“This book provides an in-depth introduction to radio resource management (RRM) in wireless networks. Through various practical examples, it demonstrates how to model and analyze RRM problems and optimize wireless network performance. It is an outstanding textbook for graduate students and an excellent scholarly reference for engineers and researchers.”

Geoffrey Li, Georgia Institute of Technology
Radio Resource Management in Wireless Networks

An Engineering Approach

Do you need to design efficient wireless communications systems? This unique text provides detailed coverage of radio resource allocation problems in wireless networks and the techniques that can be used to solve them. Covering basic principles and mathematical algorithms, with a particular focus on power control and channel allocation, you will learn how to model, analyze, and optimize the allocation of resources in both physical and data link layers for a range of different network types. Both established and emerging networks are considered, including CDMA and OFDMA wireless networks, relay-based wireless networks, and cognitive radio networks. Numerous exercises help you put knowledge into practice and provide the tools needed to address some of the current research problems in the field.

This is an essential reference whether you are a graduate student, researcher, or industry professional working in the field of wireless communication networks.

Ekram Hossain is a Professor in the Department of Electrical and Computer Engineering at the University of Manitoba, and a Fellow of the IEEE. He is the co-author or co-editor of several books, including Wireless Device-to-Device Communications and Networks (Cambridge, 2015) and Smart Grid Communications and Networking (Cambridge, 2012).

Mehdi Rasti is an Assistant Professor in the Department of Computer Engineering and Information Technology at the Amirkabir University of Technology in Iran. He previously worked at the Shiraz University of Technology (Iran) and the Wireless@KTH Royal Institute of Technology (Sweden).

Long Bao Le is an Associate Professor in the National Institute of Scientific Research (Energy, Materials, and Telecommunications) at the University of Quebec.
Contents

Preface  page xiii

Part I  Basics of Wireless Networks

1  Introduction  3

1.1  Basics of a Wireless Communication System  3
1.1.1  Electromagnetic Spectrum and Frequency Range  3
1.1.2  Signal Characterization  5
1.1.3  Modulation  8
1.1.4  Wireless Channel and Signal Propagation  12
1.1.5  Channel Capacity  21
1.1.6  SINR and Channel Model for Packet Communication Systems  22

1.2  Medium Access in Wireless Networks  27
1.2.1  Channel Partitioning–Based MAC  28
1.2.2  Random Access–Based MAC  34
1.2.3  Duplexing  34

1.3  Wireless Access Technologies  35
1.3.1  Cellular Wireless Technology  35
1.3.2  WLAN, WMAN, and WPAN Technologies  39

1.4  Exercises  45
References  46

2  Wireless Networks and Resource Allocation  49

2.1  Protocol Layers for Data Communication  49
2.1.1  Physical Layer  49
2.1.2  Data Link Layer  50
2.1.3  Network Layer  51
2.1.4  Transport Layer  51
2.1.5  Session, Presentation, and Application Layers  51

2.2  Classification of Wireless Networks  52
2.2.1  Classification Based on Infrastructure  52
2.2.2  Classification Based on Spectrum Access  55
2.2.3  Classification Based on Heterogeneity  55
2.3 Physical Layer Issues in Wireless Networks 56
  2.3.1 Basic Components 56
  2.3.2 Digital Transmission Techniques 58
  2.3.3 Link Adaptation 62
  2.3.4 Diversity Transmission Techniques 65
  2.3.5 Smart Reception/Diversity Combining Techniques 65

2.4 Radio Link Layer Issues in Wireless Networks 68
  2.4.1 Multiple Access and Scheduling Methods 68
  2.4.2 Error Control Methods 93
  2.4.3 Power Control Methods 99
  2.4.4 Cell Association, Handoff Management, and Admission Control 102

2.5 Taxonomy of Resource Allocation 105

2.6 Exercises 106

References 110

Part II Techniques for Modeling and Analysis of Radio Resource Allocation Methods in Wireless Networks

3 Optimization Techniques 117

  3.1 Basics of Optimization 117
    3.1.1 Convex Functions 117
    3.1.2 Optimality Conditions for Unconstrained Optimization 119
    3.1.3 Line Search Methods for Unconstrained Optimization 121
  3.2 Convex Optimization 123
    3.2.1 Introduction 123
    3.2.2 Duality 124
    3.2.3 KKT Conditions 125
    3.2.4 Algorithms 127
  3.3 Integer Programming 132
    3.3.1 Cutting Plane Method 132
    3.3.2 Branch and Bound Algorithm 135
  3.4 Stochastic Optimization 140
    3.4.1 Introduction 140
    3.4.2 Robust Optimization 141
  3.5 Dynamic Programming 143
    3.5.1 Introduction 143
    3.5.2 Examples of Dynamic Programming 145
  3.6 Exercises 147

References 153

4 Game Theory 154

  4.1 Fundamentals of Game Theory 154
    4.1.1 Brief History 154
4.1.2 Definition of a Game 155
4.2 Non-cooperative Game 159
  4.2.1 Static Game 159
  4.2.2 Dynamic Game 162
  4.2.3 Bayesian Game 166
  4.2.4 Evolutionary Game 168
4.3 Cooperative Game 172
  4.3.1 Nash Bargaining Solution 172
  4.3.2 Coalition Game 174
4.4 Auction Theory 178
  4.4.1 Introduction to Auction Theory 178
  4.4.2 Special Auction 180
4.5 Exercises 182
References 186

Part III Physical Layer Resource Allocation in Wireless Networks

5 General System Model and Preliminary Concepts 191
  5.1 System Model for a General Multi-Cell Wireless Network 191
    5.1.1 Modeling Path-Gains 192
    5.1.2 SINR Model 194
    5.1.3 Transmit Power Vector Corresponding to a Given SINR Vector 195
  5.2 System Model for a Single-Cell Wireless Network 198
    5.2.1 Modeling Path-Gains 198
  5.3 SINR Feasibility in Interference-Limited Wireless Networks 200
    5.3.1 Existence of a Positive Transmit Power Vector Corresponding to a Given SINR Vector 201
    5.3.2 Existence of a Constrained Transmit Power Vector Corresponding to a Given SINR Vector 203
  5.4 Exercises 205
References 206

6 Power Control in Cellular Wireless Networks 207
  6.1 Objectives of Power Control 207
    6.1.1 Performance Measure and Objective Functions 208
    6.1.2 Distributed Versus Centralized Approach 210
  6.2 Different Power Control Optimization Problems 210
  6.3 Closed-Loop and Open-Loop Power Control 215
    6.3.1 Open-Loop Power Control 215
    6.3.2 Closed-Loop Power Control 216
  6.4 Distributed Power Control Algorithms 216
    6.4.1 Criteria for Evaluation and Analysis of Distributed Power Control 217
### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4.2 Existing Theoretical Frameworks for Fixed-Point and Convergence Analysis</td>
<td>217</td>
</tr>
<tr>
<td>6.5 Distributed Target-SINR Tracking Power Control (TPC)</td>
<td>220</td>
</tr>
<tr>
<td>6.6 Distributed Opportunistic Power Control (OPC)</td>
<td>222</td>
</tr>
<tr>
<td>6.7 Distributed Dynamic Target-SINR Tracking Power Control (DTPC)</td>
<td>223</td>
</tr>
<tr>
<td>6.8 Exercises</td>
<td>227</td>
</tr>
<tr>
<td>References</td>
<td>230</td>
</tr>
<tr>
<td>7 Distributed Joint Power and Admission Control</td>
<td>231</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>231</td>
</tr>
<tr>
<td>7.2 Distributed Joint Power and Admission Control Algorithms</td>
<td>233</td>
</tr>
<tr>
<td>7.2.1 TPC with Permanent Removal (TPC-PR)</td>
<td>233</td>
</tr>
<tr>
<td>7.2.2 TPC with the Capability of Temporary Removal (TR)</td>
<td>234</td>
</tr>
<tr>
<td>7.2.3 TPC with Both Temporary and Permanent Removal (TPC-TPR)</td>
<td>235</td>
</tr>
<tr>
<td>7.2.4 TPC with the Capability of Temporary Removal and Feasibility Check (DFC)</td>
<td>235</td>
</tr>
<tr>
<td>7.2.5 TPC with Soft Removal (TPC-SR)</td>
<td>239</td>
</tr>
<tr>
<td>7.3 Exercises</td>
<td>243</td>
</tr>
<tr>
<td>References</td>
<td>244</td>
</tr>
<tr>
<td>8 Joint Power and Admission Control in Cognitive Radio Networks</td>
<td>245</td>
</tr>
<tr>
<td>8.1 Introduction</td>
<td>245</td>
</tr>
<tr>
<td>8.2 System Model and Background</td>
<td>246</td>
</tr>
<tr>
<td>8.3 Protection Constraints for Primary Users and Different JPAC Problems in CRNs</td>
<td>249</td>
</tr>
<tr>
<td>8.4 Characterization of Feasible Interference Region</td>
<td>251</td>
</tr>
<tr>
<td>8.4.1 Total Received-Power-Temperature: Expressing PUs’ Protection Constraints Based on FRPR</td>
<td>252</td>
</tr>
<tr>
<td>8.4.2 Total Inter-Cell Interference Temperature: Expressing PUs’ Protection Constraints Based on FIIR</td>
<td>253</td>
</tr>
<tr>
<td>8.4.3 Total Cognitive Interference Temperature: Expressing PUs’ Protection Constraints Based on FCIR</td>
<td>255</td>
</tr>
<tr>
<td>8.4.4 Example and Discussion</td>
<td>257</td>
</tr>
<tr>
<td>8.5 Existing Centralized JPAC Algorithms to Maximize the Number of Supported SUs Subject to PUs’ Protection Constraint</td>
<td>259</td>
</tr>
<tr>
<td>8.5.1 A General SSA</td>
<td>260</td>
</tr>
<tr>
<td>8.5.2 Interference Constraint-Aware Stepwise Maximum Interference Removal Algorithm (I-SMIRA)</td>
<td>263</td>
</tr>
<tr>
<td>8.5.3 Link-Gain Ratio Algorithm (LGRA) and Effective Link-Gain Ratio Algorithm (ELGRA)</td>
<td>264</td>
</tr>
<tr>
<td>8.6 Distributed JPAC Algorithms for CRNs</td>
<td>267</td>
</tr>
<tr>
<td>8.6.1 TPC with PU-Protection Algorithm (TPC-PP)</td>
<td>267</td>
</tr>
<tr>
<td>8.6.2 Improved TPC-PP (ITPC-PP)</td>
<td>270</td>
</tr>
</tbody>
</table>
## Contents

8.7 Exercises .................................................. 271
References ....................................................... 274

9 Cell Association in Cellular Networks .................................................. 276

9.1 Introduction .................................................................. 276
9.2 System Model and Notations ........................................... 276
9.3 Distributed Joint Cell Association and Power Control ............. 278
9.4 Distributed Cell Association Schemes in Wireless Networks ....... 281
   9.4.1 Reference Signal Received Power (RSRP)-Based Cell
        Association Scheme ................................................. 281
   9.4.2 Biasing-Based Cell Range Expansion (CRE) in Wireless
        Networks with Heterogeneous BSs ............................... 282
9.5 Open Research Issues .................................................. 284
9.6 Exercises .................................................................. 286
References .............................................................. 287

Part IV Link Layer Resource Allocation in Wireless Networks

10 Sub-Carrier/Sub-Channel Allocation in OFDMA Networks ......... 291

10.1 Introduction ................................................................ 291
10.2 OFDM-Based Multiple Access ....................................... 292
10.3 Adaptive Radio Resource Allocation in OFDM Systems ......... 293
   10.3.1 System-Centric Approaches .................................... 294
   10.3.2 Application-Centric Approaches ............................... 302
10.4 Open Research Issues .................................................. 303
References .............................................................. 304

11 Resource Allocation in Relay-Based Networks .......................... 308

11.1 Introduction ................................................................ 308
11.2 Overview of Cooperative Diversity ................................. 309
   11.2.1 Amplify-and-Forward Relaying ............................... 309
   11.2.2 Decode-and-Forward Relaying ............................... 310
   11.2.3 Selection or Opportunistic Relaying .......................... 311
   11.2.4 Incremental Relaying ............................................. 312
   11.2.5 Two-Way Relaying .............................................. 313
   11.2.6 Other Enhancements ............................................. 317
11.3 Resource Allocation for Single-Carrier Systems ................. 318
   11.3.1 Power Allocation for AF Relaying ............................ 318
   11.3.2 Power Allocation for Selection AF Relaying .............. 320
   11.3.3 Joint Relay Selection and Power Allocation for ANC Two-Way
        Relaying ................................................................. 321
x

## Contents

11.4 Resource Allocation for Multi-Carrier Systems 323
  11.4.1 Resource Allocation for AF Multi-Carrier Wireless Networks 324
  11.4.2 Resource Allocation for DF Multi-Carrier Wireless Networks 333
  11.4.3 Resource Allocation for Multi-User ANC Two-Way Relay Networks 337
11.5 Further Discussion 341
11.6 Exercises 342
References 348

12 Channel Allocation for Infrastructure-Based 802.11 WLANs 352
  12.1 Introduction 352
  12.2 System under Consideration 353
    12.2.1 Network Topology 353
    12.2.2 Channelization 353
    12.2.3 Medium Access Control 355
  12.3 Channel Assignment and AP Placement in IEEE 802.11 WLANs 356
    12.3.1 Channel Assignment 356
    12.3.2 AP Placement 356
  12.4 Challenges in Channel Assignment in IEEE 802.11 WLANs 356
  12.5 Channel Assignment Schemes in Centrally Managed Environments 358
    12.5.1 Channel Assignment with AP Placement 358
    12.5.2 Channel Assignment without AP Placement 363
  12.6 Channel Assignment Schemes in Uncoordinated Environments 365
    12.6.1 Least Congested Channel Search (LCCS) 365
    12.6.2 MinMax Approach 366
    12.6.3 MinMax II Approach 367
    12.6.4 Hminmax/Hsum: Weighted Coloring Approach 367
    12.6.5 Pick-Rand and Pick-First Approach 368
    12.6.6 Pick-Rand and Pick-First II Approach 368
    12.6.7 Channel Hopping Approach 368
    12.6.8 Measurement-Based No-Coord 369
  12.7 Comparison among Various Channel Assignment Schemes 369
  12.8 Current Practice in Channel Assignment 371
  12.9 Open Research Issues 372
References 373

Part V Cross-Layer Modeling for Resource Allocation in Wireless Networks

13 Joint PHY/RLC Design in Cellular Wireless Networks 379
  13.1 Introduction 379
  13.2 Radio Link Control (RLC) Protocols: ARQ and HARQ 379
  13.3 Link Adaptation with Adaptive Modulation and Coding (AMC) 381
Contents

13.4 Channel Modeling 383
  13.4.1 I.I.D. Channel Models 383
  13.4.2 Two-State Markov Channel Model 383
  13.4.3 Finite-State Markov Channel Model 386

13.5 ARQ Protocols with I.I.D. Errors 386

13.6 ARQ Protocols in Two-State Markov Channel 388
  13.6.1 GBN-ARQ Protocol in Two-State Markov Channel 389
  13.6.2 SR-ARQ Protocol in Two-State Markov Channel 391

13.7 Truncated ARQ Protocol with Link Adaptation under I.I.D. Channels 392

13.8 Delay Analysis of GBN-ARQ Protocol with Link Adaptation under FSMC 393
  13.8.1 System and Protocol Description 394
  13.8.2 Queuing Model 395
  13.8.3 Derivations of Matrix Blocks in (13.57) 397
  13.8.4 Delay Analysis 400
  13.8.5 Numerical Example 401

13.9 Hybrid ARQ Protocol with Transmission Size Adaptation 402

13.10 Exercises 405

References 412

Index 415
Preface

Wireless communications and networking technology are advancing at a very rapid pace. Newer technologies and standards are evolving to serve the ever-increasing number of users demanding different types of mobile applications and services. Research and development activities on wireless technology constitute one of the most important segments of research and development in the telecommunications area today. Radio resources such as the radio spectrum and transmission power are the fundamental ingredients for any wireless system. Radio resource management is a fundamental problem in wireless networks. Efficient allocation and management of radio resources to serve the mobile users with different requirements is essential for practical deployment and operation of wireless communications systems and networks.

Radio resource allocation is a very broad topic that cannot be fully covered in a single book. This book with the title *Radio Resource Management in Wireless Networks: An Engineering Approach* particularly focuses on resource management issues related to power control, interference management, joint power control and cell association, channel assignment, and multiple access control in traditional as well as emerging wireless networks such as multi-tier cellular, relay-based cellular, and cognitive radio networks (CRNs). This book intends to provide background knowledge on resource allocation in wireless networks from a system-centric or engineering point of view, review the existing related literature, discuss the research challenges, present different techniques to model and analyze the resource allocation problems, and develop both centralized or distributed resource allocation algorithms. This book includes the classical as well as recently developed models and analyses for allocation of different resources (e.g., spectrum and transmission power) in cellular wireless networks. It will be useful for graduate students (M.Sc. and junior Ph.D. students) to understand different resource allocation problems and algorithms in different types of wireless networks (i.e., both infrastructure-based and infrastructure-less wireless networks). It can be used as a textbook for a graduate course on “Wireless Networks” as well as a reference book for researchers and engineers working in the area of wireless communications and networks.

Graduate students who intend to work in this area need to familiarize themselves with the basic concepts of resource management (e.g., power control, channel allocation, error control) in wireless networks and the related mathematical models. The majority of currently existing textbooks on Wireless Communications/Mobile Networks focus on the physical layer (PHY) aspects of wireless communications and do not provide an
in-depth treatment of the resource allocation as well as medium access and radio link control problems in wireless networks. This book intends to fill in this gap and cover the state-of-the-art of research and development in this area. This book can serve as a quick reference for major radio resource management issues in wireless networks as well as related mathematical models for analysis and optimization of radio resource management techniques.

The key features of this book are as follows:

- A systematic view looking at resource management problems in wireless networks that will help readers to classify and compare different types of problem formulations and extend the existing modeling approaches and solution concepts to new system models;
- A comprehensive review of the state-of-the-art research on major resource management problems (e.g., channel and power allocation, error control) in wireless networks;
- Coverage of a wide range of techniques from optimization and game theory (along with examples) for design, analysis, and optimization of resource allocation methods in wireless networks;
- Examples and practice problems (exercises) on different resource management problems;
- Outlines for the key research issues related to resource management in wireless networks.

We have organized the book into the following parts.

Part I: Basics of Wireless Networks

In Chapter 1, starting with an introduction to wireless communications (including signal characterization and modulation, radio propagation and channel models for wireless packet communications), different wireless access technologies (e.g., cellular wireless, wireless local area network [WLAN], wireless metropolitan area network [WMAN], and wireless personal area network [WPAN] technologies) are briefly reviewed. Issues related to medium access control in wireless networks are also discussed.

In Chapter 2, basics of protocol layers for data communication, categorization of different wireless networks, and the fundamentals of modeling and analysis of different physical and radio link layer techniques (including different digital transmission and smart reception techniques, multiple access, scheduling, error control, power control, cell association, handoff management and admission control methods) are discussed. Also, a high-level taxonomy of research areas related to resource management in wireless networks is presented.

Part II: Techniques for Modeling and Analysis of Radio Resource Allocation Methods in Wireless Networks

This part presents different techniques, which can be applied to model and analyze the resource allocation problems in wireless networks. In Chapter 3, optimization techniques are discussed. Major variations of optimization techniques (e.g., unconstrained
and constrained optimization, nonlinear optimization, combinatorial optimization) are presented along with several examples of their applications in modeling and analysis of radio resource management problems.

In Chapter 4, game theory techniques are discussed in the context of optimizing radio resources. The basics of different game theoretic models, namely, non-cooperative game, auction game, and coalition game models, are presented. Several examples of applications of these game models for power control, bandwidth allocation, and sharing in wireless networks are discussed.

Part III: Physical Layer Resource Allocation in Wireless Networks
This part deals with the distributed power control and user association problems in wireless networks. Chapter 5 presents the system model, notations, basic definitions, and relationships related to uplink and downlink power control in interference-limited cellular networks. Both single-cell and multi-cell scenarios are considered. This chapter provides the basics to understand the materials presented in the subsequent chapters of this part of the book.

Starting with a discussion on the motivations of power control, Chapter 6 discusses the conventional open-loop, closed-loop, and centralized power control methods. Then it discusses different existing power control algorithms for homogeneous (single-tier) wireless cellular networks. In this context, different objectives of distributed power control (e.g., aggregate transmit power, system throughput, and outage ratio) are discussed. The existing power control algorithms are compared in terms of different criteria such as fixed-point existence and uniqueness, convergence, rate of convergence, objective functions, and implementation complexity in terms of signaling overhead.

Chapter 7 deals with the joint distributed power and admission control problem in cellular networks. The existing methods for joint admission and power control are discussed, and the state-of-the-art results are summarized.

Chapter 8 deals with the joint power and admission control problem in underlay CRNs, where primary users have a higher priority than secondary users while accessing the radio spectrum. Different optimization models for this problem are introduced, the feasible interference regions for the primary users are characterized, and the existing algorithms for centralized joint power and admission control are reviewed. To this end, two algorithms are presented for joint power and admission control in CRNs.

Chapter 9 deals with the distributed cell association problem (also called the user association problem) in cellular networks. The existing methods for distributed cell association are reviewed, and the problem of distributed joint cell association and power control is studied. Also, methods for generalizing the existing distributed power control algorithms to joint power control and base station (BS) assignment algorithms are discussed.

Part IV: Link Layer Resource Allocation in Wireless Networks
Chapter 10 provides an overview of the channel allocation methods in multi-carrier (e.g., orthogonal frequency division multiple access [OFDMA]) networks. Also, some open research issues are discussed.
Preface

Chapter 11 deals with the resource allocation problem in cooperative (e.g., relay-based) networks. The fundamental aspects of cooperative protocols and resource allocation methods for single and multi-carrier relaying networks are presented.

Chapter 12 presents a survey on the channel allocation methods for wireless local area networks (WLANs). A qualitative comparison among the different methods is made and current practices in channel allocations are discussed. Several open research directions in this area are outlined.

Part V: Cross-Layer Modeling for Resource Allocation in Wireless Networks

Chapter 13 discusses important models for performance analysis and cross-layer (e.g., physical-link) design of radio link level error control methods. In particular, it covers the fundamental aspects of analysis and design of automatic repeat request (ARQ) and hybrid ARQ (HARQ) protocols considering wireless channels with different characteristics.
Part I

Basics of Wireless Networks
1 Introduction

1.1 Basics of a Wireless Communication System

Wireless connectivity is one of the most prominent features of modern communications. A basic wireless communication system consists of three main components: transmitter, wireless medium, and receiver. The transmitter transforms the information into physical signals and transmits it over a transmission medium. The wireless medium, which is also called the communication channel, carries the signal, e.g., the free space carries the electromagnetic waves. The receiver receives physical signals from the medium and converts them into information. Usually, the medium corrupts and distorts the transmitted signal. The aim of any communication system is to obtain the transmitted information at the receiver with the lowest possible error rate. Figure 1.1 depicts the block diagram of a digital communication system. For successful multiuser communication, specific network structures and protocols need to be employed.

1.1.1 Electromagnetic Spectrum and Frequency Range

To transmit information over free space, electromagnetic waves are used. The properties of electromagnetic waves vary with frequency, and each frequency band is suitable for certain applications. The natural electromagnetic spectrum consists of a wide range of frequencies, starting from a few Hz up to $10^{22}$ Hz. Figure 1.2 lists the frequency bands in the electromagnetic spectrum along with the corresponding applications.

Wireless communication is usually possible in a specific frequency range called the radio frequency (RF) band. The RF band consists of frequencies as low as 30 MHz up to 30 GHz. One advantage of RF band is, for most of the part of this band, relatively small antennas can be used to receive and transmit electromagnetic signals. When impinging on the earth’s atmosphere, the RF signals penetrate the ionosphere and the earth’s curvature does not hinder the transmission. The low-frequency RF signals, i.e., shortwave signals, are reflected by the ionosphere and thus can be used to broadcast signals.

While military use of the RF spectrum may vary in different countries, the International Telecommunications Union (ITU) allocates and standardizes the commercial spectrum worldwide. The commercial spectrum bands can be divided into licensed and unlicensed bands. The licensed bands are usually allocated to the operators through an
Figure 1.1 Block diagram of a digital wireless communication system.

Figure 1.2 Electromagnetic spectrum.

auction process by the government authority (e.g., Federal Communications Commission [FCC] in the United States), and they are used exclusively by the corresponding operators. As an example, Table 1.1 lists the licensed commercial frequency allocation in the USA [3]. The unlicensed bands on the other hand are free and can be used...
Table 1.1 Licensed Commercial Frequency Allocation in the United States

<table>
<thead>
<tr>
<th>Application</th>
<th>Frequency Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM radio</td>
<td>535–1605 kHz</td>
</tr>
<tr>
<td>FM radio</td>
<td>88–108 MHz</td>
</tr>
<tr>
<td>Broadcast TV (channels 2–6)</td>
<td>54–88 MHz</td>
</tr>
<tr>
<td>Broadcast TV (channels 7–13)</td>
<td>174–216 MHz</td>
</tr>
<tr>
<td>Broadcast TV (UHF)</td>
<td>470–806 MHz</td>
</tr>
<tr>
<td>4G wireless (LTE, LTE-A)</td>
<td>1850–1910 MHz (uplink), 1930–1990 MHz (downlink), 1710–1755 MHz (uplink), 824–849 MHz (uplink), 869–894 MHz (downlink), 699–716 MHz (uplink), 729–746 MHz (downlink), 777–787 MHz (uplink), 746–756 MHz (downlink), 788–798 MHz (uplink), 758–768 MHz (downlink), 704–716 MHz (uplink), 734–746 MHz (downlink), 1850–1915 MHz (uplink), 1930–1995 MHz (downlink)</td>
</tr>
<tr>
<td>3G wireless</td>
<td>746–764 MHz, 776–794 MHz, 1.7–1.85 MHz, 2.5–2.69 MHz</td>
</tr>
<tr>
<td>1G and 2G cellular</td>
<td>806–902 MHz</td>
</tr>
<tr>
<td>2G cellular</td>
<td>1.85–1.99 GHz</td>
</tr>
<tr>
<td>Wireless comm. services</td>
<td>2.305–2.32 GHz, 2.345–2.36 GHz</td>
</tr>
<tr>
<td>Satellite digital radio</td>
<td>2.32–2.325 GHz</td>
</tr>
<tr>
<td>MMDS</td>
<td>2.15–2.68 GHz</td>
</tr>
<tr>
<td>Satellite TV</td>
<td>12.2–12.7 GHz</td>
</tr>
<tr>
<td>LMDS</td>
<td>27.5–29.5 GHz, 31–31.3 GHz</td>
</tr>
<tr>
<td>Fixed wireless services</td>
<td>38.6–40 GHz</td>
</tr>
</tbody>
</table>

by anyone under some regulations. Many wireless systems including Wi-Fi, Bluetooth, WiMax, and cordless phones use the unlicensed bands. Due to the high number of independent systems, the unlicensed spectrum bands almost always face interference. The licensed and unlicensed spectrum allocations in the United States for different wireless standards are shown in Figure 1.3 [3]. In Table 1.1, LMDS stands for Local Multipoint Distribution Service, and MMDS stands for Multichannel Multipoint Distribution Service. For the evolving fifth-generation (5G) systems, in the United States, the following new spectrum bands are being considered for potential use: 27.5–29.5 GHz, 37–40.5 GHz, 47.2–50.2 GHz, 50.4–52.6 GHz, and 59.3–71 GHz.

1.1.2 Signal Characterization

The information must be converted to physical signals for transmission. In this section we briefly define the signal power, energy, and spectral representation.
Signal Power and Energy

A signal is a function that carries information. It can be a function of one or several independent variables such as time, location, etc. Given a signal $x(t)$, the amount of energy a signal carries is called the signal energy and is defined as

$$ E = \int_{-\infty}^{+\infty} |x(t)|^2 dt. \quad (1.1) $$

The dimension of $|x(t)|^2$ is not necessarily consistent with the definition of energy in physics, but this nomenclature is widely used in electrical engineering and its related fields including wireless communications. While equation 1.1 calculates the total energy of a signal, the energy contained in a specific time interval can be calculated by modifying the integral limits. For example, the energy contained in the interval $[t_1, t_2]$ is obtained as $\int_{t_1}^{t_2} |x(t)|^2 dt$. In the discrete-time system, the energy of signal $x[n]$ where $n$ is an integer is defined as

$$ E = \sum_{n=-\infty}^{+\infty} |x[n]|^2. \quad (1.2) $$

For many signals, the energy could approach infinity. Therefore, the signal energy may not give any useful information. In this case, the signal power is a useful measurement. The signal power is the density of energy along the independent variable. For example, the energy density of $x(t)$ at time instance $t$ equals $|x(t)|^2$, which is the instantaneous power of the signal. Usually we are interested in the average power. The average power of signal $x(t)$ is formally defined as

$$ P = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 dt. \quad (1.3) $$
For a periodic signal, the right-hand side of the equation 1.3 reduces to \( \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 dt \), where \( T \) is the period of the signal. In a discrete-time system, the average power of signal \( x[n] \), where \( n \) is an integer, is defined as

\[
P = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{n=-N}^{N} |x[n]|^2
\]

(1.4)

where \( N \) is the fundamental period of the discrete-time signal.

A signal with finite energy and zero power is generally referred to as an energy signal, e.g., \( x(t) = \text{sinc}(t) \), and a signal with infinite energy and finite power is referred to as a power signal, e.g., \( x(t) = \sin(t) \). Note that both power and energy of a signal may be infinite, e.g., \( x(t) = e^t \).

The following two units are generally used for signal power: dBW (decibel-watt) and dBm (decibel-milliwatt). The power in dBW is given by \( P_{\text{dBW}} = 10 \log_{10} \frac{P_W}{1\text{W}} \). The power in dBm is given by \( P_{\text{dBm}} = 10 \log_{10} \frac{P_{\text{Watt}}}{1\text{mW}} \). For example, if a transmitter transmits 50 watts of power, the transmit power is equivalent to \( 10 \log_{10} 50 = 17 \text{ dBW} \), or \( 10 \log_{10}(50 \times 10^3) = 47 \text{ dBm} \). In general, the decibel is a measure of the ratio between two signal levels, which can be used to denote relative magnitudes or changes in magnitude. This can be expressed mathematically as \( G_{\text{dB}} = 10 \log_{10} \frac{P_{\text{out}}}{P_{\text{in}}} \), where \( G_{\text{dB}} \) = gain (in decibels), \( P_{\text{in}} \) = input power level, and \( P_{\text{out}} \) = output power level.

### Frequency Spectrum of a Signal

It is very useful to represent a signal by its frequency contents to understand the contribution of each frequency component in the construction of the signal. The Fourier transform is the most powerful tool for frequency-domain analysis of a signal. The Fourier transform pair of signal \( x(t) \) is defined as

\[
X(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft} dt
\]

(1.5)

\[
x(t) = \int_{-\infty}^{+\infty} X(f)e^{+j2\pi ft} df
\]

(1.6)

where \( X(f) \) is the Fourier transform of \( x(t) \). Note that \( X(f) \) represents the contribution of frequency \( f \) in construction of signal \( x(t) \). For discrete-time signals, the Fourier transform pair is defined as

\[
X(e^{j2\pi f}) = \sum_{n=-\infty}^{+\infty} x[n]e^{-j2\pi fn}
\]

(1.7)

\[
x[n] = \int_{2\pi}^{2\pi} X(e^{j2\pi f})e^{+j2\pi fn} df
\]

(1.8)

where the argument \( e^{j2\pi f} \) asserts the periodicity of \( X(e^{j2\pi f}) \) with period \( 2\pi \).
The amplitude $|X(f)|$ is a measure of contribution of frequency $f$ in the signal. Moreover, the power spectral density of a deterministic (i.e., not random) signal can be defined as $|X(f)|^2$, which represents the amount of power delivered by frequency $f$. Figure 1.4 shows a signal and the amplitude of its Fourier transform.

The power spectral density (PSD) of a power signal is given by $P_x(f) = \lim_{T_0 \to \infty} \frac{1}{T_0} |X_{T_0}(f)|^2$. In terms of this PSD, the normalized average signal power is $P = \int_{-\infty}^{\infty} P_x(\omega) d\omega$ (or $P = \int_{-\infty}^{\infty} P_x(f) df$).

**Bandwidth of a Signal**

The absolute bandwidth of a signal is defined as the range of frequencies over which the signal has a non-zero power spectral density. For example, the absolute bandwidth of a rectangular pulse is infinity. A more commonly used measure is the first null-to-null bandwidth, which is equal to the main spectral lobe.

1.1.3 Modulation

A physical sinusoidal electromagnetic wave, called the carrier, can carry the information. The process of embedding the information in the carrier is called modulation. When modulated, the properties of the carrier such as amplitude and phase are changed according to the information at hand. The modulation schemes can be divided into two categories: analog and digital.

Digital modulation enables digital communication, which has numerous advantages over its analog counterpart. In analog communication, the transmitted messages are analog waveforms embedded in the carrier phase or amplitude. However, in digital communication, messages are in the form of bit streams embedded in the carrier which are decoded at the receiver. Based on the decoded binary bit stream, the receiver reconstructs the transmitted message (e.g., bit streams). This is in contrast to analog communication where the receiver can only amplify the noise-corrupted received message waveform. This fundamental difference enables digital communication systems to work under arbitrary error rates. The other striking difference between analog and digital
1.1 Basics of a Wireless Communication System

communications is the immense difference in information rate. Through digital techniques, the information rate in modern digital communications can be orders of magnitude higher than that in analog communications.

An important concept in modulation is baseband/passband conversion. Due to the large wavelength, low-frequency signals that are referred to as baseband cannot be transmitted using practical antennas. The signal is therefore transferred to RF frequencies. The frequency contents of the signal remain unchanged but only shifted in frequency. The resultant signal is referred to as a passband signal. Baseband-to-passband conversion and passband-to-baseband conversion are called up-conversion and down-conversion, respectively.

Mathematically, a signal \( x(t) \) is called a baseband signal if \( X(f) \approx 0, |f| > B \) for some \( B > 0 \). It is said to be a passband signal if \( X(f) \approx 0, |f \pm f_c| > B \). The time domain relationship between a baseband signal \( x(t) = x_I(t) + jx_Q(t) \) (which is also called the complex envelope of a real-valued passband signal) and its passband signal \( s(t) \) is given by \( s(t) = Re\{x(t)e^{j2\pi f_c t}\} \) (which implies \( s(t) = x_I(t)\cos2\pi f_c t - x_Q(t)\sin2\pi f_c t \)). \( x_I(t) \) and \( x_Q(t) \) are also referred to as the in-phase \([I]\) component and the quadrature \([Q]\) component, respectively, of the passband signal \( s(t) \), and \( f_c \) is a frequency reference generally chosen in or around the band occupied by \( S(f) \).

### Analog Modulation

In analog modulation, the information is in the form of a continuous signal, e.g., a voice signal. This signal is called the modulating signal. The main analog modulation schemes are amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). In AM, FM, and PM, the amplitude, frequency, and phase of the carrier fluctuate according to the modulating signal. Figure 1.5 demonstrates AM, for example. Denoting the message (modulating) signal by \( x(t) \), the transmitted AM signal on carrier frequency \( f_c \) is given by \( s(t) = [1 + mx(t)]\cos(2\pi f_c t), \) where \( m \) is called the modulation index and \( 0 < m < 1 \).

---

**Figure 1.5** Modulating signal (information) and the modulated carrier in AM.
Digital Modulation

Modern wireless communications rely on digital modulation techniques. In digital modulation, the information is in the form of symbols that carry one or more bits. The information symbols are usually complex numbers whose amplitude and phase determine the amplitude and phase of the carrier signal. We now briefly overview some of the most common digital modulation schemes. The modulating bit stream is assumed to be 001011011010 for each modulation scheme.

PAM

Pulse amplitude modulation (PAM) is a digital modulation scheme based on varying the amplitude. In PAM, \( k \) bits are represented by \( 2^k \) possible amplitudes. For example, with 4-PAM, \( k = 2 \) and \( 2^k = 4 \) different amplitude levels are possible, representing 00, 01, 10, or 11. Figure 1.6 shows a 4-PAM modulated bit stream where the baseband and up-converted passband signals are shown.

PSK

Phase-shift keying (PSK) modulation is based on changing the phase of the carrier. In PSK, \( k \) bits are represented by \( 2^k \) possible complex numbers whose phases determine the phase of the carrier. For example, with \( k = 3 \), eight different phases are possible, each representing a 3-bit stream. It is useful to show the PSK symbols in a complex plane: a representation which is called the signal constellation. Figure 1.7 depicts the signal constellations for 2-PSK, 4-PSK, and 8-PSK. 2-PSK and 4-PSK are often referred to as

![Figure 1.6 4-PAM modulated signal.](https://www.cambridge.org/core/terms). https://doi.org/10.1017/9781316212493.002

![Figure 1.7 Signal constellation of PSK symbols on the complex plane.](https://www.cambridge.org/core/terms).
1.1 Basics of a Wireless Communication System

binary PSK (BPSK) and quadrature PSK (QPSK), respectively. BPSK is also a special case of PAM. Each signal constellation has a corresponding phase for the carrier. Note that the bit streams are mapped to the PSK symbols such that adjacent symbols differ only in one bit, and in this way the bit error rate is minimized. Figure 1.8 shows an 8-PSK modulated bit stream where the phase of the PSK signal is determined according to Figure 1.7.

**QAM**

Quadrature amplitude modulation (QAM) is based on changing the amplitude and phase of the carrier. In QAM, \( k \) bits are represented by \( 2^k \) possible complex numbers whose phase and amplitude determine the phase and amplitude of the carrier. For example, with \( k = 4 \), 16 different complex symbols are possible, each representing a 4-bit stream. Figure 1.9 depicts the signal constellation for 4-QAM, 8-QAM, and 16-QAM. Note that 2-QAM and 4-QAM are identical to BPSK and QPSK. Each symbol determines the phase and amplitude of the carrier. Note that the grid of QAM symbols is designed to achieve certain goals such as minimum power or minimum error rate. Figure 1.10 shows an 8-QAM modulated bit stream where the phase and amplitude of the QAM signal is determined according to Figure 1.9.

**FSK**

Frequency-shift keying (FSK) modulation is based on changing the frequency of the carrier. In FSK, \( k \) bits are represented by \( 2^k \) possible frequencies. Therefore, each bit

---

**Figure 1.8** 8-PSK modulated signal.

**Figure 1.9** Signal constellation of QAM symbols on the complex plane.
Introduction

Figure 1.10 8-QAM modulated signal.

stream determines the frequency of the carrier. For example, with $k = 2$, four different frequencies are possible, representing a 2-bit stream. While FSK occupies a larger bandwidth compared to PAM, PSK, and QAM, it can be very robust and insensitive to many destructive phenomena in the wireless propagation medium. Figure 1.11 shows an 8-FSK modulated signal.

Modulations with Memory
The modulation schemes discussed above were memoryless, where the transmitted symbol depends only on the current data. However, in modulation schemes with memory, the transmitted symbol depends on the current bit stream and also the previous bit stream. Some important examples are differential PSK (DPSK), continuous phase FSK (CPFSK), non-return to zero (NRZ), and non-return to zero inverted (NRZI) modulations. For details, interested readers are referred to [1, 2].

1.1.4 Wireless Channel and Signal Propagation
The signals sent over the wireless channels undergo several destructive phenomena. The most important effects of wireless channel on signal propagation are path-loss and shadowing, which change slowly in time and are known as the large-scale effects, and multipath fading, which changes rapidly in time and is known as the small-scale effect.

Figure 1.11 8-FSK modulated signal.
Path-Loss
An electromagnetic wave propagating through free space loses its power as it travels farther from the source. For a wave with frequency $f$, the free space path-loss is given by [3]

$$L = \frac{(4\pi d)^2}{G\lambda^2}$$  \hspace{1cm} (1.9)

where $d$ is the transmitter/receiver distance, $G$ is the product of transmitter and receiver antenna directivity gains, and $\lambda = c/f$ is the wavelength with $c$ being the propagation speed. Thus, a signal transmitted with power $P_t$, is received with power $P_r = P_t \frac{G}{(4\pi d)^2}$. Note that the free-space loss depends on the frequency of the signal, and high-frequency signals are heavily attenuated compared to low-frequency signals.

The propagation path-loss depends not only on the distance between the transmitter and the receiver and the wavelength, but also on the antenna heights of the transmitter (e.g., a base station [BS]) and the receiver (e.g., a mobile station [MS]). A simple two-ray model (ground-reflection model) can be used to take into account the effect of transmitting and receiving antenna heights. Assuming that there is a line-of-sight (LoS) signal path between the BS and the MS and there is a reflected signal path, with the height of the BS antenna (above the earth), $h_t$, and the height of the MS antenna, $h_r$, the received power for the two-ray model can be approximated as [4] $P_r = P_t G_t G_r h_t h_r^2 d^4$. Note that $P_r$ is frequency-independent and it follows an inverse fourth-power law rather than the inverse-square law.

For urban propagation environments, the path-loss cannot be accurately modeled by the free space model. Over the years, empirical models based on measurements have been proposed for different propagation environments. One such popular empirical path-loss model is the Okumura-Hata model given by [6]

$$L_{dB} = \begin{cases} 
A + B \log_{10} d, & \text{for urban area} \\
A + B \log_{10} d - C, & \text{for suburban area} \\
A + B \log_{10} d - D, & \text{for open area}
\end{cases} \hspace{1cm} (1.10)$$

where $A$, $B$, $C$, and $D$ depend on frequency and transmitter/receiver height and are obtained by empirical formulas. Interested readers may refer to [3, 6] for details on this model and other empirical path-loss models.

It is very useful to express the empirical path-loss models with simplifying approximations. Simplified models are good off-the-shelf tools for performance analysis in many wireless designs. The simplified path-loss model is given by [3]

$$L = K d^{-\eta}$$  \hspace{1cm} (1.11)

where $K$ is a constant that depends on antenna and channel characteristics and $\eta$ is the path-loss exponent. Depending on the environment, the value of $\eta$ can range from 1.6 to 6.5, but values between 2 and 4 are typical. The constant $K$ can be approximated as $K = \frac{\lambda^2 d_0^2}{(4\pi d_0)^2}$, where $d_0$ is the reference far-field distance of antenna. The far-field or
Fraunhofer distance of a transmitting antenna \( (d_f) \) is given by \( d_f = \frac{2D^2}{\lambda} \), where \( D \) is the largest physical linear dimension of the antenna and \( \lambda = \frac{c}{f_c} \) (\( c = \) light speed, m/s; \( f_c = \) carrier frequency, Hz).

Note that in practical systems the path-loss exponent \( \eta \) can be a function of distance \( d \). Therefore, multi-slope path-loss models, where different path-loss exponents hold for different distances, would be more practical for performance analysis and modeling [6].

**Shadowing**

Shadowing or shadow fading occurs when the signal path is blocked by large obstacles such as mountains or buildings, which cause heavy attenuations as long as the receiver remains in the shadow. Figure 1.12 demonstrates shadowing where a car is about to enter the shadow zone.

Shadow fading causes random fluctuations in the signal power and can be modeled statistically; i.e., received signal power due to shadowing can be described by a random variable with a probability density function (PDF). Log-normal distribution is widely used to model shadow fading. Denoting \( \Psi \) as the random variable that models the instantaneous received signal power due to path-loss and shadowing, its PDF can be given by a log-normal PDF with parameters \( \mu \) and \( \sigma \) as follows:

\[
\begin{align*}
    f_{\Psi}(\psi; \mu, \sigma) &= \frac{1}{\psi \sigma \sqrt{2\pi}} e^{-\frac{\ln\psi - \mu^2}{2\sigma^2}}, \\
    \psi &> 0
\end{align*}
\]

where \( \mu \) denotes the average power at the receiver at a certain distance \( d \) from the transmitter (also referred to as area mean power). In a macrocellular environment, the typical value of \( \sigma \) ranges between 8 and 12 dB. Note that the log-normally distributed variable \( \Psi \) has a Gaussian distribution when transformed to \( \Psi_{dB} \). The log-normal PDF is illustrated in Figure 1.13. The log-normal shadowing model has been confirmed by empirical measurements to characterize shadow fading.
Based on the above discussion, we can conclude that, in the presence of path-loss and shadowing, for a separation distance of $d$ between the transmitter and receiver, the signal attenuation $L$ is log-normally distributed (normal in dB) about the mean distance-dependent value and is given by

$$L = Kd^{-\eta}10^{z/10}$$  \hspace{1cm} (1.13)

where $z$ is a zero-mean Gaussian distributed random variable (in dB) with standard deviation $\sigma$ (also in dB). Therefore,

$$L_{\text{dB}}(d) = \bar{L}_{\text{dB}}(d) + z$$  \hspace{1cm} (1.14)

where $\bar{L}_{\text{dB}}(d) = 10 \log_{10}(K) - 10\eta \log_{10}(d)$. Consequently, $P_r(d)$ dBM = $P_t$ dBM $- L_{\text{dB}}(d)$. Note that $P_r(d)$ measured in dB has a Gaussian (normal) distribution with mean $\bar{P}_r(d) = P_t - \bar{L}$. Therefore, if we define the signal outage probability as the probability for the received signal power to fall below a threshold $\gamma$, then it can be easily shown that

$$P_{\text{outage}} = \Pr\{P_r(d) < \gamma\} = Q\left(\frac{\bar{P}_r(d) - \gamma}{\sigma}\right)$$  \hspace{1cm} (1.15)

where

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-\frac{x^2}{2}} dx.$$  \hspace{1cm} (1.16)

Equivalently, the signal coverage probability is given by

$$P_{\text{coverage}} = \Pr\{P_r(d) > \gamma\} = Q\left(\frac{\gamma - \bar{P}_r(d)}{\sigma}\right).$$  \hspace{1cm} (1.17)
Introduction

Figure 1.14 Multipath fading in wireless channels.

For a transmitter such as a BS with omnidirectional transmission with a transmission radius (e.g., cell radius) $R$, the signal area coverage (or area reliability), which is the probability that the received power by a user anywhere in the cell radius is above $\gamma$ (or equivalently, the percentage of area with a received signal that is $\geq \gamma$), can then be defined as follows:

$$P_{\text{coverage}}^{(\text{area})} = \frac{1}{\pi R^2} \int_0^{2\pi} \int_0^R \Pr\{P_r(r) > \gamma\} rdrd\theta. \tag{1.18}$$

Multipath Fading
The most unpredictable and rapid fluctuations in signal amplitude is caused by the multipath fading effect. This phenomenon occurs in multipath environments where the transmitted signal is reflected by many objects and arrives at the receiver along multiple signal paths as demonstrated in Figure 1.14. Because each signal path has a distinct delay, the signals arriving at the receiver will have different phases. This implies that the multipath components may add constructively or destructively. The random fluctuations in the signal amplitude due to the multipath effect are called multipath fading.

When the number of signal paths is large, the central limit theorem can be invoked. Under this condition, it can be shown [3] that the real and imaginary parts of the received baseband signal are independent Gaussian random variables, and the signal amplitude follows a Rayleigh distribution when the LoS signal path is not present and a Rician distribution when the LoS path is present. The Rayleigh and Rician PDFs are, respectively,
given by
\[
    f_X(x; \sigma) = \frac{x}{\sigma^2} \exp \left( -\frac{x^2}{2\sigma^2} \right), \quad x \geq 0 \tag{1.19}
\]
\[
    f_X(x; \nu, \sigma) = \frac{x}{\sigma^2} \exp \left( -\frac{-(x^2 + \nu^2)}{2\sigma^2} \right) I_0 \left( \frac{x\nu}{\sigma^2} \right), \quad x \geq 0 \tag{1.20}
\]
where \(2\sigma^2\) is the average received power of the signal due to path-loss and shadowing, \(\nu^2\) is the power of the LoS component (the non-centrality parameter), and \(I_0(\cdot)\) is the zero-order modified Bessel function of the first kind and is given by
\[
    I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} \exp(x \cos \theta) d\theta. \tag{1.21}
\]

With Rayleigh fading, the received power \((p)\) is exponentially distributed with mean \(2\sigma^2\), and the PDF of \(p\) is given by
\[
    f(p) = \frac{1}{2\sigma^2} e^{-p/2\sigma^2}. \tag{1.23}
\]

While Rayleigh and Rician distributions are suitable for several propagation scenarios, sometimes other distributions provide better models. In particular, the Nakagami-\(m\) distribution provides more flexibility and matches many experimental measurements. The Nakagami-\(m\) PDF is given by
\[
    f_X(x; m, \Omega) = \frac{2^m m^m x^{2m-1}}{\Gamma(m)\Omega^m} \exp \left( -\frac{m}{\Omega} x^2 \right), \quad x > 0, \quad m \geq 1/2 \tag{1.22}
\]
where \(m = \frac{\mathbb{E}[x^2]}{\text{Var}[x^2]}, \Omega = \mathbb{E}[x^2], \Gamma(m) = \int_0^\infty t^{m-1} e^{-t} dt\). For Nakagami-\(m\) fading, the PDF of received signal power is given by
\[
    f_p(p) = \left( \frac{m}{\overline{P}} \right)^m \left( \frac{p^{m-1}}{\Gamma(m)} \right) \exp \left( -\frac{mp}{\overline{P}} \right) \tag{1.23}
\]
where \(\overline{P} = \frac{1}{2} \mathbb{E}[x^2] = \frac{1}{2} \Omega\). A Nakagami-\(m\) PDF can model fading conditions that are either more or less severe than Rayleigh fading (e.g., \(m = 1\) implies Rayleigh fading, \(m \rightarrow \infty\) implies no fading). The Nakagami parameter \(m\) is related to the Rician statistics as
\[
    m = \frac{(1+R_f)^2}{2R_f + 1}, \quad R_f = \frac{\nu^2}{2\sigma^2} \quad \text{and} \quad \Omega = \frac{2\sigma^2}{1-\sqrt{1-1/m}}. \tag{5.12}
\]
Therefore, \(m = 1\) implies \(R_f = 0\), i.e., Rayleigh statistics. The case \(1/2 \leq m < 1\) is not covered by Rician statistics. Figure 1.15 illustrates Rayleigh, Rician, and Nakagami-\(m\) distributions.

Unlike path-loss and shadowing, which are slowly changing phenomena, multipath fading causes the signal to change very rapidly. Therefore, the signal amplitude must be modeled by a random process instead of just a random variable. Let \(\alpha(t)\) denote the time-varying amplitude of the received passband signal due to multipath fading. To characterize \(\alpha(t)\), in addition to its instantaneous PDF (Rayleigh, Rice, or Nakagami), the time variations must also be determined. The autocorrelation function and power spectral density of a random process, which are Fourier transform pairs, describe the behavior of the random process. The time-varying signal amplitude under narrowband fading can be modeled by using Jakes’s model [3]. It can be shown that when the relative speed of transmitter/receiver is \(v\), under uniform scattering, the autocorrelation function
Figure 1.15 Signal amplitude distributions due to multipath fading. All PDFs are normalized to have unit power $E[x^2] = 1$.

and power spectral density of normalized $\alpha(t)$ are, respectively, given by

$$R_\alpha(\tau) = J_0(2\pi f_m \tau)$$

$$S_\alpha(f) = \frac{1}{\pi \sqrt{f_m^2 - f^2}}, \quad |f| < f_m$$

(1.24)

(1.25)

where $f_m = \frac{v}{\lambda}$ is the maximum Doppler frequency and $J_0(\cdot)$ is the Bessel function of the first kind with order zero, which is given by

$$J_0(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{i! \Gamma(i + 1)} \left(\frac{x}{2}\right)^{2i}$$

(1.26)

in which $\Gamma(\cdot)$ is the gamma function given by $\Gamma(i + 1) = i!$, for integer values of $i$. Figure 1.16 depicts these second order statistics.

Figure 1.16 Autocorrelation function and PSD of Jakes’s model.
Channel Impulse Response and Frequency Response

A channel is effectively modeled if the path-loss, shadowing, and multipath fading are all considered. When the delay spread of the signal paths is very small relative to the inverse of signal bandwidth, the channel is said to have a frequency-flat or narrowband fading [3]. In this case, the signal is multiplied by a complex random value, and its frequency contents remain unchanged. If the transmitted signal is denoted by $s(t)$, the received signal in narrowband fading is given by

$$r(t) = \alpha e^{j\theta} \sqrt{\psi L} \times s(t) = h \times s(t) \quad (1.27)$$

where $L$ is the deterministic path-loss, $\psi$ is the random shadowing power coefficient, $\alpha$ is the random multipath fading amplitude, and $\theta$ is the uniformly distributed random multipath fading phase. Note that these parameters are statistically independent. In narrowband fading, the channel impulse response is given by $h(t) = h \times \delta(t)$ with $h$ defined as above and $\delta(\cdot)$ being the Dirac delta function.

When the delay spread of the signal paths is not very small relative to the inverse of signal bandwidth, the channel is said to have a frequency-selective or wideband fading [3]. In this case, the signal is filtered by a complex random system, and its frequency contents are distorted. If the transmitted signal is denoted by $s(t)$, the received signal in wideband fading is given by

$$r(t) = \sqrt{\psi L} \times c(t) \star s(t) \quad (1.28)$$

where $\star$ denotes the convolution operator and $c(t)$ can be modeled as

$$c(t) = \sum_{n=1}^{N} \alpha_n e^{j\theta_n} \times \delta(t - t_n) \quad (1.29)$$

where $N$ is the number of multipath clusters, $t_n$ is the delay of the $n$th cluster, $\alpha_n$ is the random multipath fading amplitude of the $n$th cluster, and $\theta$ is the uniformly distributed random multipath fading phase of the $n$th cluster. Note that in wideband fading, each cluster undergoes narrowband fading. In wideband fading, the channel impulse response is given by $h(t) = \sqrt{\psi L} \times c(t)$. Figure 1.17 depicts the channel impulse response $h(t)$ and the channel frequency response $H(f)$ for a wideband fading channel.

The above characterization of a wireless channel facilitates mathematical analysis of a wireless system. In addition, there are several experimental-based models that may be used to model wireless channels. These include the COST model and the ITU model, which recommend specific losses for a given environment (e.g., urban, rural, hilly, and vehicular, pedestrian) and condition. For details, interested readers are referred to [8, 9].

---

1 Path-loss and shadowing were characterized as power coefficients in the previous sections. Here, to consider the signal amplitude coefficients, their root values are used. Multipath fading is characterized as a signal amplitude coefficient and is used directly.
Introduction

<table>
<thead>
<tr>
<th>t</th>
<th>h(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>H(f)</td>
</tr>
</tbody>
</table>

Figure 1.17 Impulse response and frequency response for a frequency-selective fading channel.

Channel-Gain due to Path-Loss, Shadowing, and Multipath Fading

In the presence of path-loss, shadowing, and multipath fading, for transmit power $P_t$, the statistically varying received power \( P_r(d) \) is

\[
P_r(d) = \alpha^2 10^{\frac{z}{10}} K d^{-\eta} P_t G
\]

where $\alpha$ and $z$ are random variables (RVs) used to represent, respectively, the shadow fading and multipath effects. Therefore, the link-gain $g$ between the transmitter and receiver is given by

\[
g = G \alpha^2 10^{\frac{z}{10}} K d^{-\eta}.
\]

In the case of Rayleigh fading, $f_\alpha(\alpha) = \frac{\alpha}{\sigma^2} e^{-\frac{\alpha^2}{2\sigma^2}}, \alpha \geq 0$ and $E[\alpha^2] = 2\sigma^2$. Therefore,

\[
P_r(d) \text{ dBm} = 10 \log_{10} \alpha^2 + z + 10 \log_{10} K d^{-\eta} + P_t \text{ dBm} + 10 \log_{10} G
\]

\[
= \hat{P}_r(d) \text{ dBm} + 10 \log_{10} \alpha^2
\]

where $\hat{P}_r(d) \text{ dBm}$ (referred to as local mean power) is an RV representing the statistically varying long-term received power in dBm due to shadow fading and path-loss. The received power due to shadow fading is a Gaussian RV with average value $\hat{P}_r(d) \text{ dBm}$, which is referred to as area mean power. Here $\hat{P}_r(d) \text{ dBm} = 10 \log_{10} K d^{-\eta} + P_t \text{ dBm} + 10 \log_{10} G$ with $E[\alpha^2] = 1$ (i.e., $\sigma^2 = 1/2$) and $E[z] = 0$. That is,

\[
f_{\hat{P}_r(d) \text{ dBm}}(p) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(p - \hat{P}_r(d) \text{ dBm})^2}{2\sigma^2}\right).
\]

Since

\[
P_r(d) = \alpha^2 \hat{P}_r
\]

taking the expectation with respect to multipath random variation only (i.e., conditioning on the path-loss and shadow fading variations),

\[
E[P_r(d)] = E[\alpha^2] \hat{P}_r
\]

\[
= \hat{P}_r.
\]
Note that with Rayleigh fading, the PDF of instantaneous received power is exponentially distributed with the average value given by the local mean power $\tilde{P}_r$ as follows:

$$f_{P_r}(p) = \frac{1}{\tilde{P}_r} e^{-\frac{p}{\tilde{P}_r}}.$$

### 1.1.5 Channel Capacity

The maximum transmission rate (in bits per second) with an arbitrarily small error rate in a noisy channel is referred to as the channel capacity. This is defined by Shannon’s formula as

$$R_{\text{max}} = W \log_2 (1 + \text{SNR}) \quad (1.37)$$

where $W$ is the channel bandwidth and SNR is the signal-to-noise ratio defined as $\frac{P}{\sigma^2}$ with $P$ and $\sigma^2$ being the signal power and noise power, respectively, at the receiver. For example, in a channel with 10 kHz bandwidth, if the received signal has a 1 mW power and the receiver noise power is also 1 mW, then the upper bound for the transmission rate would be $R_{\text{max}} = 10^4 \log_2(1 + 1) = 10$ kb/s. This means that the best possible communication system for this channel operates below 10 kb/s.

The above formulation for the channel capacity holds only when the received signal and noise have a flat spectrum across the bandwidth. Narrowband channels usually have such a characteristic. Denoting the received signal and noise spectra as $P(f)$ and $N(f)$, respectively, SNR would be frequency-dependent given as $\text{SNR}(f) = P(f)/N(f)$. The flat spectra $P(f) = P/W$ and $N(f) = \sigma^2/W$ yield $\text{SNR}(f) = P/\sigma^2$, which is independent of frequency. The capacity of a baseband frequency selective channel with bandwidth $W$ can be calculated as

$$R_{\text{max}} = \int_{-W/2}^{W/2} \log_2 \left( 1 + \frac{P(f)}{N(f)} \right) df. \quad (1.38)$$

It is easy to see that (1.38) reduces to (1.37) for flat spectra. The noise spectrum is usually fixed. However, the spectrum of the received signal can be altered by changing the transmission power spectrum. If the channel frequency response is known, then the received signal power spectrum can be predicted because we have $P(f) = |H(f)|^2 P_T(f)$, where $H(f)$ and $P_T(f)$ are the channel frequency response and the transmitted signal power spectrum, respectively. Figure 1.18 shows these spectra along with the resultant SNR for a sample channel. Note that, in this example, the transmitter spends more power in the frequencies with smaller channel-gains to sustain the SNR across the entire bandwidth at the receiver. An important problem in any communication system is to choose the transmission power spectrum $P_T(f)$ such that the channel capacity is maximized. Of course there is always the maximum power constraint as given by $\int P_T(f) df \leq P_{\text{max}}$.

When the channel consists of several narrowband subchannels, then the spectra are approximately flat across each subchannel. In this case, the total capacity is calculated
Introduction

Figure 1.18 Channel frequency response, noise and transmitted signal power spectra, and the resultant SNR at the receiver for a sample channel.

as \[ R^{(\text{max})} = \sum_i W_i \log_2 (1 + \text{SNR}_i) \]  

(1.39)

where \( W_i \) is the bandwidth of the \( i \)-th subchannel and \( \text{SNR}_i = P_i/\sigma_i^2 \) is the SNR of the \( i \)-th subchannel with \( P_i \) and \( \sigma_i^2 \) being the signal power and noise power in the \( i \)-th subchannel, respectively.

In many wireless communication systems, the communication links are not only corrupted by noise, but also by interference from other users. In such a case, assuming that interference is similar to noise, Shannon’s formula can be written as \( C = W \log_2(1 + \text{SINR}) \), where SINR is the signal-to-interference-plus-noise ratio. Thus, for narrowband or frequency-flat fading channels, we have \( \text{SINR} = P/(\sigma^2 + I) \), where \( I \) is the interference power. For wideband or frequency-selective channels, we have \( \text{SINR}(f) = P(f)/(N(f) + I(f)) \). The interference spectrum \( I(f) \) depends on many parameters and may vary considerably across the bandwidth. A smart communication system should be able to deal with high-interference frequencies and make the most out of interference-free or low-interference frequencies.

1.1.6 SINR and Channel Model for Packet Communication Systems

Signal-to-Interference-Plus-Noise Ratio (SINR)

SINR, which is measured at the receiving end of a communication link, is one of the main performance metrics for a wireless communications system. For a receiver located at location \( y \), if the transmitter located at \( x_0 \) transmits with power \( P(x_0) \), the SINR at the...
receiver can be calculated as

$$\text{SINR}(y) = \frac{P(x_0)h_010^{\frac{\gamma_0}{10}}}{P_N + \sum_{x_i \in \Psi_x} P(x_i)h_i10^{\frac{\gamma_i}{10}} - \eta}$$

where $P_N$ is the thermal noise power, $\|x - y\|$ is the Euclidean distance between points located at $x$ and $y$, $h_i$s are multipath power gains, $z_i$s are shadowing gains, and $\Psi_x$ denotes the locations of the interfering transmitters. Note that the SINR is affected by the network geometry along with propagation environment. The thermal noise in watts ($P_N$) for a link of bandwidth of $W$ Hz can be obtained as

$$P_N = k \times T \times W \quad (1.40)$$

where $k$ = Boltzmann’s constant = $1.3803 \times 10^{-23}$ J/K, and $T$ = temperature, in kelvins (absolute temperature). Note that the noise power density in watts per 1 Hz of bandwidth is $N_0 = k \times T$.

**Example**: For a cellular network with a set of users $\mathcal{M}$ and a set of BSs $\mathcal{B}$, let $g_{m,b}$ denote the link-gain between a mobile user $m$ and BS $b$, $P_{m,c}$ denote the transmit power of user $m$ in channel $c$ in the uplink, and $P_{b,c}$ denote the transmit power of BS $b$ in channel $c$ in the downlink. Then the uplink SINR for the link between user $m$ and BS $b$ using channel $c$ is

$$\gamma_{m,c}^{(\text{uplink})} = \frac{P_{m,c}g_{m,b}}{\sum_{n \in \mathcal{M}, n \neq m} P_{n,c}g_{n,b} + P_N + P_N}.$$  \hfill (1.41)

The downlink SINR for the link between user $m$ and BS $b$ using channel $c$ is

$$\gamma_{m,c}^{(\text{downlink})} = \frac{P_{b,c}g_{m,b}}{\sum_{i \in \mathcal{B}, i \neq b} P_{b,c}g_{m,i} + P_N}.$$  \hfill (1.42)

Based on the SINR, important network performance metrics can be obtained as follows [22]:

- SINR outage probability: $\mathcal{O} = \mathbb{P}(\text{SINR} < \theta)$, where $\theta$ is the minimum required SINR,
- SINR coverage probability, $\mathcal{C} = 1 - \mathcal{O}$,
- Bandwidth normalized average rate (or spectral efficiency [SE]), $\mathbb{E}[\ln(1 + \text{SINR})]$,
- Transmission capacity, $T = \lambda(1 - \mathcal{O})$, where $\lambda$ denotes the average number of active links per unit area.

For an interference-limited wireless system, the noise power is generally assumed to be negligible (i.e., SINR $\approx$ SIR), and for a noise-limited system SINR $\approx$ SNR.

**I.I.D. Channel Models for Packet Transmissions**

In general, there is correlation in the channel status over consecutive packet transmissions, and the level of such correlation depends on how fast the channel varies over time. If the channel correlation over two consecutive packet transmissions is sufficiently small, then the independent and identically distributed (i.i.d.) channel model can be
adopted. Otherwise, Markov channels can be employed for highly correlated channel setting. In the following, we describe both i.i.d. and Markov channel models in more detail where channel evolutions over fixed-size transmission time slots are considered. Generally, it is assumed that the channel state remains static in one time slot and may change in consecutive time slots. For the i.i.d. channel model, the channel state in each time slot is one of possible channel states, which is independent from those in the previous time slots. The number of channel states depends on the physical (PHY)-layer design.

**Finite-State Markov Channel (FSMC) Model**

For performance modeling and analysis of wireless protocols, it is often very useful to use an FSMC model. Such a model can capture the correlation (i.e., memory) in the packet error process [11, 12]. For a fading channel, the corresponding FSMC model can be defined based on a finite number of states and the state transition matrix. Let \( S = \{ s_0, s_1, \ldots, s_{K-1} \} \) denote the finite set of states for this FSMC model and \( \{ S_n \}, n = 0, 1, 2, \ldots \) be a Markov process with stationary transitions. The elements of the transition probability matrix \( P \) are

\[
p_{j,k} = \text{Prob}(S_{n+1} = s_k | S_n = s_j),
\]

where \( \sum_{k=0}^{K-1} p_{k,j} = 1, \forall k \in \{0, 1, 2, \ldots, K - 1\} \). If \( \pi_k \) denotes the steady-state probability for state \( k \), then \( 0 < \pi_k \leq 1 \) and \( \sum_{k=0}^{K-1} \pi_k = 1, \forall k \in \{0, 1, 2, \ldots, K - 1\} \).

The states of the FSMC model can be defined by partitioning the range of the received SNR into a finite number of intervals. If \( A_0(=0) < A_1 < A_2 < \cdots < A_K(=\infty) \) denote the thresholds of the received SNR, the channel is said to be in state \( s_k, k \in \{0, 1, 2, \ldots, K - 1\} \), if the received SNR is in the interval \( [A_k, A_{k+1}) \). There is a bit error rate (BER) \( e_k \) associated with each channel state.

For a Rayleigh fading channel, noting that the received instantaneous SNR \( \gamma \) follows an exponential distribution with probability density function

\[
f_{\gamma}(\gamma) = \frac{1}{\gamma_0} \exp\left( -\frac{\gamma}{\gamma_0} \right), \quad \gamma \geq 0
\]

where \( \gamma_0 \) is the average SNR, the steady state probability for each state is given by

\[
\pi_k = \int_{A_k}^{A_{k+1}} f_{\gamma}(\gamma) d\gamma = \exp\left( -\frac{A_k}{\gamma_0} \right) - \exp\left( -\frac{A_{k+1}}{\gamma_0} \right). \tag{1.44}
\]

If the number of states and the corresponding SNR thresholds are given, the transition probabilities and the average BER at each state can be calculated.

For a specific digital modulation scheme, the average BER \( e_k \) is a function of the received SNR, \( \gamma \) (e.g., with BPSK with coherent demodulation, \( e_k(\gamma) = Q(\sqrt{2\gamma}) \)), where \( Q(.) \) is defined as in (1.16). The average BER for each state can then be written as

\[
e_k = \frac{\int_{A_k}^{A_{k+1}} \frac{1}{\gamma_0} \exp\left( -\frac{\gamma}{\gamma_0} \right) e_k(\gamma) d\gamma}{\pi_k}. \tag{1.45}
\]

For a packet transmission system, it can be considered that the channel state transition occurs after one packet transmission time \( t_p \). Assume a slow fading channel with the
channel states associated with consecutive packets assumed to be in neighboring states so that $p_{j,k} = 0$, $\forall |j - k| > 1$. The transition probability $p_{k,k+1}$ from state $s_k$ to state $s_{k+1}$ can be approximated by the ratio of the level crossing rate at threshold $A_{k+1}$ and the average number of packets per second staying in state $s_k$. The transition probability $p_{k,k-1}$ can be approximated by the ratio of the level crossing rate at threshold $A_k$, and the average number of packets per second the SNR falls in the interval associated with state $s_k$ as follows:

$$p_{k,k+1} \approx \frac{N(A_{k+1})\gamma_p}{\pi_k}, \quad k = 0, 1, 2, \ldots, K - 2. \quad (1.46)$$

$$p_{k,k-1} \approx \frac{N(A_k)\gamma_p}{\pi_k}, \quad k = 0, 1, 2, \ldots, K - 1. \quad (1.47)$$

The values of $p_{0,0}$, $p_{K-1,K-1}$, and $p_{k,k}$ ($k = 1, 2, 3, \ldots, K - 2$) are given as $p_{0,0} = 1 - p_{0,1}$, $p_{K-1,K-1} = 1 - p_{K-1,K-2}$, and $p_{k,k} = 1 - p_{k,k-1} - p_{k,k+1}, k = 1, 2, 3, \ldots, K - 2$. Note that, for a Rayleigh fading channel, if $N(\gamma)$ denotes the level crossing rate (in the upward direction or in the downward direction only), then

$$N(\gamma) = \sqrt{\frac{2\pi \gamma}{\gamma_0}} f_m \exp\left(\frac{-\gamma}{\gamma_0}\right) \quad (1.48)$$

where $f_m$ is the maximum Doppler frequency. For a Nakagami-$m$ fading channel,

$$f_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\gamma_0^m \Gamma(m)} \exp\left(\frac{-m\gamma}{\gamma_0}\right) \quad (1.49)$$

where $\Gamma(m) = \int_0^\infty t^{m-1} \exp(-t)dt$. Then

$$N(\gamma) = \sqrt{\frac{2\pi m \gamma}{\gamma_0}} \frac{f_m}{\Gamma(m)} \left(\frac{m\gamma}{\gamma_0}\right)^{m-1}. \quad (1.50)$$

The channel transition probability matrix, therefore, can be written as

$$\mathbf{T} = \begin{bmatrix}
p_{0,0} & p_{0,1} & \cdots & \cdots & 0 \\
p_{1,0} & p_{1,1} & p_{1,2} & \cdots & 0 \\
0 & \ddots & \ddots & \ddots & 0 \\
0 & \cdots & p_{K-1,K-2} & p_{K-1,K-1} & p_{K-1,K} \\
0 & \cdots & \cdots & p_{K,K-1} & p_{K,K}
\end{bmatrix} \quad (1.51)$$

where only a few elements in this transition matrix are non-zero.

For FSMC modeling, the SNR thresholds and the number of states need to be selected based on different criteria. Note that the SNR range of each state should be large enough to cover the SNR variation during a packet time so that for most of the time, a received packet completely falls in one state. The channel state for transmission of the next packet can be the current state or one of the two neighboring states. If the SNR interval of each state is too large, then within a packet transmission time, the received SNR values
will be distributed in a smaller range than the state SNR interval. Therefore, a packet transmitted during that state may have a BER quite different from the BER for the state.

In the following, three methods [12] for selection of SNR thresholds are discussed:

- **Selection of SNR thresholds based on equal channel state probability:** The thresholds are chosen such that \( \pi_0 = \pi_1 = \cdots = \pi_{K-1} \).

- **Selection of SNR thresholds based on time duration of each state:** Let \( \bar{\tau}_k \) denote the average duration of state \( s_k \), which is the ratio of total time the received signal remains between \( A_k \) and \( A_{k+1} \) and the total number of such signal segments, both measured during some long time interval \( T \). Let \( \tau_j \) be the duration of each signal segment. Then \( \bar{\tau}_k \) can be expressed as follows [12]:

\[
\bar{\tau}_k = \frac{\sum \tau_i}{N(A_k)T + N(A_{k+1})T} = \frac{Pr\{A_k \leq \gamma \leq A_{k+1}\}}{N(A_k) + N(A_{k+1})} = \frac{\pi_k}{N(A_k) + N(A_{k+1})}.
\]

Let us assume that we require the average duration of the \( k \)th state to be some multiple of the packet length, i.e., \( \bar{\tau}_k = c_k t_p \), for \( k = 0, 1, 2, \ldots, K - 1 \), where \( c_k \) is a constant. Therefore,

\[
c_k = \frac{\exp\left(-\frac{A_k}{\gamma_0}\right) - \exp\left(-\frac{A_{k+1}}{\gamma_0}\right)}{\sqrt{\frac{2\pi A_k}{\gamma_0}}} \cdot \sqrt{\frac{2\pi A_{k+1}}{\gamma_0}} \exp\left(-\frac{A_{k+1}}{\gamma_0}\right)} + \frac{1}{f_{mt_p}}, \quad k = 0, 1, \ldots, K - 1 \tag{1.52}
\]

with \( A_0 = 0 \) and \( A_K = \infty \).

With a given partitioning of the SNR values and a fixed \( f_{mt_p} \), the corresponding values of \( c_k \) can be obtained. If \( c_k = c \), for \( k = 0, 1, \ldots, K - 1 \), we will have \( K \) equations containing \( K - 1 \) SNR thresholds \( A_1, A_2, \ldots, A_{K-1} \) and \( c \) (and hence a total of \( K \) variables). These equations may be solved for each value of \( K \) of interest for given \( f_{mt_p} \). Based on the SNR thresholds, the transition probability matrix can be obtained.

For a fixed \( f_{mt_p} \), as \( K \) increases, \( c \) decreases, and vice versa. For a fixed value of \( c \), \( K \) increases as \( f_{mt_p} \) decreases, and vice versa. For large value of \( c \) (i.e., small value of \( K \)), packets stay in the current state long enough so that transitions occur to the current and the adjacent states most of the time. However, a smaller value of \( K \) results in a larger SNR interval for each state. For a large SNR interval, different packets in the state may fall in different SNR subintervals and have BERs quite different from the average BER. Therefore, a suitable number of states \( K \) and the value of \( c \) have to be determined for a specific application.

- **Selection of SNR thresholds (based on packet error rate (PER) at each state):** If bit errors are uncorrelated, for packet length of \( l \) bits,

\[
\text{PER} = 1 - (1 - \text{BER})^l. \tag{1.53}
\]
However, information bits may incur different error probabilities for large-size QAM constellations. Also, bit errors are usually correlated when coded transmissions are decoded in the maximum likelihood (ML) sense. Therefore, evaluation of PER based on BER (as in (1.53)) may not be accurate. Instead, the following approximation for PER can be used:

\[
\text{PER}_k(\gamma) \approx \begin{cases} 
1, & 0 < \gamma < \gamma_k \\
0, & \gamma \geq \gamma_k 
\end{cases}
\]

(1.54)

where \( k \) is the state index and \( \gamma \) is the received SNR. Parameters \( a_k, g_k, \) and \( \gamma_k \) (in (1.54)) are state-dependent and can be obtained by fitting (1.54) to the exact PER. The average PER corresponding to state \( k \) is given by

\[
\text{PER}_k = \frac{1}{\pi_k} \int_{A_k}^{A_{k+1}} \text{PER}_k(\gamma) f_\gamma(\gamma) d\gamma \\
= \frac{1}{\pi_k} \int_{A_k}^{A_{k+1}} a_k \exp(-g_k \gamma) f_\gamma(\gamma) d\gamma.
\]

Given \( \text{PER}_k = P_0, \forall k \), the thresholds \( \{A_k\}_{k=1}^{K-1} \) can be obtained as follows [13, 14]:

1. Set \( k = K - 1 \) and \( A_{k+1} = +\infty \).
2. For each \( k \), search for \( A_k \in [0, A_{k+1}) \) that satisfies \( \text{PER}_k = P_0 \).
3. If \( k > 1 \), \( k = k - 1 \) and go to step 2, otherwise go to step 4.
4. Set \( A_0 = 0 \).

The FSMC model has been extensively used for performance analysis and optimization of radio link layer protocols and cross-layer resource allocation (e.g., in [15]–[17]).

1.2 Medium Access in Wireless Networks

For wireless communications, it is possible to combine a group of signals and transmit them on the same medium. If the individual signals belong to one source, this technique is referred to as multiplexing. If each signal belongs to a different source, this technique is referred to as multiple access. Multiplexing and multiple access methods share the available resources such as frequency, time, code, and space. In a wireless network, the medium access control (MAC) protocol determines how multiple users share a common medium for communication and hence the order of transmission of packets by the wireless nodes. The exact functionality of the protocol depends on the respective system and the application at hand. The channel access methods may be divided into two groups: channel partitioning-based MAC and random access-based MAC. We discuss the two groups below.
1.2.1 Channel Partitioning–Based MAC

The channel partitioning is carried out (by a central controller) in the time, frequency, or code domain. The methods of channel allocation discussed in this section are static, whereas random access techniques (described in the next section) are dynamic.

Frequency-Division Multiplexing (FDM)

The available frequency band may be divided into small non-overlapping sub-bands. Each signal is modulated by a different carrier frequency and transmitted over one of the frequency sub-bands. Such a multiplexing method is called frequency-division multiplexing (FDM). The frequency domain-transmitted signal is $Y(f) = \sum_i X_i(f)$, where $X_i(f)$ is the Fourier transform of the $i$th signal. The time domain-transmitted signal $y(t)$ is the inverse Fourier transform of $Y(f)$. With proper demodulation and low pass filtering, the receiver is able to extract individual signals. Figure 1.19 demonstrates the FDM technique. FDM enjoys properties such as continuous transmission and low interference. It divides the channel into different frequency bands and assigns each frequency to a specific user.

Orthogonal Frequency-Division Multiplexing (OFDM)

It is possible to divide the available frequency band into overlapping but orthogonal sub-bands if the sub-band spacing is equal to the inverse of pulse width. This technique, shown in Figure 1.20, is called orthogonal frequency-division multiplexing (OFDM) and occupies smaller bandwidth compared to FDM. Comparison with Figure 1.19 reveals that in the OFDM technique, the same six sources are packed together to save spectrum. Orthogonality means that the inner product of individual signals in the time domain is zero, i.e., $\int x_i(t)x_j^*(t)dt = 0$, for $i \neq j$. For narrowband signals, orthogonality of two sinusoidal pulses of length $T$ holds when we have $f_1 - f_2 = k/T$ for integer $k$. The reason is that $\int_0^T \sin(2\pi f_1 t) \sin(2\pi f_2 t)dt = 0$ when frequency spacing is $k/T$. The receiver exploits the orthogonality property and extracts each signal.
Suppose that $s_i$ is a complex symbol corresponding to a bit stream (for example, a QPSK symbol carries 2 bits of information). The OFDM signal is formed as

$$y(t) = \sum_{i=0}^{N-1} s_i e^{j2\pi f_i t}, \quad 0 < t < T$$ (1.55)

where $f_i$ is the $i$th subcarrier, $N$ is the number of subcarriers, and $T$ is the symbol duration, and we have $f_i - f_j = \frac{k}{T}$ with integer $k$.

The OFDM signal is usually constructed by using an $N$-point discrete Fourier transform (DFT). The most popular DFT implementation method is the fast Fourier transform (FFT) method. DFT can be viewed as samples of the Fourier transform and is implemented as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}$$ (1.56)

$$x[n] = \frac{1}{N} \sum_{n=0}^{N-1} X[k] e^{+j2\pi kn/N}$$ (1.57)

where $k$ is the frequency index and both $x[n]$ and $X[k]$ are $N$ point discrete signals. Then the complex data symbol $s_i$ for $i = 1, \ldots, N$ would be the $N$-point frequency domain DFT. An $N$-point time domain signal is constructed by the inverse DFT of $s_i$ and then fed to a discrete-to-continuous converter. Figure 1.21 shows a 4-point OFDM modulated bit stream where $s_i$s are chosen from an 8-QAM constellation. Figure 1.22 depicts the simplified block diagram of an OFDM transmitter and receiver.

When individual symbols belong to different sources/destinations, this technique is referred to as orthogonal frequency-division multiple access (OFDMA). OFDMA uses the available bandwidth efficiently due to the overlap between adjacent signals, and its demodulation is fast and simple using FFT. Moreover, it can make smart use of channel condition on different subcarriers and also facilitates easy and fast channel equalization. One difficulty in OFDMA systems is the high peak-to-average power ratio (PAPR), which limits the transmission in the uplink. Interested readers are referred to [1, 3, 8] for more information.
**Time-Division Multiplexing (TDM)**

In time-division multiplexing (TDM) or time-division multiple access (TDMA), each signal is transmitted in a specific time-slot using all the available bandwidth. While the signals enjoy a full bandwidth as shown in Figure 1.23, continuous transmission is not possible. Moreover, the receiver must be fully synchronized with the transmitter time slots.

TDM divides time into separate frames, and then a time frame is further divided into a number of time slots. Each time slot is assigned to a specific node, which transmits during its assigned time. Time slot sizes are typically chosen to be long enough to transmit a single packet. With TDM, it is possible to have collision-free transmission. However, it can have poor performance in certain scenarios. For instance, even if only one node is transmitting, it has to wait until its turn to transmit while the channel is idle during the rest of the period.

**Code-Division Multiplexing**

In this case, all users use the same radio spectrum band simultaneously. However, each signal can use a specific code to distinguish itself from others. When all signals

---

**Figure 1.21** Four-point OFDM modulated bit stream where $x[n]$ and $X[k]$ are time domain and frequency domain signals respectively.

**Figure 1.22** Block diagram of the OFDM transmitter and receiver.
are transmitted simultaneously each with its own code, the technique is called code-
division multiplexing (CDM) or code-division multiple access (CDMA). The receiver
can extract the signals from a transmitter only if the corresponding code is known.

CDMA methods use direct-sequence spread spectrum (DSSS) modulation. In this
modulation, each signal is multiplied by a pseudo-noise (PN) sequence as shown in
Figure 1.24. Then all signals are transmitted simultaneously. The transmitted DSSS signal is \( y(t) = \sum_{i} s_i(t)c_i(t) \), where \( s_i(t) \) is the signal of the \( i \)th user and \( c_i(t) \) is the PN sequence of the \( i \)th user. Note that the PN sequences are orthogonal to each other, and the receiver can extract the \( i \)th signal by correlating (multiplying and summing) \( y(t) \) with \( c_i(t) \). Figure 1.25 shows the simplified block diagram of a CDMA transmitter and \( i \)th user receiver.

CDMA systems based on DSSS modulation spread the signal in the frequency
domain. The PN sequences spread each signal in frequency. As a result, the narrow-
band signals with high spectral density are transformed into wideband PN signals with
low spectral density. This feature is shown in Figure 1.26; note that only one signal is
shown here. Sometimes the CDMA signals have a spectral density lower than that of
the background noise, enabling secure and low-power communications. More details
on CDMA systems can be found in [1, 3].
Introduction

Figure 1.25  Block diagram of a QAM-based DSSS CDMA transmitter and the receiver for the \( i \)th user.

Space-Division Multiplexing (SDM)

Space-division multiplexing (SDM) or space-division multiple access (SDMA) systems employ the physical space to enable different users to access the channel simultaneously, i.e., physical space is shared among users. In SDMA, the BS or the access point (AP), knowing the geographical position of users and with the aid of spot beam antennas, transmits the signal of each user on a certain beam directed at that user. Different beams cover different cell areas as shown in Figure 1.27. SDMA avoids unnecessary interference present in the omnidirectional transmission mode. Therefore, the same resources such as time and frequency may be simultaneously used in distinct antenna beams. Moreover, other multiple access methods such as CDMA, TDMA, and OFDMA may

Figure 1.26  The DSSS signal is spread in frequency.
be employed in each direction. SDMA techniques can improve the SINR performance of wireless transmission to a significant degree.

In wireless communications, the transmitter and receiver may use several antennas to exploit the diversity and increase the capacity. Such systems are called multiple-input multiple-output (MIMO) systems. In a MIMO system, several wireless channels are formed between transmitter/receiver antenna pairs. For example, if there are $M$ transmit antennas and $N$ receive antennas, a total number of $MN$ channels are present. The statistical independence of these channels can be exploited in favor of channel capacity or multiple access methods.

When used as a diversity method, MIMO systems can increase the SNR and channel capacity significantly. This is achieved by designing a space-time block code (STBC). An STBC is usually represented by matrix $A$ whose columns correspond to the transmit antennas and whose rows correspond to the time slots. For a transmitter with two antennas, one popular STBC is the Alamouti code for which

$$A = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}$$

(1.58)

where two data symbols $s_1$ and $s_2$ are transmitted over two time slots with two transmit antennas ($(.)^*$ denotes the conjugation). The single antenna receiver is able to detect the symbols using a proper decoder. The resulting system is superior than a classic single antenna system. Higher order STBCs are designed for systems with a larger number of antennas. Figure 1.28 shows a MIMO system consisting of 3 transmitter antennas and 2 receiver antennas. Interested readers are referred to [1, 3] for more details on MIMO systems.
1.2.2 Random Access–Based MAC

In distributed wireless networks with random node locations (e.g., ad hoc/sensor networks, cognitive small cell networks), a decentralized or random access MAC protocol is required to control the access of the network nodes to the shared wireless spectrum in order to limit the mutual interference. With random access MAC protocols, the channel is not partitioned, but it is entirely allocated to a randomly chosen node. There is no scheduling for channel assignment, and this can lead to unwanted packet collisions. The random access MAC protocols should be designed to minimize collisions. Examples of random access MAC include ALOHA, carrier-sense multiple access (CSMA), and carrier-sense multiple access with collision avoidance (CSMA/CA).

In slotted ALOHA, time is divided into equal size slots, and all the nodes are synchronized. A node with a frame to send starts transmitting at the start of a time slot and does so with full link capacity if there are no collisions. If there is a collision, the transmission is terminated, and the frame is re-transmitted in the subsequent time slots until no collision occurs. Pure ALOHA is similar to the slotted ALOHA except that the node does not wait until the start of a new time slot to transmit. Here the transmissions of the nodes are not synchronized.

With CSMA protocols, each transmitter having a packet to transmit has to sense the channel before attempting transmission. If the channel is sensed to be busy, it generates a random backoff timer that decreases only when the channel is sensed idle and freezes when the channel is sensed busy. The transmitter cannot access the channel to transmit its packet until the backoff timer expires. The channel busy/idle decision is based on the carrier-sensing threshold defined by the CSMA MAC protocol. The carrier-sensing threshold defines both the contention domain for each transmitter and the minimum distance between simultaneously active transmitters.

Other than channel partitioning and random access-based protocols, there is another set of MAC protocols known as “taking turns” protocols. Two very common such protocols are polling and token passing protocols. In the polling protocol, a single master node invites slave nodes in a cyclic manner to transmit, while in the token passing protocol a control token is passed among all the nodes in a certain sequence, and the node with the token has the right to transmit. Both of these protocols avoid collisions and offer high efficiencies. However, they suffer from the common drawback of single point failure.

1.2.3 Duplexing

There are three possible modes in which communications take place between two nodes: simplex, half-duplex, and full-duplex. In the simplex mode, only one node transmits while the other node receives. This configuration does not change over time, i.e., the communication is unidirectional. A good example is radio broadcasting in which the station is the transmitter and the personal radio set is the receiver. In the half-duplex mode, the two nodes may both act as a transmitter/receiver, but not simultaneously. Therefore, at any given time, one node transmits while the other node receives; however, the nodes
may change their roles, and the communication direction may be reversed. The communication mode used by walkie-talkie or push-to-talk devices is an example of half-duplex communication mode. In the full-duplex mode, both nodes can transmit and receive at the same time. Modern cellular networks use a full-duplex mode of communication. Many full-duplex systems use different frequencies for transmission and reception. However, it is possible to use the same spectrum band for full-duplex communication, and this mode is referred to as *in-band* full-duplex communication [18]. Figure 1.29 shows the three communication methods. Note that mitigation of self-interference (SI), which is caused by the coupling of the transceiver’s own transmit signal to the receiver while attempting to receive a signal sent by another wireless node, is the major challenge in the implementation of in-band full-duplexing.

1.3 Wireless Access Technologies

1.3.1 Cellular Wireless Technology

In cellular wireless networks, the network coverage area is divided into smaller regions called cells. A BS is generally placed at the center of each cell that serves the users within that cell. Most cellular wireless networks operate using spectrum bands between 450 MHz and 3 GHz. For modeling and analysis, traditionally, in a single-tier cellular network, each cell is represented by a hexagon as shown in Figure 1.30. In a hexagonal grid, each BS is at distance $d$ from its adjacent base stations. The users are randomly scattered in the area, and a user is served by the nearest BS. All BSs are interconnected via fast wired or wireless links. However, in a practical network, due to the variation of the capacity demand across the service area and the infeasibility of deploying base stations in some locations (e.g., rivers, hills, rails, buildings, etc.), there is spatial randomness in the locations of BSs, and they do not have the same coverage area. Therefore, the assumption of hexagonal grids is considered to be very idealized. Moreover, the grid-based modeling approach does not provide tractable results for inter-cell interference (i.e., interference from different base stations), and the performance metrics of interest are usually obtained via Monte Carlo simulations.

Recently, random network models based on stochastic geometry have been used to abstract the topology of cellular wireless networks [22]. With such a model, the locations of the BSs (as well the users) are drawn from some realizations of a stochastic point process in the $\mathbb{R}^2$ plane. The most tractable and well-understood stochastic point
process in the literature is the Poisson point process (PPP). A point process in $\mathbb{R}^d$ is a PPP if and only if the number of points within any bounded region has a Poisson distribution, and the numbers of points within disjoint regions are independent. Figure 1.31 shows an example of such a model for a single-tier cellular network.

Cellular networks use the concept of frequency reuse. In order to reduce interference, each cell uses a portion of the frequency not used by its adjacent cells so that adjacent cells do not cause interference for each other. Therefore, the entire spectrum band is used by a cluster of cells, which is of course reused by other clusters. Different values for frequency reuse factor can be used. With a reuse factor of 1, each cell uses the entire spectrum band. With a reuse factor of 3, three adjacent cells form a cluster, and each uses one-third of the available spectrum band. With reuse factor of 7, seven adjacent cells form a cluster, and each uses one-seventh of the available spectrum band. With a
frequency reuse factor of \( N \), each cell uses \( 1/N \) of the available spectrum band. In a cell, the radio resources (e.g., spectrum and power) are allocated to the users such that a certain goal (e.g., capacity maximization or power minimization) is achieved. Figure 1.32 shows spectrum allocation for three different reuse factors.

To accommodate the rapidly growing user population and the associated traffic load, multi-tier cellular architecture, where small cells (e.g., femtocells, picocells, microcells) underlay the traditional macrocells, is emerging as the next generation cellular network architecture [19]. Such a network is generally referred to as a heterogeneous network (HetNet) or a small cell network (SCN). In a multi-tier network, the small cells can offload users from the congested macrocells to enhance their quality-of-service (QoS) and increase the overall system capacity. The relatively smaller link distance in small cells as well as the smaller number of users served by each small cell can increase the users’ QoS performance. An example of a two-tier HetNet model is shown in Figure 1.33. Note that the techniques for modeling and optimization of legacy cellular wireless networks and algorithms for different network functions (e.g., user association, resource allocation, admission control, power control, etc.) should be revisited and adapted to the HetNet characteristics [20].

Device-to-device (D2D) communication is another key feature of the emerging cellular wireless networks [19, 21]. It enables nearby wireless devices to communicate directly with each other bypassing their corresponding BSs. Since it uses a single-hop communication instead of a dual-hop communication, D2D communication improves the latency, spectral efficiency, and power consumption of D2D transmitters. Also, it offloads traffic from the cellular BSs and thus reduces congestion at the BSs. Potential

![Figure 1.32 Frequency reuse factor of 3, 4, and 7 in cellular wireless networks.](https://www.cambridge.org/core/asset/content/9781316212493/Figure_1.32.png)
applications of D2D communication include localized social networking and data transfer, home automation, and commerce and advertising. Public safety is another application of D2D communication where local connectivity can be ensured in the absence of BSs or hazards at the BSs. An example of a D2D-enabled cellular network is shown in Figure 1.34.
The 3GPP LTE and LTE-Advanced (LTE-A) technologies (also referred to as 4G technologies) are currently the major cellular technologies. LTE supports deployment on different frequency bandwidths such as 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz, and 20 MHz. For the downlink, LTE uses an Orthogonal Frequency Division Multiple Access (OFDMA) air interface, while for the uplink it uses a Single Carrier Frequency Division Multiple Access (SC-FDMA) scheme with a $1 \times 2$ MIMO configuration. The LTE technology can use either a Frequency Division Duplex (FDD) mode or a Time Division Duplex (TDD) mode. The FDD mode uses separate frequencies for downlink and uplink in the form of a band pair. The TDD mode uses one single frequency band. The frame duration in LTE is 10 ms. In the case of FDD, the entire frame is used for uplink or downlink transmissions. In the case of TDD, the frame is divided for uplink and downlink communications. Each frame consists of 10 sub-frames, and the transmission duration of each sub-frame is 1 ms. Each sub-frame is divided into 2 time slots. Each time slot has a duration of 0.5 ms. The resource block (RB) in LTE refers to a time slot spanned with 12 sub-carriers where the bandwidth of each sub-carrier is 15 kHz. LTE supports different types of modulation technique such as QPSK, 16-QAM, and 64-QAM. The peak data rate in uplink/downlink transmission mode depends on the modulation used between the BS and the user. The major LTE specifications are summarized in Table 1.2 [20].

The LTE standards have evolved to LTE-Advanced standards (starting from 3GPP Release 10) [27, 28]. The LTE-Advanced features are summarized in Table 1.3 [20].

For the 5G systems, the aggregate data rate in bits/s/unit area (i.e., total amount of data the network should serve) is expected to be roughly 1000 times that in 4G systems [29]. The worst data rate that a user can reasonably expect to receive when within network range (also referred to as the edge rate) will range from 100 Mbps to as much as 1 Gbps. The best-case data rate that a user can achieve (also referred to as peak data rate), assuming that the entire bandwidth being allocated to a single user, the highest modulation and coding scheme is used, and the maximum number of antennas is supported, is likely to be in the range of tens of Gbps in low-mobility scenarios. Also, in terms of latency, 5G will need to be able to support a round-trip latency of about 1 ms to support real-time control applications, which is an order of magnitude faster than 4G (which is on the order of about 15 ms).

### 1.3.2 WLAN, WMAN, and WPAN Technologies

**WLAN**

Wireless local area networks (WLANs) operate in the unlicensed frequency bands. A WLAN system allows users to stay connected within a local area, e.g., home or office. The central device is called the access point (AP), which is connected to the Internet through a wired or wireless connection. Due to the limited number of users and short-range operations, the power consumption is less when compared to the cellular networks.

A WLAN may be deployed either in infrastructure mode, which requires a central device connected to the Internet, or in ad hoc mode, which does not require a central
### Table 1.2 Major Specifications of the LTE Standard [24]–[27]

<table>
<thead>
<tr>
<th>Specifications</th>
<th>LTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>3GPP Release 8</td>
</tr>
<tr>
<td>Frequency bands</td>
<td>700 MHz, 1.5 GHz, 1.7/2.1 GHz, 2.6 GHz</td>
</tr>
<tr>
<td>Access scheme – Uplink</td>
<td>SC-FDMA</td>
</tr>
<tr>
<td>Access scheme – Downlink</td>
<td>OFDMA</td>
</tr>
<tr>
<td>Channel bandwidth (MHz)</td>
<td>1.4 3 5 10 15 20</td>
</tr>
<tr>
<td>Number of sub-channels</td>
<td>6 15 25 50 75 100</td>
</tr>
<tr>
<td>Number of sub-carriers (1 sub-channel consists of 12 sub-carriers)</td>
<td>72 180 300 600 900 1200</td>
</tr>
<tr>
<td>IDFT/DFT size</td>
<td>128 256 512 1024 1536 2048</td>
</tr>
<tr>
<td>Data modulation</td>
<td>QPSK, 16 QAM, 64 QAM</td>
</tr>
<tr>
<td>Duplexing</td>
<td>FDD</td>
</tr>
<tr>
<td></td>
<td>TDD</td>
</tr>
<tr>
<td>Frame size</td>
<td>1 ms sub-frames</td>
</tr>
<tr>
<td>Sub-carrier spacing</td>
<td>15 kHz</td>
</tr>
<tr>
<td>Channel coding</td>
<td>Convolutional and Turbo Coding</td>
</tr>
<tr>
<td></td>
<td>rate: 78/1024 to 948/1024</td>
</tr>
<tr>
<td>Cyclic prefix length – Short</td>
<td>4.7 μs</td>
</tr>
<tr>
<td>Cyclic prefix length – Long</td>
<td>16.7 μs</td>
</tr>
<tr>
<td>Peak uplink data rate</td>
<td>75 Mbps</td>
</tr>
<tr>
<td></td>
<td>(channel bandwidth: 10 MHz)</td>
</tr>
<tr>
<td>Peak downlink data rate</td>
<td>150 Mbps</td>
</tr>
<tr>
<td></td>
<td>(2 × 2 MIMO, channel bandwidth: 20 MHz)</td>
</tr>
<tr>
<td>User-plane latency</td>
<td>5–15 ms</td>
</tr>
</tbody>
</table>

### Table 1.3 Major Specifications of the LTE-Advanced Standard [27], [28]

<table>
<thead>
<tr>
<th>Standard</th>
<th>Target requirements of LTE-Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak data rate</td>
<td>Uplink: 500 Mbps</td>
</tr>
<tr>
<td></td>
<td>Downlink: 1 Gbps</td>
</tr>
<tr>
<td>(Assuming low mobility and 100 MHz channel bandwidth)</td>
<td></td>
</tr>
<tr>
<td>Peak spectral-efficiency</td>
<td>Uplink: 15 b/s/Hz (up to 4 × 4 MIMO)</td>
</tr>
<tr>
<td></td>
<td>Downlink: 30 b/s/Hz (up to 8 × 8 MIMO)</td>
</tr>
<tr>
<td>Average downlink cell</td>
<td>2.4 b/s/Hz (up to 2 × 2 MIMO)</td>
</tr>
<tr>
<td>spectral-efficiency</td>
<td>2.6 b/s/Hz (up to 4 × 2 MIMO)</td>
</tr>
<tr>
<td></td>
<td>3.7 b/s/Hz (up to 4 × 4 MIMO)</td>
</tr>
<tr>
<td>Average downlink cell</td>
<td>0.07 b/s/Hz (up to 2 × 2 MIMO)</td>
</tr>
<tr>
<td>edge spectral-efficiency</td>
<td>0.09 b/s/Hz (up to 4 × 2 MIMO)</td>
</tr>
<tr>
<td></td>
<td>0.12 b/s/Hz (up to 4 × 4 MIMO)</td>
</tr>
<tr>
<td>Mobility</td>
<td>Considered up to 500 km/h</td>
</tr>
<tr>
<td>Duplexing</td>
<td>FDD and TDD</td>
</tr>
<tr>
<td>User-plane latency</td>
<td>Less than 10 ms</td>
</tr>
</tbody>
</table>
device. All WLAN systems may face interference due to the activity of other systems in the unlicensed frequency bands. The most well-known WLAN standard is IEEE 802.11 (also known as Wi-Fi) in which methods such as OFDM, TDM, and DSSS are used to achieve speeds up to 54 Mb/s [4]. An IEEE 802.11-based WLAN uses a collection of standards that details the MAC sublayer and the physical (PHY) layer.

Two main spectrum bands were specified in the Wi-Fi standard: the 2.4 GHz Industrial, Scientific, and Medical (ISM) band and the 5 GHz Unlicensed National Information Infrastructure (UNII). In the 2.4 GHz band, the standard defines 14 overlapping channels, each of bandwidth 22 MHz whose center frequencies are 5 MHz apart. In the 5 GHz UNII band, there are three sub-bands referred to as low, middle, and high, each of which contains four non-overlapping channels, each with a bandwidth of 20 MHz. The major features of the WLAN standards are summarized in Table 1.4.

In addition to those shown in Table 1.4, there are a few other standards including 802.11ac, 802.11ad, and 802.11af. 802.11ac builds on 802.11n. It utilizes dual band wireless technology, supporting simultaneous connections on both the 2.4 GHz and 5 GHz Wi-Fi bands. 802.11ac offers a data rate up to 1300 Mbps on the 5 GHz band and up to 450 Mbps on the 2.4 GHz band. IEEE 802.11ad defines a new physical layer for 802.11 networks to operate in the 60 GHz millimeter wave spectrum and promises data transmission rates up to 7 Gbps. IEEE 802.11af, also referred to as “White-Fi” and “Super Wi-Fi,” is based on cognitive radio technology, and it allows WLAN operation in TV white space spectrum in the VHF and UHF bands between 54 and 790 MHz.

The IEEE 802.11 MAC specification recommends two spectrum access modes of operation: (i) contention-based distributed coordination function (DCF) and (ii) contention-free point coordination function (PCF). The DCF mode employs the CSMA/CA technique. In the PCF mode (which is available only in an infrastructure-based WLAN), the medium access is controlled by the APs.

---

**Table 1.4 Summary of WLAN Standards (www.ieee.org/11)**

<table>
<thead>
<tr>
<th>Feature</th>
<th>802.11</th>
<th>802.11a</th>
<th>802.11b</th>
<th>802.11g</th>
<th>802.11h</th>
<th>802.11n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum</td>
<td>2.4 GHz</td>
<td>5 GHz</td>
<td>2.4 GHz</td>
<td>2.4 GHz</td>
<td>5 GHz</td>
<td>2.4, 5 GHz</td>
</tr>
<tr>
<td>PHY layer</td>
<td>DSSS, FHSS</td>
<td>OFDM</td>
<td>DSSS</td>
<td>DSSS, OFDM</td>
<td>OFDM</td>
<td>MIMO, OFDM, DSSS</td>
</tr>
<tr>
<td>No. of non-overlapping channels</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Channel bandwidth</td>
<td>20 MHz</td>
<td>10–30 MHz</td>
<td>25 MHz</td>
<td>54 Mbps</td>
<td>54 Mbps</td>
<td>40 MHz</td>
</tr>
<tr>
<td>Maximum data rate</td>
<td>2 Mbps</td>
<td>54 Mbps</td>
<td>11 Mbps</td>
<td>54 Mbps</td>
<td>54 Mbps</td>
<td>600 Mbps</td>
</tr>
<tr>
<td>Transmit power</td>
<td>100 dBm</td>
<td>60 dBm</td>
<td>100 dBm</td>
<td>100 dBm</td>
<td>200 dBm (indoor)</td>
<td></td>
</tr>
</tbody>
</table>
Wireless metropolitan area networks (WMANs) are systems that interconnect users in a relatively large metropolitan area. Central units such as BSs with large coverage are deployed in WMANs. These networks can connect WLANs across a city as well as connect access points to the Internet. The most well-known WMAN technology is the IEEE 802.16-based WiMAX (Worldwide Interoperability for Microwave Access) technology, which enables high-speed connectivity within large metropolitan areas [23]. Many homes and offices use WiMAX to connect their WLAN APs to the Internet.

In the IEEE 802.16 architecture, user equipment is referred to as a subscriber station (SS). The SSs communicate through the BSs in the network. The network operates using the 10–66 GHz (IEEE 802.16) or 2–11 GHz (IEEE 802.16a) band and supports a data rate in the range of 32–130 Mbps depending on the bandwidth of operation as well as the modulation and coding schemes. WirelessMAN-SC is the air interface specification used in the 10–66 GHz band, in which the signal propagation between BS and SS should be line-of-sight and single-carrier modulation is used. The IEEE 802.16a air-interface operates in the 2–11 GHz band and supports non-line-of-sight communication. The following air interface specifications are defined for 802.16a: WirelessMAN-SC2 for single-carrier modulation, WirelessMAN-OFDM for OFDM with a TDMA access scheme, and WirelessMAN-OFDMA for an orthogonal OFDMA scheme.

In the 10–66 GHz band, channel bandwidth of 20, 25, or 28 MHz can be used. For modulation, QPSK, 16-QAM and 64-QAM can be used depending on the channel quality (i.e., SNR at the receiver). The system uses a frame size of 0.5, 1, or 2 ms for transmission, and a frame is divided into subframes for downlink and uplink transmissions. While TDM is used for downlink transmission, TDMA is used for uplink transmission. These subframes are composed of transmission bursts (i.e., uplink and downlink bursts), which carry MAC information and users’ data. Each transmission burst, corresponding to a particular SS, is separated from each other by a preamble field and contains several MAC protocol data units (PDUs). The major 802.16 standards are summarized in Table 1.5.

The IEEE 802.16 uses a connection-oriented MAC that provides a mechanism for requesting bandwidth, transporting, and routing data to higher layer. IEEE 802.16 MAC supports two classes of SS: grant per connection (GPC) and grant per SS (GPSS). In the case of GPC, bandwidth is granted to a connection individually. In contrast, for GPSS, a portion of the available bandwidth is granted to each of the SSs, and each SS is responsible for allocating the bandwidth among the corresponding connections. Since the BS does not need to keep track of allocations for all connections, GPSS is more scalable and efficient than GPC.

The IEEE 802.16 defines the following types of services, each of which has different QoS requirements: unsolicited grant service (UGS), real-time polling service (rt-PS), non-real-time polling service (nrt-PS), and best-effort service (BE). For UGS, the BS generally allocates a fixed amount of bandwidth to each of the connections in a static manner. For PS, the amount of bandwidth required is determined dynamically based on the required QoS performances and the traffic arrival rates. For BE service, there is no
### Table 1.5 Summary of IEEE 802.16 Standards

<table>
<thead>
<tr>
<th>Feature</th>
<th>802.16</th>
<th>802.16a-2004</th>
<th>802.16e-2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency band</td>
<td>10–66 GHz</td>
<td>2–11 GHz</td>
<td>2–11 GHz for fixed; 2–6 GHz for mobile applications</td>
</tr>
<tr>
<td>Application</td>
<td>Fixed LoS</td>
<td>Fixed NLoS</td>
<td>Fixed and mobile NLoS</td>
</tr>
<tr>
<td>MAC architecture</td>
<td>Point-to-multipoint, mesh</td>
<td>Point-to-multipoint, mesh</td>
<td>Point-to-multipoint, mesh</td>
</tr>
<tr>
<td>Transmission scheme</td>
<td>Single carrier</td>
<td>Single carrier, 256 OFDM, or 2048 OFDM</td>
<td>Single carrier, 256 OFDM, or scalable OFDM with 128, 512, 1,024, or 2048 subcarriers</td>
</tr>
<tr>
<td>Modulation</td>
<td>QPSK, 16 QAM, 64 QAM</td>
<td>QPSK, 16 QAM, 64 QAM</td>
<td>QPSK, 16 QAM, 64 QAM</td>
</tr>
<tr>
<td>Gross data rate</td>
<td>32–134.4 Mbps</td>
<td>1–75 Mbps</td>
<td>1–75 Mbps</td>
</tr>
<tr>
<td>Multiplexing</td>
<td>Burst TDM/TDMA</td>
<td>Burst TDM/TDMA/OFDMA</td>
<td>Burst TDM/TDMA/OFDMA</td>
</tr>
<tr>
<td>Duplexing</td>
<td>TDD and FDD</td>
<td>TDD and FDD</td>
<td>TDD and FDD</td>
</tr>
<tr>
<td>Channel bandwidths</td>
<td>20, 25, 28 MHz</td>
<td>1.75, 3.5, 7, 14, 1.25, 5, 10, 15, 8.75 MHz</td>
<td>1.75, 3.5, 7, 14, 1.25, 5, 10, 15, 8.75 MHz</td>
</tr>
</tbody>
</table>

QoS guarantee, and the bandwidth left after serving UGS and PS traffic is allocated for BE service.

**WPAN**

A wireless personal area network (WPAN) interconnects devices that are in close proximity (a few meters). They are used to interconnect different devices and gadgets such as headphones, keyboards, mobile phones, and tablets. A key feature of WPAN technology is that when any two WPAN-equipped devices are in close proximity, they can communicate with each other. Another important feature is the ability of each device to lock out other devices selectively, thus preventing unauthorized access.

Bluetooth and ZigBee are the most popular WPAN technologies, which are based on the IEEE 802.15 standards. The different WPAN standards defined by the IEEE are as follows: 802.15.1 (Bluetooth), 802.15.3 (high-speed WPAN), 802.15.3c (millimeter-wave WPAN), 802.15.4 (ZigBee), 802.15.4a (UWB WPAN), 802.15.5 (personal area mesh network), and 802.15.6 (body area network). Bluetooth is a low-power and low-cost technology that uses short-range radio waves to connect devices that are a maximum of 10 meters away from each other. It uses standard communication protocol targeted toward handling voice, images, and file transfer in ad-hoc networks. Bluetooth
Introduction

Table 1.6 ZigBee vs. Bluetooth

<table>
<thead>
<tr>
<th>Feature</th>
<th>ZigBee</th>
<th>Bluetooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding, modulation</td>
<td>DSSS, OQPSK</td>
<td>FHSS, GFSK</td>
</tr>
<tr>
<td>Access mechanism</td>
<td>CSMA/CA</td>
<td>TDMA</td>
</tr>
<tr>
<td>Power consumption</td>
<td>Lower</td>
<td>Higher (high duty cycle)</td>
</tr>
<tr>
<td>Data rate</td>
<td>250 kbps</td>
<td>1 Mbps</td>
</tr>
<tr>
<td>Connect time</td>
<td>30 ms</td>
<td>1 s</td>
</tr>
<tr>
<td>Audio support</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Complexity</td>
<td>28 Kb protocol stack</td>
<td>250 Kb protocol stack</td>
</tr>
<tr>
<td>Active node density</td>
<td>65536</td>
<td>8</td>
</tr>
<tr>
<td>Application</td>
<td>Suitable for small data (e.g., control data)</td>
<td>Suitable for high burst data (e.g., image, video)</td>
</tr>
</tbody>
</table>

operates in the unlicensed 2.4 GHz band (ISM band) and can use 79 RF channels (with 1 MHz carrier spacing) based on frequency hopping spread spectrum (FHSS) technology. It uses time-division duplexing (TDD) with 625 µs slot duration and 1600 hoppings per second over the 79 channels. The transmission power of a Bluetooth device can vary from 1 mW to 100 mW. It offers a data rate of 1 Mbps (up to 3 Mbps with Bluetooth version 2).

Wibree is another WPAN technology for the 2.4 GHz band, which is complementary to the Bluetooth technology, has a low duty cycle and a power consumption of 10–25% of that of a Bluetooth chip, and therefore is suitable for ultra-low-power short-range wireless connectivity. Wibree is suitable for low-burst data, while Bluetooth is suitable for high-volume data.

The IEEE 802.15.3-based WPAN technology is intended for high-speed multimedia communications in the 2.4 GHz band using 15 MHz bandwidth and 4 channels. It supports 5 data rates from 11 Mbps to 55 Mbps. The transmit power level is about 1 mW, and the transmission range is about 10 m. It uses CSMA/CA-based channel access.

The IEEE 802.15.3c standard-based millimeter-wave (mm-wave) WPAN technology operates in the 57–64 GHz unlicensed band and offers a data rate over 2 Gbps over a transmission range of about 10 m using transmission power of about 10 mW. One main challenge for these WPANs in the 60 GHz band is the propagation loss, which could be 20 dB worse than that at the 5 GHz band.

