfading channels, often known as “Foschini’s equation” [Foschini and Gans 1998]. Let us start with the capacity equation for “normal” (single-antenna) Additive White Gaussian Noise (AWGN) channels. As Shannon showed, the information-theoretic (ergodic) capacity of such a channel is (see also Chapter 14)

$$C_{\text{shannon}} = \log_2 (1 + \gamma \cdot |H|^2) \quad (20.31)$$

where γ is the SNR at the RX, and H is the normalized transfer function from the TX to the RX (as we are for now dealing with the frequency-flat case, the transfer function is just a scalar number). The key statement of this equation is that capacity increases only logarithmically with the SNR, so that boosting the transmit power is a highly ineffective way of increasing capacity.

Consider now the MIMO case, where the channel is represented by matrix Eq. (20.28). Let us then consider a *singular value decomposition*¹¹ of the channel:

$$\mathbf{H} = \mathbf{W}\mathbf{\Sigma}\mathbf{U}^\dagger \quad (20.32)$$

where $\mathbf{\Sigma}$ is a diagonal matrix containing singular values, and $\mathbf{W}$ and $\mathbf{U}^\dagger$ are unitary matrices composed of the left and right singular vectors, respectively. The received signal is then

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n} \quad (20.33)$$

$$= \mathbf{W}\mathbf{\Sigma}\mathbf{U}^\dagger\mathbf{s} + \mathbf{n} \quad (20.34)$$

Then, multiplication of the transmit data vector by matrix $\mathbf{U}$ and the received signal vector by $\mathbf{W}^\dagger$ diagonalizes the channel:

$$\begin{aligned} \mathbf{W}^\dagger\mathbf{r} &= \mathbf{W}^\dagger\mathbf{W}\mathbf{\Sigma}\mathbf{U}^\dagger\mathbf{U}\tilde{\mathbf{s}} + \mathbf{W}^\dagger\mathbf{n} \\ \tilde{\mathbf{r}} &= \mathbf{\Sigma}\tilde{\mathbf{s}} + \tilde{\mathbf{n}} \end{aligned} \quad (20.35)$$

Note that – because $\mathbf{U}$ and $\mathbf{W}$ are unitary matrices – $\tilde{\mathbf{n}}$ has the same statistical properties as $\mathbf{n}$ – i.e., it is independent identically distributed (iid) white Gaussian noise. The capacity of the system Eq. (20.35) is the same as that of system Eq. (20.29). Now, computation of the capacity of Eq. (20.35) is rather straightforward. The matrix $\mathbf{\Sigma}$ is a diagonal matrix with R_H nonzero entries σ_k , where R_H is the rank of $\mathbf{H}$ (and thus defined as the number of nonzero singular values), and σ_k is the k th singular value of $\mathbf{H}$. We have therefore R_H *parallel* channels (eigenmodes of the channel), and it is clear that the capacity of parallel channels just adds up.

The capacity of channel $\mathbf{H}$ is thus given by the sum of the capacities of the eigenmodes of the channel:

$$C = \sum_{k=1}^{R_H} \log_2 \left[1 + \frac{P_k}{\sigma_n^2} \sigma_k^2 \right] \quad (20.36)$$

where σ_n^2 is noise variance, and P_k is the power allocated to the k th eigenmode; we assume that $\sum P_k = P$ is independent of the number of antennas. This capacity expression can be shown to

¹¹ Singular value decomposition is similar to eigenvalue decomposition, but exists also for rectangular matrices (more rows than columns, or vice versa). It decomposes any matrix into a product of three matrices: a unitary matrix corresponding to the row space, a diagonal matrix describing the strength of different eigenmodes, and a unitary matrix describing the column space.

be equivalent to

$$C = \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\bar{\gamma}}{N_t} \mathbf{H} \mathbf{R}_{ss} \mathbf{H}^\dagger \right) \right] \quad (20.37)$$

where $\mathbf{I}_{N_r}$ is the $N_r \times N_r$ identity matrix, $\bar{\gamma}$ is the mean SNR per RX branch, and $\mathbf{R}_{ss}$ is the correlation matrix of the transmit data (if data at the different antenna elements are uncorrelated, it is a diagonal matrix with entries that describe the power distribution among antennas).¹² The distribution of power among the different eigenmodes (or antennas) depends on the amount of CSIT; we also assume for the moment that the RX has perfect CSI. The equations above confirm our intuitive picture that capacity increases linearly with $\min(N_t, N_r, N_s)$, as the number of nonzero singular values R_H is upper-limited by $\min(N_t, N_r, N_s)$.

No Channel State Information at the Transmitter and Full CSI at the Receiver

When the RX knows the channel perfectly, but no CSI is available at the TX, it is optimum to assign equal transmit power to all TX antennas, $P_k = P/N_t$, and use uncorrelated data streams. Capacity thus takes on the form:

$$C = \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\bar{\gamma}}{N_t} \mathbf{H} \mathbf{H}^\dagger \right) \right] \quad (20.38)$$

It is worth noting that (for sufficiently large N_s) the capacity of a MIMO system increases linearly with $\min(N_t, N_r)$, irrespective of whether the channel is known at the TX or not.

Let us now look at a few special cases. To make the discussion easier, we assume that $N_t = N_r = N$:

1. All transfer functions are identical – i.e., $h_{1,1} = h_{1,2} = \dots = h_{N,N}$. This case occurs when all antenna elements are spaced very closely together, and all waves are coming from similar directions. In such a case, the rank of the channel matrix is unity. Then, capacity is

$$C_{\text{MIMO}} = \log_2(1 + N\bar{\gamma}) \quad (20.39)$$

We see that in this case the SNR is increased by a factor of N compared with the single antenna case, due to beamforming gain at the RX. However, this only leads to a logarithmic increase in capacity with the number of antennas.

2. All transfer functions are different such that the channel matrix is full rank, and has N eigenvalues of equal magnitude. This case can occur when the antenna elements are spaced far apart and are arranged in a special way. In this case, capacity is

$$C_{\text{MIMO}} = N \log_2(1 + \bar{\gamma}) \quad (20.40)$$

and, thus, increases linearly with the number of antenna elements.

3. Parallel transmission channels – e.g., parallel cables. In this case, capacity also increases linearly with the number of antenna elements. However, the SNR per channel decreases with N , so that total capacity is

$$C_{\text{MIMO}} = N \log_2 \left(1 + \frac{\bar{\gamma}}{N} \right) \quad (20.41)$$

Figure 20.11 shows capacity as a function of N for different values of SNR.

¹² Note that the $\mathbf{H}$ and $\mathbf{R}$ must be normalized to ensure that $\bar{\gamma}$ is the mean SNR.

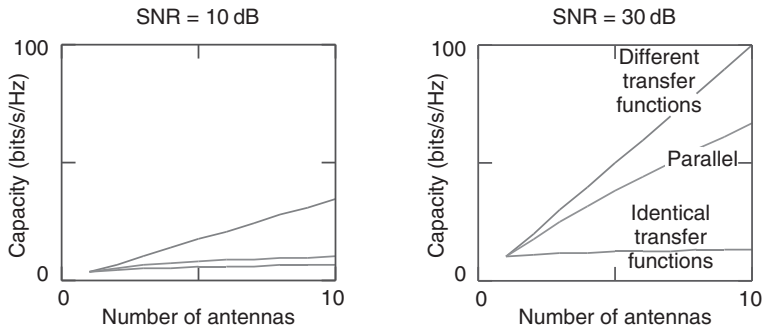


Figure 20.11 Capacity of multiple-input-multiple-output systems in additive white Gaussian noise channels.

Full Channel State Information at the Transmitter and Full CSI at the Receiver

Let us next consider the case where both the RX and TX know the channel perfectly. In such a case, it can be more advantageous to distribute power not uniformly between the different transmit antennas (or eigenmodes) but rather assign it based on the channel state. In other words, we are faced with the problem of optimally allocating power to several parallel channels, each of which has a different SNR. This is the same problem that we considered in Section 19.7.2, and therefore the answer is the same: *waterfilling*. This is another nice example of how the same mathematics can be applied to different communications problems: replace the word “subchannel” or “subcarrier in an Orthogonal Frequency Division Multiplexing (OFDM) system” by “eigenmode in a MIMO system,” and we can apply the whole discussion of Section 19.7.2 to MIMO systems with CSIT.

20.2.6 Capacity in Flat-Fading Channels

General Concepts

In the previous section, we considered capacity for one given channel realization – i.e., for one channel matrix $\mathbf{H}$. In wireless systems, we are, however, faced with channel fading. In this case, entries in channel matrix Eq. (20.28) are random variables. If the channel is Rayleigh fading, and fading is independent at different antenna elements, the h_{ij} are iid zero-mean, circularly symmetric complex Gaussian random variables with unit variance – i.e., the real and imaginary part each has variance 1/2. This is the case we will consider for now, unless stated otherwise. Consequently, the power carried by each h_{ij} is chi-square-distributed with 2 degrees of freedom. This is the simplest possible channel model; it requires the existence of “heavy multipath” – i.e., many MPCs of approximately equal strength (see Chapter 5) as well as a sufficient distance between the antenna elements. Since fading is independent, there is a high probability that the channel matrix is full rank and the eigenvalues are fairly similar to each other; consequently, capacity increases linearly with the number of antenna elements. Thus, the existence of heavy multipath, which is usually considered a drawback, becomes a major advantage in MIMO systems.

Because the entries of the channel matrix are random variables, we also have to rethink the concept of information-theoretic capacity. As a matter of fact, two different definitions of capacity exist for MIMO systems:

- *Ergodic (Shannon) capacity*: this is the expected value of the capacity, taken over all realizations of the channel. This quantity assumes an infinitely long code that extends over all the different channel realizations.
- *Outage capacity*: this is the minimum transmission rate that is achieved over a certain fraction of the time – e.g., 90% or 95%. We assume that data are encoded with a near-Shannon-limit-achieving code that extends over a period that is much shorter than the channel coherence time. Thus, each channel realization can be associated with a (Shannon) capacity value. Capacity thus becomes a random variable (rv) with an associated cumulative distribution function (cdf); see also the discussion in Section 14.9.1.

No Channel State Information at the Transmitter and Perfect CSI at the Receiver

Now, what is the capacity that we can achieve in a fading channel without CSI? Figure 20.12 shows the result for some interesting systems at a 21-dB SNR. The (1, 1) curve describes a Single Input Single Output (SISO) system. We find that the median capacity is on the order of 6 bit/s/Hz, but the 5% outage capacity is considerably lower (on the order of 3 bit/s/Hz). When using a (1, 8) system – i.e., 1 transmit antenna and 8 receive antennas – the mean capacity does not increase that significantly – from 6 to 10 bit/s/Hz. However, the 5% outage capacity increases significantly from 3 to 9 bit/s/Hz. The reason for this is the much higher resistance to fading that such a diversity system has. However, when going to a (8, 8) system – i.e., a system with 8 transmit and 8 receive antennas – both capacities increase dramatically: the mean capacity is on the order of 46 bit/s/Hz, and the 5% outage probability is more than 40 bit/s/Hz.

The exact expression for the *ergodic capacity* was derived in [Telatar 1999] as

$$E\{C\} = \int_0^\infty \log_2 \left[1 + \frac{\overline{\gamma}}{N_t} \lambda \right] \sum_{k=0}^{m-1} \frac{k!}{(k+n-m)!} [L_k^{n-m}(\lambda)]^2 \lambda^{n-m} \exp(-\lambda) d\lambda \quad (20.42)$$

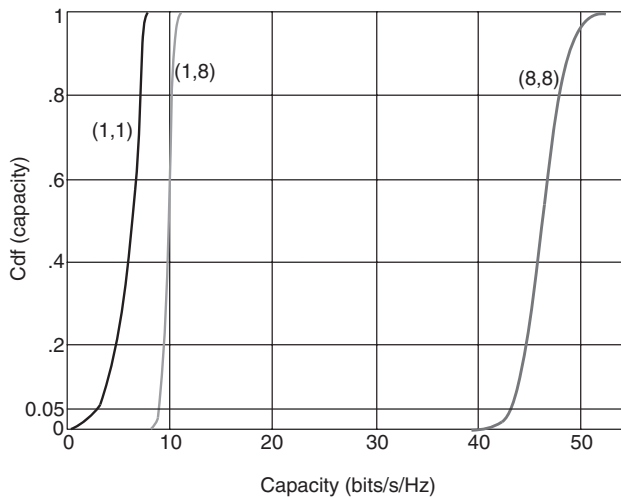


Figure 20.12 Cumulative distribution function of capacity for 1×1 , 1×8 , and the 8×8 optimum scheme.

Reproduced with permission from Foschini and Gans [1998] © Kluwer.

where $m = \min(N_t, N_r)$ and $n = \max(N_t, N_r)$ and $L_k^{n-m}(\lambda)$ are associated Laguerre polynomials. Exact analytical expressions for the cdfs of capacity are rather complicated; therefore, two approximations are in widespread use:

- Capacity can be well approximated by a Gaussian distribution, such that only the mean – i.e., the ergodic capacity given above – and variance need to be computed.
- From physical considerations, the following upper and lower bounds for capacity distribution have been derived in [Foschini and Gans 1998] for the case $N_t \geq N_r$:

$$\sum_{k=N_t-N_r+1}^{N_t} \log_2 \left[1 + \frac{\bar{\gamma}}{N_t} \chi_{2k}^2 \right] < C < \sum_{k=1}^{N_t} \log_2 \left[1 + \frac{\bar{\gamma}}{N_t} \chi_{2N_r}^2 \right] \quad (20.43)$$

where χ_{2k}^2 is a chi-square-distributed random variable with $2k$ degrees of freedom.¹³ These two bounds have very clear physical interpretations. The lower bound corresponds to capacity that can be achieved with a Bell labs LAYered Space Time (BLAST) system; this system and its operating principle are described below. The upper bound corresponds to an idealized situation where there is a separate array of receive antennas for each transmit antenna; it receives the signal in such a way that there is no interference from other transmit streams. As can be seen from Figure 20.13, the lower bound is fairly tight, while the upper bound can become rather loose especially for large number of antennas.

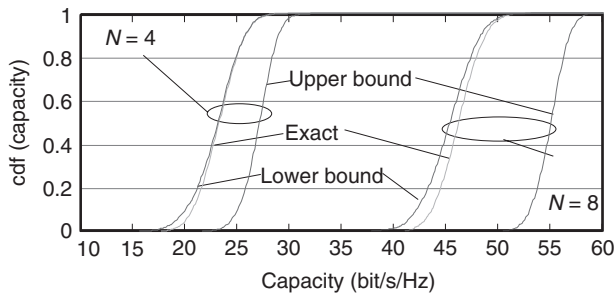


Figure 20.13 Exact capacity, upper bound, and lower bound of a multiple-input-multiple-output system in an independent identically distributed channel at a 21-dB signal-to-noise ratio, with $N_r = N_t = N$ equal to 4 and 8.

Reproduced with permission from Molisch and Tufvesson [2005] © Hindawi.

Perfect Channel State Information at the Transmitter and Receiver

The capacity gain by waterfilling (compared with equal-power distribution) is rather small when the number of transmit and receive antennas is identical. This is especially true in the limit of large SNRs: when there is a lot of water available, the height of “concrete blocks” in the vessel has little influence on the total amount that ends up in the vessels. When N_t is larger than N_r , the benefits of waterfilling become more pronounced (see Figure 20.14). We can interpret this the following way: if the TX has no channel knowledge, then there is little point in having more transmit than receive antennas – the number of data streams is limited by the number of receive antennas. Of course, we can transmit the same data stream from multiple transmit antennas, but this does not

¹³ Equation 20.43 is a slight abuse of notation, indicating as it does that the cdf of the capacity is bounded by the cdfs of the random variables given by the equations on the left and right sides.

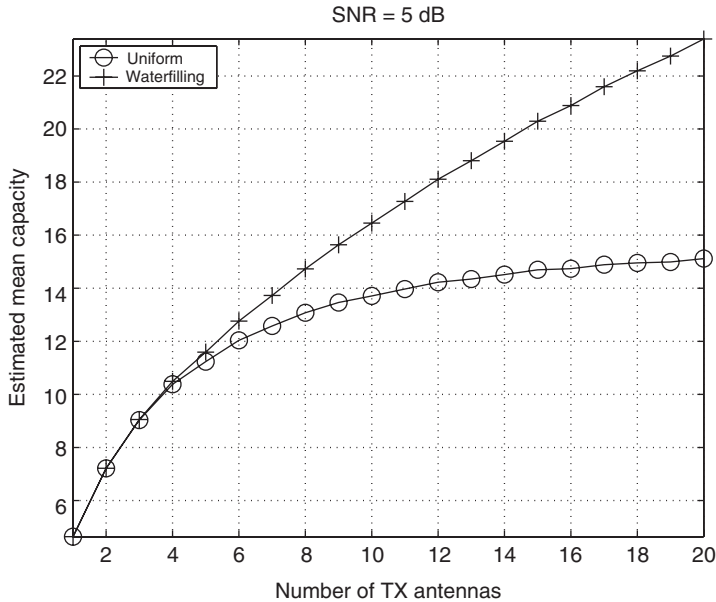


Figure 20.14 Capacity with and without channel state information at the transmitter with $N_t = 8$ antennas and a signal-to-noise ratio of 5 dB.

Reproduced with permission from Molisch and Tufvesson [2005] © Hindawi.

increase the SNR for that stream at the RX; without channel knowledge at the TX, the streams add up incoherently at the RX.

On the other hand, if the TX has full channel knowledge, it can perform beamforming, and direct the energy better toward the receive array. Thus, increasing the number of TX antennas improves the SNR, and (logarithmically) capacity. Thus, having a larger TX array improves capacity. However, this also increases the demand for channel estimation.

20.2.7 Impact of the Channel

Up to now, we have discussed the capacity of a channel with flat-fading and iid zero-mean complex Gaussian coefficients. Channels occurring in practice are more complicated, and deviations of the channel from idealized assumptions can have a significant impact on capacity. In the following, we describe some of the more important effects.

Channel Correlation

Correlation of the signals at different antenna elements can significantly reduce the capacity of a MIMO system. This can be shown the following way: the capacity is determined by the distribution of the singular values of the channel matrix. For a given SNR, maximum capacity is achieved when the channel transfer matrix has full rank and the singular values of $\mathbf{H}$ are equally strong. If the coefficients of the channel matrix are iid Rayleigh fading, then this situation is *approximately* fulfilled (though the ordered eigenvalues still have different values). But if the fading of the channel coefficients is correlated, then the singular value spread – i.e., the difference between the largest

and the smallest singular values – becomes much bigger. This, in turn, leads to a reduction in system capacity because some of the “parallel channels” of Eq. (20.36) have extremely low SNRs.

Correlation is influenced by the angular spectrum of the channel as well as the arrangement and spacing of antenna elements (see Chapter 13). For antennas that are spaced half a wavelength apart, a uniform angular power spectrum leads approximately to a decorrelation of incident signals. A smaller angular spread of the channel leads to an increase in correlation. Since we are now looking at a MIMO system, we have to consider correlation both at the TX and at the RX. A popular model (the so-called Kronecker model) assumes that correlation at the TX is independent of correlation at the RX (see Section 7.4.7).¹⁴ Realization of the channel matrix can then be obtained as

$$\mathbf{H}_{\text{kron}} = \frac{1}{E\{\text{tr}(\mathbf{H}\mathbf{H}^\dagger)\}} \mathbf{R}_{\text{RX}}^{1/2} \mathbf{G}_G \mathbf{R}_{\text{TX}}^{1/2} \quad (20.44)$$

where $\mathbf{G}_G$ is a matrix with iid complex Gaussian entries with unit variance.

Analytical computation of capacity is much more complicated in the case of correlated channels. It is easier to obtain simulation results by generating realizations of the channel matrix from Eq. (20.44), and then inserting them into Eq. (20.38) or Eq. (20.37).

Figure 20.15 shows the ergodic capacity of a 4×4 MIMO system with uniform linear arrays at the TX and RX as a function of angular spread at one link end.¹⁵ We see that a small angular spread leads to a drastic reduction in capacity.

Frequency-Selective Channels

The previous sections assumed frequency-flat channels. Fortunately, generalization to the frequency-selective case is straightforward. As Shannon has shown, the use of an OFDM-like scheme is optimum for dealing with the frequency-selective channel, converting it into a number of parallel flat-fading channels. Thus, capacity per unit bandwidth is a straightforward generalization of Eq. (20.37)

$$C = \frac{1}{B} \int_B \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\overline{\gamma}}{N_t} \mathbf{H}(f) \mathbf{R}_{ss}(f) \mathbf{H}(f)^\dagger \right) \right] df \quad (20.45)$$

where B is the bandwidth of the considered system.

This equation also implies that frequency selectivity offers additional diversity that increases the slope of the capacity cdf. If one of the frequency subchannels shows poor capacity, there is a good chance that another subchannel has high quality. Figure 20.16 shows an example of a measured capacity cdf in a microcellular environment. We see that the capacity cdf becomes steeper as the bandwidth increases.

Line-of-Sight Versus Non-Line-of-Sight

In some situations, there is an LOS connection between the TX and RX, resulting in different fading statistics. As we saw in Chapter 5, the fading statistics of an SISO link becomes Ricean instead of Rayleigh. For a MIMO system, the channel matrix can be written as

$$\mathbf{H} = \sqrt{\frac{K_{\text{LOS}}}{K_{\text{LOS}} + 1}} \hat{\mathbf{H}}_{\text{LOS}} + \sqrt{\frac{1}{K_{\text{LOS}} + 1}} \hat{\mathbf{H}}_{\text{res}} \quad (20.46)$$

¹⁴ For a discussion of this assumption, and a more general model, see Weichselberger et al. [2006]

¹⁵ Note that this figure is based on the idealized assumption that there is no mutual coupling between antenna elements; investigations have shown that mutual coupling influences capacity by introducing pattern diversity as well as changing the average power received in different antenna elements.

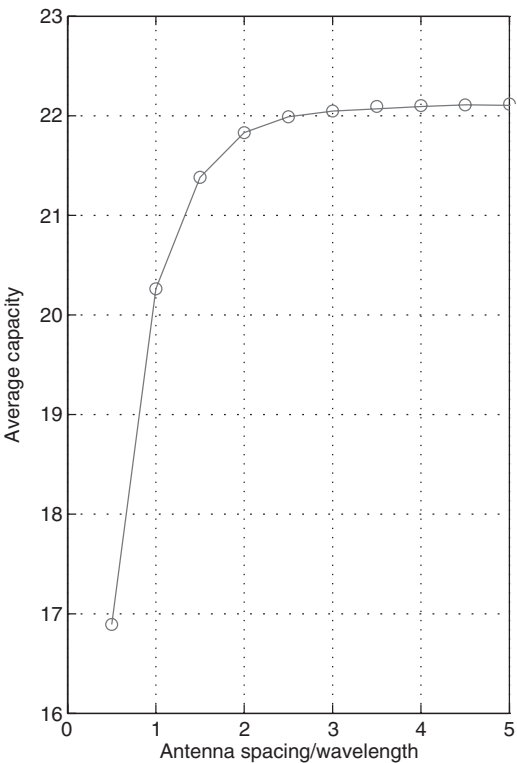


Figure 20.15 Average capacity of a 4×4 multiple-input-multiple-output system with a 10° root-mean-square angular spread as seen from the base station as a function of antenna spacing at the BS.
Reproduced with permission from Molisch and Tufvesson [2005] © Hindawi.

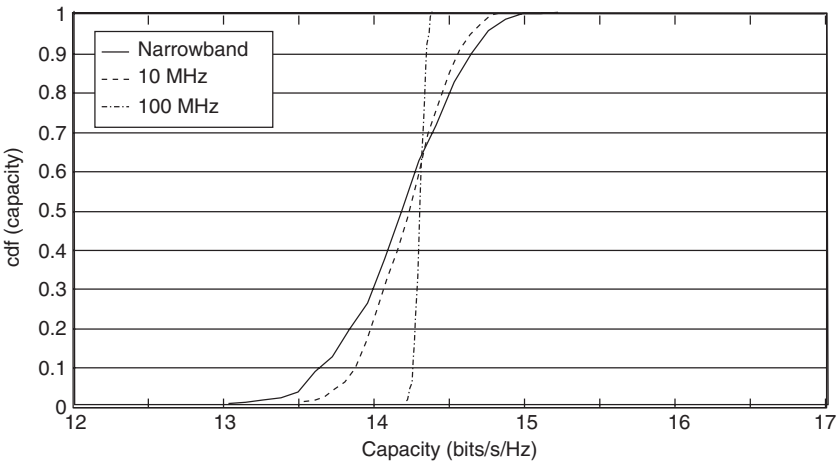


Figure 20.16 Capacity in a measured microcellular channel, for different system bandwidths: 4×4 multiple-input-multiple-output system.
Reproduced with permission from Molisch et al. [2002] © IEEE.

where K_{LOS} is the ratio of powers in the LOS to those in residual components, $\hat{\mathbf{H}}_{\text{LOS}}$ is a purely deterministic matrix, and $\hat{\mathbf{H}}_{\text{res}}$ has (uncorrelated or correlated) zero-mean Gaussian entries.¹⁶ If the distance between the TX and RX is large – i.e., larger than the Rayleigh distance (see Chapter 4) – the LOS gives rise to a matrix $\hat{\mathbf{H}}_{\text{LOS}}$ that has rank 1 (a single wave can only be associated with one singular value of the channel matrix!). This in turn implies that the singular value spread of matrix Eq. (20.46) is much larger than for a Non Line Of Sight (NLOS) matrix. Consequently, the capacity of a LOS channel is lower than for an NLOS channel *when assuming equal SNR*. One should however note that the SNR is often better in the LOS case compared with the NLOS case. For a power-limited scenario with realistic channels, the LOS case often gives the highest capacity, despite the imbalance between singular values.