The IEEE 802.15.4-based ZigBee is a low-data-rate, low-power consumption, and low-cost WPAN technology intended for home automation and control network. It operates on the unlicensed 2.5 GHz (16 channels in 2.4–2.4835 GHz band), 915 MHz (Americas), or the 868 MHz (Europe) band. The data rate is 250 kbps at 2.4 GHz, 40 kbps at 915 MHz, and 20 kbps at the 868 MHz band. The expected transmission range is 10 to 75 meters. ZigBee uses DSSS modulation with a chipping rate of 2 Mchips/sec. The OQPSK modulation scheme is used at the 2.4 GHz band, and BPSK is used in other bands. ZigBee uses CSMA/CA as the channel access mechanism, and it supports star, tree, and mesh network topologies. A comparison between ZigBee and Bluetooth technologies is shown in Table 1.6.
The short-range low-data-rate ultra-wideband (UWB)–based WPAN (IEEE 802.15.4a) is based on an amendment in physical layer of the IEEE 802.15.4 WPAN standard. It can use frequencies from 3.1 GHz to 10.6 GHz (4 channels from 3.1 GHz to 4.8 GHz and 11 channels from 6.0 GHz to 10.6 GHz) and a bandwidth of more than 500 MHz. The maximum transmit power is 41 dBm/MHz. The IEEE 802.15.4a-based WPANs offer scalable data rates of 110 kbps, 851 kbps, 6.81 Mbps, and 27.24 Mbps. Transmissions are realized by short-duration pulse modulation (implemented by an impulse radio transceiver, which has a low-cost and low-power architecture) or OFDM. Due to high bandwidth, UWB transmissions are susceptible to channel interference and narrowband jammers. For UWB WPANs, coexistence with current and possible future standards in 3.1 to 10.6 GHz band is a challenge.

The IEEE 802.15.5 standard is for mesh networking support for both high-rate and low-rate WPANs based on the amendment to the 802.15.3 and 802.15.4 MAC. The following challenges exist for 802.15.5-based wireless personal area mesh networks: MAC challenges due to mobility, hidden and exposed nodes, and interference; routing challenges; and security challenges.

The IEEE 802.15.6 standard is for wireless body area networks (e.g., Bluetooth, ZigBee, Wibree), which use ISM bands as well as frequency bands approved by national medical and/or regulatory authorities. This standard supports different QoS, extremely low power (0.1–1 mW), and data rates up to 10 Mbps while meeting medical (proximity to human tissue) and other communications (interference) regulations.

1.4 Exercises

Exercise 1.1: Using MATLAB, plot variations in path loss (in dB) with separation between transmitter and receiver, for both the free space propagation model and ground reflection (two-ray) model. Assume reasonable (realistic) values for the model parameters.

Exercise 1.2: A mobile receives many multipath signals, but there is no LoS path between the transmitter and the receiver. Assume narrowband channel fading.

i. What probability distribution will the amplitudes of the received signal follow?
ii. Find the cumulative density function of this distribution.
iii. Find the probability that the received signal amplitude will be 10 dB or more below the root mean square (RMS) value of the fading signal.

Exercise 1.3: For a Rician fading channel,

i. Find the mean squared value of the envelope of the Rician distributed signal.
ii. What does the value of fading parameter $K$ imply?
iii. Find the outage probability given the parameters $\sigma^2 = -20$ dBm, $K = 5$, and $P_{\text{min}} = -80$ dBm.
Exercise 1.4: Consider a cellular system where the received signal power is distributed according to a log-normal distribution with mean $\mu = 10$ dBm and variance equal to 8 dB. Assume that the received signal power must be above 6 dBm for acceptable performance. If the radius of the cell $R$ is 1 km, what is the transmission power needed for the base station to achieve a signal outage probability smaller than 1%?

Exercise 1.5: Show that the coverage area for a microcellular system with cell radius $R$, where channel attenuation follows $P_r = P_t \times K \times (d_0/d)^\eta - \sigma$, can be expressed as

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

where $a = \frac{\gamma - P_\text{r}(R)}{\sigma dB}$, and $b = \frac{10 \log_{10} e}{\sigma dB}$. The symbols have their usual meanings.

Following the formula above, find the coverage area for a microcellular system where $\eta = 4$, $d_0 = 1$, $K = 0$ dB, and $\sigma = 8$ dB. Assume that the received power at the cell boundary due to path loss is 10 dB higher than the minimum required received power.

Exercise 1.6: A 3-state Markov channel is used to model a channel that is a Rayleigh fading one. The expected value of the SNR at a particular place is 10 dB. The transmission carrier frequency is 900 MHz. If the mobile is moving at a velocity of 20 m/s, find

i. The SNR boundaries such that the steady state probability of each state is equal.

ii. The level crossing rate at each boundary.

Exercise 1.7: For a Rayleigh fading channel modeled as an FSMC, the channel states are chosen such that the average duration of each channel state is $cT$ where $c$ is a constant and $T$ is the packet duration. Plot the variations in $K$ with $c$ for different values of $f_m$.

Hint: The plots are similar to those in Fig. 1 in [12].

Exercise 1.8: For a Rayleigh fading channel modeled as an FSMC, the channel states are chosen such that the average packet error rate is the same as all the channel states. Obtain the SNR thresholds for the different channel states and tabulate the values for different values of $K$ and $f_m$. Assume suitable values of packet size and also assume that the radio transmitter uses MQAM ($M$-ary Quadrature Amplitude Modulation). Make other assumptions if necessary.

Hint: Use the iterative procedure discussed in Section 1.1.6. You may look at [13] and [14].

References

References


2 Wireless Networks and Resource Allocation

2.1 Protocol Layers for Data Communication

Data communication between two processes (or applications) can be implemented by performing several tasks (i.e., modules) in a hierarchical manner. These modules, when arranged in a vertical stack, form a layered protocol stack. Each layer performs a subset of functions (related to transmission and/or reception) required for communication between the two processes. Such a layer depends on more primitive functions performed by its lower layer, and it also provides services to the upper layers. During the communication process between two processes, the peer layers in the corresponding devices communicate by using a defined set of rules or conventions. This set of rules or conventions is referred to as a protocol at the corresponding layer.

The open system interconnection (OSI) model proposed by the International Organization for Standardization (ISO) [2] defines a generic protocol stack for a data communication network. The OSI model consists of the following seven layers depicted in Figure 2.1: physical, data link, network, transport, session, presentation, and application layers. The lower layers are closer to hardware-based physical data transmission procedures, while the higher layers mostly perform software-based operations. Each layer communicates with its own counterpart; for example, the network layer in the transmitter communicates with the network layer in the receiver. This independently layered structure enables very versatile and reliable networks such as the Internet. In this section, we briefly discuss each of the layers in the OSI protocol stack.

2.1.1 Physical Layer

The physical layer is concerned with the transmission of individual bits over the transmission medium. That is, it is responsible for the modulation and demodulation of the signals at the transmitter and receiver, respectively. The modulation and demodulation techniques depend on the physical medium (e.g., whether it is a “wireline” or a “wireless” medium). For wireless communications, modulation techniques such as PSK, PAM, and QAM and their variants can be used (which were briefly discussed in Chapter 1). For communications over fiber optic cables, typical transmitters are laser diodes that can be modulated to switch between ON and OFF states, which correspond
to transmission of 1 and 0, respectively. The physical layer determines the nature of signals used to encode the digital bits. It performs operations such as channel coding, digital modulation, pulse shaping, and passband/baseband conversion. For example, the bit stream may be used to generate OFDM signals or can be used to generate the baseband Manchester signals as in the case of Ethernet-based LANs.

2.1.2 Data Link Layer

The data link layer is responsible for communication over a link. Note that communication between two devices may involve multiple links (wired and/or wireless links). At the transmitting end, after receiving data from the network layer (e.g., an Internet Protocol [IP] datagram), the link layer forms frames by adding header and footer to the data and sends it to the physical layer for transmission. The physical address (also referred to as the link layer addresses or medium access control [MAC] layer addresses) of the source/destination devices (or interfaces) in the local network are included in the frames generated by the data link layer. The MAC layer addresses are unique for different devices (or interfaces). One of the main functionalities of this layer is to determine how different users should share the transmission medium (i.e., the radio spectrum in a wireless network) in a multiple access communication scenario. The MAC sublayer within the data link layer is responsible for this functionality, and it determines which node may use the shared medium at any given time. The data link layer may also need to perform error detection and/or correction. For example, error detection can be performed by using the cyclic redundancy check (CRC) codes. For error correction, forward error correction (FEC) codes can be used. While FEC detects errors and corrects them at the receiver, CRC only detects the errors. When errors are detected in a frame that are uncorrectable, a retransmission of the frame from the transmitter is generally requested. For this, different automatic repeat request (ARQ) techniques can be used.
2.1.3 Network Layer

The network layer (also called routing layer) forms data structures called datagrams by adding headers to the data received from the transport layer (which is referred to as a transport layer segment). The protocols in this layer are mainly responsible for network addressing, routing, data fragmentation, and error detection [6]. This layer is responsible for determining the routes for a datagram from the source node to the corresponding destination node and forwarding that datagram in the routers along the route, which may span multiple networks. The datagrams originating from the same host may travel different routes to the destination node. The IP is the default protocol at this layer, which acts as a glue for connecting different types of networks (e.g., optical networks, satellite networks, wireless networks). The addressing used in the network layer is different from the MAC address used in the data link layer. The network addresses identify source/destination nodes (i.e., host computers, network routers) in the network. Unlike MAC addresses, these addresses may vary depending on the network to which the host is physically attached (e.g., in a campus network or a home network). The network layer may also perform error detection, and datagrams that are detected to have errors are generally discarded.

2.1.4 Transport Layer

This layer is responsible for communication of messages generated by the applications (e.g., web surfing, file transfer, electronic mail [e-mail], video streaming) between processes running those applications at the transmitter and receiver devices. The messages received from the applications are broken by the transport layer into shorter segments. For some applications (e.g., web surfing, e-mail), the transport layer has to perform reliable and in-order data delivery. Therefore, for these type of applications, it will need to perform error detection and correction. Also, it may perform flow and congestion control (by controlling the rate of transmission at the transmitter) to ensure that the receiver node and the network do not become congested. TCP (Transmission Control Protocol) and UDP (User Datagram Protocol) are the two flagship protocols at the transport layer. While TCP provides a reliable delivery of application messages, UDP does not guarantee any reliability.

2.1.5 Session, Presentation, and Application Layers

The session layer deals with functionalities such as data compression and data encryption, while the presentation layer deals with functionalities such as authentication, authorization, synchronization (e.g., synchronization between audio and video streams), and session restoration (through checkpointing and recovery). The application layer supports different user applications (e.g., e-mail, web, Internet telephony, video conferencing, interactive games, instant messaging, file transfer, remote terminal access, streaming multimedia, network management). A few examples of application layer protocols are HTTP (HyperText Transfer Protocol) for web document request and
transfer, SMTP (Simple Mail Transfer Protocol) for e-mails, and FTP (File Transfer Protocol) for file transfer applications.

These three layers are combined into a single application layer in the Internet protocol suite, which is also referred to as TCP/IP protocol suite. Therefore, the TCP/IP protocol stack consists of the following five layers: physical, data link, network, transport, and application layers.

2.2 Classification of Wireless Networks

Wireless networks can be classified based on the following main criteria: infrastructure, spectrum access, and heterogeneity.

2.2.1 Classification Based on Infrastructure

Based on the infrastructure used to establish communication between two devices or users, wireless networks can be classified as either infrastructure-based or infrastructure-less networks. Cellular wireless networks, relay-enhanced cellular networks, and traditional WLANs are examples of infrastructure-based networks. Wireless ad-hoc and sensor networks are generally infrastructure-less networks. Wireless mesh networks fall between infrastructure-based and infrastructure-less networks. While centralized radio resource allocation methods are more suitable for infrastructure-based networks, decentralized or distributed resource allocation methods would be preferable for infrastructure-less networks.

Infrastructure-Based Networks

Infrastructure-based networks require some fixed installations such as BSs or APs to provide connectivity between mobile devices: between a station and a client of another remote BS or AP, or between a mobile device and a client of a wired network. An infrastructure-based wireless network typically divides the entire service area into a number of smaller service regions called cells (in the context of cellular networks) or basic service sets (BSSs) in the context of WLANs. In each cell/BSS, at least one BS/AP is allocated to provide network service to mobile devices. Each mobile device is associated with a particular BS/AP. Mobile devices do not directly communicate with each other. All communications to and from the mobile devices are performed through the BSs/APs. Connections among BSs/APs are generally provided by a wired backbone network. A BS/AP acts as an interface that receives, buffers, and transmits data between mobile devices in its service area, or between a device in its service area and that in the service area of another BS/AP. Handoff management is one of the major challenges in infrastructure-based wireless networks. A handoff occurs when a mobile user moves out of the transmission coverage of the current BS/AP to the coverage of another BS/AP. For seamless connectivity, reassociation, and authentication have to be performed between cells/BSSs.
2.2 Classification of Wireless Networks

Cellular wireless networks are characterized by their efficiency in large geographical areas. In a cellular wireless network, a base station communicates with portable transceivers directly. A basic schematic of a cellular wireless network is shown in Figure 2.2, where the portable transceivers are drawn as circles and are randomly distributed throughout the cell area. The base station is shown as a tower at the center of the cell. Traditionally, the cells are represented as hexagons.

Relay-based cellular networks enhance the capabilities of cellular networks by employing relay stations inside the cells. As illustrated in Figure 2.3, relays are positioned at certain points in the cell to improve the communication performance between the BS and mobile users, especially those far from the cell center (which have a relatively poor link quality with the BS). Table 2.1 compares cellular wireless network and its relay-based counterpart.

Two types of relays can be used, namely, amplify-and-forward (AF) and decode-and-forward (DF) relays. An AF relay amplifies the signal and transmits it to the destination. The noise is also amplified here. A DF relay decodes the signal to the bit level, then re-encodes the data, and transmits it to the destination.

Infrastructure-less Networks

Infrastructure-less networks are referred to as ad-hoc networks, which are formed by a collection of mobile devices without the use of any preexisting infrastructure or centralized administration (no BSs/APs and functionalities to support mobility). Ad-hoc networks have a decentralized architecture, and they use node-to-node communication. Mobile devices can communicate directly with each other as long as they are within the effective transmission range of each other. If two devices are not within the transmission range of each other, all communications between them pass through one or more relay stations.
intermediate nodes. That is, any node may act as a router to enable communication between two other nodes. Figure 2.4 illustrates the basic topology of such a network. In such a network, mobile nodes can join, leave, and move arbitrarily. Therefore, the topology of the network can change unpredictably.

Wireless sensor networks (WSNs) and wireless mesh networks (WMNs) fall between infrastructure based and ad-hoc networks. In a sensor network, the sensing nodes (with limited computation capabilities) are spread over an area of interest that is referred to as a sensor field. Each node collects data and after simple processing forwards the data to a sink node, which is connected to the Internet. In order to forward the data to the sink node, multi-hop peer-to-peer communication may be required.

A WMN consists of mesh clients and mesh routers, where the mesh routers form a wireless infrastructure/backbone. The mesh routers interwork with the wired networks to provide multihop wireless Internet connectivity to the mesh clients. Different application scenarios for WMNs include backhaul support for cellular networks, home

<table>
<thead>
<tr>
<th>Cellular Networks</th>
<th>Relay-Enhanced Cellular Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low transmission range</td>
<td>Higher transmission range</td>
</tr>
<tr>
<td>High power consumption in transmission (per unit distance)</td>
<td>Lower power consumption in transmission (per unit distance)</td>
</tr>
<tr>
<td>Require a small time window to transmit a specific sized packet</td>
<td>Require a larger time window to transmit the same packet</td>
</tr>
<tr>
<td>Low-quality connection</td>
<td>High-quality connection</td>
</tr>
<tr>
<td>Lower costs</td>
<td>Higher costs (require additional equipment)</td>
</tr>
</tbody>
</table>
2.2 Classification of Wireless Networks

An ad-hoc network.

networks, enterprise networks, community networks, and intelligent transport system networks.

2.2.2 Classification Based on Spectrum Access

Depending on the spectrum usage model, wireless networks can be categorized as either fixed spectrum access networks or dynamic spectrum access networks. The former type of networks use licensed bandwidth that is dedicated to the network and can be used anytime. However, complex frequency planning as well as methods for channel allocation among users are required to avoid co-channel interference.

In dynamic spectrum access (DSA) networks, radio resources (also referred to as spectrum holes) are used opportunistically. Figure 2.5 demonstrates spectrum holes present in a fixed (or licensed) spectrum access network. DSA networks are also referred to as cognitive radio networks (CRNs). In a CRN, the spectrum holes left behind by the primary network are used by the secondary network. Usually, the primary network is a fixed spectrum access network using a licensed bandwidth. The secondary network uses the spectrum holes without causing harmful interference to the primary network. Therefore, the spectral efficiency of the radio resources improves.

2.2.3 Classification Based on Heterogeneity

A wireless network can be either homogeneous or heterogeneous. A homogeneous cellular wireless network uses similar (or same) radio access technologies, cell sizes, and BSs with similar attributes. On the other hand, a heterogeneous cellular network
uses a mixture of network tiers with heterogeneous cell sizes, BSs with different transmit power levels, different radio access technologies (RATs) (e.g., GSM, WCDMA, cdma2000, HSDPA, LTE/LTE-A), and different backhaul connections. Interoperability among different RATs is one of the key aspects in heterogeneous networks. They cooperate with each other in order to provide the users with the best possible connectivity and ensure optimal utilization of the radio resources. A heterogeneous wireless network is also formed when a cellular network interworks with another network such as a WLAN. Such a network utilizes various technologies with different capabilities to meet user demands as much as possible. For example, 3G networks offer a wide coverage with a relatively smaller data rate. WLANs, on the other hand, provide higher data rates, however, with limited coverage. By making use of different technologies, heterogeneous networks aim at satisfying different user requirements.

Figure 2.6 shows an instance of a heterogeneous cellular network in which a macro-cell with a wide coverage area coexists with several picocells and several in-house femtocells together with a set of D2D communications links. As has been mentioned before in Chapter 1, D2D communication is an important concept in the development of heterogeneous networks. In D2D communication, the devices which are close to each other communicate directly without the aid of the BS.

### 2.3 Physical Layer Issues in Wireless Networks

#### 2.3.1 Basic Components

In a wireless network, the physical layer has three basic components: the transmitter, the channel, and the receiver. The transmitter takes the information bits from the source of information and modifies it into a form suitable for wireless transmission. Also, it shapes the signal so that it can pass reliably through the channel, while efficiently using the limited transmission medium resources (i.e., the radio spectrum). Since the devices are often mobile and limited by battery power, the transmitter must use robust
and power-efficient modulation techniques. Also, since the medium is shared with other users, the spectrum of the transmitted signals should be such that it minimizes interference (e.g., adjacent channel interference [ACI]) caused to other users.

The channel provides the physical means for transporting the signals produced by the transmitter and delivering them to the receiver. In a wireless network, the channel impairments include signal attenuation due to distance (which increases with the distance between the transmitter and receiver), fluctuation in the received signal strength due to multipath fading, which causes constructive and destructive interference among several copies of the same signal received, interference, and noise. The channel impairments could be time-varying due to either user mobility or change in the propagation conditions. The interference to a transmitted signal is caused by other transmitters that transmit signals that occupy the same frequency band (or a frequency band that overlaps with) as that of the transmitted signal. The noise is produced by the electronic devices at the front end of the receiver. A brief review on the channel impairments along with their modeling has been provided in Chapter 1. Note that power control techniques (which will be dealt with thoroughly in the later chapters of this book) can be used to compensate for the effect of long-term fading (path-loss and shadow fading). However, it is not effective to combat multipath fading. On the other hand, smart reception techniques such as the diversity reception techniques are effective methods to combat channel fading and thereby increase SNR at the receiver (and hence system capacity).

The receiver operates on the received signal to produce an estimate of the original information-bearing signal. In wireless systems, the receiver frequently estimates the time-varying nature of the channel in order to implement compensation techniques. Also, error detection and/or error correction techniques are implemented at the link layer to improve the reliability of wireless channels.

Figure 2.6 A heterogeneous cellular wireless network with different cell types.
Implementation of efficient modulation and demodulation techniques and mitigation of channel impairments (e.g., through using intelligent reception techniques such as the diversity combining techniques) are major challenges in the physical layer design in wireless networks.

2.3.2 Digital Transmission Techniques

Digital transmission techniques take digital data (i.e., data with finite number of values) as inputs and produce the modulated signals, which can be continuous or discontinuous functions of time. Digital transmission techniques in wireless networks include baseband or pulse transmission techniques (e.g., ultra wideband [UWB] transmission for short-range communication) as well as carrier modulation techniques. While in baseband transmission (e.g., UWB transmission) the signal is transmitted without modulating, with a carrier modulation technique, some characteristic of a carrier wave is varied in accordance with the data bits. Carrier modulation technologies can be categorized into two groups: amplitude, frequency, and phase modulation and spread spectrum modulation.

With UWB transmission, a pulse with a very narrow width (which occupies a spectrum of several GHz) and low power is used for short-range transmission. Pulse amplitude modulation (PAM), pulse position modulation (PPM), and pulse duration modulation (PDM) are other popular baseband modulation techniques. Among these techniques, PPM and PDM are more suitable for wireless networks when compared to PAM. Carrier modulation (narrowband modulation and wideband modulation) shifts the spectral content of a baseband signal inside the operating frequency band of the wireless communication channel and thus allows coexistence of a number of transmissions. Since with increasing carrier frequency, the wavelength and therefore the size of the antenna reduces, the length of the antenna can be reduced to a practical size. However, with increasing carrier frequency, the RF circuit design becomes more challenging, and also the penetration quality of the radio signals becomes limited. Different narrowband modulation techniques were briefly reviewed in Chapter 1.

With wideband modulation such as the spread spectrum modulation, a pseudo-noise (PN) sequence converts a narrowband signal to a noise-like wideband signal before transmission. In a multiple access environment where different transmissions occupy the same spectrum at the same time, each user is assigned a unique PN code which is approximately orthogonal to the codes of other users. At the receiver, the received signals are demodulated through cross-correlation with a locally generated version of the PN sequence. The third generation (3G) cellular systems use direct-sequence spread spectrum (DSSS) as the radio transmission technique. The OFDM technique, which is used in 4G and beyond 4G systems, first converts a block of \( K \) serial symbols into a block of \( K \) parallel symbols. For transmission of those symbols, it uses \( K \) subcarriers (hence subbands), the spectrum of which are closely spaced and overlapped. The OFDM symbols are generated using Inverse Fast Fourier Transform (IFFT). In this way, the symbol duration on the subcarriers is increased relative to the channel delay spread, which makes the transmission robust in a multipath propagation environment. Note that
different modulation techniques (e.g., BPSK, QPSK or QAM) can be used in different subcarriers.

Bandwidth efficiency (or spectral efficiency) and power efficiency are two important considerations in the design of wireless transmission techniques. If \( R \) is the data rate in bits per second and \( W \) is the bandwidth occupied by the modulated RF signal, the bandwidth efficiency \( \xi_b \) is \( \frac{R}{W} \) bps/Hz. For AWGN non-fading channels, the upper bound on the achievable bandwidth efficiency is given by Shannon’s theorem as follows:

\[
\xi_b^{(\text{max})} = \log_2 (1 + \text{SNR}) \quad (2.1)
\]

where SNR is the signal-to-noise (power) ratio at the receiver. In a multiaccess environment, the spread spectrum transmission techniques are generally very bandwidth efficient. The power efficiency \( (\xi_p) \) of a digital modulation scheme is expressed as the ratio of the signal energy per bit to noise power spectral density \( (\frac{E_b}{N_0}) \) required at the receiver input for a certain bit error probability.

The side lobes in the frequency spectrum of the modulated RF signals (and hence the signal bandwidth) can be controlled by using pulse shaping filters (PSFs) before mixing the baseband signals with the carrier, or passing the signal through a band-pass filter after mixing with the carrier. In this case, information bits are first mapped to symbols to obtain the corresponding in-phase/quadrature (I/Q) mapping (i.e., the Cartesian coordinates of the signal constellation). Then line coding is performed in the I/Q branches before the baseband signals can be passed through the PSFs. For example, in the PSFs, the baseband rectangular pulses can be changed to pulses with smoother transitions to reduce the sidelobes. After pulse shaping, the in-phase and quadrature signal components modulate the sinusoidal carriers to be summed to generate the RF signal. For instance, with rectangular pulses with duration of \( T_s = 1 \mu s \) (which corresponds to a data rate of 1 M symbols/sec), the null-to-null baseband bandwidth (i.e., width of the main spectral lobe at baseband) is \( \frac{1}{T_s} = 1 \) MHz, which implies the required RF bandwidth will be 2 MHz. With a Nyquist (or cosine) pulse of the same duration, the required baseband bandwidth is \( \frac{1}{2 T_s} = 0.5 \) MHz and hence the required RF bandwidth is 1 MHz. In general, with pulse shaping, minimum required bandwidth in baseband is \( \frac{1}{2 T_s} (1 + \beta) \), which implies the minimum required RF bandwidth of \( \frac{1}{T_s} (1 + \beta) \), where the parameter \( \beta \) is referred to as the roll-off factor \( (0 \leq \beta \leq 1) \). In other words, given an RF bandwidth of \( B_{rf} \), the maximum signaling rate is \( \frac{B_{rf}}{1+\beta} \).

As an example of pulse shaping, in case of Gaussian minimum shift keying (GMSK) modulation, to reduce the sidelobes, the baseband signals are filtered through Gaussian filters before sending to an FSK modem. Note that MSK is a special type of FSK modulation where the distance between the two frequency tones used for FSK is \( \frac{1}{T_s} \) with \( T_s \) being the duration of the transmitted symbol. In the time domain, the Gaussian filter smoothes sharp transitions of the voltage levels resulting in reduced sidelobes for the

1 Note that the amplitude, frequency, and phase of a modulating carrier wave can be manipulated by manipulating the amplitudes of separate I and Q input signals. That is, it is not required to directly vary the phase of an RF carrier wave, the hardware circuit for which would be difficult to build. The same effect can be achieved by manipulating the amplitudes of input I and Q signals.
modulated signals. The performance of GMSK modulation depends on the bandwidth $B$ of the Gaussian filter and the symbol duration $T_s$. A typical value of $BT_s$ is 0.3 (e.g., in GSM systems). Note that $BT_s = \infty$ corresponds to no filtering.

The symbol error rate ($P_s$) and bit error rate ($P_b$) performances of digital modulation techniques in a narrowband fading environment are important considerations in the selection of modulation techniques. Assuming that the modulation uses ideal Nyquist data pulses ($\text{sinc}[\frac{t}{T_s}]$) so that signal bandwidth $B$ is inverse of the symbol duration $T_s$ (i.e., $B = \frac{1}{T_s}$), let $\gamma_s = E_s/N_0$ and $\gamma_b = E_b/N_0$ denote SNR per symbol (for multi-level signaling, e.g., MPAM, MPSK) and SNR per bit (for binary signaling, e.g., BPSK), respectively, in an AWGN channel. With $M$-ary signaling, if Gray encoding is used for mapping of bits to symbols, then

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

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

The general form for $P_s$ in an AWGN channel under coherent modulation is

$$P_s(\gamma_s) \approx a_M Q(\sqrt{b_M \gamma_s})$$

where $a_M$ and $b_M$ depend on modulation type. Note that, using the following alternate expression for the $Q(.)$ function,

$$Q(z) = \frac{1}{\pi} \int_0^{\pi/2} e^{-\frac{z^2}{2\sin^2 \phi}} d\phi \quad (2.2)$$

$P_s$ can also be expressed as follows:

$$P_s(\gamma_s) \approx \frac{a_M}{\pi} \int_0^{\pi/2} e^{-\frac{b_M \gamma_s}{2\sin^2 \phi}} d\phi. \quad (2.3)$$

Table 2.2 shows the approximate symbol and bit error probabilities for different coherent modulation schemes in an AWGN channel.

In a fading channel, if the random variable $\alpha^2$ represents the instantaneous power gain (i.e., $\alpha$ represents the amplitude fading) of the fading channel, then the instantaneous bit energy of the signal is $\alpha^2 E_b$. Therefore, SNR/bit in the fading channel, $\gamma_b = \alpha^2 E_b/N_0$. Correspondingly, SNR/symbol in the fading channel, $\gamma_s = \alpha^2 E_s/N_0$. The average probability of symbol error in a slow fading channel can be evaluated as

$$P_s = \int_0^{\infty} P_s(\gamma) f(\gamma) d\gamma \quad (2.4)$$

where $P_s(\gamma)$ is the probability of symbol error in AWGN for an arbitrary modulation at a specific value of SNR/symbol $\gamma = \alpha^2 E_s/N_0$, and $f(\gamma)$ is the probability density function of $\gamma$ due to the fading channel. For example, with Rayleigh fading channels,

$$f(\gamma) = \frac{1}{\gamma_s} \exp \left(-\frac{\gamma}{\gamma_s}\right) \quad (2.5)$$
Table 2.2 Approximate Symbol and Bit Error Probabilities for Coherent Modulations [3]

<table>
<thead>
<tr>
<th>Modulation</th>
<th>$P_s(\gamma_s)$</th>
<th>$P_b(\gamma_b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFSK</td>
<td>$Q\left(\sqrt{\gamma_s}\right)$</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
</tr>
<tr>
<td>BPSK</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
</tr>
<tr>
<td>MSK</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
</tr>
<tr>
<td>GMSK</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
<td>$Q\left(\sqrt{2\gamma_b}\right)$</td>
</tr>
</tbody>
</table>

(e.g., $\delta = 0.68$ for $BT = 0.25$, $\delta = 0.85$ for $BT = \infty$)

QPSK, 4QAM

$\approx 2Q\left(\sqrt{\gamma_s}\right)$

$\approx Q\left(\sqrt{2}\gamma_b\right)$

MPAM

$\frac{2(M-1)}{M} Q\left(\sqrt{\frac{6\gamma_b}{M-1}}\right)$

$\approx \frac{2(M-1)}{M} Q\left(\sqrt{\frac{6\gamma_b}{M-1}}\right)$

MPSK

$\approx 2Q\left(\sqrt{2\gamma_b\sin\left(\frac{\pi}{M}\right)}\right)$

$\approx \frac{2}{\log_2 M} Q\left(\sqrt{\frac{2\gamma_b\log_2 M\sin\left(\frac{\pi}{M}\right)}{M-1}}\right)$

MQAM

$\approx 4Q\left(\sqrt{\frac{3\gamma_b}{M-1}}\right)$

$\approx \frac{4}{\log_2 M} Q\left(\sqrt{\frac{3\gamma_b\log_2 M}{M-1}}\right)$

where $\gamma_s = \frac{E_s}{N_0} \alpha^2$ is the average value of the SNR/symbol in the fading channel. For the general approximation $P_s \approx a_M Q\left(\sqrt{b_M \gamma_s}\right)$, the average probability of symbol error in Rayleigh fading channels can be approximated as [3]

$$
\overline{P_s} \approx \int_0^\infty a_M Q\left(\sqrt{b_M \gamma_s}\right) \times \frac{1}{\gamma_s} e^{-\gamma_s} d\gamma = \frac{a_M}{2} \left[ 1 - \sqrt{\frac{0.5b_M \gamma_s}{1 + 0.5b_M \gamma_s}} \right]
\approx \frac{a_M}{2b_M \gamma_s}
$$

where the last approximation is in the limit of high SNR.

One popular approach to compute the average symbol error probability for different modulation schemes in a fading channel is the moment generating function (MGF) approach [3]. The MGF of the probability density function $f(\gamma)$, $\gamma \geq 0$ of a non-negative random variable $\gamma$ is defined as follows:

$$
\mathcal{M}_\gamma (s) = \int_0^\infty f(\gamma) e^{\gamma s} d\gamma.
$$

As an example, for Rayleigh fading with $f(\gamma) = \frac{1}{\gamma_s} e^{-\gamma_s}$, its MGF is given as follows:

$$
\mathcal{M}_\gamma (s) = \int_0^\infty \frac{1}{\gamma_s} e^{-\gamma_s} e^{\gamma s} d\gamma = \frac{1}{1 - \gamma_s s},
$$

(2.7)

Similarly, for Nakagami-$m$ fading,

$$
\mathcal{M}_\gamma (s) = \left( 1 - \frac{s \gamma_s}{m} \right)^{-m}
$$

(2.8)

and for Rician fading with Rice factor $K$,

$$
\mathcal{M}_\gamma (s) = \frac{1 + K}{1 + K - s \gamma_s} e^{K_{\gamma_s m}}.
$$

(2.9)

Note that the Laplace transform of $f(\gamma)$ is then given by: $\mathcal{L}[f(\gamma)] = \mathcal{M}_\gamma (-s)$. 
Now, using the expression for $P_s$ in an AWGN channel as given in (2.3), the average symbol error for any modulation scheme in a fading channel, for which the probability density function of SNR is given by $f(\gamma)$, $\gamma \geq 0$ can be expressed as follows:

$$P_s \approx \frac{a_M}{\pi} \int_0^{\pi/2} \int_{0}^{\infty} e^{-\frac{b_M\gamma}{2\sin^2 \phi}} d\phi f(\gamma) d\gamma$$

where $M(\cdot)$ is the MGF for $f(\gamma)$. Therefore, the average probability of symbol error can be obtained by evaluating a single finite-range integral of the MGF of the probability density function of SNR corresponding to the fading channel. As an example, for the BPSK modulation scheme, the bit error rate in an AWGN channel is given by $P_b = Q(\sqrt{2\gamma_b})$ (i.e., $a_M = 1$ and $b_M = 2$). Based on (2.7), since for a Rayleigh fading channel $M(\gamma) = (1 + \frac{bM}{2\sin^2 \phi})^{-1}$, using (2.10), the average bit error rate of BPSK in a Rayleigh fading channel can be evaluated as follows:

$$P_b = \frac{1}{\pi} \int_0^{\pi/2} \left(1 + \frac{\gamma_b}{\sin^2 \phi}\right)^{-1} d\phi.$$  \hspace{1cm} (2.11)

Besides the symbol error rate performance, other important factors in the design of wireless transmission techniques are out-of-band radiation, resistance to multipath and interference, and constant envelope modulation. The out-of-band radiation refers to the amount of transmitted signal energy lying outside the main lobe (e.g., in the side lobes) of the signal spectrum, and this results in ACI. It is generally desirable to keep the power in the side lobes $-60$ dB below that of the main lobe. Pulse-shaping techniques, through which the baseband rectangular pulses can be changed to pulses with smoother transitions, can be used to control the sidelobes. Different modulation techniques have different degrees of resistance to multipath. For example, wideband modulation techniques such as the DSSS techniques are resistant to multipath fading and interference. Therefore, given a particular demodulator implementation, performance of a modulation scheme in a fading and interference environment is a key metric. With constant envelope modulation (e.g., FSK, GFSK, MSK, GMSK), nonlinear amplifiers (e.g., Class C amplifiers) can be used at the transmitters, which are more power efficient than linear amplifiers, which have to be used for nonconstant envelope modulations (e.g., BPSK, QPSK, QAM).

### 2.3.3 Link Adaptation

With adaptive transmission (also referred to as link adaptation), where the modulation, coding rate, and/or other signal transmission parameters can be adapted according to
the channel conditions, higher bandwidth efficiency can be achieved. For link adaptation, multiple signaling dimensions (e.g., modulation order, frequency in multi-carrier systems, antenna in MIMO systems) can be exploited. With link adaptation achieved through adaptive modulation, under favorable channel conditions, a higher order modulation scheme can be used. On the other hand, when the channel quality degrades, a lower order modulation technique can be used [11], [12], [18], [19]. Such a scheme is generally referred to as adaptive modulation and coding (AMC) scheme. In a MIMO system, link adaptation can be achieved through mapping the bits to the various signals of individual antenna elements as a function of channel characteristics. In a multicarrier system such as an OFDM system, information bits can be mapped over the different carriers with independent coding and modulation. In a MIMO-OFDM system, the adjustable transmission parameters are as follows: modulation level, coding rate, and transmission signaling scheme (spatial multiplexing or transmit diversity).

Since the systems without link adaptation capability are designed for the worst-case channel conditions, they result in inefficient channel utilization. With link adaptation, the transmission parameters are quantized and grouped as a set of modes that are optimal for use in different channel quality regions. With fast link adaptation, adaptive transmission is achieved based on mean SNR or SINR. With slow link adaptation, large-scale channel variations can be exploited based on the slow fading and distance loss or frame error rate (FER) statistics. Also, adaptation can be performed based on a combination of SNR and error statistics information. Faster adaptation leads to a higher gain in spectral efficiency, but it increases the feedback load (i.e., the number of messages to feed back the channel state information [CSI] to the transmitter). For a power-limited cellular scenario, it has been shown that an adaptive system provides a close to a 3-fold gain over a non-adaptive system in terms of spectral efficiency when averaged over the SNR range.

Most of the wireless systems predefine a set of modulation and coding schemes (MCS), each of which corresponds to unique modulation scheme and coding with specific code rate [20], [21], [22]. For example, in LTE wireless cellular systems, the data to be transmitted are encoded by turbo code with different code rates and modulated using one of the following schemes: quadrature-phase shift keying (QPSK), 16-QAM, or 64-QAM, followed by OFDM modulation. For implementation, the receiver estimates the channel quality and transmits this CSI to the transmitter to choose the suitable MCS. In other words, the link adaptation technique adapts the transmission rate over the wireless channel with the SINR.

To perform link adaptation, the received SINR $x$ is partitioned into finite number of intervals. Let $X_0 (= 0) < X_1 < X_2 < \cdots < X_{K+1} (= \infty)$ be the thresholds of the received SINR for different states. The channel is said to be in state $k$ if $X_k \leq x < X_{k+1}$ ($k = 0, 1, 2, \ldots, K$). For example, in channel state $k$, the modulation scheme $2^k$-QAM, $k = 1, 2, \ldots, K$ is chosen. In state 0, no transmission is allowed to avoid high probability of transmission error.

We now discuss the transmission errors and one particular method to choose the SINR partition points $X_k$, which are useful for performance analysis of radio link control (RLC) protocols. For coded systems, since the closed-form derivation for the packet
error rate (PER) is not easy, some suitable approximation should be proposed for design and analysis purposes. One such approximation was proposed in [10] as follows:

\[
\text{PER}_k(x) \approx \begin{cases} 
1, & \text{if } 0 < x < X_{pk} \\
 a_k \exp(-g_k x), & \text{if } x \geq X_{pk} 
\end{cases}
\]

(2.12)

where \(a_k, g_k,\) and \(X_{pk}\) are obtained by curve fitting [10]. Similar approximations for bit error rate (BER) have been used in the literature [18, 19].

Assuming that interference is negligible (e.g., when transmission occur in orthogonal channels), the average PER for each mode \(k\) can then be calculated by using the PDF of the received SNR. Under the general Nakagami-\(m\) fading channel, the PDF of SNR \(x\) at the receiver can be written as follows [19]:

\[
f_X(x) = \frac{m^m x^{m-1}}{\bar{x}^m \Gamma(m)} \exp\left(-\frac{m x}{\bar{x}}\right)
\]

(2.13)

where \(\bar{x} = \mathbb{E}[x]\) is the average SNR, \(\Gamma(m) = \int_0^\infty t^{m-1} \exp(-t) dt\) is the Gamma function, and \(m\) is the Nakagami fading parameter \((m \geq 1/2)\). The Nakagami fading model is very general, and it includes the Rayleigh fading channel as a special case when \(m = 1\) and the Rician fading can also be approximated by the Nakagami model [23].

Using this PDF of the received SNR \(x\), the average PER for each mode \(k\) can be calculated as

\[
\text{PER}_k = \int_{X_k}^{X_{k+1}} a_k \exp(-g_k x) f_X(x) dx = \frac{1}{\Pr(k) \Gamma(m)} \left( \frac{m}{\bar{x}} \right)^m \frac{\Gamma(m, b_k X_k) - \Gamma(m, b_k X_{k+1})}{(b_k)^m}
\]

(2.14)

where \(k = 1, \ldots, K\), \(b_k = m/\bar{x} + g_k\) and \(\Pr(k)\) is the probability of channel state \(k\), which can be calculated as [19]

\[
\Pr(k) = \int_{X_k}^{X_{k+1}} f_X(x) dx = \frac{\Gamma(m, m X_k/\bar{x}) - \Gamma(m, m X_{k+1}/\bar{x})}{\Gamma(m)}
\]

(2.15)

in which \(\Gamma(m, x) = \int_x^\infty t^{m-1} \exp(-t) dt\) is the complementary incomplete Gamma function. In practice, the communication performance degrades due to feedback delay and error. The impacts of feedback error and delay on the bit error rate performance of AMC schemes were investigated in [11, 12].

The packet error rate calculation presented above models the physical layer of a multi-rate wireless network. One important issue is to determine the SINR thresholds \(X_k\) to realize specific design goals. For example, these thresholds can be calculated such that all channel states have the same probability [24, 26]. In [25], a searching algorithm was proposed to find the SINR thresholds while constraining \(\text{PER}_k = P_0\). The SINR thresholds chosen by this algorithm allow to guarantee the PER in the physical layer. The algorithm is presented below.
2.3 Physical Layer Issues in Wireless Networks

Algorithm 2.1: Algorithm for searching the SINR thresholds

1. Set $k = K$ and $X_{K+1} = \infty$.
2. For each $k$, search for $X_k \in [0, X_{k+1}]$ such that PER$_k = P_0$.
3. If $k > 1$, go to 2, otherwise go to 4.
4. Set $X_0 = 0$.

Here the design parameter is $P_0$ which has to be chosen to achieve desired system performance.

2.3.4 Diversity Transmission Techniques

Diversity transmission is an effective way to combat multipath channel fading and increase SNR at the receiver and, thereby, increase link capacity. Note that methods such as power control, which can compensate for the effect of path-loss and shadow fading, is not effective to combat multipath fading. Diversity techniques use more than one independently faded version of the transmitted signal. The basic concept here is: if several replicas of the transmitted signal carrying the same information are received over multiple channels that experience independent fading, the probability that all the independently faded signal components experience deep fading simultaneously is reduced significantly. Diversity transmission can be achieved in time, frequency, or space by providing signals at different arrival times, signals at different carrier frequencies, or signals transmitted through different antennas.

In a wireless channel, the multipath signals inherently provide some time diversity and a smart receiver such as a RAKE receiver in a DSSS CDMA system exploits this diversity to improve the SNR performance. Frequency diversity exploits the fact that frequencies separated by more than the coherence bandwidth of the channel will experience uncorrelated channel fading. A multipath fading channel, which is frequency selective (i.e., symbol transmission rate is higher than the coherence bandwidth of the channel), inherently provides frequency diversity. Multicarrier modulation techniques can exploit (or implement) frequency diversity by simultaneously transmitting the modulated signals across a large bandwidth so that even if a particular frequency undergoes fading, the composite signal can still be demodulated. In addition to time and frequency diversity, multipath signals also exhibit spatial diversity signal strength, polarization of the waveform, delay, and the angle of arrival of each path. Spatial diversity can be implemented by using multiple antennas located in different locations, multiple antennas with different polarization located in the same location, sectored antenna limiting the angle of arrival of the signal, or an adaptive antenna array (smart antenna) that changes its antenna pattern adaptively. Smart antennas can reduce multiple access or co-channel interference by focusing the signals in one direction only, thereby increasing the link capacity.

2.3.5 Smart Reception/Diversity Combining Techniques

With smart reception techniques, after obtaining independently faded signal components at the output of the demodulators, these signal components are combined to detect
the transmitted signal. The popular diversity combining techniques [61] are described below.

- **Selection combining (SC):** The receiver monitors the SNR value of each diversity channel and chooses the one with the maximum SNR value for signal detection. Consider $M$ independent Rayleigh fading channels available at the receiver. Each channel is called a diversity branch. Let each branch has an instantaneous SNR $\gamma_i$. For Rayleigh fading channels, the fading amplitude $\alpha$ has a Rayleigh distribution, so the fading power $\alpha^2$ and consequently the probability distribution of the SNR is

$$f(\gamma_i) = \frac{1}{\gamma_i} e^{-\frac{\gamma_i}{\overline{\gamma}}}, \quad \gamma_i \geq 0$$

where $\overline{\gamma}$ is the mean SNR of each branch. Assume that each branch has the same average SNR given by

$$\text{SNR} = \overline{\gamma_i} = \frac{E_s}{N_0} \alpha^2 = \overline{\gamma_s}.$$  

(2.17)

The probability that a single branch has an instantaneous SNR less than some threshold $\gamma$ is

$$\Pr[\gamma_i \leq \gamma] = \int_{0}^{\gamma} f(\gamma_i) d\gamma_i = 1 - e^{-\frac{\gamma}{\overline{\gamma}}}.$$  

(2.18)

Therefore, assuming that all branches have the same average SNR $\overline{\gamma_s}$, the probability that all $M$ independent diversity branches receive signals that are simultaneously less than some specific SNR threshold $\gamma$ is

$$\Pr[\gamma_1, \ldots, \gamma_M \leq \gamma] = \left(1 - e^{-\frac{\gamma}{\overline{\gamma}_s}}\right)^M = F(\gamma).$$  

(2.19)

Therefore, the probability that SNR $> \gamma$ for one or more branches (i.e., probability of exceeding a threshold when selection diversity is used) is

$$\Pr[\gamma_i > \gamma] = 1 - \Pr[\gamma_1, \ldots, \gamma_M \leq \gamma] = 1 - \left(1 - e^{-\frac{\gamma}{\overline{\gamma}_s}}\right)^M.$$  

(2.20)

For selection diversity, the average SNR is

$$\overline{\gamma}_{\text{selection}} = \int_{0}^{\infty} \gamma f(\gamma) d\gamma$$

(2.21)

where $f(\gamma) = \frac{1}{\overline{\gamma}} e^{-\frac{\gamma}{\overline{\gamma}}}$, $F(\gamma) = \frac{M}{\overline{\gamma}} \left(1 - e^{-\frac{\gamma}{\overline{\gamma}_s}}\right)^{M-1} e^{-\frac{\gamma}{\overline{\gamma}}}$.

Equation (2.21) can be evaluated to yield the average SNR improvement offered by selection diversity as follows:

$$\frac{\overline{\gamma}_{\text{selection}}}{\overline{\gamma}_s} = \frac{1}{\sum_{k=1}^{M} \frac{1}{k}}.$$  

(2.22)

- **Maximal ratio combining (MRC):** The signals $r_i$ from each of the $M$ diversity branches are co-phased, weighted, and then summed. That is, with knowledge of
the complex channel-gains $\alpha_i(t) e^{j\theta_i(t)}$ ($i = 1, 2, \ldots, M$), the phase distortion $\theta_i(t)$ of
the received signal is removed from the $i$th branch by multiplying the signal component
by $e^{-j\theta_i(t)}$. The coherently detected signal is then weighted by the corresponding
amplitude gain $\alpha_i(t)$. The weighted received signals from all the $M$ branches are then
summed together and applied to the decision device. It can be shown that the SNR $\gamma_i$
at the output of the diversity combiner is the sum of the SNRs in each branch (i.e.,
$\gamma_i = \gamma_1 + \cdots + \gamma_M$). Also, the PDF for $\gamma_i$ can be shown to be [61]
\begin{equation}
    f(\gamma) = \frac{\gamma^{M-1} e^{-\gamma/\overline{\gamma}_s}}{\overline{\gamma}_s^M (M-1)!}, \quad \gamma \geq 0
\end{equation}
where $\overline{\gamma}_s$ is the average SNR in each diversity branch.

- **Equal gain combining (EGC):** The MRC approach requires an accurate estimate of
  the channel amplitude gain $\alpha_i(t)$, which increases the receiver complexity. EGC is
  a simpler approach that gives equal weight to all the signals equally after coherent
detection (which removes the phase distortion $\theta_i(t)$). The coherently detected signals
from all the $M$ branches are simply added and applied to the decision device. The
performance of EGC has been shown to be only marginally inferior to MRC and
superior to SC.

A RAKE receiver used in a spread spectrum communication system is a practical
example of diversity combining receiver. It exploits the diversity provided by the multi-
path components when their relative propagation delays exceed a chip period. By using a
separate correlation receiver for each of the multipath signals, it combines these signals
by using any standard diversity combiner, e.g., SC, EGC, or MRC, in order to improve
the SNR at the receiver. An MRC RAKE receiver, which weighs the received signal
from each branch by the signal amplitude at that branch, provides the optimum SNR
performance. Note that without combining, the signal arriving from each path would be
a wideband interference to the signals arriving from other paths and thus would degrade
the SNR performance of a DSSS system.

To analyze the average symbol error performance for the diversity reception tech-
niques in fading channels, the MGF approach described before can be used. For exam-
ple, with $M$-branch MRC, since the SNR at the output of the diversity combiner is the
sum of the SNRs in each branch (i.e., $\gamma_i = \gamma_1 + \cdots + \gamma_M$), assuming that the branch
SNRs are independent, the average symbol error rate for a modulation scheme can be
evaluated as follows:
\begin{align}
P_s & \approx \frac{a_M}{\pi} \int_0^{\pi/2} e^{-\frac{b_M}{\sin^2 \phi}} d\phi \int_0^\infty \cdots \int_0^\infty f(\gamma_1, \ldots, \gamma_M) d\gamma_1 \cdots d\gamma_M \\
& = \frac{a_M}{\pi} \int_0^{\pi/2} e^{-\frac{b_M}{\sin^2 \phi}} d\phi \int_0^\infty \cdots \int_0^\infty f(\gamma_1) \cdots f(\gamma_M) d\gamma_1 \cdots d\gamma_M \\
& = \frac{a_M}{\pi} \prod_{i=1}^M \mathcal{M}_{\gamma_i} \left(-\frac{b_M}{2 \sin^2 \phi}\right) d\phi.
\end{align}
Note that, using (2.24), the average symbol error rate can be computed when the branch
SNRs follow different distributions. If the branch SNRs are independent and identically
distributed (i.i.d.), then

\[ P_s = \frac{a_M}{\pi} \int_{0}^{\pi/2} \left[ \mathcal{M}_\gamma \left( -\frac{b_M}{2 \sin^2 \phi} \right)^M \right] d\phi \]  

(2.25)

where \( \mathcal{M}_\gamma(s) \) is the common MGF for the branch SNRs. As an example, the average probability of bit error for BPSK modulation under two branch MRC assuming Rayleigh fading in the first branch with \( \gamma_1 = 10 \) dB and Nakagami-\( m \) fading in the second branch with \( m = 2 \) and \( \gamma_2 = 15 \) dB, using (2.7) and (2.8),

\[ P_b = \frac{1}{\pi} \int_{0}^{\pi/2} \left( 1 + \frac{10}{\sin^2 \phi} \right)^{-1} \left( 1 + \frac{10^{1.5}}{2 \sin^2 \phi} \right)^{-2} d\phi. \]

### 2.4 Radio Link Layer Issues in Wireless Networks

In a multiuser wireless network, users share the limited radio resources (e.g., spectrum and transmission power). Therefore, one critical issue is how to share these resources among users so that the optimal network performance can be achieved. Another issue is how to provide link reliability in presence of time-varying wireless propagation and interference conditions. The resource allocation for multiple access strategy and radio link control for error recovery are two main components of radio link layer protocols.

#### 2.4.1 Multiple Access and Scheduling Methods

Multiple access methods in wireless networks can be divided into the following two main groups: contention-free channel access and contention-based random channel access schemes. In contention-free schemes, multiple users are allocated with the radio resources (e.g., time slot, channel, and code) by a central entity [27].

**Contention-Free Channel Access and Channel Allocation for Circuit-Switched Services**

Contention-free channel access can be used in time-division, frequency-division, and code-division multiple access networks.

- **Time-division multiple access (TDMA):** In TDMA, time is divided into fixed-length frames, and each frame is divided into multiple time slots which are allocated to the users. In TDMA, synchronization among the nodes is required to avoid interference [28].

- **Frequency-division multiple access (FDMA):** In FDMA, radio frequency band is divided into multiple channels that are allocated to the users. Orthogonal frequency-division multiple access (OFDMA) is an improved version of FDMA that is based on the OFDM modulation in the physical layer. In OFDMA, frequency band is divided into multiple subcarriers that are shared among the users. OFDMA is used in the IEEE 802.16-based WiMAX networks.
2.4 Radio Link Layer Issues in Wireless Networks

- **Code-division multiple access (CDMA):** In CDMA, multiple nodes can transmit on the same radio channel simultaneously using spread spectrum modulation. The spreading codes for the different users are orthogonal/near-orthogonal to each other. For link reliability, the SINR should be maintained above a threshold.

- **SDMA (Space-Division Multiple Access):** In SDMA, the spectrum is shared among the users by exploiting their spatial distribution through the use of directional antennas that minimize the interference among the terminals.

The above techniques can be also combined to have hybrid channel access techniques (e.g., hybrid FDMA/TDMA, hybrid FDMA/CDMA, hybrid TDMA/CDMA).

Allocation of channels to the users should be made such that certain quality requirements (e.g., SINR) for their transmissions (e.g., uplink transmission to a BS or downlink transmission from a BS in a cellular network) are satisfied. The same radio channel can be used simultaneously by multiple user terminals (which are referred to as co-channel terminals) as long as the terminals are spaced sufficiently apart and/or the transmit power levels in these channels are adjusted. For cellular wireless systems, different channel allocation algorithms proposed in the literature aim at achieving a minimum SINR at each mobile terminal by adjusting the “co-channel reuse distance” and/or the transmit power. A comprehensive survey on these channel allocation algorithms can be found in [29].

Depending on how the co-channel terminals are spatially separated, channel allocation schemes can be categorized as follows: fixed channel allocation (FCA), dynamic channel allocation (DCA), and hybrid channel allocation (HCA) [29]. In FCA schemes, a number of channels are assigned to each cell according to some reuse pattern, and the assignment does not adapt to traffic conditions and/or user distributions in the cellular coverage region. However, the FCA schemes are the simplest channel allocation schemes. Let $C$ denote the total number of channels in the system, which is divided into $K$ groups of approximately equal size. Each cell is assigned a group of $\eta = \lfloor C/K \rfloor$ channels. For a hexagonal cell layout, if $D$ denotes the minimum distance between two co-channel cell centers (i.e., minimum channel reuse distance) and $R$ is the radius of a cell, it can be shown that $D/R = \sqrt{3}K$. Also, there exist fully symmetric channel allocations such that $K$ can be written in the following form:

$$K = i^2 + ij + j^2,$$

where $i, j = 0, 1, 2, 3, \ldots$. Therefore, the set of possible values of $K$ is $\{1, 3, 4, 7, 9, 12, 13, \ldots\}$. For a given $R$, as $K$ decreases, $D$ decreases, and therefore, a more aggressive channel reuse is achieved. However, since the interference will increase as well, for a given SIR requirement, a minimum value of $D$ will need to be maintained.

For interference avoidance in a multitier network, a spatial channel allocation method, called fractional frequency reuse (FFR) can be also used [30, Chapter 5]. The basic idea of FFR is to partition the macrocell service area into spatial regions and assign different frequency subbands to different sub-regions. Consequently, the cell edge-zone macrocell user equipments (MUEs) will not interfere with the center-zone MUEs, and with an efficient channel allocation method, the cell edge-zone MUEs will not interfere with neighboring cell edge-zone MUEs. Different variations of the FFR scheme include the
Wireless Networks and Resource Allocation

following: \textit{strict FFR}, \textit{soft FFR}, and \textit{sectored FFR}. A quantitative comparison among these different FFR schemes can be found in [30, Chapter 5].

In DCA schemes, all channels are kept in a central pool and assigned to the users on an on-demand basis so that SIR constraints are satisfied; therefore, these schemes are adaptive to traffic variability and propagation conditions. However, these schemes may involve considerable overhead (e.g., due to measurement of interference and traffic load), and also every user terminal or BS has to be capable of transmission/reception on all channels in the system. The DCA schemes can be implemented either by a central controller or distributively. In the latter case, the channel assignment can be made either by the BS in a cell (i.e., cell-based control) or by the user him- or herself autonomously based on signal strength measurements. In the cell-based distributed DCA schemes, the BS in each cell has to maintain information about the current channel usage in neighboring cells.

In HCA schemes, the available channels are divided into \textit{fixed} and \textit{dynamic} sets. The channels in the \textit{fixed} set are assigned to cells as in the FCA schemes, while those in the \textit{dynamic} set are shared by all users in the system (e.g., using DCA strategies). The ratio of fixed to dynamic channels can be varied based on traffic load variations. Also, HCA schemes can be enhanced with channel reordering (or channel handoffs between occupied channels) to improve channel utilization [29].

Trunking Theory and Performance Analysis of Contention-Free Channel Access

Trunking theory, which studies how a large population of users can be accommodated by a limited number of resources (e.g., channels), can be used to study the performance of a circuit-switched cellular wireless system (or a trunked radio system). With a given number of channels, when a particular user requests service and all of the radio channels are in use, the user is blocked (e.g., the blocked user could be cleared from the system or the blocked user could be delayed). The related performance measure is the probability that a call is blocked (in a \textit{blocked calls cleared system}), or the probability that a call experiences a delay greater than a certain queueing time (in a \textit{blocked calls delayed system}) due to the occupancy of all channels. These measures are also referred to as the grade of service (GoS) [61].

In a trunked radio system, the traffic intensity offered by each user $A_u$, expressed in the unit of Erlangs, is given by $A_u = \lambda \cdot H$, where $\lambda$ is the call request rate and $H$ is the average duration of a call. For $U$ users, the total traffic intensity $A$ is: $A = U A_u$ Erlangs. For a blocked calls cleared system (also referred to as an \textit{Erlang loss system}) with a total of $C$ channels, Erlang-B formula relates the probability of blockage to the number of channels $C$ and the total offered traffic $A$, which is given by

$$\Pr\{\text{blocking}\} = \frac{A^C / C!}{\sum_{k=0}^{C} A^k / k!}.$$  \hfill (2.26)

The above formula is derived based on an M/M/C/C queueing system under the following assumptions: call arrivals follow a Poisson distribution (with an average of $\lambda$), there are infinite number of users, arrival of requests is memory-less, the channel occupancy
time by a user is exponentially distributed (with an average of $\mu$), and there are a finite number of channels available in the trunking pool.

For a blocked calls delayed system (also referred to as an *Erlang delay system*), Erlang-C formula can be used to obtain the probability of blocking (i.e., probability that the delay $D$ is greater than 0) as follows:

$$
Pr\{D > 0\} = \frac{AC}{AC + C! \sum_{k=0}^{C-1} \frac{A^k}{k!}}.
$$

(2.27)

Subsequently, the probability that a call is delayed beyond a time $t$ can be obtained as follows:

$$
Pr\{D > t\} = Pr\{D > 0\} \cdot Pr\{D > t | D > 0\}
$$

(2.28)

where $Pr\{D > t | D > 0\} = e^{-\frac{A^t}{\mu}}$. That is, the distribution of delay time follows an exponential distribution. This is obtained based on the assumption that the services occur in the order of their arrival. The average delay $\bar{D}$ can then be calculated as follows:

$$
\bar{D} = Pr\{D > 0\} \cdot \frac{H}{C - A}.
$$

(2.29)

Note that the Erlang-C formula is derived based on an M/M/C/D queueing analysis, where $C$ is the maximum number of simultaneous users and $D$ is the maximum number of calls that may be held in the queue ($=\infty$ in this case). For an Erlang delay system with finite capacity $n$ (i.e., when the capacity of the queue for the waiting calls is finite and equal to $n$), where $n \geq C$, the probability of blocking (i.e., probability of delay and probability of loss) can be obtained as follows:

$$
Pr\{D > 0\} = \frac{(\lambda/\mu)^C}{C!} \times \frac{C}{C - \lambda/\mu} \left[ 1 - \left( \frac{\lambda/\mu}{C} \right)^{n+1} \right] p_0
$$

(2.30)

where

$$
\frac{1}{p_0} = \sum_{k=0}^{C-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^C}{C!} \times \frac{C}{C - \lambda/\mu} \left[ 1 - \left( \frac{\lambda/\mu}{C} \right)^{n+1} \right], \quad \text{and}
$$

(2.31)

$$
Pr\{\text{loss}\} = \frac{(\lambda/\mu)^n}{C!C^{n-C}p_0}.
$$

(2.32)

This system is referred to as the *combined delay and loss system*. Note that, for $n = C$, (2.32) gives the Erlang-B formula, and for $n \to \infty$, (2.30) gives the Erlang-C formula. Also, $1/p_0$ reduces to $\sum_{k=0}^{C-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^C}{C!} \frac{1}{1 - \frac{\lambda/\mu}{C}}$. Therefore, the Erlang loss system and Erlang delay system can be regarded as special cases of the combined delay and loss system.

Note that the analyses above do not consider the SINR requirements of the users. That is, even though a user is assigned a channel, it may not obtain the minimum required SINR, and therefore, SINR outage may occur. Therefore, given a cell radius, the reuse distance may need to be selected with a considerable margin.
Scheduling for Packet-Switched Wireless Services

For packet-switched wireless services based on multiple access methods such as TDMA, FDMA, OFDMA, CDMA, or SDMA, which user(s) will be served in a particular time slot, frequency band, spreading code, or directional antenna is (are) determined based on a scheduling method. In a cellular TDMA network, where the BS serves as the scheduling agent, the scheduling method specifies a rule to determine which user(s) is (are) allowed to transmit (e.g., in the uplink of a cellular network) or allowed to receive (e.g., in the downlink of a cellular network) during a time slot. The key requirements for the design and optimization of scheduling methods for wireless networks are as follows [32]:

- **Maximize network throughput**: Throughput refers to the amount of data successfully transmitted by the users over a time period. Maximizing the overall system throughput is a key objective of most of the multiple access schemes.
- **Minimize delay**: Delay refers to the time required for a packet to be transmitted successfully since it has been received at the transmission buffer from the upper layer. Delay is a key performance metric for real-time traffic (e.g., voice and video). For such traffic, multiple channel access schemes have to be designed to minimize delay.
- **Guarantee fairness**: Fairness is a measure of whether the users are receiving an equal (or fair) share of radio resources. Multiple access schemes must guarantee a certain level of fairness to all users in the network.
- **Improve power efficiency**: Power efficiency is an important performance metric for battery-powered wireless devices. There is a tradeoff between power efficiency and network performance. To reduce power consumption, a device can be put in standby mode during which the node cannot transmit and/or receive packets (and hence not considered by the scheduler). Consequently, the throughput for the corresponding device/user decreases.

**Opportunistic schedulers** (also called greedy schedulers) such as the max-SNR schedulers exploit channel conditions to achieve a high network throughput. Therefore, they exploit the multiuser diversity. Mathematically, if $\gamma$ denotes the SNR, the scheduler chooses a user $i^*$ during time slot $t$ such that

$$i^* = \arg \max_{1 \leq i \leq n} (\gamma_i(t)).$$  \hspace{1cm} \text{(2.33)}

A variation of the max-SNR scheduling is “max-SNR scheduling with threshold” in which the user with the largest SNR above a threshold is chosen. Mathematically,

$$i^* = \begin{cases} 
\arg \max_{1 \leq i \leq n} (\gamma_i(t)), & \text{if } \exists \gamma_i(t) > \gamma_{th}(t) \\
\text{rand}(\gamma_i(t)), & \text{if } \forall \gamma_i(t) < \gamma_{th}(t).
\end{cases}$$ \hspace{1cm} \text{(2.34)}

This reduces the amount of feedback from the users, because the channel state information is fed back to the scheduling agent only if SNR is above the threshold.
With opportunistic scheduling, since the best user with regard to the channel condition is selected, it can lead to starvation of some users and thus unfairness in the system. The Jain’s fairness index [14] defined below can be used to measure the fairness among users in sharing the bandwidth:

$$F(x_1, x_2, \ldots, x_n) = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}$$

(2.35)

where there are \( n \) users and \( x_i \) is the throughput for the \( i \)th user. The value of \( F \) varies from \( \frac{1}{n} \) (worst case) to 1 (best case). In the best case, all users receive the same allocation.

To achieve an absolute fairness among users with the same resource demands, the max-min fairness criterion can be used by the scheduler in which case all the users achieve the same throughput. In general, with max-min fair allocation among users with different demands, resources are allocated in order of increasing demand, no source receives a resource share larger than its demand, and sources with unsatisfied demands receive an equal share of the resource. This maximizes the minimum share of a source whose demand is not fully satisfied. As an example, for a set of four users with demands 1, 2, 4, 5 and a channel with bandwidth 10, the max-min fair allocation can be computed as follows. First, the bandwidth is divided into four portions of size 2.5. Since this is larger than user 1’s demand, this leaves 1.5 left over for the remaining three sources, which is divided evenly among the three other users giving them 3 (=2.5 + 0.5) units each. This is larger than user 2’s demand. Therefore, the excess of 1.0 is divided evenly among the remaining two users, giving them 3.5 each. Thus, with max-min allocation, user 1 receives 1, user 2 receives 2, users 3 and 4 receive 3.5 each. With an optimization-based problem formulation for resource allocation, max-min fairness among users with equal demand can be achieved by optimizing the scheduling method to maximize the minimum throughput.

The max-min fair approach can be generalized to have a more flexible resource sharing by associating weights \( w_1, w_2, \ldots, w_n \) with users 1, 2, \ldots, \( n \), which reflect the relative resource share of the different users. Accordingly, the concept of max-min weighted fair share allocation can be defined, where (i) resources are allocated in order of increasing demand, normalized by the weight, (ii) no user receives a resource share larger than its demand, and (iii) users with unsatisfied demands receive resource shares in proportion to their weights. Therefore, the weights can be normalized so that the smallest weight is 1 (e.g., \( 1, w_2', w_3', \ldots, w_n' \)) and the resource (e.g., channel bandwidth) can be divided into \( 1 + w_2' + w_3' + \cdots + w_n' \) shares, and in each round of resource allocation, a user receives a share of resource proportional to its weight.

The proportional fairness (PF) is a popular criterion for user scheduling that takes advantage of multiuser diversity while maintaining comparable long-term throughput for all users. With PF scheduling, starvation of users is avoided. PF scheduling has been implemented in the CDMA 1xEV-DO system. If \( R_i(t) \) denotes the instantaneous data rate of user \( i \) at time slot \( t \) and \( C_i(t) \) is the average throughput for user \( i \) up to time slot
the PF scheduler will choose the user with the highest \( \frac{R_i(t)}{C_i(t)} \) at time slot \( t \):

\[
i^* = \arg \max_{1 \leq i \leq n} \left( \frac{R_i(t)}{C_i(t)} \right).
\]  

The average rate \( C_i(t) \) for all users is then updated as follows:

\[
C_i(t + 1) = \begin{cases} 
(1 - \frac{1}{t_c}) C_i(t) + \frac{1}{t_c} R_i(t), & i = i^* \\
(1 - \frac{1}{t_c}) C_i(t), & i \neq i^*
\end{cases}
\]  

(2.37)

where \( \frac{1}{t_c} \) can be viewed as the timescale over which the scheduler aims to provide proportionally fair bandwidth allocation, and it is related to the maximum time for which a user can be starved. Typical values of \( t_c \) lie between 10 and 20.

For resource allocation, we would choose a suitable objective function for the underlying optimization problem for user scheduling that could balance good overall network performance as well as fairness among the users. One such objective function, which is parameterized by a parameter \( \kappa \), is as follows [34]:

\[
U(C_1, C_2, \ldots, C_n) = \sum_{i=1}^{n} f_{\kappa}(C_i)
\]  

(2.38)

where \( f_{\kappa}(x) \) is the utility function for one user, which is given by

\[
f_{\kappa}(x) = \begin{cases} 
\ln(x), & \text{if } \kappa = 1 \\
l(x^{1/\kappa}), & \text{otherwise}
\end{cases}
\]  

(2.39)

Depending on parameter \( \kappa \), this general objective function can achieve different types of fairness. Specifically, for \( \kappa = 0 \), total throughput is maximized, while \( \kappa = 1 \) provides proportional fairness among different users [35], \( \kappa = 2 \) achieves harmonic mean fairness, and \( \kappa \to \infty \) provides max-min fairness. In general, the higher the value of \( \kappa \), the fairer the solution of the underlying optimization problem is.

The scheduling methods may also need to consider the users’ QoS requirements (e.g., minimum delay requirement, minimum throughput requirement). Accordingly, QoS-aware scheduling methods need to be designed that take into account both the channel conditions and the QoS factors. The maximum-largest weighted delay first (M-LWDF) and exponential/proportional fair (EXP/PF) are examples of such QoS-aware scheduling methods. The M-LWDF scheduler considers a probabilistic delay requirement as follows:

\[
\Pr(d_i > T_{i,\text{deadline}}) \leq \delta_i
\]  

(2.40)

where \( d_i \) is the waiting time of the head of line (HOL) packet in the queue of user \( i \), \( T_{i,\text{deadline}} \) is the delay deadline for a packet of user \( i \), and \( \delta_i \) is the target threshold for the probability of exceeding the delay deadline. Accordingly, the scheduler chooses a user for transmission as follows [37]:

\[
i^* = \arg \max_{1 \leq i \leq n} \left( \frac{a_i R_i(t) d_i(t)}{C_i(t)} \right)
\]  

(2.41)
where \( d_i(t) \) is the waiting time of the HOL packet in the queue of user \( i \) in slot \( t \), \( R_i(t) \) is the instantaneous transmission rate of user \( i \) in slot \( t \), \( C_i(t) \) is the average transmission rate of user \( i \) up to time slot \( t \), and \( a_i = \frac{-\log \delta_i}{T_i \text{deadline}} \).

The EXP-PF scheduler uses the following scheduling criterion [38]:

\[
i^* = \arg \max_{1 \leq i \leq n} \left( \exp \left( \frac{a_i D_i(t) - aD(t)}{1 + \sqrt{aD(t)}} \right) \frac{R_i(t)}{C_i(t)} \right)
\] (2.42)

where \( aD(t) = \frac{1}{n} \sum_{i=1}^{n} a_i D_i(t) \).

### Contention-Based Channel Access

For contention-based random access scheme, a user has to compete with other users to transmit over the wireless channel. A packet transmitted by a node will be received successfully if there is no collision. A collision occurs when multiple users transmit data simultaneously and the SINR at the receiver is lower than the minimum SINR required to decode the original packet correctly. If collision occurs, a user may attempt to retransmit the packet, and the specifics of the retransmission method depend on the protocol used. The most common contention-based channel access schemes are as follows [31]:

- **Pure ALOHA** and **slotted ALOHA (S-ALOHA)**: In pure ALOHA, if a node has a packet to send, it will attempt to transmit the packet immediately. If the packet collides with packets from other nodes, the node will retransmit the packet later. With slotted ALOHA (S-ALOHA), time is divided into slots, and packet transmissions are aligned with the time slots. It is known that S-ALOHA offers a better network throughput compared to the pure ALOHA protocol.

Assuming that successful transmission occurs only when there is only one transmission attempt during a time slot\(^2\) and ignoring the effects of channel fading and interference, performance of the S-ALOHA protocol can be analyzed as follows. A finite population S-ALOHA system with \( N \) users can be modeled by using a Markov chain [15] (Figure 2.7). The system moves around a finite number of states where in state \( i \), \( i \) users are backlogged. Let \( p \) denote the probability of a new packet generation, \( q \) denote the probability of retransmission, \( n \) denote total number of new packets generated by the \( N - i \) unblocked users during the current slot \((0 \leq n \leq N - i)\), \( r \) denote the number of retransmissions attempted by the \( k \) backlogged users during the same time interval \((0 \leq r \leq i)\), \( \text{Pr}\{n = 0\} \) denote the probability that no new packet is generated during the current slot, and \( \text{Pr}\{r \geq 1\} \) denote the probability that one or more of the backlogged stations attempt a retransmission during the current slot.

\(^2\) This is referred to as **protocol model of interference** as opposed to the **SINR model** in which case successful transmission can occur in presence of multiple simultaneous transmissions as long as the SINR at the receiver for that corresponding transmission is above a given threshold.
The state transition probabilities are given as follows:

\[ p_{i,j} = 0, \quad j \leq i - 2 \]
\[ p_{i,i-1} = \Pr\{n = 0\}\Pr\{r = 1\} \]
\[ p_{i,i} = \Pr\{n = 0\}\Pr\{r \neq 1\} + \Pr\{n = 1\}\Pr\{r = 0\} \]
\[ p_{i,i+1} = \Pr\{n = 1\}\Pr\{r \geq 1\} \]
\[ p_{i,j} = \Pr\{n = j - i\}, \quad j \geq i + 2, \quad \text{where} \]
\[ \Pr\{n = 0\} = (1 - p)^{N-i} \]
\[ \Pr\{n = 1\} = (N - i)p(1 - p)^{N-i-1} \]
\[ \Pr\{n = j - i\} = \binom{N - i}{j - i} p^{j-i}(1 - p)^{N-j} \]
\[ \Pr\{r = 0\} = (1 - q)^i \]
\[ \Pr\{r = 1\} = \binom{i}{j} q(1 - q)^{i-1} \]
\[ \Pr\{r \neq 1\} = 1 - \Pr\{r = 1\} = 1 - iq(1 - q)^{i-1} \]
\[ \Pr\{r \geq 1\} = 1 - (1 - q)^i. \]

The system starts in state 0 where all users are unblocked. As time proceeds, it approaches equilibrium. The equilibrium probability can be obtained by solving the following equation: \( \pi_k = \sum_{i=0}^{N} \pi_i p_{ik}, \quad k = 0, \ldots, N \) subject to \( \sum_k \pi_k = 1 \). Then the mean backlog can be obtained as \( B = \sum_{k=0}^{N} k \pi_k \) and the mean throughput \( S \) can be obtained as: \( S = \sum_{k=0}^{N} \pi_k f_k \), where \( f_k \) is the throughput in state \( k \) given as \( f_k = \Pr\{n = 1\}\Pr\{r = 0\} + \Pr\{n = 0\}\Pr\{r = 1\} \). Therefore, the mean delay \( D \) can be obtained as the ratio of mean backlog and mean throughput as follows: \( D = \frac{\sum_{k=0}^{N} k \pi_k}{\sum_{k=0}^{N} f_k \pi_k} \).

Based on the Markov chain model, the stability of the S-ALOHA protocol can be analyzed. Specifically, the stability is analyzed based on the expected drift \( d_k \) in each state, defined as the difference between the expected input and the expected output traffic in state \( k \):

\[ d_k = (N - k)p - f_k. \]
For stability, the network is expected to work in a state near an equilibrium point, i.e., where the expected drift crosses zero with negative derivative. A network in equilibrium should have the channel throughput equal to the newly generated traffic:

\[ S = (N - B)p. \]

- **Packet Reservation Multiple Access (PRMA):** It is a time-division multiplexing-based MAC protocol that allows transmission of packet voice and low bit rate data in different time slots from different users over a common channel. Time slots are grouped as frames, and each time slot is designed for a single packet transmission. The frame duration is selected such that a voice user can generate one packet per frame. In a frame, the individual slots are accessed by the mobile users for communication with the BS. A feedback signal from the BS to the mobile users concerning the previous transmitted packet is multiplexed along with the data stream from the BS. During the call setup process, a particular mobile user is given a reservation in a slot that is at the same position in succeeding frames. When a voice user has completed its communication, it halts transmission, so the BS receives a null packet and the time slot in the frame is unreserved and left available for use by the other mobile users. In the case of congestion in the network due to many mobile users, speech packets are given priority and data packets are dropped.

- **Carrier Sense Multiple Access (CSMA):** With CSMA, a user senses the status of the channel before attempting transmission. If the channel is idle, the user attempts a transmission based on a particular algorithm common to all transmitters in the system. If the transmission is unsuccessful due to a collision, the user waits for a packet retransmission interval and transmits again. The different variants of the CSMA protocol are as follows:

  1. **1-persistent CSMA:** If the channel is sensed to be busy, the user waits until it becomes idle. When the user detects an idle channel, it transmits a packet with probability 1. If a collision occurs, the user waits a random amount of time and starts all over again.

     It is possible that just after a user begins transmission, another user becomes ready to send and senses the channel. If the first user’s signal has not yet reached the second one, the latter will sense an idle channel and will also begin transmitting, resulting in a collision. The longer the propagation delay is, the more significant this effect becomes and the worse will be the performance of the protocol. Note that even if the propagation delay is 0, there can still be collisions among users. If two users become ready in the middle of a third user’s transmission, both will wait until the transmission ends, and then both will begin transmitting simultaneously resulting in a collision.

  2. **Non-persistent CSMA:** With non-persistent CSMA, if the channel is sensed to be in use, the user does not continually sense it. Instead, it waits a random period of time and then repeats the algorithm.

  3. **p-persistent CSMA:** This protocol applies to a time-slotted system where channel time is divided into contiguous slots and the duration of each time slot is equal...
to the maximum propagation delay between two users. In this system, packet transmission starts at the beginning of a time slot. When a user becomes ready to send, it senses the channel. If it is idle, it transmits with probability $p$ (i.e., it defers until the next slot with probability $q = 1 - p$).

If the next slot is also idle, it either transmits or defers again with probabilities $p$ and $q$. The process repeats until either the packet has been transmitted or another user has begun transmitting. In the latter case, it acts as if there had been a collision (i.e., it waits a random time and starts again). If the user initially senses the channel busy, it waits until the next slot and applies the above algorithm.

For a non-persistent unslotted CSMA protocol, if generation of new and retransmitted packets follows a Poisson distribution with rate $\lambda$ packets/sec, or equivalently, $G = \lambda \times t_p$ packets per packet transmission time, where $t_p$ is the packet transmission time, and the maximum propagation delay between two users is $\delta$, the average channel utilization, $S$, can be obtained as follows [47]:

$$S = \frac{\text{Average length of the time the channel is used without collision}}{\text{Average length of busy period + Average length of idle period}}$$

$$= \frac{Ge^{-aG}}{G(1 + 2a) + e^{-aG}}$$

(2.44)

where $a = \frac{\delta}{t_p}$.

For slotted non-persistent CSMA, the average channel utilization can be obtained as follows [47]:

$$S = \frac{aGe^{-aG}}{1 + a - e^{-aG}}.$$  

(2.45)

The performance of a CSMA network can drop due to the hidden node problem. This happens when users become hidden from each other (e.g., by hills, buildings, etc.), and consequently, a user may think that the channel is idle when actually a hidden user is transmitting. An improved variant of CSMA is CSMA with collision avoidance (CSMA/CA). In CSMA/CA, a handshaking process (referred to as RTS/CTS signalling) between the transmitter and receiver is performed before actual data transmission to mitigate the hidden terminal problem.

IEEE 802.11 MAC

The 802.11 MAC architecture is referred to as the distributed foundation wireless MAC (DFWMAC), which provides a distributed access control mechanism with an optional centralized control built on top of that. The lower sublayer of the MAC architecture is the distributed coordination function (DCF), which uses a simple CSMA algorithm. DCF is used by ordinary asynchronous traffic. The point coordination function (PCF) is a centralized MAC algorithm, which is built on top of the DCF and exploits features of DCF, and can be used to provide contention-free channel access.

DCF describes two techniques for packet transmission: the basic access (two-way handshaking technique) and the RTS/CTS access (four-way handshaking technique).
The RTS/CTS scheme is useful to combat the hidden terminal problem. The basic DCF mechanism works as follows. If a user has a MAC frame to transmit, it listens to the channel. If the channel is idle, it waits to see if the channel remains idle for a time equal to IFS (interframe space). If so, the station may transmit immediately. Note that there are different values of IFS as follows: SIFS (short IFS), PIFS (point coordination function IFS), and DIFS (distributed coordination function IFS), and SIFS < PIFS < DIFS. These different values of IFS are used to provide priority-based access to different types of frames.

If the channel is busy (either because the user initially finds the channel busy or because the channel becomes busy during the IFS idle time), the user defers transmission and continues to monitor the channel until the current transmission is over. Once the current transmission is over, the user delays another IFS. If the channel remains idle for this period, then the user backs off a random amount of time and senses the channel again. If the channel is still idle, the station may transmit. During the backoff time, if the channel becomes busy, the backoff timer is halted, and it resumes when the channel becomes idle.

DCF uses a discrete-time backoff scale [16]. The time immediately following an idle DIFS is slotted, and a station is allowed to transmit only at the beginning of each slot time. The slot time $\sigma$ is equal to the time needed for any user to detect the transmission of a packet from any other station. DCF uses a technique known as binary exponential backoff (BEB). At each packet transmission, the backoff time is uniformly chosen in the range $(0, w - 1)$, $w$ is the contention window, which depends on the number of transmissions failed for the packet. At the first transmission attempt, $w$ is set to $CW_{\text{min}}$, the minimum contention window. After each unsuccessful transmission, $w$ is doubled, up to a maximum value $CW_{\text{max}} = 2^m CW_{\text{min}}$. The backoff time counter is decremented as long as the channel is sensed idle, “frozen” when a transmission is detected on the channel, and reactivated when the channel is sensed idle again for more than a DIFS. The user transmits when the backoff time reaches zero.

Also, unless there is fragmentation of an upper layer data unit to multiple MAC layer frames, to avoid channel capture, a user must wait a random backoff time between two consecutive new packet transmissions, even if the medium is sensed idle in the DIFS time. In case of fragmentation, the different fragments are transmitted in sequence, with only an SIFS between them, so that only the first fragment must contend for the channel access.

An acknowledgment (ACK) is transmitted by the receiver to signal the successful packet reception. The ACK is immediately transmitted after an SIFS period of time. Since the SIFS plus the propagation delay is shorter than a DIFS, no other station is able to detect the channel idle for a DIFS until the end of the ACK. If the transmitter does not receive the ACK within a specified timeout period (ACK_Timeout), or it detects transmission on the channel, it reschedules the packet transmission according to the given backoff rules.

With the RTS/CTS scheme, a user who intends to transmit a packet waits until the channel is sensed idle for a DIFS, follows the given backoff rules, and then transmits a request to send (RTS) frame. When the receiver detects an RTS frame, it responds after
an SIFS with a clear to send (CTS) frame. The transmitter is allowed to transmit only if the CTS frame is correctly received.

The saturation throughput (i.e., system throughput in overload conditions) performance of the DCF scheme with the exponential backoff protocol details was analyzed in [16] for both the access mechanisms and for any combination of the two methods under ideal channel conditions and finite number of users. The transmission queue of each user is assumed to be always non-empty. Let $n$ denote the fixed number of contending stations and $\tau$ denote the stationary probability that a station transmits a packet in a generic (i.e., randomly chosen) slot time. Also, let $W = CW_{\text{min}}$ and the maximum backoff stage $m$ be such that $CW_{\text{max}} = 2^m W$. Also, let $W_i = 2^i W$, where $i \in (0, m)$ is the “backoff stage.” Assuming that at each transmission attempt, regardless of the number of retransmissions, each packet is assumed to collide with a constant and independent probability $p_c$, the probability $\tau$ that a station transmits in a randomly chosen slot time can be obtained as [16]

$$
\tau = \frac{2(1 - 2p_c)}{(1 - 2p_c)(W + 1) + p_c W (1 - (2p_c)^m)}.
$$

The function $\tau(p_c)$ is continuous in the range $p_c \in (0, 1)$. Also, this is a monotone decreasing function that starts from $\tau(0) = \frac{2}{W+1}$ and reduces up to $\tau(1) = \frac{2}{1+2^m W}$. Now, $p_c = 1 - (1 - \tau)^{n-1} \rightarrow \tau^*(p_c) = 1 - (1 - p_c)^{\frac{m}{n}}$. Therefore, $\tau^*(p_c)$ is a continuous and monotone increasing function in the range $p_c \in (0, 1)$ with $\tau^*(0) = 0$ and $\tau^*(1) = 1$. Therefore, a unique solution for $\tau$ can be obtained. The normalized system throughput $S$ is then defined as follows:

$$
S = \frac{E[\text{Payload information transmitted in a slot time}]}{E[\text{Length of a slot time}]} (2.47)
$$

which can be expressed as [16]

$$
S = \frac{p_{tr}p_s E[L]}{(1 - p_{tr})\sigma + p_{tr}p_s T_s + p_{tr}(1 - p_s)T_c}
= \frac{E[L]}{T_s + \sigma \frac{1-p_c}{p_{tr}} + T_c \left( \frac{1}{p_s} - 1 \right)} (2.48)
$$

where $E[L]$ is the average size of payload, $p_{tr}$ is the probability that there is at least one transmission in the considered slot time ($= 1 - (1 - \tau)^n$), and $p_s$ is the probability that transmission in the channel is successful given that at least one station transmits ($= \frac{n(1-\tau^{n-1})}{1-(1-\tau)^n}$). $T_s$ is the average time the channel is sensed busy because of a successful transmission, $T_c$ is the average time the channel is sensed busy by each station during a collision, and $\sigma$ is the duration of an empty slot time ($E[L]$, $T_s$, $T_c$, $\sigma$ are all expressed in the same unit). The denominator denotes the average amount of time spent on the channel in order to observe the successful transmission of a packet payload.

In contrast to the DCF mechanism, the PCF mechanism uses polling by a centralized point coordinator that makes use of PIFS when issuing polls. Since PIFS < DIFS, the point coordinator can seize the channel and lock out all asynchronous traffic while it issues polls and receives responses. When a poll is issued, the polled user may respond
using SIFS. If the point coordinator receives a response, it issues another poll using PIFS. If no response is received during the expected turn around time, the coordinator issues another poll.

To prevent the point coordinator from locking out completely all asynchronous traffic, an interval known as the superframe is defined. During the first part of this interval, the point coordinator issues polls in a round-robin fashion to all users configured for polling. The point coordinator then idles for the remainder of the superframe, allowing a contention period for asynchronous access.

### Packet Radio Multiple Access in Multihop Wireless Networks

For a multihop packet radio network, slotted ALOHA-type multiple access strategies with adaptive transmission range control were modeled and analyzed in [36]. The authors considered a system model where the nodes are distributed as a two-dimensional Poisson point process, and at any time each node may either transmit or receive, but not both. Two or more transmissions to the same receiving node in the same slot result in a collision, and the transmissions are considered to be unsuccessful. For a node to transmit in a slot, it must be in transmit mode (with a given probability) and able to find a receiver in the forward direction within its maximum transmission range denoted by $R$. Node $B$ is said to be node $A$’s neighbor if $B$ is within a distance $R$ from $A$. Node $B$ is said to be in transmitter $A$’s forward direction if non-negative progress is produced when node $B$ is chosen as transmitter $A$’s receiver. Otherwise, node $B$ is said to be in $A$’s backward direction. Node $A$ interferes with node $B$ in a slot if node $A$ transmits while node $B$ is receiving a packet and $B$ is within node $A$’s transmission range. A successful transmission from $A$ to $B$ will occur if $B$ does not transmit and $B$’s neighbors do not interfere with $B$. If a transmitter is unable to find a receiver from which forward progress will result, then it will not transmit in that slot.

The three following transmission strategies were analyzed [36]:

1. **MFR (Most Forward with Fixed Radius):** With an objective to minimizing the number of hops needed for a packet to reach its destination, in this strategy, a node transmits to a neighbor with the largest forward progress using a transmission radius $R$, which is the maximum transmission radius.

2. **NFP (Nearest with Forward Progress):** With an objective to reducing collision as much as possible, in this strategy, a node transmits to the nearest neighbor so that a forward progress can be made. Also, it uses adjustable transmission power just strong enough to reach the receiving node.

3. **MVR (Most Forward with Variable Radius):** This strategy aims at reducing collision to some extent while maintaining the goal of obtaining the largest progress possible. This is similar to the MFR strategy except that the transmission radius is adjusted to be equal to the distance between the transmitter and the receiver.

For the above strategies, two performance metrics were considered: one-hop (or local) throughput and normalized (with respect to average node density) average progress. The one-hop throughput $S$ at node $A$ was evaluated as follows: $S = \Pr(A+) \times \Pr(A\text{ transmits}) \times \Pr(T_{AB})$, where $\Pr(A+)$ denotes the probability that node $A$ can find
a receiving node in the forward direction and $T_{AB}$ denotes the event that transmission from $A$ to $B$ is successful given that $A$ transmits to $B$.

Numerical results showed that MVR is always better than MFR in terms of normalized average progress and one-hop throughput $S$, and NFP has the highest throughput of all. Also, NFP has the best normalized average progress when the average number of neighbors of a node (i.e., average number of nodes within the transmission radius $R$) is greater than 8. For average number of neighbors less than 8, NFP and MVR achieve almost same normalized average progress. Also, for NFP, the normalized average progress and one-hop throughput are fairly constant as the number of neighbor nodes becomes large (e.g., $>7$).

**Propagation Effects on Packet Radio Multiple Access**

Many works in the literature have focused on characterizing the propagation effects (e.g., effects of statistical propagation mechanisms such as fading, shadowing, and path loss) on radio packet multiple access protocols [39]–[46]. These works also consider different spatial distributions for the users in the network. To describe the spatial distributions of users in the network, the following models are generally used.

*Uniform distribution*: If users of the network are distributed according to a two-dimensional Poisson point process (PPP) with density $\lambda$, then the probability of finding $i$ users in an area of size $A$ is given by

$$\Pr (i \text{ users in } A) = \frac{(\lambda A)^i \exp(-\lambda A)}{i!}, \quad i \geq 0. \tag{2.49}$$

Assuming that the receiver (e.g., tagged user, BS or AP) is at the center of the network or cell with transmitters located in a circle with radius $R$, the PDF of the distance of the user $r$ from the receiver is given as

$$f(r) = \frac{2r}{R^2}, \quad 0 < r \leq R. \tag{2.50}$$

If the range is normalized, then (2.50) becomes

$$f(r) = \begin{cases} 2r, & 0 < r \leq 1 \\ 0, & r > 1. \end{cases} \tag{2.51}$$

*Log-normal spatial distribution*: The log-normal spatial distribution is given by [40]

$$f(r) = \frac{\beta}{\sqrt{2\pi\sigma_d r}} \exp\left\{ -\frac{\beta^2 \ln r^2}{2\sigma_d^2} \right\} \tag{2.52}$$

where $\sigma_d$ is the spatial logarithmic variance. This distribution allows more realistic modeling of the traffic very close to the receiver.

*Bell shape distribution*: If users are distributed according to a bell-shaped spatial density function from the BS with path-loss exponent $\beta = 4$, then the corresponding PDF is [44]

$$f(r) = 2re^{-\pi r^4/4}, \quad r \geq 0. \tag{2.53}$$
This original model of S-ALOHA is pessimistic for wireless communication environments since the received signal for a common receiver arrive with different power levels. Since packets competing to access a common receiver arrive from different distances and with independent fading levels, not necessarily all colliding packets will be destroyed. In this case, a packet may be able to *capture* the receiver (e.g., BS) in the presence of $k$ interfering packets, if the SINR at the receiver as defined in (2.54) below, where $P_{Ri}$ is the received power from user $i$, $P_N$ is the background/thermal noise power, is higher than a predetermined threshold during a certain section of the time slot. This threshold is known as the *capture ratio* $z_0$, and the certain section of a time slot is referred to as *capture window* $t_w$. Therefore, the probability of packet success can be given as in (2.55):

$$\text{SINR} = \frac{P_{R0}}{P_N + \sum_{i=1}^{k} P_{Ri}} \quad (2.54)$$

$$P_s(k) = \Pr[\text{SINR} > z_0]. \quad (2.55)$$

Suppose that $X$ sends a packet (test packet) to $Y$ in a given time slot and the received power at $Y$ is $P_{R0}$. Also assume that $Y$ does not transmit in the same time slot and there are $k$ interfering packets. The probability of successful reception of the test packet is given by

$$P_s(k) = \Pr \left[ P_{R0} > z_0 \sum_{i=1}^{k} P_{Ri} \right] \quad (2.56)$$

where the effect of noise is neglected. The capture ratio and the capture window are influenced by the particular modulation and coding technique used. The probability that a packet is successfully received or captured even when there is a number of packets transmitted at the same time is significant when the near-far effect and the random fluctuation in the received signal power due to fading and shadowing are taken into account. Fading and shadowing can improve the overall network performance [39]. In particular, in the presence of fading, the throughput is higher, and the heavier the shadowing the better the performance. They also have the effect of enhancing the “fairness” of the protocol, giving far users some chance to be captured.

Assume that the number of packets (new and retransmissions) generated follows the Poisson distribution with mean generation rate $\lambda$ packets per second. The mean offered channel traffic in packets per time slot is therefore given by [39]

$$G = \lambda \tau$$

where $\tau$ is the slot length. The probability that the test packet is overlapped by $k$ packets is

$$R_k = \frac{G^k e^{-G}}{k!}.$$
Using the capture model defined before, the probability of being able to capture the receiver in an arbitrary time slot given the capture ratio \( z_0 \) is

\[
P_{\text{cap}}(z_0) = 1 - \sum_{k=1}^{\infty} R_k (1 - P_s(k))
\]

\[
= 1 - \sum_{k=1}^{\infty} R_k Pr\left(\frac{P_{R0}}{P_k} < z_0\right)
\]

\[
= 1 - \sum_{k=1}^{\infty} R_k F_{Z_k}(z_0)
\]

where \( R_k \) is the probability that the test packet is overlapped by \( k \) packets, \( P_k = \sum_{i=1}^{k} P_{Ri} \), and \( F_{Z_k}(z) \) is the cumulative distribution function of \( Z_k \), the random signal-to-interference ratio (SIR) when noise is neglected, which is given by

\[
Z_k = \frac{P_{R0}}{P_k}, \quad 0 \leq Z_k < \infty.
\]

Note that, under Rayleigh fading, since the instantaneous received power \( p \) of the signal follows an exponential distribution about the local-mean power \( \bar{P}_s \), the PDF of \( P_{R0} \) can be obtained from

\[
f_{P_{R0}}(p) = \frac{1}{\bar{P}_s} \exp\left(-\frac{p}{\bar{P}_s}\right), \quad p \geq 0.
\]

With (2.57), the channel throughput can be calculated as the product of the offered load in a time slot and the probability of a successful transmission in a time slot as follows:

\[
S = GP_{\text{cap}}(z_0).
\]

If all packets reach the receiver with uncorrelated Rayleigh fading and equal mean power \( \bar{P}_0 \) (Ring Model), the PDF of the joint interference power \( P_k \) under coherent addition can be approximated by [39]

\[
f_{P_k}(p_k) = \frac{1}{k\bar{P}_0} \exp\left(-\frac{p_k}{k\bar{P}_0}\right)
\]

where the mean interference power of \( k \) interfering packets is \( k\bar{P}_0 \). The PDF of \( P_k \) under non-coherent addition can be given as the \( k \)-fold convolution of the Rayleigh distribution resulting in the gamma distribution as follows:

\[
f_{P_k}(p_k) = \frac{1}{\bar{P}_0} \left(\frac{p_k}{\bar{P}_0}\right)^{k-1} \exp\left(-\frac{p_k}{\bar{P}_0}\right) \frac{1}{(k-1)!}.
\]

The CDF of the random variable \( Z_k \) can be calculated as

\[
F_{Z_k}(z_0) = \int_0^{z_0} dz \int_0^{\infty} f_{P_{R0}}(zw) f_{P_k}(w) dw
\]

using (2.60) and (2.61) for the PDF of \( P_k \), for coherent and non-coherent additions, respectively.
2.4 Radio Link Layer Issues in Wireless Networks

Without receiver capture, the throughput of S-ALOHA is given by

\[
S = Ge^{-G}. \tag{2.63}
\]

Using the results above, the throughput with receiver capture under coherent and non-coherent addition of interference signals can be obtained, respectively, as follows [39]:

\[
S_{\text{coherent}} = G \exp \left( -G \frac{z_0}{z_0 + 1} \right) \tag{2.64}
\]

\[
S_{\text{non-coherent}} = Ge^{-G} \sum_{n=0}^{\infty} \frac{G^n}{n!} \frac{1}{kz_0 + 1}. \tag{2.65}
\]

Therefore, it is obvious that the throughput of S-ALOHA improves under Rayleigh fading. Also, coherent addition of interfering packets gives a higher throughput since \( P_k \) in (2.60) has higher probabilities for lower values when compared to \( P_k \) in (2.61).

The coherent-interference model can be extended by assuming that packets arrive with different mean power (e.g., when mobiles are at different distances from the BS). In this case, we need a PDF for the mean received packet power, i.e., \( f_{P_s}(\bar{p}_s) \), and a PDF for the mean interference power of \( k \) packets, \( f_{P_k}(\bar{p}_k) \), which is the convolution of \( k \) \( P_s \) random variables. As a result, the PDF of \( P_{R0} \) is given by

\[
f_{P_{R0}}(p_s) = \int_0^\infty \frac{1}{\bar{p}_s} \exp \left( \frac{p_s}{\bar{p}_s} \right) f_{P_s}(\bar{p}_s) d\bar{p}_s \tag{2.66}
\]

and the PDF of \( P_k \) is given by

\[
f_{P_k}(p_k) = \int_0^\infty \frac{1}{\bar{p}_k} \exp \left( \frac{p_k}{\bar{p}_k} \right) f_{P_k}(\bar{p}_k) d\bar{p}_k. \tag{2.67}
\]

To obtain these two PDFs, we first need to evaluate \( f_{P_s}(\bar{p}_s) \), which depends on the spatial distribution of the traffic offered to the channel. The mean received power of a packet at a distance \( r \) is \( P_s = Kr^{-\eta} \) when only path loss is considered (\( \eta = \) path-loss exponent). The normalized received power can be written as \( \bar{P}_s = \rho^{-\eta} \), where \( \rho = r/R \), \( R \) is the radius of the circular area around the common receiver, and the users are located in the unit circle \( 0 \leq \rho < 1 \). Therefore, the PDF of \( \bar{P}_s \) becomes

\[
f_{\bar{P}_s}(\bar{p}_s) = f_\rho(\rho) \left| \frac{d\rho}{d\bar{p}_s} \right| \tag{2.68}
\]

where \( f_\rho(\rho) \) is the PDF that a packet is generated within distance \( \rho \) and is given by

\[
f_\rho(\rho) = \frac{2\pi}{G_t} G(\rho) \rho \tag{2.69}
\]

with \( G(\rho) \) the spatial traffic density, i.e., the number of packets per time slot offered per unit area at a normalized distance \( \rho \), and \( G_t \) is the total traffic offered to the receiver. Then, equation (2.68) becomes

\[
f_{\bar{P}_s}(\bar{p}_s) = \frac{2\pi}{\eta G_t} \bar{p}_s^{-(1+2/\eta)} G(\bar{p}_s^{-1/\eta}). \tag{2.70}
\]
Equations (2.66) and (2.67) can be calculated given a traffic density $G(\rho)$. Then plugging this result into (2.62) to obtain the SIR distribution, the throughput can be calculated.

The analysis above considers Rayleigh fading only. Throughput performance of S-ALOHA under Rician fading and capture effect can be found in [46]. Different propagation models are adopted for the intended user and $k$ interferers. The signals for the desired (or test) packet experience Rician fading while the interfering signals may experience either uncorrelated Rician fading or uncorrelated Rayleigh fading. The channel model corresponding to Rician fading of the useful signal and Rayleigh fading of $k$ interfering signals applies within a large cell where the test user is near the BS and the interfering users are far away from the BS. The channel model corresponding to Rician fading of the useful signal as well as $k$ interfering signals (Rice + $k$ Rice model) correspond to a situation where the test user and the interfering users are within a small cell or an indoor environment. To compute throughput $S$ as given by (2.59), the CDF of the capture probability $Z_k$ given $k$ interfering users needs to be computed for the specific channel model using (2.62).

When the interfering signals experience Rayleigh fading, the PDF of joint signal powers is the gamma distribution given by (2.61) with $P_0 = \sigma^2$ being the average power of the interfering signal. Recalling that the PDF of the instantaneous received power under Rician fading is given by (2.71) [46],

\[
f_P(p) = \frac{1}{\sigma^2} e^{\left[-\frac{2p + s^2}{2\sigma^2}\right]} I_0 \left[\frac{\sqrt{2ps}}{\sigma^2}\right]
\]  

(2.71)

the PDF for the sum of $k$ interfering signals is given by the convolution of (2.71) $k$ times,

\[
f_{P_k}(p_k) = \frac{1}{2\sigma^2} \left(\frac{p_k}{P}\right)^{(k-1)/2} \exp\left(-\frac{p_k + P}{2\sigma^2}\right) I_{k-1} \left[\frac{\sqrt{Pp_k}}{\sigma^2}\right]
\]  

(2.72)

where $P = \sum_{i=1}^{k} s_i$, $s_i$ is the peak value of LoS interfering received signal, and $P_k$ is the joint interfering power.

With Rician fading, a higher throughput can be achieved [46]. As expected, the throughput decreases as $z_0$ increases, and it increases as the Rice factor $K$ increases. For the Rice + $k$ Rice model, for a fixed $z_0$ and fixed value of Rician parameter $K_d$ for the desired user, the throughput decreases as the Rician parameter $K_u$ for the interfering users increases.

The PDF of the instantaneous received power $P_s$ of a received packet in a channel with path loss, lognormal shadowing, and Rayleigh fading can be characterized by (2.73) [40] as follows:

\[
f_{P_s}(p_s) = \int_0^\infty \int_0^\infty \frac{1}{\overline{P}_s} \exp\left(\frac{p_s}{\overline{P}_s}\right) \sqrt{2\pi} \sigma_s \overline{P}_s \exp\left\{-\frac{(\ln \overline{P}_s + \eta \ln r)^2}{2\sigma_s^2}\right\} dr d\overline{P}_s
\]  

(2.73)

where $f(r)$ is the PDF for the separation distance $r$ between the transmitter and the receiver and $\eta$ is the path-loss exponent (e.g., $\eta$ has a value in the range of 3 to 4). The
Laplace transform of (2.73) is
\[
\Phi_{P_s}(v) = \int_0^\infty \exp(-v x) f_{P_s}(x) \, dx
\]
\[
= \int_0^\infty \int_0^\infty \frac{1}{1 + \frac{v}{\sqrt{2\pi \sigma_s}}} \frac{f(r)}{\sqrt{2\pi \sigma_s}} \exp \left\{ -\frac{(\ln P_s + \eta \ln r)^2}{2\sigma_s^2} \right\} \, dr \, dP_s. \quad (2.74)
\]
During the capture of a packet, the instantaneous received power \(P_s\) is assumed constant. Hence, the probability of capture, given the average mean power of the test packet \(\bar{P}_0\), and \(k\) interfering packets is
\[
P_{\text{capt}}(k, \bar{P}_0) = \Phi_{P_s}(z_0/\bar{P}_0)^k. \quad (2.75)
\]
With incoherent addition of the \(k\) interfering signals, the joint PDF \(f_{P_s}(x)\) as said before is the \(k\)-fold convolution of the PDF of the individual signal powers, and in Laplace transformation this corresponds to the product of individual characteristic functions:
\[
P_{\text{capt}}(k, \bar{P}_0) = \Phi_{P_s}(z_0/\bar{P}_0)^k. \quad (2.76)
\]
This probability must be averaged over the PDF due to shadowing and the near-far effects to make it unconditional to them, and \(P_{\text{capt}}(k)\) is given as
\[
P_{\text{capt}}(k) = \int_0^\infty \int_0^\infty \frac{f(r)}{\sqrt{2\pi \sigma_s} \bar{P}_0} \exp \left(-\frac{xz_0}{\bar{P}_0} \right) f_{P_s}(x) \, dx \, dy \]
\[
= \int_0^\infty \exp \left(-\frac{xz_0}{\bar{P}_0} \right) f_{P_s}(x) \, dx
\]
\[
= \Phi_{P_s}(z_0/\bar{P}_0). \quad (2.77)
\]
The probability that one out of \(k\) packets transmitted during a time slot (\(k - 1\) interfering packets) captures the base station is given as
\[
q_k = kP_{\text{capt}}(k - 1) \quad (2.78)
\]
assuming that the probability that two or more packets simultaneously capture the receiver is zero.

Considering a uniform spatial density of the offered traffic (2.51), the probability of capture was evaluated for various propagation models [40]. It was observed that the effect of Rayleigh fading, path loss, and shadowing leads to the highest probability of capture. Also, the expected drift for each backlog state (2.43) for various propagation models was observed. When shadowing or near-far effects are combined with Rayleigh fading, the backlog clears faster (and hence the average delay decreases). The authors also compared the expected drift when uniform spatial distribution (2.51) and log-normal spatial distribution (2.52) are considered.

The capture effects for a WLAN with a finite population of users in the presence of Rayleigh fading, shadowing, and near-far effect (i.e., path loss) are investigated in [42]. The following three types of CSMA/CA protocols are analyzed:
- Basic CSMA/CA: data packet is used only for a packet transmission.
- Stop-and-wait automatic repeat request (SW ARQ) CSMA/CA: successful reception of each packet is acknowledged.
- Four-way handshake (4-WH) CSMA/CA: uses request to send (RTS) and clear to send (CTS) packets prior to the transmission of the actual data to avoid the hidden terminal problem. A collision can occur only during the RTS packet transmission period.

This work calculates the capture probability of an access point (AP) of the channel, derives the throughput and packet delay under the three propagation mechanisms, and compares the performance with a non-fading channel by using a mathematical method based on renewal theory. A uniform spatial distribution is considered in which the terminals are uniformly distributed within a unit radius around the AP. It is observed that the performance in the fading channel model is worse than an error free channel model. On the other hand, as the traffic load increases, the throughput in the fading channel model also increases. The throughput in an error free channel model decreases starting from a specific point in the traffic load. This is due to the advantage offered by the capture effect.

**Propagation Effects on Uplink Multiple Access in Cellular CDMA Networks**

As has been mentioned previously, in a CDMA system, all users use the same radio spectrum simultaneously. Therefore, these systems are in general interference-limited. The amount of interference experienced at a receiver depends on the radio propagation environment. In the following, we will demonstrate how to obtain the SINR for an uplink transmission in a cellular CDMA scenario.

Consider an uplink transmission based on CDMA from a user to its serving BS. Let $P_r$, $E_b$, $I_0$, $P_N$, and $G_p$ denote the received power, the received energy per information bit, one-sided interference-plus-noise power spectral density, background noise power, and processing gain, respectively. The received power $P_r$ is the same for all users in a cell. Note that $R_b$ is the transmission bit rate, and $W$ is the bandwidth of the transmission channel, then $G_p = \frac{W}{R_b}$. Since $P_r = R_b E_b$ and interference-plus-noise power $I = W I_0$,

$$\text{SINR} = \frac{R_b E_b}{W I_0} = \frac{1}{G_p} \frac{E_b}{I_0}. \quad (2.79)$$

Therefore, the bit energy to interference-plus-noise spectral density ratio $E_b/I_0$ can be written as

$$\frac{E_b}{I_0} = G_p \times \text{SINR}. \quad (2.80)$$

If $K$ denotes the total number of users in a cell, for a single cell system (i.e., no intercell interference), the total interference-plus-noise power at the BS is given by
\((K - 1)P_r + P_N\). Therefore,

\[
\text{SINR} = \frac{P_r}{(K - 1)P_r + P_N}.
\]

For a multicell system, if \(\kappa\) denotes the total intracell and intercell interference normalized to the total intracell interference [48],

\[
\text{SINR} = \frac{P_r}{(K - 1)P_r\kappa + P_N} = \frac{\nu_f}{(K - 1) + \nu_fP_N/P_r}
\]

where \(\nu_f\) is termed as frequency reuse efficiency factor (\(\frac{1}{\kappa}\)). Note that \(\nu_f = 1\) for a single cell system and \(\nu_f < 1\) for a multicell system due to intercell interference. Again, if the source activity is taken into account, then

\[
\text{SINR} = \frac{\nu_f}{(K - 1) \xi + \nu_fP_N/P_r} \tag{2.81}
\]

where \(\xi\) is the “source activity factor” (e.g., \(\xi = 3/8\) for voice sources).

The work in [49] analyzes the impact of interfering users along with propagation effects on uplink SINR in a cellular CDMA network considering a hexagonal cell geometry and propagation loss and shadow-fading parameters to be the same in all the cells. Consider uplink transmission for a tagged user in a typical cell \(S_0\) with base station \(BS_0\). The region outside this cell is denoted by \(S_0^c\). Users in \(S_0^c\) transmitting to their own BSs cause interference to \(BS_0\).

Using the notations used in [5], consider an interfering mobile located at point \((x, y)\) at a distance \(d_1\) from its BS in cell 1 and at a distance \(d_0\) from \(BS_0\). Then its transmitted power \(P_1 = P_r(d_1)\eta^{-10^{z_1}/10}\), where \(\eta\) is the path-loss exponent, \(z_1\) (in dB) is a Gaussian-distributed random variable corresponding to shadow fading, with standard deviation \(\sigma\) and its probability density function is given by \(f_Z(z_1) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{z_1^2}{2\sigma^2})\). The interference power received at \(BS_0\) is \(P_1d_0^{-\eta}10^{z_0/10} = P_r(d_0)^{-\eta}10^{(z_0-z_1)/10}\).

With cell radius \(R\) and given \(K\) users per cell, since traffic density \(\rho = \frac{2K}{3\sqrt{3}R^2}\) users/m², the number of users in a differential area \(dA(x, y)\) centred at point \((x, y)\) is \(\rho dA(x, y) = \frac{2KdA(x, y)}{3\sqrt{3}R^2}\). Then, integrating over all \(S_0^c\) and then taking the expectation over the random variables, the total average interference power at \(BS_0\) due to mobiles outside \(S_0\) is given by [5, 49]

\[
I_{S_0} = \frac{2K}{3\sqrt{3}R^2}P_rE \left( \int_{S_0^c} \int \left[ \left( \frac{d_1}{d_0} \right)^{\eta} \right] 10^{(z_0-z_1)/10} dA(x, y) \right)
\]

\[
= \frac{2K}{3\sqrt{3}R^2}P_rE \int_{S_0^c} \int \left( \frac{d_1(x, y)}{d_0(x, y)} \right)^{\eta} dA(x, y) \tag{2.82}
\]

where the random variables (RVs) \(z_0\) and \(z_1\) represent power variations for transmissions from mobiles in the vicinity of point \((x, y)\), which are due to shadow fading as measured at \(BS_0\) and \(BS_1\), respectively. With \(\eta = 4\), numerical evaluation of the double integral, including the term \(\frac{2}{3\sqrt{3}R^2}\), provides a value of 0.44.
To evaluate the expectation term in (2.82), express $z_0$ and $z_1$ in the following form:

$$z_i = ah + bh_i, \quad i = 0, 1, \quad a^2 + b^2 = 1$$  \hspace{1cm} (2.83)

where $RV h$ represents the shadow fading for transmissions in the vicinity of $(x, y)$ and $h_i$ represents the shadow fading due to the independent propagation conditions along the two paths toward $BS_0$ and $BS_1$. The RVs $h, h_0,$ and $h_1$ are individually Gaussian distributed and independent for which the following relations hold true:

$$
\mathbb{E}(z_i) = 0 = \mathbb{E}(h) = \mathbb{E}(h_i) \\
\mathbb{E}(z_i^2) = \mathbb{E}(h^2) = \mathbb{E}(h_i^2) = \sigma^2 \\
\mathbb{E}(hh_i) = 0 = \mathbb{E}(h_0h_1).
$$

Let $e^y = 10^{(z_0-z_1)/10}$ so that $y = 0.23(z_0 - z_1)$ is Gaussian distributed with zero average value and variance $\sigma_y^2 = (0.23)^2 \times 2b^2 \times \sigma^2 = 0.053\sigma^2$, if $b^2 = 1/2$. Therefore,

$$
\mathbb{E}\left[10^{(z_0-z_1)/10}\right] = \mathbb{E}(e^y) = \int_{-\infty}^{\infty} e^y \frac{e^{-y^2/2\sigma_y^2}}{\sqrt{2\pi}\sigma_y} dy = e^{\sigma_y^2/2}. 
$$  \hspace{1cm} (2.84)

For $\sigma = 8$ dB, $\mathbb{E}(e^y) = 5.42$. Therefore, $I_{SS} = P_r K \times 0.44 \times 5.42 = 2.38P_rK$. For $\eta = 4$ and $\sigma = 8$ dB, total interference power $= (K - 1)P_r + 2.38P_rK = (3.38K - 1)P_r$. Therefore, using (2.80),

$$
E_b/I_0 = \frac{W}{R_b} \times \frac{P_r}{(3.38K - 1)P_r} = \frac{W}{R_b(3.38K - 1)}. \hspace{1cm} (2.85)
$$

Therefore, given a target $E_b/I_0$ requirement, the system capacity (in terms of number of circuit switched connections served simultaneously in a cell) can be evaluated.

**Interference Modeling and Multihop CDMA Packet Radio Network**

In a non-orthogonal spectrum sharing system such as a spread spectrum CDMA system, a receiver experiences interference from each active transmitter and the packet success probability depends on the SINR at the receiver. The interference power at the receiver can be modeled as a random variable that is equal to the sum of the interference powers from all transmitters. The probability of packet success would depend on the selection of transmission range, a measure of which could be the average number of user terminals that are closer to the receiver than the transmitter [50]. If there are fewer transmitters near the receiver, the higher is the chance of packet success, and vice versa. Note that the powers of the interferers will vary according to their distances from the receiver and the propagation conditions (e.g., shadowing and fading). For a multihop network, the probability of packet success increases as the link distance decreases. Note that the link distance is determined by the maximum transmission power of the transmitter and the receiver sensitivity. However, with a small link distance, the number of hops that the packet must traverse increases. If the objective is to maximize the expected forward progress per transmission (i.e., probability of packet success times link distance), then there is an optimum link distance (or transmission range). For a given range $R$, with an
average of \( \lambda \) terminals per unit area, the average number of terminals in this range is 
\[ N = \lambda \pi R^2. \]
If \( p_s \) denotes the probability of packet success, then the expected forward 
progress per slot \( Z = p_s \times R \) increases with an increase in the packet success probability 
and decreases as the number of hops increases (i.e., \( R \) decreases). Since 
\[ R = \sqrt{\frac{N}{\pi \lambda}}, \]
the transmission range can be optimized by optimizing \( N \).

In [50], the authors derive the statistics of received interference power at a terminal 
for a class of signal propagation laws and show that the optimum transmission range 
is such that on the average, the number of terminals closer to the transmitter than the 
receiver is proportional to the square root of the processing gain. A time-slotted sys-
tem is considered and the positions of terminals are modeled as a spatial Poisson point 
process with parameter \( \lambda \):

\[
\Pr\{k \text{ terminals in area } A\} = \frac{e^{-\lambda A} (\lambda A)^k}{k!} \tag{2.86}
\]

where \( \lambda \) is the average number of terminals per unit area. Interference experienced at a 
receiver is assumed to be constant over a packet transmission time and independent from 
slot to slot. Assuming a constant background noise with power spectral density \( N_0/2 \), 
if \( P_0 \) is the received signal power, and \( P_i \) denotes the interference power at the receiver 
from \( i \)th interferer, in case of DS/BPSK with rectangular chip pulses, the average symbol 
energy-to-noise ratio at the detector is given by [50]

\[
\mu = \frac{E_b}{N_0^{(\text{eff})}} = \left( \frac{2I}{3G_pP_0} + \frac{1}{\mu_0}\right)^{-1} \tag{2.87}
\]

where \( N_0^{(\text{eff})}/2 \) is the equivalent white noise power spectral density, \( G_p \) is the processing 
gain, \( I = \sum_{i=1}^{k} P_i \), and \( \mu_0 = E_b/N_0 \), and \( \mu_0 \) is the SNR at the receiver detector in the 
absence of interferers. For a given \( \mu \), assuming that symbol errors are independent and 
t-error correcting block code of length \( n \) is used, the probability of packet success \( p_s(\mu) \) 
is given by

\[
p_s(\mu) = \sum_{i=0}^{t} \binom{n}{i} \left( \frac{1}{2} \text{erfc}(\sqrt{\mu}) \right)^i \left( 1 - \frac{1}{2} \text{erfc}(\sqrt{\mu}) \right)^{n-i} \tag{2.88}
\]

where \( \frac{1}{2} \text{erfc}(\sqrt{\mu}) \) denotes the probability of symbol error in which \( \text{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} e^{-x^2} dx \). If \( f_\mu(.) \) and \( F_\mu(.) \) denote the PDF and the probability distribution function of \( \mu \), respectively, the unconditional probability of packet success can be expressed as

\[
p_s = \int_0^{\infty} p_s(x) f_\mu(x) \, dx = \int_0^{\infty} [1 - F_\mu(x)] p_s'(x) \, dx.
\]
In general, for \( 0 < \alpha < 1 \),

\[
f_I(y; \alpha) = \frac{1}{\pi y} \sum_{k=1}^{\infty} \frac{\Gamma(\alpha k + 1)}{k!} \left( \frac{\rho}{y^\alpha} \right)^k \sin k\pi (1 - \alpha) \quad (2.94)
\]
2.4 Radio Link Layer Issues in Wireless Networks

where \( \rho = \pi \lambda \Gamma(1 - \alpha) \), and

\[
F_I(y; \alpha) = \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\Gamma(\alpha k)}{k!} \left( \frac{\rho}{y^{\alpha}} \right)^k \sin k\pi (1 - \alpha).
\] (2.95)

Assuming that the distance between the transmitter and the receiver is \( R \) and \( g(r) = \frac{1}{r^4} \), the signal power \( P_0 \) is \( R^{-4} \). Using the relationship \( \mu = \left( \frac{2 I_3 G_p P_0}{3 G_p} + \frac{1}{\mu_0} \right)^{-1} \), \( F_\mu(\mu) \) can be obtained as

\[
F_\mu(\mu) = \begin{cases} 
1 - \text{erfc} \left( \frac{\rho \lambda \pi R^2}{2 \sqrt{\pi K(\mu)}} \right), & \mu < \mu_0 \\
0, & \mu > \mu_0
\end{cases}
\] (2.96)

where

\[
K(\mu) = \frac{3 G_p}{2} \left( \frac{1}{\mu} - \frac{1}{\mu_0} \right).
\] (2.97)

Now, using the relationship \( p_s = \int_0^\infty [1 - F_\mu(x)] p'_s(x) \, dx \)

\[
p_s = \int_0^{\mu_0} \text{erfc} \left( \frac{\rho \lambda \pi R^2}{2 \sqrt{\pi K(\mu)}} \right) p'_s(\mu) \, d\mu.
\] (2.98)

For a network with uniform traffic and “balanced” routing, the probability of packet success calculated above gives the local throughput, which will be the same for all user terminals. Based on the \( p_s \), for a multihop network, the expected forward progress \( Z = p_s \times R = p_s \times \sqrt{\frac{N}{\pi \kappa}} \) can be optimized so that a desired value of \( N \) can be found given \( G_p \) and \( \mu_0 \).

2.4.2 Error Control Methods

As has been mentioned before, the data link layer is also responsible to cope with the data transmission errors in a wireless link. For detecting errors in the transmitted radio link layer frames, channel coding is used. Channel coding is the process of adding redundancy to the information in order to detect and possibly correct the errors at the receiver. As an example, \( n - k \) parity symbols are added to \( k \) information symbols, and the \( n \) symbol codeword is transmitted over a noisy channel. At the receiver, the syndrome (in the case of an error detecting linear block code) is computed [17]. If the syndrome is zero, an acknowledgment (ACK) is generated. If the syndrome is non-zero, the receiver sends a negative acknowledgment (NACK).

After errors are detected, in case they need to be corrected, the two following methods can be used: ARQ (Automatic Repeat Request) and FEC (Forward Error Correction). With ARQ, the transmitter retransmits the erroneous frame. An ARQ system offers high system reliability, and it is simple to implement. However, the throughput decreases with increasing channel error rate. As the round-trip delay increases, ARQ becomes more difficult to implement. Also, the delay is not constant. ARQ is not suitable for systems without a feedback channel, systems with delay jitter constraints and maximum delay.
constraints, systems with large round-trip delays and/or buffer constraints, and systems with high channel error rates. A system implementing ARQ requires to have means of error detection, acknowledgment upon packet reception, timeout mechanism, and retransmission.

**FEC Schemes**

With FEC, redundancy added to the information is exploited at the receiver to detect and correct errors in the transmitted bits. The main features of FEC as an error control method are as follows: it provides a constant throughput regardless of the channel condition, it incurs a fixed delay equal to the processing time for encoding/decoding, and it requires more encoding overhead than ARQ to achieve the same reliability. Also, it is hard to achieve high system reliability, and decoding is hard to implement and expensive.

FEC coding techniques can be divided into the following two techniques: block and convolutional error correction techniques. In an \((n, k)\) linear block code, a sequence of \(k\) information bits is used to obtain \(n - k\) parity bits, resulting in an encoded block of \(n\) bits. Linear codes form a linear vector space, and any two code words can be added (modulo 2) to produce a third code word. The Hamming distance between two code words \(c_1\) and \(c_2\), \(d(c_1, c_2)\) is the number of positions in which they differ (for example, if \(c_1 = (11010100)\) and \(c_2 = (11100010)\), \(d(c_1, c_2) = 4\)). An \((n, k)\) block code is capable of detecting \(d_{\text{min}} - 1\) errors, where \(d_{\text{min}}\) is the minimum separation between a pair of code words. In general, a code with minimum distance \(d_{\text{min}}\) can detect \(e_d\) errors and correct \(e_c\) errors, where

\[
e_d + e_c \leq d_{\text{min}} - 1 \quad (2.99)
\]

and

\[
e_c \leq e_d. \quad (2.100)
\]

For the reliability of an \((n, k)\) error detecting linear block code, let \(P_e\) denote the probability that the received codeword is correct, \(P_d\) denote the probability that the received codeword contains a detectable error, and \(P_e\) denote the probability of an undetected error. Therefore, \(P_c + P_d + P_e = 1\). Then the probability that there are undetected errors in the packet is

\[
P(E) = P_e + P_d P_e + P_d^2 P_e + \cdots
\]

\[
= P_e \cdot \frac{1}{1 - P_d}
\]

\[
\approx P_e / P_c. \quad (2.101)
\]

For random errors, noting that, with binary symmetric channel with transition probability \(\epsilon\), \(P_c = (1 - \epsilon)^n\) and linear block codes exist with \(P_e\) satisfying the following bound [17]:

\[
P_e \leq \{1 - (1 - \epsilon)^k\} 2^{-(n-k)} \quad (2.102)
\]
the following observations can be made: (1) For low channel error rates, \( P(E) \) is constant; (2) increasing the code size decreases \( P(E) \) for high error rates; (3) increasing the number of parity bits by 10 reduces \( P(E) \) by a factor of about 1000.

Bose-Chaudhuri-Hocquenghem (BCH) code and Reed-Solomon (RS) codes are popular block codes [17]. Binary BCH codes can be constructed with the following parameters:

\[
\begin{align*}
n &= 2^m - 1 \\
n - k &\leq mt \\
d_{\text{min}} &= 2t + 1
\end{align*}
\]

(2.103)

where \( m \) (\( m \geq 3 \)) and \( t \) are arbitrary positive integers.

RS codes are non-binary block codes. For a non-binary block code, the elements of a fixed-length code word are selected from an alphabet of \( q \) symbols, denoted by \( \{0, 1, 2, \ldots, q - 1\} \). Usually, \( q = 2^k \), so that \( k \) information bits are mapped into one of the \( q \) symbols. If \( N \) denotes the length of the non-binary codeword, \( L \) denotes the number of information symbols in a codeword, and \( D_{\text{min}} \) is the minimum distance, the RS codes can be described as follows:

\[
\begin{align*}
N &= q - 1 = 2^k - 1 \\
K &= 1, 2, 3, \ldots, N - 1 \\
D_{\text{min}} &= N - L + 1 \\
R_c &= L/N.
\end{align*}
\]

(2.104)

Such a code can correct up to

\[
t = \left\lfloor \frac{1}{2} (D_{\text{min}} - 1) \right\rfloor = \left\lfloor \frac{1}{2} (N - L) \right\rfloor
\]

(2.105)

symbol errors.

With convolutional codes, redundant symbols are generated as a function of a span of preceding information symbols. Convolutional coding with Viterbi decoding is one of the most popular FEC techniques. A convolutional code of rate \( 1/v \) may be generated by a \( K \) stage shift register and \( v \) modulo-2 adders. For each input information bit, the output of the modulo-2 adders provides \( v \) channel bits. The constraint length of the code is defined as the number of shifts over which a single information bit can influence the encoder output. For binary convolutional codes, the constraint length is equal to \( K \), the length of the shift register.

One important performance metric for FEC codes is coding gain, which is defined as the difference in values of \( E_b/N_0 \) required to attain a particular error rate for a transmitted message with and without coding. Let us illustrate this by a numerical example. Let us assume that information from a source is organized in 36-bit messages that are to be transmitted over an AWGN channel using non-coherently detected BFSK modulation. If no error control coding is used, the required bit error rate (BER) \( p_b \) to produce a message
error probability of $10^{-3}$, is obtained as $p_m^u = 1 - (1 - p_b)^{36} \approx 36p_b = 10^{-3} \Rightarrow p_b = 2.87 \times 10^{-5}$. The $E_b/N_o$ required to obtain this BER is obtained as: $2.87 \times 10^{-5} = \frac{1}{2} e^{-\frac{E_b}{N_o}} \Rightarrow E_b/N_o = -2 \ln(2 \times 2.87 \times 10^{-5}) = 19.6 = 12.92$ dB. Now, if we consider the use of a (127, 36) linear block code with $d_{\text{min}} = 31$ in the transmission of these messages, the BER required corresponding to the same message error probability can be calculated as $pc_m \approx \left(\frac{127}{16}\right)p_c^{11} (1 - pc)^{11} = 10^{-3} \Rightarrow pc \approx 0.0546$. The corresponding bit energy to noise power spectra density is obtained as $0.0546 = \frac{1}{2} e^{-\frac{Ec}{No}} \Rightarrow Ec/No = 19.6 = 12.92$ dB. Therefore, $E_b/N_o = \frac{127}{36} \cdot \frac{Ec}{No} = 15.63 = 11.94$ dB. Hence, the coding gain is $12.92 - 11.94 = 0.98$ dB.

**ARQ Schemes**

The following performance measures are used for ARQ schemes: reliability (which is the expected number of delivered packets containing no error) and channel utilization (which is the ratio of the average number of information bits successfully transmitted and the maximum number of bits that could have been transmitted). The three common types of ARQ protocols are as follows.

**Stop-and-Wait (SW) ARQ:** In this method, the transmitter sends the data and then waits for either an ACK message in the case of error-free reception, or a NACK message in the case of erroneous reception. Whenever the transmitter receives a NACK, it will retransmit the corresponding packet that is incorrectly received by the receiver, and when the transmitter receives an ACK it will transmit the next packet. Therefore, the expected number of transmissions for a packet is

$$T_{SW} = P_c + 2P_c(1 - P_c) + \cdots + lP_c(1 - P_c)^{l-1} + \cdots$$

$$= 1/P_c.$$  \hfill (2.106)

If the total number of bits that could be transmitted in one period is $n + \lambda \delta$, where $\lambda$ is the idle period and $\delta$ is the bit rate of the transmitter, the channel utilization $\eta_{SW} = \frac{k}{ts_{SW}(n + \lambda \delta)} = \frac{P_c}{1 + \lambda \delta / n} \cdot \left(\frac{2}{k}\right) = \frac{P_c}{1 + \lambda / r} \cdot r$, where $r = \text{code rate} = k/n$. Therefore, as $\lambda$ and/or $\delta$ increase, $\eta$ decreases. At lower bit error rates $\epsilon$, longer codes perform better; on the other hand, at higher bit error rates, shorter codes perform better. Even at very low channel error rates, SW ARQ can provide a poor throughput.

In an ARQ system, a large $n$ minimizes the time wasted in ACKs and associated delays, while a smaller $n$ minimizes $P(E)$ and the time wasted in retransmissions. Therefore, the block length can be optimized to maximize the throughput.

Based on the relationship $\eta_{SW} = \frac{P_c r}{1 + \lambda \delta / n}$, $\frac{d\eta_{SW}}{dn} = 0$ gives

$$n_o = \frac{-\lambda \delta \pm \sqrt{\frac{\lambda \delta}{\lambda \delta \ln(1 - \epsilon)}}}{2}$$

$$= \left\{ \frac{-\lambda \delta}{2} \right\} \sqrt{1 - \frac{4}{\lambda \delta \ln(1 - \epsilon)}} - 1.$$  \hfill (2.107)

With optimum block length, significant throughput improvement is observed at high bit error rates in the wireless channel.
For an SW ARQ system with $m$ duplicate transmissions, the total number of bits transmitted per actually transmitted packet is given by:

$$n + \lambda \delta + (mn + \lambda \delta) \cdot (1 - P_c) + \frac{(mn + \lambda \delta) + (1 - (1 - P_c)^m)}{1 - (1 - P_c)^m} + \frac{(mn + \lambda \delta) + 2 \cdot (mn + \lambda \delta) + (n + \lambda \delta + 3 \cdot (mn + \lambda \delta) + (1 - P_c) + (1 - P_c) + (1 - P_c)^m + \cdots = (n + \lambda \delta) + (mn + \lambda \delta) \cdot \frac{1 - P_c}{1 - (1 - P_c)^m} \cdot \frac{1 - P_c}{1 - (1 - P_c)^m}.}

$$

The average number of transmissions per packet is therefore

$$T_{SW-m} = 1 + \frac{mn + \lambda \delta}{n + \lambda \delta},$$

(2.108)

Go-Back-N (GBN) ARQ: In this method, the transmitter sends a number of packets without waiting for an ACK from the receiver. The receiver, however, acts similarly to the case of stop and wait. Generally, this method uses a sliding window algorithm with a window size of $N$ at the transmitter and a window size of 1 at the receiver. In GBN, the receiver sends cumulative ACKs corresponding to the in order packet with the highest sequence number. If a packet is lost, the sender has to retransmit every sent packet with a higher sequence number than the lost one.

The transmitter sends $N$ packets without waiting for an ACK. The continuous transmission stops only when a repeat request is received at the transmitter. In response, the transmitter transmits the requested packet, and the following $N - 1$ packets (regardless of whether they were received correctly). $N$ is selected such that it is equal to the round trip delay. The average number of transmissions required per packet is given by

$$T_{GBN} = P_c + (N + 1)P_c(1 - P_c) + (2N + 1)P_c(1 - P_c)^2 + \cdots = 1 + \frac{N(1 - P_c)}{P_c}. \quad (2.109)$$

Therefore, channel utilization $\eta_{GBN} = \frac{k}{n(1 + \frac{N(1 - P_c)}{P_c})} = \frac{P_c}{P_c + N(1 - P_c)} \cdot \frac{k}{n}$. The channel utilization improves significantly (compared to SW ARQ) for small $\epsilon$. Also, the shorter the code length $n$, the better the throughput performance is. However, as $N$ increases, $\eta$ decreases. GBN ARQ eliminates the idle time inherent in SW ARQ. However, it is more complex than SW ARQ, and it requires buffering.

Selective Repeat (SR) ARQ: In SR ARQ, the sender transmits a certain number of packets in succession. However, unlike GBN ARQ, the receiver sends ACKs for each packet with the corresponding sequence number. The receiver also accepts out of order packets and keeps them in the receiver buffer. In case of packet loss, unlike GBN ARQ, the transmitter retransmits only the individual packets that have timed out.

With SR ARQ, the average number of transmissions per packet is given by

$$T_{SR} = P_c + 2P_c(1 - P_c) + 3P_c(1 - P_c)^2 + \cdots = \frac{1}{P_c}. \quad (2.110)$$
Therefore, channel utilization $\eta_{SR} = P_e \cdot \frac{1}{\epsilon}$. The channel utilization improves for shorter code lengths especially at higher $\epsilon$. It is the most efficient among all ARQ strategies, and the channel utilization is not a function of the round-trip delay. Even with a finite buffer constraint, SR ARQ provides a much improved channel utilization over GBN. On the other hand, it is a complex ARQ method to implement, and its reliability may suffer depending on how buffer overflow is controlled.

**Hybrid ARQ Schemes**

A hybrid ARQ system combines FEC and ARQ to overcome the drawbacks of each of these schemes [17]. It provides a lower reliability but a higher throughput than for ARQ alone, and a higher reliability but a lower throughput than those for FEC alone. It is more complex than either FEC or ARQ alone. The basic structure of the hybrid ARQ error control scheme consists of an FEC subsystem contained in an ARQ system. The FEC system reduces the number of retransmissions by correcting the error patterns that occur most frequently. When a less frequent error pattern is detected, the receiver requests a retransmission.

There are two classes of hybrid ARQ error control schemes: type-I and type-II hybrid ARQ schemes. With type-I hybrid ARQ, when a received codeword is detected in error, if the number of errors is within the designed error correcting capability of the code, the errors will be corrected. However, if an uncorrectable error pattern is detected, the receiver rejects the received codeword and requests a retransmission. This technique is suitable for links with predictable noise and interference so that the designed error correcting capability is often adequate to correct the errors. Note that with type-I hybrid ARQ, the extra parity-check digits for error correction must be included in each transmission regardless of the channel state. On the other hand, with type-II hybrid ARQ, parity-check digits are sent to the receiver only when needed.

Type-II hybrid ARQ protocols adapt to the changing channel conditions through the use of incremental redundancy. When the receiver detects a received codeword in error, it saves the erroneous codeword in a buffer and requests a retransmission. Rather than retransmitting the original codeword, a block of parity check digits (formed from the original codeword) are transmitted. This block of parity-check digits are used to correct errors in the original transmission. If this is not successful, a second retransmission is requested. For the second retransmission (and any subsequent retransmissions), either the original codeword or the parity bits may be retransmitted, depending on the error correction/detection code and the strategy employed. If after the first, second, and third decoding operations, detectable errors still persist, either the first or second transmission is discarded. Type-II hybrid ARQ scheme is better suited for noisy channels with non-stationary channel error rates.

Let us give a specific example on the implementation of a type-II hybrid ARQ scheme, which is referred to as the Lin-Yu’s type-II hybrid ARQ. In this scheme, two codes are used: one is a high rate $(n, k)$ code $c_1$, which is designed for error detection only, and the other is a half-rate invertible $(2n, n)$ code $c_2$, which is designed for simultaneous error correction and detection. A $k$-bit message is first encoded using $c_1$
to form an \( n \)-bit codeword \( P_1 \). \( P_1 \) is then encoded using \( c_2 \). The \( n \) parity bits (denote by \( P_2 \)) from the \( c_2 \) codeword are saved in a buffer, while the \( c_1 \) codeword \( P_1 \) is transmitted. The initial packet is checked for errors at the receiver. If it is found to contain errors, a retransmission request is sent back to the transmitter. The transmitter responds by sending \( P_2 \). Since \( c_2 \) is invertible, the \( n \) bits used to create the codeword \( c_2 \) can be obtained by inverting \( P_2 \). An inverted version of \( P_2 \) is created and checked for errors. If the inverted version contains errors, \( P_2 \) is appended to \( P_1 \) to create a noise-corrupted \( c_2 \) codeword. After FEC decoding, the resulting message is checked once again for errors. If there are still errors, the process continues, with the transmitter alternating transmission of \( P_1 \) and \( P_2 \) until a successful transmission is achieved.

### 2.4.3 Power Control Methods

In a wireless network, power control methods are required for transmitters for two reasons. First, for effective communication, since the received power must be such that the SINR is above a given threshold, in the presence of channel attenuation as well as noise and interference, the transmitted power has to be adapted accordingly. Second, to minimize energy consumption and thereby extend the battery life (e.g., in mobile devices) or reduce utility cost (e.g., in BSs) as well as minimize co-channel interference, power control is required. Also, note that reducing the transmission power at the mobile devices/BSs can alleviate health/environmental concerns. For systems such as the spread-spectrum (SS) systems using CDMA, since all users use the same spectrum band, the received power level from all mobile units at the BS needs to be equalized to mitigate the so called near-far effect.

Power control methods (e.g., in cellular uplinks and downlinks) can be broadly categorized as open loop power control and closed loop power control methods. The former method is based on the usage of link quality measurements in one link (e.g., uplink or downlink) to adapt power in the other link (e.g., downlink or uplink). The latter method uses a dedicated control channel in one link to send control commands for power adaptation in the other link. In practical systems, power control methods consist of outer loop power control and inner loop power control methods that achieve the dual objectives of setting of a target SIR and ensuring the above set target SIR point, respectively. In most cases outer loop power control is based on an open loop strategy, while inner loop power control is based on a closed loop strategy.

### Open Loop Power Control

In case of open loop power control, the transmitter does not use any feedback from the receiver and hence no dedicated control channel is required. For example, in a cellular network, the uplink transmit power can be set based on the channel-gain in the downlink direction, which can be obtained simply by measuring the received power of a downlink pilot channel whose transmit power is known. However, in frequency division duplex (FDD) systems, where the uplink and the downlink operate at different frequencies, the channel-gain in the uplink may be different from that in the downlink. Therefore, this method may not be highly accurate. Consequently, open loop power control is only
useful to follow long-term channel variations, and it cannot be used to compensate fast fading. For this reason, its use is limited to the cases in which a dedicated channel in the opposite direction is not available, as would be the case with initial access of a mobile station to the network. For example, in 3G networks using FDD mode, this is applied only prior to initiating the transmission on the random access channel (RACH).

**Closed Loop Power Control**

In case of closed loop power control, the transmitter adjusts the transmission power based on some performance metric such as received signal power level, received SIR, or received bit error rate (BER) or block error rate (BLER), which is fed back to the transmitter from the receiver. A dedicated channel is used to transmit power control commands to the transmitter. For example, if the receiver measures that SIR is below SIR_{target}, it sends a power control command to the transmitter indicating that it has to increase the transmit power by a certain amount. If SIR is higher than SIR_{target}, the power control command indicates that the transmitted power must be decreased. If the periodicity of these power control commands is faster than the channel variations, this strategy allows the compensation of channel-gain including fast fading. Therefore, it is a preferred method in current systems for both uplink and downlink power control.

**Outer Loop Power Control**

In the uplink direction, outer loop power control in a cellular network focuses on adjusting the target SIR point in the BS according to the needs of the individual radio link. It aims at maintaining a constant quality, usually defined as a certain target BER or BLER. The required SIR for desired BLER depends on the mobile speed and the multi path profile of the channel. Therefore, if the target SIR point is set for the worst case, it would waste much capacity for those connections at low speeds. Thus, the best strategy is to let the target SIR point to float around the minimum value that just fulfills the required target quality. The target SIR point will change over with time as the speed and propagation environment change. The uplink outer loop control is typically implemented by having the BS to tag each uplink user data frame with a frame reliability indicator, such as a CRC (Cyclic Redundancy Check) result obtained during decoding of that particular user data frame. In 3G networks, when the frame quality indicator shows that the transmission quality is decreasing, the RNC (Radio Network Controller) will instruct the NodeB (i.e., the BS) to alter the target SIR point accordingly. In the downlink direction, the mobile terminal executes the outer loop power control internally in order to meet the BLER requirement that has been set by the network during the establishment or reconfiguration procedures.

**Inner Loop Power Control**

Once the SIR target point has been set either in the uplink or in the downlink direction, the inner loop power control is responsible for adjusting the transmitted power in order to reach the receiver with the target SIR point on a fast time basis.
Table 2.3 $P_{\text{max}}$ and Tolerance Levels

<table>
<thead>
<tr>
<th>$P_{\text{max}, \text{L}} \leq P_{\text{max}} \leq P_{\text{max}, \text{H}}$</th>
<th>Tolerance $T$</th>
<th>(s.t. $P_{\text{max}, \text{L}} - T \leq P_{\text{max}} \leq P_{\text{max}, \text{H}} + T$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 dBm $\leq P_{\text{max}} \leq 23$ dBm</td>
<td>2.0 dB</td>
<td></td>
</tr>
<tr>
<td>20 dBm $\leq P_{\text{max}} &lt; 21$ dBm</td>
<td>2.5 dB</td>
<td></td>
</tr>
<tr>
<td>19 dBm $\leq P_{\text{max}} &lt; 20$ dBm</td>
<td>3.5 dB</td>
<td></td>
</tr>
<tr>
<td>18 dBm $\leq P_{\text{max}} &lt; 19$ dBm</td>
<td>4.0 dB</td>
<td></td>
</tr>
<tr>
<td>13 dBm $\leq P_{\text{max}} &lt; 18$ dBm</td>
<td>5.0 dB</td>
<td></td>
</tr>
<tr>
<td>8 dBm $\leq P_{\text{max}} &lt; 13$ dBm</td>
<td>6.0 dB</td>
<td></td>
</tr>
<tr>
<td>$-40$ dBm $\leq P_{\text{max}} &lt; 8$ dBm</td>
<td>7.0 dB</td>
<td></td>
</tr>
</tbody>
</table>

Example 1 Power Control in LTE Networks [51–53]: As an example, here we describe PUSCH (Physical Uplink Shared Channel) power control in LTE networks. The setting of the UE Transmit power $P_{\text{PUSCH}}$ for the PUSCH (Physical Uplink Shared Channel) transmission in subframe $i$ is defined by

$$P_{\text{PUSCH}} = \min\{P_{\text{max}}, 10 \log_{10} M + P_0 + \alpha P_l + \delta_{\text{mcs}} + f(\Delta_i)\} \ [\text{dBm}]$$ (2.111)

where $P_{\text{max}}$ is the maximum allowed transmit power. It depends on the UE power class and should satisfy $P_{\text{max}, \text{L}} \leq P_{\text{max}} \leq P_{\text{max}, \text{H}}$ as in Table 2.3. $M$ is the number of physical resource blocks (PRBs), $P_0$ is a cell/UE specific parameter, and $\alpha$ is the path-loss compensation factor, which is a 3-bit cell specific parameter. $P_l$ is the downlink path-loss estimate, which is calculated in the UE based on the RSRP (Reference Symbol Received Power). $\delta_{\text{mcs}}$ is a cell/UE specific modulation and coding scheme, $\Delta_i$ is a closed loop correction value, and $f$ is a function that permits to use accumulated or absolute correction value.

In the above, $P_0$ is calculated as follows:

$$P_0 = \alpha (\text{SNR}_0 + P_n) + (1 - \alpha)(P_{\text{max}} - 10 \log_{10} M_0) \ [\text{dBm}]$$ (2.112)

where $\text{SNR}_0$ is the open loop SNR target, $P_n$ is the noise power per PRB, and $M_0$ is the number of PRBs for which the SNR target is reached with full power.

Among the parameters, UE receives $\Delta_i$, $\alpha$, $P_0$, and $\delta_{\text{mcs}}$ from the eNodeB. The parameter $\Delta_i$ is signaled by the eNodeB to any UE after it sets its initial transmit power; i.e., $\Delta_i$ has no role in the initial setting of the UE transmit power. Also, the parameter $\delta_{\text{mcs}}$ is neglected at this initial stage. A UE sets its initial transmission power based on other received parameters from the eNodeB and path loss calculated by the UE as follows:

$$P_{\text{PUSCH}} = 10 \log_{10} M + P_0 + \alpha P_l \ [\text{dBm}]$$ (2.113)

The power assignment for transmission at the UE is performed on the basis of PRB and each PRB contains equal amount of power. Thus, by neglecting $M$, based on (2.111), the expression used by the UE to assign power to each PRB (i.e., transmit power spectral density in a PRB expressed in dBm/PRB) is

$$\text{PSD}_{\text{TX}} = P_0 + \alpha P_l \ [\text{dBm/PRB}]$$ (2.114)
Table 2.4 Mapping of TPC Command Values to Absolute and Accumulated $\Delta_i$ Values

<table>
<thead>
<tr>
<th>TPC Command Value</th>
<th>$f_{\text{accumulated}}(\Delta_i)$</th>
<th>$f_{\text{absolute}}(\Delta_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$-1$ dB</td>
<td>$-4$ dB</td>
</tr>
<tr>
<td>1</td>
<td>$0$ dB</td>
<td>$-1$ dB</td>
</tr>
<tr>
<td>2</td>
<td>$1$ dB</td>
<td>$1$ dB</td>
</tr>
<tr>
<td>3</td>
<td>$3$ dB</td>
<td>$4$ dB</td>
</tr>
</tbody>
</table>

Depending on $\alpha$, the power control method can be categorized as follows: $\alpha = 1$ (i.e., scenario with full compensation of path loss) leads to conventional power control; $0 < \alpha < 1$ (i.e., scenario with fractional compensation of path loss) leads to fractional power control; and $\alpha = 0$ (i.e., no compensation for path-loss) leads to no power control at all in which case the users use the maximum allowed transmission powers.

Open loop power control: A UE first measures the received power from the eNodeB and estimates the path loss, which is obtained after measuring RSRP. Parameters $P_0$ and $\alpha$ are broadcasted to the UE from the eNodeB. Then, as mentioned above, the UE sets its initial open loop power $P_{OL}$ as in (2.115). Note that $\delta_{mcs}$ and $f(\Delta_i)$ are ignored in this situation as mentioned above.

$$P_{OL} = \min\{P_{\text{max}}, 10\log_{10} M + P_0 + \alpha P_i\} \text{ [dBm]}.$$  \hspace{1cm} (2.115)

Using $\alpha = 1$ leads to conventional open loop power control while $0 < \alpha < 1$ leads to fractional open loop power control.

Closed loop power control: In this scenario UE adjusts the uplink transmit power in accordance with the closed loop correction value derived based on transmit power control (TPC) commands. TPC commands are transmitted, by the eNodeB toward the UE, based on the closed loop SINR target and measured received SINR. In a closed-loop power control system, the uplink receiver at the eNodeB estimates the SINR of the received signal, and compares it with the target SINR value. When the received SINR is below the target SINR, a TPC command is transmitted to the UE to increase the transmitter power. Otherwise, the TPC command will request to decrease the transmitter power. The closed loop correction value is applied after calculating the transmission power using the open loop power control. Obviously, $f(\Delta_i)$ has to be considered now. The relationship between the TPC command value and $f(\Delta_i)$ is illustrated in Table 2.4. The $f$ determines whether accumulated or absolute $\Delta_i$ to be used.

2.4.4 Cell Association, Handoff Management, and Admission Control

Cell Association
In a cellular network, a cell association (also referred to as user association) method determines which cell (or BS) a user associates to before data communication starts. Consequently, user association affects the users’ QoS performances (e.g., SINR outage performance) as well as spectrum and energy efficiency performances in a network.
Note that in traditional single-tier cellular networks user association for downlink and uplink communications is symmetric (or coupled). That is, a user associates to the same BS for both uplink and downlink communications. However, in multitier cellular systems (or HetNets), a user may associate to different BSs for uplink and downlink communications. Therefore, user association can be asymmetric (or decoupled) in these networks. In general, user association schemes can be developed which exploit one or more of the following information: knowledge of instantaneous channel and transmit power at the receiver (i.e., channel-aware scheme), knowledge of instantaneous interference at the receiver (i.e., interference-aware scheme), traffic load information (i.e., load-aware scheme), information about resource allocation in a BS (i.e., resource-aware scheme), and information regarding the priority of service or network tier (i.e., priority-aware). The user association schemes can be used along with the appropriate power control schemes to optimize the network performance [54]. A comprehensive survey on the user association schemes for next-generation cellular networks can be found in [55].

For single and multitier cellular networks, the common user association methods in the literature are as follows [54]: Reference Signal Received Power (RSRP)-based user association, bias-based Cell Range Expansion (CRE) scheme, user association in LTE networks based on Almost Blank Sub-frame (ABS) ratio, and association based on cell zooming. In the RSRP scheme, a user is associated to the BS from which it receives the largest average signal strength. A variant of RSRP scheme is the Reference Signal Received Quality (RSRQ)-based scheme in which a user selects a BS which gives the highest SIR. In single-tier networks with uniform traffic, the RSRP and RSRQ schemes will maximize the network throughput. However, in a multitier network with different transmit powers of different BSs, since users will associate to the BSs with high transmit powers (from which they receive higher RSRP/RSRQ), these schemes will result in traffic load imbalance.

The CRE scheme can mitigate the problem of traffic load imbalance by increasing the downlink coverage footprint of low-power BSs. This is achieved by adding a positive bias to their signal strengths (i.e., RSRP or RSRQ), which allows more users to associate with low-power or biased BSs and thereby achieves a better load balancing among BSs. However, users associating to the biased BSs may experience strong interference from the unbiased high-power BSs. The ABS technique for LTE networks uses time domain orthogonalization in which specific subframes are left blank by the unbiased BS and off-loaded users are scheduled within these subframes to avoid intertier interference. This will improve the throughput performance of the users associated to the biased BSs at the expense of throughput of the users associated to the unbiased BSs.

Cell zooming is used to adaptively adjust the cell size in order to balance traffic load among cells, i.e., the size of a given cell shrinks if its traffic load increases while it expands for low traffic load conditions. The load balancing is therefore achieved by associating more users to a lightly loaded cell. On the contrary, BS sleeping allows the lightly loaded cells to sleep by reducing their coverage to zero and shifting their load to the neighboring cells. From this perspective, a user association policy can encourage users to associate with the heavy loaded cells which leads to an increased number of sleeping BSs. While significant energy efficiency is expected in such a setup, the
possibility of coverage holes increases as well. It is therefore a challenging task to balance the traffic load with two different objectives, i.e., distributing load for improving spectral efficiency and concentrating load for enhancing energy efficiency.

**Handoff Management**

As a user moves through the coverage area of a cellular network, her ongoing call has to be transferred from one cell to another. The latter cell can be in the same network tier as the former cell (in which case it would be referred to as a *horizontal handoff*) or it can be in another network tier (in which case it would be called a *vertical handoff*). The process of transferring the call is referred to as handoff (or handover). For a cellular wireless network, designing an efficient handoff management mechanism is a significant challenge, and the problem has been extensively studied. A concise review of the early works on handoff performance and control in single-tier as well as two-tier macrocell-microcell networks can be found in [56]. The performance metrics used to evaluate the handoff methods include the following: new call blocking probability, handoff call blocking probability, average number of handoffs per call, handoff call dropping probability, duration of handoff process, and probability of unnecessary handoff. Handoff triggering can be based on one of the following criteria: highest average value of received signal (i.e., relative signal strength), relative signal strength with threshold (i.e., current signal level is less than a threshold and the signal from the target BS is stronger), relative signal strength with hysteresis (i.e., signal from the target BS is stronger by a hysteresis margin), and relative signal strength with hysteresis and threshold (i.e., current signal level falls below a threshold and signal from the target BS is stronger by a given hysteresis margin). Therefore, the key variables involved in the triggering of a handoff are the length of the signal averaging window, the threshold level, and the hysteresis margin.

To accommodate handoffs, in each cell, some channels can be reserved explicitly for handoff calls. One such scheme is the *guard channel scheme* where out of $n$ channels in a cell, $m$ are exclusively used for handoffs) [57]. Therefore, new calls are blocked if there are at least $(n - m)$ calls present. Handoff calls are dropped if all $n$ channels are occupied. Handoff requests can also be queued in the target cell if all $n$ channels are occupied. The call can be queued until a channel becomes available or the received power from the current BS falls below the minimum acceptable level. This will reduce the probability of forced termination of a handoff call. The concept of using guard channels can be also combined with that of queueing new calls. A brief discussion on the system dimensioning procedures for prioritized channel assignments for handoff calls can be found in [29].

Performance modeling of network traffic (e.g., due to new calls and handoff calls) would be required for analysis of handoff management schemes. Teletraffic modeling aspects for both homogeneous and heterogeneous (e.g., multitier) cellular networks were discussed in [58]. In a multitier network, different tiers provide alternate options for network access for the admitted calls or handoff calls that are blocked due to the congestion in a particular tier. The problem of traffic modeling and analysis and handoff management for such hierarchical networks were addressed in [59]–[63].
Admission Control
The objective of a call admission control (CAC) method is to limit the number of ongoing calls so that the target QoS requirements for the ongoing calls in the network can be guaranteed. Note that the network has to serve two types of calls: new calls and handoff calls and the QoS performances for these two types of calls are generally measured by new call blocking probability and handoff call dropping probability. Since users are more sensitive to dropping of an ongoing and handed over call than blocking a new call, a CAC scheme needs to prioritize handoff calls over new calls. Also, the new call blocking probability should be maintained below a given target level. Again, for packet-switched wireless services, in addition to the call-level performance measures, packet-level performance measures (e.g., packet dropping probability, packet transmission delay) at both the wireless interface and the wired interface (e.g., at the IP-aware wireless router/BS) will need to be considered while designing the CAC schemes. In other words, the CAC methods will need to consider the availability of the network resources by taking into account the packet-level performance statistics. Also, they should be aware of the availability of the network resources at the wireless-Internet gateway and in the wired network so that wireless resources are not wasted due to dropping of packets at the wired part of the network.

Compared to a homogeneous network, resource reservation and call admission control become more challenging in a heterogeneous wireless access environment (e.g., in a multitier cellular network with different radio interfaces used in different network tiers) in which the mobile terminals have the ability to connect to different types of networks with different radio access technologies (RATs). Also, CAC must be performed based on different QoS requirements since some calls might use real-time applications that have stricter QoS requirements than those using non-real-time applications. A taxonomy of the different CAC schemes along with a two-level hierarchical CAC scheme with both the call-level and the packet-level QoS considerations for a differentiated services cellular wireless network can be found in [64]. In [65], the CAC problem for a two-tier OFDMA-based heterogeneous network was modeled as a semi-Markov decision process (SMDP) where handoff from macrocell to femtocells was considered. A CAC method for a two-tier macrocell-femtocell network was presented in [30] (Chapter 6), which uses sector-based FFR scheme for spatial channel allocation between a macrocell and femtocells.

2.5 Taxonomy of Resource Allocation

Resource allocation problems in wireless networks can be classified according to the resources or functions in each layer and the wireless network type. To state any given resource allocation problem formally, we require to determine which resources (or functions) are to be allocated (or performed) in which layer and in which type of wireless network. As such, a taxonomy of research areas on resource allocation in wireless networks is given in Table 2.5.
### Table 2.5 A High-Level Taxonomy of Research Areas on Resource Allocation in Wireless Networks

<table>
<thead>
<tr>
<th>Resource Allocation</th>
<th>Type of Wireless Network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>From Infrastructure Point of View</td>
</tr>
<tr>
<td>Cell association</td>
<td>Infrastructure-based</td>
</tr>
<tr>
<td>Channel assignment</td>
<td>Cellular (macro, pico, femto)</td>
</tr>
<tr>
<td>Power control</td>
<td>Relay-based</td>
</tr>
<tr>
<td>Scheduling</td>
<td>Ad-hoc/sensor</td>
</tr>
<tr>
<td>Admission control</td>
<td>Mesh</td>
</tr>
<tr>
<td>Cross-layer resource allocation</td>
<td></td>
</tr>
</tbody>
</table>

2.6 Exercises

**Exercise 2.1:** Assume that a cellular wireless communication system uses QPSK modulation in a transmission channel of bandwidth 30 kHz and a carrier frequency of 900 MHz with a target BER requirement of $10^{-3}$. If the transmissions experience log-normal shadowing and Rayleigh fading and the noise power spectral density is $10^{-16}$ mW/Hz, find the cell radius (or cell size) that can achieve the required BER performance. Assume that the standard deviation of shadow fading is $\sigma = 8$ dB and the maximum transmit power at the BS is 300 mW.

**Exercise 2.2:**

i. Plot the BER performance (w.r.t. $E_b/N_0$) of different digital modulation schemes (BPSK, QPSK, BFSK, MSK, GMSK, and $\pi/4$-DQPSK) in AWGN channel. For GMSK, assume suitable values for the time-bandwidth product of the Gaussian filter.

ii. Plot the BER performance (w.r.t. $E_b/N_0$) in slow Rician fading channel for the above modulation schemes. Assume the following values for Rice factor $K = 0, 6, 12, 18$. Note that $K = 0$ corresponds to the Rayleigh fading case.

iii. For a modulation scheme, the symbol error probability in AWGN is approximately $P_s \approx \alpha_M Q(\sqrt{\beta_M} \gamma_s)$. Using this, find an approximation for the average symbol error probability for this scheme in Rayleigh fading in terms of $\gamma_s$.

iv. For selection combining and maximal ratio combining, plot the probability density function $f_X(x)$ of the normalized random variable $X = \gamma_{diversity}/\gamma_s$ for varying number of receive antennas $M$ ($\gamma_{diversity}$ denotes the instantaneous SNR at the output of the diversity combiner and $\gamma_s$ denotes the average SNR for each diversity branch). Consider a slowly varying Rayleigh fading channel.

**Hint:** Choose $M = 1, 2, \ldots, 10$ and vary $x$ over $1, 2, \ldots, 10$. 

Available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781316212493.003

Downloaded from https://www.cambridge.org/core. Boston University Theology Library, on 29 May 2017 at 22:23:21, subject to the Cambridge Core terms of use,
v. Prove that, for MRC (Maximal Ratio Combining), the probability density function
of instantaneous SNR at the receiver is given by

\[ f(\gamma_M) = \frac{\gamma_M^{M-1} e^{-\gamma_M/\gamma_s}}{\gamma_M^M (M-1)!}, \quad \gamma_M \geq 0 \]

where the symbols have their usual meanings.

vi. For selection combining and maximal ratio combining, plot the variation in BER
performance with the normalized random variable \( X = \gamma_{\text{diversity}}/\gamma_s \) for varying
number of receive antennas \( M \) for the following digital modulation schemes:
BPSK, DBPSK, BFSK. Consider a slowly varying Rayleigh fading channel. Note
that BER performance for BPSK is the same as that for QPSK and MSK.

vii. Can you generate results for cases (iii) and (iv) above considering a slowly varying
Rician fading channel?

Exercise 2.3: For a Rayleigh fading channel, a selection combining scheme is deployed
in a receiver with two diversity branches. One branch has a mean SNR of 15 dB and
the other has a mean SNR of 20 dB. Derive the CDF and the PDF of the SNR of the
combined output.

Exercise 2.4: For a selection combining diversity scheme with two branches with mean
SNRs of 15 dB and 20 dB and i.i.d. Rayleigh fading, calculate the SNR outage proba-
bility for BPSK modulation if we require a BER of \( 10^{-3} \).

Exercise 2.5: For a selection diversity scheme with three branches with equal mean
SNRs of 20 dB and i.i.d. Rayleigh fading, calculate the average SNR of the combiner
output and the average BER for a modulation scheme for which the BER is given as a
function of SNR \( \gamma \) as follows: \( P_b = \frac{1}{2} \exp(-\gamma) \).

Exercise 2.6: For a wireless communication system using 3-branch MRC diversity
receiver and QPSK modulation, a BER of \( 10^{-6} \) is required. Calculate the SNR outage
probability assuming equal mean SNR of 10 dB for each diversity branch experiencing
i.i.d. Rayleigh fading.

Exercise 2.7: Find the average bit error probability for differential-BPSK (D-
BPSK) modulation with a three-branch MRC diversity scheme where one branch has
Nakagami-4 fading with mean SNR 5 dB, one has Rayleigh fading with mean SNR
10 dB, and the other has Rician fading with \( K = 2 \) and mean SNR 15 dB.

Hint: For D-BPSK, \( P_b = \frac{1}{2} \exp(-\gamma) \).

Exercise 2.8:

i. Plot the variations in probability of blocking with traffic intensity (in Erlangs, say,
from 0.1 to 30 Erlangs) for different number of channels (say, from 1 to 50) in a
blocked calls cleared system (i.e., Erlang loss system).

ii. Plot the variations in probability of delay with traffic intensity (in Erlangs, say, from
0.1 to 30 Erlangs) for different number of channels (say, from 1 to 50) in a blocked
calls delayed system (i.e., Erlang delay system). Also, plot the waiting time distribution for such a system under several different values of traffic intensity and number of channels.

iii. For a combined delay and loss system, plot variations in probability of blocking (i.e., loss) for different values of buffer size with variations in traffic intensity under a given number of channels. Show a number of different plots for different number of channels. Also, for such a system, obtain the waiting time distribution \textit{analytically}. Then plot this distribution for different values of buffer size, number of channels, and traffic load.

\textbf{Exercise 2.9:} Prove that, in an \textbf{Erlang-C} system, the variance of call waiting time can be expressed as

$$\frac{\Pr\{D > 0\}}{(C\mu - \lambda)^2} \left[2 - \Pr\{D > 0\}\right]$$

where \(\Pr\{D > 0\}\) is the probability that a call is delayed, \(C\) is the number of channels, \(a\) is the ratio of average call arrival rate and average service rate (i.e., \(a = \frac{\lambda}{\mu}\)).

\textbf{Exercise 2.10:} A packet data source produces traffic at three different rates during a transmission interval: \textbf{low}, \textbf{medium}, and \textbf{high} (Figure 2.8). The probability that the source switches between the different data rates is indicated in the figure below.

![Figure 2.8 State transition diagram.](image)

i. Write down the transition matrix.

ii. Assuming that the source started at a \textbf{low} rate, what is the probability that it will produce at \textbf{high} rate after three transmission intervals?

iii. Determine the steady state probability for each traffic rate.

\textbf{Exercise 2.11:} For a two-user slotted ALOHA network, the probability of new packet generation by a user in a time slot is 0.1 and the probability of retransmission by a backlogged user is 0.2. Determine the average throughput and the average packet transmission delay for this network.

\textbf{Exercise 2.12:} Based on the analysis of IEEE 802.11 DCF presented in [16]

i. Plot “saturation throughput vs. number of stations” for several different values of \(W\) and \(m\) (similar to Fig. 6 in [16]).
ii. Plot “saturation throughput vs. transmission probability $\tau$” for different values of $n$ for the basic access method and the RTS/CTS method (similar to Figs. 7–8 in [16]).

**Exercise 2.13:** Following [16], show that for the basic DCF mechanism, the maximum saturation throughput can be approximated by

$$S_{\text{max}} = \frac{E[L]}{T_s + \sigma K + T_c (K(e^{1/K} - 1) - 1)}$$

where the notations have their usual meanings as in [16].

**Exercise 2.14:** Plot the “$S$ vs. $G$” characteristics for non-persistent CSMA and slotted non-persistent CSMA. Show the impact of parameter $a$ on the throughput performance.

From the results, can you comment on the relative performances of the above packet radio MAC protocols in different wireless access environments?

**Exercise 2.15:** Show that, throughput ($S$) of S-ALOHA in a Rayleigh fading channel with capture (under coherent addition of interfering signals) can be expressed as [39]:

$$S = G e^{-G} \sum_{n=0}^{\infty} \frac{G^n}{n z_0 + 1},$$

where $z_0$ is the capture ratio and $G$ is the mean offered channel traffic (packets/time slot).

What is the throughput in case of perfect capture? Provide a physical interpretation for this result.

**Exercise 2.16:** Following the procedure described in [39], plot the “$S$ vs. $G$” curves for S-ALOHA for different values of capture ratio $z_0$ (for both the coherent addition and the non-coherent addition cases).

For a “quasi-constant traffic density function” (given by (42) in [39]), plot the variations in probability of successful packet reception ($\frac{S(\rho)}{G(\rho)}$) with transmission distance $\rho$ for different values of total offered traffic, $G_t$ and capture ratio $z_0$ (similar to the plots in Fig. 4 in [39]). Also, plot variations in total throughput $S_t$ with $G_t$ for different values of capture ratio (follow equation (49) in [39]).

**Exercise 2.17:** For a cellular CDMA system, using the procedure described in [49], obtain a plot for the variation in normalized other cell interference factor ($I_{\text{O}}/P_{R}K$) with respect to the standard deviation of shadow fading, $\sigma$ (e.g., $\sigma = 0, 2, 3, 4, 6, 8 \text{ dB}$) for different values of $\eta$, the path-loss exponent (e.g., $\eta = 3, 4, 5$). Assume a hard handoff scenario and $a^2 = b^2 = 1/2$ and consider the first two tiers of cell around the tagged cell.

Based on the above results, plot the variations in uplink capacity (in terms of number of users that can be supported per cell) when the required $E_b/I_0 = 7 \text{ dB}$. Ignore the impact of background noise. (This will demonstrate the “soft-capacity” nature of a CDMA system, that is, the capacity fluctuates as the propagation and hence the interference conditions vary.)

**Exercise 2.18:** Follow the derivations in [49] and show how to obtain the bounds on the mean and variance of normalized intercell interference in the uplink (as given in equation (14) in [49] for a certain set of assumed parameters). Based on equation (16) in [49], plot the variations in outage probability with number of users per sector when
$W = 1.25 \text{ MHz}$, $R = 8 \text{ kbps}$, and voice activity factor $\alpha = 3/8$. Make other assumptions, if necessary. Obtain plots for target $E_b/I_0 = 7, 10, 12, 14, 15 \text{ dB}$. (The plots should be similar to those in Fig. 3 in [49].)

**Exercise 2.19:** Show that, for a packet radio network with non-orthogonal multiple access in the same spectrum, when radio propagation is characterized by distance-dependent loss only, the characteristic function for the total interference ($I$) experienced by a mobile (located at the center of a two-dimensional plane) can be expressed as $\phi_I(\omega) = \exp(-\pi \lambda t \Gamma(1 - 2/\gamma) e^{-\pi \gamma / \omega^2 / \gamma})$, where $\gamma$ is the path-loss exponent and $\Gamma(.)$ is the gamma function. Clearly state all the assumptions, the meaning of the notations, and the steps in the derivation.

Based on the $\phi_I(\omega)$ above, for $\gamma = 4$, show that $f_I(y) = \frac{\gamma}{2} \lambda^3 y^{-3/2} \exp\left(-\pi^3 \lambda^2 / 4y\right)$.

**Exercise 2.20:** Following the derivations in [50], plot variations in expected normalized progress per hop ($Z \sqrt{\lambda}$) with probability of transmission ($p$) for different values of $N$. Show plots for different values of $K$. (These plots should be similar to those in Figs. 3, 4, and 5 in [50].)

**Exercise 2.21:** Calculate the improvement in probability of message error relative to the uncoded transmission for a (24, 12) double error correcting linear block code. Assume that coherent BPSK modulation is used and that the received $E_b/N_0 = 10 \text{ dB}$.

**Exercise 2.22:** Compare the message error probability for a communications link with and without the use of error correction coding. Assume that the uncoded transmission characteristics are BPSK modulation, Gaussian noise, $\frac{S}{N_0} = 43776$, $R = 4800 \text{ bps}$.

For the coded case, assume a single bit error correcting (15, 11) code. Consider that the demodulator makes hard decisions and this feeds the demodulated code bits directly to the decoder, which in turn outputs an estimate of the original message.

**Exercise 2.23:** Suppose that a 10 km long wireless link has a transmission rate of 10 Mbps and a bit error rate of $10^{-4}$. Data packets of size 100 bytes (which includes a fixed header of size 100 bits) are used for communication in this link.

i. What is the channel utilization and achieved throughput in the link if SW ARQ is used for error control?

ii. What is the channel utilization and achieved throughput in the link if SR ARQ is used for error control?

iii. If GBN ARQ is used, what should be the window size? What will be the achieved throughput in the link?

**References**


References


Part II

Techniques for Modeling and Analysis of Radio Resource Allocation Methods in Wireless Networks
3 Optimization Techniques

3.1 Basics of Optimization

Optimization is a broad and growing branch of mathematics that deals with the problems of decision making. Optimization problems need to be solved for many decision-making problems in our life. For example, engineers adjust their design to assemble a machine, investors choose which stocks to buy, drivers seek a safe lane that shorten their trip time, and so on. All of these problems can be formulated as mathematical optimization problems. The measurable objective can be profit, cost, gain, or anything that can be represented by a single real number. Then we need to decide the actions or variables that we can control to optimize the objective. We may also need to consider constraints that restrict the feasibility of the variables. In the end, an optimization problem will have the following form:

\[
\begin{align*}
\text{minimize (min)} & \quad f_0(x) \\
\text{subject to (s.t.)} & \quad f_i(x) \leq b_i, \quad i = 1, \ldots, m.
\end{align*}
\]

(3.1)

After the modeling step, one can use different methods to find the solution of an optimization problem. There is no universal method to solve an optimization problem but rather a collection of algorithms for different types of problems: linear optimization, convex optimization, quadratic optimization, etc. Although high-speed computers and user-friendly mathematical softwares can solve many optimization problems quickly and accurately, we still need to learn optimization techniques to formulate optimization models, choose the correct algorithms to analyze the models, verify if the problem is tractable or not, and approximate optimization problems into tractable forms. There are many forms of optimization problems, and each has specific properties and specialized methods to solve. The tree below (Figure 3.1) provides a general representation of the range of optimization problems that we might encounter. In this chapter, we will review the basics of optimization and discuss the details of some popular optimization methods that are used in many engineering problems such as the resource management problems in wireless networks.

3.1.1 Convex Functions

The convexity of a function is very important for analyzing an optimization problem. In fact, the problem will become much simpler if it is proved to be convex. Here we provide the basic definitions and properties of a convex function.
Figure 3.1 Classification of popular optimization problems. Source: www.ece.northwestern.edu/OTC/

Definition 1 A set $C$ is convex if the line segment between any two points in $C$ also lies in $C$. That means for any two points $x_1, x_2$ of $C$ and $\theta \in \mathbb{R}$ such that $0 \leq \theta \leq 1$,

$$\theta x_1 + (1 - \theta) x_2 \in C.$$ 

Definition 2 Let $S \subset \mathbb{R}^n$ be a non-empty convex set, a function $f : S \rightarrow \mathbb{R}$ is convex if for all $x, y \in S$ and $\theta$ with $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1 - \theta) y) \leq \theta f(x) + (1 - \theta) f(y).$$

Proposition 2 Let $S$ be a convex set, then $f : S \rightarrow \mathbb{R}$ is a convex function if and only if its epigraph\(^1\) set $\text{epi}(f) = \{(x, r) \in S \times \mathbb{R} : f(x) \leq r\}$ is also a convex set.

Recall that a neighborhood of a point $x$ is an open set that contains $x$. We define the local minimizer as

Definition 3 A point $x^*$ is a local minimizer if there is a neighborhood $\mathcal{N}$ of $x$ such that $f(x^*) \leq f(x)$ $\forall x \in \mathcal{N}$. Similarly, a point $x^*$ is a global minimizer if $f(x^*) \leq f(x)$ $\forall x \in S$.

The most important feature that makes convex problems attractive is that any local minimum is also a global minimum.

Theorem 3 Consider an optimization problem

$$\min \quad f(x)$$
$$s.t. \quad x \in S.$$ 

\(^1\) The epigraph or supergraph of a function is the set of points lying on or above its graph.
If $S$ is a convex set and $f$ is a convex function, then any local minimum is also the global minimum.

Clearly, to obtain the optimum point, we only need to find a local minimum for $f$. This task can be performed by many different searching algorithms such as gradient methods and Newton’s method. We will discuss more about this later in this chapter.

**Proposition 4** If $f : S \rightarrow \mathbb{R}$ is a function that is twice differentiable, i.e., its secondary derivative (Hessian matrix) $\nabla^2 f$ exists at each point in $S$, then $f$ is convex iff $S$ is convex and its Hessian is positive semidefinite:

$$\nabla^2 f \succeq 0.$$  

Recall that in calculus the sign of derivative represents the direction of a function. If the second derivative is always non-negative, then the first derivative is a monotonically increasing function. Therefore, it can change sign at most once. That means, the original function can change direction from downward to upward at one point only and that point (if exists) will be the global minimizer.

Since convexity has a lot of desirable properties, it is desirable to convert a problem into convex form. Assume $f(x)$ is a convex function. Here is a list of operations on $f$ that preserve the convexity of the new function:

- Non-negative weighted sum
- Composition with affine function
- Pointwise maximum and supremum
- Composition
- Minimization
- Perspective $g(x, t) = tf(x/t), \quad t > 0$.

### 3.1.2 Optimality Conditions for Unconstrained Optimization

We consider the simple case when there is no constraint in (3.1). We also assume the domain of the objective function to be a set $S \in \mathbb{R}^n$ and no convexity.

**Definition 4** A non-zero vector $\mathbf{d} \in \mathbb{R}^n$ is a feasible direction at $\mathbf{x} \in S$ if there exist $\alpha_0 > 0$ such that $\mathbf{x} + \alpha \mathbf{d} \in S$ for all $\alpha \in [0, \alpha_0]$.

The basic questions are how to decide if a point is a local (or global) optimum or not and how we can find these points. Intuitively, if we are at an arbitrary point $\mathbf{x}_0$ and we want to find the local optimal nearest to $\mathbf{x}_0$, we should follow one of the feasible directions that points downward and leads us to the next smaller point $\mathbf{x}_1$. Clearly, a local optimal point is the one at which we cannot go down any further.

We have the following important theorem that can be used to study optimal points of smooth functions (i.e., the ones which have derivatives of all orders).
Theorem 5 (Taylor’s theorem) Let \( f : \mathbb{R}^n \to \mathbb{R} \) be continuously differentiable and \( d \in \mathbb{R}^n \). Then
\[
f(x + d) = f(x) + \nabla f(x + td)^T d
\]
for some \( t \in (0, 1) \). Moreover if \( f \) is twice differentiable we have
\[
\exists t \in (0, 1), \quad f(x + d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x + t d)d.
\]

Using Taylor’s theorem, we can derive the first condition for a point \( x^* \) to be a local minimizer.

Theorem 6 (First-order necessary conditions) Let \( f : S \to \mathbb{R} \) be a continuous differentiable real valued function. If \( x^* \) is a local minimizer of \( f \), then for any feasible direction \( d \) at \( x^* \), we have
\[
d^T \nabla f(x^*) \geq 0.
\]

From the theorem above, if \( x^* \) is a local optimal point and lies inside the interior of \( S \), then any vector \( d \in \mathbb{R}^n \) is also a feasible direction. That means \( \nabla f(x^*) = 0 \). We have established the first necessary condition for a local optimum of a differentiable function. Geometrically, this corollary says that, the local minimum point is where the gradient of function \( f \) changes its sign. Notice that this is only the necessary condition. For example, the quadratic function \( f(x) = -x_1^2 - (x_2 + 1)^2 \) has \( \nabla f = 0 \) at \((0, -1)\), but this is the maximum point not the minimum one. We have the following sufficient condition for the local minimum of \( f \).

Theorem 7 (Second-order sufficient conditions). Let \( f : S \to \mathbb{R} \) be a twice continuously differentiable function. If \( \nabla f(x_0) = 0 \) and \( \nabla^2 f(x_0) \) is positive definite, then \( x_0 \) is a local minimum.

Calculating and checking positive definiteness of the Hessian matrix can be very troublesome. Fortunately, if we combine the first order necessary condition with the convexity in the first section, we obtain the following corollary.

Corollary 1 Let \( f : S \to \mathbb{R} \) be a convex twice differentiable function. Then \( x^* \) is a global minimum iff \( \nabla f(x^*) = 0 \).

Obviously, by solving the equation \( \nabla f(x) = 0 \), we obtain the global optimum solution. This explains why convexity is a desired property for any optimization problem.

Example 8 Consider the quadratic problem \( f(x) = \frac{1}{2} x^T Q x + c^T x \), where \( Q \) is a positive semidefinite matrix. It is easy to see that this is a convex function. From Corollary 1, the local minimum point \( x^* \) must satisfy \( Qx^* + c = 0 \). If we know that \( Q \) is invertible, then the unique global minimum point is \( x^* = -Q^{-1} c \).
3.1 Basics of Optimization

3.1.3 Line Search Methods for Unconstrained Optimization

In this section, we discuss line search algorithms to find optimal solution for a simple optimization problem without any constraint. The general steps for line search are as follows. First, we choose an initial point $\mathbf{x}_0$. The choice of this point will depend on the problem, and a good initial point can help the algorithm to execute faster.

Next, from $\mathbf{x}_0$, the algorithm will generate a sequence of points $\mathbf{x}_k$ such that the function $f(\mathbf{x}_k)$ is a decreasing one. The algorithm terminates when we cannot find a better solution or when the solution is acceptably close to the optimal point. We will now discuss two line search methods that help us in finding the next feasible point.

Gradient Descent Method

For the gradient descent method, starting from an initial point, we find the feasible descent direction, i.e., $\mathbf{d}^T \nabla f < 0$, and then find the next point along this direction such that the new objective function achieves its minimum, i.e., we find the step length such that the next point is the steepest direction from the current point.

The logic behind this step can be shown using Taylor’s theorem. Let us say we want to find a point $\mathbf{x} + \alpha \mathbf{d}$ in the neighborhood of $\mathbf{x}$ such that $f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x})$. Using Taylor’s theorem, we approximate $f(\mathbf{x} + \alpha \mathbf{d})$ with the unit direction vector $||\mathbf{d}|| = 1$ and $\alpha$ with a very small value as

$$f(\mathbf{x} + \alpha \mathbf{d}) \approx f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})^T \mathbf{d}.$$  

Clearly, we need $f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x})$ which is equivalent to $\alpha \nabla f(\mathbf{x})^T \mathbf{d} < 0$. Further we also want to choose $\mathbf{d}$ such that $\alpha \nabla f(\mathbf{x})^T \mathbf{d}$ is as small as possible. It can be easily proven that this term achieves its minimum when $\mathbf{d} = -\frac{\nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||}$. Renaming $\alpha/||\nabla f(\mathbf{x})||$ as the new $\alpha$, the next point can be written as $\mathbf{x'} = \mathbf{x} - \alpha \nabla f(\mathbf{x})$. That means, the direction is along the gradient of the current point.

Next, to reduce the searching time, we also want to choose step length $\alpha$ such that $f(\mathbf{x} - \alpha \nabla f(\mathbf{x}))$ is minimum. Again, there are two methods to find such a value.

The first choice is called exact line search. At iteration $k$, we choose $\alpha$ such that it minimizes the current objective function:

$$\alpha_k = \arg \min_{\alpha > 0} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)). \quad (3.2)$$

The gradient descent method that uses the exact line search is called the steepest descent method. Although this method gives the best possible result, it requires us to solve another unconstrained optimization (although this time the problem is much simpler since it only contains one variable).

Another method is called backtrack line search, which is as follows:

1. Choose $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$ and $\rho = 1$.
2. Until $f(\mathbf{x}_k - \rho \nabla f(\mathbf{x}_k)) < f(\mathbf{x}_k) - \alpha \rho \nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_k)$, \hspace{1cm} $\rho = \beta \rho$.

Although the result may not be optimal like the exact line search, this method has better computing cost and is guaranteed to converge. Notice that both backtrack and
Optimization Techniques

exact line search can be applied for different feasible direction \( d \) not just the \( -\nabla f(x) \).

In summary, the algorithm for gradient descent method can be stated as follows:

1. \( x_0 = \) a starting feasible point, \( k = 0 \).
2. While \( ||\nabla f(x)|| > \epsilon \) do
   (a) Compute the descent direction \( d_k = -\nabla f(x_k) \).
   (b) Compute step size \( \alpha_k \) using either the exact or backtrack line search method.
   (c) Update \( x_{k+1} = x_k - \alpha_k \nabla f(x_k) \).

The convergence of this algorithm to a stationary point is guaranteed by the following theorem.

**Theorem 9** Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) is continuously differentiable and the sequence \( x_k \) is generated by steepest descent method using exact line search. Then, if \( \nabla f(x_k) \neq 0 \), then \( f(x_{k+1}) < f(x_k) \).

Notice that if \( \nabla f(x_k) = 0 \), then \( x_{k+1} = x_k \) and the algorithm converges. Otherwise, the sequence \( f(x_k) \) is a decreasing sequence and must converge to a value that will be the local optimum of \( f \).

**Newton Method**

In gradient descent method, we approximate the objective function \( f \) using a first order Taylor approximation. In Newton’s method, the function \( f \) is approximated by a quadratic one also using Taylor’s theorem.

\[
f(x + \alpha d) \approx \tilde{f}(x) = f(x) + \alpha \nabla f(x)^T d + \frac{\alpha^2}{2} d^T \nabla^2 f(x) d.
\]

Similar to the previous case, we want to choose the direction \( d \) such that the right-hand side is minimum. This is a convex quadratic function with respect to \( d \) and achieves minimum when \( d = -\nabla^2 f(x)^{-1} \nabla f(x) \). The algorithm is given as follows:

1. \( x_0 = \) a starting feasible point, \( k = 0 \).
2. While \( ||\nabla f(x)|| > \epsilon \) do
   (a) Compute the steepest descent direction \( d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k) \).
   (b) Compute step size \( \alpha_k \) using a backtrack or exact line search.
   (c) Update \( x_{k+1} = x_k + \alpha_k d_k \).

The convergence of the algorithm is guaranteed in some special cases.

**Theorem 10** Let \( \{x_k\} \) be the sequence generated by Newton’s method. If the Hessian matrix \( \nabla^2 f(x_k) \) is positive definite and the derivative \( x_k \neq 0 \), then for \( d = -\nabla^2 f(x)^{-1} \nabla f(x) \), there exists \( \bar{\alpha} > 0 \) such that \( \forall \alpha \in (0, \bar{\alpha}) \)

\[
f(x_k + \alpha d) < f(x_k).
\]
This theorem assures that the algorithm will find a smaller solution at each step, and thus the convergence is guaranteed.

The Newton’s method is among the fastest algorithms to compute the descent directions and obtain local optimizers. However, it is also the most computationally expensive, because at each step we need to compute the Hessian matrix. If the number of variables is large, it will take a large amount of time to finish a step.

3.2 Convex Optimization

In this section we will discuss the algorithms to obtain solutions for a convex optimization problem. As mentioned in previous section, any local solution of a convex problem is also the globally optimum solution. Later, we will see several more interesting properties of a convex problem that explain why convexity is one of the most desirable properties for any optimization problem.

3.2.1 Introduction

A convex optimization problem has the following form:

\[
\begin{align*}
\min & \quad f_0(x) \\
\text{s.t.} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad Ax = b
\end{align*}
\]  

(3.3)

where \( f_i : \mathbb{R}^n \rightarrow \mathbb{R} \) are all convex functions. We denote the optimal solution by \( p^* \).

**Example 11** *A linear programming problem with the form*

\[
\begin{align*}
\min & \quad c^T x \\
\text{s.t.} & \quad Ax \leq b
\end{align*}
\]

*is a convex optimization problem. This can be easily verified using Proposition 4.*

For a convex optimization problem, thanks to its unique properties, there are some effective methods to obtain the solutions. Interior point method is a popular choice in practice. In [1], the authors claim that interior methods can solve (3.3) in about 10–100 iterations. Even if a problem is non-convex, in many cases we can perform convexification (or relaxation) procedures so that we achieve a lower convex bound of the original problem. An example is quadratic constraints quadratic programming optimization where each non-convex constraint is replaced by a looser but convex one. One method to obtain a lower bound is to use Lagrangian dual function which is always a convex function. Duality is an important aspect of optimization and will be studied in the next section.
3.2.2 Duality

We rewrite the original optimization problem in a more general form as follows:

\[
\begin{align*}
\min & \quad f_0(x) \\
\text{s.t.} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad h_i(x) = 0, \quad i = 1, \ldots, p
\end{align*}
\]  

where \( x \in D \subset \mathbb{R}^n \). We refer to this as the **primal problem**. Here we have not assumed any convexity yet. The strategy here is to find some function as a lower bound for the solution of this problem and then try to narrow the gap between them as much as possible. We also want to remove the constraints and convert the problem into an unconstrained optimization that can be solved as discussed in the previous section. Combining these two ideas, we define the Lagrangian function \( L: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \) as

\[
L(x, \lambda, \beta) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \beta_i h_i(x)
\]

where \( \lambda_i \geq 0 \) is the Lagrangian multiplier associated with the inequality constraint \( f_i(x) \leq 0 \) and \( \beta_i \) is the Lagrangian multiplier associated with the equality constraint \( h_i(x) = 0 \). The vectors \( \lambda \) and \( \beta \) are called the Lagrangian multiplier vectors of (3.4). Clearly, if \( x \) is a solution of the primal problem then we must have \( f(x) \geq L(x, \lambda, \beta) \).

We define the Lagrange dual function of (3.4) as follows:

\[
g(\lambda, \beta) = \inf_{x \in D} L(x, \lambda, \beta).
\]

**Proposition 12** For any \( \lambda \geq 0 \) and any \( \beta \), we have

\[
g(\lambda, \beta) \leq p^*
\]

where \( p^* \) is the optimal value of the primal problem.

This proposition means the dual function is the lower bound for the optimal solution of the primal problem. If the Lagrangian function is unbounded below in \( x \), then \( g \) takes the \(-\infty\) value. Notice that \( g \) has affine form of \( \lambda \) and \( \beta \), so it is concave no matter what form the primal function has.

**Example 13** Duality of a linear program (LP)

\[
\begin{align*}
\min & \quad c^T x \\
\text{s.t.} & \quad Ax \leq b
\end{align*}
\]

The dual function is

\[
g(\lambda) = \inf_x \left( c^T x + \lambda^T (Ax - b) \right) \\
= \inf_x \left( (c^T + \lambda^T A) x - \lambda^T b \right).
\]
3.2 Convex Optimization

Clearly, if \( c^T + \lambda^T A \) is not a zero vector, \( g \) is unbounded below. We obtain the following Lagrangian dual form of the standard LP:

\[
g(\lambda) = \begin{cases} 
-\lambda^T b, & \text{if } c^T + \lambda^T A = 0 \text{ and } \lambda \geq 0 \\
-\infty, & \text{otherwise.}
\end{cases}
\]

We define the Lagrangian dual problem as follows:

\[
\max_{\lambda, \beta} g(\lambda, \beta) \\
\text{s.t. } \lambda > 0.
\]

Denote by \( d^* \) the optimal solution for this dual optimization. The inequality \( d^* \leq p^* \) is called the weak duality property. If \( d^* = -\infty \), we have a trivial bound. In the other case, if \( d^* = \infty \), we conclude that the primal problem is infeasible. The difference \( p^* - d^* \) is called the optimal duality gap. In many cases, the dual problem is easier to solve than the primal one. Therefore, it is desirable to make this gap as small as possible so that we have a more accurate bound for the primal problem. We say an optimization problem has strong duality if the gap is zero. This property is guaranteed under some specific conditions when the optimization problem is convex.

**Theorem 14 (Slater’s conditions)** The convex optimization problem of the form

\[
\min f_0(x) \\
\text{s.t. } f_i(x) \leq 0, \quad i = 1, \ldots, m \\
Ax = b
\]

has strong duality properties if there exists a strictly feasible point \( x_0 \) such that \( f_i(x_0) < 0 \).

Notice that Slater’s theorem is only a sufficient condition; therefore, a non-convex problem may also have strong duality. For example, the following single constraint quadratic problem,

\[
\min x^T A_0 x + 2b_0^T x + c_0 \\
\text{s.t. } x^T A_1 x + 2b_1^T x + c_1 \leq 0
\]

where \( A_i \) is not a positive definite matrix, possesses the strong duality property [1].

### 3.2.3 KKT Conditions

Assume that all of the functions \( f \) in (14) are differentiable. The Lagrangian function becomes a continuous function with respect to \( x \). Let \( x^* \) and \( (\lambda^*, \beta^*) \) denote the primal and dual optimal points of \( f \) and \( g \), respectively, and assume that strong duality holds. Using the first order necessary condition for \( L \) at \( x^* \) and the strong duality property, we
Optimization Techniques

have the following equations:

$$\nabla x L(x^*, \lambda^*, \beta^*) = \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \beta_i^* \nabla h_i(x^*) = 0 \quad (3.6)$$

$$f_i(x^*) \leq 0, \quad i = 1, \ldots, m \quad (3.7)$$

$$h_i(x^*) = 0, \quad i = 1, \ldots, p \quad (3.8)$$

$$\lambda_i^* \geq 0, \quad \lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m. \quad (3.9)$$

When the primal problem is convex, then the KKT conditions are also the sufficient conditions for the points to be primal and dual optimal. In some special cases, it is possible to solve directly the system of equations in (3.6) to obtain the optimum value. In general, many algorithms find the optimal value by iteratively solving the KKT equations [1].

**Example 15** A cellular BS needs to allocate its power among a set of $M$ channels in order to achieve the best throughput. Each channel has different gains and noise associated with it. Denoting by $p_i$ the power transmitted over channel $i$, we have the following convex problem:

$$\min \quad - \sum_{i=1}^{M} \log_2(1 + p_i g_i)$$

$$s.t. \quad p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{M} p_i = P.$$

Introducing Lagrange multipliers $\lambda^*$ and $\beta^* \geq 0$ and applying the KKT conditions, we have

$$p_i^* \geq 0, \quad \sum_{i=1}^{M} p_i^* = P, \quad \beta_i^* p_i^* = 0, \quad i = 1, \ldots, M$$

$$- g_i / (1 + p_i^* g_i) - \beta_i^* + \lambda^* = 0, \quad i = 1, \ldots, M.$$

Solving the equations above yields $p_i^* = [1/\lambda^* - 1/g_i]^+$. Using the sum of power constraint, we have

$$\sum_{i=1}^{M} [1/\lambda^* - 1/g_i]^+ = P.$$

This equation has unique solution. Therefore, the Lagrangian multiplier $\lambda$ can be found numerically, i.e., by using bisection method. The term $\lambda^*$ can be seen as the threshold for power allocation, and any channel-gain smaller than this threshold will receive zero power. This power allocation method is called the water-filling method.
3.2.4 Algorithms

Convex Optimization with Only Equality Constraints

To begin with, we investigate the simplest case where constraints include only linear equalities and the objective is a quadratic function as follows:

$$\min_x f(x) = \frac{1}{2} x^T P x + q^T x$$

s.t. \hspace{1em} Ax = b

where \( P \succ 0 \) and symmetric.

Denoting by \( x^* \) the optimal solution and applying the KKT conditions, we have

$$Ax^* = b, \quad Px^* + q + A^T \lambda^* = 0$$

where \( \lambda^* \) are Lagrangian multipliers. Then, we can rewrite these two equations as

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}.$$  \hspace{1em} (3.10)

This matrix equation can be solved by many methods such as Gaussian elimination. However, in this case, since we assume \( P \succ 0 \), the KKT matrix (the leftmost one) must be non-singular and thus the solution can be easily obtained by multiplying both sides with its inverse matrix. The solution in this case is unique.

For a general objective function \( f \), our strategy is to find a local optimal solution, and since this problem is convex, the locally optimal solution is also the globally optimum solution. First, we start with a feasible solution. Then we search for the next feasible point which returns a better objective function. The example above hints that if we can approximate the problem by a quadratic optimization, we can easily obtain the solution. Assume that the current point is \( x \) and the next point is \( x + d \). We use a second-order Taylor’s approximation of \( f(x + d) \) as

$$\min_{d} f(x + d) \approx \tilde{f}(x + d) = f(x) + \nabla f(x)^T d + (1/2)d^T \nabla^2 f(x)d$$

s.t. \hspace{1em} \( A(x + d) = b \).

with the direction \( d \).

Using KKT conditions, the Newton step \( d \) can be found as

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \omega \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$  \hspace{1em} (3.11)

where \( \omega \) is the Lagrangian dual variable associated with the equality constraint. From the equation, the Newton step is defined only if the KKT matrix is nonsingular, this condition can be guaranteed if \( \nabla^2 f(x) > 0 \). Also, notice that if there is no equality constraint here, the direction \( d = \nabla^2 f(x)^{-1}\nabla f(x) \). This is consistent with the Newton’s method for unconstrained problem discussed earlier in this chapter. Denoting \( \lambda(x) = (d^T \nabla^2 f(x)d)^{1/2} \), it can be shown that

$$f(x) - \inf[\tilde{f}(x + d)|Ax + d = b] = \lambda^2(x)/2.$$


Obviously, this term can be used as a threshold to approximate the gap $f(x) - p^*$ to terminate the algorithm. Finally, we see that if $d$ is a feasible direction for the problem, i.e., $A(x + d) = b$, then for any $t > 0$, the direction $td$ is also feasible. This implies we should choose $t$ such that $f(x + td)$ is the minimum. The Newton’s method for equality constraint is then given as follows:

1. $x_0 = a starting feasible point such that $Ax_0 = b$, $k = 0$.
2. While $\lambda^2(x)/2 > \epsilon$ do
   (a) Compute the steepest descent direction $d_k$.
   (b) Compute step size $\alpha_k$ using backtrack or exact line search.
   (c) Update $x_{k+1} = x_k + \alpha_k d_k$.

**Interior Point Method**

This is one of the most efficient algorithms used to solve linear and nonlinear convex optimization problems. The interior point method starts from the middle of the feasible set and then iteratively looks for a better solution.

**Convex Optimization with Inequality Constraints**

We consider an optimization problem with inequality constraints as follows:

$$\min \ f_0(x) $$

$$\text{s.t. } f_i(x) \leq 0, \ i = 1, \ldots, m$$

$$A x = b.$$  \hfill (3.12)

We try to remove the inequality constraints so that we can apply the Newton’s method discussed earlier. To do that, we “push” the inequalities into the objective functions by using logarithmic functions as follows:

$$\min \ f_0(x) + \sum_{i=1}^{m} -(1/t) \log(-f_i(x)) $$

$$\text{s.t. } A x = b.$$ \hfill (3.13)

The function $\phi(x) = - \sum_{i=1}^{m} \log(-f_i(x))$ is called log barrier of the problem in (3.12). Its domain is the set of points satisfying the inequality constraint in (3.12) strictly. If any of the functions $f_i(x) \rightarrow 0$, then the objective will reach infinity and the problem becomes unbounded. Notice that the problem here is still convex and thus can be solved using Newton’s method. We can see that by increasing $t$ we can improve the quality of the approximate result, i.e., the gap between the optimal value and the approximate one is smaller. If $t \rightarrow \infty$ and given that $x$ is strictly feasible, then the approximate result will become the optimal value.

Another way to verify the correctness of this approach is to use the KKT conditions. Assume that there exists a strictly feasible solution $x^*$ and the strong duality holds.
we write the KKT conditions as follows:

\[ \begin{align*}
Ax^* &= b, \quad f_i(x^*) \leq 0, \quad i = 1, \ldots, m \\
\lambda_i^* f_i(x^*) &= 1/t, \quad i = 1, \ldots, m \\
\lambda^* &\geq 0 \\
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + A^T \nu &= 0.
\end{align*} \]

Clearly, when \( t \to \infty \), the new KKT conditions will be similar to the KKT equations of the original problem. This implies the new optimization problem and the original one will have the same solution when \( t \) approaches infinity. The next task is to find the feasible starting point for the new log-barrier problem and apply the Newton’s algorithm to solve the new optimization.

**The Central Path**

We convert (3.13) into an equivalent problem as follows:

\[
\begin{align*}
\min & \quad tf_0(x) + \phi(x) \\
\text{s.t.} & \quad Ax = b
\end{align*}
\]

where \( \phi(x) = \sum_{i=1}^{m} -\log(-f_i(x)) \). Assume that point \( x^*(t) \) is a feasible solution of (3.14). The central path of the original problem in (3.12) is defined as the set of “central points” \( x^*(t) \) that satisfy the following conditions (KKT conditions):

\[ \begin{align*}
Ax^*(t) &= b, \quad f_i(x^*(t)) < 0, \quad i = 1, \ldots, m \\
\exists \nu, \quad t \nabla f_0(x^*(t)) + \nabla \phi(x^*(t)) + A^T \nu &= 0.
\end{align*} \]

It can be proved that for any \( x^*(t) \) belonging to the central path, \( 0 \leq f_0(x^*(t)) - p^* \leq m/t \). Therefore, by increasing \( t \), we can achieve a good approximate value for the optimal solution. However, if \( t \) is large, the number of steps to solve (3.14) is also large. We need to find a good trade-off between the total number of iterations for \( t \) and the complexity in finding the optimal solution for each \( t \). Here we increase \( t \) by a factor of \( \mu > 1 \) at each step. This method is called the log barrier method and is summarized as follows:

1. \( x \) = a strictly feasible point and \( t = t(0) > 0, \mu > 1 \), and tolerance \( \epsilon > 0 \).
2. Repeat
   (a) Compute \( x^*(t) \) of (3.14) with starting point \( x \).
   (b) Update \( x = x^*(t) \).
   (c) If \( m/t < \epsilon \) stop.
   (d) Update \( t : t = \mu t \).

The barrier method requires a strictly feasible starting point \( x(0) \). In case this point is not known, a preliminary stage is performed to compute a strictly feasible point, and then this point will be used as the starting point. To do this, we solve the following
feasible problem:

\[
\begin{align*}
\min & \quad s \\
\text{s.t.} & \quad f_i(x) \leq s, \quad i = 1, \ldots, m \\
& \quad Ax = b.
\end{align*}
\] (3.15)

The variable \( s \) is interpreted as a bound on the maximum feasibility of the inequality. Clearly, the goal is to drive the bound to below zero so that we achieve a strictly feasible solution. This problem is strictly feasible, because we can choose any \( x_0 \) that satisfies the equality constraints to be the starting point for (3.15) and choose \( s \) as the max \( f_i \). Then the barrier method can be used to find the solution.

**Subgradient Method**

One drawback of the gradient descent method is that it requires the objective function to be twice differentiable and calculation of these derivatives could be very resource-consuming especially for problems with a large number of variables. The sub-gradient method uses a simpler algorithm that can solve the problem in (3.12) even if the objective function is non-differentiable but still convex. Several key advantages of the sub-gradient method over the interior point method are as follows:

- It can be applied to non-differentiable objective function \( f_0 \).
- The step length can be a fixed number (which makes the computation simpler).
- It does not require calculation of the Hessian matrix.

The main drawback is, it is not a descent method and the function value can fluctuate up or down. Therefore, the convergence time is longer than that for the interior point method.

Recall that the subgradient of a function \( f \) at \( x \), denoted as \( \partial f(x) \), is a set of vectors \( g \) that satisfies \( f(y) \geq f(x) + g^T(y - x) \) for all \( y \). Again, the strategy is: given a starting point (which may be infeasible), at each iteration, we choose the next point that has smaller objective function (or makes the constraint that was previously infeasible to a feasible one).

First, we consider the simple case where there is no constraint:

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

where \( f \) is a convex function. Denoting \( g_k \in \partial f(x_k) \), the updated point \( x_{k+1} \) is calculated from \( x_k \) as

\[
x_{k+1} = x_k - \alpha_k g_k.
\] (3.16)

Again, since \( g_k \) may not be a descent direction at \( x \), the value function can be increasing, i.e., \( f(x_{k+1}) > f(x_k) \). To circumvent that, we keep track of the smallest value so far and its corresponding point. At each step, we set

\[
f_{\text{best}}(k) = \min\{f_{\text{best}}(k - 1), f(x_k)\}.
\] (3.17)

Clearly, \( f_{\text{best}}(k) \) is a decreasing sequence, and it has a lower bound that can be \(-\infty\).
### Selection of the Step Size

There are several ways to select the step size. Here we will discuss the most popular ones:

- **Constant step size:** \( \alpha_k = \alpha \) is a positive constant.
- **Constant step length:** \( \alpha_k = \gamma / ||g_k||_2 \), where \( \gamma > 0 \). This means the distance between two consecutive points is constant, i.e., \( ||x_{k+1} - x_k||_2 = \gamma \).
- **Square summable but not summable:** The step size satisfies the following conditions:
  \[
  \alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.
  \]

The key property of these methods for choosing the step size is that they do not require any information about the data computed during the iteration (except the second option but this is still much faster than the interior point method). This makes the subgradient method a very simple algorithm to implement.

### Convergence

The subgradient method is proved to converge when subgradient of \( f \) is bounded, i.e., \( f \) satisfies the Lipschitz condition as follows:

\[
\exists G, \quad |f(x) - f(y)| \leq G||x - y||.
\]

Denoting by \( x^* \) the optimal point and \( R \geq ||x_0 - x^*|| \) the upper bound for the distance between the initial point and the optimum, we have the following theorem.

**Proposition 16** Using sub-gradient method and assuming that \( f \) satisfies Lipschitz condition, at each step \( k \), we have the following inequality:

\[
fbest(k) - f(x^*) \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}.
\] \hspace{1cm} (3.18)

By choosing a suitable step size we can control the accuracy of the algorithm. Specifically, if the step size is square summable but not summable, the algorithm will converge to \( f(x^*) \).

### Projected Subgradient Method

The subgradient method can be extended to the general case where the constraint set is \( S \in \mathbb{R}^n \) and convex. Define the orthogonal projection \( \Pi_S(x) \) of \( x \) as the point \( x' \in S \) that is closest to \( x \), i.e., \( x' = \arg \min_{x' \in S} ||x - x'||. \) The idea is as follows: at iteration \( k \), the next point is defined as \( x_k - \alpha_k g_k \) as usual, but if this new point does not belong to the constraint set \( S \), we need to project it back to \( S \). That means

\[
x_{k+1} = \Pi_S(x_k - \alpha_k g_k)
\] \hspace{1cm} (3.19)

where \( g_k \in \partial f(x_k) \). Clearly, if \( S = \mathbb{R}^n \), we have \( \Pi_S(x) = x \) and the constrained problem reduces to an unconstrained one. Using the step sizes similar to the case without constraints, it can be proven that the convergence will always occur.
For the simple case where the constraints are linear equality constraints, the optimization formulation is given by

$$\min \ f(x)$$
$$\text{s.t.} \ Ax = b. \quad (3.20)$$

The projected subgradient update can be derived as

$$x_{k+1} = \Pi_S(x_k - \alpha_k g_k) = x_k - \alpha_k (I - A^T (AA^T)^{-1} A) g_k. \quad (3.21)$$

Then we can apply the same method to obtain the step size as before.

### 3.3 Integer Programming

In many cases, we need to solve an optimization problem that accepts only an integer solution. Since the solution $x$ must be an integer vector, the objective function $f(x)$ is not continuous and differentiable, thus we cannot directly apply previous techniques such as Lagrangian duality or Newton’s method. Instead, we can customize the global optimization methods to find the optimal value.

A general integer programming problem is defined as follows:

$$\min \ f_0(x, y)$$
$$\text{s.t.} \ f_i(x, y) \leq 0, \quad i = 1, \ldots, m$$
$$x \in \mathbb{Z}^n \quad (3.22)$$

where $f_i$ is not necessarily convex. In this section, for global optimization, we will discuss application of several popular methods such as cutting plane and branch and bound and how we can apply them in integer programming problems.

#### 3.3.1 Cutting Plane Method

In this method, we try to approximate the optimal set (set that contains optimal solutions) by a polytope. At each iteration, we pick a point in the feasible set so far. If that point belongs to the optimal set, we stop. Otherwise, we find a hyperplane that separates this point from the optimal set. We remove the half-space that contains this point and start the iteration again. This is why it is called the cutting-plane method. This algorithm is less efficient than interior point method but has several advantages as follows:

- It does not require differentiability. Each iteration only requires the computation of the subgradient,
- It can be sped up by exploiting some specific structure in large and complex problems,
- It does not require evaluation of objective and all the constraints at each iteration, and thus useful when the number of constraints is large, and
- It can be decomposed to sub-problems and can be solved in parallel.
For simplicity, we consider only the unconstrained optimization case, i.e., where \( m = 0 \).

**The General Framework**

Given point \( x_0 \), denote by \( g \) a vector belonging to the subgradient of \( f_0 \) at \( x_0 \). From the definition of subgradient, we have

\[
f_0(y) \geq f_0(x) + g^T(y - x).
\]

Obviously, if \( g^T(y - x) > 0 \), then \( y \) does not belong to the optimal set. We can remove the half-plane that satisfies \( g^T(y - x) > 0 \) and add one more constraint \( g^T(y - x) \leq 0 \) into the feasible set.

If the cutting plane contains \( x \) on its border then it is a neutral cut. If \( x \) lies inside the cutting plane then it is a deep cut. A deep cut can be performed if we know a value \( f' \) such that \( f_0(x) > f' \geq f^* \). In that case we can have a deep cut as

\[
g^T(y - x) + f_0(x) - f' < 0.
\]

In most of the cases, a deep cut is better than the neutral one, because it excludes a larger set of points at each iteration. The general framework can be presented as follows:

1. Initialize a starting polyhedron \( P_0 = \{ y | A_0 y \leq b_0 \} \) known to contain optimal set \( P^* \) and set \( k = 0 \).
2. At iteration \( k + 1 \), if the stopping criteria is not met:
   - Choose a point \( x_{k+1} \) in \( P_k \).
   - If \( x_{k+1} \) is in optimal set \( P^* \), i.e., if the derivative is equal to zero, quit.
     Else update \( P_k \) by adding new cutting plane \( P_{k+1} = P_k \cap \{ y | a_{k+1}^T y \leq b_{k+1} \} \).
   - If \( P_{k+1} = \emptyset \), quit.
   - Else \( k = k + 1 \).

The two main components in the algorithm are how to choose the next query point \( x_{k+1} \) and the stopping criterion. Usually, these tasks depend on the specific problems, but we can give some popular methods as follows.

For stopping conditions, we can show that the optimal solution belongs to a ball with radius \( r \) and prove that after \( N \) iterations we obtain a solution that belongs to that ball.

For choosing the next query point, we can follow two directions:

1. The center of gravity: The query point is chosen as the center of gravity of \( P_k \). At each iteration, the volume of the polytope \( P_k \) is reduced by at least 37% and hence convergence is guaranteed and fast. However, the implementation is extremely difficult.
2. The MVE cutting-plane method: The next query point is the center of the maximum volume ellipsoid that lies inside \( P_k \). The convergence speed will depend on the number of faces of \( P_k \).

**Integer Linear Programming Using Gomory’s Cutting-Plane Algorithm**

Here we discuss one of the cutting-plane approaches for integer linear programming, which is called the Gomory’s cutting-plane algorithm. A linear integer program is one...
where all the objective and constraint functions are affine:

\[
\begin{align*}
\text{min} & \quad c^T x \\
\text{s.t.} & \quad Ax = b \\
& \quad x \in \mathbb{Z}^n, \quad x \geq 0.
\end{align*}
\]

At each step, a set of non-integer solutions is obtained by using simplex method and are removed from the feasible set by adding constraints that exclude them out of the feasible set while keeping the integer solutions in. The process is repeated until the optimal solution is an integer vector.

First, we need to recall the definition of floor operator and the basic feasible solution for a linear program.

**Definition 5** The floor of a real number, denoted as \( \lfloor x \rfloor \), is the largest integer that is smaller or equal to \( x \).

**Definition 6** For the equality \( Ax = b \), if there are \( n \) variables and \( m \) constraints, a solution with at most \( m \) non-zero values is a basic solution. Furthermore, if the basic solution satisfies the condition \( x \geq 0 \), then it is basic feasible solution.

Using simplex method, we can obtain an optimal basic feasible solution \( x \) of the linear program \([2]\). Suppose the first \( m \) columns form the basis for the optimal basic feasible solution, then we have the corresponding canonical augmented matrix as follows:

\[
\begin{bmatrix}
  a_1 & a_2 & \ldots & a_m & a_{m+1} & \ldots & a_n & b \\
  1 & 0 & \ldots & 0 & a_{1,m+1} & \ldots & a_{1,n} & b_1 \\
  0 & 1 & \ldots & 0 & a_{2,m+1} & \ldots & a_{2,n} & b_2 \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  0 & 0 & \ldots & 1 & a_{m,m+1} & \ldots & a_{m,n} & b_m
\end{bmatrix}
\]

Consider the \( i \)-th component of the optimal basic feasible solution, \( b_i \). We consider the case where \( b_i \) is not an integer. Then from the equality constraint we have

\[ x_i + \sum_{j=m+1}^{n} a_{i,j} x_j = b_i. \]

Using the floor function and the condition that \( x \) is non-negative integer, we have the following inequality constraint:

\[ x_i + \sum_{j=m+1}^{n} \lfloor a_{i,j} \rfloor x_j \leq \lfloor b_i \rfloor. \] (3.23)

Subtracting this inequality from the equality constraint, we have a new constraint, which is called Gomory cut, as follows:

\[ \sum_{j=m+1}^{n} (a_{i,j} - \lfloor a_{i,j} \rfloor) x_j \geq b_i - \lfloor b_i \rfloor. \] (3.24)

Notice that the optimal basic feasible solution \( x \) above does not satisfy the new constraint. Since \( x \) is a basic feasible solution, we must have \( x_j = 0 \) with \( j > m \). Therefore,
the left-hand side of the inequality above is zero while the right-hand side is positive (since we assume \( b_i \) is non-integer). That means, at each step, we remove at least one basic feasible solution that is not an integer vector. If the number of basic feasible solutions is finite, then the algorithm will have a finite number of iterations.

To transform the new linear program into a standard form, we need to add new non-negative variable \( x_{n+1} \) after each step to obtain the equality constraint:

\[
\sum_{j=m+1}^{n} (a_{i,j} - \lfloor a_{i,j} \rfloor)x_j - x_{n+1} = b_i - \lfloor b_i \rfloor. \tag{3.25}
\]

Notice that this new surplus variable is not necessarily an integer. The entire algorithm can be summarized as follows:

- Step 1: Solve the continuous problem using simplex algorithm.
- Step 2: If the solution is an integer vector, algorithm stops.
- Step 3: Choose a basic variable \( x_i \) which is not an integer. Construct the new constraint (3.25) and add it to the set of the constraints. Go back to Step 1.

### 3.3.2 Branch and Bound Algorithm

Branch and bound algorithms are used for global optimization. They maintain a provable upper and lower bound on the global optimal value. At each step they add “branches,” i.e., constraints into the problem to obtain new solutions. Also, they keep track of the best solution so far and prune (remove) the branches that definitely return a worse solution. Branch and bound algorithms are usually slow, and in the worst case the computation time can grow exponentially.

#### Basics of Branch and Bound

Assume that we want to find optimal value for \( f(x) \), where \( x \) is a vector of \( m \) elements. Assume that at the beginning we know the rectangle that contains the set of feasible solutions. Denote by \( p^* \) the global minimum value for \( f \). Assume that given an \( m \)-dimensional rectangle \( Q \), we can calculate the upper and lower bounds of \( p^* \) as \( \Phi_u(Q) \) and \( \Phi_l(Q) \). The main idea of branch and bound is, at each step, we divide the current set of feasible rectangles into smaller pieces and calculate the upper and lower bounds of each piece. At time \( k \), update the minimum of upper bounds as \( U_k \) and minimum of lower bounds as \( L_k \). Clearly, we have \( L_k \leq p^* \leq U_k \), so the term \( (U_k - L_k) \) can be used as the stopping criterion. The skeleton of the algorithm is as follows:

- \( k = 0 \), \( \mathcal{L}_0 = \{Q_0\}, L_0 = \Phi_l(Q_0) \) and \( U_0 = \Phi_u(Q_0) \).
- If \( U_k - L_k < \epsilon \), then stop. Else
  - Choose \( Q \in \mathcal{L}_k \) such that \( \Phi_l(Q) = L_k \) and \( Q \) is still divisible.
  - Split \( Q \) along one of its longest edges into \( Q_1 \) and \( Q_2 \).
  - Form \( \mathcal{L}_{k+1} \) by removing \( Q \) and replace it with \( Q_1 \) and \( Q_2 \).
  - Update \( L_{k+1} = \min_{Q \in \mathcal{L}_{k+1}} \Phi_l(Q) \) and \( U_{k+1} = \min_{Q \in \mathcal{L}_{k+1}} \Phi_u(Q) \).
  - \( k = k + 1 \).
Here, divisible means $\Phi_n(Q) > \Phi_l(Q)$, otherwise there is no point in considering an area where we already know that the optimal solution exists. By dividing the rectangle $Q$ along one of its longest edges, we can guarantee the convergence of the algorithm. Also, a pruning step can be performed. Let us assume that at step $k$, if some rectangle $i$ satisfies $\Phi_l(Q_i) > U_k$, we can remove this rectangle because it certainly does not contain the optimal point. By performing pruning we can reduce the search area (and thus reduce the storage requirement). The selection rule depends on the problems, but intuitively, we pick the rectangle with the smallest lower bound because it is more likely to contain the optimum point.

**Branch and Bound for Integer Linear Programs**

The algorithm above is just a skeleton and needs to be modified when we want to apply it for a specific case. For example, we need to decide the upper bound and lower bound and the splitting rules. We illustrate this by solving an example of Knapsack problem given below:

$$\begin{align*}
\text{max} \quad & 4x_1 + x_2 + 2x_3 + x_4 \\
\text{s.t.} \quad & 6x_1 + 2x_2 + 5x_3 + 4x_4 \leq 10 \\
& x_i \in \{0, 1\}.
\end{align*}$$ (3.26)

The strategy is as follows: First, we start with a feasible solution $x^*$ and its objective value $z^*$. We call this $z^*$ as the incumbent or the "best so far." Choose this as the root node and mark it as active. At each iteration we solve the relaxed version of the linear program, record the "best so far" as the new incumbent, branch the solutions which are not integers. The steps can be summarized as follows:

While there remain active nodes

1. Step 1: Select an active node $j$ and mark it as inactive.
2. Let $x_j$ and $z_j$ denote the optimal solution and the objective function of the relaxation of this LP problem at iteration $j$.
   - Case 1: If $z^* \geq z_j$ then
     - Prune node $j$.
   - Case 2: If $z^* < z_j$ and $x_j$ is feasible for the integer program, then
     - Replace the incumbent by $x_j$.
     - Prune node $j$.
   - Case 3: If $z^* < z_j$ and $x_j$ is not feasible for IP, then
     - Choose one of the non-integer elements of $x_i$ and branch it to two descendants 0 and 1.
     - Mark these direct descendants of node $j$ as active.

For problem (3.26), first we have the easy feasible solution $x^* = (0, 0, 0, 0)$ and $z^* = 0$. We solve the relaxed version when $x_i$ can take a real value from 0 to 1. The solution is $(1, 1, 0.4, 0)$ and $z_0 = 5.8.$
Next, we pick \( x_3 \) and branch it into two branches which has \( x_3 = 0 \) and \( x_3 = 1 \).

- When \( x_3 = 0 \), we obtain solution \((1, 1, 0, 0.5)\) and \( z = 5.5 \). We further branch this into two branches where \( x_4 = 0 \) and \( x_4 = 1 \). We continue doing this and obtain new incumbent \((1, 1, 0, 0)\) and \( z^* = 5 \).
- When \( x_3 = 1 \), we obtain solution \((0.8333, 0, 1, 0)\) and \( z = 5.3333 \). We can branch this node into \( x_1 = 0 \) and \( x_1 = 1 \). The branch with \( x_1 = 1 \) has no feasible solution while the branch with \( x_1 = 0 \) returns \( z = 3.75 < 5 \). So we can prune them both.

We conclude that the solution is \((1, 1, 0, 0)\) and the maximum value is 5. The steps can be illustrated by a tree as shown in Figure 3.2, where the number inside each circle displays the optimal value of the objective function for the relaxed version of the problem.

**Example 17** Consider a scenario, where a cluster \( c_l \) of small cells is deployed overlaying an existing macrocell. Within the cluster \( c_l \), one small cell is elected as a cluster head \( CH \) and is responsible for resource allocation within the cluster. Small cell user equipments (SUEs) and macro-cell UEs (MUEs) are randomly deployed such that each small cell \( f \in c_l \) serves a single SUE \( k_f \), whereas all MUEs are served by the macro-cell. Denote by \( \gamma_{n,k_f,f}^n \), the unit power SINR of an SUE \( k_f \) served by small cell \( f \) on sub-channel \( n \) with bandwidth \( \Delta f \). \( \gamma_{n,k_f,f}^n \) is estimated based on the interference generated at an SUE \( f \) on sub-channel \( n \) by the other small cells in the previous iteration (time slot). \( P_{n,k_f,f}^n \) is the power assigned to the link between SUE \( k_f \) and small cell \( f \) on sub-channel \( n \), whereas \( \Gamma_{n,k_f,f}^n \) is the sub-channel allocation indicator that takes the value of 1 if sub-channel \( n \) is assigned to SUE \( k_f \) served by femtocell \( f \), and takes the value of 0 otherwise. We define \( g_{k_f,j}^n \) as the channel-gain between user \( k_f \) and base station \( j \) on sub-channel \( n \) and denote by \( N \) the total number of sub-channels. According to this, we can define the following optimization problem [6]:

\[
\max \sum_{f \in c_l} \sum_{n=1}^{N} \Gamma_{n,k_f,f}^n \Delta f \log_2 \left( 1 + P_{n,k_f,f}^n \gamma_{n,k_f,f}^n \right) \tag{3.27}
\]

s.t.

\[
C1 : \sum_{n=1}^{N} \Gamma_{n,k_f,f}^n \Delta f \log_2 \left( 1 + P_{n,k_f,f}^n \gamma_{n,k_f,f}^n \right) \geq R_{k_f}, \forall k_f
\]
The objective of the optimization problem in (3.27) is to maximize the sum-rate of all small cells within the cluster $c_l$ by performing sub-channel and power allocation. Constraint $C1$ is the data rate requirement for an SUE $k_f$ served by small cell $f$. $C2$ is a constraint on the total power budget available at each small cell. Constraint $C3$ puts a limit on the amount of interference introduced by small cell $f$ to MUE $k_m$ on sub-channel $n$. Finally, $C4$ restricts sub-channel $n$ to be allocated in at most one small cell $f$ in cluster $c_l$. Problem (3.27) is a mixed integer non-linear problem (MINLP). Hence, joint sub-channel and power allocation is computationally intractable. Therefore, a common approach is to solve the problem in two phases iteratively as described below.

- **Phase 1**: For a given power allocation, the CH performs sub-channel allocation. For each small cell on each sub-channel, $P_{k_f}^n$ is initialized as the minimum of either $P_{f_{\text{max}}}^n$ or $\zeta_n^{k_m}$. The idea is to keep power as uniform as possible and at the same time not to violate the interference constraints.

- **Phase 2**: Given the resulting sub-channel allocation, the CH performs power allocation.

**Sub-channel allocation by branch and bound**: For a given power allocation, we have an integer linear program (ILP) that can be optimally solved using the Branch and Bound (BnB) technique. Thus, we have the following optimization problem:

$$
\max_{\Gamma_{k_f}^n \in \{0, 1\}} \sum_{f \in c_l} \sum_{n=1}^N \Gamma_{k_f}^n \Delta_f \log_2 \left( 1 + P_{k_f}^n \gamma_{k_f}^n \right) \tag{3.28}
$$

s.t.

$$
C1 : \sum_{n=1}^N \Gamma_{k_f}^n \Delta_f \log_2 \left( 1 + P_{k_f}^n \gamma_{k_f}^n \right) \geq R_{k_f}, \forall k_f
$$

$$
C2 : \sum_{f \in c_l} \Gamma_{k_f}^n \leq 1, \forall n.
$$

BnB is guaranteed to find the optimal sub-channel allocation but its complexity in the worst case is as high as that of exhaustive search since the ILP in (3.28) is an NP-hard problem.
**Power allocation**: Given sub-channel allocation, the CH can perform power allocation in an optimal manner. We have the following optimization problem:

\[
\max_{P^n_{k_f, f} \geq 0} \sum_{f \in C_f} \sum_{n=1}^{N} \Gamma^n_{k_f, f} \Delta f \log_2 \left( 1 + P^n_{k_f, f} \gamma^n_{k_f, f} \right)
\]

(3.29)

**s.t.**

\[
C_1 : \sum_{n=1}^{N} \Gamma^n_{k_f, f} \Delta f \log_2 \left( 1 + P^n_{k_f, f} \gamma^n_{k_f, f} \right) \geq R_{k_f}, \forall k_f
\]

\[
C_2 : \sum_{n=1}^{N} \Gamma^n_{k_f, f} P^n_{k_f, f} \leq P_{f_{\text{max}}}, \forall f
\]

\[
C_3 : \Gamma^n_{k_f, f} P^n_{k_f, f} \gamma^n_{k_f, f} \leq \xi^n_{k_m, f}, \forall n, f.
\]

The optimization problem in (3.29) is a non-linear convex problem that can be efficiently solved by the interior point method. Note that, since sub-channels are already allocated, there is no coupling among the resource allocation problems for small cells. Hence, each small cell can solve its own power control problem. Define the following Lagrangian for each small cell \( f \):

\[
\mathcal{L}_f \left( P^n_{k_f, f}, \mu_{k_f}, \lambda_f, \theta^n_f \right) = \sum_{n=1}^{N} \Gamma^n_{k_f, f} \Delta f \log_2 \left( 1 + P^n_{k_f, f} \gamma^n_{k_f, f} \right)
\]

\[
+ \mu_{k_f} \left( \sum_{n=1}^{N} \Gamma^n_{k_f, f} \Delta f \log_2 \left( 1 + P^n_{k_f, f} \gamma^n_{k_f, f} \right) - R_{k_f} \right)
\]

\[
+ \lambda_f \left( P_{f_{\text{max}}} - \sum_{n=1}^{N} \Gamma^n_{k_f, f} P^n_{k_f, f} \right)
\]

\[
+ \sum_{n=1}^{N} \left( \theta^n_f \left( \xi^n_{k_m, f} - \Gamma^n_{k_f, f} P^n_{k_f, f} \gamma^n_{k_m, f} \right) \right)
\]

where \( \mu_{k_f}, \lambda_f, \text{ and } \theta^n_f \) are Lagrange multipliers. Then the Lagrange dual function can be represented as

\[
g_f \left( \mu_{k_f}, \lambda_f, \theta^n_f \right) = \max_{P^n_{k_f, f}} \mathcal{L}_f \left( P^n_{k_f, f}, \mu_{k_f}, \lambda_f, \theta^n_f \right).
\]

(3.30)

Then, from (3.30), the maximization of \( \mathcal{L}_f \left( P^n_{k_f, f}, \mu_{k_f}, \lambda_f, \theta^n_f \right) \) can be decomposed into \( N \) independent optimization problems for each sub-channel \( n \). By setting the differentiation of \( \mathcal{L}_f \left( P^n_{k_f, f}, \mu_{k_f}, \lambda_f, \theta^n_f \right) \) with respect to \( P^n_{k_f, f} \) to zero, we obtain the following expression for the optimal power:

\[
P^n_{k_f, f} = \left[ \frac{1 + \mu^n_{k_f}}{\ln 2 \left( \lambda^n_f + \theta^n_f \gamma^n_{k_f, f} \right)} - \frac{1}{\gamma^n_{k_f, f}} \right]^+
\]

(3.31)
where \([x]^+ = \max(x, 0)\). On the other hand, we have the following dual problem:

\[
\min_{\{\mu_{k_f}, \lambda_f, \theta_f^n \geq 0\}} g_f \left( \mu_{k_f}, \lambda_f, \theta_f^n \right).
\] (3.32)

In order to solve the dual problem, the Lagrange multipliers can be updated using the projected subgradient method as follows:

\[
\mu_{k_f}^*[t + 1] = \left[ \mu_{k_f}^*[t] - \delta \mu[t] \left( \sum_{n=1}^{N} \Gamma_{k_f, f}^n \Delta f \log_2 \left( 1 + P_{k_f, f}^n \right) - R_{k_f} \right) \right]^+
\]

\[
\lambda_f^*[t + 1] = \left[ \lambda_f^*[t] - \delta \lambda[t] \left( P_{f, \text{max}} - \sum_{n=1}^{N} \Gamma_{k_f, f}^n P_{k_f, f}^n \right) \right]^+
\]

\[
\theta_f^{n*}[t + 1] = \left[ \theta_f^{n*}[t] - \delta \theta[t] \left( \zeta_{k_n}^n - \Gamma_{k_f, f}^n P_{k_f, f}^n g_{k_n, f}^n \right) \right]^+, \forall n
\] (3.33)

where \(\delta \mu, \delta \lambda,\) and \(\delta \theta\) are the step sizes. Since the problem in (3.29) is convex, the gap between the lower bound offered by (3.30) and the upper bound offered by (3.32), referred to as the duality gap, will diminish with iterations if the chosen step sizes satisfy the conditions discussed earlier [7].