It is also noteworthy that a strong LOS component leads to a larger spread of eigenvalues if the LOS component is a *plane* wave. A *spherical* wave leads to a transfer function matrix that can have full rank if antenna elements are spaced appropriately. The curvature of waves is noticeable up to one Rayleigh distance – i.e., typically a few meters.

Example 20.4 *Capacity in a channel with LOS. The transmit and receive arrays are uniform linear arrays with element spacing λ between the elements and $N_r = N_t = 8$. The arrays are perpendicular to the LOS connection. The DOA of the NLOS components is uniformly distributed between 0 and 2π . Estimate the mean capacity if the TX does not have CSI for an SNR of 20 dB, and $K_{\text{LOS}} = 0$ and 20 dB.*

As a first step, we have to determine the channel matrix. Since the transmit and receive arrays are linear arrays that are oriented perpendicular to the LOS, we find that $\hat{\mathbf{H}}_{\text{LOS}}$ is the all-1's matrix if transmit and receive arrays are sufficiently far apart from each other. Furthermore, $\hat{\mathbf{H}}_{\text{res}}$ has unit energy, iid complex Gaussian entries because the angular spectrum is uniformly distributed and the antenna elements are more than $\lambda/2$ apart from each other. From Eq. (20.38), capacity is then:

$$C = \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\bar{\gamma}}{N_t} \mathbf{H} \mathbf{H}^\dagger \right) \right] \quad (20.47)$$

$$= \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\bar{\gamma}}{N_t} \left[\frac{K_{\text{LOS}}}{K_{\text{LOS}} + 1} \hat{\mathbf{H}}_{\text{LOS}} \hat{\mathbf{H}}_{\text{LOS}}^\dagger + \frac{\sqrt{K_{\text{LOS}}}}{K_{\text{LOS}} + 1} (\hat{\mathbf{H}}_{\text{res}} \hat{\mathbf{H}}_{\text{LOS}}^\dagger + \hat{\mathbf{H}}_{\text{LOS}} \hat{\mathbf{H}}_{\text{res}}^\dagger) + \frac{1}{K_{\text{LOS}} + 1} \hat{\mathbf{H}}_{\text{res}} \hat{\mathbf{H}}_{\text{res}}^\dagger \right] \right) \right] \quad (20.48)$$

Using Jensen's inequality, the expected value of the capacity can be approximated as

$$E\{C\} \simeq \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\bar{\gamma}}{N_t} \left[\frac{K_{\text{LOS}}}{K_{\text{LOS}} + 1} E\{\hat{\mathbf{H}}_{\text{LOS}} \hat{\mathbf{H}}_{\text{LOS}}^\dagger\} \right. \right. \right. \quad (20.49)$$

$$\left. \left. + \frac{\sqrt{K_{\text{LOS}}}}{K_{\text{LOS}} + 1} E\{\hat{\mathbf{H}}_{\text{res}} \hat{\mathbf{H}}_{\text{LOS}}^\dagger + \hat{\mathbf{H}}_{\text{LOS}} \hat{\mathbf{H}}_{\text{res}}^\dagger\} + \frac{1}{K_{\text{LOS}} + 1} E\{\hat{\mathbf{H}}_{\text{res}} \hat{\mathbf{H}}_{\text{res}}^\dagger\} \right] \right] \quad (20.50)$$

$$= \log_2 \left[\det \left(\mathbf{I}_{N_r} + \frac{\bar{\gamma}}{N_t} \left[\frac{K_{\text{LOS}} N_t}{K_{\text{LOS}} + 1} \mathbf{1} + \frac{1}{K_{\text{LOS}} + 1} E\{\hat{\mathbf{H}}_{\text{res}} \hat{\mathbf{H}}_{\text{res}}^\dagger\} \right] \right) \right] \quad (20.51)$$

$$= \log_2 \left[\det \left(\left[1 + \frac{\bar{\gamma}}{K_{\text{LOS}} + 1} \right] \mathbf{I}_{N_r} + \left[\bar{\gamma} \frac{K_{\text{LOS}}}{K_{\text{LOS}} + 1} \mathbf{1} \right] \right) \right] \quad (20.52)$$

¹⁶ Both $\hat{\mathbf{H}}_{\text{LOS}}$ and $\hat{\mathbf{H}}_{\text{res}}$ are normalized to $E\{|\hat{\mathbf{H}}|_F^2\} = N_t N_r$, where $|\hat{\mathbf{H}}|_F$ is the Frobenius norm of $\hat{\mathbf{H}}$.

where $\mathbf{1}$ is an $N_r \times N_r$ matrix where each entry is 1; such a matrix has only a single nonzero eigenvalue, whose magnitude is N_r . Equation (20.52) can be further approximated as

$$E\{C\} \simeq N_r \log_2 \left[1 + \frac{\bar{\gamma}}{K_{\text{LOS}} + 1} \right] + \log_2 \left[1 + \frac{\bar{\gamma} K_{\text{LOS}} N_r}{K_{\text{LOS}} + 1} \right] \quad (20.53)$$

Using Eq. (20.53), we find that capacity for $K_{\text{LOS}} = 0$ and 20 dB is 54 and 17 bit/s/Hz, respectively. Note that the above is a rather crude approximation, but it gives correct trends. Monte Carlo simulations give 40 and 15 bit/s/Hz, respectively.¹⁷

Limited Number of Interacting Objects

In the intuitive picture of Section 20.2.1, we have already seen that a certain number of IOs are required to act as relays for data streams. Relating this to the mathematics of the previous section, we note that a certain number of IOs are needed in order to guarantee that the channel coefficients in matrix $\mathbf{H}$ are independent. While the number of existing IOs is always large, the number of *significant* IOs might be limited in practice. After all, IOs that are too weak to provide appreciable SNR (and thus capacity) are not useful in carrying data streams.

Figure 20.17 shows some measurement results for capacity as a function of the number of antenna elements. Measurements were taken in a microcellular scenario where the number of IOs was rather small. We find that capacity, especially for the LOS scenario, does not increase linearly with the number of antenna elements N when N exceeds 4. This is a clear sign that the number of IOs limits the achievable capacity.

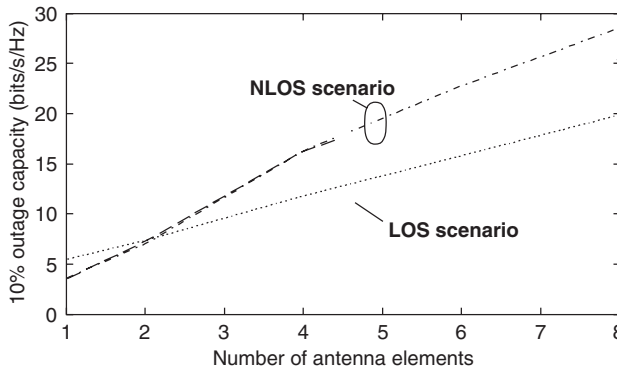


Figure 20.17 Ten percent outage capacity as a function of the number of antenna elements at the transmitter and RX in measured line-of-sight and non-line-of-sight scenarios.

Reproduced with permission from Molisch et al. [2002] © IEEE.

Keyhole Channels

There are some special cases where capacity is low even though the signals at antenna elements are uncorrelated. These cases are often referred to as *keyholes* or *pinholes*. An example of a keyhole scenario is a rich scattering environment at both the TX and RX ends; between them there is only

¹⁷ Note that the step from Eq. (20.52) to (20.53) involves approximations that are not detailed further here. It leads to a result that is somewhat similar to the result of [Ayadi et al. 2002]

one propagation path with just one degree of freedom. Such a scenario can occur when the TX and RX are surrounded by IOs, and the two IO groups are separated by a long stretch of empty space (greenfield). Another scenario is the case when IO areas are connected by a single-mode waveguide or by a diffraction edge. In all these cases, the total transfer function is proportional to

$$\mathbf{H} = \mathbf{R}_{\text{RX}}^{1/2} \mathbf{G}_{\text{G},1} \mathbf{R}_{\text{RX-TX}}^{1/2} \mathbf{G}_{\text{G},2} \mathbf{R}_{\text{TX}}^{1/2} \quad (20.54)$$

where $\mathbf{G}_{\text{G},1}$ and $\mathbf{G}_{\text{G},2}$ are both iid complex Gaussian matrices, and where $\mathbf{R}_{\text{RX-TX}}$ describes the correlation of the channel matrix between the TX and RX environments; in a keyhole channel, this is a low-rank matrix. It can also be seen from this description that the statistics of the entries are no longer Gaussian – this explains why it can be possible to have low correlation and low capacity at the same time. It should, however, be noted that keyhole channels occur very seldom in practice. While Almers et al. [2003] measured one in a controlled environment, it seems to be a rare occurrence in “normal” environments.

20.2.8 Layered Space–Time Structure

Up to now, we have only discussed the information-theoretic limits of MIMO systems. The big question is how to realize these capacities in practice. One possibility is joint encoding of the data streams that are to be transmitted from different antenna elements, combined with maximum-likelihood detection. When this technique is combined with (almost) capacity-achieving codes, it can closely approximate the capacity of a MIMO system. For a small number of antenna elements and for a small modulation alphabet (BPSK or Quadrature-Phase Shift Keying (QPSK)), such a scheme can actually be gainfully employed. However, for most practical cases, the complexity of a joint MLSE is rather high. For this reason, so-called layered space–time structures have been proposed, which allow us to break up the demodulation process into several separate pieces, each of which has lower complexity. These structures are also widely known under the name of *BLAST architectures*.

Horizontal BLAST

H-BLAST is the simplest possible layered space–time structure.¹⁸ The TX first demultiplexes the data stream into N_t parallel streams, each of which is encoded *separately*. Each encoded data stream is then transmitted from a different transmit antenna. The channel mixes up the different data streams; the RX separates them out by successive nulling and interference subtraction. In other words, the RX proceeds in the following steps (Figure 20.18):

- It considers the first data stream as the useful one, and regards the other data streams as interference. It can then use *optimum combining* for suppression of interfering streams (see Chapter 13). The RX has $N_r \geq N_t$ antenna elements available. If $N_r = N_t$, it can suppress all $N_t - 1$ interfering data streams and receive the desired data stream with diversity order 1. If the RX has more antennas, it can receive the first data stream with better quality. But in any case, interference from the other streams can be eliminated.
- The desired stream can now be demodulated and decoded. Outputs from that process are firm decisions on the bits of stream 1. Note that since we have separate encoding of different data streams, we only need knowledge of the first stream to complete the decoding process.

¹⁸ This scheme was originally called *V-BLAST* (for Vertical BLAST) but the name was changed later by Foschini et al. [2003].

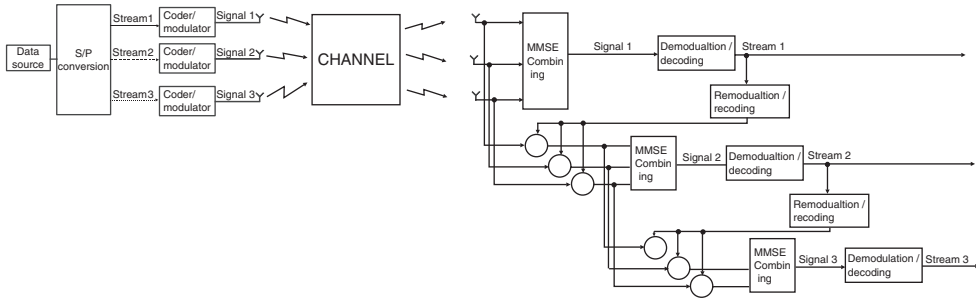


Figure 20.18 Block diagram of a horizontal BLAST transceiver.

- The bits that have thus been decoded are now re-encoded, and remodulated. Multiplying this symbol stream by the transfer function of the channel, we obtain the contribution that stream 1 has made to the total received signal at the different antenna elements.
- We subtract these contributions from the signals at the different antenna elements.
- Now we consider the “cleaned-up” signal and try to detect the second data stream. We again have N_r received signals, but only $N_t - 2$ interferers. Using optimum combining again, we can now receive the desired data stream with diversity order 2.
- The next step is again decoding, recoding, and remodulating the considered data stream (stream 2 now), and subtraction of the associated signal from the total signal at the receive antenna elements obtained in the previous step. This cleans up the received signal even more.
- The process is repeated until the last data stream is decoded.

This scheme is actually very similar to multiuser detection (Section 18.4): if different transmit streams were to come from different users, then H-BLAST would be normal serial interference cancellation. Note also that the encoding scheme does not require “cooperation” between different antenna elements (or users). Similar to serial interference cancellation, H-BLAST also faces the problem of error propagation, especially since the first decoded data stream has the worst quality. In other words: if data stream 1 is decoded incorrectly, then we subtract the “wrong” signal from the remaining signals at the antenna elements. Thus, instead of “cleaning up” the receive signal, we introduce even more interference. This in turn increases the likelihood that the second data stream is decoded incorrectly, and so on. In order to mitigate this problem, stream ordering should be used: the RX should first decode the stream that has the best SINR, then the one with the next best, and so on.

Diagonal BLAST

The main problem with H-BLAST is that it does not provide diversity. The first stream, which has diversity order 1, dominates the performance at high SNRs. A better performance can be achieved with the so-called D-BLAST scheme. In this approach, streams are cycled through the different transmit antennas, such that each stream sees all possible antenna elements. In other words, each single transmit stream is subdivided into a number of subblocks. The first subblock of stream 1 is transmitted from antenna 1, the next subblock from antenna 2, and so on (compare Figure 20.19).

Decoding can be done stream by stream; again, each decoded block can be subtracted from signals at the other antenna elements and, thus, enhances the quality of the residual signal.

The difference from H-BLAST is that each stream is sometimes in a “good” position in the sense that the other streams have already been subtracted, and thus the SINR is very high, while sometimes

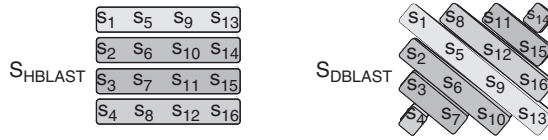


Figure 20.19 Assignment of bitstreams to different antennas for horizontal BLAST and diversity BLAST.

it is in a bad position, in the sense that it suffers full interference. Thus, each stream experiences full diversity. More precisely, each data stream alternately sees a channel with diversity order 1 (whose SNR has a pdf that is a chi-square distribution with 2 degrees of freedom χ_2^2), diversity order 2 (chi-square with 4 degrees of freedom, χ_4^2), and so on. Therefore, total capacity in the case when $N_t = N_r$:

$$C_{\text{D-BLAST}} = \sum_{k=1}^{N_t} \log_2 \left[1 + \frac{\bar{\gamma}}{N_t} \chi_{2k}^2 \right] < C \quad (20.55)$$

This actually achieves the lower capacity bound of Eq. (20.43) [Ariyavisitakul 2000].

Structures for Channel State Information at the Transmitter

When full CSI is available at the TX, transceiver schemes become much simpler, at least conceptually. By multiplying transmitted and received signal vectors by the right and left singular vectors of the channel matrix, respectively, diagonalization of the channel is achieved. Therefore, the different data streams do not interfere with each other; rather, we just have a number of parallel channels, each of which can be separately encoded and decoded. The difficulties lie instead in the practical aspects of obtaining and using CSIT (as discussed in Section 20.1.6).

20.2.9 Diversity

Multiple antennas can also be used to provide pure diversity. Again, we have to distinguish between systems that have CSIT and systems that do not. The former can achieve beamforming gain in addition to diversity gain, while the latter is restricted to achieving better resistance to fading.

Diversity with Channel State Information at the Transmitter

Having the CSIT leads to a conceptually simple transceiver structure in the diversity case, just as it did for spatial multiplexing. The transmit vector consists of weighted replica of a single data symbol, $\mathbf{s} = \mathbf{u}\mathbf{c}$. Consider again the singular value decomposition of the channel matrix, Eq. (20.34). We then find that the received signal is

$$\mathbf{r} = \mathbf{W}\mathbf{\Sigma}\mathbf{U}^\dagger \mathbf{u}\mathbf{c} + \mathbf{n} \quad (20.56)$$

The RX performs a linear combination (summation) of the signals at the different antenna elements, such that the output of the combiner is $\tilde{c} = \mathbf{w}\mathbf{r}$. Choosing $\mathbf{w} = (\mathbf{W})_1$, and $\mathbf{u} = (\mathbf{U})_1$ maximizes the SNR at the RX, where $(\mathbf{W})_1$ is the left singular vector corresponding to the largest singular value (and similarly for $\mathbf{U}$). In other words, we choose the transmit and receive weights according

to the singular value decomposition of the channel. The SNR that can be achieved this way is

$$\gamma = \frac{P}{\sigma_n^2} \tilde{\sigma}_{\max}^2 \quad (20.57)$$

where $\tilde{\sigma}_{\max}$ is the largest singular value of matrix $\mathbf{H}$. For the case when we have only one transmit antenna, the scheme reduces to maximum-ratio combining; for the case of a single receive antenna, it becomes maximum-ratio transmission. Figure 20.20 shows the cdf of the SNR that can be achieved this way. We can see that the slope of the largest eigenvalue is identical to the slope of the sum of the eigenvalue. In a 4×4 MIMO system, the diversity order is 16, and SNR distribution is almost a step function – in other words, the required fading margin is very small. We also find that there is an enhancement of the average SNR, but it is smaller than $N_r N_t = 16$ (12 dB) (see also Section 20.2.10).

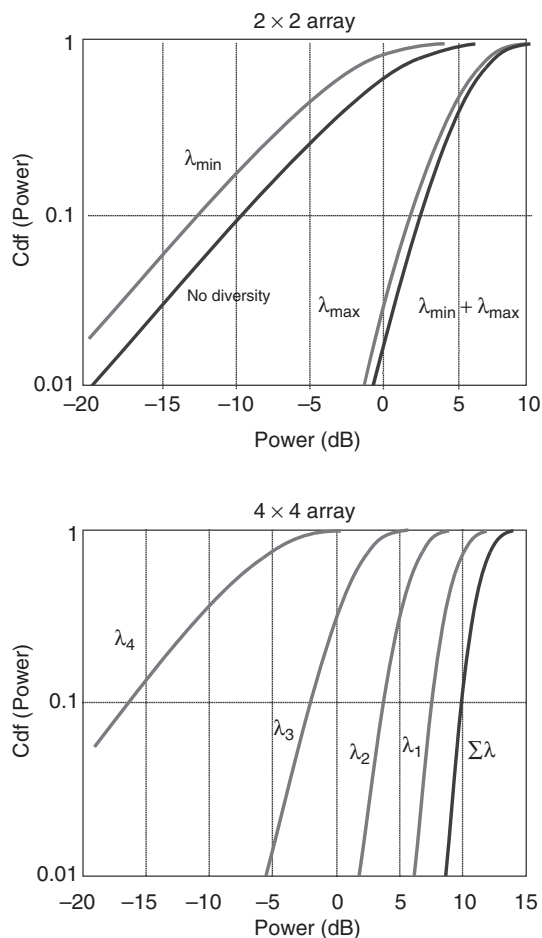


Figure 20.20 Cumulative distribution function of the signal-to-noise ratio in 2×2 (left) and 4×4 (right) multiple-input-multiple-output diversity with independent identically distributed Rayleigh fading at all antenna elements.

Example 20.5 Consider a 2×2 MIMO system with a 10-dB average SNR at each of the antenna elements. What is the probability that the SNR is smaller than 7 dB?

As shown in Andersen [2000], the pdf for the largest eigenvalue of matrix $\mathbf{H}\mathbf{H}^\dagger$ in a 2×2 system is

$$pdf_{\lambda_{\max}}(\lambda) = \exp(-\lambda)[\lambda^2 - 2\lambda + 2] - 2\exp(-2\lambda) \quad (20.58)$$

and the cdf is

$$cdf_{\lambda_{\max}}(x) = 1 - \exp(-x)(x^2 + 2) + \exp(-2x) \quad (20.59)$$

Here λ describes the normalized SNR. We now want to identify the probability that the largest eigenvalue is 3 dB below the mean value of $\mathbf{H}\mathbf{H}^\dagger$; i.e.,

$$cdf_{\lambda_{\max}}(0.5) = 3 \cdot 10^{-3} \quad (20.60)$$

Of course, we can also ascribe capacity to a diversity system. This capacity follows immediately from SNR statistics:

$$C = \log_2 \left[1 + \frac{P}{\sigma_n^2} \max_k (\tilde{\sigma}_k^2) \right] \quad (20.61)$$

Diversity Without Channel State Information at the Transmitter – Space–Time Coding

If the channel is unknown at the TX, then we try to transmit different versions of the data stream from the different transmit antennas. In Chapter 13, we have already encountered some methods that achieve this – especially, delay diversity. Another approach that has gained enormous attention is space–time coding. In this approach, redundancy is introduced by sending from each transmit antenna a differently encoded (and fully redundant) version of the same signal. There are multiple ways in which encoding can be done. In the following, we will first describe the *Alamouti* code, the most popular form of *space–time block codes*. Subsequently, we mention the basic principles of *space–time trellis codes*. The codes we consider here work independently of the number of receive antennas, and can thus be seen as a form of transmit diversity (see also Chapter 13).

Orthogonal Space–Time Block Codes

The idea behind Space Time Block Codes (STBCs) is to transmit data in a way that guarantees high diversity, while allowing a simple decoding process. The most popular STBC is the Alamouti code [Alamouti 1998]. Its idea is simple yet ingenious. Consider a flat-fading channel, where the complex channel gain from TX antenna 1 is given by h_1 and the gain from TX antenna 2 to the RX by h_2 . Now, transmit the two symbols, c_1 and c_2 , from the two transmit antennas at time instant 1:

$$\mathbf{s}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad (20.62)$$

At the second time instant, transmit the vector:

$$\mathbf{s}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -c_2^* \\ c_1^* \end{pmatrix} \quad (20.63)$$

The factor $1/\sqrt{2}$ stems from the necessity to keep the energy constant, as we are now using two transmit antennas. Then the received signal can be written as

$$\mathbf{r} = \begin{pmatrix} r_1 \\ r_2^* \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \mathbf{n} = \mathbf{H}\mathbf{c} + \mathbf{n} \quad (20.64)$$

We have thus created a “virtual” MIMO system. It is important to note that the columns of the “virtual” channel matrix $\mathbf{H}$ are orthogonal. Therefore, $\mathbf{H}^\dagger \mathbf{H}$ becomes a scaled identity matrix $\alpha \mathbf{I}$. For decoding, we thus first multiply the received signal vector by $\mathbf{H}^\dagger$. Then we get

$$\tilde{\mathbf{r}} = \mathbf{H}^\dagger \mathbf{r} = \mathbf{H}^\dagger \mathbf{H}\mathbf{c} + \mathbf{H}^\dagger \mathbf{n} = \alpha \mathbf{c} + \tilde{\mathbf{n}} \quad (20.65)$$

Since the columns of $\mathbf{H}$ are orthogonal, the components of $\tilde{\mathbf{n}}$ are still uncorrelated zero-mean Gaussian, and have variance $\alpha \sigma_n^2$. Therefore, decoding of data c_1 and c_2 becomes decoupled; which greatly reduces the computational effort in the RX.

As far as performance is concerned, we find that α describes the “effective” channel power gain, which is

$$\alpha = \frac{|h_1|^2 + |h_2|^2}{2} \quad (20.66)$$

Consequently, the diversity order is 2: both h_1 and h_2 would have to be in a fading dip for the effective channel gain to be low. We also see that the scheme can only increase diversity, but not beamforming gain, since it does not have CSIT.

There have been many attempts to generalize the Alamouti code to more than two transmit antennas. Unfortunately, it can be shown that orthogonal STBCs for more than two antennas have a rate smaller than 1 – in other words, we cannot even achieve the rate that we could have with a single-antenna system [Tarokh et al. 1999].

Space–Time Trellis Codes STBCs provide full diversity order, but no coding gain. Such coding gain can be obtained from Space Time Trellis Codes (STTCs). Given N_t transmit antennas, the STTC maps each symbol from the information source to a *vector* of N_t transmit symbols that are sent from the different antenna elements. Decoding requires a vector Viterbi decoder. The error probability – i.e., the probability of confusing one codeword $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_L)$ of length L_c with another codeword $\tilde{\mathbf{C}}$ – is upper-bounded by

$$P(\mathbf{C} \rightarrow \tilde{\mathbf{C}}) \leq \left(\prod_{i=1}^{R_c} \lambda_i \right)^{-N_t} \left(\frac{E_S}{4N_0} \right)^{-R_c N_t} \quad (20.67)$$

where R_c and λ_i are the rank and eigenvalues, respectively, of the error matrix:

$$\sum_{i=1}^{L_c} (\mathbf{c}_i - \tilde{\mathbf{c}}_i)(\mathbf{c}_i - \tilde{\mathbf{c}}_i)^\dagger \quad (20.68)$$

The term

$$\left(\prod_{i=1}^{R_c} \lambda_i \right)^{-N_t} \quad (20.69)$$

represents the coding gain, while the term

$$\left(\frac{E_S}{4N_0} \right)^{-R_c N_t} \quad (20.70)$$

describes diversity gain. In order to achieve full diversity, the rank of the error matrix should thus be as high as possible (*rank criterion*); in order to achieve high coding gain, the determinant of the error matrix should be maximized (*determinant criterion*). The rank criterion and the determinant criterion provide important guidelines on how to design STTCs.

The main drawback with STTCs is their complexity. The requirement for a vector Viterbi RX has proved to be a major obstacle in their application in practical systems.

20.2.10 Tradeoffs between Diversity, Beamforming Gain, and Spatial Multiplexing

MIMO systems can be used to achieve spatial multiplexing, diversity, and/or beamforming. However, it is not possible to attain all of those goals simultaneously at their full extent. First, we find that there is a tradeoff between beamforming and diversity gain; this tradeoff also depends on the environment in which we are operating. Consider first an LOS scenario. In this case, it is obvious that the achievable beamforming gain is $N_t N_r$: we form beams at the TX (with gain N_t) and at the RX (with gain N_r), and point them at each other. The gains thus multiply. On the other hand, there is obviously no diversity gain, since there is no fading in an LOS scenario – in other words, the slope of the SNR distribution curve does not change.