### 3.4 Stochastic Optimization

#### 3.4.1 Introduction

Consider a linear constrained optimization problem as follows:

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad Ax \leq b \\
& \quad x \geq 0.
\end{align*}
\] (3.34)

In all of the optimization models discussed so far, except the variable vector \(x\), we assume that all the parameters involved in the objective function and the constraints are always fixed. In practice, this may not be the case. The input parameters may vary, and in some cases, the fluctuation is noticeable and the result may not be feasible anymore. For instance, consider the following simple problem in a wireless communication system.

**Example 18** There are \(n\) mobile users sharing a common transmission channel for uplink communication with a cellular BS. Each user \(i\) (where \(i = 1, \ldots, m\)) transmits to the BS using transmit power \(p_i\). The channel attenuation factor \(h_i\) for user \(i\) includes both shadowing and multi-path fading. When the BS receives a signal from user \(i\), since all the users transmit in the same channel, the power received from other users will act as interference. For reliable communication, the SINR for user \(i\) must exceed a threshold value \(\gamma_i\). The problem is what power \(p_i\) should user \(i\) choose in order to minimize the energy consumption while maintaining the reliable SINR.
Since the total power consumption is $p_1 + \cdots + p_m$ and the SINR of user $i$ at the BS is

$$C_i = \frac{h_i p_i}{\sum_{j=1, j\neq i}^{m} h_j p_j + N_i} \geq \gamma_i$$

where $h_i$ is the channel-gain for the link from user $j$ to the BS, the resource allocation problem can be formulated as follows:

$$\text{minimize} \quad p_1 + p_2 + \cdots + p_m$$

$$\text{s.t.} \quad h_i p_i - \gamma_i \sum_{j \neq i}^{m} h_j p_j \geq \gamma_i N_i, \quad i = 1, \ldots, m. \quad (3.35)$$

We can solve the problem above effectively by using well-known techniques such as simplex algorithm or interior point method. But the channel-gains $h_i$ are random variables and their uncertainty range may be large. In practical applications, its average value is used to compute power allocation. But this introduces risks because there may be some realizations of $h_i$ that cause the solution to be no longer valid or even infeasible (i.e., resulting in outage for some users). Robust optimization and chance-constrained optimization are branches of operation research that deal with uncertainty in the input. For simplicity, we only discuss stochastic optimization in linear programming.

For the example above, we want to make sure that the constraint for $i$ is satisfied with probability $1 - q_i$. The new (chance-constrained) problem can then be written as follows:

$$\text{min} \quad p_1 + p_2 + \cdots + p_m$$

$$\text{s.t.} \quad \Pr \left( h_i p_i - \gamma_i \sum_{j \neq i}^{m} h_j p_j \leq \gamma_i N_i \right) \leq q_i, \quad i = 1, \ldots, m \quad (3.36)$$

$$p_i \geq 0, \quad i = 1, \ldots, m.$$

Usually $h_i$s are independent identical random variables. In some special cases, we can derive these probabilities in closed form and solve the new system to obtain the transmit power. The new optimization problem will guarantee a small violation probability of the constraints in exchange for higher transmit power.

### 3.4.2 Robust Optimization

Robust optimization is an extreme case of chance-constrained problem where the solution is immune to the uncertainty of the data input. Consider the following linear programming problem:

$$\text{min} \quad c^T x$$

$$\text{s.t.} \quad a_i^T x \leq b_i, \quad i = 1, \ldots, M.$$
The vector \( a_i \) belongs to the uncertainty set \( A_i \). We consider two special forms of \( A_i \).

The uncertainty set is a box: In this case, we can model vector \( a_i = (a_{i1}, \ldots, a_{iN}) \) as 
\[ a_i = \bar{a}_i + \xi_i^T \hat{a}_i, \]
where \( \xi_i = (\xi_{i1}, \ldots, \xi_{iN}) \) is a random vector and \(-1 \leq \xi_{ij} \leq 1\). Denote the set of all \( \xi_i \) as \( Z_i \).

Replacing the vector \( a_i \) back into the constraint, we have
\[ \bar{a}_i^T x + \xi_i^T \hat{a}_i^T x \leq b_i, \quad \forall \xi_i \in Z_i. \] (3.37)

The new robust linear optimization problem can be expressed as follows:
\[
\begin{align*}
\min_{x,y} & \quad c^T x \\
\text{s.t.} & \quad \bar{a}_i^T x + \hat{a}_i^T y_i \leq b_i, \quad i = 1, \ldots, M \\
& \quad -y \leq x \leq y \\
& \quad y \geq 0.
\end{align*}
\]

It is easy to show that if \( x^* \) is the optimal solution for this problem, then \( y^* \) is the vector of absolute values of elements in \( x \). The optimal solution is “robust” against the uncertainty set \( A_i \).

Modeling the uncertainty set as a box is simple but the resulting robust solution is usually very pessimistic, and therefore, may not be desirable. We may allow the constraints to be violated with a small uncertainty instead. If the parameter \( a_{ij} \) is a random variable and has symmetrical distribution, the problem in (3.37) can be converted to the following [3]:
\[
\begin{align*}
\min_{x,y} & \quad c^T x \\
\text{s.t.} & \quad \sum_{j=1}^N \hat{a}_{ij} x_j + \Omega_i \sqrt{\sum_{j=1}^N \hat{a}_{ij}^2 y_{ij}^2} \leq b_i, \quad i = 1, \ldots, M \\
& \quad -y_{ij} \leq x_j - z_{ij} \leq y_{ij} \\
& \quad y_{ij} \geq 0, \quad 1 \leq i \leq M, \quad 1 \leq j \leq N.
\end{align*}
\]

It can be proved that the solution of this new non-linear optimization is robust where the probability that the \( i \)-th constraint is violated is at most \( \exp(-\Omega_i^2/2) \). Thus by choosing a suitable \( \Omega \), we can obtain the solution that satisfies our requirement.

The uncertainty set is an ellipsoid: Another way to model uncertainty, which is less pessimistic, is to use the ellipsoid model [3]. In this case, for the \( i \)-th constraint, the uncertainty vector \( \xi_i \) belongs to a ball \( Z \) with radius \( \Omega_i \), i.e., \( \sqrt{\sum_{j=1}^N \xi_{ij}^2} \leq \Omega_i \). Assume \( x = (x_1, \ldots, x_N) \) to be the optimal robust solution. Since the \( i \)-th constraint must be
satisfied for all values of $\zeta$, we must have
\[
\sum_{j=1}^{N} \bar{a}_{ij}x_j + \sum_{j=1}^{N} \zeta_{ij}\hat{a}_{ij}x_j \leq b_i, \quad \forall \zeta_i \in \mathbb{Z}_i
\]
\[
\Leftrightarrow \max_{\zeta \in \mathbb{Z}_i} \sum_{j=1}^{N} \zeta_{ij}\hat{a}_{ij}x_j \leq b_i - \sum_{j=1}^{N} \bar{a}_{ij}x_j
\]
\[
\Leftrightarrow \Omega_i \sqrt{\sum_{j=1}^{N} \hat{a}_{ij}^2x_j^2} \leq b_i - \sum_{j=1}^{N} \bar{a}_{ij}x_j.
\]

The final result is obtained by using the Cauchy-Schwarz inequality. Since the new constraint is convex [3, Chapter 1], the new problem is then a convex conic quadratic optimization and can be solved using second order conic programming algorithms, i.e., interior point methods. By modeling the uncertainty by an ellipsoid, we obtain a less conservative result than that obtained using boxes for the uncertainty sets.

### 3.5 Dynamic Programming

#### 3.5.1 Introduction

Dynamic optimization or dynamic programming is a technique that transforms a complex problem into a sequence of simpler problems. In this kind of a problem, a decision is made at each stage, and we must find an optimal sequence (or policy) to minimize the total cost. For the stochastic case, we do not have the full information about the system at first but it is revealed over time. At each stage, the dynamic programming algorithm chooses the decision based on the sum of the present cost and the expected future cost, assuming that optimal decisions are made in the subsequent stages. Dynamic programming finds a large number of applications in control theory. A well-known example application of dynamic programming is the Bellman-Ford algorithm to find the shortest path between two nodes.

The general structure of a dynamic programming is described below.

- **System model**
  - The problem is discretized into $N$ stages. Each stage is solved as an ordinary optimization problem and its solution is used in the next stage of the problem. Often, the stages are understood as different time epochs of the planning horizon.
  - $x_k$ is the state of the system. Each stage may have a finite or an infinite number of states. A state represents the information that is relevant for future optimization. The number of different states should be small because it will affect the computational complexity of the problem. Defining the stages and states is an important task in solving a problem by using dynamic programming.
  - $u_k$ is the decision that has to be made at stage $k$ and usually must satisfy some constraints. For discrete cases, $u_k$ often belongs to a given finite set. For stochastic
cases, a feedback system can be used to predict the future behaviors of the system and decide the current action. We define a policy at each stage as the set of rules that maps the current state into the action.

- At stage \( n \), given \( u_n \), we define the transition rule \( f \) that the system transforms from state \( x_n \) to state \( x_{n+1} \), i.e., \( x_{n+1} = f(x_n, u_n) \).
- The solution of the problem is an optimal policy that is the collection of policies at each stage such that the overall cost (or profit) is minimized (or maximized).
- Given the current state, the optimal policy for the remaining stages is independent of the policy decisions made in previous stages. In the other words, the optimal decision at each stage depends on the current state but not how the system gets to this state. This is called the principle of optimality.
- For a stochastic system, where the next state is a random variable, we define \( w_n \) to be the random parameters or new information from the environment that comes during stage \( n \). Usually, the distribution of \( w_n \) is available. Then we have \( x_{n+1} = f_n(x_n, u_n, w_n) \).
- From \( u_n, x_n \), and the function \( f_n \), we derive the transition probability that state \( x_n \) transforms to \( x_{n+1} \) (i.e., \( \Pr(x_{n+1} = X | x_n, u_n) \) or \( \Pr(w_n = W | x_{n+1}, x_n, u_n) \)).

- The cost function \( g \) is defined for each state \( n \). We want to minimize the total cost over \( N \) stages:

\[
\min_{u_1, \ldots, u_{N-1}} \mathbb{E} \left\{ g(x_N) + \sum_{n=0}^{N-1} g_k(x_n, u_n, w_n) \right\}. \tag{3.38}
\]

- In dynamic programming, usually we find the optimal policy for the final stage and move backward and find the optimal policy of the previous stage based on the stages we have already solved. To do that, we need to find the recursive relationship that identifies the optimal decision for stage \( n \) given the optimal policy for stage \( n+1 \) is available,

\[
v_n^*(x_n) = \max_{u_n} (g(u_n, x_n, w_n) + v_{n+1}^*(x_{n+1})) \tag{3.39}
\]

where \( x_n \) is the state of stage \( n \), \( u_n \) is the action at stage \( n \) and \( x_{n+1} = f_n(x_n, u_n, w_n) \) is the state of stage \( n+1 \) if the system uses action \( u_n \) at state \( x_n \) and the random variable is \( w_n \). Finally, \( v_n^*(x_n) \) is the contribution from stage \( n \) to \( N \) if the system starts at state \( x_n \) of stage \( n \).

- We assume that the system has Markov property, that is, the next state depends only on the current state. The random parameter does not depend on its past value \( w_1, \ldots, w_{n-1} \) but may depend on \( x_n \) and \( u_n \). Usually, given \( u_n \) and \( x_n \), we can derive the distribution of the next \( w_n \) as \( \Pr(w_n | x_n, u_n) \). For simplicity, in many cases, we assume that \( w_n \) is a sequence of independent and identically distributed random variables, which means \( w_n \) is independent of both \( u_n \) and \( x_n \).
3.5 Dynamic Programming

3.5.2 Examples of Dynamic Programming

Inventory Control Problem

We illustrate dynamic programming using a simple problem where dynamic programming can be applied [4].

Example 19 Consider a problem of ordering a certain quantity of a certain item at each of \( N \) periods to meet a (stochastic) demand, while minimizing the incurred expected cost. Let us use the following notations:

- \( x_k \) is the stock available at the beginning of the \( k \)-th period.
- \( u_k \) is the stock ordered (and immediately delivered) at the beginning of the \( k \)-th period; for simplicity, we can allow negative balance, i.e., \( x_k < u_k \).
- \( w_k \) is the demand during the \( k \)-th period with given probability distribution.
- \( g(x_k, u_k, w_k) \) is the cost function during time \( k \). One simple cost function is \( g(x_k, u_k, w_k) = cu_k + r(x_k + u_k - w_k) \), which is the profit of selling \( u_k \) units and the cost of keeping the remaining.
- The state transition function \( f \) is \( x_{k+1} = f(x_k, u_k, w_k) = x_k + u_k - w_k \).
- Assume that the stock available at stage 0 is \( x_0 \).

We want to choose an optimal policy for \( u_k \) such that the total cost over \( N \) stages is minimized.

Assuming that the stock available \( x_k \) are non-negative integers, we can use them as the states for this problem. Denote by \( \{u^*_0, \ldots, u^*_N\} \) the optimal stock orders over \( N \) stages. Using the principle of optimality, we divide the problem into “tail subproblems” whereby we are at state \( x_k \) and want to minimize the future cost from time \( k \) to \( N \).

Define a value function \( v^*_k(x_k) \) which is the optimal value of all subsequent decisions, given that we are in stage \( k \) and the state is \( x_k \). The * notation implies optimal value. From the principle of optimality, we have

\[
v^*_k(x_k) = \min_{u_k} \mathbb{E}[g(x_k, u_k, w_k) + g(x_{k+1}, u_{k+1}, w_{k+1}) + \cdots + g(x_{N-1}, u_{N-1}, w_{N-1})]\]
\[
= \min_{u_k} \mathbb{E}[g(x_k, u_k, w_k) + v^*_{k+1}(x_{k+1})]
\]
\[
= \min_{u_k} \mathbb{E}[g(x_k, u_k, w_k) + v^*_{k+1}(x_k + u_k - w_k)]
\]
\[
= \min_{u_k} \mathbb{E}[g(x_k, u_k, w_k)] + \min_{u_k} v_{k+1}(x_k + u_k - w_k).
\]

The optimal value we want to find is \( v^*_0(x_0) \). We devise a backward recursion algorithm as follows:

Tail subproblem of length 1:

\[
v_{N-1}(x_{N-1}) = \min_{u_{N-1}} \mathbb{E}[cu_{N-1} + r(x_{N-1} + u_{N-1} - w_{N-1})].
\]
Tail subproblem of length $N - k$:

$$v_k(x_k) = \min_{u_k} \mathbb{E}[c u_k + r(x_k + u_k - w_k) + v_{k+1}(x_k + u_k - w_k)]. \tag{3.42}$$

Without loss of generality, assume that at stage $N$, $v_N(x_N)$ is zero, then we can perform backward recursion to find the optimal cost $v^*_N(x_0)$.

At step $N - k$ we solve the tail subproblem with length $k$ by using results of $v_{k+1}(x)$ obtained from tail subproblem with length $k + 1$. Then we use these values to calculate the next tail subproblem with length $k - 1$. It can be proved that the value function obtained using backward recursion is also the optimal value, i.e., $v_k(x_k) = v^*_k(x_k)$. The example shows that dynamic programming is actually a combination of greedy algorithm and brute-force search. At each step, we calculate all possible values for this stage and then keep the best for later steps.

**Linear Programming**

We illustrate the steps of dynamic programming by solving the following linear program [5]:

$$\max \ 4x_1 + 5x_2,$$

s.t.  
1. $0 \leq x_1 \leq 4$
2. $0 \leq x_2 \leq 6$
3. $3x_1 + 2x_2 \leq 18$. \tag{3.43}

First, we should choose the number of stages to be 2, which corresponds to the number of variables $x_1$ and $x_2$. The decision of the problem corresponds to the values of $x_1$ and $x_2$. Next, we need to decide the state that is the information required to make the decision at each stage. In this case, we define the state as the amount of slack left in the constraints at the beginning of each stage. For the first stage, the amount of slack (or state) is $s_1 = (4, 6, 18)$, which corresponds to the right-hand side of the constraints. For the second stage, the amount of slack will be $s_2 = (4 - x_1, 6, 18 - 3x_1)$.

In this case, the state is a vector with three variables, or we say the dimension of the state is 3. Since we need to consider all possible combinations of values of all state variables, a large number of dimensions for state vector will cause the complexity to increase rapidly. This phenomenon is called “curse of dimensionality.” However, in our case, the problem is simple enough and can be solved efficiently.

Next, we define $v_n(s_n, x_n)$ as the contribution of the action $x_n$ to the objective function if the system starts at state $n$ and then optimal decision is made at stage $n + 1$ (if $n < 2$).

- If $n = 2$, $v_2(s_2, x_2) = 5x_2$.
- If $n = 1$, $v_1(s_1, x_1) = 3x_1 + \max_{0 \leq x_2 \leq 6, \ x_2 \leq 18 - 3x_1} 5x_2$.

Also, define $v^*_n(s_n)$ as the optimal contribution from stage $n$ onward. Clearly, we have

$$v^*_n(s_n) = \max_{x_n} v_n(s_n, x_n). \tag{3.44}$$

Since at stage 2 we do not know the value of $x_1$, we use the general term $(R_1, R_2, R_3)$ to represent the state instead. We can use $(4 - x_1, 6, 18 - x_1)$ to represent stage 2;
however, if the problem has more than two stages and we are solving the problem backward, it is complicated to express the final stage using the previous decisions. First, we find the optimal decision at stage 2, given the action \( x_1 \). Then we use this result to solve stage 1 and obtain \( x_1 \). The details are as follows:

**Stage 1:** At stage 2, the state is \( s_2 = (R_1, R_2, R_3) \). Assuming that all of them are non-negative (otherwise the problem is infeasible), we have

\[
v_2^*(s_2) = \max_{x_2} v(s_2, x_2) = \max_{0 \leq x_2 \leq R_2} 5x_2 = \min\{R_2, R_3/2\}.
\]

(3.45)

**Stage 2:** Next we replace \((R_1, R_2, R_3)\) by \((4 - x_1, 6, 18 - 3x_1)\) and plug the solution obtained for \( v_2^*(s_2) \) into \( v_1^* \) to find \( x_1 \):

\[
v_1^*(s_1) = \max_{x_1} v(s_1, x_1) = \max_{0 \leq x_1 \leq 4} \max_{3x_1 \leq 18} 4x_1 + v_2^*(4 - x_1, 6, 18 - 3x_1)
\]

\[
= \max_{0 \leq x_1 \leq 4} \max_{3x_1 \leq 18} 4x_1 + 5 \min\{6, 9 - 3x_1/2\}.
\]

(3.46)

We consider two cases, namely, \( x_1 \geq 2 \) and \( x_1 < 2 \) to obtain the maximum value of \( v_2^* \) as 38 and \( x_1 = 2 \). From that we obtain \( x_2 = \min\{R_2, R_3/2\} = \min\{6, 9 - 3x_1/2\} = 6 \).

We can apply normal linear optimization algorithms such as simplex method to find that the solution is the same as dynamic programming method. However, when the problem requires integer solutions, dynamic programming is very useful since it does not require continuity in the solution as in linear programming.

### 3.6 Exercises

#### Basics of Optimization

**Exercise 3.1:** Prove that if \( f, g \) are convex, monotonic, and positive, then the product \( fg \) is convex. Similarly, if \( f \) is positive, convex, and increasing while \( g \) is positive, concave, and decreasing, then \( f/g \) is convex.

**Exercise 3.2:** Let \( f: \mathbb{R} \to \mathbb{R} \) be a twice differentiable function. Prove that \( f \) is convex if and only if either of the conditions below is satisfied.

- (First order condition): \( f(y) \geq f(x) + f'(x)(y - x) \) \( \forall x, y \in \mathbb{R} \)
- (Second order condition): \( f''(x) \geq 0 \).

Hint: For the first condition, apply the definition of convexity for \( z = x + t(x - y) \). For the second condition, use Taylor’s theorem.

**Exercise 3.3:** Prove that, if function \( f: \Omega \to \mathbb{R} \) is strongly convex on its domain, i.e., there exists \( m > 0 \) such that \( \nabla^2 f(x) \succeq mI \) \( \forall x \in \Omega \), then we have

\[
f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}||y - x||^2, \quad \forall x, y.
\]

Use this result to prove the stopping condition of the gradient descent method:

\[
||\nabla f(x)|| \leq (2me)^{1/2} \Rightarrow f(x) - p^* \leq \epsilon.
\]
Hint: Use Taylor’s theorem to prove the first problem and find the minimum of its right-hand side with respect to $y$.

**Exercise 3.4:** Using gradient descent and exact line search find the optimal value of the quadratic function:

$$
\min_x f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2).
$$

Prove that the optimal solution is $x_1 = 0$, $x_2 = 0$.

**Exercise 3.5:** Prove that if $x^*$ is the stationary point of a convex differentiable function, then $f(x^*)$ is the global minimizer.

**Exercise 3.6:** Prove that, if $f$ is strongly convex and twice differentiable and $\nabla^2 f(x) \preceq M/I$, then the algorithm using gradient descent with exact line search will always converge, i.e., $\forall \epsilon > 0\exists k_0$, $f(x_k) - p^k \leq \epsilon$ for all $k > k_0$.

**Exercise 3.7:** Let $f : \mathbb{R} \rightarrow \mathbb{R}$ $f(x) = (x - a)^4$ where $a \in \mathbb{R}$. Apply Newton’s method to find the minimizer of $f$.

i. Write the update equation.
ii. Let $y_k = |x_k - a|$ where $x_k$ is the $k$-iterate of the Newton’s method. Show that the sequence $y_{k+1} = \frac{2}{3}y_k$. From that find the optimal point of $f$.

**Convex Optimization**

**Exercise 3.8:** Consider the following problem:

$$
\begin{align*}
\min_x & \quad ||Ax - b|| \\
\text{s.t.} & \quad x^T 1 = 1 \\
& \quad x_i \geq 0.
\end{align*}
$$

Prove that this problem is convex.

**Exercise 3.9:** Consider the following problem:

$$
\begin{align*}
\min_x & \quad \mathbb{E}_u[f_0(x, u)] \\
\text{s.t.} & \quad \mathbb{E}_u[f_i(x, u)] \quad i = 1, \ldots, m.
\end{align*}
$$

If $f_0$ and $f_i$ convex, prove that this problem is also convex.
Exercise 3.10: Consider the following quadratic problem:

\[
\min_x f(x) = (1/2)x^T P x + q^T x
\]

where \( P \in \mathbb{R}^n \) is a symmetric matrix.

i. Show that if \( P \) is not a semidefinite matrix, then the problem is unbounded below.

ii. Suppose \( P \) is a semidefinite matrix, but if \( P x = -q \) does not have any solution, then the problem is unbounded below.

Hint: If \( P \) is not semidefinite, then \( \exists v, v^T P v < 0 \).

Exercise 3.11: Consider the following convex problem:

\[
\min_x x_1^2 + x_2^2 + x_3^2 + x_4^2
\]

s.t. \( x_1 + x_2 + x_3 + x_4 = 1 \).

Use KKT conditions to find the optimal solution.

Hint: The solution is \( 1/4, 1/4, 1/4, 1/4 \).

Exercise 3.12: Consider the following problem:

\[
\min_x tf_0(x) - \sum_{i=1}^m \log(-f_i(x))
\]

s.t. \( Ax = b \)

with the central path \( x^*(t) \). Given \( u > p^* \), let \( z^*(u) \) is the solution of

\[
\min_x - \log(u - f_0(x)) - \sum_{i=1}^m \log(-f_i(x))
\]

s.t. \( Ax = b \).

Show that the set of \( z^*(u) \) is also the central path of the first problem.

Hint: For any \( u > p^* \) find \( t \) such that \( x^*(t) = z^*(u) \).

Exercise 3.13: Consider the log-barrier problem:

\[
\min_x tf_0(x) + \phi(x)
\]

s.t. \( Ax = b \)

where \( \phi(x) = \sum_{i=1}^m \log(-f_i(x)) \). Denote \( x^* \) is the optimal solution for the log-barrier problem and \( p^* \) is the solution of the original problem. Prove that \( f_0(x^*) - p^* \leq m/t \).

Hint: Write the KKT conditions for this problem, define \( \lambda^*_i = \frac{-1}{f_i'(x^*)} \), and show that \( \lambda^*, v^*, x^* \) are also optimal solutions of the original dual problem.

Exercise 3.14: Consider the following convex optimization problem:

\[
\min_x f(x)
\]

s.t. \( Ax = b \).

Prove that, for projected subgradient method, if the starting point \( x_0 \) is feasible then for any \( k \geq 0 \), \( x_0 \) is also feasible, i.e., \( Ax_k = b \).

Hint: Prove by induction.
Integer Programming

Branch and Bound Method

Exercise 3.15: Solve the following integer programming problem:

$$\begin{align*}
\text{max} \quad U &= 5x_1 + 4x_2 \\
\text{s.t.} \quad x_1 + x_2 &\leq 5 \\
10x_1 + 6x_2 &\leq 45 \\
x_1, x_2 &\in \mathbb{N}.
\end{align*}$$

Hint: Assume $x_1$ and $x_2$ have real values. Apply linear programming to obtain the solution. From that, choose the variable that is not an integer and perform branching, i.e., add constraints using the nearest smaller integer and nearest larger integer of this variable.

Exercise 3.16: Solve the following integer problem using branch and bound:

$$\begin{align*}
\text{min} \quad x_1^2 + (x_3 - 1)^2 + (x_2 - x_3 + 2)^2 \\
\text{s.t.} \quad x_1 + x_2 + x_3 &= 5 \\
x_i &\geq 0 \quad \text{and} \quad x_i \in \mathbb{Z}.
\end{align*}$$

Hint: Assuming that $x_i$ can take real values, this problem is a convex quadratic optimization. You can use \textit{quadprog} function in MATLAB to solve the relaxed version efficiently and obtain the lower bound. The answer is $(1, 1, 3)$.

Cutting Plane Method

Exercise 3.17: Prove that the extra cut $x_1 + 2x_2 \geq 5$ is a valid constraint for the following integer linear program:

$$\begin{align*}
\text{min} \quad x_1 + 2x_2 \\
\text{s.t.} \quad x_1 + 5x_2 &\geq 5 \\
x_1 + x_2 &\geq 4 \\
x_1, x_2 &\geq 0, \quad x_1, x_2 \in \mathbb{Z}.
\end{align*}$$

From here can you guess the optimal solution.

Exercise 3.18: Apply the cutting plane method to solve the following integer linear program:

$$\begin{align*}
\text{max} \quad f(x_1, x_2) &= 7x_1 + 10x_2 \\
\text{s.t.} \quad -x_1 + 3x_2 &\leq 6 \\
7x_1 + x_2 &\leq 35 \\
x_1, x_2 &\geq 0, \quad x_1, x_2 \in \mathbb{Z}.
\end{align*}$$
Hint: Denote \( x_3 \) and \( x_4 \) as slack variables. Solve this linear program by using simplex method. Keep adding new constraints based on the solutions using Gomory’s method until an integer solution is obtained. The solution is \((x_1, x_2, x_3, x_4) = (4, 3, 1, 4)\).

**Stochastic Optimization**

**Exercise 3.19:** (Lobo 1998) Consider the following stochastic linear programming problem:

\[
\begin{align*}
\text{min} & \quad c^T x \\
\text{s.t.} & \quad \Pr(a^T x \leq b) \geq p.
\end{align*}
\]

Assume that vector \( a \) is a Gaussian random vector with mean \( \bar{a} \), covariance matrix \( \Sigma \), and \( p \leq 0.5 \). Prove that this problem is equivalent to the following second-order conic program:

\[
\begin{align*}
\text{min} & \quad c^T x \\
\text{s.t.} & \quad a^T x + \Phi^{-1}(p)||\Sigma^{1/2}x||_2 \leq b
\end{align*}
\]

where \( \Phi^{-1} \) is the inverse of the normal cumulative distribution function, i.e., \( \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt \).

Hint: Denote \( u = a^T x \), this is also a Gaussian random variable. Using the definition of normal distribution, we can obtain the probability in closed form. Since \( p \geq 0.5 \), we have \( \Phi^{-1}(p) \geq 0 \).

**Exercise 3.20:** Prove that if \( x_1, x_2, \ldots, x_N \) are independent exponential distributed random variables with mean \( \mathbb{E}[x_i] = 1/\lambda_i \), then

\[
\Pr\left(x_1 \leq \sum_{i=2}^{N} x_i\right) = 1 - \prod_{i=2}^{N} \left(1 + \frac{\lambda_i}{\lambda_1}\right).
\]

Hint: Use formula \( \Pr(x > y) = \int_{t=0}^{\infty} \Pr(x > t) f_y(t) dt \). This formula can be used to calculate the outage probability at a BS assuming Rayleigh fading channels.

**Exercise 3.21:** Consider the following inequality:

\[
f \leq \sum_{m=1}^{M} (a^T_m c) p_m.
\]

Vector \( c = (c_1, \ldots, c_N) \) is a random vector with covariance matrix \( G \) and mean \( \bar{c} \). Prove that if \( c \) belongs to the uncertainty set \( U = \{\bar{c} + G^{1/2} u, ||u||_2 \leq \Omega\} \), then the robust solution satisfies

\[
f \leq \sum_{m=1}^{M} \left(a^T_m \bar{c} - \Omega \sqrt{a^T_m Ga_m}\right) p_m.
\]
Hint: Use Cauchy-Schwarz inequality. Note that $a_m^T \bar{c}$ and $\sqrt{a_m^T G a_m}$ are actually the mean and standard deviation of $a_m^T c$. This problem shows that ellipsoid is a natural way to model the uncertainty set.

Dynamic Programming

Exercise 3.22: Using dynamic programming, write the steps and the recursive formula required to solve the Knapsack problem below:

$$\min_{m} \ z = r_1 m_1 + r_2 m_2 + \cdots + r_n m_n$$
$$\text{s.t. } w_1 m_1 + \cdots + w_n m_n \leq W \quad (3.52)$$
$$m_i \in \mathbb{N}$$

where $m_i$ is the number of units, and $r_i$ is the weight of each unit of item $i$.

Hint: Define stages as the index of item, state $x_i$ as the total weight assigned to item from $i$ to $n$. Contribution at stage $i$ is the function $f_i(x_i)$ which is the minimum total weight for items from $i$ to $n$. Show that the return at state $i$ is a function of $x_i$ only.

Exercise 3.23: A traveling salesmen needs to travel from one location A to another location H by passing some other locations in between. Given the duration of traveling between two locations, find his minimum time to reach H from A.

Exercise 3.24: Maximize the following non-linear programming problem using dynamic programming:

$$\max \ U = x_1 (1 - x_2) x_3 \quad (3.53)$$
$$\text{s.t. } x_1 - x_2 + x_3 \leq 1 \quad (3.54)$$
$$x_i \geq 0. \quad (3.55)$$

Hint: Define the state of the problem as the remaining resources in the constraint when each variable is introduced, i.e., $S_1 = 1, S_2 = 1 - x_1, S_3 = 1 - x_1 + x_2$. Apply backward recursion for the objective function. The optimal solution is $(2/3, 2/3, 2/3)$. 

Exercise 3.25: Using dynamic programming, solve the following nonlinear integer problem:

\[
\min_U \quad U = (x_1 + 2)^2 + x_2x_3 + (x_4 - 5)^2 \\
\text{s.t.} \\
x_1 + x_2 + x_3 + x_4 \leq 5 \\
x_i \geq 0, \quad x_i \in \mathbb{Z}.
\]  

(3.56) (3.57) (3.58)

Hint: Define the state of the problem as in the previous exercise. Group \( x_2 \) and \( x_3 \) into one state. The solution is \((0, 0, 0, 5)\).

References

4 Game Theory

4.1 Fundamentals of Game Theory

4.1.1 Brief History

In a typical optimization problem, we need to maximize/minimize an objective function by controlling the values of a vector that satisfies a set of constraints. In this case there is only one party that controls the system, and its actions do not depend or are not affected by other parties. However, in practice, there are many situations in which we must make decisions to optimize an objective function in presence of other parties, and their actions can change the outcome we expect. The information about the decisions of other parties may or may not be available to us at the time we make our decisions or moves. Since each party has its own objective and is usually selfish, it will try to maximize its benefit. In such a case, the solution of a normal optimization problem may not result in the best profit for every party. If any party thinks it can achieve a better payoff, it will act alone, and thus, the solution may not be useful. Therefore, we may wish to find a solution (i.e., an equilibrium) that everyone is satisfied with and hence does not want to move. Game theory is able to provide such a solution. It is a branch of applied mathematics that “uses models to study interactions with incentive structures” among different decision-makers. In game theory, we need to anticipate the opponents’ moves and reply with the best action to optimize the objective.

Game theory has a quite young history. It started with the work of Augustin Cournot’s Mathematical Principles of the Theory of Wealth in 1838 where he studied a duopoly using formal game-theoretic analysis. Emile Borel’s series of papers during 1921–1926 defined strategies of a game. In 1944, game theory was established as a separate mathematical field due to the book Theory of Games and Economic Behavior by Von Neumann and Oskar Morgenstern. This book provided much of the basic terminology and problem setup that is still in use today. Then in 1950, John Nash proved that finite games have always have an equilibrium point, at which all players choose actions that are best for them given their opponent’s choices. This definition has great importance in game analysis. Recently, game theory has found many applications in psychology and politics, then economics and engineering with several Nobel prizes awarded to game theorists. (John Nash himself was awarded the Nobel Prize in 1994.)

One classical example of game is the “Prisoners’ Dilemma.” Two criminals are arrested and suspected of a crime they did together. The police, however, have no proof
Table 4.1 Prisoners’ Dilemma

<table>
<thead>
<tr>
<th></th>
<th>Confess (C)</th>
<th>Not confess (NC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confess (C)</td>
<td>(4, 4)</td>
<td>(1, 10)</td>
</tr>
<tr>
<td>Not confess (NC)</td>
<td>(10, 1)</td>
<td>(3, 3)</td>
</tr>
</tbody>
</table>

to convict either of them for this crime. Instead, the police make an offer to each of them: If one testifies against the other, he/she receives a reduced sentence. The prisoners cannot contact each other and therefore one does not know what the other does. If in some ways we can demonstrate the benefit or “payoff” of each prisoner (i.e., player) for different scenarios, then we can analyze this game and predict what action each one will take. The payoff can be shown using Table 4.1, where the rows represent the actions for the first player and the columns represent those for the second player.

If one prisoner does not confess, the other prisoner will have a higher payoff if he/she confesses. If one prisoner decides to confess, then the other also has a higher payoff if he/she chooses to confess. Therefore, assuming that the prisoners do not care about each other and each of them always wants to maximize his/her own benefit (i.e., they are rational), they will tend to confess and receive a reduced sentence. This simple example shows how, using the assumption of rationality of players, we can analyze problems (or games) involving multiple players and predict the outcome of the game. The above example also shows that if each player pursues his/her own self-interest, it leads both players to be worse off than had they cooperated.

Recently, game theory has become a very important mathematical tool, alongside traditional optimization methods, to model and analyze different problems in wireless communications. These problems include distributed channel access, power control, sleep management of mobile devices, etc. A survey of the different game theoretical models used for modeling and analysis of multiple access problems in wireless networks can be found in [1].

4.1.2 Definition of a Game

A game is defined by the following elements: the players of the game, the actions available to each player at each decision point, and the payoffs of each player for each outcome. To predict which actions the players will choose, we need to make the following “rationality” assumption:

**Definition 7** Rational players: Each individual acts rationally in the sense that he/she always chooses the action that gives him/her a better payoff.

Note that the payoff does not need to be limited to money. However, the payoffs must be converted to real numbers so that they can be compared or used to compute the expected value. From the definition of rationality, we see that a player does not try to beat other players. He/she can either help or hurt others depending on how these actions can improve his/her own payoff.
We have the following definition of strategic form games:

**Definition 8 (Strategic Form Games)** A strategic form game is a triplet $G = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ such that

- $N$ is a finite set of players, i.e., $N = 1, \ldots, N$.
- $S_i$ is the set of available actions for player $i$; $s_i \in S_i$ is an action of player $i$. The vector of actions of all players $(s_1, s_2, \ldots, s_N)$ is called the strategy (or action) profile.
- $u_i : S \rightarrow \mathbb{R}$ is a utility function (payoff) of player $i$ where $S = \prod_{i \in N} S_i$ is the set of all action profiles.

Also, we often use the following notations:

- $s_{-i} = [s_j]_{j \neq i}$: vector of actions for all players except $i$.
- $S_{-i} = \prod_{j \neq i} S_j$: set of all action profiles for all players except $i$. Then the vector $(s_i, s_{-i}) \in S$ is a strategy profile.

A strategy of a player is a set of rules that map the information available to him/her to his/her actions. The mapping can be one-to-one if we use pure strategy or one-to-many if we use mixed strategies that involve randomizing.

**Definition 9 (Pure Strategy of a Game)** A pure strategy is a deterministic rule, which determines the only move a player will make for any situation he or she could face.

The strategic form is widely used in game theory because it is simple to understand and easy to model for both discrete and continuous strategy sets. For the prisoners’ dilemma game described above:

- Two prisoners are the players.
- Each player has the same set of actions $\{N, C\}$, where $N$ means “not-confess” and $C$ means “confess.”
- The payoff for each player is given in Table 4.1.

**Definition 10 (Dominant Strategy)** A strategy $s_i \in S_i$ is dominant for player $i$ if $u_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i})$ for all $s'_i \in S_i$ and for all $s_{-i} \in S_{-i}$.

**Definition 11 (Dominant Strategy Equilibrium)** A strategy profile $s^*$ is the dominant strategy equilibrium if for each player $i$, $s_i^*$ is a dominant strategy.

Clearly, for prisoner $i$, “confess” is the dominant strategy, thus the game has only one dominant strategy equilibrium, which is $(C, C)$.

Since the decisions are controlled by different parties and assuming that they are not cooperating, we are interested in strategies such that an equilibrium is obtained and the system becomes stable. Dominant strategy equilibrium, if it exists, is one of the best choices since all players choose their best actions independent of others. However, unlike prisoners’ dilemma, many games do not have dominant strategy equilibrium. A more general form of equilibrium was proposed by John Nash in 1950, which is called the Nash equilibrium.
Definition 12 (Pure Strategy Nash Equilibrium) A (pure strategy) Nash Equilibrium of a strategic game \( \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle \) is a strategy profile \( s^* \in S \) such that for all \( i \in N \)

\[
u_i(s^*_i, s^*_{-i}) \geq u_i(s_i, s^*_{-i}), \quad \forall s_i \in S_i.
\]

If a strategy profile is Nash equilibrium, then no player can profitably deviate given that the strategies of the other players remain unchanged.

Definition 13 The best response function \( b_i \) of player \( i \) to strategy profile \( s_{-i} \) refers to the strategies that satisfy

\[
b_i(s_i) = \{ s_i \in S_i | u_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i}), \forall s'_i \in S_i \}.
\]

Thus a strategy profile is a Nash equilibrium if the corresponding strategy of each player is his/her best response to the strategies of other players. In the prisoners’ dilemma game described before, there is only one Nash equilibrium, which is \((C, C)\).

Example 20 (Matching Pennies) The game is played between two players, Player A and Player B. Each player has a penny and must secretly turn the penny to head or tail (Table 4.2). The players then reveal their choices simultaneously. If the pennies match (both heads or both tails), Player A keeps both pennies, so he/she wins one from Player B (+1 for A, −1 for B). If the pennies do not match (one head and one tail), Player B keeps both pennies, so he/she receives one from Player A (−1 for A, +1 for B). This is an example of a zero-sum game, where one player’s gain is exactly equal to the other player’s loss. We have the payoff matrix as follows:

<table>
<thead>
<tr>
<th></th>
<th>Head</th>
<th>Tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>(1, -1)</td>
<td>(-1, 1)</td>
</tr>
<tr>
<td>Tail</td>
<td>(-1, 1)</td>
<td>(1, -1)</td>
</tr>
</tbody>
</table>

This game clearly has no Nash equilibrium with pure strategy. That means for each player no strategy can be considered as “good.” However, assume that we allow the users to randomly choose their actions, i.e., a probability is assigned to each action in \( S_i \). Then there exists a (mixed) strategy for each player such that an equilibrium is obtained. For the matching penny game, the equilibrium occurs when each player chooses head or tail with equal probability of 0.5.

Definition 14 (Mixed Strategy of a Game) A mixed strategy is an assignment of a probability to each pure strategy. Given the strategic game \( G \) defined as above, for each player \( i \), we denote the set of probability distribution over the set of pure strategies \( S_i \) of player \( i \) as \( \Sigma_i \). Then a mixed strategy \( \sigma_i : S_i \rightarrow \mathbb{R} \) and \( \sigma_i \in \Sigma_i \) is a function that assigns a probability to each \( s_i \in S_i \) such that \( \sigma_i(s_i) \geq 0 \forall s_i \in S_i \) and \( \sum_{s_i \in S_i} \sigma_i(s_i) = 1 \). Denote by \( \sigma \in \Sigma = \prod_{i \in N} \Sigma_i \) the mixed strategy profile for all players. Then, the payoff of player \( i \) is the expected...
value of utility function over the set $S$ of available actions defined as

$$u_i(\sigma) = \sum_{s \in S} \left( \prod_{j=1}^{N} \sigma_j(s_j) \right) u_i(s_i, s_{-i}).$$

Also, the support of a mixed strategy $\sigma_i$ is referred to as the set of pure strategies to which $\sigma_i$ assigns a positive probability. The Nash equilibrium can be extended to include the mixed-strategy case as stated below.

**Definition 15** (Mixed Strategy Nash Equilibrium) A (mixed strategy) $\sigma^*$ of a strategic game $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is a mixed Nash Equilibrium if for each player $i$

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(\sigma_i, \sigma_{-i}^*), \quad \forall \sigma_i \in \Sigma_i.$$  

In order to find a mixed Nash equilibrium, we can use the following proposition.

**Proposition 21** A mixed strategy profile $\sigma^*$ is a Nash equilibrium if and only if for each player $i$,

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(s_i, \sigma_{-i}^*), \quad \forall s_i \in S_i.$$  

The matching penny game above has one unique mixed Nash equilibrium where each player chooses head or tail with probability 0.5. In fact, for any strategic form game, a mixed Nash equilibrium always exists.

**Theorem 22** (Nash 1950) Every strategic game with a finite number of players and a finite number of actions per player has a mixed strategy Nash equilibrium.

In a strategic form game, there can be multiple Nash equilibrium. It is important to choose the best equilibrium that fits our objective, i.e., maximizing efficiency or fairness. One important measure of efficiency is based on the concept of Pareto-optimality defined below.

**Definition 16** (Pareto Optimality) A strategy profile $s \in S$ is Pareto-superior to another strategy profile $s' \in S$ if for every player $i \in N$ we have

$$u_i(s_i, s_{-i}) \geq u_i(s'_i, s'_{-i})$$

with strict inequality for at least one player. Accordingly, a strategy profile $s \in S$ is Pareto-optimal if there exists no other strategy profile that is Pareto-superior to $s$.

From the definition, we can see that the strategy profile of a game is Pareto-optimal if there is no other strategy profile that makes every player at least as well off and at least one player strictly better off. Hence, a Pareto-optimal strategy profile cannot be improved without hurting any player. In the prisoners’ dilemma, the strategy profile $(C, C)$ is a Nash equilibrium but not a Pareto-optimal, while the strategy profile $(NC, NC)$ is Pareto-optimal. Therefore, we should pick the Nash equilibrium point that is also Pareto optimal, if that point exists. When a Nash equilibrium is not Pareto optimal, it implies that the players’ payoffs can be increased.
4.2 Non-cooperative Game

4.2.1 Static Game

A static game is one in which a single decision is made by each player and each player has no knowledge of the decision made by the other players before making its own decision. Decisions are made simultaneously (i.e., the order is irrelevant). A static game is also called a one-shot game. The prisoners’ dilemma is an example of static game. Each player can play only once. When the strategies of the game are chosen from a continuous set, it is called a continuous game.

**Definition 17** A continuous game is a game \( \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle \), where \( N \) is a finite set while \( S_i \) are non-empty compact spaces, and \( u_i : S \rightarrow \mathbb{R} \) is a continuous function.

Similar to the Nash theorem in the finite case it can be shown that a mixed strategy Nash equilibrium always exists for a continuous game.

**Theorem 23** (Glicksberg) Every continuous game has a mixed strategy Nash equilibrium.

Consider a simple example for uplink power control in a single-cell CDMA system with \( N \) mobile users and one BS. All users share the same channel and time resources to transmit data to the BS. Each user needs to decide on its transmit power. Intuitively, increasing the transmit power will increase throughput. However, using more power also means higher cost. Also, by increasing its transmit power, each user indirectly causes more interference to other users. To make a balance between power cost and throughput, we can use the following utility function for each user \( i \) [2]:

\[
    u_i(p_i, p_{-i}) = \omega_i \log(1 + \gamma_i) - \lambda_i p_i, \quad p_i \geq 0, \forall i
\]

where \( p_i \) is the transmit power of user \( i \), \( p_{-i} \) is the power profile of all players except player \( i \), and \( \gamma_i \) is the SINR at the BS for user \( i \)'s transmission. Denote the channel-gain from user \( j \) to the BS by \( h_j \). If the additive noise power is constant and equal to \( \sigma^2 \), then the SINR for user \( i \) at the BS is

\[
    \gamma_i = G \frac{h_i p_i}{\sum_{j=1,j \neq i}^N h_j p_j + \sigma^2}
\]

where \( G \) is the spreading gain for the system (assume \( G > 1 \)). The term \( h_i p_i \) is the received power at the BS from user \( i \). Clearly, the interference powers received from other users reduce the SINR and thus the throughput of user \( i \).

The cost function above has two weight factors \( \lambda_i \) and \( \omega_i \), which are the cost per power unit and the profit for one unit of data transferred. If we assume that the battery of mobile users has infinite energy, for the power control problem, we have the static strategic form game described below.
Example 24 (Static Power Control Game) The static uplink CDMA power control game is a triplet \(\langle N, (S_i), (u_i)\rangle\).

- \(N\) is the set of \(N\) users.
- \(S_i\) is the set of actions for user mobile \(i\). In this case, \(S_i = [0, \infty)\).
- \(u_i\) is the utility function for each user, which is given by (4.1).

For this problem, a Nash equilibrium can be obtained.

First, since we can model the problem as a strategic form continuous game, using Glicksberg theorem, a Nash equilibrium exists. One popular method to find the Nash equilibrium for a static game is to investigate the best response function of each player, assuming that all other players’ actions are known. Here, for power profile \(p_{-i}\), the best response function of user \(i\) is

\[
b(p_{-i}) = \{p_i | p_i \geq 0, \quad u_i(p_i, p_{-i}) \geq u_i(p_j', p_{-i}) \quad \forall p_j' \geq 0\}.
\]

Clearly, the best response \(b(p_{-i})\) of user \(i\) is the optimal point of \(u_i\) with respect to \(p_i\). Since \(u_i\) defined in (4.1) is a continuous function of \(p_i\), using first order necessary condition, we have

\[
\nabla_{p_i} u_i = 0.
\]

To avoid lengthy equations, denote the received power at the BS as \(y_i = h_i p_i\) and the interference power as \(y_{-i} = \sum_{j=1, j \neq i}^{N} h_j p_j\). Combining the first order necessary condition with the constraint that the transmit power cannot be negative, we have

\[
p_i^* = b_i(p_{-i}) = \begin{cases} \frac{1}{h_i} \left( a_i - \frac{1}{G} y_{-i} \right), & \text{if } y_{-i} \leq G a_i \\ 0, & \text{otherwise} \end{cases}
\]

where \(a_i = \omega_i h_i - \frac{\sigma^2}{G} > 0\) can be seen as the threshold limit for interference. If interference is too large or the price \(\lambda_i\) is too high such that this \(a_i\) is negative, then it is more economical for player \(i\) to stop transmission. Let us denote by \(p^* = (p_1^*, p_{-1}^*)\) the equilibrium power profile for all players. Then \(p^*\) must be a fixed point and for every \(p_i^*\) equation (4.3) holds. We consider only the users with positive transmit power and remove any user \(i\) with zero transmit power. Replacing \(p^*\) into this, after some algebraic manipulation, we have the following linear equations:

\[
\begin{pmatrix}
1 & \frac{k_2}{G h_1} & \frac{k_3}{G h_1} & \ldots & \frac{k_N}{G h_1} \\
\frac{k_1}{G h_2} & 1 & \frac{k_3}{G h_2} & \ldots & \frac{k_N}{G h_2} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\frac{k_1}{G h_N} & \frac{k_2}{G h_N} & \ldots & 1
\end{pmatrix}
\begin{pmatrix}
p_1^* \\
p_2^* \\
p_N^*
\end{pmatrix}
= \begin{pmatrix}
a_1/h_1 \\
a_2/h_2 \\
a_N/h_N
\end{pmatrix}.
\]

Theorem 25 [2] In the uplink power control game above with \(N\) users, let us index the users using the following rules: \(a_i < a_j \Rightarrow i > j\) and the order being chosen arbitrarily if \(a_i = a_j\). Let \(N^* < N\) be the largest integer satisfying the following inequality:

\[
a_{N^*} \geq \frac{1}{G + N^* - 1} \sum_{i=1}^{N^*} a_i.
\]
Then the power control game admits a unique Nash equilibrium with the following property:

\[
p^*_i = \begin{cases} 
\frac{1}{h_i} \left\{ \frac{G}{G-1} \left[ a_i - \frac{1}{G^{N^*}+1} \sum_{j=1}^{N^*} a_j \right] \right\}, & \text{if } 0 \leq i \leq N^* \\
0, & \text{if } i > N^*.
\end{cases}
\]  

(4.5)

If there is no such \( N^* \), then the unique Nash equilibrium is the case when no user transmits data.

Note that if \( G > 1 \) and \( a_1 > 0 \), we always have \( a_1 > \frac{1}{G} a_1 \), which implies at least one user will transmit.

**Distributed Algorithm for Convergence**

Although there is a unique Nash equilibrium, for each user, obtaining the equilibrium will require a substantial amount of information exchanges. Each player will need to know the values of \( a_i \) from all other players. A distributed algorithm, which requires less information exchange, is desirable. Also, in [2], the authors propose a parallel update scheme using the best response function (4.3) as described below.

Assuming that time is discretized and the channel-gain and noise are constants, the formula to update the transmit power of user \( i \) at time \( k + 1 \) is

\[
p^{(k+1)}_i = \max \left( 0, a_i - \frac{1}{G} \sum_{j=1}^{N} y_j^{(k)} \right)
\]

(4.6)

where \( y_j^{(k)} \) is the total received power at the BS minus the received power from user \( i \) at time \( k \). We can write it in sequential form as

\[
y_i^{(k+1)} = \max \left( 0, a_i - \frac{1}{G} \sum_{j<i}^{N} y_j^{(k)} \right).
\]

(4.7)

It can be shown that the parallel update algorithm will converge to the unique Nash equilibrium from any starting point if the following condition is satisfied:

\[
\frac{N - 1}{G} < 1.
\]

(4.8)

The implementation is easy. Assume that user \( i \) knows its channel-gain \( h_i \). At each time \( k \), the BS sends a feedback to users, which is the total power it received from all users, i.e., \( \sum_{j}^{N} y_j^{(k)} \). Then user \( i \) just needs to remove its own received power from the sum and perform the update.

In [2], the authors also propose a randomized iterated method for power control that converges to the Nash equilibrium. The users update their transmit powers with a
predefined probability \( \pi_i < 1 \). The update rule is

\[
y_i^{(k+1)} = \begin{cases} 
h_i \max \left(0, a_i - \frac{1}{G} \sum_{j \neq i}^{N} y_j^{(k)} \right), & \text{with probability } \pi_i \\
y_i^{(k)}, & \text{with probability } 1 - \pi_i.
\end{cases}
\]

(4.9)

It is shown that the random update scheme converges almost surely to the Nash equilibrium point under the condition that

\[
\frac{N - 1}{G} \bar{\pi} + (1 - \pi) < 1
\]

(4.10)

where \( \bar{\pi} \) and \( \pi \) are the upper bound and lower bound of all \( \pi_i \). The condition here is less strict than (4.8). By choosing an appropriate \( \pi_i \), we can asymptotically achieve the Nash equilibrium after a finite number of steps.

### 4.2.2 Dynamic Game

In a static game, each player has only one chance to make decision and all of them act “simultaneously.” However, for many applications the game may need to be played over a number of time periods, thus making the game dynamic. A game can be dynamic in two ways. First, the interaction between players can be dynamic. One player may have the right to play first, and then others, who observe the leader, need to adapt their own actions (i.e., leader-follower model). Second, a game is dynamic if a one-shot game is repeated a number of times and the players observe the outcome of previous games before playing the game at later stages. This means they can “learn” from the past (i.e., repeated game). Then a dynamic game can be presented as a tree where the root is the start of the game. Each level of the tree is called a stage. A node of the tree, usually denoted by a filled circle, shows one possible scenario of the game. The nodes are connected by the edges, which represent possible sequences of players’ moves.

We say a sequential game has complete information if only one player acts at a time and if each player knows all other players’ actions that are performed before his/her turn. In contrast, a sequential game has incomplete information when some of the players do not know all the previous choices of other players. In many game models for wireless communication scenarios it is common that full information about other players is not available at the time one player makes his/her move.

For a dynamic game, the strategic form is not enough, and we need to have another more general form, which is called the extensive form game.

**Definition 18** (Extensive Form Game) A standard extensive form game consists of

- A finite set of players \( N \).
- A tree representation \( \mathcal{T} = (V, v_0, Z) \), where \( V \) is a set of nodes, \( v_0 \) is the root, \( Z \) is the set of terminal nodes. \( \mathcal{T} \) is also called terminal histories, i.e., a sequence of actions of every user.
- A set of functions for each node \( x \notin Z \):
  - player \( i(x) \) who moves at \( x \)
  - set \( S(x) \) of possible actions at \( x \)
  - successor node \( n(x, s) \) resulting from action \( s \) of \( x \).
4.2 Non-cooperative Game

For each node $v$, $h(v)$ is a sequence of actions taken up to the considered state. For example, in Figure 4.1, history of the first terminal node is $(D,D)$ and $H$ denotes the set of game states.

Payoff function $u_i : Z \rightarrow \mathbb{R}$ assigning payoffs to players when the terminal nodes are reached.

A set of information partition for each player which contains what he/she knows about other players’ moves at the time he/she plays. In a tree representation, a number of nodes can be circled to show which information is available to a player. In Figure 4.1(a), the dotted circle around two nodes at stage 2 means both actions of player 2 are in the same information set and player 1 cannot tell which action player 2 takes. In contrast, in Figure 4.1(b), player 1 knows exactly which point in the tree it will act upon.

We use a simple sequential multiple access game to illustrate the pure strategies of a dynamic game [5]. Two transmitters can either transmit or remain quiet. If they both transmit, the interference is high and both waste power (payoff is negative). If one of them is quiet and the other transmits, then the quiet transmitter receives nothing while the other has the maximum profit. If both of them are quiet, then nobody benefits. Assume that the first one has the higher priority to play first. The other player has lower priority and is the follower. The game tree can be represented as in Figure 4.2. $c$ is the transmit cost and assume that 1 is the benefit if there is no collision.

In the first stage, player 1 can either Transmit (T) or remain Quiet (Q). Player 2 needs to define his/her strategy based on the player 1’s move so he/she has four options: (TT), (TQ), (QT), and (QQ), where TQ means player 2 transmits if player 1 transmits.
and remains quiet if player 1 is quiet. This game has three Nash equilibriums, which are
(T,QT), (T,QQ), and (Q,TT).

**Theorem 26** (Kuhn 1953) Every finite extensive-form game with perfect information
has a pure strategy Nash equilibrium.

The proof is based on the concept of backward induction. This technique is an iterative
technique similar to dynamic programming and very useful to solve perfect inform-
ation sequential game. In backward induction, one first guesses the optimal choice
of player who makes the last move of the game. Then, from this result the preceding
player of the last one decides his/her action. The process continues backward until the
root is reached. In the case of imperfect information, backward induction is not useful
since there can be infinite number of possible cases. Instead, we can use the concept of
subgame equilibrium, which requires the strategy of each player to be optimal not only
from the root but also after every history.

**Definition 19** (Subgame) A subgame of a dynamic non-cooperative game consists of
a single node in the extensive form representation of the game (the game tree) and all
of its successors down to the terminal nodes. The information sets and the payoffs of a
subgame are inherited from the original game. Moreover, the strategies of the players
are restricted to the histories of the subgames.

**Definition 20** (Subgame Equilibrium) A strategy profile \( s \in S \) is a subgame perfect
equilibrium if the strategies of the players (restricted to the subgame) constitute a Nash
equilibrium on every subgame of the original game.

As has been mentioned before, for many cases, there is a hierarchy among the players.
One or more players have the priority to play before others and thus enforce their own
preferences to every one. Such a player is called a leader of the game, and the others are
referred to as followers.

Denote the set of strategies for leader and follower as \( S_1 \) and \( S_2 \), respectively. The set
of best responses of player 2 for strategy \( s_1 \) of player 1 is defined below.

**Definition 21** Given a two-person game with leader-follower model, if \( s_1 \) is the strategy
of player 1, then the best response of player 2 is the set \( B(\mathbf{s}_1) = \{ s_2 \in S_2 : u_2(s_1, s_2) \geq
u_2(s_1, s) \ \forall s \in S_2 \} \).

Nash equilibrium cannot be used in this case because in case of Nash equilibrium,
two strategies need to be considered simultaneously. We need a new type of equilibrium,
namely, the Stackelberg equilibrium, for the sequential case.

**Definition 22** (Stackelberg Equilibrium) In a two-person game with player 1 as leader
and player 2 as the follower, a strategy \( s^*_1 \in S_1 \) is called a Stackelberg equilibrium
strategy for the leader if

\[
\min_{s_2 \in B_2(s^*_1)} u_1(s^*_1, s_2) = \max_{s_1 \in S_1} \min_{s_2 \in B_2(s_1)} u_1(s_1, s_2) \triangleq u^*_1. \tag{4.11}
\]

Using Stackelberg equilibrium implies the leader will try to maximize his/her worst-
case payoff when the follower is hostile and tries to reduce the leader’s payoff. In the
multiple access game in Figure 4.2, the Stackelberg equilibrium strategy for player 1 is $T$ and his/her worst payoff is zero.

**Theorem 27** [3] Every two-person finite game admits a Stackelberg strategy for the leader.

**Example 28** (Power Control in a Two-Tier Macrocell-Small Cell Network) Following [6], we illustrate how the Stackelberg model can be used for downlink power control in a two-tier macrocell-small cell network with one macro base station (MBS) and many small cell base stations (SBSs) sharing a set of orthogonal downlink channels (e.g., an OFDMA scenario). Since the transmit power of MBS is large compared to SBSs, we assume it to be the leader. The MBS needs to find a Stackelberg equilibrium for its transmit power while an SBS needs to devise its best response strategy to both MBS and other SBSs.

The system model can be described as follows:

- One MBS $u_0$ and a set of SBSs $\mathcal{N} = \{u_1, \ldots, u_N\}$.
- All BSs share a set of orthogonal channels $\mathcal{L} = \{1, \ldots, L\}$.
- For any BS $(u_i, u_j)$, where $i, j \in \{1, \ldots, N\}$ and channel $k, l \in \mathcal{L}$, denote by $g_{ij}^{kl}$ the link-gain for channel $l$ of BS $j$ to channel $k$ of receiver of BS $i$. We can assume $g_{ij}^{kl} = 0$ if $k \neq l$.
- For each BS $i$ denote by $p_i = (p_1^i, \ldots, p_L^i)$ its transmit power vector for $L$ channels. We have the following power constraint:

$$p_1^i + \cdots + p_L^i \leq \bar{p}_i. \tag{4.12}$$

- For channel $k$ of receiver $i$, the noise is denoted as $n_k^i$. Then the interference for user $i$ at channel $k$ is $\mu_k^i = \sum_{j \neq i} g_{ij}^k p_j^k$. If we normalize the link-gain such that $g_{ii}^k = 1$, then the total downlink transmission capacity at a user served by BS $i$ with $0 \leq i \leq N$ is

$$C_i = C_i(p_0, \ldots, p_N) = \sum_{k=0}^{L} \log_2 \left(1 + \frac{p_k^i}{v_k^i}\right) \tag{4.13}$$

where $v_k^i = \mu_k^i + n_k^i$.

The Stackelberg strategy of the MBS (i.e., the leader) can be found by solving the following two-stage optimization problem:

$$\max_{p_0} \sum_{k=1}^{L} \log_2 \left(1 + \frac{p_k^0}{\sum_{j=1}^{N} g_{0j}^k p_j^0 + n_0^k}\right) \tag{4.14}$$

s.t. $p_0^k \geq 0, \sum_{k=1}^{L} p_0^k = \bar{p}_0 \tag{4.15}$

$$p_j \in \arg \max_{p_j} \{C_j : p_j^k \geq 0, \sum_{k=1}^{L} p_j^k = \bar{p}_j\} \quad \forall j : N \geq j \geq 0. \tag{4.16}$$
Here (4.14) and (4.15) constitute the upper sub-problem where the MBS finds the best strategy for the Stackelberg equilibrium. And (4.16) is the lower sub-problem where the SBSs find a Nash equilibrium among them. Denote by $A_f$ the set of active channels used by BS $i$ (i.e., the channels with positive transmit power). Given the power vector of the MBS, the best response power transmission for each SBS $i$ can be found by using water-filling algorithm given as follows:

$$p^k_i = \begin{cases} 
0, & \text{if } k \in \mathcal{L} \setminus A_f \\
K_i - v^k_i, & \text{otherwise.}
\end{cases} \quad (4.17)$$

Using the condition $\sum_{k \in \mathcal{L}} p^k_i = \bar{p}_i$, we can calculate the value of Lagrange multiplier $1/K_i$.

For the upper problem, we split the channel set $\mathcal{L}$ into two parts: the first $A_f$ is a set of channels such that at least one of the SBSs uses channel(s) from this set, and the second set $\mathcal{L} \setminus A_f$ consist of the channels that are not used by any SBS. Now we can solve each case separately. If we exclude the lower sub-problem, the upper sub-problem is a convex optimization problem, so we can use Lagrangian method. Denote by $A_l$ the set of channels where the leader transmits and $\mathcal{N}$ is the set of channels where both the MBS and SBSs transmit. We have the following results [6]:

$$p^k_0 = \begin{cases} 
0, & \text{if } k \in \mathcal{L} \setminus A_f \\
\omega - n^k_0, & \text{if } k \in \mathcal{N} \\
2(1 + 2a_k)\omega, & \text{if } k \in \mathcal{N} \quad (4.18)
\end{cases}$$

where $\omega$ is Lagrangian multiplier and $a_k$ is the parameter of the noise-interference function $F(p^k_0)$ at the leader over channel $k$. $F(p^k_0)$ is assumed to be a linear function of $p^k_0$: $F(p^k_0) = a_k p^k_0 + b_k$.

### 4.2.3 Bayesian Game

In the previous sections, we have assumed that all information about the game, which includes the actions and the payoffs, are available to each player. However, in many situations, this assumption is not true. For example, in a competition, rivals will not share information about their strategies and payoffs. One way to model problems with incomplete information is to use Bayesian game models.

In a Bayesian game, a player may have a type (chosen from a set of different types), and this information is available to the others. However, they do not know exactly which type this player has at the time when they make their moves. Denote by $\mathcal{T}_i$ the set of all types for player $i$, $t_i \in \mathcal{T}_i$ is the type of player $i$, and $t_{-i}$ is the vector of other players’ types. Then the belief of player $i$ is a conditional probability mass function (pmf) of the other types of all other players, given its own type, i.e., $Pr(t_{-i}|t_i)$. This belief captures the uncertainty, and this makes the Bayesian model different from a complete information game.
**Definition 23** A Bayesian game model is defined by the following elements [3]:

- A set of players $i \in \mathcal{N} = \{1, \ldots, N\}$.
- A set of actions available for player $i \colon \mathcal{A}_i$ for $i \in \mathcal{N}$ with $a_i \in \mathcal{A}_i$ denoting a typical action for player $i$.
- A set of possible types for player $i$: $\mathcal{T}_i$ for $i \in \mathcal{N}$ with $t_i \in \mathcal{T}_i$ denoting a typical type for player $i$.
- Let $\mathbf{a} = (a_1, \ldots, a_N)$, $\mathbf{t} = (t_1, \ldots, t_N)$ denote the strategy profile and type profile of the game.
- Nature’s move: $\mathbf{t}$ is selected according to pmf $\Pr(\mathbf{t})$ on $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2 \times \cdots \times \mathcal{T}_N$, which induces natural conditional probabilities.
- Strategies: $s_i : \mathcal{T} \to \mathcal{A}_i$ for $i \in \mathcal{N}$ where $s_i(t_i) \in \mathcal{A}_i$ is the action of player $i$ of type $t_i$.
- Payoff $u_i(\mathbf{t}, \mathbf{a})$ of player $i$.

Then the Bayesian Nash equilibrium is defined as the $N$-tuple strategy $\mathbf{s}^* = (s_1^*, s_2^*, \ldots, s_N^*)$ such that for every $i \in \mathcal{N}$,

$$s_i^*(t_i) = \arg \max_{a_i \in \mathcal{A}_i} \sum_{t_{-i} \in \mathcal{T}_{-i}} u_i(s_1^*(t_1), \ldots, s_{i-1}^*(t_{i-1}), a_i, s_{i+1}^*(t_{i+1}), \ldots, s_N^*(t_N); t) \Pr(t_{-i}|t_i)$$  

(4.19)

where $\mathbf{a}_{-i} = (a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$, $\mathbf{t}_{-i} = (t_1, t_2, \ldots, t_{i-1}, t_{i+1}, \ldots, t_N)$. In Nash equilibrium, we maximize the average conditional payoff for player $i$ given his/her own type $t_i$. The Bayesian Nash equilibrium ensures that if the strategies of other players are fixed, then $i$ cannot benefit by deviating from its own strategy. Notice that in the above equations, we consider only pure Nash equilibrium strategies. It is possible to extend the definition to mixed strategies. For the case of mixed strategies, a Bayesian Nash equilibrium always exists.

**Theorem 29** Consider a finite incomplete information (Bayesian) game. Then a mixed strategy Bayesian Nash equilibrium exists.

**Theorem 30** Consider a Bayesian game with continuous strategy spaces and continuous types. If the strategy sets and type sets are compact and the payoff functions are continuous and concave functions, then a pure strategy Bayesian Nash equilibrium exists.

**Example 31** (Bayesian Power Control Game) Bayesian game model can be applied for power control in cellular CDMA or OFDMA networks. In the previous examples we assumed that the channel state information (CSI) at each BS are always available at other BSs. However, this assumption may not be practical, since a significant amount of information exchange will be required for this. Therefore, in [7], the authors propose an uplink power control model in which there are $N$ users transmitting to the BS. It is assumed that each user knows exactly the channel-gain to the BS and for other users it knows only the probability distribution of channel-gains. The transmit power of player $i$ is upper bounded by $P_{\text{max}}^i$. Denote by $g_i$ the channel-gain for user $i$. Since $g_i$ is random and user $i$ does not have any other information, we can view $g_i$ as a type of user $i$. The transmit power is a function of $g_i$ and denoted by $p_i(g_i)$. The achievable data rate, which
is also the payoff of user $i$, is defined as
\[
    u_i = \log_2 \left( 1 + \frac{p_i(g_i)g_i}{\sigma^2 + \sum_{j \neq i} p_j(g_j)g_j} \right).
\]
(4.20)

A Bayesian equilibrium for the uplink power control problem for the users can be obtained.

Since the channel-gain is random, the objective of each user is the solution of the following optimization problem:
\[
    \max_{g} \mathbb{E}_g \left[ \log_2 \left( 1 + \frac{p_i(g_i)g_i}{\sigma^2 + \sum_{j \neq i} p_j(g_j)g_j} \right) \right]
\]
\[\text{s.t.} \quad \mathbb{E}_g [p_i(g_i)] \leq P_i^{\text{max}} \]
\[p_i(g_i) \geq 0 \]
(4.21)

where $g$ is the set of channel-gains for all users. Denote by $g_{-i}$ the set of channel gains for all users except user $i$. For a given strategy $p_{-i} = \{p_1, \ldots, p_{k-1}, p_{k+1}, \ldots, p_N\}$, the single-user maximization problem in (4.21) is a convex optimization problem, and therefore, we can use Lagrangian duality to obtain the result as follows:
\[
    \mathbb{E}_{g_{-i}} \left[ \log_2 \left( 1 + \frac{g_i}{\sigma^2 + p_i(g_i)g_i + \sum_{j \neq i} p_j(g_j)g_j} \right) \right] = \lambda_i
\]
(4.22)

where $\lambda_i$ is the duality variable. In [7], the authors show that a pure strategy Bayesian equilibrium of this problem exists and is unique. The proof is given by proving that the strategy set $P$ and the type set $g$ are compact, while the payoff function $u_i$ is continuous and concave. The uniqueness is proven by using diagonally strictly concave properties of the weighted sum payoff.

Application of a Bayesian game-theoretic model to solve the distributed bandwidth sharing problem among multiple mobile nodes can be found in [8]. In the considered model, the wireless nodes compete for the shared bandwidth from a wireless access point and a mobile node is unable to completely observe other mobile nodes’ behavior (e.g., speed of movement, bandwidth demand). The Bayesian Nash equilibrium is considered as the solution of this game model.

4.2.4 Evolutionary Game

Consider a simple example from biology [9]. The beetles in a population compete with each other over food. The one with a large body gets more food if he/she competes with a smaller beetle. If two beetles with the same size compete, they will get the equal share of food. The payoff table can be built as shown in Table 4.3.

As can be seen from the table, large beetles win over small beetles. However, if two large beetles battle each other, they cannot extract the full amount of resources compared to small ones. Here we try to explain which behaviors will dominate and thus what is the final outcome using an evolutionary game model.
Table 4.3 Body-Size Game

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Big</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>(5, 5)</td>
<td>(1, 8)</td>
</tr>
<tr>
<td>Big</td>
<td>(8, 1)</td>
<td>(3, 3)</td>
</tr>
</tbody>
</table>

Evolutionary games deal with long-term scenarios where players can adapt and respond to the situations based on actual behaviors of other players in a “myopic way,” i.e., their decisions do not involve full computation of optimal strategies for other players and itself. Consider a large population of players. At each instant, a player will be randomly chosen and matched with another player and both play a strategic form game. We define the fitness of a player (beetle) in a population to be the expected payoff it receives from an interaction with a random member of the population. Here we focus on a symmetric game, i.e., a game where the payoff for playing a particular strategy depends only on the other strategies employed, not on who are playing them. The beetle’s fitness then decides the genes of its future offspring (big or small). The fitness can be denoted by \( u(s, s') \), where \( s \) is the strategy of this player and \( s' \) is the strategy of its opponent.

We say that a strategy \( S' \) invades a strategy \( S \) at level \( x \) (where \( x \) is a small positive number) if a fraction \( x \) of the population uses \( S' \) and the rest uses \( S \). The mixed strategy of a random player is defined as \( xS' + (1 - x)S \).

**Evolutionary Stable Strategies**

**Definition 24** The strategy \( S \) is an evolutionary stable strategy (ESS), if for each strategy \( S' \neq S \), there exists \( \hat{x} \in (0, 1) \) such that

\[
u(S, xS' + (1 - x)S) > u(S', xS' + (1 - x)S), \quad \forall x \in (0, \hat{x}). \tag{4.23}\]

From the definition, if the entire population is using ESS, a small group of invaders using different strategies will receive a lower payoff and tend to disappear. In the other words, the ESS cannot be invaded by other strategies. Assume that the payoff function is linear and continuous over \( S' \), i.e., \( u(S, xS' + (1 - x)S) = xu(S, S') + (1 - x)u(S, S') \). By letting \( x \) approach zero, we have \( u(S, S') \geq u(S', S') \), which shows that the ESS is also the mixed strategy Nash equilibrium.

For the body-size game, assume that the population is invaded by big beetles. Denote by \( x \) the fraction of big beetles (\( x \) is a small positive number). For a small beetle, the probability that it meets a small one is \( 1 - x \) and the probability that it meets a big one is \( x \). Therefore, the expected payoff for a small beetle is \( 5(1 - x) + 1 \cdot x = 5 - 4x \). Similarly, the expected payoff for a big beetle is \( 8 - 5x \). Clearly, for small enough \( x \), the expected payoff of a big beetle will be higher. Therefore, “small” is not an ESS. Using the same method, we can check that “big” is an ESS. From Table 4.3, although we see that the population will be better off if all of them are small beetles, the players still choose to become big beetles in the long run.
If an evolutionary game is symmetric and given by Table 4.4, then we can prove that the strategy $S$ is an ESS when either (i) $a > c$, or (ii) $a = c$ and $b > d$.

**Replicator Dynamics**

In the definition of ESS, we assume that all players use the same strategy. We want to find out if a polymorphic population with many types of players can achieve the same stability as in ESS. This will be done by using replicator dynamics. Consider a population $N = 1, \ldots, N$, where everybody can choose a strategy $s \in S$ and $S$ is the set of strategies. Let $n_s(t)$ be the number of agents using strategy $s$ at time $t$, then we have $\sum_{s \in S} n_s(t) = N$. Denote by $x_s(t)$ the fraction of population that uses strategy $s$, i.e., $x_s(t) = n_s(t)/N$. The population state can be defined as a vector $x(t) = [x_1(t), \ldots, x_s(t), \ldots]$ of dimension $S$. We assume that the expected payoff of an agent using strategy $s$ is a function of $u(s, x)$.

Omitting $t$ to simplify the equations, we can find the average payoff for a random player as $\bar{u}(x) = \sum_{s \in S} x_s u(s, x)$. We compare the payoff for a strategy with the average value. Intuitively, an agent will switch to another strategy that leads to a higher payoff, and the bigger the payoff, the faster will be the switching rate. Based on this remark, we have the following replicator dynamics of the population share:

$$
\dot{x}_s = (u(s, x) - \bar{u}(s, x))x_s
$$

where $\dot{x}_s$ is the derivative of $x_s$ with respect to time. Clearly, when the system is stable, $x_s$ should not change and thus $\dot{x}_s = 0$. Then we have the following definition for evolutionary equilibrium.

**Definition 25** (Evolutionary Equilibrium) A fixed point or evolutionary equilibrium of the replicator dynamics is a population that satisfies $\dot{x}_s = 0$, $\forall s \in S$. The fixed points describe populations that are no longer evolving.

**Example 32** Consider the matrix game described by Table 4.5.

If $x_1$ and $x_2$ (where $x_1 + x_2 = 1$) are the fractions of population using $S$ and $S'$, the average payoff is $u(S, x) = x_1 + x_2$ and $u(S', x) = 2x_1$, which give $\bar{u}(x) = x_1^2 + 3x_1x_2$.

---

**Table 4.4** General Symmetric Game

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$S'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>$(a, a)$</td>
<td>$(b, c)$</td>
</tr>
<tr>
<td>$S'$</td>
<td>$(c, b)$</td>
<td>$(d, d)$</td>
</tr>
</tbody>
</table>

---

**Table 4.5** Evolutionary Equilibrium

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$S'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>$(1, 1)$</td>
<td>$(1, 2)$</td>
</tr>
<tr>
<td>$S'$</td>
<td>$(2, 1)$</td>
<td>$(0, 0)$</td>
</tr>
</tbody>
</table>
Applying the formula in (4.24), we have the following fixed points: 
\((x_1, x_2) \in \{(0, 0), (1/2, 1/2), (1, 0)\} \).

Given an initial point of replicator dynamics sufficiently close to an evolutionary equilibrium, this evolutionary equilibrium is stable (robust against small perturbation) in the following two cases [3]:

- The solution path of replicator dynamics will remain arbitrarily close to the equilibrium. This is referred to as Lyapunov stability.
- The solution path of replicator dynamics converges to equilibrium point. This is referred as asymptotic stability.

**Example 33 (Power Control Problem as an Evolutionary Game)** We consider a power control game for CDMA networks [10]. Assume that there is a large population of users that are randomly placed over a plane following a Poisson point process with density \( \lambda \). All users share the same channel. Denote by \( i \) and \( u_i \) the transmitter and its receiver, respectively. A random variable \( r \) represents the distance between a user and its receiver. If \( \zeta(r) \) is the probability density function for \( r \), then the probability that another user is within a circle \( R \) of the receiver is \( \int_0^R \zeta(r)dr \). Any user \( i \) can choose one of two power levels \( P_L \) or \( P_H \) to transmit. The problem can be modeled as an evolutionary game.

Denote by \( x \) the fraction of population choosing \( P_H \) and \( 1 - x \) as the fraction choosing \( P_L \). The total interference contributed by all nodes at the receiver is

\[
I(x) = g\lambda \pi (xP_H + (1 - x)P_L) \left( \frac{\eta}{\eta - 2} r_0^{-\eta} - 2(1 - \mu R^{-\eta}) \right) \tag{4.25}
\]

where \( \eta \) is the attenuation coefficient and \( r_0 \) is the radius of reception. The SINR at the receiver of user \( i \) can be written as

\[
\gamma_i(P_i, x, r) = \begin{cases} \frac{gP_i r_0^\beta}{\sigma_i^2 + \beta \zeta(r)}, & \text{if } r \leq r_0 \\ \frac{gP_i r_0^\beta}{\sigma_i^2 + \beta \zeta(r)}, & \text{if } r > r_0. \end{cases} \tag{4.26}
\]

The user's utility function (fitness) is defined as

\[
u(P_i, x) = w_1 \int_0^R \log (1 + \gamma_i(P_i, x, r)) \zeta(r)dr - w_2 P_i \tag{4.27}
\]

where \( w_1 \) and \( w_2 \) are the cost parameters representing the price of one unit of throughput and one unit of power, respectively.

**Proposition 34** Define function \( h : [0, 1] \rightarrow \mathbb{R} \) as [10]

\[
h(x) = \int_0^R \log \left( \frac{1 + \gamma_i(P_H, x, r)}{1 + \gamma_i(P_L, x, r)} \right) \zeta(r)dr. \tag{4.28}
\]

For all density function \( \zeta \), if \( h(1)(w_2/w_1)(P_H - P_L) < h(0) \), there exists a unique ESS \( x^* \), which is given by \( x^* = h^{-1}(w_2/w_1(P_H - P_L)) \).

It is observed that the ESS \( x^* \) decreases when \( w_2 \) increases. This means that the users become less aggressive as the cost per unit of transmit power increases.
The radio access network selection problem in a heterogeneous wireless access environment was modeled using the theory of evolutionary games [11]. In the considered model, groups of users in different service areas compete to share the limited amount of bandwidth in the available wireless access networks. The competition was modeled as a dynamic evolutionary game and the evolutionary equilibrium was considered to be the solution to this game. For a cognitive radio network, the dynamic behavior of unlicensed users in acquiring spectrum access opportunities from licensed users was modeled using an evolutionary game [12].

4.3 Cooperative Game

In economics and politics there are many cases where competitors sit together and try to reach a deal that is acceptable to all of them, knowing that failing to reach a deal will result a worse result. Also, there are cases where players form competitive groups and the players in a group cooperate with each other. In game theory, the first case is called a bargaining process, and the second case is called a coalition game. These games belong to a different branch of game theory, namely, cooperative game theory. We briefly discuss the important concepts and some examples of both kinds of games.

4.3.1 Nash Bargaining Solution

Nash’s Model of Bargaining

Two players are faced with a set \( \mathcal{X} \) of alternatives. The rules are that, if they both agree on some alternative \( x \) in \( \mathcal{X} \), then \( u_i(x) \) will be the outcome. Otherwise (i.e., if they fail to agree on an outcome) there is a fixed disagreement outcome \( d \) that will be the result. The space of feasible outcomes can be defined as

\[
S = \{(u_1(x_1), u_2(x_2)) | x = (x_1, x_2) \in \mathcal{X}\}. \tag{4.29}
\]

Denoting by \( d = (d_1, d_2) \) the disagreement utility where \( d_1 = u_1(D) \) and \( d_2 = u_2(D) \), the problem is defined as a pair \( (S, d) \), where \( S \in \mathbb{R}^2 \) and \( d \in S \) such that \( S \) is a compact and convex set and there exists \( s \in S \) such that \( s_1 > d_1 \) and \( s_2 > d_2 \) (individual rationality, no one will accept a payoff that is lower than his/her guaranteed payoff under disagreement). We assume the complete information hypothesis, when the set \( (S, d) \) is known to both bargainers it provides all the required information to each of them to make a decision. Nash provides a unique solution for the bargaining problem under the following axioms:

- Pareto efficiency: A bargaining solution \( f(S, d) \in S \) is Pareto-efficient if it could not be improved on both players’ advantages, i.e., \( \not\exists s = (s_1, s_2) \in S, s_i > f_i(S, d), i = 1, 2 \).
- Symmetry: If \( (S, d) \) is symmetric (i.e., if \( (s_1, s_2) \in S \) then \( (s_2, s_1) \in S \)), then \( f_1(S, d) = f_2(S, d) \). The bargaining solution will not discriminate among players if they are indistinguishable.
4.3 Cooperative Game

- Independence of irrelevant alternatives: Given bargaining problems \((S, d)\) and \((S', d)\), if \(S\) contains \(S'\) and \(f(S, d) \in S'\), then \(f(S', d) = f(S, d)\).
- Invariance to equivalent utility representation: If we transform \((S, d)\) to \((S', d')\) by taking \(s'_i = a_is_i + b_i\) and \(d'_i = a_id_i + b_i\), then \(f_i(S', d') = a_i f_i(S, d) + b_i\).

**Theorem 35** (Nash’s Unique Bargaining Solution) There exists a unique solution satisfying the four axioms, and this solution is the pair of utilities \((s_1^*, s_2^*)\) that solves the following optimization problem:

\[
\max_{(s_1, s_2)} (s_1 - d_1)(s_2 - d_2) \\
\text{s.t.} \quad (s_1, s_2) \in S, \quad (s_1, s_2) \geq (d_1, d_2)
\]  

where \((s_1 - d_1)(s_2 - d_2)\) is known as the Nash product.

The theorem can be extended to the \(N\)-player case by expanding \(S\) into an \(N\)-dimensional utility space. Then the Nash bargaining solution for \(N\) players is the solution of the following optimization problem:

\[
\max_{(s_1, s_2, \ldots, s_N)} (s_1 - d_1)(s_2 - d_2) \cdots (s_N - d_N) \\
\text{s.t.} \quad (s_1, s_2, \ldots, s_N) \in S, \quad (s_1, s_2, \ldots, s_N) \geq (d_1, d_2, \ldots, d_N)
\]  

where \((s_1 - d_1)(s_2 - d_2) \cdots (s_N - d_N)\) is known as the Nash product.

Choosing the disagreement point is important because it affects the “bargaining power” of the players, i.e., the one with a large disagreement point can tolerate more loss than the one with a small disagreement point. But we need to make sure to choose a point such that the player can improve upon bargaining.

In many cases, some players can have more privilege than others; if we remove the symmetry axiom, the optimization can be restated as

\[
\max \prod_i (s_i - d_i)^{\alpha_i} \\
\text{s.t.} \quad s \in S, \quad (s_1, s_2, \ldots, s_N) \geq (d_1, d_2, \ldots, d_N)
\]  

where \(\alpha_i\) is the weight of player \(i\).

**Example 36** (Bargaining Game for Bandwidth Sharing) Consider a network with \(N\) players seeking to share the transmission bandwidth for video transmission [13]. In this case, assume users cooperate to decide how to split network resources to optimize their performances. However, each user has his/her own interest and will try to negotiate to get a better share.

The bargaining game can be summarized as follows:
- The set of players \(\mathcal{N} = \{1, \ldots, N\}\).
- Each player \(i\) obtains a share \(x_i\) of bandwidth, i.e., \(\sum_{i \in \mathcal{N}} x_i = 1\). The utility of player \(i\) is defined as

\[
u_i(x_i) = \frac{255(x_i - x_{0,i})}{D_{0,i}(x_i - x_{0,i}) + \mu_i}
\]  

Available at https://www.cambridge.org/core/terms. doi:10.1017/9781316212493.005
where \( x_{0,i} \) is the bandwidth share at disagreement point, \( D_{0,i} \) and \( \mu_i \) is the rate-distortion parameters. The utility function is chosen such that the peak-signal-to-noise ratio of user \( i \), \( \text{PSNR}_i = 10 \log_{10} u_i(x_i) \).

- A disagreement point \( d \in \mathcal{S} \) represents the minimum utility that each user is guaranteed before he/she enters the bargaining.

Since at the disagreement point \( x_i = x_{0,i}, u_i = 0 \), we can define the disagreement point as the origin, i.e., \((0, 0, \ldots, 0)\). It can be shown that the utility region \( \mathcal{S} \) of this problem is convex and compact. To make the problem more general, we assign each user a weight factor \( 1 > \alpha_i > 0 \) and \( \sum_{i \in \mathcal{N}} \alpha_i = 1 \). Since the disagreement point is the origin, the Nash product is

\[
\max_{u_1, \ldots, u_N} \prod_{i \in \mathcal{N}} u_i(x_i)^{\alpha_i}. \tag{4.34}
\]

The solution will provide a resource allocation that is Pareto-optimal for all players. For this problem, bisection methods can be used to find the solution [13]. Interestingly, taking the logarithm of Nash’s product, we have

\[
10 \log_{10} \prod_{i \in \mathcal{N}} u_i(x_i)^{\alpha_i} = \sum_{i \in \mathcal{N}} \alpha_i \text{PSNR}_i^* \tag{4.35}
\]

where \( \text{PSNR}^* \) is the PSNR achieved by players in Nash bargaining solution. Therefore, we may say that the Nash bargaining maximizes the weighted sum of PSNRs for all players given the total bandwidth. Also, in [13], it is shown that the Nash bargaining solution achieves a better and fairer performance compared to equal allocation method.

### 4.3.2 Coalition Game

Another type of cooperative game is coalition game. In this game, players can form different coalitions to improve their payoffs. Within the same coalition, players can coordinate their strategies and redistribute the payoff in some specific way. A simple definition of coalition is as follows:

**Definition 26** A coalition is simply a subset of the set of players that is formed in order to coordinate strategies and to agree on how the total payoff is to be divided among the members.

A fundamental concept in coalition game is the coalition value denoted by \( v \). This quantity defines the worth of a coalition in a game.

**Definition 27** (Coalition Game) A coalition game consists of

- A set of players \( \mathcal{N} = \{1, \ldots, N\} \).
- For each coalition, a set of actions.
- For each player, preferences over the set of all actions for all coalitions in which he/she is a member.
A characteristic function $v$ is a function that assigns a value to each coalition of $\mathcal{N}$. The number $v(S)$ is called the worth of coalition $S$.

The most popular form of a coalition game is the transferable utility (TU) game where for each coalition $S$, the total utility represented by $v(S)$ can be divided in any manner among the coalition members. The value in a TU game is usually defined as monetary gain, and the value of a coalition game can be split among players in the same coalition using some fairness rule. The share $x_i$ each player $i$ in coalition $s$ receives from value $v(S)$ is called the payoff of user $i$. Consequently, the vector $x \in \mathbb{R}^{|S|}$ constitutes the payoff allocation for players in $S$.

In many cases, a coalition may not be characterized by a single value or have a specific rule on how the players share the payoff within a coalition. These games are called a non-transferable utility game (NTU). The value $v(s)$ of a coalition $s$ is a set of payoff vectors $v(s) \subseteq \mathbb{R}^s$ and an element $x_i \in x$, where $x \in v(s)$ is a payoff that player $i$ can obtain given a strategy that is played by him/her in coalition $s$. In fact, NTU is a general case for TU, where each player may have different strategies to maximize his/her own profit.

Now, we will discuss canonical coalition game. It is a type of coalition game where the formation of large coalitions is always beneficial to the players. This property is called super-additivity and defined below.

**Definition 28** (Super-additive Game) A TU game $(\mathcal{N}, v)$ is super-additive if

$$v(s_1 \cup s_2) \geq v(s_1) + v(s_2), \quad \forall s_1, s_2 \subseteq \mathcal{N} \text{ and } s_1 \cap s_2 = \emptyset. \quad (4.36)$$

For a cooperative game, super-additivity means forming a larger coalition will always provide a better or equal payoff. From the super-additivity property, it is easy to see that forming a grand coalition (coalition that includes all players) will achieve the highest value for the coalition. Then the problem now is to find a payoff allocation that guarantees no group of players has any incentive to leave the grand coalition while satisfying a set of fairness rules. We define the core of the game as the set of payoff allocations that guarantees no group of players wants to leave the grand coalition.

**Definition 29** (Group Rationality) A payoff vector $x \in \mathbb{R}^\mathcal{N}$ is group rational if

$$\sum_{i \in \mathcal{N}} x_i = v(\mathcal{N}).$$

**Definition 30** (Individual Rationality) A payoff vector $x \in \mathbb{R}^\mathcal{N}$ is individual rational if no player is better off by acting alone, i.e., $x_i \geq v(i)$.

**Definition 31** (Imputation Property) An imputation is a payoff vector that is both individual and group rational.

The formal definition of the core is as follows:

**Definition 32** (Core of TU Game) Given a TU canonical game $(\mathcal{N}, v)$, the core $C_{TU}$ is defined as a set of imputations in which no coalition $S \in \mathcal{N}$ has an incentive to leave $\mathcal{N}$.
and form $S$ instead:

$$C_{TU} = \left\{ \mathbf{x} : \sum_{i \in N} x_i = v(N) \text{ and } \sum_{i \in S} x_i \geq v(S) \forall S \subseteq N \right\}. \quad (4.37)$$

In general, given a TU coalition game $(N, v)$, the core can be found by solving an LP:

$$\min \mathbf{x} \sum_{i \in N} x_i \quad \text{s.t. } \sum_{i \in S} x_i \geq v(S), \quad \forall S \subseteq N \quad (4.38)$$

where $\mathbf{x}$ is an imputation. Clearly, the existence of the TU core is decided by the feasibility of the LP above. However, determining whether the core is non-empty or not by solving an LP is an NP-complete problem (since the number of constraints grows exponentially with $N$). In some special cases, we can use the Bondareva-Shapley theorem to determine the emptiness of the core.

**Definition 33 (Balanced Game)** A canonical TU game $(N, v)$ is known as a balanced game iff the inequality

$$\sum_{S \subseteq N} \mu(S)v(S) \leq v(N) \quad (4.39)$$

is satisfied for all non-negative weight collections $\mu = (\mu(S))_{S \subseteq N}$ that satisfy $\sum_{S \in N} \mu(S) = 1$. This set of non-negative weights is known as the balanced set.

The meaning of a balanced game is that if each coalition $S$ is assigned a small probability $\mu(S)$ to appear then its value will become $\mu(S)v(S)$. The expected value of all coalitions cannot exceed the value of the grand coalition.

**Theorem 37 (Bondareva-Shapley Theorem)** The core of a TU game is non-empty if and only if the game is balanced.

Another technique to solve the core in TU game is using a convex property. For a convex game, a non-empty core always exists; however, the reverse is not true.

**Definition 34 (Convex Game)** A TU canonical game $(N, v)$ is convex if

$$v(S_1) + v(S_2) \leq v(S_1 \cup S_2) + v(S_1 \cap S_2), \quad \forall S_1, S_2 \subseteq N. \quad (4.40)$$

The core is an important concept in coalition game. Its non-emptiness guarantees that the grand coalition is stable and optimal. However, it has the three following drawbacks:

- The core can be empty.
- The core can be large and selecting a suitable core can be difficult.
- The allocations in the core can be unfair.

To avoid these problems, Shapley suggested a unique mapping $\phi$ that associates with each coalition game $(N, v)$ and satisfies the following four axioms ($\phi_i$ is the payoff given to player $i$ by Shapley value $\phi$):
4.3 Cooperative Game

- Efficiency axiom: $\sum_{i \in \mathcal{N}} \phi_i(v) = v(\mathcal{N})$, i.e., group rationality.
- Symmetry axiom: If player $i$ and $j$ are such that $v(S \cup i) = v(S \cup j)$ for all coalition $S$ that do not contain $i$ or $j$, then $\phi_i(v) = \phi_j(v)$, i.e., if both players have the same contribution, their share should be the same.
- Dummy axiom: If player $i$ is such that $v(S \cup i) = v(S) \forall S : i \notin S$, then $\phi_i(v) = 0$, i.e., a player who does not contribute, receives nothing.
- Additivity axiom: If $u$ and $v$ are characteristic functions, then $\phi(u + v) = \phi(u) + \phi(v)$. The additivity axiom states that if we re-model the game as a single game in which each coalition $S$ achieves a payoff of $u(S) + v(S)$, the players’ payments in each coalition $S$ should be the sum of the payments they will receive for that coalition under the two separate games with $u$ and $v$.

Theorem 38 (Shapley Value) Given a coalition game $(\mathcal{N}, v)$, there is a unique payoff allocation $\phi(\mathcal{N}, v)$ that divides the full payoff of the grand coalition and that satisfies the symmetry, dummy player, and additivity axioms. This Shapley value is defined as

$$\phi_i(v) = \sum_{S : i \notin S} \frac{|S|!(N - |S| - 1)!}{N!} [v(S \cup i) - v(S)]. \quad (4.41)$$

The Shapley value is defined as a fair way of dividing the grand coalition’s payment among its members. However, this approach ignores the stability of the solution. In general, the Shapley value is unrelated to the core. But if the game is convex, it can be proved that the Shapley value belongs to the core. In such a case, the Shapley value reflects both fairness and stability.

Example 39 Consider the problem of bandwidth allocation for a set of users accessing a wireless Gaussian multiple-access channel [17]. The users try to find a mutual deal that allocates a fair share of bandwidth to everyone. Every user will try to act on its own which will reduce the rate available for remaining users if he/she does not agree with his/her share. A coalition game can be formulated as $(\mathcal{N}, v)$, where $\mathcal{N}$ is the set of wireless users and $v$ is a function of maximum sum-rate that a coalition can achieve. We assume that the other players outside a coalition $S$ will be hostile and cooperate to “jam” the transmission of users in $S$. The characteristic function $v$ represents the capacity that can be shared among the players in the same coalition, so this game is a transferable utility (TU) game. In [17], the authors prove that this game is super-additive, i.e., the sum-rates of the union of two disjoint coalitions are not less than the sum of sum-rates of two coalitions acting alone, because the set of jammers are the same in both cases. Therefore, users can form a grand coalition including all people to obtain the best payoffs, and the problem becomes a canonical coalition game.

Denoting by $P_i : i \in \mathcal{N}$ the power constraint for each user, we can define the capacity region of a coalition $S$ as $C^S = \{R \in \mathbb{R}^S \mid \sum_{i \in S} R_i \leq C(\sum_{i \in S} P_i, \sigma^2)\}$, where $C(p, \sigma^2) = \frac{1}{2} \log(1 + p/\sigma^2)$. It can be proved that all imputations belong to the capacity region $C^\mathcal{N}$ as well as to the core. However, the Shapley value does not lie in the core and so the grand coalition may not be fair. Instead, the authors propose another unique solution that satisfies dummy, efficiency, and symmetry axioms (skip the additivity axiom) named...
“envy-fairness” solution. This solution satisfies the envy-fairness axiom, which states that given two players $i$, $j$ with power constraints $P_i > P_j$, an allocation $\psi$ is envy-free if it gives the payoff $\psi_j(v)$ to player $j$ in game $v$ and payoff $\psi_i(v^{i,j})$ in the new game $v^{i,j}$, where $P_i$ is reduced to $P_j$ and both the payoffs are equal, i.e., $\psi_i(v^{i,j}) = \psi_j(v)$. The unique solution here is both envy-free and lies in the core, so it is both stable and fair.

The problem of wireless access based on a channel reservation sharing method for a group of mobile users was modeled in [14] by using a coalitional game model along with a stochastic game model. In the considered model, the mobile users form coalitions to share a reserved channel. The mobility diversity of the mobile users is exploited in order to reduce the cost of wireless access. Also, a channel access policy is required for contention resolution among mobile users belonging to the same coalition when they are at the same location. A stable coalitional structure and equilibrium channel access policy were obtained based on the game models.

Another application of a coalitional game model can be found in [15]. This work considered the problem of cooperative packet delivery to mobile nodes in a hybrid wireless mobile network, where both infrastructure-based and infrastructure-less (i.e., ad hoc mode or peer-to-peer mode) communications are used. A solution was proposed based on a coalition formation among mobile nodes to cooperatively deliver packets among these mobile nodes in the same coalition. The problem of rational coalition formation for cooperative packet delivery among mobile nodes in a wireless mobile delay-tolerant network (DTN) was considered in [16]. The mobile-nodes–forming coalitions can be either well-behaved or misbehaving in the sense that the well-behaved nodes always help each other for packet delivery, while the misbehaving nodes act selfishly and may not help the other nodes. A Bayesian coalitional game model was developed to analyze the behavior of mobile nodes during coalition formation. This model considered the presence of uncertainty of node behavior (i.e., type). Given the beliefs about the other mobile nodes’ types, each mobile node makes a decision to form a coalition. To find a stable coalitional structure in this coalitional game with incomplete information, a solution concept called Nash-stability was considered. Another solution concept called the Bayesian core was also considered that guarantees that no mobile node has an incentive to leave the grand coalition. Also, the Bayesian game model was extended to a dynamic game model for which a method was proposed for each mobile node to update its beliefs about other mobile nodes’ types when the coalitional game is played repeatedly.

### 4.4 Auction Theory

#### 4.4.1 Introduction to Auction Theory

As the name suggests, auction theory deals with how people act (bidding) in auction markets. Auctions have always been a large part of the economic landscape, with some auctions reported as early as in Babylon in about 500 B.C. There are many possible design (set of rules) for an auction, and game theory can be used to study the efficiency of a given design, the optimal and equilibrium of a bidding strategy, and the revenue comparison. Another branch of game theory related to auction is mechanism design. In this type of game, the designer choose a game structure rather than inheriting one
and investigates its outcome. Auction has important application in wireless communications. For example, we can apply it to model the spectrum auction process by the government. Applications of auction theory for resource allocation in cognitive radio networks were discussed in [18].

**Definition 35** An auction is a market mechanism in which an object service or set of objects is exchanged on the basis of bids submitted by the participants. It has a set of rules that will govern the sale or purchase of an object to the submitter of the most favourable bid. Some specific mechanisms are first price, second price, and English and Dutch auctions.

- A first-price auction is an auction in which the bidder who submitted the highest bid wins the object and needs to pay the price equal to the bid. In practice, the first price auction is often sealed-bid (participants submit bids simultaneously).
- A second-price auction is an auction in which the bidder who submitted the highest bid win and pays the price equal to the second highest bid.
- An English auction is a type of second-price auction where the auctioneer directs the participants to beat the current standing bid. The new bid must be higher than the current bid. The auction ends when no people wants to increase the bid. Then the winner, who currently places the bid, wins. All information about the bids are available to everyone in the auction house. It is termed second-price bid because the winner just needs to outbid the second person by a small amount. Thus the winner needs to pay a slightly larger amount than the second bid.
- A Dutch auction is a first-price auction in which the price keeps going down until the first one accepts it. First the “clock” (auctioneer) introduces a price that is substantially higher than the real value of the object. Then the clock gradually reduces the price until the first one buzzes in and accepts to pay.

For the bidders, their game models can be categorized into two categories. The first one is the private-value model where every bidder has a different belief in the value of the object and each participant guesses the bidding price of each of the competing bidders from a random distribution. The second type is a common-value auction where the actual value of the object is the same for every participant. However, each bidder has different private information about how much that value is. We have the following theorem.

**Theorem 40** (Revenue Equivalence Theorem) Any two auctions with the following properties:

- The bidder with highest price wins
- The bidder with lowest price receive zero profit
- Bidders are risk-neutral (a risk-neutral bidder is indifferent between choices with equal expected payoffs even if one choice is riskier)
- Value distributions are strictly increasing and atomless

have the same revenue and also the same expected profit for each bidders.

**Example 41** Consider a two-player first price auction game as follows. Two people are bidding an object. Each buyer assumes that the rival’s value of the object has uniform
distribution in the interval \([0,1]\), i.e., \(F(v) = v\). Assume that the value of the object for the seller is zero.

For this game, denote \(p\) as the bid price and \(v\) is the value player 1 believes the object is worth, the expected utility \(U\) for player 1 can be written as

\[
U(p) = (v - p)\Pr(p > B(v_0))
\]

(4.42)

where \(v_0\) is the value of the object player 2 believes and \(B\) is his/her bidding function. If \(B\) is monotonically increasing, then \(B\) has an inverse function, namely, \(Y = B^{-1}\). The utility function can be rewritten as \(U(p) = (v - p)\Pr(Y(p) > v_0)\). Using the distribution function \(F(v_0)\) we have

\[
\Pr(Y(p) > v_0) = F(Y(p)) = Y(p).
\]

(4.43)

Therefore, \(U(p) = (v - p)Y(p)\). To maximize the utility, player 1 should choose \(p\) such that \(U'(p) = 0\). Differentiating \(U\) and setting it to zero we have

\[
-Y(p) + (Y(p) - p)Y'(p) = 0.
\]

(4.44)

Solving this differential equation gives us \(Y(p)\), which is the inverse Nash equilibrium strategy for this game. This in turn gives us \(B\) the Nash equilibrium strategy for player 2. It can be proved that \(B(v) = v/2\).

### 4.4.2 Special Auction

#### Double Auction

A double auction is a process of buying and selling goods where both potential buyers and sellers submit their bids and asking prices to an auctioneer, and then an auctioneer chooses some price \(p\) that clears the market: all the sellers who asked less than \(p\) sell, and all buyers who bid more than \(p\) buy at this price \(p\).

Assume there are \(N\) buyers and \(M\) sellers. Each buyer wants to buy \(x_i\) items with cost \(p_{b_i}\) per item while each seller tries to sell \(y_j\) items with price \(p_{s_j}\). All information are publicly open. To determine the trading price, the demand quantities from all buyers are arranged in descending order and the supply quantities are arranged in ascending order. At the intersection \(T^*\) the demand meets the supply then \(m' - 1\) seller will deal with \(n' - 1\) buyers (as shown in Figure 4.3).

An application of double auction for competitive spectrum bidding and service pricing in IEEE 802.22-based wireless regional area networks (WRANs) can be found in [19]. The IEEE 802.22-based WRAN technology utilizes the TV bands, i.e., 54–862 MHz band using the dynamic-spectrum-access-based cognitive radio concept and coexists with the legacy TV services. WRAN service providers need to obtain the spectrum from the TV broadcasters by bidding for TV bands. Since there can be multiple TV broadcasters owning multiple TV bands and multiple WRAN service providers, a double auction model can be used to determine the number of TV bands procured by the service providers as well as the trading price. After procuring the TV bands, WRAN service providers compete with each other in terms of the service price charged to WRAN users. A non-cooperative game was formulated to obtain the number of TV bands to bid for and the service price for each of the WRAN service providers.
4.4 Auction Theory

Figure 4.3 Double auction game [3, p. 234].

The players of this game are the WRAN service providers. The strategy of a service provider is the number of TV bands to bid for and the service price offered to its users. The payoff of a service provider is given by its profit, which is the difference between the revenue it earns from its users and the cost of procuring TV spectrum from the TV broadcasters. The Nash equilibrium was obtained for this game.

Combinatorial Auction

Combinatorial auctions are those auctions in which bidders can place bids on combinations of items, called packages, rather than just individual items. An example of combinatorial auction is an estate sale or the FCC’s nationwide narrowband auction of spectrum rights in the United States. Here bidders were interested in different collections of spectrum licenses.

In theory, buyer may submit any possible combination of items available. However, in practice, these bids are limited by the resources of the buyers and a set of rules of the sellers, i.e., the auctioneer may limit the number of items or types of items a bidder can bid at a time. The difficult part is to decide which collections of bids to accept, which is called Combinatorial Auction Problem (CAP).

Assume there are $N$ bidders in a set $\mathcal{N}$ and $\mathcal{M}$ is the set of $m$ distinct objects. For every subset $S$ of $\mathcal{M}$, denote $b_i(S)$ as the bid that player $i$ willing to pay for $S$. Clearly, $b_i(S) \geq 0$. Assume that each player cares about only her own profit and not on what others receive. Denote $y(S, i) = 1$ if the bundle $S$ is sold to player $i$ and zero otherwise. We can formulate the CAP as the revenue maximization problem for the auctioneer as follows:

$$\max \sum_{i \in \mathcal{N}} \sum_{S \subseteq \mathcal{M}} b_i(S) y(S, i)$$  \hspace{1cm} (4.45)

$$\text{s.t} \sum_{S \ni j \in \mathcal{N}} \sum_{i \in \mathcal{N}} y(S, i) \leq 1, \quad \forall j \in \mathcal{M}$$  \hspace{1cm} (4.46)

$$\sum_{S \subseteq \mathcal{M}} y(S, i) \leq 1, \quad \forall i \in \mathcal{N}$$  \hspace{1cm} (4.47)

$$y(S, i) \in \{0, 1\}, \quad \forall i \in \mathcal{N}, S \subseteq \mathcal{M}.$$  \hspace{1cm} (4.48)
The first constraint ensures that overlapping sets of goods are never assigned ($S \ni j$ means we sum over all subsets $S$ that contain item $j$). The second ensures that no bidder receives more than one subset. This is a linear integer programming and is solvable although very time consuming. If we assume that bid functions are super-additive, a simpler problem can be derived. Denote $b(S) = \max_{i \in N} b_i(S)$ is the highest bid for set $S$. Denote $x_S = 1,$ if the highest bid for $S$ wins $S$ (it may not win if someone bids for $S'$ that contains $S$ with a higher price). The CAP can be formulated as follows:

$$\begin{align*}
\text{max} & \quad \sum_{S \subseteq M} b(S)x_S \\
\text{s.t.} & \quad \sum_{S \ni i} x_S, \quad \forall \ i \in M \\
& \quad x_S = \{0, 1\} \leq 1, \quad \forall \ S \subseteq M.
\end{align*}$$

(4.49)

A thorough review on CAP can be found in [20]. An application of CAP in the context of resource allocation for wireless network virtualization can be found in [21].

### 4.5 Exercises

**Exercise 4.1:** (Bertrand Competition) Suppose that there are two firms in the market that produce the same product. This product has the same marginal cost $c$ per each unit for both firms. Assume that the consumers will buy from the firms with lower price. In case, both prices are the same then the profit is split evenly. Here each firm try to maximize the profit they make, which is given as follows:

$$u(p_i, p_{-i},) = \begin{cases} 
Q(p_i)(p_i - c), & \text{if } p_{-i} > p_i \\
\frac{1}{2}Q(p_i)(p_i - c), & \text{if } p_{-i} = p_i \\
0, & \text{otherwise}
\end{cases}$$

where $Q(p)$ is the demand of the market when the price is at $p$. Prove that there exist a unique Nash equilibrium for this game model given by $p_1 = p_2 = c$.

**Hint:** Check all possible cases of $p_1, p_2, c$, i.e., $p_1 > p_2 > c \ldots$ to see if there exists another Nash equilibrium.

**Exercise 4.2:** (Mixed Strategies for Zero-Sum Matrix Game) Consider two players, player 1 makes a choice $k \in K = \{1, \ldots, m\}$ and player 2’s choice is $l \in L = \{1, \ldots, n\}$. If the player 1 chooses $k$ and player 2 chooses $l$ then the payoff are $P_{k,l}$ (gain) and $-P_{k,l}$ (loss) for the first and second players, respectively, where $P_{m \times n}$ is the payoff matrix. Both of them using mixed strategy to play the game. Assume the probability that player 1 chooses action $k = i$ is $u_i = P(k = i)$ and for player 2 denote $v_j = P(l = j)$. Clearly player 1 tries to maximize its expected payoff while player 2 tries to minimize it. Prove that there exist value $V$ such that:
• There exists a mixed strategy for player 1 such that his/her average gain is at least $V$ no matter what player 2 does, and
• There exists a mixed strategy for player 2 such that his/her average loss is at most $-V$ no matter what player 1 does.

**Hint**: Use this formula: $\min\{x^T y | x \succeq 0, x^T 1 = 1\} = \min_{i=1, \ldots, m} y_i$ where $y_1, \ldots, y_m$ are elements of vector $y$. Convert the problems for players 1 and 2 into LP problems and show that they are primal and dual problems of each other.

**Exercise 4.3**: (Second Price Auction) Consider the following auction mechanism:

• There is an object bid by $n$ players.
• Each player has his/her own valuation $v_i$ for the object. Assume that $v_1 > v_2 > \cdots > v_n > 0$ and the information about $v_i$ is available to all of the players.
• All of the players simultaneously submit their bids.
• The object is given to the highest bidder (or to a random player with highest bid). However, he/she needs to pay only the second highest bid.
• The winner’s payoff is $v_i - b_i$ where $b_i$ is the price he/she will pay. Others receive zero payoff. Prove that in Second Price Auction, the truthful bidding, i.e., everyone will bid with $v_i$, is a Nash equilibrium. Can you find another Nash equilibrium for this model?

**Hint**: Prove that nobody has an incentive to deviate. Check if $(v_2, v_1, 0, \ldots, 0)$ is a Nash equilibrium or not.

**Exercise 4.4**: Let $G = (\mathcal{I}, (S_i)_{i \in \mathcal{I}}, (u_i)_{i \in \mathcal{I}})$ be a game in strategic form. Prove that the mixed strategy $\sigma^*$ is a Nash equilibrium if and only if for each player $i \in \mathcal{I}$ every pure strategy in the support of $\sigma^*_i$ is the best response to $\sigma^*_i - i$.

**Hint**: Assume there is a strategy with positive probability and not the best response, then shifting that probability to a best response will improve the expected utility.

**Exercise 4.5**: (Crime Report) [4] A crime is observed by $n$ people. Each person wants to inform the police but also prefers other persons do it (to avoid paperwork). Denote $v$ as the benefit to inform the police and $c$ as the cost to make the call. Clearly $v > c > 0$; otherwise, no one will inform. The payoff for a person is $0$ if no one calls, $v - c$ if he/she calls, and $v$ if someone calls. Let us assume further that none of them can communicate to each other and each of them has the same probability $p$ to call the police. Find the Nash equilibrium of this game (i.e., find $p$).

**Hint**: Use conditional probability. Show that the larger group is, the lesser is the chance that this crime will be reported!

**Exercise 4.6**: We extend the Bertrand Competition as follows: Assume the demand is fixed as one unit while each firm has capacity constraint of $2/3$ of the unit demand. That means $2/3$ of the demand will be supplied by the player with the cheaper price and $1/3$ will be supplied by the other player. Assume that both firms use the same mixed strategy and $c = 0$ for simplicity. Find the Nash equilibrium in this case.
Hint: Denote by $F(p)$ the cumulative distribution of price for each player. Calculate the expected payoff for player 1 if he/she uses price $p_1$. Use the proposition from Exercise 4.3 that every action in the support of a mixed strategy must yield the same payoff to a player at equilibrium (since all of them are best responses) to obtain $F(p)$.