In a heavily scattering environment, the slope of SNR distribution changes drastically due to the use of multiantenna elements. If the channel is known at the TX, we can choose our antenna weights in such a way that it is unlikely that the receive signals are in a fading dip; furthermore, even if this happens, it is unlikely that *all* the receive signals are in a fading dip. In other words, it is easy to make sure that *at least one* receive signal has good quality. As discussed in Chapter 13, the diversity order is the slope of the BER versus SNR curve for very high SNRs:

$$d_{\text{div}} = - \lim_{\bar{\gamma} \rightarrow \infty} \frac{\log[BER(\bar{\gamma})]}{\log(\bar{\gamma})} \quad (20.71)$$

it can also be related to the slope of the SNR distribution at very low values of the SNR, and is thus a measure for the likelihood that all signals are in a bad fading dip simultaneously. The maximum diversity order in a heavily scattering environment can be shown to be $N_t N_r$. On the other hand, it also turns out that the maximum beamforming gain in such a heavily scattering environment is upper limited by $(\sqrt{N_t} + \sqrt{N_r})^2$. The reason is that it is not possible to form the transmit beam pattern in such a way that MPCs overlap constructively *at all receive antenna elements simultaneously*. Note the key tradeoff here between achieving full diversity order and beamforming gain.

There is also a fundamental tradeoff between spatial multiplexing and diversity. Zheng and Tse [2003] showed that the optimum tradeoff curve between diversity order and rate r is piecewise linear, connecting the points:

$$d_{\text{div}}(r) = (N_t - r)(N_r - r), \quad r = 0, \dots, \min(N_t, N_r) \quad (20.72)$$

This implies the maximum diversity order, $N_t N_r$, and maximum rate, $\min(N_t, N_r)$, cannot be achieved simultaneously.

20.2.11 Feedback for MIMO

As we have discussed previously, performance of MIMO systems can be improved by CSIT. Such channel knowledge might be obtainable from the principle of reciprocity in TDD systems, but requires significant amount of feedback (and thus overhead) in FDD systems, see also Section 20.1.6.

A naive approach to CSI feedback requires quantizing the entries of the transfer function matrix, resulting in a large number of bits in the feedback message.

The problem is compounded by the fact that quantization of the channel coefficients has to be done with a fine resolution, because the eigenstructure of a channel matrix is fairly sensitive to changes in the coefficients. In other words, if some of the entries of the channel matrix change even by a small amount (due to the quantization), then the eigenvectors computed from this perturbed matrix can be quite different from the eigenvectors of the original matrix. Since the optimal beamforming vectors are usually along the eigenvectors of the matrix, we thus find a strong sensitivity of beamforming vectors (and resulting SNR) to the quantization.

Example 20.6 *We consider a typical example for a cellular MIMO system: let the BS have 8 antenna elements, and each MS have 2 antenna elements. The system has 5 MHz bandwidth centered at 2 GHz carrier frequency, and operates in a channel with 250 kHz coherence bandwidth. The coherence time is 5 ms, corresponding to typical vehicular speeds. With 30 users in the cell, what is the total overhead data rate for the feedback?*

Assume that real and imaginary part are quantized with 6 bits each, and a rate 2/3 code is used to protect the feedback information. The total number of feedback bits per second is

$$2 \cdot 6 \cdot \frac{1}{2/3} \cdot K \cdot N_t \cdot N_r \cdot \frac{B}{B_{\text{coh}}} \cdot \frac{1}{T_{\text{coh}}} \quad (20.73)$$

$$= 2 \cdot 6 \cdot \frac{1}{2/3} \cdot 30 \cdot 8 \cdot 2 \cdot \frac{5000}{250} \cdot \frac{1}{0.005} \quad (20.74)$$

$$= 34.6 \cdot 10^6 \quad (20.75)$$

Thus, the total feedback load is tens of Mbit/s, which is clearly unacceptable. Note that we have assumed here that the CSI from all the MSs has to be fed back, even though at a given point in time the number of data streams is limited to 8 by the dimensionality constraint mentioned previously; the feedback of all MSs' CSI might be required for the purpose of scheduling as discussed below. But even when considering only 4 active users, the overall feedback data rate is 4.6 Mbit/s – about as much as the typical cell capacity for a 5 MHz system. This example shows clearly that feedback reduction techniques are extremely important.

If the TX uses linear precoding (beamforming), an effective way for reducing feedback is the use of limited *feedback codebooks*. In this approach the RX does not feed back the channel transfer function matrix itself, but rather the index of a specific setting of the precoder matrix that it wants the TX to use. Codebook sizes (i.e., number of possible precoder settings) are usually very small (8 or 16 entries, requiring 3 or 4 bits to communicate the index of the entry), and thus the savings can be significant: instead of feeding back $12 \cdot N_t \cdot N_r$ bits in each coherence time and each coherence bandwidth (192 bits in the example above), only 3 or 4 bits need to be transmitted – a saving of almost two orders of magnitude.

The advantages of using codebooks are compounded if the number of spatial streams is smaller than the number of antenna elements. The channel coefficients are the entries into a matrix, whose size is proportional to $N_t N_r$. On the other hand, the number of codebook indices that have to be fed back is L , i.e., the BS has to know for each spatial stream what codeword (precoder setting) to use.

Good performance can, however, only be achieved with a good design of the entries of the codebook. The basic idea of codebook design is to exploit the intrinsic geometry of the underlying vector spaces. Note that the correlation structure of the channel (which in turn depends on the antenna array and the propagation channel) influences the characteristics of the vector space.

In the following, we consider only feedback for a single user, and assume a flat-fading channel. A generalization to frequency-selective channels is given at the end of this section.

Feedback for Single-Data-Stream Transmission

We first discuss the situation where the BS transmits a single data stream to an MS with multiple antenna elements. Assume we have a codebook $\mathcal{C}$ (set of quantized precoding vectors $\mathbf{t}$ with unit norm) of size Q , so that feedback of one codebook entry takes $\log_2 Q$ bits. For a given $\mathbf{t}^{(q)}$, the SNR available at the MS is

$$SNR_q = \gamma \|\mathbf{H}\mathbf{t}^{(q)}\|_2^2 \quad (20.76)$$

where γ is again the SNR obtained in a SISO AWGN channel of unit gain, and $\mathbf{H}$ is the $N_r \times N_t$ transfer function matrix. The RX measures the transfer function matrix, and defines the best possible precoding vector as

$$\tilde{\mathbf{t}} = \arg \max_{\mathbf{t}^{(q)} \in \mathcal{C}} SNR_q \quad (20.77)$$

where $\tilde{\mathbf{t}}$ depends on $\mathbf{H}$, even though we do not explicitly write this dependence. The MS can determine the desired $\tilde{\mathbf{t}}$ for a given (measured) $\mathbf{H}$, e.g., by a brute-force search over the whole codebook.

But before the codebook can be used, we first have to define what its entries are (clearly, the codebook has to be known a priori to the TX and the RX). It is desirable to have a codebook that minimizes – on average – the loss of SNR at the RX

$$\zeta = E_{\mathbf{H}} \left\{ \gamma \|\mathbf{H}\mathbf{t}^{(\text{opt})}\|_2^2 - \gamma \|\mathbf{H}\tilde{\mathbf{t}}\|_2^2 \right\} \quad (20.78)$$

Now it can be shown that this loss can be upper-bounded as

$$\zeta \leq \hat{\zeta} \leq E_{\mathbf{H}} \left\{ \gamma \|\mathbf{H}\|_2^2 \right\} E_{\mathbf{H}} \left\{ 1 - \max_{\mathbf{t}^{(q)} \in \mathcal{C}} |\mathbf{v}^* \mathbf{t}^{(q)}|^2 \right\} \quad (20.79)$$

where $\mathbf{v}$ is the right singular vector of $\mathbf{H}$ corresponding to the largest singular value. Furthermore $\hat{\zeta}$ can be upper-limited as

$$\hat{\zeta} \leq E_{\mathbf{H}} \left\{ \gamma \|\mathbf{H}\|_2^2 \right\} \left(\left(\frac{\delta}{2} \right)^2 \left(\frac{\delta}{2} \right)^{2(N_t-1)} Q + \left[1 - \left(\frac{\delta}{2} \right)^{2(N_t-1)} Q \right] \right) \quad (20.80)$$

where

$$\delta = \min_{p \neq q} \sqrt{1 - |[\mathbf{t}^{(q)}]^\dagger \mathbf{t}^{(p)}|^2} \quad (20.81)$$

i.e., the minimum over the subspace distance between $\mathbf{t}^{(p)}$ and $\mathbf{t}^{(q)}$. Finding the optimum codebook thus means that we have to maximize the minimum of the subspace distances – a problem that is related to so-called *Grassmannian line packing*. An example codebook for $N_t = 2$ and $Q = 4$ is given in Table 20.2.

Note that the codebook design is not unique – a multiplication with a constant phase (for all entries) does not change the SNR (and that multiplication is undone at the RX anyway).

The above derivation derives the codebook for iid complex Gaussian fading channels. Let us now consider the case where the channel is correlated at TX and RX, and let us consider the Kronecker model of Section 7.4.7.

$$\mathbf{H} = \mathbf{R}_{\text{RX}}^{1/2} \mathbf{G} \mathbf{R}_{\text{TX}}^{1/2}, \quad (20.82)$$

Table 20.2 Codebook entries for limited feedback with $N_t = 2$ and $Q = 4$. [Love et al. 2003]

$q = \backslash$	$t_1^{(q)}$	$t_2^{(q)}$
1	$-0.1612 - 0.7348j$	$-0.5135 - 0.4128j$
2	$-0.0787 - 0.3192j$	$-0.2506 + 0.9106j$
3	$-0.2399 + 0.5985j$	$-0.7641 - 0.0212j$
4	-0.9541	0.2996

where $\mathbf{G}_G$ is a matrix with independent identically distributed (iid) complex Gaussian entries. It is intuitive that the transmit correlation should have a strong impact on the codebook: there is a relationship between the correlation matrix and the angles in which the radiation is transmitted. Clearly, there is no sense in quantizing finely in directions which are never used to transmit energy. This statement can be put into a more quantitative form by the following precoder design recipe: the codebook vectors $\hat{\mathbf{t}}^{(q)}$ are the codebook vectors from the uncorrelated case, scaled by the correlation matrix

$$\hat{\mathbf{t}}^{(q)} = \frac{\mathbf{R}_{\text{TX}}^{\dagger/2} \mathbf{t}^{(q)}}{\|\mathbf{R}_{\text{TX}}^{\dagger/2} \mathbf{t}^{(q)}\|_2}. \quad (20.83)$$

Note, however, that this approach requires the TX and the RX to know the TX correlation matrix.

An even simpler codebook consists simply of the coefficients of a Discrete Fourier Transform (DFT) matrix. In that case, the entries of the codebook are just beams pointing to different directions. Clearly such a codebook works well if the angular spread of the signal is small, as often happens at macrocellular BSs.

Feedback for Space–Time Block Codes and Spatial Multiplexing

The design of the precoding codebook becomes more complicated when the TX employs orthogonal space–time block coding. In that case, the precoding vector $\mathbf{t}$ has to be replaced by a precoding matrix $\mathbf{T}$ of size $N_t \times \tilde{L}$, where $\tilde{L}$ is a dimension of the space–time code (e.g., 2 for an Alamouti code). The optimization goal is still the maximization of the overall SNR at the RX (note that there is only a single data stream, even though it is space–time coded)

$$\text{SNR}_q = \gamma \|\mathbf{H}\mathbf{T}^{(q)}\|_2^2 \quad (20.84)$$

A bound to the SNR loss due to the quantization can be derived as

$$\zeta = E_{\mathbf{H}} \left\{ \gamma \|\mathbf{H}\mathbf{T}^{(\text{opt})}\|_2^2 - \gamma \|\mathbf{H}\tilde{\mathbf{T}}\|_2^2 \right\} \leq E_{\mathbf{H}} \left\{ \gamma \lambda_1^2(\mathbf{H}) \right\} \left\{ \tilde{L} + \left(\frac{\delta}{2\sqrt{\tilde{L}}} \right)^{2N_t\tilde{L}+O(N_t)} \mathcal{Q} \left[\left(\frac{\delta}{2} \right)^2 - \tilde{L} \right] \right\} \quad (20.85)$$

where the minimum subspace distance is now defined as the minimum chordal distance

$$\delta = \min_{p \neq q} \frac{1}{\sqrt{2}} \|\mathbf{T}^{(p)}(\mathbf{T}^{(p)})^\dagger - \mathbf{T}^{(q)}(\mathbf{T}^{(q)})^\dagger\|_F \quad (20.86)$$

The codebook entries are not unique – clearly a multiplication with a unitary matrix does not change the characteristics of $\mathbf{T}$. This is similar to the case of a single data stream, where multiplication with a constant phase does not change the optimality.