**Exercise 4.7:** (Stackelberg’s duopoly game) Consider a market with two players that produce the same product. Assume player 1 has more priority and can declare his/her quantity of produced units first $q_1$, and after that player 2 can choose her own $q_2$. Assume $C(q) = cq$ is the cost of production $q$ units where $c > 0$. Denote $Q = q_1 + q_2$ the price per unit is decided by the market using the following formula:

$$P(Q) = \begin{cases} M - bQ, & \text{if } Q \leq M \\ 0, & \text{otherwise} \end{cases}$$

The payoff is the profit that each player receives. Find the subgame perfect equilibrium of this model.

**Hint:** Find best response of firm 2 given the produced units of firm 1.

**Exercise 4.8:** (The Ultimatum Game) Assume that there are two players picking up some money $c$ on the road. Player 1 has more priority and offer the amount of $b$ to player 2 while he/she keeps the rest. If player 2 rejects the offer, then both the players bring the money to the police and thus receive nothing. Find the subgame perfect equilibrium for this game. Assume that $c$ is an integer and player 1 can only offer an integer amount of money. Then what is the new subgame perfect equilibrium?

**Hint:** Using backward induction. Player 1 offers 0 and player 2 accepts all offer.

**Exercise 4.9:** (The Pirate Game) We extends the Ultimate Game. Assume there are 5 pirates $A, B, C, D, E$ that find a chest of 100 gold pieces. We assume that $A > B > C > D > E$ where the sign $>$ means higher priority. The pirates decide to split the gold by using following procedure: The gold piece cannot be divided smaller so the distribution must use integer values. The highest priority person will propose the distribution, then they all will vote for this. If the votes with “Yes” are larger or equal to the “No” votes, then the distribution is accepted; otherwise, the proposer will be killed and the next priority person will start the process again. Assume that all pirates want to survive (we say that if a person is dead her payoff is $-\infty$) and each of them wants to maximize her received gold. Also, assume that the pirates do not trust each other and prefer to kill the proposer if the results are the same. Find the subgame perfect equilibrium.

**Hint:** Use backward induction, i.e., assume only D and E are left. Show that the amount of gold distributed is $(98, 0, 1, 0, 1)$. The captain receives most of the gold.

**Exercise 4.10:** (Cournot Duopoly with Incomplete Information) Two firms compete against each other to sell the same product. Each firm needs to determine the level of output to maximize the profit. Assume firm 1 has marginal cost to produce 1 unit as $c_1 = 0$, which is known to both firms, while firm 2’s marginal cost $c_2$ is private. However, firm 1 knows that $c_2$ can take two values: $c_L$ with probability $\theta$ and $c_H$ with
4.5 Exercises

Table 4.6 Battle of Sexes

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1, 2</td>
<td>0, 0</td>
</tr>
<tr>
<td>S</td>
<td>0, 0</td>
<td>2, 1</td>
</tr>
</tbody>
</table>

probability $1 - \theta$. Assume $q_1$ and $q_2$ are the level of output for firm 1 and 2, the price is determined by the market using the formula $p = M - q_1 - q_2$. Find the Bayesian Nash equilibrium (BNE) for this game.

**Hint:** The BNE is a triple $(q_1^*, q_2(c_L)^*, q_2(c_H)^*)$. Given $q_1^*$, consider the best response $q_2$ of player 2 for the cases when the prices are $c_H$ and $c_L$. Then derive the expected payoff for player 1 and find its best response.

**Exercise 4.11:** (Auction with incomplete information [4]) Two players are bidding for a product where the one with higher price wins. Assume that the bidding price of each player is private. However, player 1 knows about the distribution of the bidding price of player 2 and vice versa. Assume that both of the distributions are uniform distributions in the interval $[0, 1]$, i.e., the probability that player $i$’s valuation of the product is at most $v$ is, $F(v) = v, \forall v \in [0, 1]$. Prove that the bidding price strategy $\beta_i(v) = v/2$ for $i = 1, 2$ is the Bayesian Nash equilibrium for this auction.

**Hint:** Assume player 2 uses this strategy, derive the expected payoff function of user 1 if user 1’s bid is $b_1$. Then derive user 1’s best response.

**Exercise 4.12:** Consider an evolutionary game with linear utility function $u(s, s')$, prove that a strategy $\sigma^*$ is evolutionarily stable if and only if for any strategy $\sigma \neq \sigma^*$ we have $u(\sigma^*, \sigma^*) \geq u(\sigma, \sigma^*)$ also if $u(\sigma^*, \sigma^*) = u(\sigma, \sigma^*)$, then we have $u(\sigma^*, \sigma) > u(\sigma, \sigma)$.

**Hint:** Use the fact that $u$ is linear in its argument.

**Exercise 4.13:** (Battle of Sexes) A couple always go out for a concert every weekend. The husband prefers Bach while the wife prefers Stravinsky. If both of them cannot agree, then they both will stay at home. The payoff is given in Table 4.6 where the first number is the wife’s payoff while the second is the husband’s. Prove that the strategy in which each player chooses his/her favourite concert with probability 2/3 is a mixed strategy equilibrium. Prove that this strategy is also an ESS.

**Hint:** Use Exercise 4.12.

**Exercise 4.14:** (Landowner-workers game [4]) A landowner has a estate and wants to hire more persons to work with him/her. If he/she hires $k \geq 0$ workers, then the whole team can produce an output of $f(k + 1)$ where $f$ is an increasing function. Assume that he/she can hire at most two persons. Find the core of the game.

**Hint:** Assume $x_1, x_2, x_3$ are the distribution of profit of the team where player 1 is the owner. Apply the definition of core.
Exercise 4.15: (The Gloves Game) There are three persons playing a game. Player 1 and 2 have the right gloves and player 3 has the left glove. They want to form a coalition such that they can have a pair of glove. If that is the case, then the payoff is 1; otherwise, the payoff is 0. List the characteristic function \( v \) for all the coalitions and calculate the Shapley value for each person.

Hint: The Shapley values are \((1/6, 1/6, 2/3)\).

References


Part III

Physical Layer Resource Allocation in Wireless Networks
5 General System Model and Preliminary Concepts

We consider a sufficiently general model for wireless networks. The network could be either a cellular or an ad hoc network, and it corresponds to a collection of interfering radio links in a single channel. Each link corresponds to a single-hop radio transmission from a transmitter node to an intended receiver node. In the cellular network paradigm, a link corresponds to an up-stream or a down-stream transmission between a mobile and the BS it is associated with. In the ad hoc network paradigm, a link corresponds to single-hop transmission between mobile nodes.

In an FDMA system, the channels are non-overlapping frequency bands, and in a CDMA system, the entire spectrum is viewed as a single channel. In a non-orthogonal uplink transmission scenario, such as in a CDMA system, the transmit powers from all links appear as interference. In an orthogonal uplink transmission scenario, such as in an OFDMA system, transmit powers from the links terminating to the same BS are orthogonal and do not cause interference to one another. In this scheme, the transmit power from a link terminating on a given BS appears as interference only to the links terminating to different BSs using the same channels. In other words, there is no intra-cell interference and only inter-cell interference may occur, which is the case when the carriers are reused.

5.1 System Model for a General Multi-Cell Wireless Network

Now consider a multi-cell wireless network with K BSs (cells) and M active user equipments (UEs) denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{M} = \{1, 2, \ldots, M\}$, respectively. We will use the terms “user” and “UE” interchangeably. Let $p_i$ be the transmit power of user $i$. Noise is assumed to be additive white Gaussian whose power at the receiver of BS $k \in \mathcal{K}$ is $\sigma_k^2$, and at the receiver of UE $i \in \mathcal{M}$ is $\tilde{\sigma}_i^2$. Let $b_i$, where $b_i \in \mathcal{K}$, denote the BS corresponding to user $i$. We assume that one user is associated with only one BS for both uplink and downlink communications.\(^1\) Denote the set of UEs served by BS $k \in \mathcal{K}$ by $C_k = \{i \in \mathcal{M} | b_i = k\}$.

We will first introduce the notations for uplink and downlink path-gains and then the SINR at uplink and downlink. Then we will proceed to derive a relation between

\(^1\) Simultaneous connections to multiple BSs and different BS association for uplink and downlink can be also considered. Such a general scheme would increase the degrees of freedom that can be exploited to further improve network capacity and balance the load among different BSs in different cells [1].
the transmit power vector and SINR vector, based on which we will introduce one of the most important concepts in wireless networks, namely, the concept of SINR feasibility.

5.1.1 Modeling Path-Gains

In wireless networks with multiple receiving points such as multi-cell and ad hoc wireless networks, all links do not terminate to a single receiver. Hence, for such a general system model, not only the path-gain between each transmitter and its corresponding receiver, but also the path-gain between that transmitter and other receivers should also be determined. In multi-cell cellular and ad hoc networks, it is required to represent the path-gain (at either uplink or downlink) with two indexes, one for determining the transmitting point and the other one for the receiving point. Therefore, we need to index all transmitters and receivers. In cellular wireless networks, at the uplink, the transmitters are UEs and the receivers are BSs, and at the downlink, the BSs are transmitting points and the UEs are the receiving points. Since the transmitters and receivers are interchanged in the uplink and downlink, it would be more convenient if we index mobile UEs and BSs and use these indexes for describing both uplink and downlink path-gains, as explained below.

Let $h_{rt}$ denote the path-gain from the transmitter point with index $t$ to the receiver point with index $r$. With this description of path-gains, among two indexes for path-gains, the first one denotes the index of receiving point (i.e., $r$ in $h_{rt}$) and the second one denotes the index of transmitter point (i.e., $t$ in $h_{rt}$). In uplink, the indexes $t$ and $r$ correspond to UEs and BSs, respectively, i.e., $t \in \mathcal{M}$ and $r \in \mathcal{K}$. In the downlink, the indexes $t$ and $r$ correspond to BSs and UEs, respectively, i.e., $t \in \mathcal{K}$ and $r \in \mathcal{M}$. Note that these notations can also be used for mobile ad hoc networks where the transmitters and receivers are mobile UEs. In such a network, $h_{rt}$ would represent the path-gain between the transmitter of user $t$ and the receiver of user $r$. However, to facilitate exposition, from here on, we adopt a cellular terminology, although all discussions are also valid for mobile ad hoc wireless networks.

Although the notations introduced for describing path-gains can be used for both uplink and downlink by noting which (mobile user or BS) nodes are transmitters or receivers, in order to avoid any possible confusion, we use different notations for uplink and downlink path-gains. We use the superscript “$\sim$” for all parameters and variables related to downlink. In doing so, and in order to differentiate between uplink and downlink path-gains, we use the notation $h_{ij}$ to denote the uplink path-gain between the transmitter of user $j \in \mathcal{M}$ and the receiver of BS $i \in \mathcal{K}$, and use the notation $\tilde{h}_{ij}$, to denote the downlink path-gain between the transmitter of BS $j \in \mathcal{K}$ and the receiver of user $i \in \mathcal{M}$. Note that for both notations $h_{ij}$ and $\tilde{h}_{ij}$, the first and second indexes correspond to the indexes of receivers and transmitters, respectively.

Using the notations introduced above, the uplink and downlink path-gains between user $i \in \mathcal{M}$ and BS $b_j \in \mathcal{K}$ (i.e., the BS assigned to user $j$) are represented by $h_{bi,j}$ and $\tilde{h}_{bj,i}$, respectively. Also, $h_{bi}$ and $\tilde{h}_{bi}$ are the uplink and downlink path-gains, respectively, for user $i$ and its assigned BS $b_i$. In general, we have $h_{b_i,j} \neq \tilde{h}_{b_j,i}$.
The path-gains are usually described by a path-gain matrix. Let \( H_{K \times M} = [h_{ij}] \) and \( \tilde{H}_{M \times K} = [\tilde{h}_{ij}] \) denote the uplink and downlink path-gain matrices, respectively:

\[
H_{K \times M} = \begin{bmatrix}
h_{11} & h_{12} & \ldots & h_{1M} \\
h_{21} & h_{22} & \ldots & h_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
h_{K1} & h_{K2} & \ldots & h_{KM}
\end{bmatrix}
\]

(5.1)

and

\[
\tilde{H}_{M \times K} = \begin{bmatrix}
\tilde{h}_{11} & \tilde{h}_{12} & \ldots & \tilde{h}_{1K} \\
\tilde{h}_{21} & \tilde{h}_{22} & \ldots & \tilde{h}_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{h}_{M1} & \tilde{h}_{M2} & \ldots & \tilde{h}_{MK}
\end{bmatrix}.
\]

(5.2)

Note that the uplink path-gain matrix \( H \) is a \( K \times M \) matrix, while the downlink path-gain matrix \( \tilde{H} \) is an \( M \times K \) matrix. This is because we have \( M \) transmitters and \( K \) receivers in the uplink, while there are \( K \) transmitters and \( M \) receivers in the downlink.

Sometimes, in order to represent the relation between the transmit power and SINR vectors in matrix format, it is more convenient to represent either uplink or downlink path-gain matrices with an \( M \times M \) matrix. We now introduce an \( M \times M \) path-gain matrix, which is commonly used in related literatures.

To explain this kind of description of path-gains and differentiate it from our earlier introduced notations for path-gains, let us use a different notation, say, \( h'_{ij} \) to denote the uplink path-gain from user \( j \) to the BS assigned to user \( i \):

\[
h'_{ij} = h_{bi}. \tag{5.3}
\]

Similarly, let \( \tilde{h}'_{ij} \) denote the downlink path-gain from the BS assigned to user \( j \) to user \( i \):

\[
\tilde{h}'_{ij} = \tilde{h}_{ib}. \tag{5.4}
\]

In this kind of description of path-gains, both indexes are UEs’ indexes, i.e., for \( h_{ij} \) and \( \tilde{h}'_{ij} \), we have \( i, j \in \mathcal{M} \), but one for transmitter of UEs and the other one for the corresponding BS assigned to the UEs. The \( M \times M \) uplink and downlink path-gain matrices, respectively, corresponding to this kind of notation are

\[
H'_{M \times M} = \begin{bmatrix}
h'_{11} & h'_{12} & \ldots & h'_{1M} \\
h'_{21} & h'_{22} & \ldots & h'_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
h'_{M1} & h'_{M2} & \ldots & h'_{MM}
\end{bmatrix} = \begin{bmatrix}
h_{b1} & h_{b2} & \ldots & h_{bM} \\
h_{b2} & h_{b2} & \ldots & h_{b2M} \\
\vdots & \vdots & \ddots & \vdots \\
h_{bM} & h_{bM} & \ldots & h_{bM}
\end{bmatrix} \tag{5.5}
\]
and

\[
\mathbf{H}'_{M \times M} = \begin{bmatrix}
\tilde{h}'_{11} & \tilde{h}'_{12} & \cdots & \tilde{h}'_{1M} \\
\tilde{h}'_{21} & \tilde{h}'_{22} & \cdots & \tilde{h}'_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{h}'_{M1} & \tilde{h}'_{M2} & \cdots & \tilde{h}'_{MM}
\end{bmatrix} = \begin{bmatrix}
\tilde{h}_{1b_1} & \tilde{h}_{1b_2} & \cdots & \tilde{h}_{1b_M} \\
\tilde{h}_{2b_1} & \tilde{h}_{2b_2} & \cdots & \tilde{h}_{2b_M} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{h}_{Mb_1} & \tilde{h}_{Mb_2} & \cdots & \tilde{h}_{Mb_M}
\end{bmatrix}.
\] (5.6)

Note that there are redundant components in each column and row of the matrices \( \mathbf{H}'_{M \times M} \) and \( \mathbf{H}_{M \times M}' \), respectively, in contrast to that of the matrices \( \mathbf{H}_{K \times M} \) and \( \mathbf{H}_{M \times K} \), because for each pair of UEs \( i \) and \( l \), which are served by the same BS, i.e., \( b_i = b_l \), we have \( h'_{ij} = h'_{lj} \) and \( \tilde{h}'_{ji} = \tilde{h}'_{lj} \). Also note that the path-gains for different links in an ad hoc wireless network can be similarly described by a path-gain matrix like \( \mathbf{H}'_{M \times M} \) in which \( h'_{ij} \) denotes the path-gain between transmitter of mobile user \( j \) and receiver of mobile user \( i \). Note that for the path-gains described by \( \mathbf{H}_{K \times M} \), a cellular wireless network with a \( K \) BSs and \( M \) UEs is implicitly assumed.

In this book, depending on the case we may use one of the two above notations. In fact, there is no significant difference between these two notations, since two uplink path-gains matrices \( \mathbf{H}'_{M \times M} \) and \( \mathbf{H}_{K \times M} \) as well as two downlink path-gains matrices \( \mathbf{H}_{M \times M} \) and \( \mathbf{H}_{M \times K} \) are related to each other by the relations (5.3) and (5.4), respectively. However, when we deal with an ad hoc or a cellular wireless network with an already known BS assignment (for example, see Chapters 6 and 7), using the matrices \( \mathbf{H}'_{M \times M} \) and \( \mathbf{H}_{M \times M} \) is more common and convenient. For the case in which the BS assignment is not fixed and is going to be dynamically assigned (i.e., \( b_i \) is variable, as will be discussed in Chapter 9), using path-gain descriptions \( \mathbf{H}_{K \times M} \) and \( \mathbf{H}_{M \times K} \) is more convenient.

### 5.1.2 SINR Model

#### Uplink SINR

For a given uplink transmit power vector \( \mathbf{p} = [p_1, p_2, \ldots, p_M]^T \), the corresponding uplink SINR of a user \( i \), denoted by \( \gamma_i \) is

\[
\gamma_i(\mathbf{p}) = \frac{h_{bi}p_i}{\sum_{j \neq i} h_{bj}p_j + \sigma^2_{b_i}}. \quad (5.7)
\]

The uplink interference experienced by user \( i \) at its assigned BS is \( I_i(\mathbf{p}) = \sum_{j \neq i} h_{bj}p_j + \sigma^2_{b_i} \). The uplink interference \( I_i(\mathbf{p}) \) can be rewritten as \( I_i(\mathbf{p}) = I_i^{\text{int}} + I_i^{\text{ext}} + \sigma^2_{b_i} \), where \( I_i^{\text{int}} = \sum_{j \in C_{b_i}, j \neq i} p_jh_{bj} \) and \( I_i^{\text{ext}} = \sum_{j \notin C_{b_i}} p_jh_{bj} \) are the uplink intra-cell interference and the uplink inter-cell interference, respectively.

The total uplink received power plus noise at the BS \( k \in K \) is

\[
\psi_k^T(\mathbf{p}) = \sum_{i \in M} p_ih_{ki} + \sigma^2_k. \quad (5.8)
\]

Let us define the effective uplink SINR of user \( i \) by

\[
\theta_i(\mathbf{p}) = \frac{\gamma_i(\mathbf{p})}{\gamma_i(\mathbf{p}) + 1}. \quad (5.9)
\]
By putting (5.7) into (5.9), we have
\[ \theta_i(p) = \frac{p_i h_{bi}}{\varphi_i^T(p)}. \] (5.10)

In fact, the effective uplink SINR is the ratio of the uplink received power of user \( i \) to the total uplink received power plus noise at its assigned BS.

**Downlink SINR**

For a given downlink transmit power vector \( \tilde{p} = [\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_M]^T \), the corresponding downlink SINR of a user \( i \), denoted by \( \tilde{\gamma}_i \), is
\[ \tilde{\gamma}_i(p) = \frac{\tilde{p}_i h_{bi}}{\tilde{\varphi}_i^T(p)} + \tilde{\sigma}_i^2. \] (5.11)

The downlink interference experienced by user \( i \) at its receiver is \( \tilde{I}_i(p) = \sum_{j \neq i} \tilde{h}_{ib_j} \tilde{p}_j + \tilde{\sigma}_i^2 \). The downlink interference \( \tilde{I}_i(p) \) can be rewritten as \( \tilde{I}_i(p) = \tilde{I}_{i}^{\text{int}} + \tilde{I}_{i}^{\text{ext}} + \tilde{\sigma}_i^2 \), where \( \tilde{I}_{i}^{\text{int}} = \tilde{h}_{ib_i} \sum_{j \in C_{bi}, j \neq i} \tilde{p}_j \) is the downlink intra-cell interference (note that \( b_j = b_i \) for all \( j \in C_{bi} \)), and \( \tilde{I}_{i}^{\text{ext}} = \sum_{j \not\in C_{bi}} \tilde{p}_j \tilde{h}_{ib_j} \) is the downlink inter-cell interference.

The total downlink received power plus noise at user \( i \in M \) is
\[ \tilde{\varphi}_i^T(\tilde{p}) = \sum_{j \in M} \tilde{p}_j \tilde{h}_{ib_j} + \tilde{\sigma}_i^2. \] (5.12)

Similar to the uplink, the effective downlink SINR of user \( i \) is defined by
\[ \tilde{\theta}_i(\tilde{p}) = \frac{\tilde{\gamma}_i(\tilde{p})}{\tilde{\gamma}_i(\tilde{p}) + 1}, \] (5.13)

and it represents the ratio of the downlink received power of BS \( b_i \) to the total downlink received power plus noise at the user \( i \)'s receiver:
\[ \tilde{\theta}_i(\tilde{p}) = \frac{\tilde{p}_i h_{bi}}{\tilde{\varphi}_i^T(\tilde{p})}. \] (5.14)

### 5.1.3 Transmit Power Vector Corresponding to a Given SINR Vector

There is a one-to-one relation between a transmit power vector \( p = [p_1, p_2, \ldots, p_M]^T \) and the actual SINR vector \( \gamma = [\gamma_1, \gamma_2, \ldots, \gamma_M]^T \), which are given by (5.7) and (5.11) for the uplink and downlink, respectively. In other words, the SINR vectors corresponding to different transmit power vectors are different, i.e., if \( p \neq p' \) then we have \( \gamma \neq \gamma' \), where \( \gamma \) and \( \gamma' \) are SINR vectors corresponding to \( p \) and \( p' \), respectively.

As has been mentioned previously, for any given uplink or downlink transmit power vector, its corresponding SINR vector is easily obtained by (5.7) or (5.11), respectively. Now assume that an SINR vector is given and we are interested in knowing its corresponding transmit power vector. What is the transmit power vector corresponding to a given SINR vector? And how is it obtained? This is a basic question, because for a given QoS requirement for UEs (e.g., the SINR values of UEs), we would need to know
what transmit power levels for the UEs would result in those required SINR values. To answer this question, it is required to obtain the transmit power vector as a function of the received SINR vector, i.e., the inverse functions of (5.7) and (5.11) for uplink and downlink cases, respectively. For a general multi-cell wireless network, the transmit power level of each UE is derived as a function of its corresponding SINR vector by taking a matrix inversion, and for a single-cell wireless network, it is derived as a simple closed-form function of a given SINR vector for uplink and downlink cases as described below.

**Uplink Transmit Power Vector as a Function of Uplink SINR Vector**

To obtain a linear equation for relation between the uplink power vector and its corresponding SINR vector, we rewrite (5.7) as

$$p_i = \sum_{j \neq i} h_{b_i j} \gamma_j p_j + \gamma_i \frac{\sigma^2_{b_i}}{h_{b_i i}}.$$  

(5.15)

Rewriting the above linear relation for all $i \in M$ in matrix format, we have

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_M \end{bmatrix} = \begin{bmatrix} 0 & \gamma_1 \frac{h_{b_1 2}}{h_{b_1 1}} & \cdots & \gamma_1 \frac{h_{b_1 M}}{h_{b_1 1}} \\ \gamma_2 \frac{h_{b_2 1}}{h_{b_2 2}} & 0 & \cdots & \gamma_2 \frac{h_{b_2 M}}{h_{b_2 2}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_M \frac{h_{b_M 1}}{h_{b_M M}} & \gamma_M \frac{h_{b_M 2}}{h_{b_M M}} & \cdots & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_M \end{bmatrix} + \begin{bmatrix} \gamma_1 \frac{\sigma^2_{b_1}}{h_{b_1 1}} \\ \gamma_2 \frac{\sigma^2_{b_2}}{h_{b_2 2}} \\ \vdots \\ \gamma_M \frac{\sigma^2_{b_M}}{h_{b_M M}} \end{bmatrix}.$$  

(5.16)

The above matrix format equation can also be rewritten as

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_M \end{bmatrix} = \begin{bmatrix} \gamma_1 & 0 & \cdots & 0 \\ 0 & \gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_M \end{bmatrix} \begin{bmatrix} 0 & \frac{h_{b_1 2}}{h_{b_1 1}} & \cdots & \frac{h_{b_1 M}}{h_{b_1 1}} \\ \frac{h_{b_2 1}}{h_{b_2 2}} & 0 & \cdots & \frac{h_{b_2 M}}{h_{b_2 2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{h_{b_M 1}}{h_{b_M M}} & \frac{h_{b_M 2}}{h_{b_M M}} & \cdots & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_M \end{bmatrix} + \begin{bmatrix} \gamma_1 \frac{\sigma^2_{b_1}}{h_{b_1 1}} \\ \gamma_2 \frac{\sigma^2_{b_2}}{h_{b_2 2}} \\ \vdots \\ \gamma_M \frac{\sigma^2_{b_M}}{h_{b_M M}} \end{bmatrix}.$$  

(5.17)

Using matrix notations, the relation between the uplink transmit power vector and its corresponding uplink SINR vector can be rewritten as

$$\mathbf{p} = \mathbf{D}(\gamma)\mathbf{Gp} + \mathbf{D}(\gamma)\eta$$  

(5.18)
5.1 System Model for a General Multi-Cell Wireless Network

where $\mathbf{D}(\gamma)$ denotes a diagonal matrix whose diagonal elements are the corresponding components of the SINR vector $\gamma$, the $(i, j)$ component of $\mathbf{G}$ is

$$G_{ij} = \begin{cases} \frac{h_{bj}}{h_{bi}}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}$$  \hspace{1cm} (5.19)$$

and the $(i)$ component of $\eta$ is

$$\eta_i = \frac{\sigma^2_{bi}}{h_{bi}}.$$  \hspace{1cm} (5.20)

From (5.18) we know that

$$(\mathbf{I} - \mathbf{D}(\gamma)\mathbf{G})\mathbf{p} = \mathbf{D}(\gamma)\eta$$  \hspace{1cm} (5.21)$$

where $\mathbf{I}$ is an $M \times M$ identity matrix. Given an uplink SINR vector, the corresponding uplink transmit power vector is thus computed as

$$\mathbf{p} = (\mathbf{I} - \mathbf{D}(\gamma)\mathbf{G})^{-1}\mathbf{D}(\gamma)\eta.$$  \hspace{1cm} (5.22)$$

Therefore, to obtain the uplink transmit power vector corresponding to a given uplink SINR vector $\gamma$, we require calculating the inverse of matrix $\mathbf{I} - \mathbf{GD}(\gamma)$.

**Downlink Transmit Power Vector as a Function of Downlink SINR Vector**

To obtain a linear equation for relation between the downlink power vector and its corresponding downlink SINR vector, we rewrite (5.11) as

$$\tilde{p}_i = \sum_{j \neq i} \frac{\tilde{h}_{ibj}}{\tilde{h}_{ibi}} \tilde{\gamma}_j \tilde{p}_j + \tilde{\gamma}_i \frac{\tilde{\sigma}_i^2}{\tilde{h}_{ibi}}.$$  \hspace{1cm} (5.23)$$

Rewriting the above linear relation for all $i \in \mathcal{M}$ in matrix format, we have

$$\begin{bmatrix} \tilde{p}_1 \\ \tilde{p}_2 \\ \vdots \\ \tilde{p}_m \end{bmatrix} = \begin{bmatrix} 0 & \frac{\tilde{h}_{ib1}}{\tilde{h}_{ibi}} & \cdots & \frac{\tilde{h}_{ibM}}{\tilde{h}_{ibi}} \\ \frac{\tilde{h}_{2b1}}{\tilde{h}_{2bi}} & 0 & \cdots & \frac{\tilde{h}_{2bM}}{\tilde{h}_{2bi}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\tilde{h}_{Mb1}}{\tilde{h}_{Mbi}} & \frac{\tilde{h}_{Mb2}}{\tilde{h}_{Mbi}} & \cdots & 0 \end{bmatrix} \begin{bmatrix} \tilde{\gamma}_1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\gamma}_M \end{bmatrix} \begin{bmatrix} \tilde{p}_1 \\ \tilde{p}_2 \\ \vdots \\ \tilde{p}_m \end{bmatrix} + \begin{bmatrix} \frac{\tilde{\sigma}_1^2}{\tilde{h}_{ibi}} \\ \frac{\tilde{\sigma}_2^2}{\tilde{h}_{2bi}} \\ \vdots \\ \frac{\tilde{\sigma}_M^2}{\tilde{h}_{Mbi}} \end{bmatrix}.$$  \hspace{1cm} (5.24)$$

2 It can be now seen why describing the path-gain matrix as an $M \times M$ matrix makes it easier to write the relationship between transmit power and SINR vectors.
Using matrix notations, the relation between the downlink transmit power vector and its corresponding downlink SINR vector can be rewritten as

\[ \tilde{p} = D(\tilde{\gamma})\tilde{G}p + D(\tilde{\gamma})\tilde{\eta} \]  
(5.25)

where the \((i, j)\) component of \(\tilde{G}\) is

\[ \tilde{G}_{i,j} = \begin{cases} \tilde{h}_{jb}, & \text{if } i \neq j \\ \tilde{h}_{ib}, & \text{if } i = j \end{cases} \]  
(5.26)

and the \((i)\) component of \(\tilde{\eta}\) is

\[ \tilde{\eta}_i = \frac{\tilde{\sigma}_i^2}{\tilde{h}_{ib}}. \]  
(5.27)

From (5.25),

\[ (I - D(\tilde{\gamma})\tilde{G})\tilde{p} = D(\tilde{\gamma})\tilde{\eta} \]  
(5.28)

where \(I\) is an \(M \times M\) identity matrix. Given a downlink SINR vector, the corresponding downlink transmit power vector is thus obtained as

\[ \tilde{p} = (I - D(\tilde{\gamma})\tilde{G})^{-1}D(\tilde{\gamma})\tilde{\eta}. \]  
(5.29)

Therefore, to obtain the downlink transmit power vector corresponding to a given downlink SINR vector \(\tilde{\gamma}\), we require calculating the inverse of matrix \(I - D(\tilde{\gamma})\tilde{G}\).

5.2 System Model for a Single-Cell Wireless Network

The relations in (5.22) and (5.29) give the transmit power vector corresponding to a given SINR vector for uplink and downlink cases, respectively. These relations are for a general multi-cell wireless network. However, for the uplink of a single-cell wireless network in which there exists a single BS, a simpler relation for the transmit power vector as a function of the SINR vector can be derived, as explained below. First, for a single-cell scenario, we will discuss how the uplink and downlink path-gains can be described in a simpler manner.

5.2.1 Modeling Path-Gains

A single-cell system model for a wireless network corresponds to the case in which only a single cell is considered and the interference from other cells on the cell under consideration is ignored or assumed to be fixed. For such a single-cell system model, the uplink and downlink path-gains for each user can be described with a single index of that user. Since only a single BS is considered, it does not need to be indexed. Let \(h_i\) and \(\tilde{h}_i\) denote the uplink and downlink path-gain between user \(i\) and the BS, respectively. In fact, for a single-cell system model, we can omit the terms related to assigned BS
(i.e., $b_i$) in all the relations mentioned for the general system model of multi-cell wireless networks.

**An Uplink Single-Cell Network**

For a given uplink transmit power vector $p$, the uplink SINR of a user $i$ is

$$\gamma_i(p) \triangleq \frac{p_i h_i}{\sum_{j \neq i} p_j h_j + \nu}$$  \hspace{1cm} (5.30)

where $\nu = I^{\text{int}} + \sigma^2$ in which $I^{\text{int}}$ is the intracell interference and $\sigma^2$ is the noise power at the BS. The total received power plus noise at the BS is

$$\varphi^T(p) = \sum_{i \in M} p_i h_i + \nu.$$  \hspace{1cm} (5.31)

The effective uplink SINR of user $i$ is

$$\theta_i(p) = \frac{p_i h_i}{\varphi^T(p)}.$$  \hspace{1cm} (5.32)

It is seen from (5.32) that by deriving the total received power as a function of the SINR vector, the uplink transmit power level for each user would be derived as a function of the uplink SINR vector. For the single-cell model, we can derive the total uplink received power as a simple function of the uplink SINR vector as explained below.

From (5.32), we have

$$p_i h_i = \theta_i(p) \varphi^T(p).$$  \hspace{1cm} (5.33)

Thus by summing the above term over the set of all UEs, we have

$$\sum_{i \in M} p_i h_i = \sum_{i \in M} \theta_i(p) \varphi^T(p).$$  \hspace{1cm} (5.34)

Since from (5.31) we know $\sum_{i \in M} p_i h_i = \varphi^T(p) - \nu$, by putting it into (5.34) we have

$$\varphi^T(p) - \nu = \varphi^T(p) \sum_{i \in M} \theta_i(p),$$

and consequently, the total received power is obtained as the following function of the SINR vector:

$$\varphi^T(p) = \frac{\nu}{1 - \sum_{k \in M} \theta_k}.$$  \hspace{1cm} (5.35)

Thus the transmit power level for each user $i$ is obtained as a function of the SINR vector:

$$p_i = \frac{\theta_i}{h_i} \times \frac{\nu}{1 - \sum_{k \in M} \theta_k} = \frac{\gamma_i}{h_i(\gamma_i + 1)} \times \frac{\nu}{1 - \sum_{k \in M} \frac{\gamma_k}{\gamma_k + 1}}, \quad \forall i \in M.$$  \hspace{1cm} (5.36)

In fact, the transmit power level given by (5.36) is the same as that obtained by (5.22) for a single-cell system model.
A Downlink Single-Cell Network

For a given downlink transmit power vector $\tilde{p}$, the uplink SINR of a user $i$ is

$$\tilde{\gamma}_i(\tilde{p}) = \frac{\tilde{p}_i \tilde{h}_i}{\tilde{h}_i \sum_{j \neq i} \tilde{p}_j + \tilde{v}_i}$$

(5.37)

where $\tilde{v}_i = \tilde{I}_{\text{int}} + \tilde{\sigma}_i^2$ in which $\tilde{I}_{\text{int}}$ is the intracell interference and $\tilde{\sigma}_i^2$ is the noise power at the user $i$'s receiver. The total downlink transmit power is

$$\tilde{P}_T(\tilde{p}) = \sum_{i \in M} \tilde{p}_i.$$  

(5.38)

The effective downlink SINR of user $i$ is

$$\tilde{\theta}_i(\tilde{p}) = \frac{\tilde{p}_i \tilde{h}_i}{\tilde{h}_i \tilde{P}_T(\tilde{p}) + \tilde{v}_i}.$$  

(5.39)

Thus we have

$$\tilde{p}_i = \tilde{\theta}_i P_T + \tilde{\theta}_i \tilde{v}_i.$$  

(5.40)

By taking sum of two sides of (5.40) over the set of all UEs, we have

$$\sum_{j \in M} \tilde{p}_j = P_T \sum_{j \in M} \tilde{\theta}_j + \sum_{j \in M} \tilde{\theta}_j \tilde{v}_j.$$  

(5.41)

Since $\tilde{P}_T = \sum_{j \in M} \tilde{p}_j$, from the above we conclude that

$$P_T = \frac{\sum_{j \in M} \tilde{\theta}_j \tilde{v}_j}{1 - \sum_{j \in M} \tilde{\theta}_j}.$$  

(5.42)

Thus the transmit power level for each user $i$ is obtained as a function of the downlink SINR vector:

$$\tilde{p}_i = \frac{\tilde{\theta}_i}{\tilde{h}_i} \left( \frac{\tilde{h}_i \sum_{j \in M} \tilde{\theta}_j \tilde{v}_j}{1 - \sum_{j \in M} \tilde{\theta}_j} + \tilde{v}_i \right) = \frac{\tilde{\gamma}_i}{\tilde{h}_i} \left( \frac{\tilde{h}_i \sum_{j \in M} \tilde{\gamma}_j \tilde{v}_j}{\tilde{h}_i \tilde{\gamma}_i + 1} + \tilde{v}_i \right), \quad \forall i \in M.$$  

(5.43)

In fact, the transmit power level given by (5.43) is the same as that obtained by (5.29) for a single-cell system model.

5.3 SINR Feasibility in Interference-Limited Wireless Networks

In what follows, in order to introduce the fundamental concept of SINR feasibility in an interference-limited wireless network, we first examine under what condition a positive
transmit power vector exists corresponding to a given SINR vector and then consider additional constraints (such as the maximum transmit power constraint) on transmit power (other than the positivity) and finally introduce the concept of SINR feasibility, which is of great importance in many resource allocation problems in wireless networks. In order to make the understanding of these concepts easier, we consider first a single-cell and then a general multi-cell wireless network.

5.3.1 Existence of a Positive Transmit Power Vector Corresponding to a Given SINR Vector

It is obvious that the transmit power level should have a positive value. Now an important question is: For any arbitrary positive SINR vector, is there any positive transmit power vector that corresponds to that SINR? It might seem that when there is no upper limit for transmit power levels, any arbitrary SINR vector is achievable by a positive transmit power vector. However, this is not true. In fact, any arbitrary SINR vector is not necessarily achievable, even if the transmit power levels are allowed to be increased without any upper limitation. To see this, let us consider two UEs in a single-cell network. The relation between the uplink SINR and transmit power of these two UEs is as follows:

\[
\gamma_1 = \frac{p_1 h_1}{p_2 h_2 + v} \Rightarrow p_1 = \gamma_1 (p_2 h_2 + v) \quad (5.44)
\]

\[
\gamma_2 = \frac{p_2 h_2}{p_1 h_1 + v} \Rightarrow p_2 = \gamma_2 (p_1 h_1 + v). \quad (5.45)
\]

Figure 5.1 illustrates two lines, one for \(p_1\) versus \(p_2\) and one for \(p_2\) versus \(p_1\). When \(\gamma_1 \gamma_2 < 1\), these two lines intersect with each other in a positive region, and when \(\gamma_1 \gamma_2 > 1\), they intersect with each other in a negative region. This means that if \(\gamma_1 \gamma_2 > 1\), the SINR values of \(\gamma_1\) and \(\gamma_2\) are not simultaneously reachable, i.e., there exist no positive values for \(p_1\) and \(p_2\) to make the SINR values of \(\gamma_1\) and \(\gamma_2\) reachable. The achievable region of SINR levels for an uplink case with two UEs is illustrated in Figure 5.2. In fact,
the transmit power space $\{[p_1, p_2] | p_1 \geq 0, p_2 \geq 0\}$ is mapped into the corresponding SINR space, $\{[\gamma_1, \gamma_2] | \gamma_1 \gamma_2 < 1\}$.

In general, from (5.36), we conclude that, for a single-cell wireless network, for a given uplink target-SINR $\gamma \geq 0$, if and only if the following condition holds:

$$\sum_{j \in M} \frac{\gamma_j}{\gamma_j + 1} < 1 \quad (5.46)$$

then there exists a positive uplink transmit power vector corresponding to the uplink SINR vector $\gamma$. Similarly, if and only if

$$\sum_{j \in M} \frac{\tilde{\gamma}_j}{\tilde{\gamma}_j + 1} < 1 \quad (5.47)$$

then there exists a positive downlink transmit power vector corresponding to the downlink SINR $\tilde{\gamma} \geq 0$.

For a general multi-cell wireless network, the following theorem shown by J. Zander in 1992 [2] states under what condition a given SINR vector is achievable by a positive transmit power vector. Let $\rho(A)$ denote the spectral radius of matrix $A$. The spectral radius of a matrix is the maximum of the absolute value of its eigenvalues.

**Theorem 42** For a given uplink SINR vector $\gamma \geq 0$, there exists a corresponding positive power vector $p \geq 0$, if $\rho(GD(\gamma)) < 1$.

The same results can be concluded for the downlink case.

**Theorem 43** For a given downlink SINR vector $\tilde{\gamma} \geq 0$, there exists a corresponding positive power vector $\tilde{p} \geq 0$, if $\rho(\tilde{GD}(\tilde{\gamma})) < 1$. 

---

**Figure 5.2** The feasible region of SINR levels for an uplink case with two UEs.
With some mathematical operations similar to the derived relations for a single-cell network (i.e., the relations in (5.30)–(5.36)), one may obtain a relation for the transmit power level of each user $i \in \mathcal{M}$ as a function of the achieved SINR values for the UEs in $\mathcal{C}_h$ as

$$p_i = \frac{\gamma_i}{h_{bi}(\gamma_i + 1)} \times \frac{s_i^2 + I_{ext}^i}{1 - \sum_{j \in \mathcal{C}_h} \frac{\gamma_j}{\gamma_j + 1}}, \quad \forall i \in \mathcal{M}. \quad (5.48)$$

Similarly, for the downlink, we have

$$\tilde{p}_i = \frac{\tilde{\gamma}_i}{\tilde{h}_{bi}(\tilde{\gamma}_i + 1)} \left( \frac{\tilde{\gamma}_j \tilde{h}_{bj} + 1}{1 - \sum_{j \in \mathcal{C}_h} \frac{\tilde{\gamma}_j}{\tilde{\gamma}_j + 1}} + \tilde{\nu}_i \right), \quad \forall i \in \mathcal{M}. \quad (5.49)$$

Thus for a given uplink SINR vector $\gamma \geq 0$, if there exists a corresponding transmit power vector $\mathbf{p} \geq 0$, then we have

$$\sum_{j \in \mathcal{C}_k} \frac{\gamma_j}{\gamma_j + 1} < 1, \quad \forall k \in \mathcal{K}. \quad (5.50)$$

Also, for a given downlink SINR vector $\tilde{\gamma} \geq 0$, if there exists a corresponding transmit power vector $\tilde{\mathbf{p}} \geq 0$, then we have

$$\sum_{j \in \mathcal{C}_k} \frac{\tilde{\gamma}_j}{\tilde{\gamma}_j + 1} < 1, \quad \forall k \in \mathcal{K}. \quad (5.51)$$

If the conditions $\rho(\text{GD}(\gamma)) < 1$ and $\rho(\tilde{\text{GD}}(\tilde{\gamma})) < 1$ (stated in Theorems 42 and 43, respectively) hold, then the conditions in (5.50) and (5.51) hold, respectively.

A wireless network where multiple transmitters cause interference to each other is called an interference-limited wireless network, which implicitly means even if there is no constraint on maximum transmit power, a given SINR vector may not be still achievable, due to the inherently limited interference that transmitters can cause to each other.

### 5.3.2 Existence of a Constrained Transmit Power Vector Corresponding to a Given SINR Vector

What we have discussed so far can be summarized as follows. Assuming no constraint on upper limit of the transmit power levels, as long as the conditions stated in Theorems 42 and 43 hold for uplink and downlink, respectively, the corresponding SINR vector is simultaneously achievable by a positive transmit power vector. However, if there is an additional constraint on transmit power levels, say, an upper bound on the maximum transmit power levels, under what condition is a given SINR reachable? We are now ready to introduce the concept of SINR feasibility. In what follows, we first define the feasible set of transmit power vectors and then introduce the concept of SINR feasibility and explain how it can be checked.
Besides the positivity of transmit power levels, there are usually additional constraints on transmit power vector. These additional constraints are imposed due to hardware or regulatory limitations. They may include maximum transmit power for each user, which is usually imposed on uplink transmit power, or a constraint on the maximum aggregate (total) power, which is usually imposed on downlink transmit power by the BS. Different kinds of constraints bring different levels of difficulty in solving the power control problems in wireless networks. If a positive transmit power vector also satisfies the other imposed constraints on transmit power vector, we say that it is feasible.

Definition 36  A given uplink transmit power vector $p$ is called feasible if $p \in P$ where $P = \{p | 0 \leq p \leq \tilde{p}\}$, in which $\tilde{p} = [\tilde{p}_1, \ldots, \tilde{p}_M]$ is the maximum uplink transmit power vector. A given downlink transmit power vector $\tilde{p}$ is called feasible if $\tilde{p} \in \tilde{P}$, where $\tilde{P} = \{\tilde{p} | \tilde{p} \geq 0, \sum_{i \in C_k \tilde{p}_i \leq \tilde{P}_k, \forall k \in K}\}$ in which $\tilde{P}_k$ is the maximum total transmit power for BS $k$.

Note that while the above constraints on uplink and downlink transmit power levels are more common, one may consider other constraints and hence define different feasible sets of transmit power vectors. For example, $\tilde{P} = \{p | 0 \leq \tilde{p} \leq \tilde{p}, \sum_{i \in C_k} \tilde{p}_i \leq \tilde{P}_k, \forall k \in K\}$, where $\tilde{p}$ is the maximum downlink transmit power vector, $\tilde{P} = \{p | 0 \leq p \leq \tilde{p}, \sum_{i \in C_k} p_i h_{ki} \leq P_k, \forall k \in K\}$ in which $P_k$ is the maximum value for the total uplink received power at the BS $k$.

Now an important question is: For any arbitrary SINR vector, does there exist a corresponding feasible transmit power vector? What is the SINR feasibility region? Prior to answering this question, we define the SINR feasibility as follows.

Definition 37  A given (uplink or downlink) SINR vector is feasible if a feasible transmit power vector exists that corresponds to the SINR vector.

To examine the feasibility of a given uplink or downlink SINR vector in a multi-cell wireless network, we need to obtain the corresponding positive transmit power vector (if it exists) corresponding to the SINR vector, by using (5.22) and (5.29) for uplink and downlink cases, respectively, and check whether the obtained transmit power vector is feasible or not. This means that a given uplink SINR vector $\gamma$ is feasible if $p \in P$, where $p = (I - D(\gamma)G)^{-1}D(\gamma)\eta$. Similarly, a given downlink SINR vector $\tilde{\gamma}$ is feasible if $\tilde{p} \in \tilde{P}$, where $\tilde{p} = (I - D(\tilde{\gamma})\tilde{G})^{-1}D(\tilde{\gamma})\tilde{\eta}$.

For a single-cell wireless network, for a given SINR, we can find the corresponding transmit power vector by using the simpler functions of (5.36) and (5.43), for uplink and downlink cases, respectively. A given uplink SINR vector $\gamma$ is feasible if

$$0 \leq \frac{\gamma_i}{h_i(\gamma_i + 1)} \times \frac{v}{1 - \sum_{j \in M} \gamma_j + 1} \leq \tilde{p}_i, \quad \forall i \in M. \quad (5.52)$$

Note that the left side inequality is equivalent to the inequality in (5.46). A given downlink SINR vector $\gamma$ is feasible if (5.47) holds and $\sum_{i \in M} \tilde{p}_i \leq \tilde{P}$, where $\tilde{P}$ is the maximum possible value of aggregate transmit power consumed by the BS, and $\tilde{p}_i$ is given...
by (5.43), that is,
\[
\sum_{i \in \mathcal{M}} \left( \frac{\tilde{y}_i}{h_i(\tilde{y}_i + 1)} \left( \tilde{h}_i \sum_{j \in \mathcal{M}} \frac{\tilde{y}_j}{\tilde{y}_j + 1} + \tilde{v}_i \right) \right) \leq \tilde{P}
\]
(5.53)
or equivalently,
\[
\frac{\sum_{j \in \mathcal{M}} \frac{\tilde{y}_j}{\tilde{y}_j + 1} \frac{\tilde{v}_j}{\tilde{y}_j + 1}}{1 - \sum_{j \in \mathcal{M}} \frac{\tilde{y}_j}{\tilde{y}_j + 1}} \sum_{j \in \mathcal{M}} \frac{\tilde{y}_j}{\tilde{y}_j + 1} + \sum_{j \in \mathcal{M}} \frac{\tilde{y}_j}{\tilde{y}_j + 1} \frac{\tilde{v}_j}{\tilde{h}_j} \leq \tilde{P}.
\]
(5.54)

5.4 Exercises

Exercise 5.1: Consider a CDMA cellular wireless network with \(M\) active UEs denoted by \(\mathcal{M} = \{1, 2, \ldots, M\}\). Let \(p_i\) denote the transmit power of UE \(i\), where \(0 \leq p_i \leq \overline{p}_i\), and \(\overline{p}_i\) denotes the maximum transmit power of UE \(i\).

i. Assume two UEs \(i, j \in \mathcal{M}\) served by the same BS, i.e., \(b_i = b_j\). Show that the uplink SINR values for two UEs \(i\) and \(j\) are the same (i.e., \(\gamma_i(p) = \gamma_j(p)\) holds), if and only if their received power levels at that BS are also equal (i.e., \(p_i h_{bi} = p_j h_{b,j}\) holds).

ii. Use the result obtained above to find the maximum value of equal SINR that is obtained by solving the following power control problem referred to as the max-equal SINR power control problem:

\[
\text{max-equal SINR: } \max_{0 \leq p \leq \overline{p}} \{ \gamma | \gamma_i(p) = \gamma, \text{ for all } i \in \mathcal{M} \}.
\]

iii. Obtain the maximum value of equal SINR defined above for the unlimited transmit power case (i.e., \(\overline{p}_i \to \infty\)).

Exercise 5.2: Consider two SINR vectors \(\gamma\) and \(\gamma'\), which correspond to the power vectors \(p\) and \(p'\), respectively. Show that if \(\gamma \geq \gamma'\) then \(p \geq p'\) holds. But the reverse is not necessarily true, i.e., \(p \geq p'\) does not result in \(\gamma \geq \gamma'\) in general.

Exercise 5.3: For any two SINR vectors \(\gamma\) and \(\gamma'\) that satisfies \(0 \leq \gamma' \leq \gamma\), show that if \(\gamma\) is feasible then \(\gamma'\) is feasible as well, i.e., the feasible SINR region is downward comprehensive.

Exercise 5.4: Show that the feasible SINR region is not convex in general, i.e., for any two feasible SINR vectors \(\gamma\) and \(\gamma'\), the SINR vector \(\lambda \gamma + (1 - \lambda)\gamma'\), where \(0 < \lambda < 1\), may not be feasible [3]. Also, show that in the case of unlimited power, the feasible SINR region is log-convex. (This was shown in [4].)
References


6 Power Control in Cellular Wireless Networks

For a wireless network, the transmit power is one of the main radio resources. Two major objectives of power control in a wireless network are to extend UEs’ battery life and to maintain an acceptable QoS (e.g., in terms of the SINR or throughput) for all UEs by minimizing interferences to UEs. Data services require a higher SINR (as a measure of QoS) as compared to the voice service, because the latter is more tolerant to bit errors. In contrast to the voice service for which the QoS is measured by a step function of the SINR [1], the commonly used QoS measure for the data service is, in general, an increasing function of the SINR.

A distributed scheme for power control is preferred to a centralized one, because in the former, the transmit power level of a user is decided by that user by utilizing the locally available information and uses minimal feedback from the BS. In this way, the need for frequent power setting commands by the BS are avoided, and the processing capabilities at the BS needed to obtain the instantaneous uplink power levels of all UEs are substantially reduced. In contrast, a centralized approach needs to have information about path-gains and throughput requirements for all UEs at the BS.

In this chapter, we first discuss why power control is needed and state the objectives of power control, followed by a discussion on conventional open/closed loop and centralized power control algorithms. Then various existing distributed power control algorithms are presented and evaluated according to different criteria.

6.1 Objectives of Power Control

In Chapter 5, we studied the relation between transmit power and SINR vectors. We know that the achieved SINRs by UEs at uplink or downlink determine their experienced QoS. Now the question is how the predetermined target QoS for users can be achieved. The transmit power cannot be set at random or at a fixed level by the UEs. For example, if the UEs served by the same cell transmit at the same fixed power level, the SINR for the UEs with good path-gains (e.g., those near the BS) are high, whereas for the far UEs it will be low. This is called the near-far problem, which can be addressed if the transmit power is dynamically controlled.

The transmit power and consequently the interference in interference-limited wireless networks are required to be controlled according to a power control algorithm. In addition to interference management, power control would extend the battery life of UEs.
and reduce energy consumption. In other words, two major functionalities of power control in a wireless network are energy management to extend UEs’ battery life and interference management to optimize a given network performance measure (objective function).

Design of a power control algorithm depends on the network objective, i.e., different power control schemes are designed and adopted for different objectives. For example, from a user’s point of view, the objective of power control may be to support a UE with its minimum acceptable throughput, whereas from a system’s point of view it may be to maximize the aggregate throughput. As will be seen in next sections, these two objectives are orthogonal to one another. In the former, it is required to compensate for the near-far effect by allocating higher power levels to UEs with poor channels as compared to UEs with good channels. In the latter, high power levels are allocated to a few UEs with the best channels and very low (even zero) power levels to others.

In the next section, we will introduce the performance measures and objective functions, based on which the corresponding power control optimization problems will be formally stated and their optimal solutions will be presented, followed by a discussion about comparing centralized versus distributed schemes for addressing a given power control optimization problem.

6.1.1 Performance Measure and Objective Functions

Different power control schemes are evaluated and compared against each other in terms of some specific performance measures. The most important measures to compare the performance of different power control schemes are aggregate transmit power, outage ratio, fairness, and the aggregate throughput (i.e., the sum of achievable rates by the UEs), which will be explained below.

Aggregate Transmit Power
The aggregate transmit powers for uplink and downlink are simply defined as \( \sum_{i \in \mathcal{M}} P_i \) and \( \sum_{k \in \mathcal{K}} \tilde{P}_i \), respectively. This performance measure is more important at the uplink in comparison with the downlink. Under similar conditions, for two power control algorithms, the one with less aggregate transmit power is preferred to the other.

Outage Ratio
For any modulation scheme employed in the network, the values of the minimum acceptable transmit data rate and maximum tolerable bit-error-rate (BER) for each user \( i \in \mathcal{M} \), corresponds to a minimum acceptable SINR for that user, which we call the target-SINR. In other words, in order to have a data rate higher than a threshold or a BER lower than a threshold for a given user, the SINR of that user needs to be greater than a so called target-SINR. The target-SINR of user \( i \) is denoted by \( \hat{\gamma}_i \), which usually corresponds to a maximum tolerable BER. Data services require a higher SINR as compared to the voice service, because the latter is more tolerant to bit errors.
6.1 Objectives of Power Control

Given a transmit power vector \( \mathbf{p} \), user \( i \) is supported if \( \gamma_i(\mathbf{p}) \geq \gamma_i \). Let us denote the set of supported UEs by a given a transmit power vector \( \mathbf{p} \), as \( S(\mathbf{p}) = \{ i \in \mathcal{M} | \gamma_i(\mathbf{p}) \geq \gamma_i \} \). Its complementary set is \( S'(\mathbf{p}) = \mathcal{M} - S(\mathbf{p}) \). The cardinality of a given set \( \mathcal{A} \) is denoted by \( |\mathcal{A}| \). Given a transmit power vector \( \mathbf{p} \), the outage-ratio for all UEs is given by \( O(\mathbf{p}) \), as follows:

\[
O(\mathbf{p}) = \frac{|S'(\mathbf{p})|}{|\mathcal{M}|}.
\] (6.1)

It is obvious that when the target-SINRs of UEs are feasible, there exists a transmit power vector \( \mathbf{p} \) such that \( O(\mathbf{p}) = 0 \), otherwise there is no such transmit power vector. Similarly, given a downlink transmit power vector \( \tilde{\mathbf{p}} \), the outage-ratio \( O(\tilde{\mathbf{p}}) \) for downlink is

\[
O(\tilde{\mathbf{p}}) = \frac{|\tilde{S}(\tilde{\mathbf{p}})|}{|\mathcal{M}|},
\] where \( \tilde{S}(\tilde{\mathbf{p}}) = \{ i \in \mathcal{M} | \tilde{\gamma}_i(\tilde{\mathbf{p}}) \geq \tilde{\gamma}_i \} \).

**Throughput**

In contrast to voice service for which the throughput is measured by a step function of the SINR, the commonly used throughput measure for data service is, in general, an increasing function of the SINR. This is because reaching an SINR value higher than a given target value has no practical effect on the service quality of voice, whereas for data services, a higher SINR results in a better throughput.

An information theoretic approach is commonly used to define the (data) throughput for each user \( i \) by the channel capacity as the highest rate at which user \( i \)'s information can be sent through the channel with an arbitrary low probability of error, i.e., assuming that the interference process is similar to the noise process,

\[
T_i(\mathbf{p}) = k \log_2(1 + \gamma_i(\mathbf{p}))
\] (6.2)

where \( k \) is a constant. This logarithmic function is the capacity of a Gaussian channel, provided that the noise plus interference for each user is Gaussian [2] \( k = W/2 \) for a discrete-time channel, and \( k = W \) for a continuous-time channel [2], where \( W \) is the channel bandwidth). Also, restricting ourselves to no specific channel model, modulation, and coding scheme, the throughput for user \( i \), i.e., \( T_i \), can be in general defined as a function of SINR, as \( T_i(\gamma_i) \). \( T_i(\gamma_i) \) is increasing and concave with respect to \( \gamma_i \) for every channel model with the average power constraint [3].

The system (aggregate) throughput defined as the sum of achievable rates by UEs is given by

\[
T(\mathbf{p}) = \sum_i T_i(\mathbf{p}).
\] (6.3)

**Fairness**

Fairness is an important notion in allocating resources in wireless data networks. There is no unique criterion for fairness, since it is highly application dependent. To compare a set of allocated transmit power vectors in terms of fairness, one may regard the transmit power vector with a higher value of minimum SINR achieved by UEs as more fair compared to the other ones. For example, to compare two transmit power vectors, say, \( \mathbf{p} \) and \( \mathbf{p}' \), in terms of fairness, we say \( \mathbf{p} \) is more fair than \( \mathbf{p}' \) if \( \min_{i \in \mathcal{M}} \gamma_i(\mathbf{p}) > \min_{i \in \mathcal{M}} \gamma_i(\mathbf{p}') \).
6.1.2 Distributed Versus Centralized Approach

A given power control problem can be addressed in either a centralized or a distributed manner, by using the conventional optimization techniques and designing a distributed iterative power-updating algorithm, respectively. In the centralized approach, it is assumed that the full knowledge of all required parameters is available at a single node where the problem is to be solved. In the distributed approach, it is assumed that each transmitter sets its own transmit power level using its own locally available information, and a few nonlocally available parameters are obtained via minimal feedback and message passing mechanisms with other nodes. In fact, in many distributed power control algorithms, in order to optimize any of the aforementioned global objective functions, in addition to using the locally available information, minimal feedback and message passing between users may be required. As will be seen in Section 6.4 where we discuss the existing distributed power control algorithms, one criterion for comparing distributed algorithms is how much nonlocally available information a distributed algorithm requires to be provided with via feedback from the receivers and/or message passing between nodes.

A distributed power control, specially at the uplink, is practically preferred to a centralized one, because in the former, the transmit power level of a transmitter is decided by the transmitter utilizing the local information and minimal feedback from its corresponding receiver. In contrast, a centralized approach needs to have information about path-gain and throughput requirements for all transmitters. In distributed power control schemes, available resources including the frequency spectrum and processing capabilities are more efficiently used. This is because the need for frequent feedback from the receivers and messages (e.g., power levels, path-gains, and throughput requirements) passing among transmitters is avoided, and the processing capabilities at each transmitter needed for sending and receiving instantaneous messages to and from other transmitters, which may be quite significant, are substantially reduced.

6.2 Different Power Control Optimization Problems

There is a wide variety of power control problems with different objective functions, an exhaustive coverage of which is beyond the scope of this book. In this chapter, we consider the most important state-of-the-art power control optimization problems, which are chosen for the following reasons. First, these problems and their corresponding solutions reflect the main characteristics of the state-of-the-art power control problems and their centralized or distributed solutions. Second, their corresponding objective functions and constraints are also common in the majority of other resource allocation problems such as cell association, carrier assignation, and scheduling. Third, they are good samples through which a systematic view of looking at resource allocation problems in wireless networks can be provided. This would help readers understand how to classify and compare the problems to find a research gap and how to extend the existing solutions to new network paradigms. For these reasons, and within the two functionalities
for power control, namely, energy and interference management, several kinds of power control problems are considered. These include the problems of minimizing aggregate power required to maintain a given target QoS (in terms of a given target-SINR) for UEs, maximizing aggregate throughput, minimizing outage ratio when the required target QoSs for all UEs are infeasible, maximizing aggregate throughput subject to a given target-SINR, and providing max-min fairness.

**Aggregate Power Minimization Subject to Given Target-SINR Constraints**

The problem of minimizing aggregate power subject to the constraint of supporting all UEs with their given target-SINRs is formally defined as

$$\min \sum_i p_i$$

s.t. $\gamma_i(p) \geq \hat{\gamma}_i, \forall i \in M$

$$0 \leq p \leq \bar{p}$$

variable $p$

where $0 \leq p \leq \bar{p}$ implies $0 \leq p_i \leq \bar{p}_i$ for all $i \in M$, in which $\bar{p}_i$ is the maximum transmit power for UE $i$. In this problem, it is implicitly assumed that the target-SINRs for all users are feasible, and the objective is to support all users with their target-SINRs, consuming the minimum aggregate transmit power. When the target-SINRs are feasible, the optimal (power vector) solution to problem (6.4) is given by the transmit power vector corresponding to the SINR vector, obtained by $p = (I - D(\gamma)G)^{-1} D(\gamma)\eta$, derived in Chapter 5 (via relation (5.22)).

**Minimization of Outage Ratio**

When the required target-SINRs for users are infeasible (i.e., the system is infeasible), the problem in (6.4) has no solution, because there is no feasible transmit power vector that can satisfy SINR requirements for all users. In an infeasible system, the minimum number of users should be removed to make the target-SINRs feasible for remaining users. This corresponds to minimizing the outage ratio, which is formally defined as

$$\min O(P)$$

s.t. $0 \leq p \leq \bar{p}$

variable $p$.

As has been stated earlier, in general, the minimum-outage problem is NP-complete. In what follows, for the special case of a single-cell system in which the target-SINRs of all users are the same, the solution to the minimum-outage problem is provided for a single cell.

The target-SINRs of users in a given subset $B \subseteq M$ are feasible if a power vector $0 \leq p \leq \bar{p}$ exists that satisfies the target-SINRs of users in $B$. Given the target-SINR of each user in the set $B \subseteq M$, by using the relation (5.36) obtained in Chapter 5, we
conclude that their target-SINRs are feasible if
\[
0 \leq \frac{\hat{\gamma}_i}{h_i(\hat{\gamma}_i + 1)} \times \frac{v}{1 - \sum_{k \in B} \frac{\gamma_k}{\gamma_k + 1}} \leq \bar{p}_i, \quad \forall i \in B. \tag{6.6}
\]

Suppose all users have a common target-SINR, i.e., \(\hat{\gamma}_i = \hat{\gamma}\), where \(\hat{\gamma}\) is a positive constant. From (6.6) we conclude that target-SINRs of users in \(B \subseteq M\) are feasible if
\[
0 \leq \frac{v\hat{\gamma}}{h_i(1 - (|B| - 1)\hat{\gamma})} \leq \bar{p}_i, \quad \forall i \in B. \tag{6.7}
\]

When \(|B|\) users are active and their SINRs are \(\hat{\gamma}\), the received power at the BS for all of them is \(\bar{\phi} = \frac{v\hat{\gamma}}{1 - (|B| - 1)\hat{\gamma}}\).

**Theorem 44 [4]** Let \(\bar{\varphi}_i = \bar{p}_i h_i\) denote the max-received power for user \(i\), and without loss of generality, suppose that users are indexed in an increasing order of their max-received power, i.e., \(\bar{\varphi}_1 < \bar{\varphi}_2 < \cdots < \bar{\varphi}_M\). Define
\[
\Gamma_i \triangleq \frac{\bar{\varphi}_i}{(M - i)\bar{\varphi}_i + v}, \quad i = 1, 2, \ldots, M. \tag{6.8}
\]

The minimum-outage, denoted by \(O^*\) for different values of the common target-SINR \(\hat{\gamma}\), is
\[
O^* = \begin{cases} 
0, & \text{if } 0 \leq \hat{\gamma} \leq \Gamma_1 \\
\frac{l}{M}, & \text{if } \Gamma_1 < \hat{\gamma} \leq \Gamma_{l+1}, \text{ where } l \in \{1, 2, \ldots, M - 1\} \\
1, & \text{if } \hat{\gamma} > \Gamma_M.
\end{cases} \tag{6.9}
\]

**Proof** Note that \(\Gamma_i\) is the achieved SINR by each user indexed from \(i\) to \(M\) when users indexed from \(1\) to \(i - 1\) switch off and users indexed from \(i\) to \(M\) transmit at a level so that their received power at the BS is \(\bar{\varphi}_i\). We have \(\Gamma_i < \Gamma_j\) for all \(i < j\). It can be easily shown that when \(0 \leq \hat{\gamma} \leq \Gamma_1\), the target-SINR \(\hat{\gamma}\) is reachable by each user \(i\) by transmitting at \(p_i = \frac{v\hat{\gamma}}{h_i(1 - (|B| - 1)\hat{\gamma})} \leq \bar{p}_i\), meaning that the outage ratio is zero. For \(\hat{\gamma} > \Gamma_M\), we have \(\hat{\gamma}_i > \frac{\bar{\varphi}_i}{v}\), and thus \(\hat{\gamma}_i > \frac{\bar{\varphi}_i}{v}\) for all \(i \in M\), which means no user can reach his/her target-SINR and thus the minimum-outage-ratio is 1. If \(\Gamma_1 < \hat{\gamma} \leq \Gamma_M\), then there exists an \(l \in \{1, 2, \ldots, M - 1\}\) for which \(\Gamma_l < \hat{\gamma} \leq \Gamma_{l+1}\). Now we prove that in this case, the minimum number of removals (non-supported users) is \(l\), meaning that the minimum-outage-ratio is \(\frac{l}{M}\). To prove this, we first demonstrate that by removing \(l\) users, the target-SINRs of all remaining users are reachable, and then show that the minimum-number of removals cannot be lower than \(l\). From the right side and the left side of \(\Gamma_l < \hat{\gamma} \leq \Gamma_{l+1}\), we have
\[
\frac{v\hat{\gamma}}{1 - (M - l - 1)\hat{\gamma}} \leq \bar{\varphi}_{l+1} \tag{6.10}
\]
and
\[
\frac{v\hat{\gamma}}{1 - (M - l)\hat{\gamma}} > \bar{\varphi}_l \tag{6.11}
\]
respectively. From (6.10) we conclude that \( \frac{\bar{\psi}}{1-(M-l-1)\hat{\gamma}} \leq \bar{\psi}_i \) for all \( i \in \{l+1, l+2, \ldots, M\} \). Therefore by using (6.7), we conclude that the target-SINRs of users in \( \{l+1, l+2, \ldots, M\} \) are feasible (i.e., they are supported with the common target-SINR, if users 1 to \( l \) switch off and the transmit power for each remaining user \( i \) is set at \( p_i = \frac{\bar{\psi}}{h_i(1-(M-l-1)\hat{\gamma})} \), that also satisfies the power constraint by noting (6.10)). When \( \Gamma_l < \hat{\gamma} \leq \Gamma_{l+1} \), if the number of non-supported users can be lower than \( l \), then at least the target-SINRs of users in \( \{l, l+1, \ldots, M\} \) should be feasible. If this is true, then from (6.7) we have \( \frac{\bar{\psi}}{1-(M-l)\hat{\gamma}} \leq \bar{\psi}_i \) for all \( i \in \{l, l+1, \ldots, M\} \), which contradicts (6.11).

Maximization of Aggregate Throughput
From a system point of view, the goal of power control is to optimize the aggregate throughput subject to the peak transmit power constraint. In the problem of maximizing the aggregate throughput, no predetermined target-SINRs for users are assumed, and the objective is to maximize the aggregate throughput which is formally defined as

\[
\max \sum_i T_i(p) \quad \text{s.t. } 0 \leq p \leq \bar{p}
\]

(6.12)

As has been mentioned before, the throughput for each user is in general an increasing function of the SINR such as \( \log_2(1 + \gamma_i(p)) \). For the general case, the problem of maximizing aggregate throughput is non-convex, and thus it is hard to solve it. Alternatively, one may provide few properties of the optimal solutions and solve it for special cases, as stated in two following theorems, respectively.

Theorem 45 Let \( p^* \) be the optimal power vector to problem (6.12). The optimal transmit power vector will have at least one component equal to the maximum transmit power, i.e., there exists at least one \( i \) for which \( p_i^* = \bar{p}_i \).

As a special case, the value of SINR can be regarded as the throughput, as defined in [5], by considering

\[
\max \sum_i \gamma_i(p) \quad \text{s.t. } 0 \leq p \leq \bar{p}
\]

(6.13)

variable \( p \).

In [6], the aggregate throughput is defined as the aggregate of the variable transmission rates for a given SINR, which is equivalent to (6.13) [5]. The following theorem is proved in [7] (Propositions 2–3).

Theorem 46 [7] Similar to Theorem 44, let \( \bar{\psi}_i = \bar{p}_i h_i \) denote the max-received power for user \( i \), and without loss of generality, suppose that users are indexed in an increasing order of their max-received power, i.e., \( \bar{\psi}_1 < \bar{\psi}_2 < \cdots < \bar{\psi}_M \). If \( \frac{\bar{\psi}_M}{\sigma^2} \geq 1 \), then the
optimal solution to the optimum aggregate SINR problem is \( p_M^* = \bar{p}_M \) and \( p_i^* = 0 \) for \( i = 1, 2, \ldots, M - 1 \).

Note that \( \frac{\mu u}{\sigma^2} \geq 1 \) is usually satisfied, which we assume here as well. Thus for the optimum aggregate SINR, only the UE with the highest maximum received-power transmits at its maximum power while the remaining UEs do not transmit at all. Although this strategy maximizes the aggregate throughput, it may be extremely unfair to UEs with low maximum-received power, who may never achieve a satisfactory SINR (QoS). Thus in general, the optimum aggregate SINR (6.13) and the max-min QoS, which will be defined in (6.14) as a measure of fairness, do not have the same solution. Usually, a higher aggregate throughput is achieved at the expense of fairness and vice versa.

Although no minimum QoS is guaranteed for UEs, this optimization problem can be justified where a given operator wants to maximize his/her revenue by maximizing the total aggregate data throughput per used carrier (channel). Considering a single channel only, this optimization problem magnifies the unfairness in that channel. However, it is worth noting that maximizing the aggregate throughput per channel may not result in overall unfairness, since due to channel diversity, a given UE is likely to have the best channel condition in one of the available channels.

### Fairness
The formulation of a fair power control problem can be defined as the max-equal QoS to find a transmit power vector so that the minimum achievable QoS is maximized:

\[
\begin{align*}
\max \min_i T_i(p) \\
\text{s.t. } 0 \leq p & \leq \bar{p} \\
\text{variable } p.
\end{align*}
\]  

(6.14)

Since the throughput is an increasing function of the SINR, the above problem corresponds to

\[
\begin{align*}
\max \min_i \gamma_i(p) \\
\text{s.t. } 0 \leq p & \leq \bar{p} \\
\text{variable } p.
\end{align*}
\]  

(6.15)

### Maximization of Aggregate Throughput Subject to Given Target-SINR Constraints
On one hand, addressing the problem of aggregate power minimization subject to users’ target-SINRs constraint, i.e., (6.4), causes users to exactly hit their fixed target-SINRs in feasible systems even if additional resources are still available that can otherwise be used to achieve higher SINRs (and thus better throughputs). In addition, the fixed-target-SINR assignment is suitable usually for voice service, for which reaching an SINR value higher than the given target value has no practical effect on service quality (due to characteristics of the service and human ears). In contrast, for data services, a higher SINR results in a better throughput, which is desirable. Thus, it is important to design a
power control algorithm for wireless data networks by which the minimum acceptable target-SINRs (which are assumed to be feasible) are guaranteed for all users, and at the same time, the system throughput is increased to the extent that the required resources are available by increasing the actual SINRs received by some users. On the other hand, although the system throughput is maximized in the aggregate throughput maximization problem in (6.12), no minimum acceptable SINR is guaranteed for users (unfairness). Motivated by the above observations, one may define the problem of system throughput maximization subject to a given feasible lower bound for the achieved SINRs of all users in wireless cellular networks:

$$\max \sum_i T_i(p)$$

s.t. $$\gamma_i(p) \geq \hat{\gamma}_i, \forall i \in \mathcal{M}$$

$$0 \leq p \leq \bar{p}$$

variable p.

In this problem, assuming feasibility of the minimum acceptable target-SINRs, the actual SINRs received by some users are increased (to a value higher than their minimum acceptable target-SINRs), so far as the required resources are available and the system remains feasible (meaning that reaching the minimum target-SINRs are guaranteed for the remaining users). This would enhance the system throughput (at the cost of higher power consumption) as compared to the problem in (6.4).

6.3 Closed-Loop and Open-Loop Power Control

6.3.1 Open-Loop Power Control

As has been mentioned before in Chapter 2, in open-loop power control schemes, the transmitting node tunes its transmit power without using any feedback information from the receiver. In other words, the transmitter does not possess any information about the interference and noise levels at the receiver. The transmitter measures the channel-gain (using the pilot signal) and its own local interference and assumes that the receiver shares the same values. Based on these channel parameters, the transmit power is chosen to meet a target SINR. When the channel condition is poor, transmission power is increased, and when the channel condition is good, transmission power is decreased.

Open-loop power control is fast and distributed and has low complexity. However, it suffers from a serious drawback. The channel-gain in the downlink may not be the same as that in the uplink. Therefore, local measurements of the channel-gain are not necessarily close to the actual value at the receiver. This problem is also present in interference measurement. Moreover, in many cases, the uplink and downlink transmissions take place in different frequencies, in which case the channel-gains and interference at the transmitting and receiving ends are not only different but also uncorrelated. Open-loop power control algorithms are a good choice in reciprocal channels, i.e., channels that have approximately the same gain in the uplink and downlink.
6.3.2 Closed-Loop Power Control

As has been mentioned in Chapter 2, in the closed-loop power control method, the channel parameters are measured at the receiving node and fed back to the transmitter so that the transmit power may be adjusted based on the receiver channel-gain and interference. Clearly, this method is much more accurate than open-loop power control. As the name suggests, closed-loop power control is based on a feedback system. The transmitter regularly receives feedback information from the receiver containing information about the channel condition. As in all feedback systems, the closed-loop power control must never diverge; otherwise, the system will face a critical failure. Fast convergence speed is very desirable because the powers must be adjusted before the channel state changes again. Signaling overhead, which is caused due to the feedback information, must be minimal. The algorithm should also be computationally efficient.

Due to the feedback delay, the system may not be able to follow fast variations in the channel state. When the fading is very fast, a closed-loop power control fails to tune proper power levels. Sometimes, a fast inner-loop control is employed to cope with this problem. The inner-loop control is a fast low-performance closed-loop power control, which almost instantly follows the fast variations of the channel. The accuracy is gradually improved by the slow high-performance outer loop. Figure 6.1 depicts a simplified block diagram for closed-loop power control.

6.4 Distributed Power Control Algorithms

In a distributed power control algorithm, each user $i$ updates her transmit power by a power-updating (set-valued) function $f_i(p)$, as $p_i(t + 1) = f_i(p(t))$, where $p(t)$ is the transmit power vector at time $t$. Let $f : \mathcal{P} \to \mathcal{P}$, where $\mathcal{P}$ is the set of feasible transmit power vectors, denote the power-updating function set; that is, $f(p) = [f_1(p), f_2(p), \ldots, f_M(p)]^T$. We say a distributed power-updating function is convergent if $\lim_{t \to \infty} |p(t + 1) - p(t)| = 0$, or equivalently, $\lim_{t \to \infty} |f(p(t)) - p(t)| = 0$. The necessary condition for convergence of a distributed power updating function is the existence of a solution (so-called fixed-point) for the set of equations $f(p) = p$, or equivalently,
\[ p_i = f_i(p) \text{ for all } i \in \mathcal{M}. \]

Let us denote the fixed-point of the power-updating function set \( f(p) \) by \( p^* \), which is obtained by solving \( p^* = f(p^*) \). If a distributed power control algorithm converges to an equilibrium state, it will be a fixed-point of the corresponding power-update function.

### 6.4.1 Criteria for Evaluation and Analysis of Distributed Power Control

The distributed power control algorithm (or in general any distributed resource allocation algorithm) can be evaluated and compared against each other according to the following criteria:

- Existence of fixed-point
- Convergence
- Optimality/sub-optimality
- Distributed implementation

Ideally, we wish to design a distributed power-update function so that (1) it has a (preferably unique) fixed-point, (2) starting from any arbitrary initial transmit power vector, it converges to its fixed point, (3) its fixed point is the optimal solution for a given power control optimization problem (i.e., it solves the optimization problem in a distributed manner), and (4) it uses only locally available information and a minimum amount of non-local information.

In fact, to thoroughly analyze a given distributed power-update function, we should examine the following:

- Whether there exists a (unique or non-unique) fixed-point for it,
- Starting from any initial transmit power vector, whether it converges to its fixed-point,
- Whether the fixed-point of the distributed power-update function solves the power control problem optimally or sub-optimally, and
- How much non-locally available information is required by each transmitter to update its transmit power and how this information is provided.

In what follows, we first introduce the existing theoretical frameworks to analyze the existence and uniqueness of a fixed-point and convergence properties for general classes of distributed power-update functions. Then we present existing power control algorithms with different objectives and evaluate them based on the aforementioned criteria.

### 6.4.2 Existing Theoretical Frameworks for Fixed-Point and Convergence Analysis

To analyze the existence of fixed-point(s) and the convergence of distributed power control algorithms, there are three well-known theoretical frameworks that present sufficient conditions for guaranteeing the existence and uniqueness of a fixed-point and convergence for general classes of distributed power-update functions. In [8], Yates introduced a framework for standard functions, and in [9], Sung and Leung introduced a new framework for standard type II functions, and also extended both frameworks for...
standard and standard type II functions to a framework for general two-sided scalable functions. These frameworks are formally defined below. For ease of reference, let us call the first framework presented by Yates in [8] as “standard type-I” (and the corresponding functions as “standard type-I” functions), although it was called by the author the framework for “standard” functions.

**Definition 38 [8]** A power-update function \( f(p) = [f_1(p), f_2(p), \ldots, f_M(p)]^T \) is standard type-I if for all \( p \in P \), the following properties hold:

- **Monotonicity:** if \( p \geq p' \), then \( f(p) \geq f(p') \), and
- **Scalability:** for all \( a > 1 \), \( f(ap) < af(p) \).

**Definition 39 [9]** A power-update function \( f(p) = [f_1(p), f_2(p), \ldots, f_M(p)]^T \) is standard type II if for all \( p \in P \), the following properties hold:

- **Monotonicity:** if \( p \geq p' \), then \( f(p) \leq f(p') \), and
- **Scalability:** for all \( a > 1 \), \( f(ap) > \frac{1}{a}f(p) \).

**Definition 40 [9]** A power-update function \( f(p) = [f_1(p), f_2(p), \ldots, f_M(p)]^T \) is two-sided scalable if for all \( a > 1 \), \( \frac{1}{a}p \leq p' \leq ap \) implies

\[
\frac{1}{a}f_i(p) \leq f_i(p') \leq af_i(p), \forall i \in M. \quad (6.17)
\]

The framework in [9] includes both standard type-I and type-II functions as well as some other functions that are neither standard type-I nor type-II, as will be discussed in this section. In more detail, standard type-I and type-II functions are disjoint (as can be seen from their conflicting monotonicity properties), and both are subsets of the two-sided scalable functions. This is explained below by introducing the concept of a wide-sense standard function.

**Definition 41 [9]** A given power-update function \( f(p) \) is said to be wide-sense standard if each of its components is either standard type-I or type-II.

**Theorem 47 [9]** A wide-sense standard function \( f(p) \) is two-sided scalable.

The following example shows that the converse is not true. The iterative function

\[
f(x) = \begin{cases} 
  x, & \text{if } 0 < x \leq 1 \\
  \frac{1}{x}, & \text{if } x > 1
\end{cases} \quad (6.18)
\]

is a two-sided scalable function. However, it is neither standard type-I nor type-II, since the monotonicity properties do not hold.

For the given two functions \( f_1(p) \) and \( f_2(p) \), let us define

\[
f_{\text{max}}(p) = \max\{f_1(p), f_2(p)\} \quad (6.19)
\]

and

\[
f_{\text{min}}(p) = \min\{f_1(p), f_2(p)\} \quad (6.20)
\]
where the minimum or maximum operation is applied component-wise. We can easily show that taking component-wise maximum or minimum of two-sided scalable functions conserves the two-sided scalability.

**Lemma 1** [9] For the given two-sided scalable functions \( f_1(p) \) and \( f_2(p) \), their component-wise maximum or minimum, i.e., \( f^{\max}(p) \) or \( f^{\min}(p) \), respectively, are two-sided scalable.

The following two theorems enable us to prove the existence and uniqueness of the fixed-point and convergence of many of the existing power-update functions proposed in the literature.

**Theorem 48** [9] Assume a given power-update function \( f(p) \) that is standard type-I or standard type-II or two-sided scalable. If there exists a fixed point for \( f(p) \), then

I. This fixed point is unique and
II. Starting from any initial transmit power vector, the power control algorithm \( p(t + 1) = f(p(t)) \) converges to its unique fixed-point, in both synchronous and asynchronous power-updating cases.

**Theorem 49** [9] If a given power-update function \( f(p) \) is two-sided scalable and continuous, and there exists \( l, u > 0 \) such that \( l \leq f(p) \leq u \) for all \( p \in \mathcal{P} \) (i.e., \( f(p) \) is upper and lower bounded), then a fixed point for \( f \) exists.

Based on Theorems 48 and 49, we conclude the following important corollary, which indeed summarizes the main result of the theoretical frameworks introduced above for convergence and fixed-point analysis.

**Corollary 2** [9] If a given power-updating function \( f(p) \) is two-sided scalable and continuous, and there exists \( l, u > 0 \) such that \( l \leq f(p) \leq u \) for all \( p \in \mathcal{P} \), then a unique fixed point for \( f(p) \) exists, and the power control algorithm \( p(t + 1) = f(p(t)) \) converges to this unique fixed-point, starting from any initial transmit power vector, in both synchronous and asynchronous power updating cases.

Note that the aforementioned theoretical frameworks provide us with the sufficient condition for existence and uniqueness of fixed-point and convergence. If a given power-update function cannot be accommodated in these frameworks, it does not necessarily mean non-existence or non-uniqueness of fixed-point or non-convergence of that power-update function. There may be some convergent distributed power-update function with a unique fixed-point (or with possible multiple fixed-points) that does not fall into these frameworks. Thus it is possible to expand the existing frameworks into a broader one or develop new frameworks to cover functions which are not two-sided scalable functions.
6.5 Distributed Target-SINR Tracking Power Control (TPC)

In this section, we introduce the most well-known distributed power control algorithm, which we call the distributed target-SINR-tracking power control algorithm (TPC). In [10], Foschini and Milijanic proposed the TPC for addressing the problem of minimizing the aggregate power subject to the constraint of supporting all UEs with their given target-SINRs (formally stated in (6.4)), in a distributed manner. In this problem, it is implicitly assumed that the target-SINRs for all users are feasible, and the objective is to support all users with their target-SINRs, consuming the minimum aggregate transmit power. In what follows, we present the TPC algorithm and then analyze it based on the criteria introduced in previous section.

To determine how good a channel is for a given user $i$, not only the corresponding path-gain but also the amount of interference caused to it should be considered. The channel-goodness for a given user $i$ can be measured by the effective interference for user $i$, which is defined as the ratio of her experienced interference to her path-gain, denoted by $R_i$ as follows:

$$R_i(p) = \frac{I_i(p)}{h_{ii}}.$$  \hspace{1cm} (6.21)

The value of $R_i$ represents the channel status for user $i$, i.e., a higher interference and a lower path-gain result in a higher $R_i$, implying a poor channel as compared to a lower interference and a higher path-gain, which result in a lower $R_i$, implying a good channel.

The idea behind the TPC algorithm is very simple. Given the interference at time $t$, $I_i(p(t))$, the transmit power for user $i$ at time $t + 1$ is set to $p_i(t + 1) = \frac{I_i(p(t))}{h_{ii}} \gamma_i$. The unconstrained power-update function for the TPC is

$$p_i(t + 1) = \gamma_i R_i(p(t))$$ \hspace{1cm} (6.22)

where $R_i(p(t))$ is the effective interference experienced by user $i$ at time $t$. This causes that, at each iteration, the target-SINR value of each user is rigidly tracked by that user. In what follows, we analyze the TPC according to the criteria introduced in Section 6.4.1.

Existence of Fixed-Point and Convergence

Lemma 2 TPC’s power update function given by (6.22) is a standard type-I function.

It was shown in [10] and [8] that if and only if the target-SINR vector $\gamma_i$ is feasible, then the problem in (6.4) has a unique solution. In this case, for both synchronous and asynchronous power updating, the fixed-point of the unconstrained-TPC is unique, and it converges to this unique fixed-point. And the users attain their target-SINRs with the minimum aggregate transmit power. In other words, in a feasible system, the fixed point of unconstrained-TPC solves the optimization problem of minimizing the aggregate transmit power while supporting all UEs with their given target-SINRs.
Theorem 50  In a feasible system, the power-update function of unconstrained-TPC given by (6.22) has a unique fixed-point, to which it converges, for both synchronous and asynchronous power updates.

When the system is infeasible, the problem in (6.4) has no solution, because there is no transmit power vector that can satisfy the SINR requirements for all users. When the unconstrained-TPC, which was originally designed to solve the problem in (6.4) in a distributed manner assuming feasibility of the system, is used in an infeasible system, since the target-SINRs are rigidly tracked, all users increase their transmit powers at each step. Thus, the unconstrained-TPC diverges. Thus, the following observation is made.

Observation 6.5.1  In an infeasible system, the unconstrained-TPC does not converge, because the infeasible target-SINRs are rigidly tracked by all users, and consequently, their transmit powers increase at each step.

In order to deal with the divergence of unconstrained-TPC in infeasible systems, in [11] and [12], the transmit power is assumed to be constrained:

$$p_i(t+1) = \min \{\bar{p}_i, \hat{\gamma}_i R_i(p_i(t))\}. \quad (6.23)$$

Constraining the transmit power of UEs guarantees the convergence of power control even in an infeasible system, as stated in the following theorem.

Theorem 51  For both feasible and infeasible systems, the power-update function of constrained-TPC given by (6.23) has a unique fixed-point to which it converges for both synchronous and asynchronous power updates.

The above theorem can be easily proven by applying Corollary 2, and noting that the power-updating function of constrained TPC given by (6.23) is a type-I standard function, and it is lower and upper bounded, i.e., $0 < \hat{\gamma}_i R_i(0) \leq \min \{\bar{p}_i, \hat{\gamma}_i R_i(p(t))\} \leq \bar{p}_i$, for all $p$ and $i$, irrespective of feasibility or infeasibility of the system.

In summary, when the fixed-point exists for the TPC (which is the case for unconstrained-TPC when the system is feasible, and is the case for constrained-TPC when the system is either feasible or infeasible), the convergence to this fixed-point is also guaranteed. When the system is infeasible, the unconstrained-TPC diverges, whereas the constrained-TPC converges to its fixed-point.

Optimality/Sub-Optimality of TPC

As has been mentioned before, if and only if the target-SINR vector $\hat{\gamma}$ is feasible (with unconstrained or constrained transmit power), then the corresponding TPC (unconstrained and constrained-TPC, respectively) converges to a fixed point (for both synchronous and asynchronous power updates). In such a case, the UEs attain their target-SINRs with the minimum aggregate transmit power. That is, the fixed point solves the optimization problem of minimizing the aggregate power while supporting all of the UEs with their given target-SINRs.
In the case of unconstrained transmit power, when the system is infeasible (i.e., when there exists no positive transmit power that corresponds to the target-SINR vector), the unconstrained-TPC diverges. In the case of constrained transmit power, when the system is infeasible (i.e., when there exists no constrained transmit power vector that corresponds to the target-SINR vector), the constrained-TPC converges. However, at the equilibrium (i.e., at the convergence point), there exist some users who have not achieved their target-SINRs. This is because, in this case, the optimization problem in (6.4) has no feasible solution although the UEs transmit at their maximum power levels.

**Distributed Implementation and Signaling Overhead**

In the power-update function of the TPC given by (6.22), at each time $t$, each UE needs to know the values of $I_i(p(t))$ and $h_{ii}$, or their ratio, i.e., $R_i(p(t))$. The value of path-gain needs to be measured or estimated by using specific techniques (for example, by sending a signal with known transmit power and measuring it at the receiver). The value of $I_i(p(t))$ or $R_i(p(t))$ needs to be broadcast by the BS to its serving UEs. The need for knowing the path-gains and (effective) interference can be avoided as explained below.

Noting that $R_i(p(t)) = \frac{\gamma_i(p(t))}{\gamma_i(p(t))}$, the power-update function in TPC can be rewritten as

$$f_i^{(T)}(p(t)) = \frac{\gamma_i}{\gamma_i(p(t))} p_i(t) \quad (6.24)$$

where $\gamma_i(p(t))$ is the actual SINR of user $i$ at iteration $t$. By using this equivalent power-update function for the TPC, each UE needs only to know its own previous SINR value, which results in minimal feedback information as compared to the update function of the TPC given by (6.22).

Furthermore, (6.24) reveals that the TPC algorithm is indeed the same as the closed-loop power control algorithm, since the ratio of $\gamma_i(p(t))$ in the TPC algorithm can be viewed as the command of increasing or decreasing the power in the closed-loop power control algorithm, corresponding to $\gamma_i(p(t)) < \gamma_i$ and $\gamma_i(p(t)) > \gamma_i$, respectively.

### 6.6 Distributed Opportunistic Power Control (OPC)

In this section, we introduce another well-known distributed power control algorithm called OPC proposed in [9] to address the problem of distributed power control for maximizing system throughput, assuming unconstrained transmit power. This problem is formally defined as follows:

$$\max \sum_i T_i(p)$$

s.t. $p \geq 0$

variable $p$. (6.25)

The OPC algorithm exploits the fact that the system throughput can be enhanced if the UEs with good channels transmit at high power levels and the UEs with poor channels transmit at low power levels. Given the interference $R_i(p(t))$ at time $t$, the transmit power for UE $i$ at time $t + 1$ is set to a value proportional to the inverse of effective interference at time $t$. This will cause each UE with a good channel (low effective interference)
to transmit at a high power level, and each UE with a poor channel (high effective interference) to transmit at a low power level. The power-update function for OPC is given as follows [9]:

$$f_i^{(O)}(p(t)) = \frac{\eta_i}{R_i(p(t))}$$  \hspace{1cm} (6.26)

where $\eta_i$ is a constant for UE $i$. In this algorithm, each UE $i$ tries to keep the product of its transmit power and its effective interference to a constant $\eta_i$, called the target signal-interference product (SIP). The transmit power level in OPC is indeed updated in a manner opposite to that in TPC, i.e., it is increased when the channel is good and is decreased when the channel is poor.

**Existence of Fixed-Point and Convergence**

**Lemma 3** The OPC's power update function given by (6.26) is a standard type-II function.

From the above lemma and by applying Corollary 2, and noting that the power-update function of OPC given by (6.26) is lower and upper bounded, i.e., $0 < \frac{\eta_i}{R_i(p)} \leq \frac{\eta_i}{R_i(0)}$, for all $p$ and $i$, the following theorem can be proven [9].

**Theorem 52** The power-update function of OPC given by (6.26) has a unique fixed-point to which it converges for both synchronous and asynchronous power updates.

**Optimality/Sub-Optimality of OPC**

As stated earlier, although OPC does not result in optimum system throughput, it significantly enhances the system throughput by allowing the UEs with good channels to transmit at high power levels and the UEs with poor channels to transmit at low power levels.

**Distributed Implementation and Signaling Overhead**

In the power update function of the OPC given by (6.26), at each time $t$, each user needs to know the values of $I_i(p(t))$ and $h_{ii}$, or their ratio, i.e., $R_i(p(t))$. Similar to the TPC, the power-update function in OPC can be rewritten as

$$f_i^{(O)}(p(t)) = \eta_i \frac{\gamma_i(p(t))}{p_i(t)}$$  \hspace{1cm} (6.27)

where $\gamma_i(p(t))$ is the actual SINR of user $i$ at iteration $t$. By using this equivalent power-update function, each UE needs only to know its own previous SINR value. This results in minimal feedback information since each UE sets its own target-SINR locally.

### 6.7 Distributed Dynamic Target-SINR Tracking Power Control (DTPC)

In TPC, each UE tracks its own predefined fixed target-SINR. The TPC that was proposed in [10] enables users to achieve their fixed target-SINRs at minimal aggregate transmit power if the target-SINRs are feasible. However, there is a major drawback in the original TPC [10]. It causes users to exactly hit their fixed target-SINRs in feasible systems even if additional resources are still available that could otherwise be used to
achieve higher SINRs (and thus better throughputs). In addition, the fixed-target-SINR assignment is suitable only for voice service for which reaching an SINR value higher than the given target value has no practical effect on service quality (due to characteristics of the service and human ears). In contrast, for data services, a higher SINR results in a better throughput, which is desirable. Thus, it is important to design a power control algorithm for wireless data networks by which the minimum acceptable target-SINRs (which are assumed to be feasible) are guaranteed for all users, and at the same time, the system throughput is increased to the extent that the required resources are available by increasing the actual SINRs received by some users.

From the system’s point of view, OPC allocates high power levels to users with good channels (experiencing high path-gains and low interference levels), and very low power to users with poor channels. In this algorithm, a small difference in path-gains between two users may lead to a large difference in their actual throughputs [9]. Since an opportunistic algorithm always favors those users with better channels, it magnifies unfairness. For users with low-mobility (when their channels vary slowly or are static), this might lead to long-term unfairness.

The characteristics of existing distributed power control schemes are summarized as follows.

**Observation 6.7.1** TPC can provide all users with their fixed target-SINRs when the system is feasible but cannot improve the system throughput further even if additional resources are available. OPC significantly improves the system throughput but cannot guarantee the minimum acceptable SINRs for some users (unfairness).

Motivated by the above observations, in [13], the distributed power control problem is defined with an objective of maximizing system throughput while providing a minimum SINR for each UE, formally defined in (6.16).

It can be shown that the optimization problem (6.16) is non-convex, and thus it cannot be solved by using the conventional methods. When the minimum acceptable target-SINRs are feasible, the actual target-SINRs tracked by some users with better channels should be dynamically set to a value higher than their minimum acceptable target-SINRs to the extent that the required resources are available and the system remains feasible (i.e., the minimum target-SINR constraint in (6.16) is satisfied). This will enhance the system throughput, but at the cost of more power consumption.

It is well-known that for enhancing the system throughput, the UEs with good channels should transmit at higher power levels compared to other UEs [9] [14]. On the other hand, reducing the outage necessitates that UEs with poor channels also transmit at just enough power to reach their minimum acceptable SINRs. Based on these observations, [13] proposes that the UEs with good channels set their transmit power levels in an opportunistic manner and the UEs with poor channels transmit to track their minimum-acceptable-target-SINR. This can be done in a distributed manner if each UE $i$ updates its transmit power as follows:

$$f_i^{(D)}(p(t + 1)) = \begin{cases} \frac{\eta_i}{R_i(p(t))}, & \text{if } R_i(p(t)) < R_i^{th} \\ \hat{\gamma}_i R_i(p(t)), & \text{if } R_i(p(t)) \geq R_i^{th} \end{cases}$$

(6.28)
where $R_i^{th}$ is the effective interference threshold, $\eta_i$ is a constant, and $\hat{\gamma}_i$ is the minimum-acceptable-target-SINR for UE $i$. Note that this algorithm is a generalized selective scheme of either OPC or TPC. At the extreme cases where $R_i^{th} \to 0$ or $R_i^{th} \to \infty$, the proposed power-update function turns into either TPC or OPC, respectively.

The dynamic target-SINR tracking power control algorithm (DTPC) is explained below. Rewrite (6.28) as the DTPC power-update function

$$f_i^{(D)}(p(t + 1)) = \hat{\gamma}_i(p(t))R_i(p(t))$$

(6.29)

in which $\hat{\gamma}_i(p(t))$ is a target-SINR dynamically set as

$$\hat{\gamma}_i(p(t)) = \begin{cases} \eta_i R_i(p(t)), & \text{if } R_i(p(t)) < R_i^{th} \\ \hat{\gamma}_i, & \text{if } R_i(p(t)) \geq R_i^{th}. \end{cases}$$

(6.30)

For DTPC to be continuous, its three parameters are adjusted as follows. For a given $\hat{\gamma}_i$ and $\eta_i$, the value of $R_i^{th}$ is set as

$$R_i^{th} = \sqrt{\eta_i \hat{\gamma}_i}.$$  

(6.31)

Under this setting, we have indeed $f_i^{(D)}(p(t)) = \max \left\{ \frac{\eta_i}{R_i^2(p(t))}, \hat{\gamma}_i R_i(p(t)) \right\}$.

The minimum acceptable target-SINR in (6.29) is set exactly to the same value of the fixed-target-SINR that was set in the TPC. The terms fixed target-SINR and minimum acceptable target-SINR are used here interchangeably. As seen in Figure 6.2, by using DTPC, a given UE sets its target-SINR at the minimum acceptable value when the channel is not good (the value of the effective interference is higher than the threshold). When its channel is good, it opportunistically sets its target-SINR at a value higher than the minimum acceptable value. When the set of minimum target-SINRs are feasible in a non-Pareto efficient manner (i.e., additional resources are available that can be used to further enhance users’ SINRs, meaning that the sum of $\frac{\gamma_j}{\gamma_j+1}$ is far from 1), some UEs with better channels reach higher SINRs as compared to the minimum acceptable value, and the remaining UEs exactly hit their minimum target-SINRs. This means that the system throughput is increased while the UEs’ minimum target-SINRs are guaranteed. However, note that this enhancement in system throughput causes all UEs to consume more power as compared to the fixed (minimum-acceptable) target-SINR tracking scheme.

**Existence of Fixed-Point and Convergence**

The DTPC’s power-update function is a two-sided scalable function. Since the power-update functions corresponding to the fixed-target-SINR tracking and the opportunistic power control algorithms are two-sided-scalable [9], and since the proposed power-updating function can be rewritten as $f_i^{(D)}(p(t)) = \max \left\{ \frac{\eta_i}{R_i^2(p(t))}, \hat{\gamma}_i R_i(p(t)) \right\}$, then from Lemma 1 the following lemma can be proven.

**Lemma 4** The DTPC’s power update function given by (6.29) is a two-sided scalable function.

From the above lemma and by applying Corollary 2, the following theorem is proven in [13].
Figure 6.2 Dynamic target-SINR vs. effective interference in DTPC [13].

**Theorem 53 [13]** There exists a fixed-point for the DTPC’s power-updating function $f(D)(p)$ to which it converges, if the minimum acceptable target-SINR vector $\hat{\gamma}$ is feasible.

**Optimality/Sub-Optimality of DTPC**

In the following theorem, it will be shown that by using DTPC for a given feasible set of minimum target-SINRs, the system throughput is improved as compared to TPC, while guaranteeing that all users are supported at least with their minimum target-SINRs.

**Theorem 54 [13]** Given a feasible target-SINR vector $\hat{\gamma}$, let $p^{*(T)}$ and $p^{*(D)}$ denote the fixed points of TPC and DTPC, respectively. We have $\gamma_i(p^{*(D)}) \geq \hat{\gamma}_i$ for all $i \in M$ and thus $T(p^{*(D)}) \geq T(p^{*(T)})$. More specifically,

I. If and only if $R_i(p^{*(T)}) \geq R_i^{th}$ for all $i$, then $\gamma_i(p^{*(D)}) = \hat{\gamma}_i$ for all $i \in M$, which proves Part I. For Part II, the (non-empty) set of users with $R_i(p^{*(T)}) < R_i^{th}$ is denoted by $N \neq \emptyset$. In this case, there exists a non-empty subset of users in $N$, denoted by $L$, for which we have $\hat{\gamma}_i(p^{*(D)}) > \hat{\gamma}_i$, i.e., $R_i(p^{*(D)}) < R_i^{th}$ for all $i \in L$. If this is not true, then we have $R_i(p^{*(D)}) \geq R_i^{th}$ for all $i \in M$, meaning that $p^{*(D)} = p^{*(T)}$, and consequently, $R_i(p^{*(T)}) \geq R_i^{th}$ for all $i \in M$, which contradicts the existence of at least one user with $R_i(p^{*(T)}) < R_i^{th}$. Thus from this and from (6.22) and (6.29), we conclude that $\gamma_i(p^{*(D)}) > \gamma_i(p^{*(T)})$ for each user $i \in L$, and $\gamma_i(p^{*(D)}) = \gamma_i(p^{*(T)})$ for each user $i \in M \setminus L$, which proves Part II.
Distributed Implementation and Signaling Overhead

Similar to the TPC and OPC algorithms, by using $R_i(p(t)) = \frac{p_i(t)}{\gamma_i(p(t))}$, the DTPC’s power-update function can be rewritten as

$$f_i^{(D)}(p(t)) = \max \left\{ \gamma_i(p(t)) \frac{\gamma_i(p(t))}{\gamma_i(p(t))}, \frac{\gamma_i(p(t))}{p_i(t)} \right\}. \quad (6.32)$$

Thus from signaling overhead point of view, the DTPC needs the same minimal feedback information as either TPC or OPC does, in which each UE needs only to know its own previous SINR value.

6.8 Exercises

Exercise 6.1: Consider a single cell in a CDMA wireless network with $M$ active UEs denoted by $\mathcal{M} = \{1, 2, \ldots, M\}$. Let $p_i$ denote the transmit power of UE $i$, where $0 \leq p_i \leq \bar{p}_i$, and $\bar{p}_i$ denotes the maximum transmit power of UE $i$.

i. Show that if two UEs achieve the same SINR, then their received power levels at the BS are also equal.

ii. Use the result obtained above to find the maximum value of equal SINR, which is obtained by solving the following power control problem referred to as max-equal SINR power control problem:

$$\text{max-equal SINR: } \max_{0 \leq p \leq \bar{p}} \left\{ \gamma_i(p(t)) \right\} \quad \text{s.t. } \forall i \in \mathcal{M}.$$

iii. Can you provide a distributed solution for the max-equal SINR power control problem?

Exercise 6.2: Let $\hat{\gamma}_i$ denote the minimum acceptable target-SINR for each user $i$. Answer the following questions for dynamic target-SINR tracking power control algorithm (DTPC) for unconstrained and constrained cases.

i. Unconstrained DTPC: Consider a case in which the transmit power level for each UE is unconstrained. Suppose that the target-SINR vector $\hat{\gamma} = [\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_M]$ is feasible.

(a) Show that if all UEs employ unconstrained DTPC, at the fixed-point, to which the algorithm converges all UEs are supported with their minimum target-SINRs.

(b) Do all users obtain an SINR greater than their minimum target-SINR in DTPC? If not, which user(s) obtain an SINR greater than the minimum target-SINR in DTPC?

(c) Now suppose that we want to fairly increase the achieved SINRs for all UEs, formally state this problem and try to solve it. Could you modify the unconstrained DTPC so that all users benefit from available resources (i.e., all users obtain an SINR greater than the minimum target-SINR).

ii. Constrained DTPC: Now consider the case for constrained DTPC. Again suppose that the minimum acceptable target-SINRs for UEs are feasible. Are all UEs still
supported with their minimum acceptable target-SINRs when constrained DTPC is used? Why? If all UEs are not supported with their target-SINRs, how this issue can be solved?

**Exercise 6.3:** Consider a single cell in a CDMA wireless network with $M$ active UEs denoted by $\mathcal{M} = \{1, 2, \ldots, M\}$. Let $p_i$ denote the transmit power of UE $i$, where $0 \leq p_i \leq \overline{p}_i$, and $\overline{p}_i$ denotes the maximum transmit power of UE $i$.

i. Show that if two UEs achieve the same SINR, then their received power levels at the BS are also equal.

ii. Use the result obtained above to find the maximum value of equal SINR, which is obtained by solving the following power control problem referred to as max-equal SINR power control problem:

$$\text{max-equal SINR: } \max_{0 \leq \mathbf{p} \leq \overline{\mathbf{p}}} \{\gamma | \gamma_i(\mathbf{p}) = \gamma, \forall i \in \mathcal{M}\}.$$

iii. Can you provide a distributed solution for the max-equal SINR power control problem?

**Exercise 6.4:** Apply the constrained TPC algorithm to a simple CDMA single-cell network within which the UEs are uniformly distributed in a 500 m $\times$ 500 m square cell. The BS is in the middle of cell. The network conditions are chosen such that the system is feasible:

- Number of users $N_u = 5$
- Background noise power, $\nu = 10^{-10}$
- Maximum power of each user, $P_i = 1$ mW
- Target SINR, $\hat{\gamma} = 0.05$
- Path-gain, $h_i = 0.09d^{-3}$.

i. **Feasible system:** Check if the system (target-SINR vector) is feasible. If the system is infeasible, again distribute users in the cell so that the target-SINRs become feasible. Simulate the TPC for the feasible system explained above.

   - Plot the SINR and the transmit power of the users versus the number of iterations (as a measure of time).
   - Change the initial transmit power of users. Does it change the equilibrium transmit power vector (i.e., where the TPC converges to)?
   - Do all users reach their target-SINRs at the equilibrium transmit power vector?

ii. **Infeasible system:** Now change one of the simulation parameters (for instance, the number of users, users’ locations, target-SINRs, or noise power level) to make the system infeasible.

   - Plot SINR and power of the users versus the number of iterations under the infeasible system setting.
   - Do all users reach their target-SINRs at the equilibrium transmit power vector?
   - How can we check the feasibility or infeasibility of system by observing the equilibrium transmit power vector or SINR vector of the TPC?
iii. Performance comparison of the TPC in feasible and infeasible systems
   - Compare the performance of TPC for two cases of a feasible and infeasible system simulated above. Based on your simulation results, explain the positive and negative aspects of the TPC algorithm.
   - Discuss how we can reduce the number of unsupported users (those who have not reached their target-SINRs) in infeasible system.

Exercise 6.5: Apply the OPC algorithm to a simple single-cell wireless network with the following parameters:

- Background noise power, $\sigma^2 = 10^{-10}$
- OPC constant, $\eta = 0.05$
- Path-gain, $h_i = 0.1d^{-3}$.

Simulate the system under the above conditions, for five users and cell radius of 250 m. The users should be uniformly distributed in the cell.

- Plot SINR and power of each user versus the number of iterations.
- Which users transmit at high power levels? Does it depend on the initial transmit power vector?
- Increase or decrease the OPC constant and observe its impact on performance of the OPC algorithm.

Exercise 6.6: To study how TPC, OPC, and DTPC algorithms work when users move, assume a two-cell wireless network with nine users as shown in Figure 6.3. Suppose that users 1 to 6, 8, and 9 are fixed, and user 7 at $t = 0$ starts moving from its starting point $x = 700$ m in cell no. 2 toward the end point $x = 300$ m in cell no. 1 along the illustrated line at a uniform speed of 20 m/s (72 km/h). The movement of user 7 from the starting point to the end point lasts 20 seconds. Each user updates its transmit power every 1 ms (1 kHz) employing either TPC, or OPC, or DTPC. When user 7 enters into cell no. 1, i.e., at $t = 10$, BS 1 is assigned to it. Excluding the moving user 7, note that user 2 in cell no. 1 and user 8 in cell no. 2 are the closest users to the BS in their corresponding cells. Suppose that the transmit power is unconstrained.

- Plot the transmit-power levels and the received SIRs versus time for users 2, 7, and 8.
- Discuss and interpret the results.
References


7 Distributed Joint Power and Admission Control

7.1 Introduction

Although the existence of a fixed-point and convergence is guaranteed for the constrained TPC in both feasible and infeasible systems, it still suffers from a severe drawback, in infeasible systems, as explained below. In this chapter, TPC implies the constrained-TPC given by (6.23) unless stated otherwise.

In feasible systems, TPC can achieve target-SINRs of UEs while minimizing the aggregate transmit power. The following example illustrates a severe drawback of TPC. Consider an infeasible system with four UEs for which the path-gain and target-SINR vectors are given by \([0.15, 0.40, 0.50, 0.60]\) and \([0.50, 0.40, 0.50, 0.40]\), respectively. Suppose the AWGN power is 0.1 Watt, and the maximum transmit power for each user is 1 Watt. The fixed-point power-vector of the constrained TPC is \([1, 1, 1, 0.767]\) Watts and thus their achieved SINRs at the equilibrium are \([0.10, 0.33, 0.45, 0.40]\), meaning that only one user (user 4) attains his/her target-SINR using TPC, and the other users do not reach their target-SINRs, while they are transmitting at their maximum transmit power level. In general, by employing the TPC in an infeasible system, at the corresponding equilibrium transmit power vector \(\mathbf{p}\), for each non-supported user \(i\), we have \(\gamma_i(\mathbf{p}) < \hat{\gamma}_i\), and thus \(p_i < \hat{\gamma}_i R_i(\mathbf{p})\) holds, which means \(\min\{p_i, \hat{\gamma}_i R_i(\mathbf{p})\} = p_i\). From this, the following observation is made.

**Observation 7.1.1** In an infeasible system, TPC causes all non-supported UEs to transmit at their maximum powers while they do not achieve their target-SINRs. This is because all UEs (including the non-supported ones) rigidly track their respective SINRs. This unnecessarily drains the batteries of non-supported UEs and increases interference to other UEs.

From the example given above, one can easily see that the transmit power vector \([0.22, 1, 0.93, 0.67]\) Watts results in the target-SINR vector \([0.024, 0.40, 0.50, 0.40]\). This means UEs 2, 3, and 4 can be supported, and their required transmit power to attain their target-SINRs is also reduced if UE 1 refrains from transmitting at maximum power and instead transmits at 0.22 Watt or less.

**Observation 7.1.2** When the non-supported UEs transmit at their maximum transmit power, it increases the number of non-supported UEs, which can be avoided if some non-supported UEs are removed or reduce their transmit power.
Figure 7.1 A system with two users. The solid lines are $p_1 = f_1^{(T)}(p_2) = \hat{\gamma}_1(p_2 h_2 + \sigma^2)/h_1$ and $p_2 = f_2^{(T)}(p_1) = \hat{\gamma}_2(p_1 h_1 + \sigma^2)/h_2$. The parameters are $\tilde{p}_1 = f_1^{(T)}(0)$, $\tilde{p}_2 = f_2^{(T)}(0)$, and $\tilde{p}_1 = f_1^{(T)}(\tilde{p}_2)$ and $\tilde{p}_2 = f_2^{(T)}(\tilde{p}_1)$.

To further explain the above observations, let us consider a system with two active UEs. Figure 7.1 is a two-dimensional space whose $x$-axis and $y$-axis are the transmit power levels of UE 1 and UE 2, respectively. The constrained power levels of the two UEs are upper bounded by the dashed lines. The solid lines are $p_1 = f_1^{(T)}(p_2) = \hat{\gamma}_1(p_2 h_2 + \sigma^2)/h_1$ and $p_2 = f_2^{(T)}(p_1) = \hat{\gamma}_2(p_1 h_1 + \sigma^2)/h_2$, which represent the transmit power levels required for reaching the target-SINRs of UE 1 and UE 2, respectively, as a function of the other UE’s power. If these two lines intersect, the intersection point $A$, i.e., $p = (p_1, p_2)$, is the fixed-point of the unconstrained-TPC. If this intersection point satisfies the power constraint, it is also the fixed-point of the power-update function of the constrained-TPC; otherwise, the system is infeasible, and the fixed point for the constrained-TPC is $p^* = \left( \min\{p_1, \tilde{p}_1\}, \min\{p_2, \tilde{p}_2\} \right)$. In this case, if $p_1 > \tilde{p}_1$ and $p_2 > \tilde{p}_2$, then the point $B$, i.e., $p^* = (\tilde{p}_1, \tilde{p}_2)$, is the fixed-point, which is the worst case in the sense that both UEs transmit at their maximum power levels without reaching their target-SINRs. In this case, if a UE is removed, the other UE can reach its target-SINR by transmitting at a lower power level.

The two observations above show that, in an infeasible system, the policy of either removing some non-supported UEs or reducing their transmit power has at least two advantages: (1) the remaining ones can be supported, and consequently, the number of supported UEs is increased, and (2) the aggregate transmit power is reduced. Motivated by these observations and the advantages of removing a portion of non-supported UEs in infeasible systems, we now formally state the admission control (or gradual removal) problem.

Given a transmit power vector $p$, we denote the set of supported UEs by $S(p) = \{ j \in M | \gamma_j(p) \geq \hat{\gamma}_j \}$, and the cardinality of this set by $|S(p)|$. The outage probability
is $(|\mathcal{M}| - |S(p)|)/|\mathcal{M}|$. We define the problem of minimum-outage-based removal or equivalently, the problem of maximizing the number of supported UEs by

$$
\min O(p) \\
s.t. \ 0 \leq p \leq \bar{p}
$$

(7.1)

which is applicable to both feasible and infeasible systems. When the system is feasible, all UEs can be supported (i.e., the minimum outage probability is zero), and its solution is given by TPC consuming minimal transmit power. When the system is infeasible, a minimal number of UEs should be removed. In this case, the problem is an NP-complete problem [1].

### 7.2 Distributed Joint Power and Admission Control Algorithms

In the following sections, we describe the existing distributed power control algorithms proposed in the literature that address the gradual removal problem. One way to classify the existing joint power and admission control algorithms is as follows: (1) Those algorithms under which some UEs with poor channels remove themselves permanently (e.g., as in [1], [2], and [3] (the converged version)), and (2) those in which some UEs with poor channels temporarily remove themselves (e.g., as in [3] (the not-converged version)). In the former case, each removed UE either tries another channel (if one exists) or remains disconnected, as opposed to the latter case in which each removed UE (or each non-supported UE that decreases its transmit power) stays on the same channel and resumes its transmission if its effective interference becomes acceptable again. Generally, for two algorithms with the same performance, the one with the capability of resuming transmission for removed users is preferred to the one with permanent removal, because the former adapts to interference and can either remove or resume UEs if the conditions warrant. On the other hand, temporary removals may cause oscillations between removing and resuming states, and the algorithm may not converge in an infeasible system.

#### 7.2.1 TPC with Permanent Removal (TPC-PR)

In [2], an interesting idea is applied in which the UEs update their transmit powers according to the TPC, but when the instantaneous effective interference for a given UE $i$ exceeds the threshold $R_i^{th}$, then UE $i$ removes itself permanently with a given probability $\delta$ and stays with probability $1 - \delta$. The probability of permanent removal is set so that the probability of exactly one removal at each step is maximized. The algorithm can be stated as follows. There are two states (State 1 and State 2) for each UE, and all UEs start with State 1. During iteration $t$, each UE $i$ updates its transmit power based on its state using the following rules:
State 1:

a: For \( R_i(t) \leq R_{th}^i \), let \( p_i(t + 1) = \tilde{\gamma}_i R_i(t) \), increment \( t \), and stay in State 1.

b: For \( R_i(t) > R_{th}^i \), with probability \( \delta \) go to State 2, or with probability \( 1 - \delta \) let \( p_i(t + 1) = \tilde{p}_i \), increment \( t \), and stay in State 1.

State 2:

Permanently stop transmitting, i.e., \( p_i(t) = 0 \) for all \( t \geq t_s \), where \( t_s \) is the first time that UE \( i \) comes to State 2.

7.2.2 TPC with the Capability of Temporary Removal (TR)

To deal with the stated drawback of TPC in infeasible cases, in [3], the power-update function of TPC is revised as follows:

\[
\text{TPC-TR: } p_i(t + 1) = \begin{cases} 
\tilde{\gamma}_i R_i(t), & \text{if } R_i(t) \leq R_{th}^i \\
0, & \text{if } R_i(t) > R_{th}^i 
\end{cases}
\]  

(7.2)

in which a UE with a poor channel refrains from transmitting at its maximum power level and removes itself temporarily. If this algorithm converges, it reduces the outage probability and the total consumed power as compared to TPC.

In contrast to TPC that has a fixed-point (even in infeasible cases), the existence of a fixed-point for the algorithm in (7.2), and consequently, its convergence cannot be guaranteed in an infeasible system, and some UEs may oscillate between switch-off and switch-on (target-SINR-tracking) modes. The reason for such oscillations is that when a given UE experiences an effective interference greater than a threshold, it switches itself off, thereby reducing its interference to others. This in turn causes others to reduce their transmit powers, resulting in a reduced interference to that UE. If the effective interference is lower than the threshold value, the UE resumes its transmission, and the same event is repeated.

The possibility of power oscillations in (7.2) arises due to two reasons. The first and the more important one is the possibility of non-existence of a fixed-point for the power-update function in (7.2), implying that it is impossible for the power-update process to converge. The second reason that may cause oscillation, even if some fixed-points exist (it may have several fixed-points), is the initial transmit power. When there is at least one fixed-point for the power-update function in (7.2), depending on the initial transmit power, some UEs may oscillate between target-SINR-tracking and switch-off modes [3], specifically for the synchronous case, as we discuss below.

Consider again Figure 7.1. When the system is infeasible, only \( p = (\tilde{p}_1, 0) \) or \( p = (0, \tilde{p}_2) \) can be the fixed-point of (7.2), if \( \tilde{\gamma}_1 > \tilde{p}_1 \), or \( \tilde{\gamma}_2 > \tilde{p}_2 \), respectively (\( \tilde{p}_i \) and \( \tilde{\gamma}_i \) for \( i = 1, 2 \) are shown in Figure 7.1). Otherwise (i.e., when \( \tilde{p}_1 \leq \tilde{p}_1 \) and \( \tilde{p}_2 \leq \tilde{p}_2 \)), there is no fixed point for the power-update function (7.2), because \( f_1^{(T)}(\tilde{p}_2) \leq \tilde{p}_1 \) and \( f_2^{(T)}(\tilde{p}_1) \leq \tilde{p}_2 \). It is obvious that when there is no fixed-point, the algorithm never converges. When there is at least one fixed-point, we note that both UEs, depending on their initial transmit power, may oscillate among a sequence of power vectors (e.g.,
\[ p = (0, 0), \quad p = (\tilde{p}_1, \tilde{p}_2), \quad p = (\tilde{\tilde{p}}_1, \tilde{\tilde{p}}_2), \ldots, \quad p = (0, 0) \] or converge to their fixed point. In summary, when there is no fixed-point, the power-update process based on (7.2) does not converge (i.e., there is no equilibrium). But when a fixed-point exists, depending on the initial value of the transmit power vector, the power-update process may or may not converge (i.e., reaching an equilibrium depends on the initial value of the transmit power vector).

In Section 7.2.4, we will describe a distributed power control algorithm with the capabilities of temporary removal and distributed feasibility check \[4\] for which a fixed point is guaranteed to exist, and a heuristic solution is proposed so that the power update process converges to a fixed point when there is at least one fixed point.

### 7.2.3 TPC with Both Temporary and Permanent Removal (TPC-TPR)

The same method used in the TPC-PR is combined with the TPC-TR in \[3\] to deal with the problem of oscillations in the TPC-TR algorithm. This algorithm is exactly the same as the TPC-PR, except for State 1-b, which incorporates both temporary and permanent removals (TPC-TPR). State 1-b is revised as follows:

**State 1:**

b: For \( R_i(t) > R_{th} \), with a probability of \( \delta \) go to State 2, or with a probability of \( 1 - \delta \), let \( p_i(t + 1) = 0 \), increment \( t \), and stay in State 1.

### 7.2.4 TPC with the Capability of Temporary Removal and Feasibility Check (DFC)

In \[4\], for a single-cell wireless network, a distributed power control algorithm with the capability of temporary gradual removal and feasibility check (abbreviated as DFC) is proposed, whose fixed-point is guaranteed to exist. And at its equilibrium, all transmitting UEs attain their target-SINRs while the aggregate transmit power is minimized and no UE is unnecessarily removed. This is called a Pareto-efficient equilibrium. Under a Pareto- and energy-efficient power allocation, the aggregate transmit power to support a given subset of UEs, whose target-SINRs are reachable, is minimized, but no additional UE (if one exists) can be supported at the same time.

To compare the efficiency of the two transmit power control schemes, the concept of Pareto dominance (as defined below) is used.

**Definition 42** A transmit power vector \( p \) Pareto dominates another vector \( p' \) if \( S(p') \subseteq S(p) \). A power vector \( p \) is Pareto-efficient if there is no power vector that Pareto dominates \( p \).

Similar to the gradual removal problem, it can be shown that the problem of Pareto-efficient power control is also an NP-complete problem.

One can easily prove the following lemmas.

**Lemma 5** \[4\] The transmit power vector \( p \) is Pareto-efficient if and only if for any \( i \in S'(p) \), the target-SINRs of UEs in the set \( S(p') \cup \{i\} \) are not feasible.
Lemma 6 [4] \textit{If the transmit power vector} $\textbf{p}$ \textit{solves the minimum-outage problem (7.1), then it is Pareto-efficient. The converse is not necessarily true, i.e., a Pareto-efficient transmit power vector does not necessarily minimize the outage.}

As implied by Lemma 6, the set of power vectors that minimizes the outage is a subset of all Pareto-efficient power vectors. The following example shows that there exists a Pareto-efficient power vector that does not minimize the outage. Consider an infeasible system with three users with the same target-SINR value of 0.5 and the path-gain vector of $[0.005, 0.01, 0.009]$. Suppose the AWGN power is 0.0005 Watt and the maximum transmit power for each user is 0.1 Watt. It is straightforward to verify that the target-SINRs are feasible for each set of UEs $\{1\}$, $\{2\}$, $\{3\}$, and $\{2, 3\}$, and infeasible for any other set of UEs. Thus the minimum outage is $1/3$, which is obtained by a transmit power vector that satisfies target-SINRs of $\{2, 3\}$. Only the transmit power vectors that satisfy the target-SINRs of $\{1\}$ or $\{2, 3\}$ are Pareto-efficient, among which the transmit power vectors that satisfy the target-SINR of $\{1\}$ does not minimize the outage. Furthermore, as implied by this example, the Pareto-efficient transmit power vector(s) as well as the transmit power vector(s) that minimizes outage are not unique in general.

Under Pareto-efficient transmit power control, all transmitting UEs are supported with their target-SINRs (by consuming the minimum required power), and no UE is unnecessarily removed (i.e., no UE, if one exists, can be even theoretically supported while the others are supported). In other words, the least number of UEs that cannot be supported along with the given corresponding supported UEs (due to the infeasibility of the system) are switched off. This policy has two main advantages: (1) All UEs save energy in the following manner. The removed UEs save their energy by switching off, when they cannot reach their target-SINRs, and hence the aggregate consumed power by transmitting UEs is minimized. (2) When some UEs are removed, no interference is introduced to others, which can increase the number of supported UEs. Note that a Pareto-efficient transmit power vector is not generally unique. When the system is feasible, all UEs can be supported, meaning that the minimum outage probability is zero and the transmit power vector is Pareto-efficient. In this case, TPC results in zero-outage (minimum-outage), consuming the minimum amount of aggregate transmit power. When the system is infeasible, at least one Pareto-efficient transmit power vector exists, but in general is not unique. These can be seen by considering again Figure 7.1. When the system is feasible, the intersection of two solid-lines is uniquely Pareto- and energy-efficient. When the system is infeasible, there are two Pareto- and energy-efficient transmit power vectors $\textbf{p} = [\tilde{p}_1, 0]$ and $\textbf{p} = [0, \tilde{p}_2]$.

The problem considered in [4] is how to identify and remove the least number of UEs in a distributed and Pareto-efficient manner (i.e., no UE is unnecessarily removed) in infeasible systems. This would make the target-SINRs of the remaining UEs feasible, and hence, they can use TPC to attain their target-SINRs in an energy-efficient manner. To address this problem, a Pareto-efficient distributed target-SINR-tracking power control algorithm with the capabilities of temporary removal and feasibility check (abbreviated as DFC) was proposed in [4]. It was shown that DFC has at least one fixed-point,
and all of its fixed-points are Pareto-efficient. Furthermore, when the target-SINRs are the same for all UEs, the DFC guarantees the minimum outage probability.

Define

\[ g_i(p(t), p_{th}^i) \triangleq \begin{cases} f_i^T(p(t)), & \text{if } f_i(p(t)) \leq p_{th}^i \\ 0, & \text{if } f_i(p(t)) > p_{th}^i \end{cases} \] (7.3)

where \( p_{th}^i \) is a predefined threshold, and \( f_i(p(t)) = \tilde{R}_i(p(t)) \). This function is a generalized version of (7.2). DFC has the following distributed power-update function:

\[ p_i(t+1) = f_i^E(p(t)) \triangleq \begin{cases} g_i(p(t), p_{th1}^i), & \text{if } p_i(t) \neq 0 \\ g_i(p(t), p_{th2}^i), & \text{if } p_i(t) = 0 \end{cases} \] (7.4)

where \( p_{th1}^i \) and \( p_{th2}^i \) are the two thresholds utilized by user \( i \) and given by

\[ p_{th1}^i \triangleq \bar{p}_i \] (7.5)

and

\[ p_{th2}^i \triangleq \frac{\nu(\tilde{\gamma}_i + 1)}{\bar{p}_i h_{ii} + \nu} \bar{p}_i \] (7.6)

where \( \nu \) denotes the inter-cell-interference plus noise experienced by each user associating with a single BS.

These two thresholds are determined by each UE \( i \) in a distributed manner by using the minimal information pertinent to that UE only. DFC can be interpreted as follows. There are two states for each UE at any given time. If a UE is transmitting (i.e., its transmit power is greater than zero), it operates in TPC as long as its required transmit power for reaching its target-SINR is lower than \( p_{th1}^i \); otherwise, it temporarily removes itself. A removed UE resumes its transmission if the required power to reach its target-SINR becomes lower than \( p_{th2}^i \) (this threshold is lower than \( p_{th1}^i \) by assuming \( \tilde{\gamma}_i \leq \frac{p_{th2}^i}{\nu} \)).

**Existence of Fixed-Point in DFC and Its Convergence**

For analysis of the existence of fixed-point and convergence of DFC, a single-cell system is considered in [4] in which the inter-cell interference (i.e., \( \nu \)) is assumed to be fixed. The following analyses are accomplished based on this assumption. It was shown in [4] that using two different thresholds (7.5) and (7.6) by each UE to decide between target-SINR-tracking or switching-off (in contrast to [3], where a single threshold is employed), the existence of at least one fixed-point is guaranteed.

**Theorem 55 [4]** There exists at least one fixed-point for the power-update function of DFC.

The fixed-point of the DFC’s power-update function is not generally unique. It is obvious that when there is no fixed-point, the algorithm never converges. If a distributed power update function has a fixed point, then that algorithm can potentially (not necessarily) converge. Based on the framework developed in [5] and [6], when there exists a fixed-point for a continuous power-update function, under certain conditions, the fixed-point is unique and the corresponding algorithm converges to the fixed-point for any
initial transmit power vector. The frameworks in [5] and [6] are suitable for the algorithm without removal. However, since in DFC some UEs switch off, the corresponding power-update function is not continuous, and thus these frameworks cannot be applied to study its convergence.

In DFC, the existence of at least one fixed-point is guaranteed. However, depending on the initial value of the transmit power vector, the power-update process may or may not converge, i.e., reaching an equilibrium depends on the initial value of the transmit power vector. This problem, i.e., the existence of a fixed point and no convergence due to the initial value of a transmit power vector, can be resolved (as opposed to the case in which no fixed point exists), as explained below.

For some initial transmit power vectors, some UEs oscillate between two modes of switching off and target-SINR-tracking (due to discontinuity of the power-update function). This causes the algorithm to oscillate among a sequence of power vectors (including the no fixed-point), especially in synchronous cases. This possible oscillation can be prevented if the UEs avoid such sequences. To resolve this, a heuristic solution was proposed in [4]. Each UE counts the number of times it switches between the switch-off and transmit states. When it exceeds a predefined threshold, it randomly and independently sets its transmit power level for the next iteration, resets the counter, and sets its transmit power according to DFC for the forthcoming iteration. This is equivalent to a new initial power vector that guarantees that the equilibrium (a fixed-point) will be eventually reached.

**Pareto-Efficiency of DFC**

It is shown in [4] that all fixed-points in DFC are Pareto-efficient.

**Theorem 56 [4]** Any fixed-point \( \mathbf{p}^* \) for the power-update function of DFC is Pareto-efficient.

As has been stated earlier, in general, the minimum-outage problem is NP-complete. In what follows, for the special case in which the target-SINRs of all UEs are the same, the solution to the minimum-outage problem for a single cell is first obtained, and then it is shown that DFC minimizes the outage.

**Theorem 57 [4]** When the target-SINRs of all UEs are the same (\( \hat{\gamma} \)), any fixed-point of the power-update function of DFC minimizes the outage probability, i.e., for any fixed-point \( \mathbf{p}^* \), \( O(\mathbf{p}^*) = O^* \), where \( O(\mathbf{p}^*) = \frac{|S(\mathbf{p}^*)|}{|\mathcal{M}|} \) and \( O^* \) is the minimum value of the outage given by (6.9).

**Proof** Similar to Theorem 44, suppose \( \bar{\varphi}_1 < \bar{\varphi}_2 < \cdots < \bar{\varphi}_M \), and define \( \Gamma_i \) as in (6.8). When \( 0 \leq \hat{\gamma} \leq \Gamma_1 \) the system is feasible and it can be easily seen that the fixed-points of TPC and DFC are the same, implying that they minimize the outage-ratio to zero. When \( \hat{\gamma} > \Gamma_M \), the fixed-point for the power-update function of DFC is no transmission by each UE (and thus the outage-ratio is 1), because in this case, we have \( \hat{\gamma} > \frac{\bar{\varphi}_i}{v_i} \) for all \( i \in \mathcal{M} \), and consequently, \( f_i(\mathbf{T})(\mathbf{0}) = \hat{\gamma} v_i / h_i > p_i^{th2} \) for all \( i \in \mathcal{M} \). Now we prove the theorem for the case of \( \Gamma_1 < \hat{\gamma} \leq \Gamma_M \), i.e., we prove that for any given fixed-point for the
power-update function \( p^* \) of DFC, we have \( |S'(p^*)| = l \), where \( l \in \{1, 2, \ldots, M - 1\} \) for which \( \Gamma_l < \widehat{\gamma} \leq \Gamma_{l+1} \). It is obvious that \( |S'(p^*)| > l \). We now prove that \( |S'(p^*)| > l \) contradicts the following fixed-point constraint:

\[
f_i^{(T)}(p^*) > p_i^{\text{th}2}, \quad \forall i \in S'(p^*)
\] (7.7)

and thus prove \( |S'(p^*)| = l \). As all UEs in \( S(p^*) \) reach the same target-SINR, their received power are the same \( \widehat{\gamma} = \frac{\psi_i}{\eta_i + \nu} \). Thus we have \( f_i^{(T)}(p^*) = \frac{\psi_i}{\eta_i + \nu} \) for all \( i \in S'(p^*) \), and consequently, the fixed-point constraint (7.7) can be rewritten as

\[
\widehat{\gamma} > \frac{\psi_i}{|S(p^*)|\psi_i + \nu}, \quad \forall i \in S'(p^*).
\] (7.8)

If \( |S'(p^*)| > l \), then we have \( |S(p^*)| < M - l \), and thus from (7.8) we have

\[
\widehat{\gamma} > \frac{\psi_i}{(M - l)\psi_i + \nu}, \quad \forall i \in S'(p^*).
\] (7.9)

For \( |S'(p^*)| > l \), we also conclude that the maximum-index among UEs in \( S'(p^*) \) is greater than \( l \), or equivalently, there exists a UE \( i \in S'(p^*) \) for which \( i \geq l + 1 \). From this and from (7.9), we have

\[
\widehat{\gamma} > \frac{\psi_{l+1}}{(M - l)\psi_{l+1} + \nu} > \frac{\psi_{l+1}}{(M - l + 1)\psi_{l+1} + \nu} = \Gamma_{l+1}
\] (7.10)

which contradicts \( \Gamma_l < \widehat{\gamma} \leq \Gamma_{l+1} \). This completes the proof.

### 7.2.5 TPC with Soft Removal (TPC-SR)

Based on the observations made on TPC and OPC, a distributed power control algorithm is proposed in [7] in which the transmit powers of UEs are adjusted according to either TPC or OPC, depending on the channel conditions to achieve the objectives of gradual removal. We call this strategy TPC-SR, as defined and formulated below.

In TPC-SR, when a given UE \( i \) operates in the target-SINR-tracking mode, an increase in \( R_i \) would increase its transmit power, until the UE’s threshold for \( R_i \) is reached, upon which further increase in \( R_i \) would reduce the UE’s transmit power (opportunistic mode) (see Figure 7.2).

The TPC-SR’s power-update function is as follows [7]:

\[
p_i(t+1) = f_i^{(C)}(p(t)) \triangleq \begin{cases} 
    \widehat{\gamma}_i R_i(p(t)), & \text{if } R_i(p(t)) \leq R_i^{\text{th}} \\
    \frac{\eta_i}{R_i(p(t))}, & \text{if } R_i(p(t)) \geq R_i^{\text{th}} 
\end{cases}
\] (7.11)

where \( R_i^{\text{th}} \) is the effective interference threshold, \( \eta_i \) is a constant, and \( \widehat{\gamma}_i \) is the target-SINR for UE \( i \) (see Figure 7.2). These three parameters of TPC-SR are adjusted so that \( f_i^{(C)}(p(t)) \) is continuous:

\[
\eta_i = \widehat{\gamma}_i (R_i^{\text{th}})^2.
\] (7.12)

Obviously, the performance of TPC-SR highly depends on the threshold values of the effective interference set by users.
There are three parameters in TPC-SR given by (7.11): $\hat{\gamma}_i$, $R_i^{th}$, and $\eta_i$. The values of these parameters depend on one another via (7.12). We consider $\hat{\gamma}_i$ and $R_i^{th}$ as the two tunable parameters and obtain $\eta_i$ from (7.12). The value of $\hat{\gamma}_i$ for each UE $i$ is simply decided by its target-SINR. Thus we need only to propose a distributed tuning of the effective interference threshold $R_i^{th}$.

Assume that each UE chooses its target-SINR $\hat{\gamma}_i$, and adjusts its $R_i^{th}$ by

$$ R_i^{th} = \frac{P_i}{\hat{\gamma}_i} $$

(7.13)

where $\frac{P_i}{\hat{\gamma}_i}$ is the maximum value of $R_i$ for which the target-SINR for UE $i$ is achievable (by transmitting at its maximum power). Note that if $R_i^{th}$ is set to a value higher than the above, it may result in maximum transmit power without reaching target-SINRs for some UEs when the system is infeasible (similar to TPC), which is not desirable. If it is set to a value lower than the above, the target-SINRs may not be achieved even when the system is feasible (because the UE may unnecessarily go to the OPC mode). By substituting (7.12) into (7.13), we obtain $\eta_i = \frac{P_i^{2}}{\hat{\gamma}_i}$.

**Fixed-Point and Convergence Analysis of TPC-SR**

In what follows, for a given $R_i^{th}$ for each user $i$, it was shown in [7] that TPC-SR converges to a unique fixed point in both feasible and infeasible systems.

**Lemma 7 [7]** TPC-SR’s power update function $f^{(C)}(p)$ in (7.11) is two-sided scalable.

**Theorem 58 [7]** (a) TPC-SR’s power update function $f^{(C)}(p)$ has a fixed point, i.e., there exists a transmit power vector $p^*$ such that $p^* = f^{(C)}(p^*)$. In addition, the fixed point is unique.

(b) For any given initial power vector, TPC-SR converges to the fixed point of $f^{(C)}(p)$ in both synchronous and asynchronous cases.

**Proof** From Corollary 2, we conclude that if $f(p)$ is two-sided scalable and continuous, and there is a $u > 0$ such that $f(p) \leq u$ for all $p$, then the fixed point $p^* = f(p^*)$.
exists and is unique. Also, for any initial power vector, the power control algorithm $p(t + 1) = f(p(t))$ converges to $p^*$. Thus, since $f^C(p)$ is continuous (with constraint (7.12)) and is upper bounded ($f_i^C(p) \leq \hat{\gamma}_i R_i^t$ for all $i$), and since $f^C(p)$ is two-sided scalable, this theorem is proved.

Optimality/Sub-Optimality of TPC-SR

So far we have shown that, given $R_i^t$ for each UE $i$, TPC-SR converges to a unique fixed point. We now show how the UEs can adjust the parameters of TPC-SR in a distributed manner to reduce both the outage probability and power consumption as compared to TPC. In the following two theorems we will show that by using TPC-SR (as compared to TPC), each UE consumes less energy, resulting in a reduced total power consumption, and a lower outage-probability.

**Theorem 59** [7] Given $\hat{\gamma}_i$ for all $i$, let $p^{*(T)}$ and $p^{*(C)}$ denote the fixed points of TPC and TPC-SR tuned by (7.13), respectively.

(a) If and only if the system (i.e., the target-SINR vector $\hat{\gamma}$) is feasible, then $p^{*(C)} = p^{*(T)}$.

(b) If and only if the system is infeasible, then $p^{*(C)} < p^{*(T)}$, and consequently, $\sum_i p_i^{*(C)} < \sum_i p_i^{*(T)}$.

**Proof** The first part can be easily proved, so we prove only the second part. We know that if $p^{(C)}(t) \leq p^{(T)}(t)$ for a given $t$, then from (6.21) we have $R_i(p^{(C)}(t)) \leq R_i(p^{(T)}(t))$ for all $i$ and thus $p^{(C)}(t + 1) \leq p^{(T)}(t + 1)$ obtained by comparing TPC-SR’s power-update function in (7.11) tuned by (7.13) (which can be rewritten as $p^{(C)}_i(t + 1) = \min(\hat{\gamma}_i R_i(p^{(C)}(t)), \frac{\eta R_i}{p_i^{(T)})})$ with the TPC’s power-update function in (6.23). Thus, if $p^{(C)}(t_0) \leq p^{(T)}(t_0)$ for a given time $t_0$, then $p^{(C)}(t) \leq p^{(T)}(t)$ for all $t \geq t_0$. Since for any given initial transmit power $p(0)$, TPC and TPC-SR converge to their own unique fixed point (Theorems 51 and 58), we assume that the initial transmit power $p(0)$ is the same for both. Therefore, we have $p^{(C)}(t) \leq p^{(T)}(t)$ for all $t \geq 0$, and consequently, $p^{*(C)} \leq p^{*(T)}$. In what follows, we prove that for an infeasible system, the strict inequality holds, i.e., $p^{*(C)} < p^{*(T)}$. As a direct result of the first part of this theorem, and from $p^{*(C)} \leq p^{*(T)}$, there is at least one $j$ for which $p_j^{*(C)} < p_j^{*(T)}$ in an infeasible case, because otherwise, it contradicts the first part. Now we prove that if there is at least one $j$ for which $p_j^{*(C)} < p_j^{*(T)}$, then $p_i^{*(C)} < p_i^{*(T)}$ for all $i$. This is true because from $p_j^{*(C)} < p_j^{*(T)}$ and $p^{*(C)} \leq p^{*(T)}$ we conclude that for all $i \neq j$, $R_i(p^{*(C)}) < R_i(p^{*(T)})$, which implies $p_i^{*(C)} < p_i^{*(T)}$. Thus, if the system is infeasible, then $p^{*(C)} < p^{*(T)}$. Also, from this and from the first part, one can conclude that if $p^{*(C)} < p^{*(T)}$, then the system is infeasible. The two latter statements complete the proof of the second part of this theorem.

Theorem 59 states that in a feasible system, the performance of TPC-SR tuned by (7.13) is the same as that of TPC. This implies that, in a feasible system, under TPC-SR, the outage probability is zero (i.e., target-SINRs are reached and all users operate in the target-SINR-tracking mode), and the minimum transmit power vector required to
satisfy the users’ target-SINRs is obtained, i.e., both the optimization problems in (6.4) and (7.1) are solved by TPC-SR and TPC in a similar manner.

In TPC-SR tuned by (7.13), if the system is infeasible, then at the fixed point, there exists at least one UE operating in the opportunistic mode and vice versa. This is because, at the fixed point of TPC-SR in an infeasible system, if all UEs operate in TPC, Theorem 3 is contradicted. In this case, the UEs operating in the opportunistic mode are not supported (do not reach their target SINRs). Such UEs transmit at less than their maximum powers, as opposed to TPC where all non-supported UEs transmit at their maximum power. Thus the total powers for all UEs are strictly less than those in case of TPC.

**Theorem 60**

Given \( \hat{\gamma}_i \) for all \( i \), \( S(p^{*(T)}) \subseteq S(p^{*(C)}) \), which implies \( |S(p^{*(T)})| \leq |S(p^{*(C)})| \), where \( p^{*(T)} \) and \( p^{*(C)} \) are the fixed points of TPC and TPC-SR tuned by (7.13), respectively. In other words, if a UE reaches its target-SINR using TPC, then that UE also reaches its target-SINR using TPC-SR tuned by (7.13). The inverse is not necessarily true, i.e., there may exist a UE that does not reach its target-SINR by using TPC, although its target-SINR is reachable by using TPC-SR.

**Proof** From Theorem 59, we have \( p^{*(C)} \leq p^{*(T)} \) and thus \( R_i(p^{*(C)}) \leq R_i(p^{*(T)}) \) for all \( i \). Suppose a given UE \( i \) reaches its target-SINR using TPC, i.e., \( \gamma_i(p^{*(T)}) = \hat{\gamma}_i \), which implies \( R_i(p^{*(T)}) \leq \frac{P}{\gamma_i} \). In this case, UE \( i \) reaches its target-SINR when TPC-SR is used, i.e., \( \gamma_i(p^{*(C)}) = \hat{\gamma}_i \). This is because, if \( \gamma_i(p^{*(C)}) < \hat{\gamma}_i \), then we must have \( R_i(p^{*(C)}) > \frac{P}{\gamma_i} \), and thus \( R_i(p^{*(T)}) < R_i(p^{*(C)}) \), which contradicts \( R_i(p^{*(T)}) \leq R_i(p^{*(C)}) \).

The following example shows that the inverse is not true. Consider again the example of an infeasible system with four users discussed before in this chapter. At the fixed point of TPC, only one UE is supported. But in TPC-SR, one can easily verify that at the fixed point, the UEs’ transmit power vector is \([0.22, 1, 0.93, 0.67] \) Watts and their achieved SINR vector is \([0.024, 0.04, 0.05, 0.04] \), respectively. This means that UE 1 (the non-supported UE with the worst channel) senses that the system is infeasible and reduces its transmit power, instead of transmitting at its maximum power, and thus UE 2 and UE 3 can attain their target-SINRs in addition to UE 4. In addition the transmit power consumed by each UE is reduced as well.

Using TPC-SR (instead of TPC) in an infeasible system causes some UEs to operate in the opportunistic mode, meaning that they start reducing their transmit powers as the effective interference increases. This makes more resources available to other UEs to reach their target-SINRs. Hence, the number of UEs that reach their target-SINRs by using TPC-SR is always equal to or higher than that due to using TPC. Equivalently, the outage probability in TPC-SR tuned by (7.13) is less than that of TPC. Thus, although in general TPC-SR does not guarantee the minimum-outage probability defined in (7.1), it has a unique equilibrium to which the algorithm converges. This equilibrium is closer to the optimum solution of (7.1) (in the sense that the outage probability as well as the power consumption are less) as compared to TPC’s equilibrium.

Now we discuss which UEs in an infeasible system may operate in the opportunistic mode in favor of others by adjusting \( R_i^h \) according to (7.13). Consider the case that
\( \pi \) and \( \hat{\gamma}_i \) are the same for all UEs and the system is infeasible. In this case, \( R_i^{th} \) is the same for all UEs. Thus, those UEs that operate in the opportunistic mode have high values of \( R_i \) (have low path-gains and/or high interference levels). This means that when the system is infeasible, a UE with a poor channel operates in the opportunistic mode (it starts reducing its transmit power as its effective interference increases), thus favoring UEs with a good channel, which is desirable in this case. Similarly, for the case that target-SINRs are not the same for different UEs, in an infeasible system, those UEs with low path-gains and/or with high values for target-SINRs may operate in the opportunistic mode.

### Distributed Implementation and Signaling Overhead

In the power-update function of TPC-SR given by (7.11), at each time \( t \), each UE sets its \( \hat{\gamma}_i \) and \( \eta_i \), and consequently, it needs to know the value of \( R_i(p(t)) \), which is given by \( \pi_i(p(t)) \). \( \gamma_i(p(t)) \). Thus, TPC-SR’s power-update function can be rewritten as

\[
 f_i^{(C)}(p(t)) = \min \left\{ \eta_i \frac{\gamma_i(p(t))}{p_i(t)}, \frac{\hat{\gamma}_i}{\gamma_i(p(t))} \pi_i(p(t)) \right\}. \tag{7.14}
\]

Therefore, TPC-SR needs the same minimal feedback information as either TPC or OPC does, in which each UE needs only to know its own previous SINR value. Note that each UE sets its \( \hat{\gamma}_i \) and \( \eta_i \) values locally.

### 7.3 Exercises

**Exercise 7.1**: To solve the gradual removal problem in an infeasible system where the target-SINRs of UEs are not reachable, Mr. Power Control has proposed the following distributed power-update function:

\[
p_i(t + 1) = \begin{cases} 
\hat{\gamma}_i R_i(p(t)), & \text{if } R_i(p(t)) \leq R_i^{th} \\
\eta_i \left( \frac{\gamma_i(p(t))}{p_i(t)} \right)^2, & \text{if } R_i(p(t)) \geq R_i^{th}
\end{cases} \tag{7.15}
\]

where \( R_i^{th} \) is the effective interference threshold, \( \eta_i \) is a constant, and \( \hat{\gamma}_i \) is the target-SINR for user \( i \). These three parameters are adjusted as

\[
 \eta_i = \hat{\gamma}_i (R_i^{th})^3. \tag{7.16}
\]

Analyze this power-update function based on the criteria used for analyzing distributed power control algorithms.

**Exercise 7.2**: Apply TPC and TPC-SR algorithms to a simple four-cell wireless network in which users are distributed in a 1000 m \( \times \) 1000 m area covered by four base stations and have the same target-SINR of \( \hat{\gamma}_i = 0.2 \). The users should be uniformly distributed in the cell, where \( h = 0.09 d^{-3} \). Each cell covers 500 m \( \times \) 500 m, and each user is assigned to its nearest base station. Consider different total number of users, ranging from four users (1 user/cell) to 28 users (seven users/cell) with step size of four users. To do so, for
each total number of users, locate uniformly and randomly the users in four cells, apply
the TPC and TPC-SR algorithm separately, and compute their outage probability and
aggregate transmit power at the equilibrium where the algorithms converge. Assume
a constraint on maximum power ($\bar{p} = 1$ mW) and background noise power of $\sigma^2 = 10^{-12}$.

- Plot the outage probability and the aggregate transmit power versus the total number
  of users, respectively, for TPC and TPC-SR.
- Compare TPC and TPC-SR algorithms in feasible and infeasible systems.

References

power control and noise,” *ACM/Baltzar Wireless Networks Journal*, vol. 2, no. 1, 1996,
pp. 27–43.
tation, Royal Institute of Technology (KTH), Sweden, 1996.
with capability of temporary removal,” *IEEE Transactions on Vehicular Technology*, vol. 50,
trol algorithm with feasibility check in wireless networks,” *IEEE Transactions on Information
on Selected Areas in Communications*, vol. 13, no. 7, 1995, pp. 1341–1347.
2635.
8 Joint Power and Admission Control in Cognitive Radio Networks

8.1 Introduction

Due to increasing demand for wireless access services, efficient utilization of the limited frequency spectrum has become crucial. Exclusive licensing of spectrum bands to specific users or services is very inefficient from the viewpoint of spectrum utilization, and it lacks the agility needed to support new applications. Cognitive radio networks (CRNs) have thus emerged as an adaptive cohabitation paradigm for wireless communication. The primary radio networks (PRNs) can dynamically share the spectrum with the secondary users (SUs) so that the SUs achieve their minimum acceptable quality-of-service (QoS), and at the same time, all the primary users (PUs) are protected in the sense that the SUs do not violate the QoS requirements of the PUs.

The key concept in cognitive radio networks is opportunistic or dynamic spectrum access, which allows SUs to opportunistically access the band licensed to the PUs. There are two approaches for opportunistic spectrum access: spectrum overlay and spectrum underlay. In the overlay spectrum access strategy, the channels that are unused by the PUs are detected by the CRN through spectrum-sensing mechanisms and are assigned to the SUs. With overlay spectrum access, a channel-sharing method such as orthogonal frequency division multiple access (OFDMA) or time division multiple access (TDMA) is employed where spectrum holes (e.g., unused frequency or time slots) are detected and accessed in an opportunistic manner by SUs. In the underlay scenario, the available frequency spectrum is shared by all of the PUs and SUs, and since the admission of any of the SUs causes interference to all of the PUs’ receiving points, the interference caused by the SUs must be controlled through power control strategies in a way that all PUs are protected (i.e., all PUs achieve their target signal-to-interference-plus-noise ratio [SINR]). With underlay spectrum users employ channel sharing methods such as code-division multiple access (CDMA) or OFDMA in a way that the interference imposed by the SUs remains below a specified threshold and the QoS requirements of all of the PUs are supported. Therefore, with underlay spectrum access, which we focus on in this chapter, the overall spectrum can be utilized more effectively at the expense of higher complexity in controlling the QoS of SUs and the aggregate interference caused to the primary receivers.

For underlay cognitive radio networks, the same problems discussed in Chapters 6 and 7 for conventional cellular networks can be similarly stated and addressed as long as the PUs’ protection is taken into account. In other words, from the optimization point
of view, the same objective function, constraints, and variables considered in traditional wireless networks can be also considered in CRN, but with additional constraint on PUs’ protection. Accordingly, the existing solutions may be also extended to be employed in CRNs so that the protection of PUs is guaranteed.

It is worth noting that emerging wireless networks such as the multi-tier cellular networks and/or device-to-device communication networks face the same problem of prioritized uplink power control and interference management where all users in different tiers share the same licensed spectrum but with different priorities of access. Thus, the joint power and admission control (JAPC) algorithms discussed in this chapter can also be employed in such networks for cross-tier interference management. Also, note that these methods can be used for both OFDMA and CDMA-based CRNs. While in the former case uplink power control is performed for transmission over different sub-channels shared among PUs and SUs over space and time, in the latter case, uplink power control is performed for transmission over the entire spectrum (i.e., a single channel).

In this chapter, first, the general system model and notations are presented in Section 8.2. Different optimization problems corresponding to the JAPC problems in underlay CR are formally stated in Section 8.3. In Section 8.4, different possible ways for defining the protection for PUs are reviewed and used to characterize the feasible interference region in underlay CR networks. Existing centralized JAPC algorithms are categorized, and existing sequential searching algorithms are introduced in Section 8.5. A distributed JAPC algorithm for underlay cognitive radio networks is presented in Section 8.6.

8.2 System Model and Background

In this section, the system model and the notations utilized throughout this chapter are briefly introduced. Also, the relation between the uplink transmit power vector and its corresponding SINR vector and the mechanisms for checking the feasibility of a given SINR discussed in Chapters 5 and 6 are reviewed.

Consider the uplink model of a cellular CDMA cognitive radio network with a collection of PUs and SUs located randomly in the coverage areas.1 The network consists of a set of $M = M^p + M^s$ users denoted by $\mathcal{M} = M^p \cup M^s$ including $M^p$ PUs denoted by set $M^p = \{1, \ldots , M^p\}$ and $M^s$ SUs denoted by set $M^s = \{M^p + 1, \ldots , M^p + M^s\}$. Also, assume that there exists a set of $B = B^p + B^s$ base stations (BSs) denoted by $B = \{1, 2, \cdots , B\}$ including a set of $B^p$ primary BSs (PBSs) denoted by $B^p = \{1, 2, \ldots , B^p\}$ serving the PUs and a set of $B^s$ cognitive radio (secondary) base-stations (SBSs) denoted by $B^s = \{B^p + 1, B^p + 2, \ldots , B^p + B^s\}$ serving the SUs.

1 Although the cellular model is considered, the system model is sufficiently general to be easily applied to an ad hoc network model.
Let \( b_i \in B \) be the serving BS of user \( i \) and let \( \mathcal{M}_m^p \) and \( \mathcal{M}_n^s \) be the sub-set of PUs and SUs associated to base-stations \( k \in B^p \) and \( n \in B^s \), respectively:

\[
\mathcal{M}_m^p = \{ i \in \mathcal{M}^p | b_i = m \}, \\
\mathcal{M}_n^s = \{ i \in \mathcal{M}^s | b_i = n \}.
\]

Let \( p_i \) denote the transmit power of user \( i \) and assume that \( h_{ki}, h_{bi}, \) and \( h_{bj} \) denote the uplink path-gain from user \( i \) toward BS \( k \in B \) and the corresponding BSs of user \( i \) and \( j \), respectively. The noise power at the corresponding BS of user \( i \) is denoted by \( \sigma^2_{b_i} \), which is assumed to be additive white Gaussian. The transmit power \( p_i \) is always limited to a maximum threshold denoted by \( \bar{p}_i \) (i.e., \( p_i \in [0, \bar{p}_i] \)). Considering the receiver to be a conventional matched filter, for any given uplink transmit power vector \( \mathbf{p} = [p_1, p_2, \ldots, p_M]^T \), the uplink SINR of user \( i \) at its BS, which is denoted by \( \gamma_i \), is

\[
\gamma_i(\mathbf{p}) = \frac{h_{bi}p_i}{\sum_{j \in \mathcal{M}} h_{bj}p_j + \sigma^2_{b_i}},
\]

where

\[
\gamma_i(\mathbf{p}) = \begin{cases} 
\frac{h_{bi}p_i}{\sum_{j \in \mathcal{M}^p} h_{bj}p_j + \sum_{j \in \mathcal{M}^s} h_{bj}p_j + \sigma^2_{b_i}}, & \text{if } i \in \mathcal{M}^p \\
\frac{h_{bi}p_i}{\sum_{j \in \mathcal{M}^p} h_{bj}p_j + \sum_{j \in \mathcal{M}^s} h_{bj}p_j + \sigma^2_{b_i}}, & \text{if } i \in \mathcal{M}^s.
\end{cases}
\] (8.1)

To write a linear equation for the relation between the uplink power vector and its corresponding SINR vector, we rewrite (8.1) as

\[
p_i = \sum_{j \neq i} \frac{h_{bj}p_j}{h_{bi}} \gamma_j p_j + \gamma_i \frac{\sigma^2_{b_i}}{h_{bi}}.
\] (8.2)

Rewriting the above linear relation for all \( i \in \mathcal{M} \) in matrix format, we have

\[
\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix} = \begin{bmatrix} 0 & \gamma_1 & \frac{h_{b1}}{h_{b2}} & \ldots & \gamma_1 & \frac{h_{b1}}{h_{bM}} \\ \gamma_2 & 0 & \ldots & \gamma_2 & \frac{h_{b2}}{h_{bM}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_M & \frac{h_{bM}}{h_{b1}} & \ldots & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{bmatrix} + \begin{bmatrix} \gamma_1 & \frac{\sigma^2_{b1}}{h_{b1}} \\ \gamma_2 & \frac{\sigma^2_{b2}}{h_{b2}} \\ \vdots & \vdots \\ \gamma_M & \frac{\sigma^2_{bM}}{h_{bM}} \end{bmatrix}
\] (8.3)
The above equation in matrix format can also be written as

\[
\begin{bmatrix}
    p_1 \\
    p_2 \\
    \vdots \\
    p_m \\
\end{bmatrix}
= \begin{bmatrix}
    \gamma_1 & 0 & \ldots & 0 \\
    0 & \gamma_2 & \ldots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \ldots & \gamma_M \\
\end{bmatrix}
D(\gamma)
\begin{bmatrix}
    h_{b1,2} & \ldots & h_{b1,M} \\
    h_{b2,1} & \ldots & h_{b2,M} \\
    \vdots & \ddots & \vdots \\
    h_{bM,1} & \ldots & h_{bM,M} \\
\end{bmatrix}
\begin{bmatrix}
    p_1 \\
    p_2 \\
    \vdots \\
    p_m \\
\end{bmatrix}
\]

Using matrix notations, the relation between the uplink transmit power vector and its corresponding uplink SINR vector can be rewritten as

\[
p = D(\gamma)Gp + D(\gamma)\eta \tag{8.5}
\]

where \(D(\gamma)\) denotes a diagonal matrix whose diagonal elements are the corresponding components of the SINR vector \(\gamma\), the \((i, j)\) component of \(G_{M \times M}\) is

\[
G_{ij} = \begin{cases} 
    \frac{h_{b_i,j}}{h_{b_i,i}}, & \text{if } i \neq j \\
    0, & \text{if } i = j 
\end{cases} \tag{8.6}
\]

and the \(i\)th component of \(\eta\) is \(\eta_i = \frac{\sigma_{b_i}^2}{h_{b_i,i}}\). From (8.5) we know that

\[
(I - D(\gamma)G)p = D(\gamma)\eta \tag{8.7}
\]

where \(I\) is a \(M \times M\) identity matrix. Given an uplink SINR vector, its corresponding uplink transmit power vector is thus computed by

\[
p = (I - D(\gamma)G)^{-1}D(\gamma)\eta. \tag{8.8}
\]

Let \(\gamma_i\) be the minimum acceptable SINR (known as target-SINR) of user \(i\). \(\gamma_i\) usually corresponds to the maximum tolerable bit-error rate, below which the user is not satisfied with its acceptable QoS.

**Definition 43** The SINR vector \(\gamma = [\gamma_1, \ldots, \gamma_M]^T\) is feasible if there exists a power vector \(0 \leq p \leq \bar{p}\) satisfying the target-SINR vector \(\gamma\). Moreover, the system is called
feasible if the target-SINR vector $\hat{\gamma}$ is feasible; otherwise, the system is called infeasible.

8.3 Protection Constraints for Primary Users and Different JPAC Problems in CRNs

We first define the concept of user protection. Then, by adding a new constraint on PUs protection to the problems discussed in Chapters 6 and 7, for conventional cellular networks, we formulate and review various optimization problems in underlay CRN, followed by a discussion on the need to characterize PUs’ protection in such optimization problems.

**Definition 44** For a given power vector $0 \leq p \leq \bar{p}$, a user $i \in M$ is said to be protected if $\gamma_i(p) \geq \hat{\gamma}_i$, where $\gamma_i(p)$ is obtained from (8.1). Correspondingly, for a given SINR vector $\gamma \geq \hat{\gamma}$, a user $i \in M$ is said to be protected if $0 \leq p_i(\gamma) \leq \bar{p}_i$ where $p_i(\gamma)$ is obtained from (8.8).

Assuming that all PUs can be protected in the absence of the SUs, it is desirable to design a scheme to admit all or a subset of SUs into the set of active users so that a given objective function $f_o(p)$ is optimized subject to the PUs’ protection constraints as well as the constraint that the transmit power levels for the SUs are feasible and the QoS requirements of the admitted SUs are met. This corresponds to the following general optimization problem [6]:

\[
\begin{align*}
\text{optimize} & \quad f_o(p) \\
\text{s.t.} & \quad \text{protection of PUs,} \\
& \quad \text{feasibility of transmit power level, and} \\
& \quad \text{QoS requirements for the admitted SUs.} \\
\text{variable} & \quad 0 \leq p \leq \bar{p}.
\end{align*}
\]

(8.9)

As implied by the above general optimization problem, for underlay cognitive radio networks, the same problems discussed in Chapters 6 and 7 for conventional cellular networks can be similarly stated and addressed as long as the protection of PUs is taken into account. In other words, from the optimization point of view, the same objective function, constraints, and variables considered in traditional wireless networks can be also considered in CRN, but with additional constraint on the protection of PUs. Accordingly, the existing solutions may be also extended to be employed in CRNs so that the PUs’ protection is guaranteed. Inspired by the power control problems studied in Chapters 6 and 7, one may similarly define different power control problems for underlay CRNs, three examples of which are given in what follows.

**Minimizing total transmit power subject to target-SINR constraint:** The optimization problem, corresponding to minimizing the total transmit power in underlay CRNs,
Joint Power and Admission Control in Cognitive Radio Networks

is formulated as follows:

\[
\begin{align*}
\min & \quad \sum_{i \in \mathcal{M}_p} p_i + \sum_{j \in \mathcal{M}_s} p_j \\
\text{s.t.} & \quad \gamma_i \geq \hat{\gamma}_i, \quad \forall i \in \mathcal{M}_p \\
& \quad \gamma_j \geq \hat{\gamma}_j, \quad \forall j \in \mathcal{M}_s \\
& \quad 0 \leq p_i \leq \bar{p}_i, \quad \forall i \in \mathcal{M}_p \\
& \quad 0 \leq p_j \leq \bar{p}_j, \quad \forall j \in \mathcal{M}_s
\end{align*}
\] (8.10a)

The constraints in (8.10b) and (8.10d) correspond to the PUs’ protection, and the constraints in (8.10c) and (8.10e) correspond to the QoS requirements and the feasibility of transmit power level for the SUs, respectively. In a feasible system where all PUs and SUs can be simultaneously protected, the TPC algorithm is employed to find the optimal solution for (8.10) (in a similar way for the conventional cellular networks); otherwise (i.e., when there exists no power vector for simultaneous protection of all PUs and SUs), the problem (8.10) has no solution, and alternatively the problem of minimizing SUs’ outage ratio needs to be considered, as explained in what follows.

Maximizing the number of protected SUs (minimizing SUs’ outage ratio) subject to PUs’ protection constraints in an infeasible system: In an infeasible system there exists no feasible power vector for protecting all PUs and SUs simultaneously. Given a power vector \( p \), let \( S^s \) denote the set of supported SUs achieving their desired target-SINRs:

\[
S^s(p) = \{ j \in \mathcal{M}_s | \gamma_j(p) \geq \hat{\gamma}_j \}.
\]

Correspondingly, let \( O^s(p) \) be the outage ratio of the SUs given by

\[
O^s(p) = \frac{|S^s(p)|}{M_s}
\]

where \( S^s(p) = \mathcal{M}_s \setminus S^s(p) \) and \( |.| \) denotes the cardinality of the corresponding vector. The problem of minimizing SUs’ outage ratio subject to PUs’ protection constraints is formally stated as follows:

\[
\begin{align*}
\min & \quad O^s(p) \\
\text{s.t.} & \quad \gamma_i \geq \hat{\gamma}_i, \quad \forall i \in \mathcal{M}_p \\
& \quad 0 \leq p_i \leq \bar{p}_i, \quad \forall i \in \mathcal{M}_p \\
& \quad 0 \leq p_j \leq \bar{p}_j, \quad \forall j \in \mathcal{M}_s
\end{align*}
\] (8.11)

In (8.11), the first constraint corresponds to the protection of PUs. Additionally, the feasibility of transmit power level and QoS requirements of the admitted SUs are satisfied through the constraints \( 0 \leq p_j \leq \bar{p}_j \) and \( \gamma_j \geq \hat{\gamma}_j, \forall j \in \mathcal{M}_s \), respectively.

The above optimization problem is NP-hard. There may also be many power vectors belonging to the set of optimal solutions for (8.11) and result in minimum outage ratio.
Among them, those solutions that correspond to the minimum aggregate transmit power of the supported users are of the most preference.

Maximizing the aggregate throughput of SUs in a feasible system: In a feasible system, where all PUs and SUs can simultaneously be protected, it may be preferred to assign SUs as high SINR values as possible. In other words, the objective may be defined as to maximize the aggregate throughput of SUs, as expressed in the following corresponding problem:

$$\max \sum_{j \in \mathcal{M}^s} \log(1 + \gamma_j(p))$$

subject to

$$\gamma_i(p) \geq \hat{\gamma}_i, \quad \forall i \in \mathcal{M}^p$$
$$0 \leq p_i \leq \bar{p}_i, \quad \forall i \in \mathcal{M}^p$$
$$0 \leq p_j \leq \bar{p}_j, \quad \forall j \in \mathcal{M}^s$$

variable $p$.

In (8.12), the first constraint corresponds to the protection of PUs, and the constraint $0 \leq p_j \leq \bar{p}_j, \forall j \in \mathcal{M}^s$ corresponds to the feasibility of transmit power level for the SUs.

8.4 Characterization of Feasible Interference Region

In general, as can be seen in (8.11) and (8.12), the objective function is a function of the SINRs of SUs. It is desirable to state and solve the optimization problem at the CRN level with minimal involvement of the PRN. That is, the corresponding objective and constraint functions for the CRN should depend on the transmit power levels of the SUs and the PUs’ protection constraints. These protection constraints should preferably not depend on the PRN variables such as the transmit power levels of the PUs. But it is observed that the PUs’ protection constraints in the form of $\gamma_i(p) \geq \hat{\gamma}_i$ for all $i \in \mathcal{M}^p$ expressed in the first constraint in all aforementioned problems (8.11) and (8.12) depend on the instantaneous power vector (or correspondingly SINR vector) of PUs. Therefore, to avoid this coupling (i.e., to decouple the optimization problem for CRN from that of PRN), we need to obtain an equivalent constraint for the PUs’ protection constraints so that it depends on the transmit power levels of SUs and minimal information pertaining to the PRN, and it is independent of the optimization variables (power levels) related to PUs. This constraint that has been derived in [6] based on the concept of a feasible cognitive interference region enables us to formulate and address the optimization problems at the CRN level, while the protection of PUs is guaranteed with minimal information feedback from the PRN.

As has been mentioned before, in the state-of-the-art interference management schemes for underlay CRNs (e.g., [3, 8–19]), it is assumed that the protection constraints for the PUs are satisfied if the cognitive interference for each primary receiving-point is lower than the interference temperature limit (ITL) of the corresponding
receiving-point. This corresponds to a box-like region, which is not really the case, as shown in [5] and [6] and explained below.

Let us now state the PUs’ protection constraints in terms of the constraints on the maximum interference that can be caused to PBSs. For each PBS, this can be defined in three ways, depending on the interference imposed (by all PUs and SUs, or all PUs and SUs from other cells, or by all SUs) on the primary receiving points as follows:

- Total received power by the corresponding PBS [7],
- Total inter-cell-interference imposed from all SUs together with PUs not being served by the corresponding PBS [5], and
- Cognitive interference imposed by all SUs on the corresponding PBS [5].

Corresponding to the above three items, three different presentations of the feasible interference region for the primary users protection are introduced as

- The feasible total received-power region (FRPR),
- The feasible total inter-cell interference region (FIIR), and
- The feasible total cognitive interference region (FCIR).

In what follows, the above feasible interference regions are characterized, and it is shown that while the FRPR and FIIR are box-like, the FCIR is a polyhedron (i.e., the maximum feasible cognitive interference threshold for each primary receiving-point is not a constant, and it depends on that for each of the other primary receiving-points).

### 8.4.1 Total Received-Power-Temperature: Expressing PUs’ Protection Constraints Based on FRPR

Given the uplink power vector \( \mathbf{p} \) corresponding to an SINR vector \( \gamma \), let \( \varphi_{m} \) denote the total received power plus noise at the BS \( k \in B^0 \):

\[
\varphi_k(\mathbf{p}) = \sum_{i \in M} p_i h_{k,i} + \sigma_k^2.
\]

Thus, we have

\[
\gamma_i = \frac{h_{b_i} p_i}{\varphi_{b_i} - h_{b_i} p_i}, \quad \forall i \in M,
\]

which results in the following lemma.

**Lemma 8**  Given the SINR vector of the PU \( i \) and the total received power by the PBS \( b_i \), the transmit power for each PU \( i \in M^0 \) (corresponding to \( \gamma_i \) and \( \varphi_{b_i} \)), denoted by \( p_i(\gamma_i, \varphi_{b_i}) \) is obtained as

\[
p_i(\gamma_i, \varphi_{b_i}) = \frac{\gamma_i}{(\gamma_i + 1)} \frac{\varphi_{b_i}}{h_{b_i,i}}
\]

(8.13)

where \( \varphi_{b_i} \) is the total received power plus noise at the BS of user \( i \), i.e., \( \varphi_{b_i} = \sum_{j \in M} h_{b_i,j} p_j + \sigma_{b_i}^2 \).

It is clear that if \( p_i(\gamma_i, \varphi_{b_i}) \leq \overline{p}_i \), then PU \( i \) is protected. Based on this, we define the feasible total received power region (FRPR) as follows.
**Definition 45** Let $\phi^p = [\varphi_1, \varphi_2, \ldots, \varphi_B]^T$ denote the vector of total received power by the PBSs. Given the target SINR vector of PUs $\widehat{\gamma}^p$, the FRPR $F_{\phi^p} \subset \mathbb{R}_+^B$ is defined as the vector space of total received power by the PBSs for which all the PUs are protected:

$$F_{\phi^p} = \{ \phi^p | 0 \leq p_i(\widehat{\gamma}_i, \varphi_{b_i}) \leq \overline{p}_i, \quad \forall i \in M_p, \quad \forall k \in B^p \}$$  \hspace{1cm} (8.14)

where $p_i(\widehat{\gamma}_i, \varphi_{b_i})$ is obtained from (8.13). In addition, given $\widehat{\gamma}^p$, we say that $\phi^p$ is feasible if $\phi^p \in F_{\phi^p}$.

Given $\widehat{\gamma}^p$, let $\delta_k$ denote the maximum value of the total received power plus noise at BS $k \in B^p$ that can be tolerated by all of its associated PUs (i.e., all of its associated PUs can be protected). We call $\delta_k$ as the total received-power temperature for PBS $k$, which is formally defined and obtained as follows [7]:

$$\delta_k = \max \{ \varphi | 0 \leq p_i(\widehat{\gamma}_i, \varphi) \leq \overline{p}_i, \quad \forall i \in M_p \} = \min_{i \in M_p} \left\{ \frac{p_i h_i}{\overline{p}_i} \right\}$$  \hspace{1cm} (8.15)

where $p_i(\gamma_i, \varphi_{b_i})$ is obtained from (8.13). From this, the following theorem is directly obtained.

**Theorem 61** The FRPR is a closed box of the following form:

$$0 \leq \varphi_k \leq \overline{\varphi}_k, \quad \forall k \in B^p,$$  \hspace{1cm} (8.16)

where $\overline{\varphi}_k$ is given by (8.15).

As can be seen, the total received-power temperature for each PBS $k \in B^p$, i.e., $\overline{\varphi}_k$, is a dynamic function of noise level, target-SINRs, channel-gains, and maximum transmit power levels for users associated with PBS $k$. In fact, the values of $\overline{\varphi}_k$ for all $k \in B^p$ indicate the amount of interference tolerability of PBSs at the primary network in the underlay spectrum access strategy. The value of the total received power temperature for each PBS dynamically decreases (increases) as the number of its associated PUs increases (decreases) and/or the channel status of primary network becomes weaker (stronger) [7].

**Lemma 9** If the transmit power vector $p$ satisfies the SINR requirements of all PUs, then we have $\varphi_k(p) \leq \overline{\varphi}_k$, for all $k \in B^p$, or equivalently, $\max_{k \in B^p} \{ \frac{\varphi_k(p)}{\overline{\varphi}_k} \} \leq 1$.

### 8.4.2 Total Inter-Cell Interference Temperature: Expressing PUs' Protection Constraints Based on FIIR

Given the uplink power vector $p$ corresponding to an SINR vector $\gamma$, let $I^p_k$ denote the total inter-cell interference imposed by all SUs together with PUs $i \notin M_p^k$ on the PBS
Joint Power and Admission Control in Cognitive Radio Networks

\[ k \in \mathcal{B}^p: \]
\[
I_k^p = \sum_{i \in \mathcal{M}_{k}}^{h_{ki}} p_i h_{ki} = \sum_{i \in \mathcal{M}_{k}^p}^{i \notin \mathcal{M}_{k}} p_i h_{ki} + \sum_{i \in \mathcal{M}_k}^{h_{ki}} p_i h_{ki}, \quad \forall k \in \mathcal{B}^p. \tag{8.17}
\]

**Lemma 10** For each PBS \( k \in \mathcal{B}^p \), given the SINRs of the PUs in the corresponding cell (i.e., \( \gamma_p^k \)) and the total inter-cell interference caused by other cells to that cell (i.e., \( I_p^k \)), the transmit power of each user \( i \in \mathcal{M}_k^p \), denoted by \( p_i(\gamma_p^k, I_p^k) \), is obtained as

\[
p_i(\gamma_p^k, I_p^k) = \frac{1}{h_{ki}} \frac{\gamma_i}{(\gamma_i + 1)} \left( \frac{I_p^k + \sigma^2_k}{1 - \sum_{j \in \mathcal{M}_k^p}^{j \neq i} \frac{\gamma_j}{\gamma_j + 1}} \right). \tag{8.18}
\]

**Proof** Let \( \varphi_{h_i} \) be the total received power plus noise at the BS of user \( i \), i.e., \( \varphi_{h_i} = \sum_{j \in \mathcal{M}_k} h_{h_{ij}} p_j + \sigma^2_{h_i} \). Thus, we have \( \gamma_i = \frac{h_{h_{ij}} p_i}{\varphi_{h_i} - h_{h_{ij}} p_i}, \forall i \in \mathcal{M}_k \), which results in

\[
p_i = \frac{\gamma_i}{(\gamma_i + 1)} h_{h_{ji}}. \tag{8.19}
\]

From (8.19) for each \( k \in \mathcal{B} \) and \( n \in \mathcal{B} \), the following is obtained:

\[
\sum_{i \in \mathcal{M}_n}^{h_{ki}} p_i h_{ki} = \varphi_n \sum_{i \in \mathcal{M}_n}^{h_{ki}} h_{ki} \frac{\gamma_i}{h_{h_{ki}} + 1}. \tag{8.20}
\]

By letting \( k = n \) and adding \( \sum_{i \notin \mathcal{M}_k}^{h_{ki}} p_i h_{ki} + \sigma^2_k \) to both sides of (8.20), \( \varphi_k \) is obtained as

\[
\varphi_k = \frac{\sum_{i \notin \mathcal{M}_k}^{h_{ki}} (p_i h_{ki}) + \sigma^2_k}{1 - \sum_{i \in \mathcal{M}_k}^{h_{ki}} (\gamma_i / \gamma_i + 1)}, \quad \forall k \in \mathcal{B}. \tag{8.21}
\]

From (8.19) and (8.21) and the fact that \( I_k^p = \sum_{i \notin \mathcal{M}_k}^{h_{ki}} (p_i h_{ki}) \), (8.18) is concluded.

**Definition 46** Let \( \mathbf{I}^p = [I_1^p, I_2^p, \ldots, I_{\mathcal{B}_p}^p]^T \) denote the vector of total inter-cell interference imposed on the PBSs. Given the target SINR vector of PUs \( \hat{\gamma}_p \), the FIIR \( \mathbf{F}_I^p \subset \mathbb{R}^{\mathcal{B}_p} \) is defined as the vector space of total inter-cell interference imposed on the PBSs for which all the PUs are protected:

\[
\mathbf{F}_I^p = \{ \mathbf{I}^p | 0 \leq p_i(\hat{\gamma}_p, I_k^p) \leq \bar{p}_i, \forall i \in \mathcal{M}_k, \forall k \in \mathcal{B}^p \} \tag{8.22}
\]

where \( p_i(\hat{\gamma}_p, I_k^p) \) is obtained from (8.18). In addition, given \( \hat{\gamma}_p \), we say that \( \mathbf{I}^p \) is feasible if \( \mathbf{I}^p \in \mathbf{F}_I^p \).

Given \( \hat{\gamma}_p \), let \( \overline{I}_k^p \) denote the maximum value of the total inter-cell interference at BS \( k \in \mathcal{B}^p \) that can be tolerated by all of its associated PUs (i.e., all of its associated PUs can be protected). We call \( \overline{I}_k^p \) the total inter-cell interference temperature limit (TITL).
8.4 Characterization of Feasible Interference Region

of the PBS $k \in B^p$, which is formally defined and obtained as follows:

$$\bar{T}_k^p = \max \{ I \mid 0 \leq p_i(\gamma^p, I) \leq \bar{p}_i, \forall i \in M_k^p \} = \bar{\varphi}_k \times \left( 1 - \sum_{j \in M_k} \frac{\gamma_j}{\gamma_j + 1} \right) - \sigma_k^2$$

(8.23)

where $p_i(\gamma^p, I)$ is obtained from (8.18) and $\bar{\varphi}_k$ is the total received-power temperature limit given by (8.15). From this, the following theorem is directly obtained.

**Theorem 62** [6] The FIIR is a closed box of the following form:

$$0 \leq I_k^p \leq \bar{T}_k^p, \quad \forall k \in B^p$$

(8.24)

where $\bar{T}_k^p$ is given by (8.23).

8.4.3 Total Cognitive Interference Temperature: Expressing PUs’ Protection Constraints Based on FCIR

Given the uplink power vector $\mathbf{p}$ corresponding to an SINR vector $\gamma$, let $I_k^{\rightarrow p}$ denote the interference caused by all of the SUs to the PBS $k \in B^p$, called the cognitive interference caused to the PBS $k \in B^p$:

$$I_k^{\rightarrow p} = \sum_{i \in M} (p_i h_{ki}), \quad \forall k \in B^p.$$  

(8.25)

**Lemma 11** [6] Given the SINR vector of the PUs $\gamma^p$ and the cognitive interference vector imposed on the PBSs by the SUs ($I^{\rightarrow p} = [I_1^{\rightarrow p}, I_2^{\rightarrow p}, \ldots, I_{B^p}^{\rightarrow p}]^T$), the transmit power for each PU $i \in M^p$ (corresponding to $\gamma^p$ and $I^{\rightarrow p}$), denoted by $p_i(\gamma^p, I^{\rightarrow p})$, is obtained as

$$p_i(\gamma^p, I^{\rightarrow p}) = \frac{\gamma_i}{(\gamma_i + 1)} \times \frac{\varphi_{hi} (\gamma^p, I^{\rightarrow p})}{h_{hi}}, \quad \forall i \in M^p$$

(8.26)

where $\phi^p = [\varphi_1, \varphi_2, \ldots, \varphi_{B^p}]^T$ is given by

$$\phi^p (\gamma^p, I^{\rightarrow p}) = (I - \mathbf{H}(\gamma^p))^{-1} (\mathbf{N}^p + I^{\rightarrow p})$$

(8.27)

in which $\mathbf{I}$ is a $B^p \times B^p$ identity matrix, $\mathbf{N}^p = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_{B^p}^2]^T$ and $\mathbf{H}(\gamma^p)$ is a $B^p \times B^p$ matrix whose elements are given by

$$H_{mn} = \begin{cases} \sum_{i \in M_n} \frac{\gamma_i}{\gamma_i + 1}, & \text{if } m = n, \\ \sum_{i \in M_n} \frac{\nu_i}{\nu_i + 1}, & \text{if } m \neq n. \end{cases}$$

(8.28)
Joint Power and Admission Control in Cognitive Radio Networks

Proof From (8.20) and (8.21), for any PBS \( m \in B^p \), we have

\[
\varphi_m = \frac{\sum_{i \notin M_m} (p_i h_{mi}) + \sigma_m^2}{1 - \sum_{i \in M_m} \left( \frac{\gamma_i}{\gamma_i + 1} \right)} - \sum_{i \in M_m} \frac{h_{mi}}{h_{mi}} \left( \frac{\gamma_i}{\gamma_i + 1} \right)
\]

\[
= \frac{\sum_{n \in B^p \setminus n \neq m} (\varphi_n \sum_{i \in M_n} h_{mi} \left( \frac{\gamma_i}{\gamma_i + 1} \right)) + \sigma_m^2}{1 - \sum_{i \in M_m} \left( \frac{\gamma_i}{\gamma_i + 1} \right)}, \quad \forall m \in B^p.
\]

From (8.29), we conclude that

\[
\varphi_m \left( 1 - \sum_{i \in M_k} \left( \frac{\gamma_i}{\gamma_i + 1} \right) \right) - \sum_{n \in B^p \setminus n \neq m} \varphi_n \sum_{i \in M_n} h_{mi} \left( \frac{\gamma_i}{\gamma_i + 1} \right) = \mathbf{I}^{s \to p}_m + \sigma_m^2, \quad m = 1, 2, \ldots, B^p.
\]

Writing (8.30) in matrix form results in (8.27), which together with (8.19) results in (8.26).

Definition 47 Given the target SINR vector of PUs \( \hat{\gamma}^p \), the feasible cognitive interference region (FCIR) \( F_{s \to p}^{I} \subset \mathbb{R}^{B^p} \) is defined as the space of interference vectors imposed by the CRN (i.e., SUs) on the PBSs for which all the PUs are protected,

\[
F_{s \to p}^{I} = \{ \mathbf{I}^{s \to p} | 0 \leq p_i(\hat{\gamma}^p, \mathbf{I}^{s \to p}) \leq \overline{\gamma}_i, \quad \forall i \in M_k, \quad \forall k \in B^p \}
\]

where \( p_i(\gamma^p, \mathbf{I}^{s \to p}) \) is obtained from (8.26). Furthermore, we say that a given cognitive interference vector \( \mathbf{I}^{s \to p} \) is feasible if \( \mathbf{I}^{s \to p} \in F_{s \to p}^{I} \).

As can be seen from (8.26), the transmit power of PU \( i \) depends not only on cognitive interference at its serving PBS \( b_i \), but also on cognitive interference at the other PBSs. This is in contrast to relations (8.13) and (8.18), which give the transmit power of PU \( i \) as a function of the total received power and inter-cell interference, respectively, at its serving PBS \( b_i \) only (independent of those at other PBSs). For this reason, in contrast to FRPR and FIIR, both of which are box-like, the FCIR is a polyhedron, which means that the values of cognitive interference temperature limits depend on each other, as shown in the following theorem.

Theorem 63 [6] The FCIR is a polyhedron given by the following matrix inequalities:

\[
0 \leq \mathbf{I}^{s \to p}
\]
8.4 Characterization of Feasible Interference Region

Figure 8.1 (a) FRPR, (b) FIIR, and (c) FCIR of a CRN having two PBSs.

and

\[ \text{AI}^{s\rightarrow p} \leq \mathbf{C}^p \] (8.33)

where

\[ \mathbf{A}^p = (\mathbf{I} - \mathbf{H}(\gamma^p))^{-1} \] (8.34)

and

\[ \mathbf{C}^p = \overline{\phi}^p - (\mathbf{I} - \mathbf{H}(\gamma^p))^{-1} \mathbf{N}^p \] (8.35)

in which \( \mathbf{I} \) is a \( B^p \times B^p \) identity matrix and \( \overline{\phi}^p = [\overline{\varphi}_1, \overline{\varphi}_2, \ldots, \overline{\varphi}_{B^p}]^T \), where \( \overline{\varphi}_k \) is given by (8.15).

**Proof** Since we have assumed that the feasibility of the system holds in the absence of SUs (i.e., when \( I_k^{s\rightarrow p} = 0, \forall k \in B^p \)), we have \( 0 \leq p_i(\gamma^p, I^{s\rightarrow p})|_{I_{s\rightarrow p}=0} \). Therefore, from (8.26) and (8.27), we conclude that \((\mathbf{I} - \mathbf{H}(\gamma^p))^{-1}\) exists, and from the Perron-Frobenius theorem, it is positive component-wise. The feasibility of \( I^{s\rightarrow p} \) leads to the protection of all PUs for all \( k \in B^p \). Therefore, from Definition 47 and from (8.26) and (8.27) it is concluded that

\[ 0 \leq (\mathbf{I} - \mathbf{H}(\gamma^p))^{-1} \mathbf{N}^p \leq \overline{\phi}^p. \] (8.36)

The inequality in the right-hand side of (8.36) directly results in (8.32). The inequality in the left-hand side of (8.36) together with the fact that the feasible cognitive interference imposed on any PBS is non-negative results in the following:

\[ \max\{-(\mathbf{I} - \mathbf{H}(\gamma^p))^{-1} \mathbf{N}^p, 0\} \leq I^{s\rightarrow p} \]

which leads to (8.33). This completes the proof.

8.4.4 Example and Discussion

In the following, an example of a two-dimensional FRPR, FIIR and FCIR is illustrated for a simple PRN consisting of two PBSs coexisting with a CRN (see Figure 8.1), and then it is argued which of the FRPR, FIIR, or FCIR is preferred to express the PUs’ protection in practical resource allocation problems for CRNs.
Example 64 [6] Consider a CRN wherein the primary tier consists of two base-stations ($B^p = \{1, 2\}$). From Theorems 61 and 62, the FRPR and FIIR is the box enclosed by the inequalities

\[
0 \leq \varphi_1 \leq \overline{\varphi}_1 \\
0 \leq \varphi_2 \leq \overline{\varphi}_2
\]

and

\[
0 \leq I^p_1 \leq \overline{I}^p_1 \\
0 \leq I^p_2 \leq \overline{I}^p_2
\]

respectively, where $\overline{\varphi}_1$ and $\overline{\varphi}_2$, and $\overline{I}^p_1$ and $\overline{I}^p_2$ are obtained from (8.15) and (8.23), respectively. From Theorem 63, the FCIR is the space enclosed by the following inequalities:

\[
A^p_{11}I^p_{1\rightarrow p} + A^p_{12}I^p_{2\rightarrow p} \leq C^p_1 \\
A^p_{21}I^p_{1\rightarrow p} + A^p_{22}I^p_{2\rightarrow p} \leq C^p_2 \\
0 \leq I^p_{1\rightarrow p} \\
0 \leq I^p_{2\rightarrow p}
\]

where the coefficients $A^p$ and $C^p$ are obtained from (8.34) and (8.35), respectively, as follows:

\[
\begin{pmatrix}
A^p_{11} & A^p_{12} \\
A^p_{21} & A^p_{22}
\end{pmatrix} = \frac{1}{K(\gamma^p)} \times \begin{pmatrix}
1 - \sum_{i \in M_2} \frac{\gamma_i}{\gamma_i + 1} & \sum_{i \in M_2} \frac{h_{1i}}{h_{2i}} \cdot \frac{\gamma_i}{\gamma_i + 1} \\
\sum_{i \in M_1} \frac{\gamma_i}{\gamma_i + 1} & 1 - \sum_{i \in M_1} \frac{\gamma_i}{\gamma_i + 1}
\end{pmatrix}
\]

and

\[
\begin{pmatrix}
C^p_1 \\
C^p_2
\end{pmatrix} = \begin{pmatrix}
\overline{\varphi}_1 - A^p_{11}\sigma^2_1 - A^p_{12}\sigma^2_2 \\
\overline{\varphi}_2 - A^p_{21}\sigma^2_1 - A^p_{22}\sigma^2_2
\end{pmatrix}
\]

where

\[
K(\gamma^p) = \left(1 - \sum_{i \in M_2} \frac{\gamma_i}{\gamma_i + 1}\right) \left(1 - \sum_{i \in M_1} \frac{\gamma_i}{\gamma_i + 1}\right) - \left(\sum_{i \in M_1} \frac{h_{1i}}{h_{1i} + 1}\right) \left(\sum_{i \in M_2} \frac{h_{1i}}{h_{2i} + 1}\right).
\]

(8.37)

Theorems 61, 62, and 63 derive an equivalent constraint for PUs’ protection in terms of FRPR, FIIR, and FCIR, respectively. In other words, the PUs’ protection constraints expressed as $0 \leq p_i \leq \overline{p}_i$ and $\gamma_i(p) \geq \overline{\gamma}_i$, for all $i \in \mathcal{M}^p$ (for example, in the optimization problems (8.11) and (8.12)) can now be replaced with the either of the constraints in (8.16) or (8.24) or the constraints in (8.32) and (8.33).
Note that in the FCIR characterized by (8.32) and (8.33), the feasible cognitive interference vector $I^{\rightarrow P}$ is a function of the transmit power level for SUs (as defined in (8.25)), and the coefficients $A_p$ and $C_p$ are a function of $\gamma_p$ (as seen in (8.34) and (8.35)), and both are independent of instantaneous transmit power levels of PUs. This is in contrast to FRPR and FIIR, which provide an equivalent constraint for PUs’ protection that depends on instantaneous transmit power levels of PUs as seen in (8.16) or (8.24). Therefore, in contrast to FRPR and FIIR, FCIR enables us to state the interference optimization problems at the CRN level with minimal involvement of PRN, i.e., only the coefficients $A_p$ and $C_p$, which are a function of $\gamma_p$ (as seen in (8.34) and (8.35)), are required to be given by PRN to the CRN.

### 8.5 Existing Centralized JPAC Algorithms to Maximize the Number of Supported SUs Subject to PUs’ Protection Constraint

Subject to the constraint of guaranteeing the QoS of the currently active PUs in CRNs, the existing JPAC algorithms try to find and admit a subset of SUs to optimize different objectives. These objectives include maximizing the number of admitted users, maximizing the aggregate throughput, and minimizing the aggregate transmit power. Considering the objective of maximizing the number of admitted SUs, ideally, we would like to find a globally optimal solution to the admission control problem. This problem is generally NP-hard and requires an exhaustive search through all possible sub-sets of admitted SUs. To overcome this difficulty, most of the existing JPAC algorithms focus on obtaining a sub-optimal solution in either a centralized or a distributed manner.

The existing centralized JPAC algorithms are categorized according to the procedures of finding the maximum feasible set. The existing centralized JPAC algorithms are categorized as (see Figure 8.2) optimal exhaustive searching algorithms (ESAs), random searching algorithms (RSAs), and sequential searching algorithms (SSAs) [3]. The exhaustive search algorithm does search over all possible set of admitted SUs whose target-SINRs along with target-SINRs of all PUs are feasible. This algorithm obtains the globally optimal solution, but it would lead to unaffordable computational complexity. The second group of algorithms are based on probabilistic mechanisms used for the SUs to access the channel. Such algorithms suffer from low speed of convergence. In the last group of JPAC algorithms, the opportunity to access the spectrum is quantified by assigning some removal metric (or some admission metric) to each SU. In other words, the greater the value for removal criterion, the lower the opportunity to access the spectrum. Therefore, the SUs with greater values of the removal criterion are gradually removed until the network becomes feasible for the remaining SUs along with the PUs.

Depending on whether the removal criterion of SUs is determined before the search procedure or not, the SSAs may be sub-categorized into the following classes: the SSAs using *short-sighted* removal metrics and the SSAs using *fore-sighted* removal metrics [3]. In the latter algorithms, the removal metric of each SU is available before the search procedure. However, in the former algorithms, the removal metrics of the SUs are not
known in advance and are updated as any of the SUs are removed at each iteration of the removal process.

As has been mentioned before, the ESA and RSA suffer from computational complexity and low speed of convergence, respectively; hence, the main focus of the current chapter is on SSAs. In what follows, the general procedure for an SSA is first explained, and then three of existing SSAs are introduced.

8.5.1 A General SSA

The pseudocode for a general SSA is given below.

**Algorithm 1 Pseudocode for the SSA**

**Initialization phase:**
Let $\mathcal{M}^p$, and $\mathcal{M}^s$ be sets of the PUs, and SUs, respectively and let $A^s = \mathcal{M}^s$ be the admitted SUs

**Step 1 (Feasibility checking phase):**
Check the feasibility of the target-SINRs vector of the set $A^s \cup \mathcal{M}^p$
if the target-SINRs for the set $A^s \cup \mathcal{M}^p$ is feasible
Admit all SUs in $A^s$
Terminate the algorithm
else
Go to Step 2

**Step 2 (Removal phase):**
$i^* = \arg\max_{i \in A^s} RC_i$, where $RC_i$ is a removal criterion measured either at this step (short-sighted) or at initialization phase (fore-sighted)
$A^s \leftarrow A^s \setminus i^*$
Go back Step 1

In this algorithm, at first (initialization phase), all SUs together with PUs are admitted. Then at the feasibility checking phase, the required target SINRs for all PUs and
admitted SUs are checked. If it is feasible, the algorithm terminates, else the SUs are sequentially removed based on a certain removal (or admission) metric, until the target SINRs for all PUs together with remained SUs become feasible. As can be seen, a sequential searching algorithm consists of two main phases of feasibility checking and SU removal, which are explained below.

**Feasibility Checking Phase**

There are two mechanisms for feasibility checking of a given SINR vector in a conventional cellular wireless network. They are centralized and distributed feasibility checking mechanisms, which may be also used in underlay CRNs.

**Centralized feasibility checking:** As derived in Chapter 5, the power vector $\mathbf{p}$ corresponding to a given SINR vector $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_M]^T$ is obtained as

$$\mathbf{p} = (\mathbf{I} - \mathbf{D}(\gamma))^{-1} \mathbf{D}(\gamma) \eta.$$  

Therefore, the target-SINR vector $\hat{\gamma}$ is feasible, if

$$0 \leq (\mathbf{I} - \mathbf{D}(\hat{\gamma}) \mathbf{G})^{-1} \mathbf{D}(\hat{\gamma}) \eta \leq \bar{p}. \quad (8.38)$$

This enables us to check the feasibility of a given target SINR vector.

**Distributed feasibility checking:** We know that when the target-SINRs for a set of user are feasible, the distributed Target-SINR tracking Power Control (TPC) algorithm proposed in [1] has the following constrained power updating function:

$$p_i(t + 1) = \max \left\{ p_i, \frac{\hat{\gamma}_i}{\gamma_i(\mathbf{p}(t))} p_i(t) \right\}, \quad \forall i \in \mathcal{M} \quad (8.39)$$

where $\gamma_i(\mathbf{p}(t))$ is the actual SINR of user $i$ at iteration $t$, and solves the following optimization problem in a distributed manner:

$$\begin{align*}
\min_{\mathbf{p}} & \sum_{i \in \mathcal{M}} p_i \\
\text{s.t.} & \gamma_i \geq \hat{\gamma}_i, & \forall i \in \mathcal{M} \quad (8.40b) \\
& 0 \leq p_i \leq \bar{p}_i, & \forall i \in \mathcal{M} \quad (8.40c) \\
\text{variable} & \mathbf{p}. \quad (8.40d)
\end{align*}$$

It can be shown that, if the target-SINR vector is feasible, the optimal transmit power vector of (8.40) is given by

$$\mathbf{p}^* = (\mathbf{I} - \mathbf{D}(\hat{\gamma}) \mathbf{G})^{-1} \mathbf{D}(\hat{\gamma}) \eta$$  

which is obtained by the TPC in a distributed manner. In other words, when the target-SINR vector is feasible, for the optimal solution to the problem (8.40), equality in constraint (8.40b) holds, which implies all users meet their target-SINRs exactly. In this case, the TPC algorithm converges to a unique fixed-point that is the optimal solution of the problem (8.40). Consequently, the TPC algorithm can be employed as a distributed feasibility checking scheme. In fact, when all users employ the TPC, depending on the
feasibility or infeasibility of the target-SINR vector, either of the following two cases may happen, respectively, at the fixed-point to which the algorithm converges:

- If and only if the target-SINR vector is feasible, then all users (employing the TPC) reach their target-SINRs at the equilibrium.
- If and only if the target-SINR vector is infeasible, then there are some (at least one) users at the equilibrium that do not obtain their target-SINRs, while they are transmitting at their maximum power.

### Removal Phase

As will be seen, the existing SSAs are mostly different in kind of removal metric adopted for identifying SU candidates for sequentially removing. In a general SSA (Algorithm 1), the removal (or admission) criterion may be either short-sighted or fore-sighted. In SSAs with a short-sighted removal criterion, the admission metrics for SUs depend on instantaneous transmit power such as interference, effective interference, and SINR. Therefore, the removal metric of each SU is dynamically updated as any of the SUs are removed from the admitted set (and thus the transmit powers of users are updated) at each iteration. On the other hand, in SSAs with a fore-sighted removal criterion, the removal criterion is independent of the instantaneous transmit power and is expressed as a function of parameters such as the users’ path-gains, target-SINR, and maximum transmit power, which causes the removal criterion of SUs to remain fixed as any SUs are removed from the admitted set at each iteration. Thus the short-sighted removal criterion is iteratively updated at removal phase; and on the other hand, the fore-sighted removal criterion is calculated at the initialization phase, and it does not change at the removal phase (as long as the path-gains are fixed during the convergence time of the SSA). For this reason, the SSAs with a short-sighted removal criterion have a higher order of complexity because the removal criteria (or admission metrics) for the SUs are not known in advance and are updated as any of SUs are removed at each iteration. In what follows, three SSAs are introduced: one SSA with short-sighted and two SSAs with fore-sighted removal criteria.

In [2] and [3], the optimization problem of minimizing the SUs’ outage ratio subject to PUs’ protection constraints is formally stated as follows:

\[
\begin{align*}
\max & \quad |A^e| \\
\text{s.t.} & \quad \gamma_i \geq \gamma_i^T, \quad \forall i \in A^e, \quad (8.42a) \\
& \quad \sum_{j \in A^P} h_{kj} p_j \leq ITL_k, \quad \forall k \in B^p, \quad (8.42b) \\
& \quad 0 \leq p_i \leq \bar{p}_i, \quad \forall i \in M^p, \quad (8.42c) \\
& \quad 0 \leq p_i \leq \bar{p}_i, \quad \forall i \in A^e, \quad (8.42d) \\
\text{variable} & \quad p \quad (8.42e)
\end{align*}
\]

where $A^e \subseteq M^e$ is the set of admitted SUs, i.e., $A^e = \{ j \in M^e \mid p_j \neq 0 \}$. The constraint (8.42c) is assumed to be corresponding to PUs’ protection, and those in (8.42e)
and (8.42b) correspond to the feasibility of the transmit power level and QoS requirements of admitted SUs, respectively. In (8.42c), ITL\(_k\) is the cognitive interference temperature limit at PBS \(k\), as defined in Section 8.4.3 as the maximum value of the total interference caused by the SUs to the PBS \(k\), that can be tolerated by all of its PUs. Although, in problem (8.42), the ITL value at each primary receiving-point is implicitly assumed fixed and independent of the ITL values for other primary receiving-points (i.e., box-like ITL values are assumed), it is not really the case, as shown in Section 8.4.3. However, since, the existing JPAC algorithms in literature mostly consider the problem of maximizing the number of supported SUs subject to PUs’ protection constraint in the form of (8.42c), we also introduce them assuming the fixed and box-like ITL values.

### 8.5.2 Interference Constraint-Aware Stepwise Maximum Interference Removal Algorithm (I-SMIRA)

Let us define the dynamic parameters \(\alpha_i(p)\), \(\beta_i(p)\) and \(D(p)\) as follows:

\[
\alpha_i(p) = \left[ p_i \sum_{j \in M', j \neq i} h_{b,j} + N_{b,i} \right] - \frac{h_{b,i}}{\gamma_i} p_i \tag{8.43}
\]

\[
\beta_i(p) = \left[ \sum_{j \in M', j \neq i} h_{b,j} p_j + N_{b,i} \right] - \frac{h_{b,i}}{\gamma_i} p_i \tag{8.44}
\]

\[
D(p) = \sum_{i \in A'} \beta_i(p) \tag{8.45}
\]

where \(N_{b,i}\) is the total noise plus interference power caused by PUs at the SBS corresponding to the SU \(i\) (i.e., \(N_{b,i} = \sum_{k \in M'} h_{b,k} p_k + \sigma_k^2\)). Intuitively, \(\alpha_i\) quantifies the aggregate relative interference that SU \(i\) causes to other SUs, and \(\beta_i\) reflects the degree by which the QoS constraint for SU \(i\) is violated. It can be easily found that

\[
D = \sum_{j \in A'} \beta_j(p) = \sum_{j \in A'} \alpha_j(p).
\]

Additionally, if the QoS constraint for SU \(j\) is satisfied with equality, then \(\beta_j(p) = 0\), and therefore, we have \(D = 0\) if and only if all admitted SUs in \(A'\) are supported.

The I-SMIRA algorithm [2] decides the set of admitted SUs as follows. The target SINRs for all PUS and admitted SUs are checked using the TPC algorithm. If all PUs and the admitted SUs obtain their target-SINRs, the algorithm is terminated. Otherwise, two cases are considered, as explained below:

- **Case I**: The interference limits of the PBSs are not violated (i.e., (8.42c) holds), but there exists at least one admitted SU with a violated QoS requirement (i.e., (8.42b) does not hold).
In this case, an SU is removed based on the following removal criterion:

\[ RC_i = \max(\alpha_i(p), \beta_i(p)). \]  

(8.46)

- **Case II: The interference limits of PBSs are violated.** In this case, the I-SMIRA algorithm would remove the SU which violates both QoS and interference constraints the most in each step. Now, let us define the measure that quantifies degree of violation at primary receiving point \( k \) as follows:

\[ \eta_k = ITL_k - \sum_{j \in A_s} h_{kj} p_j. \]  

(8.47)

In this case, an SU is removed based on the following removal criterion that quantifies the aggregate interference effect that each SU creates to other SUs.

In the I-SMIRA algorithm, the removal criterion for each user \( i \) (RC\( _i \)) is expressed as follows:

\[
RC_i = \frac{D(p)}{D(p) + \sum_{k \in B^p} \eta_k} \times \max \left[ \sum_{j \in A', j \neq i} h_{b_j,j} p_j, \sum_{j \in A', j \neq i} h_{b_j,j} p_l \right] + \sum_{k \in B^p} \frac{\eta_k}{D(p) + \sum_{j \in B^p} \eta_j} h_{ki} p_i.
\]  

(8.48)

Note that \( \sum_{j \in A', j \neq i} h_{b_j,j} p_j \) is the total interference that the admitted SU \( i \) creates to other SUs, while \( \sum_{j \in A', j \neq i} h_{b_j,j} p_l \) is the total interference caused for the admitted SU \( i \). Moreover, \( h_{ki} p_i \) denotes the interference that the SU \( i \) creates for PBS \( k \). Recall that \( D(p) \) quantifies the degree of violation for QoS constraints, and \( \eta_k \) quantifies the degree of violation for the interference constraint of PBS \( k \). Therefore, the removal criterion (8.48) removes in each step the SU that creates the largest amount of interference for PBS and receives or creates the largest amount of interference from or for other SUs in the weighted average sense [2] (see Algorithm 2).

### 8.5.3 Link-Gain Ratio Algorithm (LGRA) and Effective Link-Gain Ratio Algorithm (ELGRA)

To solve the optimization problem in (8.42) sub-optimally, the LGRA is proposed in [3]. LGRA proposes a fore-sighted admission metric for each SU \( j \) that is defined as follows:

\[
\min_{k \in B} ITL_k \frac{h_{b_j,j}}{h_{kj}}.
\]  

(8.49)

The admission metric of each SU \( j \), (8.49), is developed from the constraints on its transmit power. In more detail, substituting \( \gamma_j = \frac{p_j h_{b_j,j}}{I_j(p)} \) into the constraint (8.42b) gives

\[
p_j \geq \frac{\gamma_j}{h_{b_j,j}} \left( N_{b_j} + \sum_{i \in M', i \neq j} p_i h_{b_i,j} \right).
\]  

(8.50)
Algorithm 2 I-SMIRA [2]

1: **Initialization:**
2: \( \mathcal{A}^s \leftarrow \mathcal{M}^s \)
3: **Step 1 (Feasibility checking phase):**
4: Starting from a non-negative power vector, let \( \mathbf{p} \) be the equilibrium obtained through the following iterative power update function:
5: \[
\begin{align*}
    p_i^{t+1} &= \begin{cases} 
        \min\{\bar{p}_i, \frac{\gamma_i(\mathbf{p}_t)}{\gamma_i(\mathbf{p})} p_i^t\}, & \text{if } i \in \mathcal{M}^p \cup \mathcal{A}^s \\
        0, & \text{if } i \in \mathcal{M}^s \setminus \mathcal{A}^s
    \end{cases}
\end{align*}
\]
6: Check the feasibility of the target SINR vector of \( \mathcal{A}^s \cup \mathcal{M}^p \) (if all PUs are protected and all admitted SUs are supported with their target-SINRs)
7: **if** the target SINR vector is feasible
8: Admit all SUs in \( \mathcal{A}^s \)
9: Terminate the algorithm
10: **else**
11: **Go to Step 2**
12: **Step 2 (Removal phase):**
13: If the interference limits of PBSs are not violated (i.e., \( (8.42c) \) holds) but \( \exists i \in \mathcal{A}^s \) such that \( \gamma_i(\mathbf{p}) < \hat{\gamma}_j \), let \( i^* = \arg\max_{i \in \mathcal{A}^s} \{\max(\alpha_i(\mathbf{p}), \beta_i(\mathbf{p}))\} \)
14: If the interference limits of PBSs are violated, let \( i^* = \arg\max_{i \in \mathcal{A}^s} RC_i \):
15: where \( RC_i \) is given by \( (8.48) \)
16: \( \mathcal{A}^s \leftarrow \mathcal{A}^s \setminus i^* \)
17: **Go back to Step 1**

From \( (8.42c) \), we have
\[
    p_j \leq \frac{\text{ITL}_k}{h_{k,j}} - \sum_{i \in \mathcal{M'}, i \neq j} p_i h_{k,i} h_{k,j}, \quad \forall k \in \mathcal{B}^p. \tag{8.51}
\]

By combining \( (8.50) \) and \( (8.51) \), we obtain
\[
    \frac{\hat{\gamma}_j}{h_{b,j}} \left( N_{b,j} + \sum_{i \in \mathcal{M'}, i \neq j} p_i h_{b,i} \right) \leq p_j \leq \frac{\text{ITL}_k}{h_{k,j}} - \sum_{i \in \mathcal{M'}, i \neq j} p_i h_{k,i} h_{k,j}, \quad \forall k \in \mathcal{B}^p. \tag{8.52}
\]

Therefore, it is concluded that
\[
    f_{k,j} \leq \text{ITL}_k \times \frac{h_{b,j} h_{k,j}}{h_{k,j}}, \quad \forall k \in \mathcal{B}^p \tag{8.53}
\]

where \( f_{k,j} = \gamma_j \times N_{b,j} + \sum_{i \in \mathcal{M'}, i \neq j} (\hat{\gamma}_j h_{b,i} + \frac{h_{b,j}}{h_{k,j}} h_{b,j}) p_i \), is the interference caused by other secondary users to secondary user \( j \); hence, \( \text{ITL}_k \times \frac{h_{b,j}}{h_{k,j}} \) reflects the interference-resistance capability of the secondary user \( j \) [3]. Therefore, according to
the LGRA, SUs with the lowest admission metric (i.e., argmin\(_{k \in B^p} ITL_k \times \frac{h_{b,j}}{h_{b,j}}\), or equivalently, argmax\(_{k \in B^p} \frac{1}{\prod_{j} \frac{h_{b,j}}{h_{b,j}}} \times \frac{h_{b,j}}{h_{b,j}}\)) are gradually removed until the target-SINRs for all PUs together with the admitted SUs become feasible. In other words, the user who has the lower link-gain ratio with respect to a PBS that has a low interference temperature is more likely to be removed. The LGRA is given by Algorithm 3, where the removal criterion \(RC_i\) is defined as

\[
RC_i = \max_{k \in B^p} \frac{1}{ITL_k} \times \frac{h_{b,j}}{h_{b,j}}.
\]  

(8.54)

**Algorithm 3** LGRA [3] and ELGRA [4]

1: Initialization:
2: \(A^s \leftarrow M^s\)
3: Measure the removal criterion \(RC_i\) \(\forall i \in M^s\),
   according to (8.54) for LGRA, or (8.55) for ELGRA
4: **Step 1 (Feasibility checking phase):**
5: Employ the centralized feasibility checking method and check if
   the target-SINRs vector of the set \(A^s \cup M^p\) is feasible or not;
6: if it is feasible
7: Admit all SUs in \(A^s\)
8: Terminate the algorithm
9: else
10: Go to Step 2
11: **Step 2 (Removal phase):**
12: \(i^* = \arg\max_{i \in A^s} RC_i\)
13: \(A^s \leftarrow A^s \setminus \{i^*\}\)
14: Go back to Step 1

A drawback of LGRA is that it does not consider different values of the target-SINRs in the admission of SUs. This has been resolved in [4], where an SSA with a fore-sighted removal criterion is proposed for addressing the JPAC problem for network containing one PBS and one SBS.

The SSA proposed in [4] is given by Algorithm 3 in which the users are sorted in an increasing order of effective link-gain ratios (ELGRs) \(\hat{\theta}_j \times \frac{h_j^{(p)}}{h_j^{(s)}}\), where \(h_j^{(p)}\) and \(h_j^{(s)}\) are the uplink path-gains between SU \(j\) and the PBS and the SBS, respectively, and \(\hat{\theta}_i = \frac{\gamma_i}{1+\gamma_i}\) is the effective target-SINR of user \(i\) (it is assumed in [4] that only one SBS and one PBS exist). The SUs with highest ELGR are gradually removed until the target-SINRs for all PU and admitted SUs become feasible. Thus, the removal criterion \(RC_i\) proposed in [4] is

\[
RC_i = \hat{\theta}_j \times \frac{h_j^{(p)}}{h_j^{(s)}}.
\]  

(8.55)
Different values of target-SINRs are not taken into account by LGRA [3], since it assumes the same value of the target-SINR for all SUs. On the other hand, although ELGRA [4] takes into account different values of target-SINRs for users, it does not consider different values of interference temperature at different PBSs (since it assumes a single PBS and SBS). Inspired by this observation, we can consider a new admission criterion as

$$ RC_i = \max_{k \in B^p} \frac{\hat{\theta}_j}{\mathrm{ITL}_k} \times \frac{h_{ki}}{h_{bi}}, \quad \forall i \in M $$

which takes into account the different values of interference temperature and target-SINRs.

### 8.6 Distributed JPAC Algorithms for CRNs

The centralized JPAC algorithms require full knowledge of channel-gain information, and their complexity is high. The existing centralized power and admission control algorithms may not be easily employed in practice. In this section, we introduce the distributed JPAC algorithm proposed in [7] for addressing the problem of distributed uplink power control in cellular cognitive radio networks.

The existing distributed interference management algorithms in conventional cellular wireless networks introduced in Chapters 6 and 7 do not guarantee that the total interference caused to PUs by SUs does not exceed a given threshold, which result in the outage of some PUs (i.e., some PUs are not supported with their required SINRs). However, these algorithms can be used by the SUs, provided that the interference caused by them to the PUs does not exceed a given threshold. In particular, if the SUs limit their transmit power levels so that the total interference caused to the PUs does not exceed a given threshold (which each primary receiver can broadcast to all SUs), each PU is able to reach its target-SINR by the TPC, and the SUs could minimize their outage ratio by employing an existing distributed power control algorithm. This is the main idea used in [7] to develop distributed uplink power control algorithms as discussed below.

#### 8.6.1 TPC with PU-Protection Algorithm (TPC-PP)

The following power update function is proposed in [7] to address the JPAC problem in CRNs:

$$ p_i(t + 1) = \begin{cases} 
\min \left\{ \bar{p}_i, \frac{\gamma_i}{\gamma_i(\mathbf{p}(t))} p_i(t) \right\}, & \text{for all } i \in M^p \\
\min \left\{ \bar{p}_i, \beta(t) p_i(t), \frac{\gamma_i}{\gamma_i(\mathbf{p}(t))} p_i(t) \right\}, & \text{for all } i \in M^s 
\end{cases} $$ \hspace{1cm} (8.56)

where \( \beta(t) = \min_{k \in B^p} \{ \frac{\bar{\gamma}_k}{\bar{\gamma}_k(\mathbf{p}(t))} \} \) in which \( \bar{\gamma}_k \) is given by (8.15). According to (8.56), the TPC with PU-protection algorithm (TPC-PP) is proposed as in Algorithm 4.
Algorithm 4 TPC with PU-protection (TPC-PP) [7]

1: Set $t = 1$, for each user $i \in \mathcal{M}$, initialize the transmit power randomly $p_i(t) \in (0, \overline{p}_i]$. 
2: repeat
3: for each PU $i \in \mathcal{M}^p$ do
4: Obtain the parameter $\frac{\overline{p}_i}{\gamma_i(p(t))}$ from his/her own PBS.
5: Update the power as $p_i(t + 1) = \min \left\{ \overline{p}_i, \frac{\overline{p}_i}{\gamma_i(p(t))} p_i(t) \right\}$.
6: end for
7: Each PBS $k \in \mathcal{B}^p$ multicasts the parameter $\frac{\overline{p}_k}{\psi_i(t)}$ to all SU $i \in \mathcal{M}^s$.
8: for each SU $i \in \mathcal{M}^s$ do
9: Obtain the parameter $\frac{\overline{p}_i}{\gamma_i(p(t))}$ from her own SBS.
10: Find $\beta(t) = \min_{k \in \mathcal{B}^p} \left\{ \frac{\overline{p}_k}{\psi_i(t)} \right\}$.
11: Update the power as $p_i(t + 1) = \min \left\{ \overline{p}_i, \beta(t) p_i(t), \frac{\overline{p}_i}{\gamma_i(p(t))} p_i(t) \right\}$.
12: end for
13: Update the power vector $p(t + 1) = [p_i(t + 1)]_{\forall i \in \mathcal{M}}$.
14: Update $t = t + 1$.
15: until $t = T_{\text{max}}$ or convergence to any fixed point.

In TPC-PP, each PU employs the TPC, and each SU employs the TPC as long as the total received power plus noise power at each PBS $k$, i.e., $\varphi_k(t)$ is less than the corresponding total-received power temperature $\overline{\varphi}_k$, otherwise it updates its transmit power proportional to $\frac{\overline{p}_k}{\psi_i(t)} p_i(t)$, which is equivalent to setting the transmit power $p_i(t + 1)$ to $\min_{k \in \mathcal{B}^p} \left\{ \frac{\overline{p}_k}{\psi_i(t)} \right\} p_i(t)$. As mentioned in Chapter 6, the TPC algorithm is indeed the same as the closed-loop power control algorithm, since the ratio of $\frac{\overline{p}_i}{\gamma_i(p(t))} p_i(t)$ in the TPC algorithm can be viewed as the commands of increasing or decreasing the power in closed-loop power control algorithm, corresponding to $\gamma_i(p(t)) < \overline{p}_i$ and $\gamma_i(p(t)) > \overline{p}_i$, respectively. Similarly, the term $\frac{\overline{p}_k}{\psi_i(t)} p_i(t)$ can also be viewed as a power-updating command issued by the PBS to SUs. The TPC-PP algorithm for the SUs can be interpreted as follows. Each SU receives two power-updating commands at each iteration: one is unicast from her own receiver, in terms of $\frac{\overline{p}_i}{\gamma_i(p(t))}$, and the other ones are multicast from each primary BS to all SUs, in terms of $\frac{\overline{p}_k}{\psi_i(t)}$.

Indeed, the TPC-PP algorithm uses a mixed-strategy for spectrum access. When there are many PUs with large target-SINR requirements associated with a PBS and/or the corresponding channel-gains are poor, the total received-power temperature for that primary BS is set to a very small value (according to (8.15)). This corresponds to spectrum overlay strategy. On the other hand, when the number and/or the target-SINR requirements of the PUs actively associated with each PBS is moderate and/or the channel-gains are good, the values of total received-power-temperature for the primary BSs can be non-zero. These values would indicate the amount of interference tolerability of the primary network in the spectrum underlay strategy. Therefore, by dynamically setting
the value of the total received-power temperature for each primary BS in an optimum manner using (8.15), a mixed strategy is adopted.

Analysis of the signaling overhead and fixed-point analysis of the TPC-PP are provided below.

**Analysis of Signaling Overhead**

In the TPC-PP algorithm, in addition to information that each user requires to update his/her transmit power using the TPC at each iteration, each SU needs to know the ratio of the total received-power temperature to the instantaneous total received power plus noise for each primary BS, i.e., \( \frac{\bar{\varphi}_k}{\varphi_k(t)} \), which is provided by the primary base stations. Thus, in comparison with TPC, the additional signaling overhead that TPC-PP incurs is that it requires each PBS \( k \) to iteratively provide the secondary users with the value of \( \frac{\bar{\varphi}_k}{\varphi_k(t)} \) (via a broadcast message in the control channel). Each primary BS \( k \) may broadcast the values of \( \varphi_k \) and \( \varphi_k(t) \), individually, or the ratios, i.e., \( \frac{\bar{\varphi}_k}{\varphi_k(t)} \), collectively, to the SUs.

Note that the value of \( \varphi_k \) needs to be updated by PBS \( k \) only when one of its associated PUs, who has the minimum value of \( \varphi_k(h_k) \) among all associated PUs, leaves or enters the system. However, in contrast, the value of \( \varphi_k(t) \) needs to be updated at each iteration. Since in practice each SU may cause severe interference only to its nearby primary BS, each PBS should inform only its nearby SUs of the values of \( \varphi_k \) and \( \varphi_k(t) \). Alternatively, each secondary BS can collect the values of \( \varphi_k \) and \( \varphi_k(t) \) from all the nearby PBSs and feed back only its minimum ratio, i.e., \( \min_{k \in \mathcal{B}_p} \{ \frac{\bar{\varphi}_k}{\varphi_k(t)} \} \) to its associated SUs. The feedback information can be quantized, and this quantized feedback information (bits) can be multicast in practical implementation. This is similar to CSI quantization and feedback commonly used in practice. With this implementation, we can control the feedback overhead and performance trade-off by choosing appropriate feedback bits. This feedback information can also be sent to secondary BSs by primary BSs via a possible wired network between them, and then SBSs send this feedback to their own SUs.

**Existence of Fixed-Point in TPC-PP and Its Properties**

It was shown in [7] that there exists at least one fixed-point for the TPC-PP power-update function, the and all of its fixed-points guarantee zero-outage for the PUs. It was shown in [7] that when the target-SINRs for all PUs and SUs are feasible, the fixed-point of TPC-PP is unique and is the same as that of the TPC. However, in an infeasible system, the TPC-PP power update-function may have multiple fixed-points. Given any fixed-point \( p^* \) of the power-update function of the TPC-PP, we have \( \gamma_i(p^*) \geq \bar{\gamma}_i \), \( \forall i \in \mathcal{M}^p \).

The key properties of TPC-PP algorithm are summarized as follows [7]:

1. The TPC-PP algorithm keeps the total received power plus noise at each PBS below the threshold given by (8.15), i.e., \( \varphi_k(p^*) \leq \bar{\varphi}_k \), so that all the PUs attain their target-SINRs. In other words, the TPC-PP guarantees that the existence of the SUs does not cause an outage to any PU. When the system is infeasible, all the PUs together with some SUs attain their target-SINRs, and the remaining SUs are unable to obtain their target-SINRs.
2. When the system is feasible, the fixed-point of TPC-PP is unique and the same as that of the TPC power update function at which all users attain their target-SINRs consuming minimum aggregate transmit power.

8.6.2 Improved TPC-PP (ITPC-PP)

Although all fixed-points of the power-update function TPC-PP result in a zero-outage ratio for PUs, the outage ratio for SUs is not necessarily the same for all fixed-points. Therefore, among all possible fixed-points of the TPC-PP algorithm, the fixed-points with minimal outage ratio of SUs are of most importance. The TPC-PP algorithm may converge to any of its fixed-points, depending on its initial transmit power vector. Now, an important question is how to lead the TPC-PP to converge to a desired fixed-point.

According to the TPC-PP power update function in (8.56), when

\[ \phi_l(\mathbf{p}(t)) < 1 \] for any iteration \( t \), where \( l = \arg \min_{k \in B^p} \left\{ \frac{\phi_l}{\phi_k(\mathbf{p}(t))} \right\} \), each SU, whether it has high or low path-gain with PBS \( l \), decreases its transmit power proportional to \( \frac{\phi_l}{\phi_k(\mathbf{p}(t))} \) in order to make the interference caused by SUs to PBSs lower than the threshold value. However, it is more efficient if the SUs that cause a lot of interference to PBS \( l \) (such SUs have high channel-gains to PBS \( l \)) decrease their transmit power levels more than the other SUs. Thus, if a SU causes a very low interference to PBS \( l \) (such an SU has a low channel-gain with PBS \( l \)), that user should not decrease its transmit power. This is because its power reduction may make it unsupported while not reducing the interference caused to the PBS \( l \) significantly. Accordingly, we propose the following improved TPC-PP (ITPC-PP) power update-function:

\[
p_i(t+1) = \begin{cases} 
\min \left\{ \frac{\gamma_i}{\gamma(\mathbf{p}(t))} p_i(t) \right\}, & \forall i \in U^p \\
\min \left\{ \frac{\gamma_i}{\gamma(\mathbf{p}(t))} p_i(t) \right\}, & \forall i \in U^s 
\end{cases} \tag{8.57}
\]

where

\[
\beta_i(t) = \begin{cases} 
\beta(t), & \text{if } \beta(t) \geq 1 \\
\beta(t) \left( 1 + |\phi_i - \phi_l(\mathbf{p}(t))| \right) \frac{\phi_i(\mathbf{p}(t)) - p_i(t) h_{li}}{h_{li}} & \text{if } \beta(t) < 1 
\end{cases} \tag{8.58}
\]

where \( \beta(t) = \min_{k \in B^p} \left\{ \frac{\phi_l}{\phi_k(\mathbf{p}(t))} \right\} \) and \( l = \arg \min_{k \in B^p} \left\{ \frac{\phi_l}{\phi_k(\mathbf{p}(t))} \right\} \).

From the viewpoint of signaling overhead, in ITPC-PP, in addition to the information required in TPC-PP, each SU needs to know (estimate) its channel gain with primary BS \( l \). In fact, the only difference between ITPC-PP and TPC-PP is that when \( \min_{k \in B^p} \left\{ \frac{\phi_l}{\phi_k(\mathbf{p}(t))} \right\} < 1 \), ITPC-PP causes each SU \( i \) to decrease its transmit power level proportional to \( \beta(t)(1 + |\phi_i - \phi_l(\mathbf{p}(t))|) \frac{\phi_i(\mathbf{p}(t)) - p_i(t) h_{li}}{h_{li}} \). On the other hand, in TPC-PP, all SUs decrease their transmit power proportional to \( \beta(t) \). If the effective interference experienced by a given SU \( i \) at primary BS \( l \) is lower than that of SU \( j \), i.e., if
\[
\frac{\phi_l(p(t))-p_i(t)h_{li}}{h_{li}} < \frac{\phi_l(p(t))-p_j(t)h_{lj}}{h_{lj}},
\]
the channel-gain of SU \(i\) toward primary BS \(l\) is better than that of SU \(j\), and consequently, SU \(i\) causes more interference toward primary BS \(l\) as compared to SU \(j\). In this case, if \(1 - \frac{\phi_l(p(t))}{\phi_l(p(t))} < 1\), SU \(i\) should reduce its transmit power more than SU \(j\). This is done by adjusting \(\beta_i(t)\) according to (8.58), because \(\frac{\phi_l(p(t))-p_i(t)h_{li}}{h_{li}} < \frac{\phi_l(p(t))-p_j(t)h_{lj}}{h_{lj}}\) results in \(\beta_i(t) < \beta_j(t)\), which causes SU \(i\) to decrease its power more in comparison with SU \(j\). Therefore, with the ITPC-PP, the SUs close to primary BS \(l\) reduce their transmit power more as compared to SUs far from primary BS \(l\).