One might now ask the question why to consider space–time block codes, in particular Alamouti codes, in a setting where feedback is available – after all, Alamouti codes perform worse than maximum-ratio transmission (which is approximated by the beamforming based on the feedback). However, there are situations where CSI becomes uncertain (due to noise) or quickly outdated. In this case, it is advantageous to supplement the partial-CSI-based beamforming with extra diversity from the space–time coding. If average CSI is available, then the precoder should be used to appropriately “color” the transmission signal (i.e., align with the correlation matrix information) before transmission.

When spatial multiplexing is used in the transmission, finding the optimum precoding codebook becomes even more difficult, because – depending on the RX type – different criteria can be used for optimization.

Finally, we note that antenna selection at the transmitter can be interpreted as precoding with very simple precoder settings (entries of $\mathbf{T}$ are 1 or 0. Consequently, feedback of the precoder settings is also very simple.

Feedback in Frequency-Selective Channels

Consider now an OFDM system operating in a frequency-selective channel. The brute-force approach would be to feed back the codebook entry index for each subcarrier separately. However, this would be highly wasteful, since channels on adjacent subcarriers are correlated, and therefore also the optimum beamformers are correlated. As an intuitive improvement is thus easy to group the subcarriers, and just feed back one index for each group (e.g., the index for the subcarrier at the center of the group). More sophisticated approaches use interpolation (each group might have to be multiplied with a different phase factor; this phase factor also has to be fed back).

20.3 Multiuser MIMO

We now turn our attention to the question of how MIMO systems work in a cellular scenario where the BS communicates with multiple users at the same time. As we see in the following, this situation requires some new paradigms for the usage of the degrees of freedom provided by the multiple antenna elements. As a matter of fact, many of the issues are more related to SDMA (see Section 20.1.5 and 20.1.7) than to the single-user MIMO discussed previously in Section 20.2.

The system model we consider in the following is outlined in Figure 20.21. A single BS with N_{BS} antenna elements communicates with K MSs with $N_{MS}^{(k)}$ antenna elements each. We assume that $N_{BS} > N_{MS}^{(k)}$, but that $N_{BS} < \sum_k N_{MS}^{(k)}$. This is the case normally occurring in a cellular network, and also the one with the most interesting effects for multiuser MIMO. A typical fourth-generation

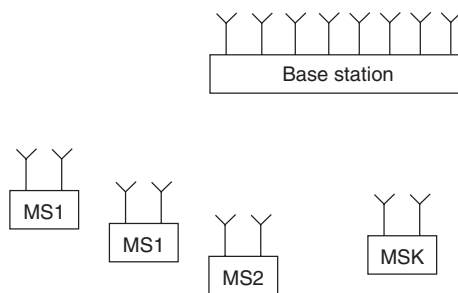


Figure 20.21 Multiuser MIMO system setup.

cell would have $N_{BS} = 8$, $N_{MS}^{(k)} = 2$, and $K = 20$. There is no requirement that all MSs are active all the time; rather the BS can schedule the communication with specific users according to their channel state. However, certain fairness and delay criteria must be fulfilled for the different users.

It is tempting to think of a multiuser MIMO system simply as a single-user system where the antennas of the RX are distributed over different locations. This point of view is helpful in many aspects but must be taken with a grain of salt, since it can lead to some erroneous conclusions especially for the downlink case. The key differences between single-user MIMO and multiuser MIMO are as follows:

- Cross-layer design (scheduling) is critical, and can greatly reduce the probability of a breakdown of the overall capacity due to peculiarities of the channel state. If there are more MSs than there are BS antennas (a typical case in a cellular setting), then a judicious scheduling algorithm can greatly reduce the possibility that the channels to two users are (almost) linearly dependent (which would make spatial multiplexing to them inefficient).
- CSIT has a dramatic impact on the performance of a multiuser MIMO system in the downlink. One might wonder why the CSIT is so critical for the performance of this situation – after all, in the single-user case, the capacity difference between systems with and without CSIT is marginal. If the different MSs could collaborate in the reception of the signals, then indeed the capacity with and without CSIT would hardly differ; since by assumption $N_{BS} < \sum_k N_{MS}^{(k)}$, the RX could decode all the data streams that the BS sends out. However, in reality, the MSs *cannot* collaborate, and each MS can decode at most $N_{MS}^{(k)}$ data streams. Thus, without CSIT, the following situations could occur: (i) the BS transmits only $\min_k (N_{MS}^{(k)})$ streams, which drastically limits the capacity of the downlink; (ii) the BS transmits N_{BS} data streams, but since the MSs cannot decode that many data streams, the SINR at each MS is extremely bad, and thus the capacity is low. With CSIT, however, the BS can “null out” undesired streams in the direction of the different MSs, and thus transmit more data streams while still enabling a good SINR at each MS.

20.3.1 Performance Limits

In a first step, we want to explore the fundamental limits of multiuser MIMO, i.e., how many users can be supported, and at what rate. We will not discuss in detail concepts of information theory, but mostly restrict ourselves to intuitive results that we can draw from the analogy with single-user MIMO and smart antennas.

Uplink

First consider the uplink: We have K MSs with $N_{MS}^{(k)}$ antenna elements each, and a BS with N_{BS} antenna elements. Each of the MSs is capable of transmitting multiple data streams, or of transmitting one data stream from multiple antenna elements. It is clear that such an arrangement is very similar to a single-user MIMO system with $\sum_k N_{MS}^{(k)}$ transmit antenna elements and N_{BS} receive antenna elements. The only difference lies in the fact that the different MSs cannot cooperate in transmitting their information; however, this hardly influences the capacity: consider the extreme case that each MS has only a single antenna: then the system is identical to an H-BLAST system; it is known that as long as the transmit rates from the different antenna elements are allowed to be different, an H-BLAST system can achieve capacity. Thus the capacity of the uplink system grows with N_{BS} , the number of BS antenna elements (remember that we assume $N_{BS} < \sum_k N_{MS}^{(k)}$). The RX at the BS can be, e.g., an H-BLAST RX that combines MMSE reception with serial interference cancellation.

The optimum power assignments for the different users follows from the principle of “iterative waterfilling” [Yu et al. 2004], in which we iterate the following algorithm to convergence:

- for $k = 1$ to K
 - for a given user k , compute the covariance matrix of the noise and the interference from all other users;
 - obtain the optimum signal correlation matrix from the waterfilling principle as described in Section 20.2.5.

Downlink

Let us now turn to the downlink. The received signal $\mathbf{r}^{(k)}$ at the k -th MS is

$$\mathbf{r}^{(k)} = \mathbf{H}^{(k)}\tilde{\mathbf{s}} + \mathbf{n} \quad (20.87)$$

Here $\tilde{\mathbf{s}}$ is the signal transmitted to the users; it consists of a superposition of the data intended for all the different users $\tilde{\mathbf{s}} = \sum_k \tilde{\mathbf{s}}^{(k)}$. The $\tilde{\mathbf{s}}^{(k)}$ are linearly or nonlinearly encoded versions of the signal streams intended for the different users. The RXs, i.e., the MSs, cannot cooperate for the *decoding* of the datastreams. This has much more significant consequences for the practical implementation than the inability to cooperate for the encoding: access of the decoder to all the data streams is critical for interference suppression. Consequently, the only way to avoid interference is via a *TX precoding* that eliminates interference at the RX (and which requires CSIT).

In the 1980s, a remarkable result was discovered: if interference is known at the TX, then the desired signal can be precoded in such a way that the RX does not “see” any interference. This result is also known as “writing on dirty paper,” and codes that achieve or approximate such interference suppression are called “dirty paper codes” or “Costas codes.” The name derives from the principle that if one knows the position of dirt on a paper, one can write letters in such a way that the dirt does not hinder the reading of the letters [Peel 2003]. There are numerous ways on how such dirty paper codes can be approximated, e.g., through Tomlinson-Harashima precoding.

When performing the encoding of multiple data streams, the order in which the streams are encoded is very important. The encoder for the k -th MS treats the interference from data streams for MS $1, \dots, k-1$ as known (which can therefore be completely suppressed), while the interference for users $k+1, \dots, K$ has the same effect as noise. Consequently, the first encoded stream has a poor SINR, while later encoded streams have good SINR. This situation is dual to the one in the uplink (with a Successive Interface Cancellation (SIC) RX), where the first-decoded data stream has a poor SINR, while the later-decoded streams (which are derived from the “cleaned-up” signals) have a good SINR.

20.3.2 Scheduling

If the number of data streams for users is larger than N_{BS} , then the BS has to determine which users it wants to serve at a given time. The criterion for the serving might be the maximization of the overall throughput (without regard to fairness), or it might want to ensure that each user is served with a certain minimum data rate. Different ways of selecting users (and possibly, grouping users together in certain timeslots) lead to different overall capacity and quality of service. Optimizing such criteria in principle requires an exhaustive search over all scheduling assignments which becomes infeasible if the number of users is large. It is thus popular to perform a “greedy” search, where the first user to be selected is the one that gives, e.g., the highest overall capacity. The next

user that is selected from the remaining ones is the user that increases the capacity by the most significant amount, and so on. Simulations indicate that such a greedy algorithm gives performance that is quite close to the optimum scheduling.

20.3.3 Linear Precoding – Uplink

The simplest practical implementation of multiuser MIMO processing is based on linear processing at TX and RX. The RX (the BS) has all signals available, and can thus perform optimum (linear) processing for interference suppression. The optimum transmit and receive weights depend on each other (changing the TX weights at the MS changes the amount of useful power and interference for all the users at the BS), therefore an iterative determination of the optimum weights is necessary. The details depend on the optimization criterion; in the following we consider the minimization of the overall Mean Square Error (MSE).

The MSE can be written as

$$\text{MSE} = \text{tr} \left\{ \sum_{k=1}^K \left\{ \sum_{j=1}^K (\mathbf{T}^{(j)})^\dagger (\mathbf{H}^{(j)})^\dagger \mathbf{W}^{(k)} (\mathbf{W}^{(k)})^\dagger \mathbf{H}^{(j)} \mathbf{T}^{(j)} - (\mathbf{T}^{(k)})^\dagger (\mathbf{H}^{(k)})^\dagger \mathbf{W}^{(k)} \right. \right. \\ \left. \left. - (\mathbf{W}^{(k)})^\dagger \mathbf{H}^{(k)} \mathbf{T}^{(k)} + \mathbf{I} + \sigma_n^2 (\mathbf{W}^{(k)})^\dagger \mathbf{W}^{(k)} \right\} \right\} \quad (20.88)$$

where $(\mathbf{W}^{(k)})^\dagger$ is the receive matrix for the k -th user. The goal is then to minimize this MSE under the power constraints

$$\text{tr} \{ (\mathbf{T}^{(k)})^\dagger \mathbf{T}^{(k)} \} \leq P_k^{\max} \quad (20.89)$$

The TX weights are functions of the receive weights of *all* users, while the optimum RX weights depend on the transmit weights of *all* users. Thus, the RX weights and transmit weights can be computed with the following iteration:

1. Update for all users ($k = 1, \dots, K$)

$$(\mathbf{W}^{(k)})^\dagger = (\mathbf{T}^{(k)})^\dagger (\mathbf{H}^{(k)})^\dagger \left[\sigma_n^2 \mathbf{I} + \sum_{j=1}^K \mathbf{H}^{(j)} \mathbf{T}^{(j)} (\mathbf{T}^{(j)})^\dagger (\mathbf{H}^{(j)})^\dagger \right]^{-1} \quad (20.90)$$

2. Update for all users ($k = 1, \dots, K$)

$$\mathbf{X}^{(k)}(\mu'_k) = \left[\mu'_k \mathbf{I} + \sum_{j=1}^K (\mathbf{H}^{(k)})^\dagger \mathbf{W}^{(j)} (\mathbf{W}^{(j)})^\dagger \mathbf{H}^{(k)} \right]^{-1} (\mathbf{H}^{(k)})^\dagger \mathbf{W}^{(k)} \quad (20.91)$$

$$\mu_k = \max \left[\arg_{\mu'_k} \left(\text{tr} \left\{ \mathbf{X}^{(k)}(\mu'_k) (\mathbf{X}^{(k)}(\mu'_k))^\dagger \right\} \right) = P_k^{\max} \right], 0 \quad (20.92)$$

$$\mathbf{T}^{(k)} = \left[\mu_k \mathbf{I} + \sum_{j=1}^K (\mathbf{H}^{(k)})^\dagger \mathbf{W}^{(j)} (\mathbf{W}^{(j)})^\dagger \mathbf{H}^{(k)} \right]^{-1} (\mathbf{H}^{(k)})^\dagger \mathbf{W}^{(k)} \quad (20.93)$$

Typically, 10 to 20 iterations are sufficient for convergence.

20.3.4 Linear Precoding – Downlink

Also for the downlink, the linear precoding is the method that can be implemented most easily. In the case of beamforming (linear precoding) at the TX, the total transmit signal intended for the k -th user is the source signal, multiplied with a beamforming matrix $\mathbf{T}^{(k)}$

$$\tilde{\mathbf{s}}^{(k)} = \mathbf{T}^{(k)} \mathbf{s}^{(k)} \quad (20.94)$$

where the $\mathbf{T}^{(k)}$ is the precoding matrix (beamformer) for the k -th user. The correlation matrix of the k -th user's signal is $\mathbf{R}_{\tilde{\mathbf{s}}\tilde{\mathbf{s}}}^{(k)} = E\{\tilde{\mathbf{s}}^{(k)}\tilde{\mathbf{s}}^{(k)\dagger}\}$, and the power allocated to the k -th user is $P_k = \text{tr}\{\mathbf{R}_{\tilde{\mathbf{s}}\tilde{\mathbf{s}}}^{(k)}\} = \text{Trace}\{\mathbf{T}^{(k)}\mathbf{T}^{(k)\dagger}\}$, since we assume that the modulation symbols have unit energy $\mathbf{R}_{\mathbf{s}\mathbf{s}}^{(k)} = \mathbf{I}$. Usually the BS has a constraint on the total transmit power $\sum_k P_k \leq P_{\max}$.

Block Diagonalization

Consider the case where each MS has multiple antenna elements ($N_r^{(k)} \geq 1$) and the BS transmits $N_r^{(k)}$ data streams to each user. A simple solution can be achieved by imposing the dimensionality constraint $\sum_k N_r^{(k)} = N_t$. This constraint is very much along the lines of interpreting the multiple MSs as a “distributed array,” and the dimensionality constraint ensures that the number of data streams does not exceed the number of RX chains that can demodulate the signal. To couch it in the language of Section 20.1: each antenna is treated as a separate (single-antenna) user, and the BS preprocessing makes sure that data streams arriving at the antennas are well separated. This eases the design of the MS, which does not have to do any additional processing.

However, the constraint $\sum_k N_r^{(k)} = N_t$ is actually too strict. According to this constraint, adding receive antennas at the MS decreases the number of users that can be transmitted to at a given point in time; this is obviously not a reasonable constraint, as additional receive antennas serve to *improve* the reception quality. Rather, it is the *number of data streams* that intended for the MSs that is limited, and data streams intended for different MSs have to be kept apart by appropriate beamforming at the BS side.

Keeping the data streams of different MSs apart is achieved by a *block diagonalization* technique [Spencer et al. 2004b]. Define for each user an interference channel matrix $\tilde{\mathbf{H}}^{(k)}$

$$\tilde{\mathbf{H}}^{(k)} = [(\mathbf{H}^{(1)})^T, \dots, (\mathbf{H}^{(k-1)})^T, (\mathbf{H}^{(k+1)})^T, \dots, (\mathbf{H}^{(K)})^T]^T \quad (20.95)$$

Any precoding matrix $\mathbf{T}^{(k)}$ that lies in the nullspace of $\tilde{\mathbf{H}}^{(k)}$ leads to a situation where the data streams intended for the other MSs do not influence the k -th MS. Consequently, there is now a new dimensionality constraint. Let J_k be the rank of $\tilde{\mathbf{H}}^{(k)}$. Then the block diagonalization can be achieved if

$$N_t > \max_k (J_1, J_2, \dots, J_K) \quad (20.96)$$

If the propagation channels are all full rank, this means that the number of transmit antennas must be no smaller than the number of data streams. Define then the SVD of $\tilde{\mathbf{H}}^{(k)}$

$$\tilde{\mathbf{H}}^{(k)} = \tilde{\mathbf{U}}^{(k)} \tilde{\Sigma}^{(k)} [\tilde{\mathbf{V}}_{\text{fc}}^{(k)} \quad \tilde{\mathbf{V}}_{\text{lc}}^{(k)}]^\dagger \quad (20.97)$$

where $\tilde{\mathbf{V}}_{\text{fc}}^{(k)}$ contains the first J_k and $\tilde{\mathbf{V}}_{\text{lc}}^{(k)}$ the last $N_t - J_k$ right singular vectors; $\tilde{\mathbf{V}}_{\text{lc}}^{(k)}$ thus forms an orthonormal basis of the nullspace of $\tilde{\mathbf{H}}^{(k)}$. We define a new “overall effective channel matrix”

$$\hat{\mathbf{H}} = \begin{bmatrix} \mathbf{H}^{(1)} \tilde{\mathbf{V}}_{\text{lc}}^{(1)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}^{(K)} \tilde{\mathbf{V}}_{\text{lc}}^{(K)} \end{bmatrix} \quad (20.98)$$

We next find the SVD of this overall effective matrix; due to the block-diagonal structure of $\hat{\mathbf{H}}$, this can be done in an efficient (block-by-block) manner. For the k -th block,

$$\hat{\mathbf{H}}^{(k)} = \hat{\mathbf{U}}^{(k)} \begin{bmatrix} \hat{\Sigma}^{(k)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [\hat{\mathbf{V}}_{\text{fc}}^{(k)} \quad \hat{\mathbf{V}}_{\text{lc}}^{(k)}]^\dagger \quad (20.99)$$

The total beamforming/power allocation matrix $\mathbf{T}$ is then

$$\mathbf{T} = \begin{bmatrix} \tilde{\mathbf{V}}_{\text{lc}}^{(1)} \hat{\mathbf{V}}_{\text{fc}}^{(1)} & \tilde{\mathbf{V}}_{\text{lc}}^{(2)} \hat{\mathbf{V}}_{\text{fc}}^{(2)} & \dots & \tilde{\mathbf{V}}_{\text{lc}}^{(K)} \hat{\mathbf{V}}_{\text{fc}}^{(K)} \end{bmatrix} \Lambda^{1/2} \quad (20.100)$$

where Λ is a diagonal matrix that performs waterfilling on the elements of

$$\begin{bmatrix} \hat{\Sigma}^{(1)} & & \mathbf{0} \\ & \hat{\Sigma}^{(2)} & \\ \mathbf{0} & \dots & \hat{\Sigma}^{(K)} \end{bmatrix} \quad (20.101)$$

Coordinated Beamforming

Note that if the RX has a linear filter, then $\mathbf{H}^{(j)}$ should be replaced by the concatenation of filter and channel matrix. Consequently, the rank of $\tilde{\mathbf{H}}^{(k)}$ is influenced by all the RX filters. Let signals at the k -th MS be multiplied with a matrix $\mathbf{W}^{(k)\dagger}$. The overall received signal is then

$$\mathbf{r}^{(k)} = \mathbf{W}^{(k)\dagger} \mathbf{H}^{(k)} \mathbf{T}^{(k)} \mathbf{s}^{(k)} + \mathbf{W}^{(k)\dagger} \mathbf{H}^{(k)} \sum_{l \neq k} \mathbf{T}^{(l)} \mathbf{s}^{(l)} + \mathbf{W}^{(k)\dagger} \mathbf{n} \quad (20.102)$$

The second term on the r.h.s. now describes the interference. Essentially, the RX processing means the creation of an “effective” channel $\mathbf{W}^{(k)\dagger} \mathbf{H}^{(k)}$ for each user. Changing this effective channel indirectly requires changes in the transmit beamforming matrix as well. The computation of the weights $\mathbf{T}$ and $\mathbf{W}$ thus has to be done in an iterative fashion: start out with a set $\mathbf{W}^{(k)\dagger}$ (e.g., the weights for maximum-ratio combining at the k -th MS), and compute the optimum transmit weights according to one of the strategies described above. With the new $\mathbf{T}$ compute the signal statistics at the MSs, and recompute the optimum linear weights according to a certain criterion (e.g., MMSE). These new RX weights create a new set of effective channel, which then forms the basis for a recomputation of $\mathbf{T}$, and so on. The iteration is stopped when the weights (or the performance measures) do not change appreciably anymore.

Figure 20.22 shows the capacity for various types of precoding. We see that with a pure channel inversion, the capacity first increases as the number of users increases, but then starts to decrease again as K starts to approach the number of antenna elements at the BS. This is due to the increased probability of an ill-conditioned channel matrix. When the RX has more antennas, the difference between channel inversion and regularized channel inversion (see below) becomes much smaller (see Figure 20.23).

Joint Wiener Filtering

Block diagonalization can be seen as a generalization of zero-forcing. Similarly, channel regularization (which aims to minimize the MSE at the RX) can be generalized to *joint Wiener filtering*. The MSE (MSE) metric is, as usual, defined as

$$\text{MSE}_k = \mathbb{E} [\|\mathbf{r}_k - \mathbf{z}_k\|^2] \quad (20.103)$$

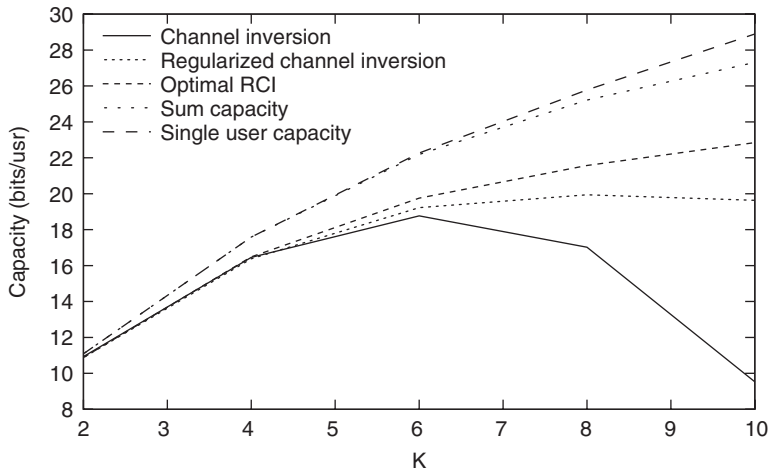


Figure 20.22 Multiuser MIMO capacity as function of the number of users K when each MS has only one antennas. $N_t = 10$, $\gamma = 10$ dB. Different precoding methods are shown.

Reproduced from Tsoulos [2006] © CRC Press.

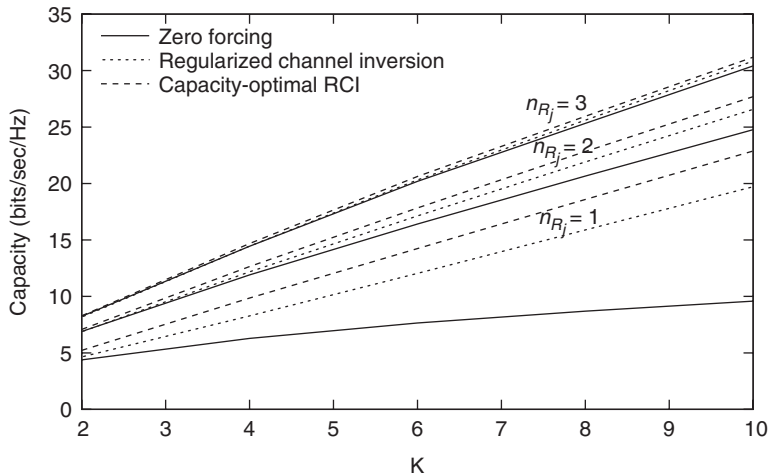


Figure 20.23 Multiuser MIMO capacity as function of the number of users K when each MS has multiple antenna; $N_t = 10$, $\gamma = 10$ dB. Different precoding methods are shown.

Reproduced from Tsoulos [2006] © CRC Press.

where the expectation is over the random data vectors, $\{\mathbf{s}_k\}_{k=1}^K$, and the noise, $\{\mathbf{n}_k\}_{k=1}^K$. The optimization problem is then

$$\begin{aligned} \left\{ \mathbf{T}_k^{\text{opt}} \right\}_{k=1}^K &= \arg \min_{\{\mathbf{T}_k\}_{k=1}^K} \sum_k \text{MSE}_k \quad \text{for } k = 1, \dots, K. \\ \text{s. t.} \quad &\text{Tr} \left\{ \sum_k \mathbf{T}_k^\dagger \mathbf{T}_k \right\} \leq P_{\max} \end{aligned} \quad (20.104)$$

To solve this optimization problem, take an eigendecomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\dagger$ of the following matrix

$$\mathbf{H}_k^\dagger \left[\sigma_n^2 \mathbf{I} + \sum_{j \neq k} \mathbf{H}_j \mathbf{T}_j \mathbf{T}_j^\dagger \mathbf{H}_j^\dagger \right] \mathbf{H}_k = [\mathbf{V}_k \quad \bar{\mathbf{V}}_k] \begin{bmatrix} \Lambda_k & \\ & \bar{\Lambda}_k \end{bmatrix} [\mathbf{V}_k \quad \bar{\mathbf{V}}_k]^\dagger \quad (20.105)$$

where $\mathbf{V}_k$ is a square matrix of dimension $\text{rank}(\mathbf{H}_k)$. Then the optimum precoding matrix is

$$\mathbf{T}_k = \mathbf{V}_k \begin{bmatrix} 0 & 0 & \dots & 0 & \xi_{1,k} & 0 & 0 \\ \dots & \dots & \dots & \dots & 0 & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 & 0 & \xi_{L,k} \end{bmatrix} \quad (20.106)$$

where L is the number of spatial streams and

$$\xi_{i,k} = \max \left[\frac{1}{\mu^{1/2} [\Lambda_k]_{i,i}^{1/2}} - \frac{1}{[\Lambda_k]_{i,i}}, 0 \right] \quad (20.107)$$

and μ is chosen to satisfy the total power constraint.