It was shown in [7] that any fixed-point \(p^*\) for the TPC-PP power-update function (8.56) is also a fixed-point for the ITPC-PP power-update function (8.57). Note that although any fixed-point of TPC-PP is also a fixed-point of ITPC-PP, since ITPC-PP (when \(1 - \frac{\phi_l(p(t))}{\phi_l(p(t))} < 1\)) causes the SUs with high channel-gains toward PBS \(l\) to decrease their transmit power levels more aggressively, a fixed-point with improved outage ratio for SUs is eventually reached for ITPC-PP, while a zero-outage ratio for PUs is still guaranteed, as demonstrated via the simulation results presented in [7].

### 8.7 Exercises

**Exercise 8.1:** Consider a network where six PUs and six SUs are fixed and served by two PBSs and two secondary BSs, respectively, in an area of 1000 m × 1000 m, as illustrated in Figure 8.3. In this network, each primary (secondary) BS serves three primary (secondary) users. For simplicity, suppose that the target SINRs for all PUs and SUs is 0.10, and the uplink channel-gain from each user \(i\) to each BS \(k\) is given by \(0.1d_{k,i}^{-3}\), where \(d_{k,i}\) is the distance. The upper bound on the transmit power for all users is 1 Watt.

![Figure 8.3](https://example.com/figure83.png)
Joint Power and Admission Control in Cognitive Radio Networks

- Examine the feasibility of target-SINRs for all PUs and SUs. If it is infeasible, again distribute users in the cell so that their target-SINRs become feasible.
- Obtain the protection constraints for PUs in terms of total received-power temperature, total intra-cell interference temperature and total cognitive interference temperature levels at two primary BSs and depict their corresponding feasible regions (i.e., FRPR, FIIR, and FCIR).
- What is the optimal solution for the problem of maximizing the number of supported SUs subject to PUs’ protection constraint?
- Simulate two cases in which users iteratively update their transmit power levels using TPC or TPC-PP, and for each PBS, illustrate the total received power, total intra-cell interference, and total cognitive interference levels, versus iteration number, for TPC and TPC-PP. Examine if their corresponding equilibrium points (fixed-point) fall into the FRPR, FIIR, and FCIR.

Exercise 8.2: Consider the same network and system parameters explained in Exercise 8.1 and suppose that the target SINR for all PUs and SUs is 0.20.

- Examine the feasibility of target-SINRs for all PUs and SUs. If it is feasible, again distribute users in the cell so that the target-SINRs become infeasible.
- Obtain the protection constraints for PUs in terms of Total Received-Power Temperature, Total Intra-Cell Interference Temperature, and Total Cognitive Interference Temperature levels at two primary BSs, and depict their corresponding feasible regions (i.e., FRPR, FIIR, and FCIR).
- Apply the Centralized JPAC Algorithms of I-SMIRA, LGRA, and ELGRA, and compare their outage ratios for SUs with each other. In I-SMIRA, LGRA, and ELGRA, the value of the cognitive interference temperature limits at each PBS \( k \) is assumed to be fixed and independent of the ITL values for other primary receiving-points, whereas it is dynamic. For this reason, to simulate I-SMIRA, LGRA, and ELGRA, you may use the total received power temperature (which is a box-like) instead of the total cognitive interference temperature, i.e., replace ITL\(_ k \) with \( \varphi_k \) in I-SMIRA, LGRA, and ELGRA.
- Simulate two cases in which users iteratively update their transmit power levels using TPC or TPC-PP, and illustrate the total received power, total intra-cell interference, and total cognitive interference, for each PBS, versus iteration number, for TPC and TPC-PP. Examine if their corresponding equilibrium points (fixed-point) fall into the FRPR, FIIR, and FCIR.

Exercise 8.3: For comparing the performances of centralized JPAC algorithms (i.e., LGRA, ELGRA, and ISMIRA) with distributed ones (i.e., TPC-PP and ITPC-PP algorithms) and TPC, for different snapshots of users’ locations and for different values of target-SINRs, consider a primary network with 3 × 3 cells where each primary cell covers an area of 1000 m × 1000 m. Each primary (secondary) user is associated with only one primary (secondary) BS. Each PBS is located at the center of its corresponding cell and serves five PUs. Consider a macro cognitive radio network with four large cells each of which serves five SUs uniformly located at a radius of 1000 m around it within
8.7 Exercises

Figure 8.4 An example of network topology for a primary network with $3 \times 3$ cells with five PUs per primary cell which coexists with a secondary network with small cells (Figure 8.4(a)) and large cells (Figure 8.4(b)), for Exercise 8.3 and Exercise 8.4, respectively. The former includes three secondary BSs within each primary cell and five SUs per each secondary BS, and the latter includes four secondary BSs and five SUs per each secondary BS.

Exercise 8.3: A secondary radio network with $3 \times 3$ cells has five PUs per primary cell which coexists with a secondary network with small cells (e.g., cognitive femtocells with small transmission radius), for Exercise 8.3 and Exercise 8.4, respectively. The former includes three secondary BSs within each primary cell and five SUs per each secondary BS, and the latter includes four secondary BSs and five SUs per each secondary BS.

The uplink channel gain from each user $i$ to each BS $k$ is given by $0.1 d_{k,i}^{-3}$ where $d_{k,i}$ is the distance, and the upper bound on the transmit power for all users is 1 Watt. The target-SINRs are considered to be the same for all users, ranging from 0.02 to 0.16 with step size of 0.02. The initial transmit power for each user is uniformly set from the interval $[0, 1]$ for each snapshot.

For each target-SINR and each snapshot, apply the centralized and distributed JPAC algorithms, and average the corresponding values of outage ratios for the PUs and SUs for TPC, TPC-PP, ITPC-PP, LGRA, ELGRA, and I-SMIRA algorithms for 200 independent snapshots for a uniform distribution of BSs and users’ locations. Illustrate the average outage ratio versus target-SINR, for each algorithm, over independent snapshots of uniformly distributed locations of users. Compare the performances of the algorithms in terms of outage ratio for SUs and PUs.

Exercise 8.4: Repeat Exercise 8.3 (which was for a secondary radio network with large cells (e.g., cognitive macro-cells with larger transmission radius), for a secondary radio network with small cells (e.g., cognitive femtocells with small transmission radius). At each primary cell, three secondary BSs are uniformly located, each of which serves eight SUs uniformly located at a radius of 200 m around it. Thus, the entire network consists of nine PBSs, 45 PUs, 27 secondary BSs, and 135 SUs. An example of such a network setting is shown in Figure 8.4(b). Using the same parameters as in Exercise 8.3, illustrate the average outage ratio versus target-SINR, for TPC, TPC-PP, ITPC-PP, and
LGR, respectively, over 200 independent snapshots of uniformly distributed locations of users and secondary BSs, and compare and discuss the simulations results.

References


9 Cell Association in Cellular Networks

9.1 Introduction

One of the important issues in infrastructure-based multi-cell wireless networks is properly associating mobile user equipments (UEs) to the serving BSs. In the literature, this is usually referred to as user association, cell association, cell selection, or BS assignment. We will use the term “cell association” in this chapter. Obviously, in a wireless network with dense deployment of the BSs, the number of potential BSs with which a UE can be associated is increased. The network densification necessitates the need for designing optimal and/or distributed cell association (BSs assignment to UEs) schemes. This is because, if the UEs are not properly associated with BSs, it may result in reduced throughput, increased interference, inefficient energy consumption, and load imbalance, in uplink and/or downlink.

In Chapters 6 and 7, it was assumed that cell association is already performed (i.e., fixed BS assignment). In fact, in those chapters a fixed cell association is assumed, under which the power control and joint power and admission control problems were defined and addressed. In this chapter, we assume the BS assigned to each UE is not fixed and can be dynamically determined.

Cell association can be performed separately or jointly with other resource allocation schemes. For instance, cell association can be performed based on some metric such as the received (pilot) signal strength, or it can be performed jointly with power control or channel allocation. In this chapter, we first briefly present the system model introduced in Chapter 5 and make a little change in notations to make it suitable for studying the problem of dynamic cell association. Then the joint cell association and power control (CAPC) schemes are studied, followed by a review of the existing approaches for distributed cell association schemes (where cell association is performed separately and independently from the power control). Finally, open challenges and problems are discussed.

9.2 System Model and Notations

We consider the same system model presented in Chapter 5, which is briefly introduced again in this chapter. Consider a multi-cell wireless network with $K$ base stations (cells) and $M$ active UEs denoted by $\mathcal{K} = \{1, 2, ..., K\}$ and $\mathcal{M} = \{1, 2, ..., M\}$, respectively.
Let $p_i$ be the uplink transmit power of UE $i$. Noise is assumed to be additive white Gaussian whose power at the receiver of the base station $k \in K$ is denoted by $\sigma^2_k$. The uplink path-gains between UE $i \in M$ and BS $b_j \in K$ (i.e., the BS assigned to UE $j$) is represented by $h_{b_ji}$. A given uplink transmit power vector $\mathbf{p}$ is called feasible if $\mathbf{p} \in \mathcal{P}$ where $\mathcal{P} = \{\mathbf{p} | 0 \leq \mathbf{p} \leq \overline{\mathbf{p}}\}$, where $\overline{\mathbf{p}}$ is the upper-bound transmit power vector.

In this chapter, we assume that only one cell is associated to each UE, and the uplink and downlink cell-association is coupled, i.e., the same BS is assigned to each UE at both uplink and downlink. Let $b_i \in K$ denote the base station associated (assigned) to the UE $i$ at uplink and downlink. Given the BS $k \in K$, denote the set of its served UEs at uplink by $C_k = \{i \in M | b_i = k\}$. The BS association vector is denoted by $\mathbf{b} = [b_1, b_2, ..., b_M]^T$. It is evident that there exist $K^M$ number of different BS assignment vectors. Let us number the different BS assignment vectors as $l = 1, 2, \ldots, K^M$, and denote the $l$-th BS assignment vector by $\mathbf{b}(l) = [b_1(l), b_2(l), ..., b_M(l)]^T$.

In this chapter, since the BS assignment and transmit power vectors are both considered variable, the SINR of a UE $i$ at the BS $k \in K$ is denoted by $\gamma_i(\mathbf{p}, k)$ as

$$\gamma_i(\mathbf{p}, k) = \frac{h_{ki}p_i}{\sum_{j \neq i} h_{kj}p_j + \sigma^2_k}. \quad (9.1)$$

The uplink interference experienced by UE $i$ at BS $k \in K$ is $I_i(\mathbf{p}, k) = \sum_{j \neq i} h_{kj}p_j + \sigma^2_k$.

The SINR of a UE $i$ at its assigned BS $b_i \in K$ is given by $\gamma_i(\mathbf{p}, b_i) = \frac{h_{bi}p_i}{\sum_{j \neq i} h_{bj}p_j + \sigma^2_{b_i}}$.

**Definition 48** A given uplink SINR vector is feasible under the BS assignment $\mathbf{b}(l)$ if a feasible uplink transmit power vector exists that corresponds to the SINR vector.

As derived in Chapter 5, given an uplink SINR vector, the corresponding uplink transmit power vector under the BS assignment vector $\mathbf{b}(l)$ is computed as

$$\mathbf{p}(l) = (\mathbf{I} - \mathbf{D}(\gamma) \mathbf{G}(l))^{-1} \mathbf{D}(\gamma) \eta(l) \quad (9.2)$$

where $\mathbf{I}$ is a $M \times M$ identity matrix, $\mathbf{D}(\gamma)$ denotes a diagonal matrix whose diagonal elements are the corresponding components of the SINR vector $\gamma$, the $(i, j)$ component of $\mathbf{G}(l)_{M \times M}$ is

$$G_{ij}(l) = \begin{cases} \frac{h_{b_i(l)}j}{h_{b_j(l)i}}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \quad (9.3)$$

and the $(i)$ component of $\eta(l)$ is $\eta_i(l) = \frac{\sigma^2_{b_i}}{h_{b_i(l)i}}$.

The following theorem states under what condition a given SINR vector is achievable by a positive transmit power vector [2].

---

1 These assumptions can be relaxed and generalized, as will be discussed in Section 9.5.
Theorem 65 For a given uplink SINR vector $\gamma \geq 0$ and under a given BS assignment $b(l)$, there exists a corresponding positive uplink power vector $p \geq 0$ if $\rho(G(l)D(\gamma)) < 1$.2

To examine the feasibility of a given uplink SINR vector under the base-station assignment $b(l)$, the corresponding positive transmit power vector corresponding to the SINR vector is obtained (if it exists) by using (9.2), and the obtained transmit power vector is checked if it is feasible or not. This means that a given uplink SINR vector $\gamma$ is feasible under the base station assignment $b(l)$ if $p(l) \in P$, where $p(l) = (I - D(\gamma)G(l))^{-1}D(\gamma)\eta(l)$.

9.3 Distributed Joint Cell Association and Power Control

In this section, we study the joint uplink cell-association and power control problem of minimizing aggregate power subject to the target-SINR constraint, which is formally stated as

$$\min \sum_{i \in M} p_i$$

s.t. $\gamma_i(p, b_i) \geq \hat{\gamma}_i, \quad \forall i \in M$ \hspace{1cm} (9.4)

variables $p \geq 0$

$b \in K^M$.

It was shown in [2] that among all feasible uplink transmit power vectors corresponding to a target-SINR vector, there exists one that is component-wise less than the other ones, and it is the optimal transmit power vector solution to this problem, as stated in the following theorem.

Theorem 66 Among all feasible uplink transmit power vectors $p(l)$ (for $l \in \{1, 2, \ldots, K^M\}$) corresponding to a given target-SINR vector, there exists an uplink transmit power vector $p^*$ so that $p^* \leq p(l)$. Such a transmit power vector (i.e., $p^*$) and its corresponding base-station assignment vector solves the joint uplink cell association and power control problem (9.4).

As an example, let us consider two UEs and two BSs denoted by $M = \{1, 2\}$ and $K = \{1, 2\}$, respectively, in a two-cell network. In this network, there exist four possible base-assignment vectors $b = [1, 1], [1, 2], [2, 1], [2, 2]$. The relation between the transmit power of each UE as a function of other UE’s transmit power, depending on its assigned BS, is obtained as

$$\begin{cases} 
\text{if } b_1 = 1 \Rightarrow \gamma_1 = \frac{p_1h_{11}}{p_2h_{12} + \sigma_1^2} \Rightarrow p_1 = \frac{\gamma_1}{h_{11}}(p_2h_{12} + \sigma_1^2) \\
\text{if } b_1 = 2 \Rightarrow \gamma_1 = \frac{p_1h_{21}}{p_2h_{22} + \sigma_2^2} \Rightarrow p_1 = \frac{\gamma_1}{h_{21}}(p_2h_{22} + \sigma_2^2)
\end{cases}$$

(9.5)

$^2$ $\rho(A)$ denotes the spectral radius (maximum of the absolute value of eigenvalues) of matrix $A$. 

---

2. $\rho(A)$ denotes the spectral radius (maximum of the absolute value of eigenvalues) of matrix $A$. 

---
and
\[
\begin{cases}
\text{if } b_2 = 1 & \Rightarrow \gamma_2 = \frac{p_2 h_{12}}{p_1 h_{11} + \sigma_1^2} \Rightarrow p_2 = \frac{\gamma_2}{h_{12}} (p_1 h_{11} + \sigma_1^2) \\
\text{if } b_2 = 2 & \Rightarrow \gamma_2 = \frac{p_2 h_{22}}{p_1 h_{21} + \sigma_2^2} \Rightarrow p_2 = \frac{\gamma_2}{h_{22}} (p_1 h_{21} + \sigma_2^2)
\end{cases}
\tag{9.6}
\]

for two UEs, respectively.

Figure 9.1 illustrates four lines, two lines for \(p_1\) versus \(p_2\) (for two cases of \(b_1 = 1\) and \(b_1 = 2\) characterized by (9.5)) and two lines for \(p_2\) versus \(p_1\) (for two cases of \(b_2 = 1\) and \(b_2 = 2\), characterized by (9.6)). Depending on the path-gains and noise level, the SINR vector may be feasible or infeasible under a given BS assignment vector. If one of these two lines intersects with one of the other two lines, i.e., if a line of \(p_1\) versus \(p_2\) intersects a line of \(p_2\) versus \(p_1\), then the intersection point is a feasible transmit power vector corresponding to the SINR vector, provided that it is in the feasible region. As can be seen in Figure 9.1, among all such feasible transmit power vectors, there exists one that is component-wise less than the other feasible ones for which this transmit power vector and its corresponding BS assignment are the optimal solution to the joint uplink cell association and power control problem given in (9.4).

Now, the question is how to obtain the optimal solution power and BS assignment vector, characterized in Theorem 66, in a distributed manner. To address the joint uplink cell association and power control problem (9.4) in a distributed manner, a joint uplink cell association and TPC (UCA-TPC) algorithm have been proposed in [1] and [2] wherein at each step the UE \(i\) is assigned to the BS, which results in the minimum effective interference as compared to other BSs. In doing so, the power update function is proposed as
\[
p_i(t + 1) = \min \left\{ \bar{p}_i, \min_{k \in \mathcal{K}} \left\{ \frac{I(p(t), k)}{h_{ki}} \right\} \right\}, \quad \forall i \in \mathcal{M} \tag{9.7}
\]
where \( I_i(p(t), k) = \sum_{j \neq i} p_j(t)h_{kj} + \sigma_k^2 \), is the instantaneous total interference caused to the UE \( i \) at the BS \( k \), and the assignment update function is given by

\[
b_i(t + 1) = \arg \min_{k \in K} \left\{ \frac{I_i(p(t), k)}{h_{ki}} \right\}, \quad \forall i \in \mathcal{M}.
\] (9.8)

**Algorithm 5** UCA-TPC Algorithm

\begin{itemize}
  \item[1:] **Initialization:**
  \item Let \( p_0 \) be the initial transmit power vector;
  \item \( t \leftarrow 1; \)
  \item **Power control algorithm:**
  \item Each UE \( i \) updates its transmit power as
  \[
p_i(t + 1) = \min\{\bar{p}_i, \min_{k \in K} \left\{ \frac{\hat{\gamma}_i I_i(p(t), k)}{h_{ki}} \right\} \}
  \]
  \item **Cell association algorithm:**
  \item Each UE \( i \) is associated with a BS as
  \[
b_i(t + 1) = \arg \min_{k \in K} \left\{ \frac{I_i(p(t), k)}{h_{ki}} \right\}
  \]
  \item If \( |p(t + 1) - p(t)| \leq \epsilon \)
  \item The current association is optimal and terminate;
  \item \( t \leftarrow t + 1; \) Go to power control algorithm.
\end{itemize}

One can show that the power update function given by (9.7) is a standard type-I function and thus it has a unique fixed-point in a feasible system. In a similar way for convergence analysis of the TPC, for any given initial transmit power vector, the distributed power control algorithm converges to its unique fixed-point (for both synchronous and asynchronous power updates), which is the optimal solution to the optimization problem (9.4) [1], as stated in the following theorem shown in [1] and [2].

**Theorem 67** Assumption the feasibility of the target-SINR vector \( \hat{\gamma} \), there exists a fixed-point \((p^*, b^*)\) for the UCA-TPC cell and power updating function, in which \( p^* \) is unique. For any given initial transmit power vector, the distributed power control algorithm (9.7) converges to its unique fixed point \( p^* \) (for both synchronous and asynchronous power updates), which is the optimal solution to the optimization problem (9.4).

Note that while the equilibrium transmit power vector \( p^* \) is unique, its corresponding base-station assignment vector may not be unique. In fact, given the unique fixed-point \( p^* \), \( \arg \min_{k \in K} \left\{ \frac{I_i(p^*, k)}{h_{ki}} \right\} \) may not be unique for some \( i \). From the signaling overhead point of view, in the power-update function given by (9.7), at each time \( t \), each UE needs to be provided with the instantaneous values of \( I_i(p(t), k) \) and \( h_{ki} \), or their ratio, i.e., \( \frac{I_i(p(t), k)}{h_{ki}} \), by each BS \( k \in K \), even if it is not currently assigned to the UE. In
practice, each UE requires only that we consider its nearby BSs, because the far BSs result in high effective interference due to significant attenuation of their transmitted signals.

9.4 Distributed Cell Association Schemes in Wireless Networks

The signaling overhead of the joint cell association and TPC algorithm introduced in the previous section is high, since it requires the instantaneous values of effective interference for each UE to be unicast by any of its nearby BS iteratively to the UE, even if the UE is not currently assigned to that BS. Alternatively, to avoid such a large signaling overhead, the cell association scheme can be performed independently from the power control scheme, and based on the reference (pilot) signal broadcast by each BS to all UEs, as will be discussed in the next two subsections, for homogeneous and heterogeneous cellular networks, respectively.

9.4.1 Reference Signal Received Power (RSRP)-Based Cell Association Scheme

In the reference signal received power (RSRP)-based cell association scheme, each BS iteratively broadcasts a reference (pilot) signal. Let $p_{k}^{RS}$ denote the power level of reference signal sent by the BS $k$ in a downlink transmission. Each UE $i$ measures the received power level from the reference signal transmitted by any of its nearby BSs $k \in K$, which is called the reference signal received power (RSRP). Let us denote the RSRP measured by UE $i$ from the BS $k$ by $\varphi_{ki}$, which is given by

$$\varphi_{ki} = p_{k}^{RS} \times \tilde{h}_{ki}$$

(9.9)

where $\tilde{h}_{ki}$ denotes the downlink path-gain from the BS $k$ toward UE $i$. A UE is associated with the BS, whose reference (pilot) signal is received with the largest average strength. In other words, in a RSRP-based cell-association scheme, the UE $i$ selects its cell with maximum RSRP as follows:

$$b_i = \arg \max_{k \in K} \varphi_{ki}.$$  

(9.10)

Traditionally, cell association is performed based on downlink attributes, as is the case in the RSRP-based cell association given by (9.10), and the same associated cell at downlink is also associated at uplink for each UE. This is due to two reasons as explained in what follows. First, designing distributed downlink-based cell association is easier, as compared to uplink-based cell association, because in downlink, all BSs can transmit (broadcast) their reference signals, which enable each UE to measure some attributes of the downlink channel with respect to its corresponding nearby BSs, and decide the appropriate BS for association in a distributed manner. This is very simple and has low complexity in contrast to the joint uplink cell association and power control, which requires the effective interference experienced by each UE to be individually
measured at each BS and then unicast to each UE (as performed in UCA-TPC). Furthermore, if an RSRP-based cell association scheme is used at uplink (i.e., each UE sends a reference signal to all nearby BSs), all nearby BSs should measure the uplink channel for a given UE and then share their measurement information to determine which of them is appropriate for serving that UE, which increases signaling overhead among BSs.

The RSRP-based cell association scheme is mostly appropriate for homogeneous wireless cellular networks, because the homogeneous BSs transmit with the same power level and thus have the same converge area. Thus, in homogeneous (single-tier) networks with uniform traffic, the RSRP-based cell association scheme results in load balance. On the other hand in a heterogeneous wireless cellular network, the RSRP-based cell-association scheme does not work well and would result in load imbalance, as discussed in the next subsection.

9.4.2 Biasing-Based Cell Range Expansion (CRE) in Wireless Networks with Heterogeneous BSs

To satisfy the ever-increasing demand for mobile broadband communications, BS deployments have become increasingly dense and heterogeneous. The network wherein the macro-BSs are overlaid by some low-power BSs such as pico- and femto-BSs, is called the multi-tier or heterogeneous network (HetNet). In homogeneous (single-tier) networks with uniform traffic, the RSRP-based cell-association scheme may maximize the system throughput. However, due to varying transmit power levels of different BSs in the downlink of multi-tier networks, such cell association policies may create a severe traffic load imbalance. This is because, the reference signal sent by the macro-BSs are stronger than that of pico- and femto-BSs, and thus BSs with higher RSRP are typically the ones with large transmit power capabilities. This results in the downlink coverage of the macro-BSs becoming very large and more UEs associating to them as compared to the number of UEs associated to low-power BSs. This would lead to overloading of high-power tiers while leaving low-power tiers underutilized.

To avoid such a load imbalance, a UE should be associated with a pico- or femto-BS even though the RSRP from the closest macro-BS in downlink is higher. This idea, which is called the Cell Range Expansion (CRE) [4], has emerged as a remedy to the problem of load imbalance in heterogeneous networks. CRE aims to increase the downlink coverage footprint of low-power BSs and is performed by adding a positive bias to the RSRP strengths or other RSRP-dependent criteria, as discussed in what follows. Such BSs are referred to as Biased BSs.

Let \( \mathcal{K} = \mathcal{K}_{\text{Small}} \cup \mathcal{K}_{\text{Macro}} \), where \( \mathcal{K}_{\text{Small}} \) and \( \mathcal{K}_{\text{Macro}} \) denote the set of low-power BSs and macro-BSs, respectively. Let \( \alpha_k^{\text{RSRP}} \) be the offset value added to the RSRP of the low-power BS \( k \in \mathcal{K}_{\text{Small}} \). The biased RSRP experienced by UE \( i \) corresponding to the reference signal transmitted by the BS \( k \) is defined as

\[
\varphi'_{ki} = \begin{cases} 
\varphi_{ki}, & \text{if } k \in \mathcal{K}_{\text{Macro}} \\
\varphi_{ki} + \alpha_k^{\text{RSRP}}, & \text{if } k \in \mathcal{K}_{\text{Small}}
\end{cases}
\]
where $\varphi_{ki}$ is the RSRP from the BS $k$ to the UE $i$ given by (9.9). In a biased RSRP-based cell association scheme, the UE $i$ determines its serving BS $b_i$ according to

$$b_i = \arg \max_{k \in K} (\varphi'_{ki}).$$  \hspace{1cm} (9.12)

This biasing allows more UEs to associate with low-power or biased BSs and thereby achieve a better cell load balancing.

Instead of biasing the RSRP criterion, one may use a different criterion for CRE in which the positive bias is added to the average received SINRs and the Reference Signal Received Quality (RSRQ) criteria, as defined and explained below. The average received SINR and the average RSRQ experienced by UE $i$ corresponding to the reference signal transmitted by the BS $k$ are defined as

$$\bar{\gamma}_{ki}(p_{RS}, k) = \frac{p_{RS}h_{ki}}{\sum_{j \in K, j \neq k} p_{RS}h_{ji} + \tilde{\sigma}_i^2}$$ \hspace{1cm} (9.13)

and

$$\tilde{\theta}_{ki} = \frac{\bar{\gamma}_{ki}(p_{RS}, k)}{\bar{\gamma}_{ki}(p_{RS}, k) + 1}$$ \hspace{1cm} (9.14)

respectively, where $\tilde{\sigma}_i^2$ is the noise at the receiver of UE $i \in \mathcal{M}$.

In biased SINR-based and RSRQ-based cell association schemes, a given offset value is added to the average received SINR and RSRQ for low-power BSs, respectively. Let us denote the offset value added to the average received SINR and RSRQ corresponding to the reference signal transmitted by each low-power BS $k \in \mathcal{K}_{Small}$ by $\alpha_{k}^{SINR}$ and $\alpha_{k}^{RSRQ}$, respectively. Given $\alpha_{k}^{SINR}$ and $\alpha_{k}^{RSRQ}$ for all $k \in \mathcal{K}_{Small}$, the biased received SINR and biased RSRQ are defined as

$$\bar{\gamma}'_{ki}(p_{RS}, k) = \begin{cases} \bar{\gamma}_{ki}(p_{RS}, k), & \text{if } k \in \mathcal{K}_{Macro} \\ \bar{\gamma}_{ki}(p_{RS}, k) + \alpha_{k}^{SINR}, & \text{if } k \in \mathcal{K}_{Small} \end{cases}$$ \hspace{1cm} (9.15)

and

$$\tilde{\theta}'_{ki} = \begin{cases} \tilde{\theta}_{ki}, & \text{if } k \in \mathcal{K}_{Macro} \\ \tilde{\theta}_{ki} + \alpha_{k}^{RSRQ}, & \text{if } k \in \mathcal{K}_{Small} \end{cases}$$ \hspace{1cm} (9.16)

respectively.

In biased SINR-based and RSRQ-based cell association schemes, the UE $i$ determines its serving BS $b_i$ according to

$$b_i = \arg \max_{k \in B} (\bar{\gamma}'_{ki})$$ \hspace{1cm} (9.17)

and

$$b_i = \arg \max_{k \in K} (\tilde{\theta}'_{ki})$$ \hspace{1cm} (9.18)

respectively. Depending on the values of offsets $\alpha_{k}^{RS}$, $\alpha_{k}^{SINR}$, and $\alpha_{k}^{RSRQ}$, the BS assigned to UE $i$ by (9.12), (9.17), and (9.18) may be different. However, it is worth noting that
if no bias is used (i.e., when there is no low-power BS (i.e., \( \mathcal{K}_{\text{small}} = \emptyset \)) as is the case in homogeneous wireless networks, or the offset values are zero, i.e., \( c_k^{\text{RS}} = c_k^{\text{SINR}} = c_k^{\text{RSRQ}} = 0 \)), the RSRP-based, SINR-based and RSRQ-based cell association schemes are identical and result in the same cell association, i.e., the assigned BS given by (9.12), (9.17), and (9.18) is the same.

9.5 Open Research Issues

There are several open research issues with respect to cell association schemes that need further investigation:

- Any of the problems and schemes studied in Chapters 6 and 7 can be extended to joint cell-association and power control ones, in which, in addition to the power level for each UE, its assigned BS needs to be adjusted. In particular, the problems of minimizing outage ratio or maximizing system throughput over power control and BS assignment have not been addressed in the literature and are still open problems. In contrast to the problem of minimizing aggregate transmit power subject to a minimum target-SINR, which is optimally addressed by the strategy of combining cell association based on minimum effective-interference, and TPC (as the distributed power control scheme corresponding to the objective), presented in Section 9.3, the strategy of combining the corresponding distributed power control scheme (i.e., gradual removal and opportunistic power control [OPC] algorithms for minimizing outage ratio or maximizing system throughput, respectively) and the cell association based on minimum effective-interference is not necessarily the optimal/suboptimal solution to the stated open problems, as discussed below for OPC and dynamic target-SINR tracking power control (DTPC) schemes, respectively.

  - For instance, joint minimum effective-interference-based cell association and OPC do not address the problem of maximizing system throughput, over power control and BS assignment, because it causes all UEs, whether with a poor or good channel condition, to be associated to the cell that provides them with the minimum effective-interference, which results in higher transmit power for all UEs, in comparison to the case that UEs are associated to the cell with maximum effective-interference, since they employ the OPC. However, the system throughput is improved when UEs with good channel conditions increase their transmit power, but it is degraded when UEs with poor channel conditions increase their transmit power, which is the case with joint minimum effective-interference-based cell association and OPC. Thus, since by employing the OPC the minimum effective-interference-based cell association results in increased transmit power by all UEs whether with poor or good channel condition, it is not appropriate for the problem of maximizing the system throughput, which requires low and high transmit power by UEs with poor and good channel condition, respectively. Consequently, for addressing the open problem of maximizing system throughput, over power control and
9.5 Open Research Issues

BS assignment, the criteria for cell association for UEs with good or poor channel condition should be different; that could be minimum or maximum effective-interference for the former and the latter, respectively. However, this would be still an open challenge that of how UEs with good channel and poor channel condition could be identified in a distributed manner.

- As an another example, in the DTPC algorithm presented in [11] and explained in Chapter 6, for a given UE, when the effective interference (the ratio of the received interference to that UE’s path-gain) is less than a given threshold, that UE opportunistically sets its target-SINR (which is a decreasing function of the effective interference) to a value higher than its minimum acceptable target-SIR; otherwise it keeps its target-SIR fixed at its minimum acceptable level. The DTPC can be extended to a joint power control and BS assignment scheme in which each UE chooses a BS that brings the minimum effective interference for that UE. Upon assigning a BS to UEs at each iteration, UEs set their power level according to DTPC. The same approach can be taken for extending the existing gradual removal scheme [12] to a joint gradual UE removal and cell-association scheme to minimize the outage ratio over power control and BS assignment.

- In this chapter, it has been assumed that only one cell is associated to each UE, and the uplink and downlink cell-association is coupled, i.e., the same BS is assigned to each UE at both uplink and downlink. However, these assumptions can be relaxed and generalized. For instance, simultaneous connections to multiple BSs for uplink or downlink, and/or different cell association at uplink and downlink, can be also considered. Such a general scheme would increase the degrees of freedom that can be exploited to further improve network capacity and balance the load among different BSs in different cells, as discussed below:

  - Existing criteria for cell association could be generalized to support simultaneous connections to multiple BSs. For instance, the minimum effective-interference-based (RSRP-based) cell association could be generalized so that when the difference among effective-interference (RSRP) levels between a given UE and some BSs that offer that UE the lowest effective-interference (largest RSRP) levels is not large, that UE can simultaneously connect to a number of those BSs. Other existing criteria for cell association could be similarly generalized.

  - In heterogeneous wireless networks, unlike downlink where the transmit power levels of BSs are different, the transmission power in the uplink depends on the UE’s battery power, which does not vary significantly from UE to UE. Therefore, the problems of coverage and traffic load imbalance may not exist in the uplink. This leads to considerable asymmetries between the uplink and downlink cell association policies in heterogeneous wireless networks [9]. Consequently, the optimal solutions for downlink cell association problems may not be optimal for the uplink and vice versa. To deal with this issue of asymmetry, it is therefore necessary that (1) for the coupled cell-association, where the same BS is associated to each UE at both uplink and downlink, a joint optimization framework is developed that can provide near-optimal, if not optimal, solutions for both uplink and downlink, and (2) for decoupled cell association, where the BS associated at uplink and downlink can
be different, individual optional/suboptimal uplink and downlink cell association are developed. For the former, instead of the traditional approach of coupled cell association in which cell association is based on either uplink or downlink parameters, a joint optimization of resources in both uplink and downlink would give better system performance. Decoupled cell association comes with several practical challenges, including among others the need for coordination between BSs for handling both uplink and downlink data [13–15].

9.6 Exercises

Exercise 9.1: Consider a wireless network including a single BS and two users. The maximum transmit power for each user is 1 Watt. The path-gain between user 1 and BS is 0.2, and path-gain between user 2 and BS is 0.3. Assume that the noise level is 0.1 Watt.

- If system throughput is denoted by $T$ and defined as the aggregate SINR of two users, i.e., $T = \sum_{i=1}^{2} \gamma_i(p)$, find the transmit power levels for each user so that the system throughput is maximized.
- Suppose that the target SINRs for users 1 and 2 are $\hat{\gamma}_1 = 1$ and $\hat{\gamma}_2 = 1.5$, respectively. Show that these values of SINR for two users are not simultaneously achievable (i.e., show that the system is infeasible). What is the minimum outage ratio? Find the power levels for two users corresponding to the minimum outage ratio.
- In order to make the system explained above feasible, a new BS is installed. For ease of reference, suppose that the previous BS and new BS are indexed as BS no. 1 and BS no. 2, respectively. The path-gains toward BS no. 2 for users 1 and 2 are 0.4 and 0.1, respectively (the path-gains toward BS no. 1 for users 1 and 2 are still 0.2 and 0.3, respectively). The noise power at each of the two BSs is equal to 0.1 Watt. Assign the BSs to two users so that the target-SINRs of $\hat{\gamma}_1 = 1$ and $\hat{\gamma}_2 = 1.5$ are feasible under the assigned BSs. Also, find the corresponding transmit power levels for users 1 and 2 for obtaining their target-SINRs.
- For the three problems mentioned above, their centralized solutions had been asked. Now, for each of the problems above, among existing distributed algorithms for power control or for joint power control and BS assignment, suggest an appropriate distributed algorithm(s).

Exercise 9.2: For comparing the performances of different cell association schemes, for different snapshots of users' locations, consider a two-tier network with 3 × 3 cells where each macro-cell covers an area of 1000 m × 1000 m. Each user is associated with only one BS. Each macro-BS is located at the center of its corresponding cell. At each
macro-cell, three femto-BSs are uniformly located. Thus, the entire network consists of nine macro-BSs, and 27 femto-BSs. For simplicity, suppose that the uplink channel-gain from each user \( i \) to each BS \( k \) is given by \( 0.1 d_{k,i}^{-3} \) where \( d_{k,i} \) is the distance. The upper bound on the transmit power for all users is 1 Watt.

- Apply the RSRP-based and SINR-based cell association schemes, and compare their results with the biased RSRP, biased SINR-based, and RSRQ-based cell association schemes, for different snapshots of users’ locations and for different offset values and power level of reference signal sent by the femto-BS and macro-BS. For each case, depict the different cell associations obtained by different schemes and different parameters.
- Given each of cell associations obtained by either RSRP-based, SINR-based, biased RSRP, biased SINR-based, and RSRQ-based cell association schemes, employ the TPC algorithm for updating the transmit power by all users, assuming a target SINR of 0.15 for all users. For each case, check if the system is feasible, given the corresponding cell association.

**Exercise 9.3:** Considering the same network and system parameters explained in Exercise 9.2, apply the UCA-TPC algorithm, and compare the results with the ones obtained in Exercise 9.2.

**References**


Part IV

Link Layer Resource Allocation in Wireless Networks
10 Sub-Carrier/Sub-Channel Allocation in OFDMA Networks

10.1 Introduction

OFDM has become the multicarrier transmission technique of choice in broadband transmission over wireless channels. This has been adopted for several wireless access technologies including IEEE 802.11a/g, IEEE 802.16, Digital Video Broadcasting (DVB), and Digital Audio Broadcasting (DAB). What makes OFDM an interesting choice for next generation broadband wireless transmission is its ability in combating frequency selective fading. Instead of transmitting digital symbols sequentially over a single wideband channel, OFDM divides the channel into many narrowband sub-channels or sub-carriers\(^1\) and then simultaneously transmits digital symbols in parallel over these sub-carriers. A transmitted digital symbol over a sub-carrier then experiences a flat fading channel.

In a multiuser scenario, a particular sub-carrier at a particular instant may appear differently, in terms of fading characteristics, to different users due to the varying nature of wireless channels and users’ locations. This provides an opportunity to assign certain sub-carriers to users who can utilize them best at that particular moment. The resulting mechanism of such sub-carrier allocation can be viewed as an OFDM-based multiple access scheme called Orthogonal Frequency Division Multiple Access (OFDMA) in which each user is assigned a subset of sub-carriers for exclusive use at any given time. In assigning sub-carriers to users, other resources such as power and modulation format also can be allocated to each assigned sub-carrier. As the number of users increases, there will be more freedom in allocating sub-carriers, transmission power, and modulation format per sub-carrier to different users.\(^2\) To tailor the OFDMA system to users’ needs in terms of desired data rate, maximum available transmission power, or utility, it is natural then to devise a resource allocation scheme that adapts to users’ varying channel conditions on a temporal basis. Adaptive radio resource allocation is thus essential to the performance of OFDMA systems. In recent years, many researchers have tried to explore this idea of adaptively assigning radio resources to users in OFDMA systems in order to optimize a certain metric of interest such as data rate, transmission power, and utility subject to certain constraints.

\(^1\) Although in some context a sub-channel may be defined as a group of sub-carriers, we do not differentiate the two terms in here to avoid any confusion that may arise.

\(^2\) This effect is referred to as multiuser diversity.
In this chapter, we provide an overview of the research in this area and identify some open research issues. In Section 10.2, we discuss the basic principles of OFDM as well as existing OFDM-based multiple access schemes. In Section 10.3, we provide an overview of the major approaches to adaptive radio resource allocation in OFDMA systems. Section 10.4 highlights some interesting open research issues. This chapter does not discuss the resource allocation issues for cooperative (e.g., relay-enhanced), MIMO, and distributed MIMO-based OFDMA cellular networks. A review of the major issues related to radio resource allocation in these types of networks can be found in [2].

10.2 OFDM-Based Multiple Access

Before we proceed any further, let us review the basic principles of OFDM (as discussed in Chapter 1) and multiple access schemes based on OFDM. OFDM uses the concepts of frequency division multiplexing to allow multiple data streams to be transmitted over a single radio channel [3–5]. Whereas typical single-carrier modulation systems usually suffer from intersymbol interference (ISI) as data rate increases (i.e., the symbol duration decreases), in a frequency-selective fading channel, OFDM is quite robust to ISI due to its parallelization technique. That is, instead of transmitting digital symbols sequentially over a single wideband channel, OFDM splits this channel into many narrowband sub-channels and transmits the digital symbols of longer duration in parallel over these sub-channels. OFDM symbols then undergo only flat fading in each sub-channel, avoid the ISI problem, and still maintain the high overall data rate.

More formally, let $B$ denote the total available bandwidth of the channel. An OFDM system then divides this bandwidth into $N$ parallel and ideally independent sub-channels, also called sub-carriers. Each sub-channel is of bandwidth $B/N$. Assuming a point-to-point transmission, an incoming stream of transmission data bits is first serial-to-parallel converted into $N$ parallel substreams. Each sub-carrier is then modulated by a group of bits in each of these substreams, possibly, with different modulation types. This operation results in $N$ parallel frequency-domain symbols. After passing these $N$ symbols through the inverse Fast Fourier Transform (IFFT) operation, an $N$-point time-domain representation of an OFDM symbol is obtained. A guard period, also called cyclic prefix, is then added to the OFDM symbol to further prevent ISI. At the receiving end, reverse operations occur. The guard period is first removed. The OFDM symbol then undergoes the Fast Fourier Transform (FFT) operation, which transforms the time domain representation of the transmitted OFDM symbol into the frequency domain representation of the digital symbol for each sub-carrier. Each sub-carrier is then demodulated accordingly to recover the group of bits for each substream. For more details on the principles of OFDM, interested readers are referred to [3–6] and [8].

---

3 This chapter is primarily based on [1].
4 In the discussion that follows, we do not differentiate between the terms sub-channel and sub-carrier unless an ambiguity exists.
In multiuser OFDM-based systems, one of the following three channelization schemes can be implemented to accommodate multiple access: (1) OFDM-TDMA (OFDM-TD), (2) OFDM-CDMA (MC-CDMA), and (3) OFDM-FDMA (OFDMA) [24]. We briefly describe each of them as follows:

1. **OFDM-TD**: OFDM-TD offers a straightforward means for multiple access by assigning all the sub-carriers to a particular user for the duration of allocated time slots. Multiple users then share the medium by means of time-division multiplexing. Different data rates are supported by allocating a different number of time slots to users. One drawback of OFDM-TD, however, is access delay, especially when the number of users in the system grows. Multiuser diversity is thus not fully exploited in OFDM-TD.

2. **MC-CDMA**: In MC-CDMA, at any given time all sub-carriers may be shared simultaneously by multiple users who are assigned with orthogonal codes. One obvious shortcoming of MC-CDMA lies in the complexity of receiver design due to the fact that all users’ signals must be detected jointly at a receiver in which the orthogonality of the codes is usually not preserved because of fading. An MC-CDMA receiver thus requires a highly complex multiuser-detection scheme to mitigate interference caused by the poor orthogonality of the codes.

3. **OFDMA**: In OFDMA, each user is assigned a subset of sub-carriers for the duration of allocated time slots. That is, multiple users occupy different disjoint portions of the entire bandwidth for the duration of allocated time slots. Both time and frequency dimensions thus are exploited. Unlike OFDM-TD, OFDMA is quite flexible because resource allocation to each user can be performed in a finer-grained manner, e.g., the small number of sub-carriers can be allocated to low-rate users. The amount of transmission power and the type of modulation also can be assigned on a per sub-carrier basis to each user. Further, not fully exploited in OFDM-TD and MC-CDMA, multiuser diversity can be exploited in OFDMA by allocating different portions of bandwidth along with transmission power and modulation type for a particular duration of allocated time slots to those users who can best utilize the resources. Resource allocation thus has a crucial role to play in OFDMA systems. In this article we focus only on OFDMA systems.

### 10.3 Adaptive Radio Resource Allocation in OFDM Systems

From the radio resource allocation point of view, the performance of OFDMA transmission systems can be optimized through three main mechanisms: (1) sub-carrier assignment, (2) bit loading, and (3) power loading. Optimization through sub-carrier assignment determines the best set of sub-carriers to which each user should be assigned for a temporal snapshot of, say, one OFDM symbol duration. Optimization through bit- and power-loading determines the best allocation of (1) the number of bits (i.e., modulation type) and (2) the amount of power to each sub-carrier, respectively. Optimization is therefore an important mathematical tool in designing OFDMA systems. All the
above three cases of optimization depend heavily on channel state information (CSI) for making an allocation decision. The accuracy of channel estimation is then usually taken into account when performing such optimization. We define an allocation mechanism as adaptive if its allocation or assignment adapts to available CSI on a temporal basis. Any combination of these three adaptive mechanisms therefore constitutes adaptive radio resource allocation.

In this section, we will discuss the major approaches to adaptive radio resource allocation in OFDMA systems. Recently, much effort has been spent in making radio resource allocation in OFDM systems as adaptive and efficient as possible. The problem of adaptive radio resource allocation thus has been tackled from many different angles, resulting in a long list of approaches and ideas. In this chapter, we aim at providing a unified view of research in this field. Discussions on pros and cons of the major approaches will be presented as we progress.

10.3.1 System-Centric Approaches

Most of the adaptive radio resource allocation schemes for OFDMA systems proposed in the literature fall into one of the following two categories: (1) Rate Adaptive (RA) and (2) Margin Adaptive (MA) schemes. In an RA scheme, the objective is to maximize the number of bits per symbol (i.e., data rate) subject to a fixed total transmission power. On the other hand, the objective of an MA method is to minimize the total transmission power subject to a fixed data rate required by each user. We refer to both methods as system-centric methods because their main focus is to optimize system-centric measures such as data rate and transmission power only. We review both methods in the following.

Rate Adaptive (RA) Approach

Consider a single-user scenario in which the system has $N$ parallel sub-carriers and total available bandwidth of $B$ Hz. The objective of the RA approach is to assign an amount of power $p_n$ to each sub-carrier $n \in \{1, 2, \ldots, N\}$ such that the number of conveyable bits in each transmitted OFDM symbol (i.e., data rate) is maximized:

$$\max_{p_n} \, \sum_{\forall n} \frac{B}{N} \log_2 \left( 1 + \frac{p_n \cdot |h_n|^2}{\Gamma \cdot \sigma^2} \right)$$

(10.1a)

subject to:

$$\sum_{\forall n} p_n \leq P_{\text{max}}$$

(10.1b)

$$p_n \geq 0, \forall n$$

(10.1c)

where $h_n$ denotes the channel-gain as perceived by the user on sub-carrier $n$. $\Gamma$ is the SNR gap defined as the difference between the SNR needed to achieve a certain data transmission rate for a practical system and the SNR needed to reach the theoretical limit. The required bit error rate (BER) is embedded in this SNR gap. Here $\sigma^2$ represents noise power per sub-carrier. Constraint (10.1b) ensures that the total transmission power does not exceed the maximum available value $P_{\text{max}}$. The constraint in (10.1c) ensures that the value of allocated power is non-negative.
Now, consider a multiuser scenario. In particular, consider a single-cell downlink OFDMA system consisting of one BS and $K$ UEs. The system employs $N$ sub-carriers and $M$ different available modulation/coding schemes. At a particular instant, each UE experiences different channel-gains on different sub-carriers. Let $h_{k,n}$ denote the channel-gain as perceived by UE $k$ on sub-carrier $n$ at time $t$. Each UE will be assigned different subsets of sub-carriers. Define a binary assignment variable $d_{k,n}$ which is one if sub-carrier $n$ is assigned to UE $k$ at time $t$, and zero if sub-carrier $n$ is not assigned to user $k$ at time $t$. Based on the power allocation $p_n$ for each sub-carrier $n$ at time $t$, one of $M$ modulation/coding schemes then must be selected for each sub-carrier such that the number of bits per symbol, or equivalently, the data rate, is maximized, and that a required level of QoS such as BER is achieved. One straightforward RA approach as proposed in [10], [12], [13], [17], and [38] is to maximize the total raw data rate of a cell:

$$\max_{d_{k,n},p_n} \sum_k \sum_n \mathcal{F}(\gamma_{k,n}, P_e) \cdot d_{k,n}$$

(10.2a)

subject to

$$\sum_k d_{k,n} \leq 1, \forall n$$

(10.2b)

$$\sum_n p_n \leq P_{\text{max}}$$

(10.2c)

$$p_n \geq 0, \forall n$$

(10.2d)

where $\mathcal{F}(\gamma_{k,n}, P_e)$ is a rate-power function that relates data rate per sub-carrier to the signal-to-noise ratio (SNR) $\gamma_{k,n} = \frac{p_n |h_{k,n}|^2}{\sigma^2}$ of UE $k$ on sub-carrier $n$ and bit error probability $P_e$. The constraint (10.2b) ensures that each sub-carrier is assigned to one and only one UE at a time. Although it is possible for multiple UEs to share a single sub-carrier (for each OFDM symbol duration), it was proved in [17] that the total data rate of the multiuser OFDMA system is maximized when only one user with the best channel-gain occupies each sub-carrier. Although the proof cannot be generalized to other modulation schemes besides QAM, the result is quite significant since it proves the fundamental assumption made in most of the works discussed in this chapter. Equations (10.2c) and (10.2d) impose the usual constraints as appeared in the single user formulation. One additional constraint usually imposed in the multiuser RA formulation is to ensure that the data rate of each user is equal or above some required minimum:

$$R_k \triangleq \mathcal{F}(\gamma_{k,n}, P_e) \geq R_k^{\text{min}}, \forall k, \forall n \in A_k'$$

(10.3)

where $R_k^{\text{min}}$ is a minimum rate required by UE $k$, and $A_k'$ is the set of sub-carriers assigned to UE $k$.

As an example, the function $\mathcal{F}(\cdot)$ was defined in [17] as follows:

$$\mathcal{F}(\gamma_{k,n}, P_e) = \frac{B}{N} \log_2 \left( 1 + \frac{\gamma_{k,n}}{\Gamma} \right)$$

where the SNR gap $\Gamma$ is expressed as

$$\Gamma = \frac{-\ln(5 \cdot P_e)}{1.5}.$$
As the first work that employed the RA approach, [10] dealt with resource allocation for a single-cell/downlink/multiuser OFDMA system in two separate steps. In the first step, the number of sub-carriers to be assigned to each UE as well as the amount of power to be allocated to each sub-carrier are determined. See Algorithm 6 for details. The idea is to gradually increase the number of sub-carriers assigned to each UE as well as the amount of power allocated for each sub-carrier until its minimum rate requirement is satisfied. The algorithm assumes that the amount of power allocated to each UE is proportional to the number of sub-carriers allocated to that UE. In the second step, the sub-carrier assignment (using the Hungarian method) and bit loading are performed.

Algorithm 6 First step of resource allocation proposed in [10]

1: Assumptions: $N_k \geq 0$ and $p_n \geq 0$, where $N_k$ is the number of sub-carriers assigned to UE $k$. Each UE $k$ requires a data rate of $R_k$, and has a channel-gain of $h_k$ averaged over all the sub-carriers assigned to the UE.
2: For each UE $k \in \{1, 2, 3, \ldots, K\}$, set $N_k = 1$.
3: Calculate $N_a = \sum_{k} N_k$.
4: Calculate each UE's power requirement: $P_k = \frac{N_k}{h_k} \cdot \mathcal{F}^{-1}\left(\frac{R_k}{N_k}\right)$.
5: Compute total transmission power requirement: $P_a = \sum_{k} P_k$.
6: while $P_a \geq \frac{N_a}{N} P_{\text{max}}$ do
7: Let $\Delta P_k = \left[\frac{N_k}{h_k} \cdot \mathcal{F}^{-1}\left(\frac{R_k}{N_k}\right) - \left(\frac{N_k+1}{h_k}\right) \cdot \mathcal{F}^{-1}\left(\frac{R_k}{N_k+1}\right)\right]$.
8: Choose $\hat{k} = \arg \min_k \Delta P_k$.
9: Update $N_{\hat{k}} = N_{\hat{k}} + 1$ and $P_{\hat{k}} = P_{\hat{k}} - \Delta P_{\hat{k}}$.
10: end while

Instead of maximizing the total data rate, one interesting RA approach [9] is to maximize the minimum of all UEs’ data rates. The authors relax the binary assignment variable $d_{k,n}$ to assume a real value in $[0, 1]$, instead of binary 0 or 1. This in effect turns the problem into a typical convex optimization problem, which can be solved for optimal solutions by any standard package. Computational complexity is, however, still high. The authors then resort to a heuristic that achieves sub-optimal solutions. The heuristic simply allocates an equal amount of power to each sub-carrier, and then uses the procedure in Algorithm 7 to perform sub-carrier assignment. The resulting suboptimal solutions have been shown to be very close to the optimal ones while the complexity is greatly reduced.

Other variants of the RA approach were also proposed. In [17], the transmit power adaptation was performed using a two-step approach. As multiple UEs can share a single carrier, the first step is to determine a set of UEs who should transmit data on a specific sub-carrier such that the data rate on that sub-carrier is maximized. The second step then involves allocating power to each sub-carrier such that the overall data rate is maximized. Implicitly, after sub-carrier assignment, this approach treats the system as a single-user OFDM system. No fairness or individual rate constraints, however, were considered in this work. In [13], the overall data rate was maximized...
while maintaining proportional fairness. A two-step approach was also adopted in this work, i.e., sub-carrier assignment is performed prior to power allocation. While the sub-carrier assignment phase was borrowed from [9], the power allocation step was performed by solving a system of $K$ non-linear equations. The authors thus aimed for sub-optimal solutions that are computationally efficient to achieve.

A generalized proportional fairness based on Nash bargaining solutions and coalitions was considered for an uplink single-cell scenario in [26]. That is, a BS acts as a market through which users (customers) gather to bargain cooperatively with one another for sub-carriers. Typically, the main objective is to maximize the overall data rate, subject to individual required data rates and maximum power available, while maintaining fairness through the Nash bargaining solutions and coalitions. The simulation results showed that the proposed scheme can provide fair resource allocation while still maintaining the performance in terms of overall uplink data rate, comparable to that of a typical RA scheme in which fairness is ignored. The overall data rate of the proposed scheme was also observed to be higher than that of the max-min fairness scheme. Another RA approach that considers proportional fairness can be found in [28].

In [30], both RA and MA approaches were jointly considered. The main contribution of this approach is to reformulate a non-linear optimization problem, as formulated in [7] and [9], into an integer program. Conversion of the non-linear optimization into an integer program, which can be solved by any standard package, was done by introducing a new indicator variable that is related to the sub-carrier assignment variable $d_{k,n}$ and the number of bits loaded onto each sub-carrier. Even with this reformulation, solving for optimal solutions is still computationally expensive. Based on [30], the work by Mao and Wang in [38] formulated the RA problem as an integer linear programming problem, and then tries to find the optimal solution via the Branch-and-Bound approach. Complexity, however, is still too high. The authors then proposed two suboptimal algorithms that trade complexity for performance (data rate) degradation. The performance

Algorithm 7 Sub-Carrier Assignment proposed in [9]

1: For each UE $k \in \{1, 2, 3, \ldots, K\}$, set $R_k = 0$, where $R_k$ is a data rate of UE $k$. Also let $A = \{1, 2, 3, \ldots, N\}$ be a set of sub-carriers.
2: Calculate $N_d = \sum_{k \in A} N_k$.
3: for $k = 1$ to $K$ do
   4: Find $\hat{n}$ such that $|h_{k,\hat{n}}| \geq |h_{k,n}|$ for all $n \in A, n \neq \hat{n}$.
   5: Update $R_k = \mathcal{F}(h_{k,\hat{n}}, P_k)$, and $A = A - \hat{n}$.
4: end for
5: while $A$ is not empty do
   6: Find $\hat{k}$ such that $R_{\hat{k}} \leq R_k$ for all $k \neq \hat{k}$.
   7: For $\hat{k}$, find $\hat{n}$ such that $|h_{\hat{k},\hat{n}}| \geq |h_{\hat{k},n}|$ for all $n \in A, n \neq \hat{n}$.
   8: Update $R_{\hat{k}} = R_{\hat{k}} + \mathcal{F}(h_{\hat{k},\hat{n}}, P_k)$, and $A = A - \{\hat{n}\}$.
7: end while
of this approach in terms of maximum data rate was shown to outperform that of [30] and [10].

In [20], a cross-layer approach for multi-input multi-output OFDM (MIMO-OFDM) was proposed. At the PHY layer, the proposed scheme allocates a set of transmit antennas and modulation levels for both real-time and non-real-time (traffic) users, using the transmit antenna selection (TAS) technique and adaptive modulation, respectively. At the MAC layer, time slots are allocated for real-time users to ensure bounded delay. Sub-carriers are assigned to non-real-time users to maximize the overall throughput while maintaining proportional fairness among users. Unrealistic in this approach is the assumption that the QAM modulation scheme employs only a continuous signal-constellation size. Another cross-layer RA approach was proposed in [32], which jointly solves the problem of sub-carrier allocation, adaptive modulation, and power control in a multi-cell setting. Intercell interference was therefore taken into account in the formulation of the optimization problem. Only one QoS requirement, the bit error rate (BER), was considered in this work.

**Margin Adaptive (MA) Approach**

Consider a single-user scenario in which the system has $N$ parallel sub-carriers and total available bandwidth of $B$ Hz. The objective of the MA approach in this case is to assign an amount of power $p_n$ as minimal as possible to each sub-carrier $n \in \{1, 2, \ldots, N\}$ such that the number of conveyable bits in each transmitted OFDM symbol (i.e., data rate) is greater than or equal to some fixed number $R$:

\[
\begin{align*}
\min_{p_n} & \quad \sum_{\forall n} p_n \\
\text{s.t.} & \quad \sum_{\forall n} \frac{B}{N} \log_2 \left(1 + \frac{p_n \cdot |h_n|^2}{\Gamma \cdot \sigma^2}\right) \geq R \\
& \quad p_n \geq 0, \forall n
\end{align*}
\]

where $R$ represents the data rate required by this particular user, and all the other notations convey the same meanings as previously defined in the RA section.

For the downlink multiuser scenario, each user is also assumed to require a certain data rate $R_k$ (for a certain BER), which then translates into (1) the number of bits to allocate to each sub-carrier assigned to that user, and (2) the modulation/coding scheme to be used (i.e., how much received power is required to achieve a certain data rate). The objective of the MA approach, in general, is as follows:

\[
\begin{align*}
\min_{d_{k,n}, p_n} & \quad \sum_{\forall k} \sum_{\forall n} p_n \cdot d_{k,n} \\
\text{s.t.} & \quad \sum_{\forall k} d_{k,n} \leq 1, \forall n \\
& \quad \sum_{\forall n} \mathcal{F}(\gamma_{k,n}, P_e) \cdot d_{k,n} \geq R_k, \forall k
\end{align*}
\]

where all the notations still carry the same meanings as in the RA case.
The first work that studied MA resource allocation in OFDMA networks is [7]. The optimization problem in this case is

\[
\begin{align*}
\min_{d_{k,n},c_{k,n}} & \sum_{\forall k} \sum_{\forall n} f_k(c_{k,n}) \cdot d_{k,n} \\
\text{s.t.} & \sum_{\forall n} d_{k,n} \cdot c_{k,n} = R_k, \forall k \in \{1, \ldots, K\} \\
& \sum_{\forall k} d_{k,n} = 1, \forall n \in \{1, \ldots, N\}
\end{align*}
\tag{10.6a}
\]

where \( f_k(c_{k,n}) \) is UE \( k \)'s required received power (in joule per symbol) in sub-carrier \( n \) for reliable reception of \( c_{k,n} \) bits/symbol. \( f_k(c_{k,n}) \) therefore depends on the modulation/coding scheme selected. When relaxing \( c_{k,n} \) and \( d_{k,n} \) to assume real values in \([0, M]\) and \([0, 1]\), respectively, the above optimization problem becomes convex and thus tractable. The optimization assumes perfect knowledge of CSI. Although the complexity involved in solving for optimal solutions in this case is still very high, this work serves as a baseline for other works that follow.

Following [7], Kivanc and Liu [11] proposed a suite of more computationally efficient algorithms for suboptimal power-and-sub-carrier allocation. A two-step approach was adopted. In the first step, the algorithms determine how many sub-carriers each UE will receive, based on that UE’s minimum required data rate. Based on CSI, a set of sub-carriers is assigned to each UE in the second step. The performances of these algorithms in terms of transmit energy per symbol were shown to be quite close to the optimal solutions as obtained in [7]. The details of each algorithm are shown in Algorithms 8 and 9, where all the notations carry their usual meanings. In the BABS (Bandwidth Assignment Based on SNR) algorithm, the number of sub-carriers allocated to each user is determined by assuming equal fading in all sub-carriers. The ACG (Amplitude Craving Greedy) algorithm simply allocates sub-carriers to those UEs with the highest channel-gains. The number of sub-carriers assigned according to this rule, however, must not exceed the number specified in the first step. No rate constraint is considered in this algorithm.

All the previous works treat each sub-carrier independently. In [15], the authors exploited the coherence bandwidth of the channel, which usually spans several highly correlated sub-carriers. The idea is to group multiple highly-correlated sub-carriers into blocks of equal spectral size before assigning each block to each user. The main objective of this approach is to support as many users as possible while keeping the overall transmission power within a constraint, and rate requirements \( R_k \) for each user \( k \) satisfied. The performances in terms of average bit SNR and the number of users supported were shown to outperform those of the static OFDMA scheme. The tradeoff between the number of users supported and the number of sub-carriers grouped in each block

\footnote{Recall that \( M \) here is the number of modulation/coding schemes available.}
Algorithm 8 Bandwidth Assignment Based on SNR (BABS) [11]

1: Let $R_{\text{max}}$ be the maximum rate (bits/unit time) a sub-carrier can transmit. Also define $R_k^{\text{min}}$ as the minimum rate required by UE $k$.

2: Calculate how many sub-carriers to allocate to UE $k$: $N_k = \left\lceil \frac{R_k^{\text{min}}}{R_{\text{max}}} \right\rceil$, $\forall k$.

3: while $\sum_{k} N_k > N$ do
4:   Search for $\hat{k} = \arg\min_{\forall k} N_k$.
5:   Set $N_{\hat{k}} = 0$.
6: end while

7: Calculate the average channel-gain for UE $k$ over all sub-carriers: $\bar{h}_k$.

8: while $\sum_{k} N_k < N$ do
9:   Calculate the difference in transmission power when one more sub-carrier is assigned to $k$: $\Delta p_k = N_k + 1 \cdot \mathcal{F} \left( \frac{R_k}{N_k + 1} \right) - N_k \cdot \mathcal{F} \left( \frac{R_k}{N_k} \right)$, $\forall k$.
10:  Search for $\hat{k} = \arg\min_{\forall k} \Delta p_k$.
11:  Set $N_{\hat{k}} = N_{\hat{k}} + 1$.
12: end while

Algorithm 9 Amplitude Craving Greedy (ACG) [11]

1: First set $A_k = \{\}$, $\forall k$.

2: for each sub-carrier $n$ do
3:   Search for $\hat{k} = \arg\max_{\forall k} |h_{k,n}|^2$.
4:   while $|A_k| = N_{\hat{k}}$ do
5:     $h_{\hat{k},n} = 0$.
6:   Search for $\hat{k} = \arg\max_{\forall k} |h_{k,n}|^2$.
7: end while
8: Set $A_{\hat{k}} = A_{\hat{k}} \cup \{n\}$.
9: end for

was also noted. Other contributions that employ this idea of sub-carrier grouping can be found in [16], [18], [29], [35], and [36].

Instead of focusing only on the PHY layer, several contributions in the literature concentrate on a cross-layer approach in which the performances of both MAC and PHY layers are taken into account. In [19], scheduling and power-and-sub-carrier allocation in a packet-switched OFDMA system were considered. The main objective is to maximize power efficiency while ensuring fairness. Queuing was also considered at the input of a packet scheduler. A frame is divided into a control phase and a transmission phase. The transmission phase may span several OFDM symbols. The optimization is then performed over a temporal snapshot of a frame (or, equivalently, a snapshot of several OFDM symbols), instead of a snapshot of one OFDM symbol as discussed earlier. The formulation of the optimization problem follows that in [30], which involves solving an integer linear programming problem. The objective is to minimize the
overall transmission power given the data rate requirements and the packet error rate for each user:

\[
\begin{align*}
\min_{d_{k,n,s},p_{k,n,s}} & \sum_{k} \sum_{n} \sum_{s \in S} d_{k,n,s} \cdot p_{k,n,s} \\
\text{s.t.} & \sum_{k} d_{k,n,s} \leq 1, \forall n, s \\
\frac{p_{k,n,s}|h_{k,n,s}|^2}{\sigma^2} & = \gamma_k, \forall k, n, s \\
\sum_{n} \sum_{s} d_{k,n,s} \cdot c \cdot r & = g_k \cdot b, \forall k \\
\sum_{k} \sum_{n} \sum_{s \in S} d_{k,n,s} \cdot p_{k,n,s} & \leq P_{\text{max}}
\end{align*}
\]

where \( S \) is a set of symbols embedded in the transmission phase of a frame and \( d_{k,n,s} \) and \( p_{k,n,s} \) represent the (sub-carrier) assignment variable and the allocated transmission power, respectively, for UE \( k \) on sub-carrier \( n \) in symbol \( s \) of the transmission phase of the frame. Constraint (10.7b) imposes the usual restriction that only one sub-carrier must be assigned to each UE at a single point in time. Similarly, constraint (10.7e) imposes the transmission power limit. Constraint (10.7c) ensures QoS in terms of packet-error-rate (PER), where \( \gamma_k \) is the SNR required to achieve a certain PER for UE \( k \). Proportional fairness is dictated by constraint (10.7d), assuming that the bandwidth allocated to the UEs is proportional to the number of packets the users will transmit. Every packet has \( b \) bits and is coded by a rate \( r \) convolutional code. A symbol on each sub-carrier carries \( c \) bits. The notation \( g_k \) denotes the number of packets selected from UE \( k \)’s queue. One drawback in this approach is that only one modulation scheme is allowed in all the sub-carriers. A similar approach was proposed by the same authors in [31]. The assignment variable \( d_{k,n,s} \) was relaxed, however, to assume a real value in order to reduce the complexity of finding the optimal solution. Another similar cross-layer MA approach can be found in [23].

All the MA approaches discussed so far consider only downlink transmission. An uplink scenario was first considered in [33], which formulated the resource allocation problem as an integer linear programming problem. The multiple access interference (MAI) in the uplink transmission was considered. After performing the sub-carrier assignment, a modulation scheme is selected for each sub-carrier, and finally the amount of power allocated to each sub-carrier is calculated according to the signal-to-interference (SIR) ratio constraint. The optimization problem was first solved by an exhaustive search. The complexity of this search was then reduced by assuming that the assignment variable \( d_{k,n,s} \) can take on a real value in the interval \([0, 1]\). The solution was shown to depend on traffic load. While convergence to optimal solution is guaranteed under light traffic conditions, instability and oscillations were observed in a high traffic scenario. A heuristic method was therefore proposed to eject some of the users in order to keep traffic to a stable level.
In [37], resource allocation in an uplink multi-cell scenario was performed in a distributed manner using a non-cooperative game approach. That is, UEs compete for radio resources without cooperating with one another. The objective of each UE (or player) is to minimize its transmission power subject to its required data rate and maximum power available to it. The objective of the system is then to minimize the sum of UEs’ transmission powers subject to the same constraints. Acting as a mediator, a BS is in charge of overseeing and improving the outcome of the non-cooperative game since the competition gives no guarantee of convergence or Nash equilibrium with optimal performances. That is, if the outcome is undesirable (i.e., non-convergence or Nash equilibrium with low performance measures), the mediator will either eject some UEs or reduce their transmission rates in order to keep the rest of UEs satisfied. The proposed scheme is shown to improve over the fixed channel assignment algorithm and the iterative water-filling algorithm in terms of total transmission power and achievable transmission rates.

10.3.2 Application-Centric Approaches

Instead of optimizing system-centric metrics such as the raw rate and overall transmit power directly, one interesting approach is to look at the resource allocation problem from users’ perspective, i.e., try to maximize the metric that quantifies users’ satisfaction. This metric is application-centric and known as utility. The utility of certain resource can be perceived differently by each user. Viewing radio resource allocation from this perspective is thus interesting in that, given each user’s utility function for a certain resource, an appropriate amount of that resource can be allocated to users who really need it. Excessive allocation thus can be avoided while fairness is inherently ensured among users. The objective of this approach in general is to assign radio resources (sub-carriers, transmission power, and modulation type) to users such that the average utility of the system is maximized. Since utility is usually expressed in terms of data rate, the RA method can be regarded as a special case of the utility-based approach.

In [21] and [22], a complete framework for cross-layer optimization of downlink, multiuser, and single-cell OFDMA systems was presented. Utility, as a function of transmission throughput, was used to model interactions between the PHY layer and the MAC layer. Moreover, fairness was traded for efficiency of resource use through the utility function. Three schemes for (1) adaptive sub-carrier assignment, (2) adaptive power allocation, and (3) joint adaptive sub-carrier and power allocation were proposed in [21]. The necessary and sufficient conditions for finding optimal allocation for each case were also derived. One major assumption in this work is that the system has an infinite number of sub-carriers, i.e., the bandwidth of each sub-carrier is infinitesimal. The system performance, in terms of average utility as a function of average SNR, obtained by this assumption then serves well as an upper bound for comparison because of the infinitesimal granularity of resource allocation. Based on [21], the algorithms for adaptive sub-carrier assignment and adaptive power allocation were proposed in [22] for practical realization. The simulation results of both contributions (in terms
of average utility) were shown to outperform those obtained by the fixed resource allocation scheme. Higher performance gain was observed when the number of users in the system increases. In [36], a utility-based approach was also used. A specific form of utility function was defined. The estimation error of channel conditions was also considered and embedded into the problem formulation. The objective of this approach is to assign a group of sub-carriers to each user and to allocate power to each group in order to maximize the defined utility subject to QoS, long-term fairness, and total transmission power. The resulting optimization problem is non-convex, finite-dimensional, and stochastic. The assignment variable $d_{k,n}$ was relaxed to assume a real value in $[0, 1]$ for tractability. Lagrangian dual decomposition was then applied to the original problem to derive an algorithm that can solve for optimal solutions in an efficient manner.

### 10.4 Open Research Issues

Adaptive resource allocation can significantly improve the performance of OFDMA systems over that of the static or fixed allocation case. Several issues, however, remain open for further investigation. These open research issues are as follows:

- Adaptive channel allocation in multi-tier OFDMA cellular networks under co-channel deployment as well as in cloud-based RANs in heterogeneous networks is a challenging research problem. Some recent works (e.g., [39, 40]) have addressed this problem for two-tier networks, which can be extended for general $K$-tier (e.g., D2D-enabled as well as relay-enhanced macrocell-small cell networks). Also, OFDM/OFDMA-based adaptive channel allocation for wireless self-backhauling [41] in multi-tier networks is an open problem. Distributed and/or low-complexity adaptive resource allocation methods would be preferable for OFDMA-based future generation ultra-dense cellular networks.

- Recent developments mostly focus on fixed or portable applications in which slow fading is usually assumed. Slow varying channels thus do not trigger adaptive resource allocation so often, resulting in low signaling overhead and quite accurate channel estimation. But it is challenging when it comes to fast fading, which is usually induced by mobility-based applications. In this scenario, channel estimation must be performed and fed back to the resource allocator very often in order to keep up with fast fading channels. The duration of a temporal snapshot in which optimization must be performed thus varies with time and speed of applications. Based on this duration, the question is what layer (MAC or PHY) performs adaptive resource allocation in which fading environments (slow or fast fading)? Mobility is therefore one interesting issue to explore in the context of adaptive OFDMA resource allocation.

- Most schemes discussed so far require the accurate knowledge of channel conditions to perform adaptive resource allocation. A significant performance gain resulting from these approaches comes with a price in signaling overhead, especially in the frequency division duplex (FDD) mode of operation in which channel measurements must be fed back to the resource allocator. Although many attempts ([15], [16], [18],...
[29], [35], [36]) have tried to reduce signaling overhead through sub-carrier grouping, it is not clear how much performance degradation must be traded for this reduced signaling overhead. The issue of signaling overhead in a fast fading scenario induced by mobility-based applications is also still open for further investigation. Embedding signaling overhead in the optimization problem is interesting from a practical point of view.

- Most schemes discussed earlier focus mainly on the PHY layer. This concept of decoupling layers from one another eases the analysis of traditional wireless networks. For next generation wireless networks, however, interaction between layers is key for handling the integration of various services with highly diverse QoS requirements. For example, in order to support integrated voice and data services, the MAC layer needs to obtain information such as type of service and associated utility, QoS, or priorities from upper layers. This information then can be used to map each type of service to appropriate PHY layer parameters. A cross-layer model that takes these interactions into account will provide insights into designing modern wireless networks. Additionally, notably open for further investigation is adaptive resource allocation in OFDMA systems characterized by a MIMO transmission technique and robust error correction strategies. A complete cross-layer study that integrates these two components for a packet-based wireless network with diverse QoS requirements may reveal interesting practical findings.

- In the application-centric approach, the question is how to model a utility function for each application that really reflects the need and satisfaction of each user. Another question to ask is how to prevent users’ misbehaving in submitting to the resource allocator their false utility functions that greedily ask for more resources than they need. Since this is a higher-layer issue, an entirely new cross-layer optimization model is needed.

References


References


11 Resource Allocation in Relay-Based Networks

11.1 Introduction

Cooperative diversity via relaying communications is an important technology to enhance the robustness and network capacity of wireless systems. Cooperative transmission between a source node and a destination node is performed with the assistance of a number of relay nodes where the source and relay nodes collaboratively transmit information to the destination node [1–5]. It is intuitive that to make cooperative transmission efficient or even possible, the source node has to carefully choose one or several good relays. Then the resource first forwards its data to those relays. After that the source and selected relays can coordinate their transmissions in such a way that maximum multiplexing or diversity gains can be achieved at the destination node.

Although cooperative diversity is simple in concept, there are many technical issues to be resolved for practical implementation. First, protocol design for cooperative diversity is one important research issue [1, 2, 4–6]. Second, most practical cooperative diversity protocols have multiple phases. For example, the source node broadcasts its message to assisting relays in the first phase, and then the relays collaboratively transmit the received information to the destination in the second phase. Therefore, cooperative transmission may not be always beneficial or even necessary because direct transmission from the source to the destination node may already be successful.

Adaptive cooperative protocols, where nodes cooperate only when necessary or they cooperate using incremental transmissions, usually result in significantly better performance than straightforward protocols [6, 7]. In addition, emerging technology such as network coding can be employed to design even smarter cooperative protocols [8]. Finally, other important issues such as relay selection and synchronization among relays’ transmissions need to be considered for practical implementation [9–15]. Research work in cooperative communication has covered different aspects of design and performance analysis of cooperative protocols as well as resource allocation.

In general, resource allocation has significant impacts on system performance [16–20]. In practice, source nodes and assisting relays usually have limited radio resources such as bandwidth and power, and these resources must be shared by several wireless nodes. Therefore, a smart radio resource allocation for wireless relay networks guarantees both fair access to available relays and good overall network performance. In addition, by using a proper relay selection strategy where each source-destination pair selects only one or a small number of good relays, efficient resource utilization can be achieved with low implementation complexity.
11.2 Overview of Cooperative Diversity

Cooperative relaying protocols allow a number of users to relay signals for one another in such a way that a diversity gain can be achieved. In fact, the information theoretic capacity of such a network setting, named a relay channel, has been investigated a few decades ago [21]. Deep understanding of multiple-input multiple-output (MIMO) systems from both information theoretic and practical system design viewpoints over the past decade has stimulated and attracted significant research efforts in cooperative diversity. In this section, we discuss fundamental aspects of cooperative diversity.

Consider a source node $s$ communicating with a destination node $d$ with the assistance of $M$ relays, $r_1, r_2, \ldots, r_M$. Let $P_i$ be the transmission power of node $i$, $a_{ij}$ be the channel-gain between nodes $i$ and $j$, and $h_{i,j} = |a_{ij}|^2$ denote the corresponding power channel-gain. The signal is corrupted by the white Gaussian noise. We assume that $N_i$ is the white Gaussian noise power measured in the signal bandwidth at node $i$, which is set equal to $N$ for all nodes in some cases to gain insights into the system performance. We assume that cooperation among wireless nodes is realized in multiple phases (i.e, time slots) with perfect synchronization by using a common system clock.

Figure 11.1 shows a general relaying system where one source-destination pair is assisted by multiple relays. In the following, we describe some popular cooperative relaying protocols and their corresponding performances.

### 11.2.1 Amplify-and-Forward Relaying

In this cooperative protocol, the source broadcasts message $x_i$ in the first phase, which is received by the destination and relays. Each relay $r_i$ amplifies the signal it received in the
first phase and transmits to the destination in the second phase. The noise powers at the relay and destination nodes are assumed to be equal to $N$ for simplicity. The destination combines the signals received in both phases to decode the message. Specifically, the received signal by relay $r_i$ in the first phase can be written as

$$ y_{r_i} = a_{sr} x_s + z_{r_i} $$

where $z_{r_i}$ denotes the white Gaussian noise at relay $r_i$. Suppose that each relay normalizes the received signal from the source before transmitting to the destination. Then the transmitted signal from relay $r_i$ can be written as

$$ x_{r_i} = g_{r_i} y_{r_i} $$

where $g_{r_i}$ is the amplifying gain, which is given as

$$ g_{r_i} = \sqrt{\frac{P_{r_i}}{h_{sr} P_s + N}}. $$

Assuming that a maximum ratio combiner (MRC) is employed at the destination, the end-to-end rate achieved by this protocol is given as [13]

$$ C_{AF} = \frac{1}{M+1} \log \left( 1 + \frac{\sum_{i=1}^{M} \gamma_i h_{sr} h_{rd} + 1}{\gamma_i h_{sr} + h_{rd} + 1} \right) $$

where $\gamma_i = P_i / N$ is the signal-to-noise ratio (SNR) at node $i$.

Another important performance measure which is extensively used for investigating performance of different cooperative protocols is outage probability. In the Rayleigh fading channel, the outage probability of the amplify-and-forward (AF) cooperative protocol can be approximated as [13]

$$ p_{AF}^{out} (SNR, R) \triangleq \Pr [C_{AF} < R] \propto \left( \frac{2^R - 1}{\gamma} \right)^{M+1} $$

where $\gamma = P/N$, and it is assumed that all nodes transmit at power level $P$. This outage probability expression shows that AF cooperative protocol achieves diversity order of $M + 1$ with $M$ relays.

### 11.2.2 Decode-and-Forward Relaying

For the decode-and-forward (DF) cooperative protocol, relay nodes apply some forms of detection and/or decoding before encoding the information and forwarding it to the destination. Such a cooperative protocol also has two phases (i.e., time slots). In the first phase, the source broadcasts the signal to the relays, which subsequently detect and/or decode it. In the second phase, the relays transmit re-encoded signals to the destination using repetition or space-time codes.

For protocols that require relays to fully decode the received signal in the first phase, the set of relays that successfully decode the signal at the end of the first phase is only a subset of all available relays. Let $D(s)$ denote the set of successfully decoding relays,
which will be called a decoding set in the following. For repetition-based coding, the destination receives separate retransmission from each relay $r_j \in D(s)$. Hence, we can write the received signal at the destination from relay $r_i$ as

$$y_d = a_{r_i}x_{r_i} + z_d$$  \hspace{1cm} (11.6)

where $x_{r_i}$ denotes the signal transmitted by relay node $r_i$ and $z_d$ denotes the Gaussian noise at the destination. If space-time coding is used, the destination will receive the superimposed signals from all relays $r_i \in D(s)$ simultaneously. Hence, the received signal at the destination in the second phase can be expressed as

$$y_d = \sum_{r_i \in D(s)} a_{r_i}x_{r_i} + z_d.$$  \hspace{1cm} (11.7)

It has been shown that both repetition-based or space-time-coding-based DF protocols achieve full diversity order of $M + 1$ in the low rate regime [2]. This diversity gain can be achieved by a distributed linear dispersion codes [22] and randomized space-time codes [23]. Although both AF and repetition-coding-based DF protocols achieve a full diversity gain, their throughput may degrade because each transmitting relay spends one time slot to transmit to the destination. This limitation can be overcome by enhancing cooperative protocols, namely, by using selection or opportunistic or incremental relaying protocol, which will be described in the following.

### 11.2.3 Selection or Opportunistic Relaying

Suppose there are $M$ relays available to assist the transmission from the source to the destination. Instead of allowing all relays in the AF protocol or only the relays in the decoding set in the DF protocol to transmit in the second phase, selection/opportunistic relay protocols choose one “best” relay to transmit in the second phase [9–11], [13–15]. Interestingly, cooperative protocols using smart relay selection strategies usually achieve full diversity order while providing higher throughput than the standard protocols. In fact, the superior throughput performance of selection-relaying protocols stems from the fact that they use radio resources (i.e., power and bandwidth) more efficiently than basic cooperative protocols presented in the previous sections.

Some typical relay selection strategies for both AF- and DF-based protocols are presented next. Consider an AF protocol with one selected relay, say, $r_i$. From (11.4), the achievable rate of the source-destination channel with one relay is

$$C_{SAF} = \frac{1}{2} \log \left( 1 + \frac{\gamma_s h_{sd} + \gamma_s h_{sr} \gamma_r h_{rd}}{\gamma_s h_{sr} + \gamma_r h_{rd} + 1} \right).$$  \hspace{1cm} (11.8)

Therefore, to maximize the achievable rate, a relay selection strategy would choose relay $r_i^*$, which maximizes the following metric [13]

$$r_i^* = \arg\max_i \frac{\gamma_s h_{sr} \gamma_r h_{rd}}{\gamma_s h_{sr} + \gamma_r h_{rd} + 1}.$$  \hspace{1cm} (11.9)

For the DF protocol, there is a set of relays that successfully decode the signal in the first phase (i.e., in the decoding set $D(s)$). If relay $r_i \in D(s)$ is chosen for transmission
Resource Allocation in Relay-Based Networks

in the second phase, the achievable rate of the source-destination channel is

\[ C_{\text{SDF}} = \frac{1}{2} \log \left( 1 + \gamma_s h_{sl} + \gamma_r h_{rd} \right). \]  

(11.10)

Therefore, to maximize the source-destination rate, an opportunistic relay selection strategy would choose relay \( r_i^* \) as follows:

\[ r_i^* = \arg\max_{i \in D(s)} \gamma_r h_{rd}. \]  

(11.11)

In [9] and [13], it has been shown that relay selection strategies in (11.9) and (11.11) achieve the full diversity order. Note that these selection metrics require the estimates of \( \gamma_r h_{sr} \) and \( \lambda_r h_{rd} \). In [10], the authors proposed two simpler relay selection metrics that require only channel-gains \( h_{sr} \) and \( h_{rd} \). Specifically, they have proposed relay selection strategies that choose a relay as

\[ r_i^* = \arg\max_{r_i} \left\{ \min \left\{ h_{sr}, h_{rd} \right\} \right\} \]  

(11.12)

\[ r_i^* = \arg\max_{r_i} \left\{ \frac{2 h_{rd} h_{sr}}{h_{sr} + h_{rd}} \right\}. \]  

(11.13)

It has been shown in [10] that these relay selection criteria achieve the optimum diversity-multiplexing tradeoff attained by the distributed space-time cooperative protocol [2]. Other relay selection strategies for OFDMA-based wireless cellular relay networks and ad hoc networks can be found in [11] and [15].

11.2.4 Incremental Relaying

Although selection relaying uses radio resources more efficiently than fixed relaying, both fixed and selection relaying protocols have to always repeat transmission. In fact, direct transmission from the source to the destination may be successful if the corresponding channel is in a favorable condition. Therefore, it can be more efficient if relay transmission is invoked only when direct transmission from the source to the destination in the first phase fails.

One simple incremental relaying protocol using the AF principle that exploits this aspect works as follows [1]. The destination upon decoding its received signal at the end of the first phase broadcasts the decoding outcome to the source and relays. If the destination succeeds in decoding the message in the first phase, the source and relays do nothing. Otherwise, all or selected relays amplify their received signals and transmit to the destination. The destination combines all received signals and decodes again.

In fact, incremental relay can be implemented as an extension of H-ARQ protocol [6, 7, 24]. One possible implementation of ARQ-based relaying can be described as follows [7]. Initially, the source node encodes \( b \) bits of information into a code-word with length \( n \) symbols. The code-word is broken into \( M \) blocks, each of which has length \( n/M \). The code can be a simple repetition code where all blocks are identical, or the blocks can be obtained by puncturing a mother code.
The protocol starts by transmitting the first block from the source node. The destination upon decoding the message broadcasts the decoding outcome to all other nodes. If the decoding at the destination is successful, the source proceeds to transmit a new message. Otherwise, either all or one selected relay in the decoding set (i.e., relays which successfully decode the message) re-encodes the message and transmits the second block to the destination. The destination combines all the received blocks and attempts to decode again. This procedure continues until the destination is successful in decoding the message or all $M$ blocks have been transmitted and the message is discarded.

Incremental relaying has both diversity and throughput advantages because relaying is invoked only when necessary. In [1], it has been shown that incremental relaying using AF principle as presented above achieves the full diversity order. In addition, it can be seen that ARQ-based incremental relaying allows many different code designs where well-investigated hybrid ARQ protocol can be adapted to the relaying system. Also, a combination of incremental relaying, hybrid ARQ, and relay selection achieves throughput and energy improvement compared to the standard protocols while still achieving the full diversity gain.

11.2.5 Two-Way Relaying

In two-way relaying, two source nodes $s_1$ and $s_2$ are interested in exchanging information with each other while exploiting the assistance of one or multiple relays. Therefore, there are two traffic flows in two opposite directions. Different designs are possible to support the two-way traffic between the source nodes depending on the relaying strategy employed at the relays as well as the transmissions of the source nodes. We describe some popular two-way relaying schemes in the following.

Straightforward designs can be adopted where each source node employs the existing two time-slot AF or DF relaying strategy to deliver its data to the other node sequentially. These designs require four time slots for data transmissions, which, therefore, severely degrade the system spectrum efficiency. More efficient schemes using only three or even two time slots are indeed possible. We now discuss these potential designs where either the AF or DF relay strategy is employed at the relays.

If the AF strategy is employed at the relays, two popular two-way relaying schemes based on analog network coding (ANC) and time division broadcast (TDBC) have been developed [25]. Moreover, if the relays employ the DF strategy, then the physical-layer network coding technique can be utilized to combine the signals received from the source nodes [26]. Detailed descriptions of these two-way relaying schemes are given in the following.

**TDBC Two-Way Relaying**

The TDBC scheme requires three time slots to complete the data transmissions. In the first and second time slots, source nodes 1 and 2 sequentially transmit their signals, which are received by the relays, nodes 2 and 1, respectively. Then, each relay combines the signals received in the first two time slots and broadcasts the resulting signal in the third time slot.
Let the signals transmitted by the two source nodes $s_1$ and $s_2$ be $x_{s_1}$ and $x_{s_2}$, respectively. The signals received by nodes $s_2$ and $s_1$ in the first and second time slots can be written as

$$y_{s_2} = a_{s_1 s_2} x_{s_1} + z_{s_2}$$
$$y_{s_1} = a_{s_2 s_1} x_{s_2} + z_{s_1}$$

(11.14)
(11.15)

where $a_{s_i s_j}$ denote the channel-gains from node $s_i$ to node $s_j$ and $z_{s_i}$ represents the white Gaussian noise at node $s_i$.

Moreover, the signals received by relay $r_i$ in the first and second time slots are, respectively,

$$y_{r_i,1} = a_{s_1 r_i} x_{s_1} + z_{r_i,1}$$
$$y_{r_i,2} = a_{s_2 r_i} x_{s_2} + z_{r_i,2}.$$  

(11.16)
(11.17)

In the third time slot, relay $r_i$ combines the signals received in the first two slots as follows:

$$x_{r_i} = \zeta_{r_i,1} y_{r_i,1} + \zeta_{r_i,2} y_{r_i,2}$$

(11.18)

where the parameters $\zeta_{r_i,j}$ can be set as

$$\zeta_{r_i,j} = \frac{\phi_{r_i,j} P_{r_i}}{h_{s_j r_i} P_{s_j} + N_{r_i}}$$

(11.19)

where the parameters $\phi_{r_i,j}$ satisfy $\phi_{r_i,1} + \phi_{r_i,2} = 1$ and $0 < \phi_{r_i,j} < 1$. Each relay $r_i$ then broadcasts the combined signal $x_{r_i}$ to $s_1$ and $s_2$. The signal received by $s_1$ and $s_2$ can be expressed as

$$y_{s_1, r_i} = a_{r_i s_1} x_{r_i} + v_{s_1}$$
$$y_{s_2, r_i} = a_{r_i s_2} x_{r_i} + v_{s_2}.$$  

(11.20)
(11.21)

where $a_{r_i s_j}$ describe the channel-gains between relay $r_i$ and node $s_j$, and $v_{s_1}$ and $v_{s_2}$ represent the white Gaussian noise at nodes $s_1$ and $s_2$, respectively.

Since node $s_i$ knows its transmitted signal $x_{s_i}$ and all channel-gains, it can remove the self-interference and obtain the interference-free signals as

$$\hat{y}_{s_1, r_i} = \zeta_{r_i,1} a_{r_i s_1} z_{r_i,1} + \zeta_{r_i,2} a_{r_i s_2} a_{r_i s_1} x_{s_2} + \zeta_{r_i,2} a_{r_i s_1} z_{r_i,2} + v_{s_1}$$

(11.22)

$$\hat{y}_{s_2, r_i} = \zeta_{r_i,2} a_{r_i s_2} z_{r_i,2} + \zeta_{r_i,1} a_{r_i s_1} a_{r_i s_2} x_{s_1} + \zeta_{r_i,1} a_{r_i s_2} z_{r_i,1} + v_{s_2}.$$  

(11.23)

We can approximate the amplification coefficients $\zeta_{r_i,j}$ given in (11.19) assuming the high SNR condition as follows:

$$\zeta_{r_i,j} = \sqrt{\frac{\phi_{r_i,j} P_{r_i}}{h_{s_j r_i} P_{s_j} + N_{r_i}}}.$$  

(11.24)

Assume that the noise powers at the relay $r_i$ and nodes $s_i$ are all equal to $N$ for simplicity. Moreover, we define the following SNRs $\gamma_{s_i} = P_{s_i}/N = P_{s_2}/N$ and $\gamma_{r_i} = P_{r_i}/N$ as before.
Then the achieved SNR corresponding to the signals $\hat{y}_{s_1,r_i}$ and $\hat{y}_{s_2,r_i}$ can be expressed as

$$\text{SNR}_{1l} = \frac{\phi_{r_i,2} h_{s_2,r_i} h_{r_i,s_1}}{\phi_{r_i,2} \gamma_{r_i} h_{r_i,s_1} + \left( \gamma_s + \phi_{r_i,1} \gamma_{r_i} \right) h_{s_2,r_i}}$$

(11.25)

$$\text{SNR}_{2l} = \frac{\phi_{r_i,1} h_{s_2,r_i} h_{r_i,s_2}}{\phi_{r_i,1} \gamma_{r_i} h_{r_i,s_2} + \left( \gamma_s + \phi_{r_i,2} \gamma_{r_i} \right) h_{s_1,r_i}}.$$  

(11.26)

With the same power $P_s$ used by nodes $s_1$ and $s_2$, the achieved SNRs at nodes $s_2$ and $s_1$ due to the direct transmissions between nodes $s_1$ and $s_2$ in the first two time slots can be expressed as $\text{SNR}_{1,0} = h_{s_2,s_1} P_s / N$ and $\text{SNR}_{2,0} = h_{s_1,s_2} P_s / N$, respectively. Assuming that the maximal ratio combiner is employed at nodes $s_1$ and $s_2$ for combining the signals received by the direct and relaying paths, the end-to-end achievable rates at nodes $s_1$ and $s_2$ can be written as

$$C_{TDBC,1i} = \frac{1}{3} \log \left( 1 + \text{SNR}_{1l} + \text{SNR}_{1,0} \right)$$

(11.27)

$$C_{TDBC,2i} = \frac{1}{3} \log \left( 1 + \text{SNR}_{2l} + \text{SNR}_{2,0} \right).$$  

(11.28)

where the factor $1/3$ accounts for the fact that the end-to-end data transmissions are completed in three time slots.

**ANC Two-Way Relaying**

The ANC two-way relaying scheme requires only two time slots for data transmissions. In the first time slot, nodes $s_1$ and $s_2$ transmit simultaneously their data symbols $x_{s_1}$ and $x_{s_2}$, which are received by the relays. Each relay $r_i$ then amplifies the received signal and broadcasts the resulting signal to $s_1$ and $s_2$.

Specifically, the signal received by relay $r_i$ in the first time slot can be written as

$$y_{r_i} = a_{s_1,r_i} x_{s_1} + a_{s_2,r_i} x_{s_2} + z_{r_i}$$

(11.29)

where $a_{s_1,r_i}$ and $a_{s_2,r_i}$ denote the channel-gains while $z_{r_i}$ represents the white Gaussian noise at the relay. Each relay $r_i$ amplifies the received signal by a factor

$$\phi_{r_i} = \sqrt{\frac{P_{r_i}}{h_{s_1,r_i} P_{s_1} + h_{s_2,r_i} P_{s_2} + N_{r_i}}}$$

(11.30)

which ensures that the power of the relayed signal equals $P_{r_i}$. In the high SNR regime, we can approximate this amplification factor as

$$\phi_{r_i} \approx \sqrt{\frac{P_{r_i}}{h_{s_1,r_i} P_{s_1} / N_{r_i} + h_{s_2,r_i} P_{s_2} / N_{r_i}}}$$

(11.31)

$$= \sqrt{\frac{\gamma_{r_i}}{(h_{s_1,r_i} + h_{s_2,r_i}) \gamma_s}}$$

(11.32)

where we have assumed that the powers of nodes $s_1$ and $s_2$ are equal to $P_s$; the noise powers at relay $r_i$ and nodes $s_1$ and $s_2$ are the same, i.e., $N_{r_i} = N_{s_1} = N_{s_2} = N$, and $\gamma_s = P_s / N$ and $\gamma_{r_i} = P_{r_i} / N$. The signal received by node $s_j$ due to the broadcast of signal
Resource Allocation in Relay-Based Networks

\( \varphi_{r_i} y_{r_i} \) from relay \( r_i \) can be expressed as

\[
y_{s_j, r_i} = \varphi_{r_i} a_{r_i s_j} y_{r_i} + v_{s_j} = \varphi_{r_i} a_{r_i s_j} s_{s_j} x_{s_j} + \varphi_{r_i} a_{r_i s_j} x_{s_j} + \varphi_{r_i} a_{r_i s_j} z_{r_i} + v_{s_j} \tag{11.33}
\]

where \( a_{r_i s_j} \) denotes the channel-gain from relay \( r_i \) to node \( s_j \) and \( v_{s_j} \) represents the Gaussian noise at node \( s_j \).

Since node \( s_j \) knows its own information symbol \( x_{s_j} \) and all channel-gains, it can remove the self-interference. Specifically, node \( s_1 \) can remove the interference \( \varphi_{r_1} a_{r_1 s_1} x_{s_1} \) while node \( s_2 \) can cancel the interference \( \varphi_{r_2} a_{r_2 s_2} x_{s_2} \) from their received signals, respectively. Therefore, nodes \( s_1 \) and \( s_2 \) obtain the following signals after the self-interference cancellation:

\[
\hat{y}_{s_1, r_i} = \varphi_{r_i} a_{r_i s_1} s_{s_1} x_{s_1} + \varphi_{r_i} a_{r_i s_1} z_{r_i} + v_{s_1} \tag{11.35}
\]

\[
\hat{y}_{s_2, r_i} = \varphi_{r_i} a_{r_i s_2} s_{s_1} x_{s_1} + \varphi_{r_i} a_{r_i s_2} z_{r_i} + v_{s_2}. \tag{11.36}
\]

Using the approximation in (11.31), the SNRs achieved by the signals \( \hat{y}_{s_1, r_i} \) and \( \hat{y}_{s_2, r_i} \) can be expressed, respectively, as

\[
\text{SNR}_1 = \frac{\frac{1}{2} \log \left( 1 + \text{SNR}_1 \right)}{\gamma_{r_i} h_{r_i s_1} h_{s_1 r_1} y_{s_1 r_1} y_{r_i}} \tag{11.37}
\]

\[
\text{SNR}_2 = \frac{\frac{1}{2} \log \left( 1 + \text{SNR}_2 \right)}{\gamma_{r_i} h_{r_i s_2} h_{s_2 r_2} y_{s_2 r_2} y_{r_i}}. \tag{11.38}
\]

The end-to-end achievable rates at nodes \( s_1 \) and \( s_2 \) can be expressed as

\[
C_{\text{ANC}, 1i} = \frac{1}{2} \log \left( 1 + \text{SNR}_1 \right) \tag{11.39}
\]

\[
C_{\text{ANC}, 2i} = \frac{1}{2} \log \left( 1 + \text{SNR}_2 \right) \tag{11.40}
\]

where the factor \( 1/2 \) accounts for the fact that end-to-end data transmissions are completed in two time slots.

**DF-Based Two-Way Relaying**

In general, we can view the DF-based two-way relaying as the realization of multiple access channels (MACs) in the first phase in which two source nodes transmit their signals to the relay, and the broadcast channel (BC) in the second phase in which the relay broadcasts the signal to the two nodes. Studies of these channels from the information theoretic perspectives have been conducted in [27], [28]. We instead discuss a practical DF-based two-way relaying strategy [26] in the following.

The DF relaying strategy is employed at each relay after nodes \( s_1 \) and \( s_2 \) transmit their information symbols \( x_{s_1} \) and \( x_{s_2} \) to each relay \( r_i \) in two different time slots. The relay decodes the received signals as in the standard DF relaying scheme [26]. Assume that the decoded bits at the relay corresponding to \( x_{s_1} \) and \( x_{s_2} \) are \( b_1(k) \) and \( b_2(k) \), respectively.

In the third time slot, the relay performs network coding operations [8] where it performs Exclusive-OR (XOR) on the bits \( b_1(k) \) and \( b_2(k) \), obtaining \( b_i(k) = b_1(k) \oplus b_2(k) \), available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781316212493.012
modulates the resulting bits \( b_r(k) \), and then broadcasts the resulting information symbols \( x_{r_i} \) to nodes \( s_1 \) and \( s_2 \) in the third time slot. The signals received at nodes \( s_1 \) and \( s_2 \) can be represented, respectively, as

\[
y_{s_1,r_i} = a_{r_i,r_i} x_{r_i} + v_{s_1} \tag{11.41}
\]

\[
y_{s_2,r_i} = a_{r_2,r_i} x_{r_i} + v_{s_2}. \tag{11.42}
\]

Nodes \( s_1 \) and \( s_2 \) then attempt to decode the corresponding received signals. These decoding attempts result in two independent estimates of \( b_r(k) \), which are denoted as \( b_{r,1}(k) \) and \( b_{r,2}(k) \) for nodes \( s_1 \) and \( s_2 \), respectively. Because each node \( s_1 \) (\( s_2 \)) already knows its own information bits \( b_1(k) \) (\( b_2(k) \)), it can obtain the estimated information bits transmitted from node \( s_2 \) (\( s_1 \)) as follows:

\[
\hat{b}_2(k) = b_{r,1}(k) \oplus b_1(k) \tag{11.43}
\]

\[
\hat{b}_1(k) = b_{r,2}(k) \oplus b_2(k). \tag{11.44}
\]

It can be verified that if the decoded bits \( b_{r,1}(k) \) and \( b_{r,2}(k) \) are error-free, then the nodes \( s_1 \) and \( s_2 \) can correctly obtain the information bits from nodes \( s_2 \) and \( s_1 \), respectively.

### 11.2.6 Other Enhancements

In the previous sections, we have discussed only scenarios where each node in the system is equipped with a single antenna. If wireless nodes have multiple antennas, then the joint power allocation and beamforming can be applied at source, destination, and relays to better mitigate the interference and enhance the network performance. In particular, the relay beamforming vectors can be designed in both MAC and broadcast phases of the two-way relaying system to optimize different design objectives such as minimization of total transmission power or maximization of the minimum SINR achieved by the two traffic flows [29]. Such beamforming design can also enable to better mitigate the interference in the multiuser two-way relaying system [30].

All relaying schemes described in previous sections are applicable to half-duplex relays, which can only transmit or receive signals on the same frequency band at any time. Recent advances on interference cancellation techniques have made the full-duplex radio communications feasible. While the full-duplex radio can transmit and receive different signals simultaneously on the same frequency band, interference leakage from the transmitter to the receiver of the full-duplex radio is inevitable because of imperfect interference cancellation in practice [31].

Full-duplex radios can be employed at relays in a full-duplex relaying system [32]. In such a system, each relay can receive the signal from a source node while transmitting the relayed signal to the destination node at the same time by using traditional relaying strategies such as amplify-and-forward and decode-and-forward relaying. In addition, dynamic relay selection and power allocation can be employed to mitigate the self-interference at the full-duplex relays [32, 33]. A hybrid communication scheme that dynamically switches between the half-duplex and full-duplex modes depending on the channel state information can further improve the system performance [33]. Finally,
when relays, source, and destination nodes are equipped with antenna arrays, the joint beamforming and power control can be applied to efficiently manage the interference [34].

Other possible enhancements include combination of adaptive modulation and coding into cooperative protocols [35], employing coding in cooperative protocols [36] and power and scheduling design for selection of a group of active retransmitting nodes [37]. In [12], a detection technique for wireless networks where perfect synchronization of users are not possible has been proposed. This design mimics an equalization technique employed in a frequency-selective fading channel.

11.3 Resource Allocation for Single-Carrier Systems

Resource allocation for single-user cooperative wireless networks has received significant research interest over the past few years [16–20]. From the information theoretic viewpoint, only capacity upper and lower bounds can be derived in the literature, because the capacity of a general relay channel is still unknown. In [20], the capacity lower and upper bounds for different cooperation strategies including time-division relaying and compress-and-forward relaying have been derived. Optimal power allocation schemes that aim at maximizing these capacity bounds have also been studied. In [18], the capacity bounds for parallel relay channels with degraded subchannels have been derived and optimized through power allocation.

For the practical AF and DF protocols, optimal power allocation and relay selection methods have been developed in different existing works such as [13] and [17]. In the following, we discuss some important resource allocation designs for single-carrier wireless systems.

11.3.1 Power Allocation for AF Relaying

Assume that the source transmission power \( P_s \) is fixed. With the SNR expression for the AF protocol with \( M \) relays given in (11.4), the SNR maximization problem under the total relay power constraint can be written as

\[
\max_{\{P_i\}} \gamma_s h_{sd} + \sum_{i=1}^{M} \frac{\gamma_s h_{sr_i} \gamma_r h_{rd} h_{ri}}{\gamma_s h_{sr_i} + \gamma_r h_{rd} + 1}
\]

subject to

\[
\sum_{i=1}^{M} P_{ri} \leq P_R.
\]

To derive the optimal solution of this optimization problem, we introduce some new notations and transform the considered problem into a more tractable form. Let \( N_d \) and \( N_r_i \) denote the powers of the white Gaussian noise at the destination and relay \( i \), respectively. Moreover, we also define the normalized power channel-gains as \( \alpha_{sr} = h_{sr}/N_r, \alpha_{rd} = h_{rd}/N_d, \) and \( \alpha_{sr} = h_{rd}/N_d \). Then, the achieved SNR at the destination...
node can be rewritten as
\[\gamma_d = \gamma_s h_{sd} + \sum_{i=1}^{M} \gamma_s h_{sr_i} + \gamma_r h_{rd} + 1\]
\[= P_s \alpha_{s,d} + \frac{P_s P_r \alpha_{s,r_d} \alpha_{r_d}}{P_s \alpha_{s,r_i} + P_r \alpha_{r_d} + 1}\]
\[= P_s \left( \alpha_{s,d} + \sum_{i=1}^{M} \alpha_{s,r_i} \right) - \sum_{i=1}^{M} \frac{P_s^2 \alpha_{s,r_i} + P_s \alpha_{s,r_d} + 1}{P_s \alpha_{s,r_i} + P_r \alpha_{r_i} + 1}.\]

Therefore, the SNR maximization problem is equivalent to the following problem:
\[
\min_{\{P_r\}} \sum_{i=1}^{M} \frac{P_s^2 \alpha_{s,r_i} + P_s \alpha_{s,r_d} + 1}{P_s \alpha_{s,r_i} + P_r \alpha_{r_i} + 1}
\]
\[
s.t. \sum_{i=1}^{M} P_r \leq P_R.\]

It can be verified that the objective function (11.50) is convex with respect to relay powers \(P_r\). Therefore, the optimal power allocation can be derived from the Karush-Kuhn-Tucker (KKT) optimality conditions. Toward this end, the Lagrangian obtained by relaxing the relay power constraint can be written as
\[
L \left( \{P_r\}, \lambda \right) = \sum_{i=1}^{M} \frac{P_s^2 \alpha_{s,r_i} + P_s \alpha_{s,r_d} + 1}{P_s \alpha_{s,r_i} + P_r \alpha_{r_i} + 1} + \frac{1}{\lambda^2} \left( \sum_{i=1}^{M} P_r - P_R \right)
\]
where the Lagrange multiplier is set equal to \(\frac{1}{\lambda^2}\). From the KKT conditions, we can derive and set the derivative with respect to each relay power \(P_r\) to zero as [13]
\[
-\alpha_{r_i} \left( \frac{P_s^2 \alpha_{s,r_i} + P_s \alpha_{s,r_d} + 1}{P_s \alpha_{s,r_i} + P_r \alpha_{r_i} + 1} \right) + \frac{1}{\lambda^2} = 0.
\]

From this, we obtain
\[
P_r = \sqrt{\frac{P_s^2 \alpha_{s,r_i} + P_s \alpha_{s,r_d} + 1}{\alpha_{r_i} \lambda - \alpha_{r_d}}}.
\]

Since the relay power must be non-negative, it can be verified that the optimal relay power is
\[
P_r = \left( \sqrt{\frac{P_s^2 \alpha_{s,r_i} + P_s \alpha_{s,r_d} + 1}{\alpha_{r_i} \lambda - \alpha_{r_d}}} \right)^{+}
\]
where \((x)^{+}\) denotes the projection of \(x\) to the set of non-negative real numbers and \(\lambda\) can be determined so that the relay power constraint is met with equality.

Suppose that we impose additional individual relay power constraints \(P_r \leq P_{r_i}^{\text{max}}\) beside the considered total relay power constraint. Since the objective function is monotonically decreasing and convex, the optimal relay power allocation with both total and
individual relay power constraint is

\[ P_{r_i} = \left( \frac{P_s^2 \alpha_{s,r_i}^2 + P_s \alpha_{s,r_i} \lambda}{\alpha_{r_i,d}} - \frac{P_s \alpha_{s,r_i} + 1}{\alpha_{r_i,d}} \right)^{p_{r_i}^{\max}} 0 \]

where \((.,)_0^{p_{r_i}^{\max}}\) denotes the projection of relay power \(P_{r_i}\) to the \([0, P_{r_i}^{\max}]\) interval.

### 11.3.2 Power Allocation for Selection AF Relaying

For the selection AF scheme, the optimal relay should be chosen to maximize the metric given in (11.9). For simplicity, suppose that relay \(i\) is chosen for certain given channel gains. Moreover, we assume that the total source and relay power is \(P_T\). The power allocation then requires us to determine the fractions \(\rho\) and \(1 - \rho\) of the total power for the source and the (selected) relay, i.e., the source and relay powers are \(\rho P_T\) and \((1 - \rho)P_T\), respectively.

Let us define the following quantities \(A_0 = \alpha_{s,d} P_T, A_i = \alpha_{s,r_i} P_T,\) and \(B_i = \alpha_{r_i,d} P_T\). Then the achieved SNR at the destination node can be rewritten as

\[ \gamma_d = A_0 \rho + \frac{A_i B_i (1 - \rho)}{A_i \rho + B_i (1 - \rho) + 1}. \] (11.57)

Our goal is then to determine the optimal \(\rho \in [0, 1]\) to maximize the SNR \(\gamma_d\). Toward this end, we derive and set the derivative of \(\gamma_d\) with respect to \(\rho\) to zero. This yields [13]

\[ D_i (B_i - A_i) \rho^2 - 2D_i (B_i + 1) \rho + (A_0 B_i + A_i B_i + A_0) (B_i + 1) = 0 \] (11.58)

where \(D_i = A_i B_i + A_0 B_i - A_0 A_i\).

The above equation has no real number solution when \(D_i < 0\). Moreover, the derivative of the objective function does not change sign in this case. Since we have

\[ \gamma_d(\rho = 1) = A_0 > \gamma_d(\rho = 0) = 0 \] (11.59)

the objective function is monotonically increasing. Therefore, we have the optimal value of \(\rho\) equal to 1, i.e., \(\rho^* = 1\).

If \(D_i > 0\), it can be shown that one solution of the above solution is outside the interval of interest \((0, 1)\). Thus, the optimal \(\rho\) must be equal to the other solution, which is

\[ \rho^* = \frac{B_i + 1}{B_i - A_i} - \frac{C_i}{D_i (B_i - A_i)} \] (11.60)

where \(C_i = A_i B_i (A_i + 1) (B_i + 1)\). If this value is outside the interval \((0, 1)\), then the optimal value of \(\rho\) equal to 1, i.e., \(\rho^* = 1\).

In summary, the optimal value of \(\rho\) corresponding to the optimal power allocation can be written as

\[ \rho^* = \begin{cases} 1, & \text{if } D_i < 0 \\ \min \left( 1, \frac{B_i + 1}{B_i - A_i} - \frac{C_i}{D_i (B_i - A_i)} \right), & \text{if } D_i > 0. \end{cases} \] (11.61)
Joint Relay Selection and Power Allocation for ANC Two-Way Relaying

Since there are two traffic flows in opposite directions between two nodes $s_1$ and $s_2$ in two-way relaying systems, relay selection should be performed to enhance the performance of both traffic flows. One potential design objective is to maximize the minimum achieved SNR of the two flows. Specifically, for the TDBC- and ANC-based two-way relaying systems with fixed power allocation, relay selection can be performed as

$$\max_i \min \{ \text{SNR}_{1,i}, \text{SNR}_{2,i} \}$$

(11.62)

where the $\text{SNR}_{1,i}$ and $\text{SNR}_{2,i}$ for relay $r_i$ have been derived in the previous sections.

We now discuss the more general joint relay selection and power allocation optimization for the ANC two-way relaying system [45]. Specifically, we will derive the optimal power allocation for two source nodes and one selected relay based on which we can optimize the relay selection. Without loss of generality, we assume the unit power for the Gaussian noise at both source nodes $s_1$ and $s_2$ and any relay $r_i$. Moreover, we assume the reciprocal channel-gains between any pair of nodes.

Under these assumptions, the signals obtained at nodes $s_1$ and $s_2$ after self-interference removal given in (11.35) and (11.36) can be rewritten as

$$\hat{y}_{s_1,r_i} = \varphi_r a_{r_1s_1} a_{r_2s_2} x_{s_2} + \varphi_r a_{r_2s_2} z_{r_i} + v_{s_1}$$

(11.63)

$$\hat{y}_{s_2,r_i} = \varphi_r a_{r_2s_2} a_{r_1s_1} x_{s_1} + \varphi_r a_{r_1s_1} z_{r_i} + v_{s_2}$$

(11.64)

where we have used $a_{r_1s_1}$ and $a_{r_2s_2}$ to denote the channel-gains between nodes $s_1$, $s_2$ and relay $r_i$, respectively. Similarly, the corresponding power channel gains are defined as $h_{r_1s_1} = |a_{r_1s_1}|^2$ and $h_{r_2s_2} = |a_{r_2s_2}|^2$.