For a given transmit precoder, the optimum RX filter matrix is the Wiener filter

$$\mathbf{W}_k = \left[\mathbf{H}_k \mathbf{T}_k \mathbf{T}_k^\dagger \mathbf{H}_k^\dagger + \left[\sigma_n^2 \mathbf{I} + \sum_{j \neq k} \mathbf{H}_j \mathbf{T}_j \mathbf{T}_j^\dagger \mathbf{H}_j^\dagger \right] \right]^{-1} \mathbf{H}_k \mathbf{T}_k \quad (20.108)$$

Joint Leakage Suppression

In joint leakage suppression, the precoding matrices are designed to maximize the ratio of the power of the desired signal received by the k -th MS and the sum of the noise and the total interference power (leakage) due to the transmit signal intended for the k -th user at all the other MSs (note that this is different from the SINR at the k -th MS). The key motivation for this criterion is that it allows an easier optimization (often in closed form) than the SINR at the RX. However, performance of systems that maximize Signal-to-Leakage and Noise Ratio (SLNR) is slightly worse than systems that maximize SINR. The SLNR of the k -th user is

$$SLNR_k = \frac{E\{\mathbf{s}_k \mathbf{T}_k^\dagger \mathbf{H}_k^\dagger \mathbf{H}_k \mathbf{T}_k \mathbf{s}_k\}}{N_r \sigma_n^2 + E\{\sum_{i \neq k} \sum_{j \neq k} \mathbf{s}_k \mathbf{T}_k^\dagger \mathbf{H}_i^\dagger \mathbf{H}_j \mathbf{T}_k \mathbf{s}_k\}} \quad (20.109)$$

which can be simplified to

$$SLNR_k = \frac{\text{Tr}\{\mathbf{T}_k^\dagger \mathbf{H}_k^\dagger \mathbf{H}_k \mathbf{T}_k\}}{\text{Tr}\{\mathbf{T}_k^\dagger [N_r \sigma_n^2 \mathbf{I} + \tilde{\mathbf{H}}_k^\dagger \tilde{\mathbf{H}}_k] \mathbf{T}_k\}} \quad (20.110)$$

The BS thus needs to find precoding matrices $\mathbf{T}_k$ that maximize the $SLNR_k$ subject to a transmit power constraint $\text{Tr}\{\mathbf{T}_k^\dagger \mathbf{T}_k\} = P_{\max}$. Note that we assume here an absence of power control, so that the powers for each user are fixed to a certain value – this decouples the optimization of the different users. We then define an auxiliary matrix $\mathbf{A}$ as a matrix that fulfills

$$\begin{aligned} \mathbf{A}_k^\dagger \mathbf{H}_k^\dagger \mathbf{H}_k \mathbf{A}_k &= \Lambda_k \\ \mathbf{A}_k^\dagger [N_r \sigma_n^2 \mathbf{I} + \tilde{\mathbf{H}}_k^\dagger \tilde{\mathbf{H}}_k] \mathbf{A}_k &= \mathbf{I} \end{aligned} \quad (20.111)$$

where Λ_k is a diagonal matrix with (arbitrary) nonnegative entries, sorted in descending order. The precoding matrix $\mathbf{T}$ is then the first L_k columns of $\mathbf{A}$.

For the special case that each user has only one data stream, $L_k = 1$, the beamforming matrix becomes a vector, and can be obtained in a simple way: perform an eigendecomposition of the matrix

$$\left[N_r \sigma_n^2 \mathbf{I} + \tilde{\mathbf{H}}_k^\dagger \tilde{\mathbf{H}}_k \right]^{-1} \mathbf{H}_k^\dagger \mathbf{H}_k \quad (20.112)$$

The optimum precoding vector is the eigenvector associated with the largest eigenvalue of this matrix.

Iterative Waterfilling

If we want to maximize the sum rate over all users, the optimization problem is

$$\begin{aligned} \left\{ \mathbf{T}_k^{\text{opt}} \right\}_{k=1}^K &= \arg \max_{\{\mathbf{T}_k\}_{k=1}^K} \sum_{k=1}^K R_k, \\ \text{s. t. } \quad \text{Tr} \left\{ \sum_{k=1}^K \mathbf{T}_k^\dagger \mathbf{T}_k \right\} &\leq P_{\max}, \quad \text{for } k = 1, \dots, K. \end{aligned} \quad (20.113)$$

The bandwidth-normalized information rate, R_k , of MS k is given as

$$R_k = \log \left| \mathbf{I}_{N_R} + \left[\sigma_n^2 \mathbf{I} + \sum_{j \neq k} \mathbf{H}_j \mathbf{T}_j \mathbf{T}_j^\dagger \mathbf{H}_j^\dagger \right]^{-1} \mathbf{H}_k \mathbf{T}_k \mathbf{T}_k^\dagger \mathbf{H}_k^\dagger \right|. \quad (20.114)$$

There is no closed-form solution for this problem, but it can be tackled by iterative solution techniques.

20.3.5 Closed-Loop Systems and Quantized Feedback

As discussed in Section 20.2.11, limited-feedback codebooks provide a good way of reducing the feedback overhead for transmit beamforming. However, for multiuser MIMO, the technique is much more difficult. First of all, the quantization of the transmit precoding vector has to be much finer than in the single-user case. This can be explained the following way:¹⁹ in a single-user case the BS forms a beam pattern that shows a maximum in the direction of the targeted MS. Slight deviations from the optimum beam pattern, due to the quantization effects, do not lead to a significant loss of SNR. In the multiuser case, in order to achieve good SINR at the MSs, we need to guarantee low interference by signals from other users; in other words, each transmit beam pattern needs to place nulls toward the users it should *not* cover. Now we know already from Section 20.1 that nulls are much more sensitive to perturbations of the antenna weights than “main beams.” Thus, quantization in multiuser settings must be finer.

Another complication arises from the fact that the optimum settings of the precoder depend on the channels of *all* users, and the MS therefore cannot easily compute the precoding matrix settings that it needs to optimize its performance. A feasible way out of this dilemma are *projection techniques*. Here the BS first sends out a number of training signals on $Q = K$ different beams $\mathbf{g}^{(q)}$. The k -th MS then forms the quantities

$$\gamma_k = \max_q \frac{|\mathbf{h}^{(k)}(\mathbf{g}^{(q)})^T|^2}{\sigma_n^2 + \sum_{p \neq q} |\mathbf{h}^{(k)}(\mathbf{g}^{(p)})^T|^2} \quad (20.115)$$

¹⁹ Compare also the description of smart-antenna beamforming in Section 20.1

which is simply the maximum SINR it can achieve if the BS transmits to this MS on the beam that is optimum for it, and to all the other MSs on different beams, without any power control. The k -th MS then just feeds back the index q with which it achieves this SINR. Of course, problems can occur when two MSs claim the same beam for themselves. This problem can be tackled, e.g., by having each MS transmit not just the index for the best beam but also for the second best, so that the BS is aware of a “fallback solution” that gives a certain MS an acceptable (though not optimum) SINR.

In the spirit of the “random beamforming” approach described in Section 20.1.9, it can be advantageous to change the beams onto which the projection takes place. In other words, the $\mathbf{g}^{(q)}$ can be generated randomly. Then, the probability is more evenly distributed for each MS to “see” a channel that maximizes its SINR and thus increase its throughput.

If the BS knows the long-term CSI (correlation matrix), then it can also adjust the vectors $\mathbf{g}^{(q)}$ to reduce the probability that a beam is not aligned with any user. This is especially important if there are few users, and they are not spatially uniformly distributed.

20.3.6 Base Station Cooperation

In a cellular system, the existence of multiple users — and thus interferers — influences the capacity, and decreases the data rate that is possible for a single user. A first investigation into this topic by [Catreux et al. 2001] looked at a cellular TDMA system with MMSE detection. Under these assumptions, it was shown that the *cellular* capacity of a MIMO system is hardly larger than that of a system with multiple antennas at the BS only (as described in Section 20.1). The reason for this somewhat astonishing result is that in a cellular system with multiple antennas at the BS only, those antennas can be used to suppress adjacent cell interference, and thus decrease the reuse distance. For a cellular MIMO system with $N_r = N_t$, the degrees of freedom created by the multiple BS antennas are all used for the separation of the multiple data streams from a single user, and none for the suppression of interfering users. Thus, we can use neither SDMA nor SFIR for the increase of the cellular capacity (remember that those can easily double or triple the capacity).

However it seems possible to combine MIMO with other techniques for multiple-access interference. For example, multiuser interference can be eliminated by BS cooperation. For the uplink, cooperating BSs can be viewed as a giant MIMO system with $N_r N_{BS}$ antenna elements at one link end, where N_{BS} is the number of cooperating BSs. The capacity for such a system can be approximated by inserting the generalized channel matrix into Eq. (20.36).

For the downlink, the effect of interfering BSs can be minimized by appropriate preprocessing if the BSs cooperate and each BS has the CSI for all MSs. Those methods are mathematically similar to the ones applied in multiuser MIMO. One fundamental distinction of multi-BS systems arises from the fact that the interference arriving at the undesired MSs is fundamentally asynchronous. Assuming perfect timing synchronization among cooperative BSs, the timing-advance mechanisms can ensure that the *desired* signals for an MS that are transmitted from multiple BSs reach the MS at the same time. However, the undesired signals do not arrive synchronously anymore; this leads to a change in the interference statistics, and requires modifications of how to compute the precoding coefficients. Another key difference to multiuser MIMO is that we have multiple power constraints — one for each BS; this complicates the precoder design.

Base Station Cooperation is also known in the literature as “network MIMO” or “Cooperative Multi-Point” (CoMP). Since it can greatly enhance performance in particular at the cell edge, it is envisioned for use in fourth-generation cellular systems.

Further Reading

An overview of smart antennas is given in the paper by Godara [1997], and the paper collections [Rappaport 1998] and [Tsoulos 2001]. The topic of space–time processing, which forms a basis

for smart antenna systems, involves many topics we have already discussed in previous chapters (diversity in Chapter 13, Rake RXs in Chapter 18; an overview can be found in Paulraj and Papadias [1997]. Smart antennas for CDMA systems are discussed in Liberti and Rappaport [1999].

For the topic of MIMO systems, the books by Paulraj et al. [2003] and Tse and Viswanath [2005] give a good introduction, especially into the spatial-multiplexing aspects. A number of recent books cover a wide range of MIMO-related topics, from fundamental considerations to space–time codes and RX designs. In particular the monographs [Oestges and Clerckx 2007], [Biglieri et al. 2007] and the edited books [Tsoulos 2006] and [Boelcskei et al. 2008] and an upcoming book by Heath [2011]. Also the review papers by Gesbert et al. [2003] and Diggavi et al. [2004] are “must-reads” for an introduction to the topic. For understanding the basic concept of capacity in MIMO systems and the derivations of the capacity distributions and bounds, the original papers by Foschini and Gans [1998] and Telatar [1999] are still worth reading. Also, Andersen [2000] gives important intuitive insights both into capacity and diversity aspects. For the case of channels that are known at the TX, Raleigh and Cioffi [1998] is the first paper, and still very interesting to read. The possibility for a signaling scheme that allows MIMO communications without CSI at the RX was first pointed out by Marzetta and Hochwald [1999]. Goldsmith et al. [2003] review the information-theoretic capacity for different assumptions about the CSI. The impact of channel correlation on the capacity is described in Shiu et al. [2000] and Chuah et al. [2002]. In frequency-selective channels, MIMO is most often combined with OFDM, see Stueber et al. [2004] and Jiang and Hanzo [2007]. The book chapter [Lozano et al. 2008] debunks several common myths about MIMO capacity.

Foschini et al. [2003] is devoted to reviewing different layered space–time structures. MIMO systems with antenna selection are reviewed in Molisch and Win [2004]. For space–time coding, Diggavi et al. [2004] gives an extensive introduction to space–time coding, while Tarokh et al. [1998] and Tarokh et al. [1999] give more mathematical details for space–time trellis codes and space–time block codes, respectively. The books of Jafarkhani [2005] and Larsson and Stoica [2008] are mostly dedicated to space–time codes. Limited-feedback considerations for MIMO were initiated by the seminal paper of Love et al. [2003]; a comprehensive review of limited-feedback can be found in Love et al. [2008]. Overviews of the multiuser downlink can be found in Spencer et al. [2004a], [2006] and Gesbert et al. [2007]. Leakage-based precoding was introduced by Sayed and coworkers, see, e.g., Sadek et al. [2007]. Iterative waterfilling was first suggested in the context of Asymmetric Digital Subscriber Line (ADSL) in Yu et al. [2002]; see also the discussion in Kobayashi and Caire [2006] for MMSE-based multiuser downlinks, see Zhang and Lu [2006] as well as Shi et al. [2007] and references therein. An early (though nonpublic) suggestion of BS cooperation is Molisch [2001]. A number of algorithms for BS cooperation are reviewed in Zhang and Dai [2004].

For updates and errata for this chapter, see wides.usc.edu/teaching/textbook