Note that with relay power optimization, we do not set the power amplification factor $\varphi_{r_i}$ as in (11.30), but $\varphi_{r_i}$ is now an optimization variable. The relay power with the amplification factor $\varphi_{r_i}$ is equal to

$$P_{r_i} = |\varphi_{r_i}|^2 (h_{r_1s_1} P_{s_1} + h_{r_2s_2} P_{s_2} + 1) .$$

(11.65)

Therefore, the total consumed power of all nodes is

$$P_{T,i} = P_{s_1} (1 + |\varphi_{r_i}|^2 h_{r_1s_1}) + P_{s_2} (1 + |\varphi_{r_i}|^2 h_{r_2s_2}) + |\varphi_{r_i}|^2 .$$

(11.66)

Let us define $q_{r_i} = h_{r_1s_1} h_{r_2s_2}$. Then the SNRs of the signals $\hat{y}_{s_1,r_i}$ and $\hat{y}_{s_2,r_i}$ can be expressed as

$$\text{SNR}_{1,i} = \frac{P_{s_2} q_{r_i} |\varphi_{r_i}|^2}{h_{r_1s_1} |\varphi_{r_i}|^2 + 1}$$

(11.67)

$$\text{SNR}_{2,i} = \frac{P_{s_2} q_{r_i} |\varphi_{r_i}|^2}{h_{r_2s_2} |\varphi_{r_i}|^2 + 1} .$$

(11.68)
The joint relay selection and power allocation optimization problem can then be stated as

$$
\max_{P_{s1}, P_{s2}, \varphi_{r_i}, t} \min \{\text{SNR}_{1,i}, \text{SNR}_{2,i}\} \tag{11.69}
$$

s.t. \hspace{1em} P_{T,i} \leq P_T^{\max} \tag{11.70}

where $P_T^{\max}$ denotes the maximum total power. This problem can be solved by first optimizing the power allocation for nodes $s_1, s_2$ and relay $r_i$, and then optimizing the relay selection. The power allocation problem for a given relay $r_i$ can be stated as

$$
\max_{P_{s1}, P_{s2}, \varphi_{r_i}, t} \min \{\text{SNR}_{1,i}, \text{SNR}_{2,i}\} \tag{11.71}
$$

s.t. \hspace{1em} P_{T,i} \leq P_T^{\max} \tag{11.72}

It can be verified that we must have $\text{SNR}_{1,i} = \text{SNR}_{2,i}$ at optimality. This is because if $\text{SNR}_{1,i} > \text{SNR}_{2,i}$, then we can slightly reduce $P_{s2}$ and increase $P_{s1}$ to increase the objective function while not violating the total power constraint [29]. Substituting (11.67) and (11.68) to the condition $\text{SNR}_{1,i} = \text{SNR}_{2,i}$ yields

$$
P_{s2} \left( h_{r_s} |\varphi_{r_i}|^2 + 1 \right) = P_{s1} \left( h_{r_s} |\varphi_{r_i}|^2 + 1 \right). \tag{11.73}
$$

Substituting this result in (11.66), the power constraint (11.72) can be rewritten as

$$
2P_{s1} \left( 1 + |\varphi_{r_i}|^2 h_{r_s} |\varphi_{r_i}|^2 \right) + |\varphi_{r_i}|^2 \leq P_T^{\max}. \tag{11.74}
$$

By introducing an auxiliary variable $t$, the optimization problem (11.71) and (11.72) can be reformulated as

$$
\max_{P_{s1}, P_{s2}, \varphi_{r_i}, t} \hspace{1em} t \tag{11.75}
$$

s.t. \hspace{1em} \text{SNR}_{1,i} \geq t, \hspace{1em} \text{SNR}_{2,i} \geq t \tag{11.76}

$$
2P_{s1} \left( 1 + |\varphi_{r_i}|^2 h_{r_s} |\varphi_{r_i}|^2 \right) + |\varphi_{r_i}|^2 \leq P_T^{\max}. \tag{11.77}
$$

Using the optimality condition $\text{SNR}_{1,i} = \text{SNR}_{2,i}$, this optimization problem is equivalent to

$$
\max_{P_{s1}, P_{s2}, \varphi_{r_i}, t} \hspace{1em} t \tag{11.78}
$$

s.t. \hspace{1em} \text{SNR}_{2,i} = \frac{P_{s1} q_{r_i} |\varphi_{r_i}|^2}{h_{r_s} |\varphi_{r_i}|^2 + 1} = t \tag{11.79}

$$
2P_{s1} \left( 1 + |\varphi_{r_i}|^2 h_{r_s} |\varphi_{r_i}|^2 \right) + |\varphi_{r_i}|^2 \leq P_T^{\max}. \tag{11.80}
$$

Combining constraints (11.79) and (11.80), this problem is equivalent to

$$
\max_{P_{s1}, P_{s2}, \varphi_{r_i}, t} \hspace{1em} t \tag{11.81}
$$

s.t. \hspace{1em} \frac{2t \left( 1 + |\varphi_{r_i}|^2 h_{r_s} |\varphi_{r_i}|^2 \right) \left( 1 + |\varphi_{r_i}|^2 h_{r_s} |\varphi_{r_i}|^2 \right)}{q_{r_i} |\varphi_{r_i}|^2} + |\varphi_{r_i}|^2 \leq P_T^{\max}. \tag{11.82}
$$
It can be verified that the power constraint (11.80) must be met with equality at optimality. This is because otherwise we can increase $t$ in the left-hand side of (11.80) while still maintaining this constraint. This equality constraint results in

$$t = \frac{(p_{\text{max}} - |\varphi_r|^2) q_r |\varphi_r|^2}{2(1 + |\varphi_r|^2 h_{r,s_1}) (1 + |\varphi_r|^2 h_{r,s_2})}.$$  \hfill (11.83)

Since $t$ does not depend on the phase of $\varphi_r$ according to (11.83), the optimization problem (11.81) and (11.82) becomes equivalent to

$$\max_{|\varphi_r|^2 \geq 0} \frac{(p_{\text{max}} - |\varphi_r|^2) q_r |\varphi_r|^2}{(1 + |\varphi_r|^2 h_{r,s_1}) (1 + |\varphi_r|^2 h_{r,s_2})}.$$ \hfill (11.84)

where we only need to determine the optimal $|\varphi_r|^2$. The optimal value of $|\varphi_r|^2$ can be determined by setting the derivative of the objective function in (11.84) to zero. Then we can use the following relations given by the total power constraints to calculate the optimal powers for nodes $s_1$ and $s_2$:

$$P_{\text{max}}^* = 2P_{s_1} (1 + |\varphi_r|^2 h_{r,s_1}) + |\varphi_r|^2$$  

$$= 2P_{s_2} (1 + |\varphi_r|^2 h_{r,s_2}) + |\varphi_r|^2.$$ \hfill (11.85) \hfill (11.86)

After some manipulation, we can obtain the optimal power allocation for nodes $s_1$ and $s_2$ as follows:

$$P_{s_1}^* = \frac{P_{\text{max}}^* \sqrt{1 + P_{\text{max}}^* h_{r,s_2}}}{2\sqrt{1 + P_{\text{max}}^* h_{r,s_1} + 2\sqrt{1 + P_{\text{max}}^* h_{r,s_2}}}} \hfill (11.87)$$

$$P_{s_2}^* = \frac{P_{\text{max}}^* \sqrt{1 + P_{\text{max}}^* h_{r,s_1}}}{2\sqrt{1 + P_{\text{max}}^* h_{r,s_1} + 2\sqrt{1 + P_{\text{max}}^* h_{r,s_2}}}}.$$ \hfill (11.88)

Interestingly, we have $P_{s_1}^* + P_{s_2}^* = \frac{P_{\text{max}}^*}{2}$; therefore, the optimal power for relay is $P_{r_i}^* = \frac{P_{\text{max}}^*}{2}$.

After applying this optimal power allocation for each potential relay $r_i$, we can then compare the achieved SNR due to each relay, and the optimal relay selection resulting in the maximum SNR can be determined.

## 11.4 Resource Allocation for Multi-Carrier Systems

In this section, we discuss the resource allocation for the multi-carrier OFDM wireless systems where one source node $s$ communicates with its destination node $d$ with the assistance of multiple relays. We consider the general resource allocation design where each relay can be assigned multiple subcarriers to assist the transmissions on the resource-relay and relay-destination links. Moreover, a signal transmitted over the source-relay link by using one particular subcarrier can be forwarded to the destination over the relay-destination link by using a different subcarrier. Therefore, resource allocation optimization involves pairing of the subcarriers, assignment of subcarrier pairs
to relays, and power allocation. A survey on the resource allocation methods for relay-enhanced OFDMA cellular networks can be found in [41].

11.4.1 Resource Allocation for AF Multi-Carrier Wireless Networks

Suppose that a source node $s$ communicates with a destination node $d$ with the help of $M$ relays, $r_1, r_2, \ldots, r_M$. Furthermore, it is assumed that there are $K$ subcarriers, which can be shared by the links, and the wireless channel on each subcarrier experiences the flat fading. Moreover, suppose that relay $i$ is associated with subcarriers $k$ and $k'$ to transmit signals on the source-relay and relay-destination links in two time slots, respectively. To avoid strong interference, each subcarrier pair $(k, k')$ is assigned to only one relay. The subcarrier pairing for the two hops of the relaying system is illustrated in Figure 11.2 where the numbers inside boxes represent the indices of subcarriers allocated to different communication links.

Let $N_d$ and $N_{ri}$ denote the noise powers of the white Gaussian noise at the destination and relay $i$, respectively. In addition, the (power) channel-gains of source-relay, relay-destination, and source-destination links corresponding to relay $i$ and subcarrier $k$ are denoted as $a_{k,i,1} (h_{k,i,1} = |a_{k,i,1}|^2)$, $a_{k,i,2} (h_{k,i,2} = |a_{k,i,2}|^2)$, and $a_{k,3} (h_{k,3} = |a_{k,3}|^2)$, respectively. Moreover, the powers used by source and relay $i$ for transmissions over subcarrier pair $(k, k')$ are represented by $P_{k,i,1}$ and $P_{k',i,2}$, respectively. For convenience, we also define $\alpha_{k,i,1} = h_{k,i,1}/N_r$, $\alpha_{k,i,2} = h_{k,i,2}/N_d$, and $\alpha_{k,3} = h_{k,3}/N_d$. In the following, the system model and resource allocation designs for the popular AF and DF relaying protocols are discussed.

The end-to-end communication rate corresponding to relay $i$ and subcarrier pair $(k, k')$ under the AF relaying strategy, which is a special case of (11.4) with a single relay, can be written as

$$R_{k,k',i} = \frac{1}{2} \log \left( 1 + \frac{\alpha_{k,i,1} \alpha_{k',i,2} P_{k,i,1} P_{k',i,2}}{\alpha_{k,i,1} P_{k,i,1} + \alpha_{k',i,2} P_{k',i,2} + 1} \right).$$  \tag{11.89}
It can be verified that $R_{k,k',i}$ is not jointly concave in $P_{k,i,1}$ and $P_{k',i,2}$. To make the optimization tractable, the rate is approximated as follows:

$$R_{k,k',i} \approx \frac{1}{2} \log \left( 1 + \alpha_{k,i,1} \alpha_{k',i,2} P_{k,i,1} P_{k',i,2} / (\alpha_{k,i,1} P_{k,i,1} + \alpha_{k',i,2} P_{k',i,2}) \right) .$$  

(11.90)

This approximation is tight in the high SNR regime, and it has been commonly adopted in the resource allocation literature.

We will jointly optimize the subcarrier pairing and assignment to relays as well as power allocation for sum rate maximization. Specifically, the dual-based resource allocation algorithm proposed in [42] will be described in the following. Toward this end, let us define binary variables $\rho_{k,k'} \in \{0, 1\}$ capturing the subcarrier pairing decisions where $\rho_{k,k'} = 1$ if subcarriers $k$ and $k'$ are paired and $\rho_{k,k'} = 0$, otherwise. Since each subcarrier can be paired only with exactly one other subcarrier, the subcarrier pairing variables $\rho_{k,k'}$ are subject to the following constraints:

$$\sum_{k=1}^{K} \rho_{k,k'} = 1, \quad \forall k', \quad \sum_{k'=1}^{K} \rho_{k,k'} = 1, \quad \forall k.$$  

(11.91)

In addition, let $t_{k,k',i} \in \{0, 1\}$ represent the assignment of subcarrier pair $(k, k')$ for relay $i$ where $t_{k,k',i} = 1$ if the subcarrier pair $(k, k')$ is assigned to relay $i$ and $t_{k,k',i} = 0$, otherwise. Because each subcarrier pair can be assigned to only one relay to avoid the strong interference, we have the following constraints:

$$\sum_{i=1}^{M} t_{k,k',i} = 1, \quad \forall k, k'.$$  

(11.92)

We will consider the resource allocation design under separate power constraints for the source and individual relays. Specifically, the source power constraint can be stated as

$$\sum_{k=1}^{K} \sum_{i=1}^{M} P_{k,i,1} \leq P_S$$  

(11.93)

where $P_S$ is the total power available at the source node. Moreover, the power constraint at each relay $i$ can be expressed as

$$\sum_{k'=1}^{K} P_{k',i,2} \leq P_{R,i}, \quad \forall i$$  

(11.94)

where $P_{R,i}$ denotes the total power of relay $i$.

The sum-rate maximization problem can be then formulated as

$$\max_{\rho, t, P} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i=1}^{M} 2 \rho_{k,k'} t_{k,k',i} R_{k,k',i}$$  

s.t. (11.91), (11.92), (11.93), (11.94)  

(11.95)
where the factor of two is introduced in the objective function to cancel out the ratio 1/2 in the rate expression. Moreover, \( \rho, t, \) and \( P \) are the vectors whose elements capture the corresponding optimization variables \( \rho_{k,k'}, t_{k,k',i}, \) and \( P_{k,i,1}, P_{k',i,2} \), respectively.

This is a mixed integer non-linear optimization problem, which is difficult to solve in general. It has been shown in [38] that by solving this type of resource allocation in the dual domain, i.e., solving the dual optimization problem, we can achieve the asymptotically optimal solution if the number of subcarriers is sufficiently large. This is because the duality gap tends to zero as the number of subcarriers tends to infinity, for which the so-called time-sharing property is satisfied.

Moreover, it is known from optimization theory that if the duality gap is zero, then the solution obtained by solving the dual optimization problem also gives the optimal solution of the original primal optimization problem [39, 40]. We describe how the dual approach can be employed to solve this problem in the following.

We first define the Lagrangian by relaxing all power constraints as follows:

\[
L(\rho, t, P, \beta) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i=1}^{M} R_{k,k',i} + \beta_S \left( P_S - \sum_{k=1}^{K} \sum_{i=1}^{M} P_{k,i,1} \right) \\
+ \sum_{i=1}^{M} \beta_{R,i} \left( P_{R,i} - \sum_{k'=1}^{K} P_{k',i,2} \right) 
\]  

(11.96)

where \( \beta = (\beta_S, \beta_{R,1}, \beta_{R,2}, \ldots, \beta_{R,M}) \geq 0 \) represents the vector whose elements are the Lagrange multipliers associated with the corresponding power constraints. Let \( D \) represent the set of subcarrier pairs \( \rho \) and subcarrier-relay assignments \( t \) satisfying the constraints (11.91) and (11.92) and \( P \) denote the set of power vectors \( P \) satisfying constraints (11.93) and (11.94). Then the dual function can be defined as

\[
g(\beta) = \max_{(\rho, t) \in D, P \in P} L(\rho, t, P, \beta). 
\]  

(11.97)

The dual optimization problem can be then expressed as

\[
\min_{\beta \geq 0} g(\beta). 
\]  

(11.98)

The Lagrange multipliers captured in the vector \( \beta \), which are also called dual variables, are the optimization variables of the dual problem.

The dual-based resource allocation algorithm iteratively optimizes the Lagrangian function in (11.97) for a given dual point \( \beta \) and then updates the dual variables \( \beta \) toward solving the dual problem in (11.98) by using the subgradient algorithm. These iterative primal-dual updates converge to the optimal solution if the duality gap is zero [39], [40]. In the following, we describe how to solve problems (11.97) and (11.98) in detail.
11.4 Resource Allocation for Multi-Carrier Systems

Optimization of Primal Variables

We first describe how to maximize the Lagrangian function at a given dual point $\beta$. The dual function given in (11.97) can be rewritten as follows:

$$g(\beta) = \max_{(\rho, t) \in D, P \in P} L(\rho, t, P, \beta)$$

$$= \max_{(\rho, t) \in D, P \in P} \sum_{k=1}^{K} \sum_{k'}^{K} \sum_{i=1}^{M} \rho_{k,k'} t_{k,k',i} L_{k,k',i}(\beta) + \beta_S P_S + \sum_{i=1}^{M} \beta_{R,i} P_{R,i} \tag{11.100}$$

where

$$L_{k,k',i}(\beta) = \log \left(1 + \alpha_{k,1} \rho_{k,k',i} + \frac{\alpha_{k,1} \rho_{k,k',i} P_{k,1,1} P_{k',1,2}}{\alpha_{k,1} P_{k,1,1} + \alpha_{k',1,2} P_{k',1,2}}\right) - \beta_S P_{k,1,1} - \beta_{R,i} P_{k',1,2}. \tag{11.101}$$

In order to solve problem (11.100), we consider the optimization of power allocation, relay assignment for given a subcarrier pairing, and optimization of subcarrier pairing separately in the following.

Power Allocation for Given Subcarrier Pairing and Relay Assignment

We study the optimal power allocation solution $P^*$ for a given subcarrier pairing and relay assignment solution $(\rho, t)$. Specifically, suppose that the subcarrier pair $(k, k')$ is valid, and it is assigned to relay $i$, i.e., we have $\rho_{k,k'} t_{k,k',i} = 1$. Then the power allocation problem for this subcarrier-pairing and relay assignment solution can be written as

$$\max_{P_{k,1,1}, P_{k',1,2}} L_{k,k',i}(\beta) \tag{11.102}$$

s.t. $P_{k,1,1} \geq 0$, $P_{k',1,2} \geq 0. \tag{11.103}$

Since the objective function of this problem is concave, the optimal solution can be determined from the Karush-Kuhn-Tucker (KKT) optimality conditions. The optimal power allocation is characterized in the following theorem.

**Theorem 68** The optimal power allocation of problem (11.102)–(11.103) can be expressed as

$$P^*_{k,1,1} = \begin{cases} c_{k,k',i} P^*_{k',1,2}, & \text{if } P^*_{k',1,2} > 0 \\ \left(\frac{1}{\beta_S} - \frac{1}{\alpha_{k,1}}\right)^+, & \text{if } P^*_{k',1,2} = 0 \end{cases} \tag{11.104}$$
where
\[
P_{k',i,2}^{*} = \begin{cases} 
\frac{\alpha_{k,i,1}^2 \alpha_{k',i,2}^2 + (\alpha_{k,i,1} \alpha_{k',i,2})^2}{c_{k',k,i} \beta_S (\alpha_{k,i,1} c_{k',k,i} + \alpha_{k,i,1} \alpha_{k',i,2} + \alpha_{k',i,2} \alpha_{k,i,1})} , & \text{if } \alpha_{k',i,2} \beta_S > \alpha_{k,3} \beta_{R,i} \\
0, & \text{if } \alpha_{k',i,2} \beta_S \leq \alpha_{k,3} \beta_{R,i}
\end{cases} 
\] (11.105)

Proof To derive the KKT conditions, we set the derivative of \( L_{k,k',i} \) with respect to each optimization variable to zero. For simplicity, we omit the indices \( k, k', \) and \( i \) in all notations \( P_{k,i,1}, P_{k',i,2}, \alpha_{k,i,1}, \alpha_{k',i,2}, \alpha_{k,3}, \) and \( \beta_{R,i}:
\[
\frac{\partial L_{k,k',i}}{\partial P_1} = -\beta_S + \frac{\alpha_3 \left( \alpha_1 P_1 + \alpha_2 P_2 \right)^2 + \alpha_1 \alpha_2^2 P_2^2}{(\alpha_1 P_1 + \alpha_2 P_2)} = 0 
\] (11.107)
\[
\frac{\partial L_{k,k',i}}{\partial P_2} = -\beta_R + \frac{\alpha_2^2 \alpha_2 P_1^2}{(\alpha_1 P_1 + \alpha_2 P_2)} = 0. 
\] (11.108)

We first consider the case where both \( P_1 \) and \( P_2 \) are positive. From these two equations, we obtain
\[
P_1^* = c P_2^* 
\] (11.109)

where
\[
c = \frac{\alpha_2}{\alpha_1 (\alpha_2 \beta_S - \alpha_3 \beta_R)} \left( \sqrt{\beta_R (\alpha_1 \alpha_2 \beta_S - \alpha_1 \alpha_3 \beta_R + \alpha_2 \alpha_3 \beta_S)} + \alpha_3 \beta_R \right). 
\] (11.110)

To ensure that \( P_1^* > 0 \) we must have \( \alpha_2 \beta_S > \alpha_3 \beta_R \). Moreover, substituting the result in (11.109) into (11.107) yields
\[
P_2^* = \frac{\alpha_1 \alpha_2^2 + (\alpha_3 - \beta_S) (\alpha_1 c + \alpha_2)^2}{c \beta_S (\alpha_1 c + \alpha_2) (\alpha_3 \alpha_1 c + \alpha_3 \alpha_2 + \alpha_1 \alpha_2)}. 
\] (11.111)

If the value of \( P_2^* \) given in (11.111) is negative, we should set \( P_2^* = 0 \). Moreover, if \( \alpha_2 \beta_S \leq \alpha_3 \beta_R \), then it can be proved that \( P_2^* = 0 \). For both cases, the relay is not employed (i.e., the direct transmission over the source-destination link is adopted); therefore, the standard water-filling approach can be employed for power allocation at the source node,
which yields

\[ P^* = \left( \frac{1}{\beta_S} - \frac{1}{\alpha_3} \right)^+. \] (11.112)

This completes the proof of the theorem.

**Optimal Relay Assignment for Given Subcarrier Pairing Solution**

Substitute the optimal powers in (11.104) and (11.105) into (11.101), and then into (11.100), and we obtain

\[
g(\beta) = \max_{(\rho, i) \in D} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i=1}^{M} \rho_{k,k'} t_{k,k',i} H_{k,k',i}(\beta)
+ \beta_S P_S + \sum_{i=1}^{M} \beta_{R,i} P_{R,i} \tag{11.113}
\]

where \( H_{k,k',i}(\beta) \) depends on whether direct or relay-based communication is employed, which is specified in the following. Specifically, if the direct communication is employed, then

\[
H_{k,k',i}(\beta) = \left[ \log \left( \frac{\alpha_{k,3}}{\beta_S} \right) \right]^+ - \beta_S \left( \frac{1}{\beta_S} - \frac{1}{\alpha_{k,3}} \right)^+. \tag{11.114}
\]

Otherwise, if the AF relaying communication is employed, then we have

\[
H_{k,k',i}(\beta) = \log \left( \frac{\alpha_{k,3} (\alpha_{k,i,1} c_{k,k',i} + \alpha_{k',i,2})^2 + \alpha_{k,i,1} \alpha_{k',i,2}^2}{\beta_S \left( \alpha_{k,i,1} c_{k,k',i} + \alpha_{k',i,2} \right)^2} \right) - \left( \beta_S c_{k,k',i} + \beta_{R,i} \right)
\times \left( \frac{\alpha_{k,i,1} \alpha_{k',i,2}^2 + (\alpha_{k,3} - \beta_S) (\alpha_{k,i,1} c_{k,k',i} + \alpha_{k',i,2})^2}{c_{k,k',i} \beta_S \left( \alpha_{k,i,1} c_{k,k',i} + \alpha_{k',i,2} \right)^2 (\alpha_{k,3} \alpha_{k,i,1} c_{k,k',i} + \alpha_{k,3} \alpha_{k',i,2} + \alpha_{k,i,1} \alpha_{k',i,2})} \right). \tag{11.115}
\]

More detailed study of the optimization (11.113) suggests the following relay assignment strategy for the given subcarrier pairing solution. Specifically, suppose that \((k, k')\) is a valid subcarrier pair, i.e., \(\rho_{k,k'} = 1\), then we should assign each subcarrier pair \((k, k')\) to relay \(i(k, k')\) as follows:

\[
t_{k,k',i} = \begin{cases} 1, & i = i(k, k') = \arg\max_i H_{k,k',i} \\ 0, & \text{otherwise} \end{cases} \tag{11.116}
\]

Therefore, the weight parameters \(H_{k,k',i}\) play important roles in determining the optimal relay assignment for a given subcarrier pairing solution.
Optimal Subcarrier Pairing
Substituting the result in (11.116) in (11.113) yields

\[ g(\beta) = \max_{\rho} \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} H_{k,k'}, (\beta) + \beta S P_S + \sum_{i=1}^{M} \beta_{R,i} P_{R,i} \quad (11.117) \]

where \( H_{k,k'} = H_{k,k'}, i(k,k') \), and \( i(k,k') = \arg\max_i H_{k,k'}, i \).

In order to determine the optimal subcarrier pairing solution, we define the \( K \times K \) weight matrix \( H = [H_{k,k'}] \). Then problem (11.117) becomes equivalent to

\[ \max_{\rho} \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} H_{k,k'}, i \quad (11.118) \]

s.t. (11.91).

This problem requires that we choose exactly one element in each row and each column of matrix \( H \) to maximize the sum weight. This is a standard linear assignment problem, which can be solved by the Hungarian algorithm [44].

Let \( \pi(k) \) denote the subcarrier index in the second hop, which is optimally paired with subcarrier \( k \) in the first hop. Then the optimal subcarrier pairing solution can be expressed as

\[ \rho_{k,k'}^* = \begin{cases} 1, & \text{if } k' = \pi(k) \\ 0, & \text{otherwise.} \end{cases} \quad (11.120) \]

Combining the optimization in all aforementioned steps, we can obtain the optimal primal variables \( (\rho^*, t^*, P^*) \) for a given dual point \( \beta \).

Optimization of Dual Variables
We now solve the dual optimization problem in (11.98). Because the dual function is always convex according to the optimization theory [40], the dual optimization problem can be solved by the standard sub-gradient algorithm. Specifically, suppose that \( P^*(\beta) \) is the optimal power allocation vector for a given \( \beta \). It can be verified that the sub-gradients of the dual function \( g(\beta) \) can be expressed as

\[ \Delta \beta_S = P_S - \sum_{k=1}^{K} \sum_{i=1}^{M} P_{k,i,1}^* (\beta) \quad (11.121) \]

\[ \Delta \beta_{R,i} = P_{R,i} - \sum_{k'=1}^{K} P_{k',i,2}^* (\beta), \quad \forall i. \quad (11.122) \]

We define the sub-gradient vector as \( \Delta \beta = (\Delta \beta_S, \Delta \beta_{R,1}, \Delta \beta_{R,2}, \ldots, \Delta \beta_{R,M}) \) for convenience. Then the sub-gradient algorithm can be used to update the dual vector in iteration \( l + 1 \) as

\[ \beta^{(l+1)} = \left[ \beta^{(l)} - \epsilon^{(l)} \Delta \beta^{(l)} \right]^+ \quad (11.123) \]
where \( \epsilon^{(t)} = \kappa / \sqrt{t} \) with \( \kappa > 0 \) denotes the diminishing step-size to guarantee the convergence of this algorithm [39].

**Refinement of Power Allocation**

In practice, the number of subcarriers \( K \) is finite, which could lead to a non-zero duality gap. Therefore, the obtained solution \((\rho^* (\beta^*) \, , \, t^* (\beta^*) \, , \, P^* (\beta^*))\) for final dual point \( \beta^* \) may not be feasible. To address this issue, we can fix \( \rho, t, \) and \( \beta \) at the obtained values \( \rho^*, t^*, \) and \( \beta^* \) while re-optimizing the power allocation to achieve a feasible solution.

Let \( \mathcal{S}_i \) denote the set of subcarrier pairs \((k, \pi (k))\) assigned to relay \( i \). Then, by substituting the solution corresponding to \((\rho^* (\beta^*) \, , \, t^* (\beta^*))\) into (11.139), the power allocation solution can be determined by solving the following problem:

\[
\begin{align}
\text{max} & \quad \sum_{i=1}^{M} \sum_{(k, \pi (k)) \in \mathcal{S}_i} 2R_{k, \pi (k), i} \\
\text{s.t.} & \quad \sum_{i=1}^{M} \sum_{(k, \pi (k)) \in \mathcal{S}_i} P_{k, i, 1} \leq P_S \\
& \quad \sum_{(k, \pi (k)) \in \mathcal{S}_i} P_{\pi (k), i, 2} \leq P_{R, i}, \quad \forall i \quad \text{with} \quad \mathcal{S}_i \neq \emptyset.
\end{align}
\]  

(11.124)–(11.126)

Define \( \mathcal{P} \) as the set of all feasible values of \( P_{k, i, 1} \) and \( P_{\pi (k), i, 2} \). The Lagrangian of problem (11.124)–(11.126) can be expressed as

\[
J (P, \eta) = \sum_{i=1}^{M} \sum_{(k, \pi (k)) \in \mathcal{S}_i} \log \left( 1 + \alpha_{k, 3} P_{k, i, 1} + \frac{\alpha_{k, 1} P_{k, i, 1} \alpha_{\pi (k), i, 2} P_{\pi (k), i, 2}}{\alpha_{k, 1} P_{k, i, 1} + \alpha_{\pi (k), i, 2} P_{\pi (k), i, 2}} \right) \\
+ \eta_S \left( P_S - \sum_{i=1}^{M} \sum_{(k, \pi (k)) \in \mathcal{S}_i} P_{k, i, 1} \right) + \sum_{i=1}^{M} \eta_{R, i} \left( P_{R, i} - \sum_{(k, \pi (k)) \in \mathcal{S}_i} P_{\pi (k), i, 2} \right).
\]  

(11.127)

Then the dual function can be defined as

\[
g (\eta) = \max_{P \in \mathcal{P}} J (P, \eta)
\]  

(11.128)

where \( \eta = (\eta_S, \eta_{R, 1}, \eta_{R, 2}, \ldots, \eta_{R, M}) \) denotes the vector of Lagrange multipliers associated with different power constraints. The dual optimization problem can be written as

\[
\min_{\eta \geq 0} g (\eta).
\]  

(11.129)

Applying the KKT conditions by following a similar procedure as above for a given dual point \( \eta \), the optimal power allocation can be expressed as

\[
P_{k, i, 1}^* = \begin{cases} 
\frac{c_{k, \pi (k), i} P_{\pi (k), i, 2}^*}{\eta_S - \frac{1}{\alpha_{k, 3}}}, & \text{if } P_{\pi (k), i, 2}^* > 0 \\
\frac{1}{\eta_S - \frac{1}{\alpha_{k, 3}}}, & \text{if } P_{\pi (k), i, 2}^* = 0 
\end{cases}
\]  

(11.130)
Resource Allocation in Relay-Based Networks

The resource allocation algorithm for multi-carrier and multi-relay AF relaying system is summarized in Algorithm 10. This algorithm performs the iterative primal-dual

Algorithm 10 Resource allocation for multi-relay multi-carrier AF relaying

1. Initialize the non-negative dual variables $\beta^{(0)}$ and iteration index $l \leftarrow 0$.
2. repeat
3. Compute $H_{k,k',i}$ as in (11.114) and (11.115) by using the optimal power allocation solution given in (11.104) and (11.105).
4. Obtain the optimal subcarrier pairing solution $\rho^*$ from (11.118)–(11.119) and the optimal assignment of subcarrier pairs to relays $t^*$ according to (11.116).
5. Update $\beta^{(l)}$ by using (11.123) then increase the iteration index $l \leftarrow l + 1$.
6. until convergence of $\beta$
7. Fix the variables $\rho$ and $t$ at the obtained values $\rho^*$ and $t^*$. Then solve the problem (11.124)–(11.126) to obtain the final feasible power allocation $P^*$.

where

\[
P^*_{\pi(k),i,2} = \begin{cases} 
\frac{\alpha_{\pi(k),i,2}}{c_{k,\pi(k),i}\eta_S(\alpha_{\pi(k),i,2}\eta_S + \alpha_{k,3}\eta_{R,i})^2} + \frac{(\alpha_{k,3} - \eta_S)(\alpha_{k,1} + \alpha_{k,2})}{c_{k,\pi(k),i}} & \text{if } \alpha_{\pi(k),i,2}\eta_S > \alpha_{k,3}\eta_{R,i} \\
0, & \text{if } \alpha_{\pi(k),i,2}\eta_S \leq \alpha_{k,3}\eta_{R,i}
\end{cases}
\]

(11.131)

The refined power allocation solution can be achieved by iteratively updating the power allocation solution using (11.130) and (11.131) for a given dual point $\eta^{(l)}$ and then updating the dual variables according to (11.132)–(11.133). The power allocation solution obtained at convergence $P^*$ from this iterative primal-dual algorithm will be adopted in the final solution.

The resource allocation algorithm for multi-carrier and multi-relay AF relaying system is summarized in Algorithm 10. This algorithm performs the iterative primal-dual

The Lagrange multipliers $\eta_S$ and $\eta_{R,i}$ can be determined to solve the dual problem (11.129) by using the iterative subgradient algorithm similarly with (11.123) as follows:

\[
\eta^{(l+1)}_S = \left[ \eta^{(l)}_S - \epsilon^{(l)} \left( P_S - \sum_{i=1}^M \sum_{(k,\pi(k)) \in S_i} P_{k,i,1} \right) \right]^+ \quad (11.132)
\]

\[
\eta^{(l+1)}_{R,i} = \left[ \eta^{(l)}_{R,i} - \epsilon^{(l)} \left( P_{R,i} - \sum_{(k,\pi(k)) \in S_i} P_{\pi(k),i,2} \right) \right]^+ , \forall i \quad (11.133)
\]

where $\epsilon^{(l)} = \kappa/\sqrt{l}$ with $\kappa > 0$ denotes the diminishing step-size to guarantee the convergence of this algorithm [39].
updates to solve the resource allocation problems in lines 2–6. Then, the refined power allocation is conducted to achieve a final feasible solution in line 7.

11.4.2 Resource Allocation for DF Multi-Carrier Wireless Networks

We now consider the resource allocation for the DF-based multi-carrier wireless system [43]. We assume that there is a single half-duplex relay, which assists the transmissions on K subcarriers from a source node to a destination node by using the DF relaying strategy. Each subcarrier \( k \) employed on the source-relay link is paired with another subcarrier \( k' \) on relay-destination link, and we again denote such subcarrier pair as \( (k, k') \).

We use similar notations for the channel-gains, transmission, and noise powers as in the previous section, but relay indices are not needed since there is a single relay. Specifically, the power channel-gains of source-relay, relay-destination, and source-destination links corresponding to subcarrier \( k \) are denoted as \( h_{k,1} \), \( h_{k,2} \) and \( h_{k,3} \), respectively. Moreover, the powers used by the source and relay for transmissions over subcarrier pair \( (k, k') \) are represented by \( P_{k,1} \) and \( P_{k',2} \), respectively. We also define \( \alpha_{k,1} = h_{k,1}/N_r \), \( \alpha_{k,2} = h_{k,2}/N_d \), and \( \alpha_{k,3} = h_{k,3}/N_d \) as normalized channel-gains where \( N_r \) and \( N_d \) denote the noise powers at the relay and destination, respectively.

The end-to-end communication rate achieved by the DF relaying over subcarrier pair \( (k, k') \) depends on whether the relay is active (i.e., the relay transmission power is positive) or not. In other words, the achievable rate depends on the operation mode, namely, relay mode or direct mode, employed on each subcarrier pair \( (k, k') \), which can be expressed as

\[
R_{k,k'} = \begin{cases} 
\frac{1}{2} \log (1 + P_{k,1} \alpha_{k,3}), & \text{direct mode} \\
\frac{1}{2} \min \left\{ \log (1 + P_{k,1} \alpha_{k,1}), \log (1 + P_{k,1} \alpha_{k,3} + P_{k',2} \alpha_{k',2}) \right\}, & \text{relay mode}
\end{cases}
\]

where we have assumed that we utilize the whole power budget at the source node in the first time slot while the source is idle in the second time slot if the direct mode is employed. Under the total power constraint for each subcarrier pair \( (k, k') \) \( P_{k,1} + P_{k',2} = P_{k,k'} \), using the relay mode is more advantageous in terms of achievable rate than the direct mode if

\[
\alpha_{k,1} > \alpha_{k,3}, \text{ and } \alpha_{k,2} > \alpha_{k,3}. \tag{11.134}
\]

These conditions basically mean that the source-relay and relay-destination channels are better than the source-destination channel. Moreover, it can be verified that the achievable rate is maximized when the rates achieved by the relay and by the destination are the same in the two communication phases under the relay mode. Under this condition, we can express the achievable end-to-end rate as

\[
R_{k,k'} = \log \left( 1 + \alpha_{k,k'} P_{k,k} \right) \tag{11.135}
\]
where $\alpha_{k,k'}$ is the equivalent channel-gain, which can be expressed as

$$
\alpha_{k,k'} = \begin{cases} 
\frac{\alpha_{k,1}\alpha_{k,2}}{\alpha_{k,1}+\alpha_{k,2}-\alpha_{k,3}}, & \text{relay mode} \\
\alpha_{k,3}, & \text{direct mode}.
\end{cases} \quad (11.136)
$$

And this rate is achieved by the following power allocation:

$$
P_{k,1} = \begin{cases} 
\frac{\alpha_{k,2}}{\alpha_{k,1}+\alpha_{k,2}-\alpha_{k,3}} P_{k,k'}, & \text{relay mode} \\
\rho_{k,k'}, & \text{direct mode}
\end{cases} \quad (11.137)
$$

$$
P_{k,2} = \begin{cases} 
\frac{\alpha_{k,1}-\alpha_{k,3}}{\alpha_{k,1}+\alpha_{k,2}-\alpha_{k,3}} P_{k,k'}, & \text{relay mode} \\
0, & \text{direct mode}
\end{cases} \quad (11.138)
$$

Therefore, the transmission mode, i.e., relay or direct mode, and the optimal achievable rate can be derived for any subcarrier pair $(k, k')$ given the channel state information. Let $\rho_{k,k'} \in \{0, 1\}$ denote the subcarrier pairing decisions as in the previous section. Then the sum-rate maximization problem under the total power constraint can be formulated as

$$
\max_{\rho, P} \sum_{k=1}^{K} \sum_{k'=1}^{K} 2\rho_{k,k'} R_{k,k'}
$$

s.t. (11.140), (11.141), (11.142), (11.143) (11.146)

$$
\rho_{k,k'} \geq 0, \quad \forall k, k'.
$$

where $P$ denotes the total network power, and $\rho$, $P$ denote the vectors whose elements are $\rho_{k,k'}$ and $P_{k,k'}$, respectively.

This is a mixed integer non-linear optimization problem, which is difficult to address. To solve this problem, we relax the integer variables $\rho_{k,k'}$ into the corresponding real variables, which represent the corresponding time-sharing factors of the subcarrier pairs. The relaxed optimization problem can be expressed as

$$
\max_{\rho, P} \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} \log \left(1 + \frac{P_{k,k'}}{\rho_{k,k'}} \frac{\alpha_{k,k'}}{\alpha_{k,3}}\right)
$$

s.t. (11.140), (11.141), (11.142), (11.143) (11.146)

$$
\rho_{k,k'} \geq 0, \quad \forall k, k'.
$$
With this relaxation, this objective function is concave with respect to both optimization variables $\rho_{k,k'}$ and $P_{k,k'}$. Therefore, the corresponding relaxed problem is a convex optimization problem. We will solve this problem in the dual domain and obtain the integer values of subcarrier pairing variables $\rho_{k,k'}$ during this process. The Lagrangian obtained by relaxing the power and subcarrier pairing constraints can be written as

$$L(\rho, P, \varphi) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} \log \left( 1 + \alpha_{k,k'} \frac{P_{k,k'}}{\rho_{k,k'}} \right)$$

$$+ \varphi \left( P - \sum_{k=1}^{K} \sum_{k'=1}^{K} P_{k,k'} \right) + \sum_{k'=1}^{K} \varphi_{k'} \left( 1 - \sum_{k=1}^{K} \rho_{k,k'} \right)$$

(11.148)

where $\varphi$ and $\varphi_{k'}$ denote the Lagrange multipliers associated with the power and subcarrier pairing constraints (11.140) and (11.141), respectively. In addition, we have defined the dual vector $\varphi = (\varphi, \varphi_1, \varphi_2, \ldots, \varphi_K)$. Other constraints are not relaxed in the Lagrangian, but they will be accounted for when we develop the resource allocation algorithm. The dual function can be determined as

$$g(\varphi) = \max_{\rho, P} L(\rho, P, \varphi) \quad \text{s.t.} \quad (11.142), (11.143), (11.147).$$

(11.149)

The dual optimization problem is

$$\min_{\varphi} \quad g(\varphi) \quad \text{s.t.} \quad \varphi \geq 0.$$

(11.150)

We now solve problem (11.149) to determine the dual function for a given dual point $\varphi$. Toward this end, we first derive the derivative of $L(\varphi, P, \varphi)$ with respect to $P_{k,k'}$ and set it equal to zero:

$$\frac{\partial L}{\partial P_{k,k'}} = \rho_{k,k'} \frac{\alpha_{k,k'} P_{k,k'}}{\rho_{k,k'}} - \varphi$$

(11.151)

$$= \frac{1}{\alpha_{k,k'}} + \frac{P_{k,k'}}{\rho_{k,k'}} - \varphi = 0.$$  

(11.152)

From this, we can obtain the optimal power as

$$P^*_{k,k'} = \rho_{k,k'} \left[ \frac{1}{\varphi} - \frac{1}{\alpha_{k,k'}} \right]^+.$$  

(11.153)

Substituting this result into (11.148), we can rewrite the Lagrangian as

$$L(\rho, P^*, \varphi) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} X_{k,k'} + \varphi P + \sum_{k'=1}^{K} \varphi_{k'}.$$  

(11.154)
where

\[ X_{k,k'} = \log \left( 1 + \alpha_{k,k'} \left[ \frac{1}{\varphi} - \frac{1}{\alpha_{k,k'}} \right]^+ \right) - \varphi \left[ \frac{1}{\varphi} - \frac{1}{\alpha_{k,k'}} \right]^+ - \varphi_{k'} . \]  

From this, we can determine the optimal subcarrier pairing \( \rho_{k,k'} \) by solving the following problem:

\[
\max_{\rho} \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} X_{k,k'} \\
\text{s.t.} \quad (11.141), (11.142).
\]  

This is again the standard assignment problem, which can be solved by using the Hungarian algorithm [44].

Moreover, the dual variables, i.e., Lagrange multipliers, can be updated by using the subgradient algorithm as follows:

\[
\varphi^{(l+1)} = \left[ \varphi^{(l)} - \epsilon^{(l)} \left( P - \sum_{k=1}^{K} \sum_{k'=1}^{K} P_{k,k'}^{(l)} \right) \right]^+ \\
\varphi_{k'}^{(l+1)} = \varphi_{k'}^{(l)} - \epsilon^{(l)} \left( 1 - \sum_{k=1}^{K} \rho_{k,k'}^{(l)} \right)
\]

where \( \epsilon^{(l)} = \kappa / \sqrt{l} \) with \( \kappa > 0 \) denotes the diminishing step-size to guarantee the convergence of this algorithm [39].

**Algorithm 11** Resource allocation for multi-carrier DF relaying

1: Initialize the dual variables \( \varphi^{(0)} > 0, \varphi_{k'}^{(0)}, \forall k' \), and iteration index \( l \leftarrow 0 \).
2: \textbf{repeat}
3: Compute \( X_{k,k'} \) for all \( (k,k') \) using \( \varphi^{(l)} \) and \( \varphi_{k'}^{(l)} \), \( \forall k' \) as in (11.155).
4: Compute the subcarrier pairing variables \( \rho_{k,k'} \) using \( X_{k,k'} \) by solving the assignment problem (11.156)–(11.157).
5: Compute the power allocation \( P_{k,k'} \) for all subcarrier pairs using \( \varphi^{(l)} \) and \( \rho_{k,k'}^{(l)} \) as in (11.153).
6: Update dual variables \( \varphi^{(l)} \) and \( \varphi_{k'}^{(l)} \), \( \forall k' \) by using (11.158)–(11.159) then increase the iteration index \( l \leftarrow l + 1 \).
7: \textbf{until} convergence of \( \varphi \)
8: Obtain the final subcarrier pairing solution \( \rho^* \) and power allocation solution \( P^* \) at the convergence.

The resource allocation for the multi-carrier DF relaying system is summarized in Algorithm 11. In each iteration of this algorithm, the subcarrier pairing and power allocation variables are computed as in lines 3–5, and then the Lagrange multipliers are updated by using the power allocation and subcarrier pairing solutions as in line 6. These operations are repeated until convergence.
11.4 Resource Allocation for Multi-Carrier Systems

11.4.3 Resource Allocation for Multi-User ANC Two-Way Relay Networks

In this section, we discuss the resource allocation optimization for the multi-user ANC two-way relay network. The resource allocation aims at optimizing the joint subcarrier pairing, relay power allocation, relay selection, and subcarrier assignment in the single-cellular setting with multiple users and multiple relays [46]. We assume that each user has both uplink and downlink traffic flows to exchange with the base station (BS) via one selected relay. This would be the case for cell-edge users whose direct communications with the BS is not possible.

Let \( U = \{1, 2, \ldots, U\} \) be the set of users, \( M = \{1, 2, \ldots, M\} \) denote the set of relays, and \( K = \{1, 2, \ldots, K\} \) denote the set of subcarriers. As described above, the ANC two-way relay communications between each user and the BS are completed in two time slots where the user and BS transmit their information symbols in the first time slot, and the relay amplifies the received signal and broadcasts the resulting signal in the second time slot.

Let \( h_{u,i,k} \) and \( h_{b,i,k} \) denote the power channel-gains from user \( u \) to relay \( i \) and from the BS to relay \( i \) on subcarrier \( k \), respectively. This would be the case for cell-edge users whose direct communications with the BS is not possible.

Let \( U = \{1, 2, \ldots, U\} \) be the set of users, \( M = \{1, 2, \ldots, M\} \) denote the set of relays, and \( K = \{1, 2, \ldots, K\} \) denote the set of subcarriers. As described above, the ANC two-way relay communications between each user and the BS are completed in two time slots where the user and BS transmit their information symbols in the first time slot, and the relay amplifies the received signal and broadcasts the resulting signal in the second time slot.

By deriving the SNRs in a similar manner with (11.67) and (11.68), the SNRs achieved at user \( u \) and BS for the two-way communications of user \( u \) exploiting relay \( i \) over the subcarrier pair \((k, k^{'})\) can be expressed as

\[
\text{SNR}_{u,i,k,k^{'}} = \frac{h_{b,i,k}h_{u,u,k}P_{b,i,k}P_{i,u,k^{'}}}{P_{i,u,k}h_{i,u,k^{'}} + m_{u,i,k}} \quad (11.160)
\]

\[
\text{SNR}_{b,i,k,k^{'}} = \frac{h_{u,i,k}h_{b,b,k}P_{u,i,k}P_{i,u,k^{'}}}{P_{i,u,k}h_{i,b,k^{'}} + m_{u,i,k}} \quad (11.161)
\]

where \( m_{u,i,k} = 1 + P_{b,i,k}h_{b,i,k} + P_{u,i,k}h_{u,i,k} \). Hence, the total achieved rates of two opposite flows over the subcarrier pair \((k, k^{'})\) can be written as

\[
R_{u,i,k,k^{'}} = \frac{1}{2} \log \left(1 + \text{SNR}_{u,i,k,k^{'}}\right) + \frac{1}{2} \log \left(1 + \text{SNR}_{b,i,k,k^{'}}\right) \quad (11.162)
\]

We assume that the powers of each user \( u \) and BS on any allocated subcarriers are fixed while we will optimize the relay transmit powers. Then it can be verified that the rate function \( R_{u,i,k,k^{'}} \) is concave in the relay power \( P_{i,u,k^{'}} \). We introduce binary variables
$x_{u,i,k,k'}$ to represent the relay selection and assignment decisions of the subcarrier pair $(k, k')$ to relays and users where $x_{u,i,k,k'} = 1$ if the subcarrier pair $(k, k')$ is assigned to support the communications between user $u$ and the BS with the assistance of relay $i$ and $x_{u,i,k,k'} = 0$, otherwise. To ensure the interference-free among users, we impose the following constraints:

$$\sum_{u=1}^{U} \sum_{i=1}^{M} \sum_{k=1}^{K} x_{u,i,k,k'} \leq 1, \quad \forall k' \quad (11.163)$$

$$\sum_{u=1}^{U} \sum_{i=1}^{M} \sum_{k=1}^{K} x_{u,i,k,k'} \leq 1, \quad \forall k. \quad (11.164)$$

Moreover, we need to impose the total power constraint for each relay as follows:

$$\sum_{u=1}^{U} \sum_{k'=1}^{K} P_{i,u,k'} \leq P_i, \quad \forall i \quad (11.165)$$

where $P_i$ denotes the maximum power of relay $i$.

We study the joint optimization of the subcarrier pairing, relay selection, relay power allocation, and subcarrier assignment to maximize the weighted sum rate. This problem can be formulated as follows:

$$\max_{x, P} \sum_{u=1}^{U} w_u \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{k'=1}^{K} x_{u,i,k,k'} R_{u,i,k,k'}$$

s.t. (11.163), (11.164), (11.165) (11.167)

where $w_u$ is the weight that represents the priority of user $u$.

Let $X$ denote the set of feasible relay selection and subcarrier pairing and allocation variables $x$ and $P(x)$ denote the set of feasible power allocation $P$ for a given $x$. We can develop an asymptotically optimal resource allocation by solving the problem (11.166)–(11.167) in the dual domain. Toward this end, we formulate the Lagrangian by relaxing the relay power constraints as follows:

$$L(P, x, \lambda) = \sum_{u=1}^{U} w_u \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{k'=1}^{K} x_{u,i,k,k'} R_{u,i,k,k'}$$

$$+ \sum_{i=1}^{M} \lambda_i \left( P_i - \sum_{u=1}^{U} \sum_{k'=1}^{K} P_{i,u,k'} \right)$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_M)$ represents the vector of Lagrange multipliers associated with the relay power constraints. Then the dual function can be defined as

$$g(\lambda) = \max_{x \in X, P \in P(x)} L(P, x, \lambda). \quad (11.169)$$

The dual optimization problem can be then stated as

$$\min_{\lambda \geq 0} g(\lambda). \quad (11.170)$$
The dual-based resource allocation algorithm can be developed by iteratively computing the dual function in (11.169) for a given dual point \( \lambda \) and updating the dual variables \( \lambda \) to solve the dual problem (11.170) by using the subgradient algorithm. We will pursue these design steps in the following.

To compute the dual function in (11.169), let us rewrite the Lagrangian function as

\[
L(P, x, \lambda) = \sum_{u=1}^{U} w_u \sum_{i=1}^{M} \sum_{k=1}^{K} L_{u,i,k'} (P_{i,u,k'}) + \sum_{i=1}^{M} \lambda_i P_i \tag{11.171}
\]

where

\[
L_{u,i,k'} (P_{i,u,k'}) = w_u \sum_{k=1}^{K} R_{u,i,k,k'} - \lambda_i P_{i,u,k'} \tag{11.172}
\]

In order to compute the dual function in (11.169), we first optimize the Lagrangian with respect to \( P \) for a given \( x \) then we optimize over \( x \). Suppose that we have \( x_{u,i,k,k'} = 1 \) for a certain \((u, i, k, k')\), then there is only one rate element \( R_{u,i,k,k'} \) corresponding to subcarrier \( k \) in (11.172) that is non-zero due to the subcarrier assignment constraint (11.163).

Because of the decomposed structure of the Lagrangian function in (11.171), we can optimize each \( L_{u,i,k'} (P_{i,u,k'}) \) in (11.172) independently. It can be verified that \( L_{u,i,k'} (P_{i,u,k'}) \) is concave in \( P_{i,u,k'} \). Therefore, the optimal \( P_{i,u,k'}^* \) can be determined from the KKT conditions, i.e., setting the derivative of \( L_{u,i,k'} (P_{i,u,k'}) \) with respect to \( P_{i,u,k'} \) to zero. More specifically, the optimal \( P_{i,u,k'}^* \) can be obtained from

\[
a P_{i,u,k'}^* + b P_{i,u,k'}^3 + c P_{i,u,k'}^2 + d P_{i,u,k'} + e = 0 \tag{11.173}
\]

where

\[
a = 2 \ln 2 \lambda_i h_{b,i,k}^2 h_{u,i,k'}^2 / m_{u,i,k}
\]

\[
b = 4 \ln 2 \lambda_i h_{b,i,k} h_{u,i,k'} (h_{b,i,k'} + h_{u,i,k})
\]

\[
c = 2 m_{u,i,k} \ln 2 \lambda_i (h_{b,i,k'}^2 + h_{u,i,k'}^2 + 4 h_{b,i,k'} h_{u,i,k})
\]

\[
d = 4 m_{u,i,k}^2 \ln 2 \lambda_i (h_{u,i,k'} + h_{b,i,k'})
\]

\[
e = 2 m_{u,i,k}^3 \ln 2 \lambda_i (P_{u,i,k} h_{u,i,k} h_{b,i,k'} + P_{b,i,k} h_{b,i,k} h_{u,i,k'})
\]

Substituting the optimal \( P_{i,u,k'}^* \) into (11.169), the dual function can be rewritten as

\[
g(\lambda) = \max_{x \in \mathcal{X}} \sum_{u=1}^{U} \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{k'=1}^{K} x_{u,i,k,k'} H_{u,i,k,k'} + \sum_{i=1}^{M} \lambda_i P_i \tag{11.174}
\]

where

\[
H_{u,i,k,k'} = w_u R_{u,i,k,k'} (P_{i,u,k'}^* (\lambda_i)) - \lambda_i P_{i,u,k'}^* (\lambda_i) \tag{11.175}
\]
Using these optimal power allocations for different combinations \((u, i, k, k')\), we can calculate the corresponding quantities \(H_{u,i,k,k'}\), which will be used to optimize variables \(x\) in the following. According to (11.174), each subcarrier pair \((k, k')\) must be allocated to the user and relay pair \((u, i)\) as follows:

\[
(u^*, i^*)_{k,k'} = \arg\max_{u \in U, i \in M} H_{u,i,k,k'}.
\]  

(11.176)

We also define

\[
Y_{k,k'} = \max_{u \in U, i \in M} H_{u,i,k,k'}.
\]

(11.177)

Hence, the dual function can be rewritten as

\[
g(\lambda) = \max_{x \in \mathcal{X}} K \sum_{k = 1}^{K} x_{u^*,i^*,k,k'} Y_{k,k'} + \sum_{i=1}^{M} \lambda_i P_i.
\]

(11.178)

Therefore, the optimal \(x\) can be obtained from the following problem:

\[
\max_{x} K \sum_{k = 1}^{K} \sum_{k' = 1}^{K} x_{u^*,i^*,k,k'} Y_{k,k'}
\]  

s.t. \( K \sum_{k = 1}^{K} x_{u^*,i^*,k,k'} \leq 1, \ \forall k' \)  

(11.180)

\[
K \sum_{k' = 1}^{K} x_{u^*,i^*,k,k'} \leq 1, \ \forall k.
\]

(11.181)

Here we need to match each subcarrier \(k\) in the first phase with a subcarrier \(k'\) in the second phase of the two-way relaying strategy while the corresponding user and relay pair \((u^*, i^*)\) is determined from (11.176). This is a standard job assignment problem, which can be solved in polynomial time by using the Hungarian algorithm [44].

We have shown how to obtain the optimal \(P\) and \(x\) for a given dual point \(\lambda\). The remaining task is to solve the dual problem (11.170) by using the subgradient algorithm. Specifically, it can be verified that the subgradient of the dual function \(g(\lambda)\) at \(\lambda_i\) is

\[
\Delta \lambda_i = P_i - \sum_{u=1}^{U} \sum_{k=1}^{K} P_{u,i,k}^* (\lambda_j), \ \forall i.
\]

(11.182)

Let us define the subgradient vector as \(\Delta \lambda^{(l)} = (\Delta \lambda_1^{(l)}, \Delta \lambda_2^{(l)}, \ldots, \Delta \lambda_M^{(l)})\) in a particular iteration \(l\). Then the subgradient algorithm to optimize the dual variables \(\lambda\) in iteration \(l + 1\) is given as

\[
\lambda^{(l+1)} = \left[\lambda^{(l)} - \epsilon^{(l)} \Delta \lambda^{(l)}\right]^{+}
\]

(11.183)

where \(\epsilon^{(l)} = \kappa/\sqrt{l}\) with \(\kappa > 0\) denotes the diminishing step-size to guarantee the convergence of this algorithm [39].
Due to the non-zero duality gap with a finite number of subcarriers \( K \), the obtained solution \( P^* \) and \( x^* \) may not be feasible. To resolve this issue, one can fix the obtained relay selection, subcarrier pairing, and assignment variables \( x^* \) and re-optimize the power allocation to achieve a feasible solution. Specifically, let \( A_{u,i} \) denote the set of subcarrier pairs \((k, k')\) allocated to the user and relay \((u, i)\) according to \( x^* \). Then the power allocation can be re-optimized by solving the following problem:

\[
\max_{x, P} \quad \sum_{u=1}^{U} w_u \sum_{i=1}^{M} \sum_{(k, k') \in A_{u,i}} R_{u,i,k,k'}(P_{u,i,k,k'})
\]

\[
\text{s.t.} \quad \sum_{u=1}^{U} \sum_{(k, k') \in A_{u,i}} P_{u,i,k,k'} \leq P_t, \quad \forall i.
\]

This problem is a convex optimization problem, which can be solved by using a standard interior point method.

**Algorithm 12** Resource allocation for ANC two-way relaying

1. Initialize the non-negative dual variables \( \lambda^{(0)} \) and iteration index \( l \leftarrow 0 \).
2. **repeat**
   3. Compute \( H_{u,i,k,k'} \) for all \((u, i, k, k')\) by using (11.175) where the optimal relay power \( P^*_{u,i,k,k'} \) is a positive root of (11.173).
   4. Compute \( Y_{k,k'} \) for all subcarrier pairs \((k, k')\) as in (11.177) then obtain the optimal assignments \( x^{(\lambda)} \) by using the Hungarian method.
   5. Update \( \lambda^{(l)} \) by using (11.183) then increase the iteration index \( l \leftarrow l + 1 \).
   6. **until** convergence of \( \lambda \)
   7. Fix the variables \( x^* \) and \( \lambda^* \) then solve the problem (11.184)–(11.185) to obtain the final feasible power allocation \( P^* \).

The overall resource allocation algorithm for the multiuser ANC two-way relaying system is summarized in Algorithm 12. The algorithm comprises the iterative procedure, in which we calculate the optimal resource allocation solution \( P(\lambda^{(l)}) \) and \( x(\lambda^{(l)}) \) for a given dual point \( \lambda^{(l)} \) in lines 3–4 and update the dual variables \( \lambda^{(l)} \) by using the subgradient algorithm in line 5, and the final step to re-optimize the relay power allocation for the obtained \( x^* \) and \( \lambda^* \) to achieve a feasible power allocation solution in line 7.

### 11.5 Further Discussion

While we have mainly discussed the fundamental relaying techniques and relaying resource allocation for simple wireless systems in this chapter, there have been significant research efforts in applications of wireless relaying to more practical and complicated wireless network settings. In particular, the multi-user single cellular wireless...
system with both uplink and downlink traffic flows has been considered in [15]. This work studies the joint optimization of the relay selection, relay strategy, power, and bandwidth allocation to maximize the sum network utility. The dual-based resource allocation framework is developed where multiple sub-problems involving the resource allocation optimization of each relay strategy on individual subcarriers are solved independently, and these sub-problems are coupled through different Lagrange multipliers, which are iteratively updated by using the sub-gradient algorithm.

A similar joint optimization for mode selection among direct, DF, and AF communication modes and power and subcarrier allocation for two-way communications in the single cellular setting has also been studied [47]. The dual decomposition approach has been employed to tackle the per-subcarrier resource allocation where the power allocation optimization for each communication node is tackled where the general rate regions achieved in the two time slots of the DF and AF relaying strategies are considered. This design also enables us to achieve the asymptotically optimal solution as the number of subcarriers is sufficiently large [38]. These research works [15] and [47], however, do not consider the subcarrier pairing optimization.

There have been significant research interests in applying relaying techniques to design the multi-hop wireless networks [48, 49]. Such design has to address the joint engineering of resource allocation and other wireless protocols such as routing and congestion control. Finally, resource allocation for relaying communications in the multi-cell multi-user cellular network has been also studied [50]. This design is challenging since resource allocation in individual cells must be jointly optimized with the resource re-use and inter-cell interference mitigation. Therefore, development of distributed, scalable, and efficient resource allocation algorithms for this setting is very challenging.

11.6 Exercises

The following exercises are based on the results in [51–54].

Exercise 11.1: Consider the multi-carrier DF-based wireless system as in Section 11.3.2 except that the same subcarrier on the source-relay link is employed on the relay-destination link for the relay mode (i.e., subcarrier pairing optimization is not considered). Suppose there are \( K \) subcarriers and a single relay. Moreover, communications from a source node to its destination node can be achieved by using either the direct communication mode or DF relay mode. Then the achievable rate on subcarrier \( k \) can be written as

\[
R_k = \begin{cases} 
\frac{1}{2} \log(1 + P_k \alpha_k), & \text{direct mode} \\
\frac{1}{2} \min\{\log(1 + P_k \alpha_{k,1}), \log(1 + P_{k,1} \alpha_{k,3} + P_{k,2} \alpha_{k,2})\}, & \text{relay mode.}
\end{cases}
\]

We are interested in optimizing the mode selection (i.e., direct or relay mode), subcarrier, and power allocation to maximize the system sum rate under the total source
and relay power constraint. In particular, let $K_s$ and $K_r$ be sets of subcarriers operating in the direct and relay modes, respectively. Then the power constraint can be stated as $\sum_{k \in K_s} P_{k,1} + \sum_{k \in K_r} (P_{k,1} + P_{k,2}) \leq P$ where $P$ represents the total power. Moreover, the system sum rate can be written as $\sum_{k \in K_s} R_k + \sum_{k \in K_r} R_k$.

i. Prove that the sets $K_s$ and $K_r$ can be determined by using the power channel-gains as follows:

- If subcarrier $k$ satisfies $\alpha_{k,3} > \alpha_{k,1}$ then $k \in K_s$.
- If subcarrier $k$ satisfies $\alpha_{k,3} \leq \alpha_{k,1}$ and $\alpha_{k,3} > \alpha_{k,2}$ then $k \in K_s$.
- If subcarrier $k$ satisfies $\alpha_{k,3} \leq \alpha_{k,1}$ and $\alpha_{k,3} \leq \alpha_{k,2}$ then $k \in K_r$.

ii. Prove that for each subcarrier $k \in K_r$, the optimal powers should satisfy $P_{k,1} = \beta_k P_{k,2}$ where $\beta_k = \frac{\alpha_{k,2}}{\alpha_{k,1} - \alpha_{k,3}}$.

Exercise 11.2: For the same sum-rate maximization considered in Exercise 11.1, let us define the total power consumption for subcarrier $k \in K_r$ in the relay mode as $P_k = P_{k,1} + P_{k,2}$. Prove that the optimal power allocations under the sum power constraint $\sum_{k \in K_s} P_{k,1} + \sum_{k \in K_r} P_k \leq P$ are as follows:

- For subcarrier $k \in K_s$, the optimal power satisfies
  $$P_{k,1} = \left[\frac{1}{\mu} - \frac{1}{\alpha_{k,3}}\right]^+$$

- For subcarrier $k \in K_r$, the optimal power satisfies
  $$P_k = \left[\frac{1}{\mu} - \frac{1}{\alpha_{k,1}} - \frac{1 + \beta_k}{\beta_k}\right]^+$$

where $[x]^+ = \max\{0, x\}$, $\beta_k$ is defined as in Exercise 11.1, and $\mu > 0$ should be chosen so that $\sum_{k \in K_s} P_{k,1} + \sum_{k \in K_r} P_k = P$.

Hint: Using the results in Exercise 11.1, the system sum rate can be expressed as

$$R = \sum_{k \in K_s} \frac{1}{2} \log (1 + P_{k,1} \alpha_{k,3}) + \sum_{k \in K_r} \frac{1}{2} \log \left(1 + P_k \alpha_{k,1} \frac{\beta_k}{1 + \beta_k}\right).$$

The optimal power allocation can be derived by defining the Lagrangian via relaxing the total power constraint and then applying the KKT optimality condition (i.e., set the derivative of the Lagrangian function with respect to different power variables to zero).

Exercise 11.3: Consider again a DF-based wireless system without sub-carrier pairing optimization as in Exercise 11.1. However, we study the resource allocation for sum-rate maximization under the source power constraint (i.e., the relay power is assumed to be sufficiently large). This power constraint could hold in the uplink communication of the cellular network where the relays have abundant power. The source power constraint can be stated as $\sum_{k \in K_s} P_{k,1} \leq P_S$. 
i. Prove that the mode selection for the two subcarrier sets $\mathcal{K}_s$ and $\mathcal{K}_r$ can be determined as follows:
   - If subcarrier $k$ satisfies $\alpha_{k,3} > \alpha_{k,1}$ then $k \in \mathcal{K}_s$; otherwise,
   - If subcarrier $k$ satisfies $\alpha_{k,3} \leq \alpha_{k,1}$ then $k \in \mathcal{K}_r$.

Hint: The relay power can be set sufficiently large in the relay mode for any subcarrier $k \in \mathcal{K}_r$.

ii. Prove that the optimal power allocation for the source node and subcarrier $k$ is as follows:
   - For subcarrier $k \in \mathcal{K}_s$, the optimal power satisfies
     \[ P_{k,1} = \left[ \frac{1}{\mu} - \frac{1}{\alpha_{k,3}} \right]^+ \]
   - For subcarrier $k \in \mathcal{K}_r$, the optimal power satisfies
     \[ P_{k,1} = \left[ \frac{1}{\mu} - \frac{1}{\alpha_{k,1}} \right]^+ \]

where $\mu > 0$ should be chosen so that $\sum_{k \in \mathcal{K}_s \cup \mathcal{K}_r} P_{k,1} = P_s$.

**Exercise 11.4:** Consider a single-carrier AF-based wireless relaying system as in Section 11.3.1 with a single relay. Suppose that the source-destination link is significantly weaker than the source-relay and relay-destination links. Let $P_s$ and $P_r$ denote the source and relay transmit powers. Moreover, let $\alpha_1$ and $\alpha_2$ denote the normalized power channel-gains with the noise powers of the source-relay and relay-destination links. For this system, the achieved SNR at the destination can be approximated as

\[ \text{SINR} = \frac{\alpha_1 \alpha_2 P_s P_r}{1 + P_s \alpha_1 + P_r \alpha_1} \approx \frac{\alpha_1 \alpha_2 P_s P_r}{P_s \alpha_1 + P_r \alpha_1} \]

where the second approximation holds in the high SNR or power regime. Suppose that we are interested in the optimal source and relay power allocation to maximize the SNR subject to the following total power constraint $P_s + P_r \leq P_T$. Prove that the optimal power solution satisfies

\[ P_s = \begin{cases} \frac{-\alpha_2 + \sqrt{\alpha_1 \alpha_2}}{\alpha_1 - \alpha_2} P_T, & \text{if } \alpha_1 \neq \alpha_2 \\ \frac{P_r}{2}, & \text{if } \alpha_1 = \alpha_2 \end{cases} \]  

(11.186)

\[ P_r = \begin{cases} \frac{\alpha_1 - \sqrt{\alpha_1 \alpha_2}}{\alpha_1 - \alpha_2} P_T, & \text{if } \alpha_1 \neq \alpha_2 \\ \frac{P_r}{2}, & \text{if } \alpha_1 = \alpha_2 \end{cases} \]  

(11.187)

**Hint:** The optimal power allocation can be obtained from the KKT optimality conditions.
**Exercise 11.5:** Consider the multi-carrier AF-based wireless relaying system with subcarrier pairing optimization as in Section 11.4.1. Use the same notations for the normalized channel-gains of the source-relay, relay-destination, and source destination links for subcarrier \( k \) and relay \( i \) as \( \alpha_{k,i,1}, \alpha_{k,i,2}, \) and \( \alpha_{k,3} \) and the source and relay powers as \( P_{k,i,1} \) and \( P_{k,i,2} \), respectively.

As described in Section 11.4.1, the achievable rate on subcarrier pair \((k, k')\) can be expressed as

\[
R_{k,k',i} \approx \frac{1}{2} \log \left( 1 + \alpha_{k,3} P_{k,i,1} + \frac{\alpha_{k,i,1} \alpha_{k',i,2} P_{k,i,1} P_{k',i,2}}{\alpha_{k,i,1} P_{k,i,1} + \alpha_{k',i,2} P_{k',i,2}} \right).
\]

Suppose that we are interested in performing power allocation for the source and relay \( i \) to maximize the achieved rate \( R_{k,k',i} \) subject to the total power constraint \( P_{k,i,1} + P_{k',i,2} = P_{k,k'} \) where \( P_{k,k'} \) denotes the total power. Prove the following statements.

i. The optimal power allocation for the source and relay is

\[
P_{k,i,1} = \begin{cases} 
\frac{a_{k,i,2} (c_{k,i} + \alpha_{k,3})}{c_{k,i} (c_{k,i} + \alpha_{k',2})}, & \text{if } \alpha_{k',i,2} > \alpha_{k,3} \\
\frac{1}{c_{k,i}} (c_{k',i} + \alpha_{k',2}), & \text{if } \alpha_{k',i,2} \leq \alpha_{k,3}
\end{cases}
\]

\[
P_{k',i,2} = \begin{cases} 
\frac{\alpha_{k,i,1} (c_{k,i} - \alpha_{k,3})}{c_{k,i} (c_{k,i} + \alpha_{k',2})}, & \text{if } \alpha_{k',i,2} > \alpha_{k,3} \\
0, & \text{if } \alpha_{k',i,2} \leq \alpha_{k,3}
\end{cases}
\]

where \( c_{k,i} = \sqrt{\alpha_{k,i,1} \alpha_{k',i,2} - \alpha_{k,i,1} \alpha_{k,3} + \alpha_{k',i,2} \alpha_{k,3}} \).

ii. The achieved rate under the optimal power allocation given in question 11.5.i can be written as

\[
R_{k,k',i} = \begin{cases} 
\frac{1}{2} \log \left( 1 + \frac{a_{k',i,2} (c_{k,i} + \alpha_{k,3})^2}{(c_{k,i} + \alpha_{k',2}) P_{k,k'}} \right), & \text{if } \alpha_{k',i,2} > \alpha_{k,3} \\
\frac{1}{2} \log \left( 1 + \alpha_{k,3} P_{k,k'} \right), & \text{if } \alpha_{k',i,2} \leq \alpha_{k,3}
\end{cases}
\]  \hspace{1cm} (11.188)

**Hint:** The optimal power allocation given in 11.5.i can be derived from the KKT optimality conditions of the constrained optimization problem.

**Exercise 11.6:** Consider the joint subcarrier pairing, subcarrier and relay assignment, and power allocation for the multi-carrier AF-based wireless system similar to the problem in Section 11.4.1, but we assume the total power constraint instead of individual source and relay power constraints. Specifically, we study the following resource
allocation problem:

\[
\max_{\rho, P} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i=1}^{M} 2\rho_{k,k'} R_{k,k',i}
\]

s.t.

\[
\sum_{k=1}^{K} \rho_{k,k'} = 1, \quad \forall k' ; \quad \sum_{k'=1}^{K} \rho_{k,k'} = 1, \quad \forall k
\]

\[
\sum_{i=1}^{M} t_{k,k',i} = 1, \quad \forall k, k'
\]

\[
P_{k,i,1} + P_1'_{k,i,2} = P_{k,k'}
\]

\[
\sum_{k=1}^{K} \sum_{k'=1}^{K} P_{k,k'} \leq P_T.
\]

Prove the following statements.

i. The optimal assignment of each subcarrier pair \((k, k')\) to relays can be performed as follows:

\[
t_{k,k',i} = \begin{cases} 
1, & \text{for } i = i(k, k') = \arg \max_i \alpha_{k,k',i} \\
0, & \text{otherwise}
\end{cases}
\]  

(11.189)

where

\[
\alpha_{k,k',i} = \begin{cases} 
\frac{c_{k,k',i} + \alpha_{k,k',i}}{c_{k,k',i} + \alpha_{k,k',i} + \alpha_{k,k',i}^2}, & \text{if } \alpha_{k,k',i} > \alpha_{k,k',i}^2 \\
\alpha_{k,k',i}, & \text{if } \alpha_{k,k',i} \leq \alpha_{k,k',i}^2
\end{cases}
\]  

(11.190)

where the definition of \(c_{k,k',i}\) is given in Exercise 11.5. This means that the optimal assignment decisions of subcarrier pairs to relays only require the channel state information.

**Hint:** Using the result in (11.188) in Exercise 11.5.ii to obtain the optimal assignment of each subcarrier pair \((k, k')\) to relays.

ii. The considered resource allocation is equivalent to the following problem:

\[
\max_{\rho, P} \sum_{k=1}^{K} \sum_{k'=1}^{K} \rho_{k,k'} \log \left( 1 + \alpha_{k,k'} P_{k,k'} \right)
\]

s.t.

\[
\sum_{k=1}^{K} \rho_{k,k'} = 1, \quad \forall k' ; \quad \sum_{k'=1}^{K} \rho_{k,k'} = 1, \quad \forall k
\]

\[
\sum_{k=1}^{K} \sum_{k'=1}^{K} P_{k,k'} \leq P_T
\]

where \(\alpha_{k,k'} = \alpha_{k,k',i(k, k')}\) with \(i(k, k')\) being the relay assigned for subcarrier pair \((k, k')\).

**Exercise 11.7:** Consider a full-duplex (FD) relay that can receive and forward signals simultaneously for relaying communications from a source node to a destination node as illustrated in Figure 11.3. The relay can employ the standard AF or DF relaying scheme.
Let $P_s$ and $P_r$ denote the source and relay transmit powers. Suppose that the noise powers at the relay and destination are both equal to $N$.

Moreover, let $\alpha_1$, $\alpha_2$, $\alpha_3$, and $\alpha_4$ denote the normalized power channel-gains with the noise powers of the source-relay, relay-destination, source-destination links, and the look-back self-interfering link at the relay. For simplicity, we appropriately normalize the parameters $\alpha_i$ so that $P_s$ and $P_r$ are constrained to be smaller than or equal to 1. We also define the following quantities:

$$
\gamma_1 = \frac{P_s \alpha_1}{P_r \alpha_4 + 1}, \quad \gamma_2 = \frac{P_r \alpha_2}{P_s \alpha_3 + 1}.
$$

Then the SNR achieved by the AF- and DF-based FD system can be expressed as

$$
\gamma_{\text{FD,AF}} = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2 + 1}, \quad \text{AF relaying}
$$

$$
\gamma_{\text{FD,DF}} = \min \{\gamma_1, \gamma_2\}, \quad \text{DF relaying}.
$$

Suppose that we fix the source power $P_s = 1$. Prove that the optimal relay power allocation to maximize the SNR satisfies

$$
P_r = \min \left\{ 1, \frac{1}{\alpha_4} \sqrt{\frac{(\alpha_1 + 1)(\alpha_3 + 1)\alpha_4}{\alpha_2}} \right\}, \quad \text{AF relaying}
$$

$$
P_r = \min \left\{ 1, \frac{1}{\alpha_4} \left( \sqrt{\frac{\alpha_4 (\alpha_3 + 1)}{\alpha_2}} + \frac{1}{4} - \frac{1}{2} \right) \right\}, \quad \text{DF relaying}.
$$

Hint: For the FD AF relaying strategy, the optimal relay power is derived from $\frac{\partial \gamma_{\text{FD,AF}}}{\partial P_r} = 0$, while for the FD DF relaying strategy, the optimal relay power is obtained from $\gamma_1 = \gamma_2$.

**Exercise 11.8:** Consider an FD wireless relaying system as in Exercise 11.7 and suppose that the maximum powers $P_s = 1$ and $P_r = 1$ are used at the source and relay. In this case, the SNR achieved by the direct transmission (DT) mode is equal to $\alpha_3$. Prove that
the DT mode is preferred over the FD mode if we have

\[
\alpha_3 > \Gamma_3 = \begin{cases} 
\frac{\gamma_1 + \alpha_2 + 1}{\gamma_1 + 1} \left( \sqrt{\frac{\gamma_1 + 1}{(\gamma_1 + \alpha_2 + 1)^2}} + \frac{1}{4} - \frac{1}{2} \right), & \text{AF relaying} \\
\min \left\{ \gamma_1, \sqrt{\alpha_2 + \frac{1}{4}} - \frac{1}{2} \right\}, & \text{DF relaying.}
\end{cases}
\]

Hint: Compare the achievable rates of corresponding relaying schemes under the DT and FD modes to derive the condition under which the DT mode is preferred over the FD mode.

**Exercise 11.9:** For same parameters in Exercise 11.5, the SNRs achieved by the half-duplex (HD) DF and AF relaying schemes with \( P_s = 1 \) and \( P_r = 1 \) are

\[
\gamma_{\text{HD,AF}} = \alpha_3 + \frac{\alpha_1 \alpha_2}{\alpha_1 + \alpha_2 + 1}, \quad \text{AF relaying}
\]

\[
\gamma_{\text{HD,DF}} = \min \{ \alpha_1, \alpha_2 + \alpha_3 \}, \quad \text{DF relaying}
\]

where we have assumed that the maximal ratio combiner is used to combine the signals in two phases at the destination.

Prove that as the maximum powers are used at the source and relay, the FD mode is preferred over the HD mode if we have

\[
\alpha_4 < \Gamma_4 = \frac{\alpha_1 (\gamma_2 - \hat{\gamma}_{\text{HD,AF}})}{\hat{\gamma}_{\text{HD,AF}} (\gamma_2 + 1)} - 1, \quad \text{AF relaying}
\]

\[
\frac{\alpha_1}{\hat{\gamma}_{\text{HD,DF}}} - 1, \quad \text{DF relaying}
\]

where \( \hat{\gamma}_{\text{HD,AF}} = \sqrt{1 + \gamma_{\text{HD,AF}}} - 1 \) and \( \hat{\gamma}_{\text{HD,DF}} = \sqrt{1 + \gamma_{\text{HD,DF}}} - 1 \). Otherwise, the HD mode is preferred over the FD mode.

Hint: Compare the achievable rates of HD and FD schemes to derive the condition under which the FD mode is preferred over the HD mode. Moreover, the rates achieved by the FD AF scheme and HD AF scheme are \( \log (1 + \gamma_{\text{HD,AF}}) \) and \( 1/2 \log (1 + \gamma_{\text{HD,AF}}) \). Hence, we can use the equivalent SNR \( \hat{\gamma}_{\text{HD,AF}} \) to cancel out the factor \( 1/2 \), i.e., we have \( 1/2 \log (1 + \gamma_{\text{FD,AF}}) = \log (1 + \hat{\gamma}_{\text{FD,AF}}) \). The same argument is also applied to the case of the DF scheme.

**References**


12 Channel Allocation for Infrastructure-Based 802.11 WLANs

12.1 Introduction

Due primarily to its unlicensed frequency band of operation and low-cost equipment, the IEEE 802.11-based wireless access technology, also known as WiFi, has been widely deployed in local area networks (LANs). A typical deployment of this technology is shown in Figure 12.1. Based on how they are managed, wireless LANs (WLANs) can be categorized into one of the following: (1) Centrally managed or (2) Uncoordinated [2]. Centrally managed deployments are usually seen in places such as university campuses, offices, or airports where all access points (APs) and associated clients are managed by a central entity. On the other hand, uncoordinated WLANs operate in the absence of a central control and are typical in places such as residential neighborhoods or private hotspots managed by different service providers (e.g., restaurants, coffee shops, etc.).

Successful deployment in either case requires efficient mechanisms for addressing performance issues such as excessive interference, which usually translates into low throughputs. In the literature, several techniques have been proposed to address such performance issues. In particular, association control (or load balancing), in which a central entity associates (respectively, disassociates) clients with (respectively, from) APs in order to balance traffic in a network, is usually proposed for the centrally managed deployments [3]. Proposed for the uncoordinated deployments, on the other hand, are such techniques as power control [4] and careful carrier-sensing [5], in which transmission power is dynamically tuned and unnecessary carrier sensing is avoided, respectively. One other technique that is extensively considered and applicable to both centrally managed and uncoordinated environments is channel assignment, in which a frequency channel is assigned to each AP for use for a certain duration of time. In this chapter, we present a survey of such channel assignment techniques.¹ We identify and discuss several major approaches applicable to the different deployment scenarios. Subsequently, a qualitative comparison is made among these approaches. Some comments on current practice in channel assignment are also presented. Finally, several important future research directions are outlined.

¹ This chapter is primarily based on [1].
12.2 System under Consideration

12.2.1 Network Topology

We focus only on an IEEE 802.11 WLAN with an infrastructure network topology as shown in Figure 12.1, where APs and clients resort to existing communication infrastructures such as legacy LANs to facilitate their communication. A fixed AP operating on a certain channel interconnects its associated clients to the infrastructure. All communication activities must be facilitated via this AP. A single instance of such a topology is referred to as a basic service set (BSS) or cell. If more BSSs exist in the same infrastructure, the system is referred to as an extended service set (ESS). In this chapter, we focus on either a single ESS managed by one particular administrator or multiple ESSs, each managed by a different administrator.

12.2.2 Channelization

Currently, two unlicensed frequency spectrum bands are available for use in IEEE 802.11 WLANs: (1) 2.4 GHz Industrial, Scientific, and Medical (ISM) band, and (2) 5 GHz Unlicensed National Information Infrastructure (UNII) band, [6] and [7]. While the legacy IEEE 802.11 and enhanced IEEE 802.11b/g WLANs operate on the 2.4-GHz band, the IEEE 802.11a WLANs employ the 5-GHz band. Both of the bands are available internationally. The number of allowable channels, however, varies from country to country due to each country’s regulations on radio spectrum allocation. In particular,
Figure 12.2 802.11 channels in the 2.4-GHz ISM band.

while most European countries and Australia allow channels 1 up to 13 in the 802.11b/g band, most North, Central, and South American countries only allow up to channel 11 in the same band [8]. In Japan, all 14 channels are allowed. These regulations are, however, subject to change.

As shown in Figure 12.2, the 2.4-GHz band consists of 14 overlapping channels each of which occupies 22 MHz. Due to the spectral overlaps of channels within this band, the standard also specifies the allowable levels of power overlaps between overlapping channels. Specifically, as shown in Figure 12.3, the signal must drop 30 dB and 50 dB below its peak power when operating at ±11 MHz and ±22 MHz apart from the center frequency, respectively [6].

The 5-GHz UNII band contains three subbands referred to as low, middle, and high, each of which contains four non-overlapping channels as shown in Figure 12.4. Each channel occupies 20 MHz. Currently the 5-GHz band is still significantly less populated than the 2.4-GHz band because (1) 802.11a equipment is not so widespread as 802.11b/g equipment and (2) due to the higher frequency, 802.11a signals cannot penetrate as far as 802.11b/g signals and are absorbed more readily by obstacles. As in the 2.4-GHz band, the power spectrum mask is also specified in the 5-GHz band [8]. This is shown in Figure 12.5, where the signal must drop 20 dB, 28 dB, and 40 dB at 11 MHz, 20 MHz, and 30 MHz apart from the center frequency, respectively.

Figure 12.3 Power spectrum mask of the 2.4-GHz ISM channels illustrated with a particular signal spectrum density $|\sin(x)|/x$. 

https://doi.org/10.1017/9781316212493.013
12.2 System under Consideration

12.2.3 Medium Access Control

To accommodate multiple clients in a WLAN, a mode of contention-based medium access called distributed coordination function (DCF) is employed [6]. DCF uses the CSMA/CA technique (discussed in Chapter 2) in which a station, either an AP or its respective associated clients, senses the wireless medium for transmission opportunity. If the medium is idle, the station starts its transmission. Otherwise it backs off and waits for a random period of time before contending again for the medium. In the case two stations sense an idle channel and start their transmissions simultaneously, collision is said to occur. If collided, both stations will have to back off for a random period of time and then retry, reducing the probability of further collision. In order to lower collision probability, clients may adopt an optional handshake-based medium access mechanism known as Request-To-Send/Clear-To-Send (RTS/CTS) signaling. Prior to transmitting packets, a station broadcasts an RTS packet to reserve the medium. If the medium is idle, the destination responds with a CTS signal. The station then seizes the medium and starts transmitting.

![Figure 12.5](https://www.cambridge.org/core) Power spectrum mask of the 5-GHz UNII channel illustrated with some random signal spectrum density.
12.3 Channel Assignment and AP Placement in IEEE 802.11 WLANs

12.3.1 Channel Assignment

We now define channel assignment in the context of IEEE 802.11 WLANs. Consider a WLAN consisting of a set of $k$ APs pre-installed in a given geographical area as shown in Figure 12.1. Each AP supports all wireless clients residing in a BSS. A pool of $j$ channels, either overlapping or non-overlapping, is available for this WLAN (see Figures 12.2 and 12.4). Channel assignment is then defined as a strategy in which one of the $j$ channels is allocated to each AP such that the interference generated as a result of such assignment is minimized. In other words, capacity required to handle the traffic load generated by stations (APs and clients) within the WLAN should be maximized as a result of such channel assignment. All the existing channel assignment strategies considered in this survey are developed based on this underlying concept. The formulation of all the channel assignment schemes are based on optimization theory. The way interference is modeled, however, differs from one scheme to another.

As an example, let us consider Figure 12.1. Each BSS has different coverage. Assuming that only three non-overlapping channels from the 2.4-GHz band are used, one simple solution for assigning channels to APs in Figure 12.1 would be to assign channels 1, 6, 11, and 1 to AP1, AP2, AP3, and AP4, respectively. In this case, there would be no interference. Another possibility could be to assign channel 6, 1, 11, and 6 to AP1, AP2, AP3, and AP4, respectively. In reality, the system is more complicated, of course, with many more APs or BSSs coexisting either in the same management domain (centrally managed) or in different management domains (uncoordinated). In this chapter, we identify and explain channel assignment schemes that fall into one of these two categories: (1) centrally managed and (2) uncoordinated.

12.3.2 AP Placement

We define AP placement as a strategy in which APs are installed in particular geographical locations so as to provide maximum radio coverage to clients subject to certain QoS requirements. Such a strategy is usually performed during the initial phase of network planning and coupled with most channel assignment strategies developed specifically for a centrally managed network. In an uncoordinated network, however, AP placement completely disappears from the framework of channel assignment due to the lack of the centralized control over all APs managed by different network administrators. In this case, AP locations are simply taken as given, and the focus is only on channel assignment.

12.4 Challenges in Channel Assignment in IEEE 802.11 WLANs

In most cases, channel assignment is more complicated than the example given in the previous section, especially when the number of BSSs and clients grows. Reusing channels in a WLAN is more challenging compared to that in a cellular network whose
coverage area is typically well planned and has regular cell shapes. Such regularity does not exist in WLANs whose deployment is usually done indoors, where building layout and construction materials usually complicate the coverage areas and have a significant effect on the overall network performance. Moreover, as WLAN deployments start to move outdoors (e.g., metropolitan hotspots in big cities), WLANs will most likely experience the same network dynamics as cellular networks. Even worse is the situation where cell coverage cannot be planned or controlled at all, i.e., uncoordinated environments.

However, even if geographically disjoint cell coverage can be planned in such area, as shown in Figure 12.6, co-channel (or adjacent-channel) interference still remains due to the nature of CSMA protocol. To illustrate this, let us consider Figure 12.6. Due to disjoint coverage, both APs can be assigned the same channel. Client 1 and Client 2 are associated with AP1 and AP2, respectively. Client 1 sitting at the boundaries of both cells is, however, within the transmission ranges of both AP1 and AP2. Assume further that AP2 is transmitting to Client 2. Even if the channel assigned to AP1 is idle, Client 1, who wants to transmit to AP1, will always sense the channel as busy because AP2’s transmission interferes with Client 1’s sensing. Client 1 will have to defer its transmission, just as it would when its own channel is busy. Client 1 therefore wastes transmission opportunity unnecessarily, resulting in reduced throughput. Similarly, suppose Client 1 is transmitting to AP1, and at the same time the channel assigned to AP2 is idle. Even residing outside AP1’s coverage, Client 2, who wants to transmit to AP2, still suffers from co-channel interference from Client 1’s transmission because Client 2 is in a transmission range of Client 1 (not drawn). Therefore, both random channel access mechanisms and random locations of clients complicate channel assignment in WLANs.

Finally, channel assignment techniques as employed in cellular mobile systems cannot be applied directly in WLAN scenarios. In cellular networks, data traffic and control signaling traffic are usually carried in separate channels. That is, while a certain set of channels is devoted to data transmission, a common channel is usually initialized to convey control information (e.g., information related to channel assignment and reassignment) within the network. Various channel assignment strategies in the cellular domain are designed and implemented just around this concept [9]. In WLANs,
however, both data and control traffic have to share the same channel. For more information on classical frequency assignment problems, which appeared in wireless communication systems as early as 1960, interested readers are referred to a classic paper by Hale [10].

12.5 Channel Assignment Schemes in Centrally Managed Environments

We describe channel assignment schemes as applied to a centrally managed network in this section. In a centrally managed network, there exists a central entity that decides and assigns a channel to each AP such that a certain performance metric of the network is optimized. A typical metric of interest is interference, which usually translates to a capacity measure. In addition to channel assignment, the placement of APs may also be controlled by this central entity to maximize the radio coverage of the network. Two subcategories of channel assignment schemes as applied to a centrally managed network are in order: (1) channel assignment with AP placement and (2) channel assignment without AP placement.

The first sub-category reflects the early developments in this field, which usually assumes that a network administrator has complete control over the placement of APs and the assignment of channels to APs. This assumption makes sense since, in the early days, WLANs were meant to provide data communications only within organizations, just like a typical legacy LAN does. Deployment of WLANs in nearby sites, if there exist any at all, is thus not much of a concern. The main challenge the schemes under this sub-category try to address is how to overcome the irregularity in cell shapes of BSSs as well as the varying traffic demands over a given area, by means of channel assignment and AP placement. (This is actually the first challenge mentioned in Section 12.4.) One common requirement of the schemes under this sub-category is thus an accurate estimation of traffic demands in a given area in which APs are to be installed. The way each scheme estimates and exploits the traffic demands, however, differs from scheme to scheme. We describe each scheme in this sub-category in Section 12.5.1.

On the other hand, the approaches under the second sub-category ignore the placement of APs and focus only on channel assignment. The main challenge these approaches try to address is the interference induced by MAC contention, which is the second challenge, and partly the third challenge, described in Section 12.4. While [11] focuses only on the interference induced by MAC contention among APs only, the recent approaches ([12] and [13]) take into account the interference induced by MAC contention among APs and clients. We describe each scheme under this sub-category in Section 12.5.2.

12.5.1 Channel Assignment with AP Placement

Traditional Approach

In the first-generation WLANs, the channel assignment problem is usually solved as part of the initial network planning. That is, assigning channels to APs is performed...
after the possible AP locations are determined. In specifying the trial locations of APs, network parameters and requirements such as mobility, user population, surrounding physical infrastructure, prospective applications, and security levels are taken into consideration [14]. After that, the planning design is refined via a site survey, which usually involves measuring signal levels at various traffic demand locations to generate a radio coverage layout and optimal AP locations. Implicitly, the site survey helps to discover actual unforeseen interference and re-adjust the channel assignment and, perhaps, AP locations accordingly. Using this approach, the work in [15] treats the channel assignment problem as a map coloring problem in which each AP represents a vertex and a non-overlapping channel by a color. Its objective is to assign one of the non-overlapping channels to an AP such that the co-channel coverage overlap between adjacent cells is minimized.

The scheme recommends no specific map coloring algorithm to use, but does suggest that channel assignment be done first for those areas with high traffic demand followed by those with light traffic. In the high traffic demand areas, multiple channels are usually provided through multiple APs located densely close to one another to boost network capacity. For example, three APs with three non-overlapping channels (1, 6, and 11) may be provided to support traffic in a small but busy conference room. With all the high traffic demand areas covered, subsequent channel assignment to those APs located in the adjacent, light traffic areas can now be performed by taking into consideration co-channel signal spillovers from those APs situated in the high traffic areas. These spillovers thus represent (co-channel) interference, which in turn creates a co-channel coverage overlap – a quantity this scheme aims to minimize.

One other work similar to this approach can be found in [16], where the real experience on WLAN design and capacity planning is reported.

**Integer Linear Programming Approach**

In [17], the problems of channel assignment and AP placement are solved simultaneously by using Integer Linear Programming (ILP). The approach considers not only radio coverage but also load balancing among APs because the authors argue that the number of active wireless clients connected to the APs affects network performance, i.e., traffic congestion at the APs degrades the network performance such as throughput. The basic idea is therefore to distribute clients to the APs in a WLAN such that congestion at APs is minimized. Correspondingly, the throughput is maximized.

A floor plan is assumed to consist of traffic demand points, each of which is given an expected traffic demand volume. A set of AP candidate locations is also given. If a signal from the AP to the demand point is above a certain threshold, an edge is drawn between a traffic demand point and a particular AP. Similarly, an edge is drawn between two APs whenever they are within a co-channel interference distance defined as a transmission range at which, if assigned the same channel, these two APs can interfere to some extent with one another. The objective is to minimize the maximum channel utilization (a measure of congestion) at each AP, while keeping a certain level of traffic demand satisfied at each demand point. Each demand point is assigned to exactly one AP. If at least one demand point is assigned to an AP, that AP will be included in the
solution set. If an edge exists between two APs, each AP will be assigned a different non-overlapping channel.

As mentioned earlier, the goal is to distribute clients throughout the network such that the overall network throughput is maximized. This requires an accurate network layout containing the descriptions of demand points with estimated traffic, client distribution, and received signal levels at each demand location. In general, since such a network layout is very dynamic, new assignment of demand points to APs and channels to APs is necessary. The new assignment, however, may cause a certain amount of disruption to client traffic. The authors propose another ILP that aims to minimize the amount of client traffic disruption due to the new assignment process, while maintaining the resulting channel utilization below that of the previous assignment.

Priority-Map Approach
In [18], the channel assignment problem was solved in tandem with the AP placement problem. A floor plan of interest is first divided into pixels. Each pixel is prioritized based on its traffic requirements, i.e., how much capacity the pixel may need and for how long. A highest-priority pixel is thus designated as one having the highest demand of capacity and availability, and similarly a lowest-priority pixel having the lowest demand of capacity and availability. Intermediate levels of priority are possible.2

With the priority map created, the strategy is to first come up with a set of possible AP locations that can provide adequate radio coverage to every pixel of the floor plan. The adequate coverage means a minimum level of capacity and availability required by each pixel. To achieve this, a wave propagation prediction model such as a ray-tracing technique can be used to predict the coverage area generated by each candidate AP. As can be expected, more than one set of possible AP locations may result from this process. For each set of possible AP locations, the next step is to eliminate those APs that create unacceptably large coverage overlap3 between their adjacent neighbors, provided that adequate capacity and availability are still supplied to every pixel affected by this elimination. To quantify this overlap, the mean difference between the received power of two adjacent APs is used. That is, if the difference in received power averaged over every pixel in the overlap area falls below a certain threshold, one of the APs should be eliminated. This step thus eliminates the possible interference created by the radio coverage overlap of adjacent APs.

After the above elimination process, the channel assignment is now applied to each set of possible AP locations. The assignment starts in a greedy manner in which a non-overlapping channel is assigned first to the AP that covers the area (a group of pixels) with highest priority. The assignment process always continues to the next AP that exhibits the lowest signal propagation path-loss (i.e., the lowest amount of signal power drop) with respect to the previous AP. In other words, continue with the AP that might have caused greatest interference if assigned the same channel. Another

2 Note that, since in practice traffic demand at a particular pixel may vary from time to time, this prioritization of pixels can only serve as a rough estimate of traffic requirements in a long run.
3 Overlap is defined as a common area covered by two or more adjacent APs.
12.5 Channel Assignment Schemes in Centrally Managed Environments

A non-overlapping channel is then assigned to this AP. The process continues until the number of non-overlapping channels is exhausted. At this point, the process may follow one of the two proposed algorithms. In the first algorithm called Mutual Interference Algorithm, the carrier-to-interference ratios (CIRs) of the signals from all the APs already assigned a channel are evaluated at the next AP. The channel (or carrier) whose signal is received with the lowest CIR is then assigned to this AP. The process continues to the next of remaining APs and so on.

In the second algorithm called Received Power Algorithm, after the number of non-overlapping channels is exhausted, the received power of the signals from all the APs already assigned a channel is evaluated at the next AP. The channel whose signal is received with the lowest power is then assigned to this AP. The process continues to the next of remaining APs and so on. The final solution is that set of AP locations (with channel assignment) that gives the highest value to the following objective function:

\[ f = w_{cir}A_{cir} + w_{cov}A_{cov} \]  

(12.1)

where \( A_{cir} \) and \( A_{cov} \) are the mean CIR and received power averaged over all the pixels in the floor, respectively, \( w_{cir} \) and \( w_{cov} \) are the respective weighing factors. While the first term in (12.1) represents the capacity requirement, the second term serves as the radio coverage measure.

**Patching Algorithm**

AP placement and channel assignment were jointly optimized in [19]. The objective is to jointly maximize throughput and fairness among wireless clients. Uplink traffic was considered. The throughput was estimated for each client based on [20], whose throughput expression is valid only for clients residing in a single cell (or BSS) and depends only on the number of clients within that cell. The estimated throughput used in this approach is however extended to include the effect of co-channel interference generated by neighboring cells. Precisely, the throughput of client \( i \) depends not only on the number of wireless clients within its cell but also on a metric called the number of its restrainers. Restrainers of client \( i \) are defined as those clients residing in client \( i \)’s neighboring cells, whose transmission can cause enough co-channel interference for client \( i \) to sense its own channel busy. The higher the metric, the lower the throughput that client \( i \) will experience. The assignment of channels to client \( i \)’s neighboring cells (APs) thus affects client \( i \)’s throughput. The objective function is calculated as the product of the sum of these estimated throughputs (i.e., aggregate throughput) and the fairness index. The fairness index captures the deviations of the individual estimated throughputs from one client to another. The fairness index is calculated as follows:

\[ F = \frac{\left( \sum_{i}^{N} Th_i \right)^2}{N \times \sum_{i}^{N} (Th_i)^2} \]  

(12.2)

where \( N \) is the total number of clients in the network and \( Th_i \) is the throughput of client \( i \). Clearly, the fairness index is one if all the clients have exactly the same...
Channel Allocation for Infrastructure-Based 802.11 WLANs

throughput, whereas the index approaches $1/N$ when the individual throughputs are heavily unbalanced.

A unique global optimal solution requires high computational complexity to be found. Therefore, a heuristic algorithm called \textit{Patching} was proposed. The algorithm starts with a set of candidate AP locations on the floor plan and a set of non-overlapping channels. The algorithm then tests to see which AP assigned to which channel yields the largest objective function value. Such an AP with an appropriate channel is then selected, placed (or patched) on the floor, and removed from the candidate set. The process repeats with the remaining AP candidates and the same set of non-overlapping channels until a predetermined number of APs is reached. In each iteration, since a newly placed AP may cause some clients to reassociate with it, the individual estimated throughputs of those clients affected by this reassociation will change and should be reestimated accordingly.

\textbf{Coverage-Oriented Approach}

In [21], AP placement and channel assignment were optimized both sequentially and jointly. Using the integer linear programming model, the objective of the AP placement problem is to maximize a total throughput over all the service area while satisfying the specified number of APs. The net throughput function was obtained by fitting the throughput measurements to a polynomial function as the received signal power varies. The measurements were performed using a specialized network performance tool that measures the average downlink data throughput while streaming TCP traffic from an AP to only one active client who sits at the various locations or pixels of the entire coverage area. The received signal power level associated with each throughput measurement was then recorded. Given the net throughput function, the AP placement problem was formulated as follows:

$$\max \left[ \sum_{a,j} \Phi(p_{aj})x_{aj} \right]$$  \hspace{1cm} (12.3)

where $\Phi(p_{aj})$ is the net throughput function that depends on received signal strength $p_{aj}$ at pixel $j$ located at some distance away from AP $a$. The binary variable $x_{aj}$ is 1 if pixel $j$ is associated with AP $a$, and 0 otherwise. The above objective function is evaluated such that each pixel is assigned to only one AP, and the number of APs to be installed is specified beforehand.

The objective of the channel assignment problem is to minimize the physical overlapping area supported by the APs operating on either the same channel or adjacent channels. The overlap area metric is obtained by actually counting the pixels that hear interfering transmissions from the other neighboring APs. This metric therefore depends on a receiver sensitivity threshold defined for each pixel as the minimum received signal power required for sensing a channel busy. The objective function for the co-channel interference case is as follows:

$$\min \left[ \sum_{ab} v_{ab}w_{ab} \right]$$  \hspace{1cm} (12.4)

where $v_{ab}$ is the number of pixels located between APs $a$ and $b$, and interfered by either $a$ or $b$. The binary variable $w_{ab}$ is 1 if APs $a$ and $b$ operate on the same channel,
and 0 otherwise. For the adjacent channel interference case, the weighting factor which depends on channel distance between two adjacent channels can be integrated into (12.4).

As mentioned earlier, both of the objectives shown in (12.3) and (12.4) can be solved sequentially: AP placement followed by channel assignment. For joint optimization, both objectives are linearly combined. A weighting factor is used for each objective to prioritize the two separate goals, i.e., to trade off the total throughput against the channel overlapping area, which in turn represents interference. This weight is actually a design parameter that, if chosen appropriately, can drive the joint approach to outperform the sequential one in terms of the net throughput and the amount of the physical overlapping area. For a small number of APs, the problem can be solved for optimality. For a larger network, the approach has to resort to heuristic algorithms.

12.5.2 Channel Assignment without AP Placement

**DSATUR: Vertex Coloring-Based Approach**

In graph coloring, APs are treated as vertices of a graph, and a single edge of the graph represents potential interference induced by a pair of adjacent interfering APs. A set of colors corresponds to a set of overlapping channels. The problem of frequency assignment is then reduced to coloring the vertices such that the number of colors used is minimal and no interfering (connected) nodes have the same color. Optimal coloring is NP-hard. In [11], an algorithm based on the concept of a saturation degree (so called DSATUR) was introduced to obtain a sub-optimal solution. The saturation degree of a vertex was defined as the number of differently colored vertices to which the vertex is adjacent. These different colors used by the neighboring vertices then constitute a set of non-admissible colors for the vertex in question. The basic idea behind this algorithm is to choose for each iteration a vertex with the highest saturation degree and color it with the least admissible color. For example, assume that AP \(i\) (vertex \(i\)) with the highest saturation degree is surrounded by two APs, one of which is colored with color (channel) 1 and the other with color (channel) 6. These two colors then constitute a set of non-admissible colors for AP \(i\). For non-overlapping 802.11b channels, the next possible color (channel) that can be assigned to AP \(i\) is therefore color (channel) 11, which is first on the list of admissible colors. The coloring is greedy in the sense that the first available color in the admissible set is always selected. The algorithm continues in this manner until all the vertices are colored.

Under this algorithm, construction of an accurate graph is the key to the channel assignment. The Inter-AP Protocol (IAPP) is then needed for APs to cooperate and construct such a graph. One way of constructing such a graph is to have all the APs in the service area listen to the beacons generated by their neighboring APs. In the beacons, information such as MAC addresses of their respective senders, signal-to-noise ratios, and received signal strength can be extracted. After identifying its neighbors, each AP sends this information to the other APs in the network via a distribution system such as a legacy LAN. The algorithm then can be applied at any AP in the network either periodically or whenever new information about the network topology arrives.
make this cooperation among APs possible, recommended practice is as specified in the 802.11F. Similar approaches are reported in [23] and [24].

**CFAssign-RaC: Conflict-Free Set Coloring**

With a detailed definition and classification of channel interference, the work in [12] solved the problem of channel assignment in conjunction with the problem of load balancing. The authors emphasized the interference model that tries to capture total interference as seen by clients while employing all the available channels in a network. In effect, the model is able to capture the so-called *hidden* interference not seen from the AP perspective. In this model, interference as seen by clients originates from two different sources: (1) from those APs located within a communication range of the client of interest, regardless of channel choices of operation, and (2) from those stations, either APs or wireless clients (in neighboring BSSs), located within a one-hop distance of the AP-client link of interest. In both cases, the client is said to suffer from channel conflict.

The main idea of this strategy is to assign channels to APs so that the clients are distributed (associated to APs) in such a way that conflict is minimized. Under this scheme, conflict is used to denote scenarios where any two stations (APs or clients) belonging to different BSSs interfere with each other by the virtue of sharing the same channel [12]. By this, the problem of load balancing is implicitly addressed. Based on this formulation, a centralized algorithm CFAssign-RaC is proposed for channel assignment. The algorithm attempts to maximize the number of clients with zero conflict. The efficiency of the algorithm with respect to the number of clients with zero conflict depends on how the range and the interference sets formed by all the stations really reflect the real channel conditions in the network. Thus, accurate site reports consisting of the list of channels each client is able to hear at its present location must be submitted by all the clients to their respective APs either periodically or dynamically.

**Measurement-Based Local-Coord**

In this measurement-based approach [13], the cost function is the weighted interference that captures interference as seen by both APs and wireless clients in a local area network. To obtain such interference, wireless clients are required to physically measure the in situ interference power on all the frequency channels used in the network whenever their associated APs are idle. The clients then average this measured interference power over all the channels and report the metric to their respective APs. The APs themselves also have to measure and average this in situ interference power. The weighted interference now can be calculated for each BSS or cell. The weighing factors may include such metrics as the average traffic volume and average received signal power of the clients within the cell. That is, higher traffic volume should contribute more to the interference metric, whereas higher received signal power should contribute less because that implies higher tolerance to interference.

---

4 The 802.11F standard was administratively withdrawn, and only the documentation on recommended practice is available [22].
Clearly, when one cell switches from channel $k$ to operate on channel $k'$, this cell itself and its neighbors who operate on either $k$ or $k'$ will see changes in their respective weighted interference metrics. Based on these changes, the Local-Coord algorithm is proposed. The basic idea is that a particular cell switches from channel $k$ to channel $k'$ if and only if the switching does not increase the maximum weighted interference seen by all the neighboring cells that operate on either $k$ or $k'$. This operation therefore implies that some kind of coordination among the local cells is necessary. That is, before this particular cell can switch to a new channel, it needs to know the weighted interference metrics of its neighbors who reside on channel $k$ and $k'$. This process of channel switching continues at different cells (APs) in the network until the channel assignment converges.

One variant of this algorithm, called Global-Coord, triggers a channel switching at an AP if and only if the overall co-channel weighted interference, as defined for the Local-Coord algorithm, in a network operating on a new set of channels is lower than that on a current set of channels. As a result, this algorithm needs to be executed at a central agent that retrieves interference information from all APs in the network.

12.6 Channel Assignment Schemes in Uncoordinated Environments

Generally, IEEE 802.11 deployments are uncoordinated, where hotspots and private WLANs managed by different network administrators coexist in various densities throughout the geographical area of interest. Taking the locations of APs as given, many research works focus only on the channel assignment problem and completely ignore the AP placement problem. This concept of improving the network performance by adjusting channel assignment alone is attractive, especially in a residential environment where an ISP providing a managed wireless service will likely have little control over the locations of APs in different homes. Moreover, even if the homeowners were open to having their APs moved, it would be very expensive for the ISP to send out technicians to move them. Several channel assignment strategies have been developed just around these constraints. The main feature of such strategies in this category is the distributed execution in which channel assignment is performed distributively by each AP instead of a central controller.

12.6.1 Least Congested Channel Search (LCCS)

In Least Congested Channel Search (LCCS) [25], each AP autonomously searches for the most lightly loaded channel, e.g., the channel with the fewest number of wireless clients. It switches to operate on that channel until the next scan finds a less congested channel. To achieve this, an AP first scans each channel for distinguishable beacons published by neighboring APs, i.e., each distinguishable beacon corresponds to each individual AP. Since each beacon contains such information as the number of wireless clients associated with each AP, the AP then determines from each beacon how many clients are associated with each AP. After scanning all available channels, the AP knows
how many clients are associated with each channel and will choose to operate on the channel with the fewest number of associated clients. When using this criterion for channel assignment, the LCCS algorithm also implicitly deals with load balancing with the unrealistic assumption that every wireless client carries the same amount of traffic. That is, the higher number of associated clients indicates the higher amount of traffic.

Instead of choosing the channel with the fewest number of associated clients, [25] also suggests to use traffic-related information, also obtained from beacons, in choosing a channel for an AP. In this case, the AP will “choose to operate on the channel with the least amount of traffic, irrespective of how many clients are associated with” [25]. Such traffic-related information may include the average number of packets handled by an AP during, say, the past 5 minutes. Finally, it should be noted that these specific fields of a beacon frame, specifying the number of wireless clients associated with each AP and the average amount of traffic, are proprietary to Cisco.

12.6.2 MinMax Approach

In the MinMax scheme [26], a channel assignment problem is treated from the AP point of view based on the assumption that too heavily loaded APs to which certain channels are assigned can potentially degrade the network performance. A set of APs is assumed to be installed in an area of interest. Only downlink traffic is considered. The objective function is an AP’s effective channel utilization defined as the fraction of time a channel assigned to an AP is used for transmission by the AP or is sensed busy because of interference from its co-channel neighbors. In the proposed strategy, two classes of neighboring interferers are considered. For each AP $i$, co-channel APs are said to be in a class-1 interferer set $C_i(1)$ if their interfering signals are above a certain threshold that can cause enough interference for AP $i$ to sense its channel busy. A class-2 interferer set $C_i(2)$ is defined as a set of pairs of co-channel interfering APs in which transmission by any pair of APs in the set can cause enough interference for AP $i$ to detect its channel busy. To determine these interferer sets, the estimation of signal propagation path-loss between each pair of APs is required. Given $C_i(1)$, $C_i(2)$, $N$ non-overlapping channels and fixed traffic load $\rho_i$ at each AP $i$, the effective utilization is calculated as

$$U_i = \rho_i + \sum_{k=1}^{N} x_{ik} \left( \sum_{j=C_i(1)} \rho_j x_{jk} + \sum_{(m,n)=C_i(2)} \rho_m \rho_n x_{mk} x_{nk} \right) \tag{12.5}$$

where $x_{ij} = 1$ if a channel $j$ is assigned to AP $i$.

The objective of this approach is to minimize the maximum effective channel utilization at the most heavily loaded AP, given a different fixed traffic load at each AP. The problem is shown to be NP-complete. A heuristic algorithm is proposed to minimize this effective channel utilization of the most heavily loaded AP. Starting with a random channel assignment to all the APs in the network, the algorithm randomly readjusts the channel assignment of the bottleneck AP’s interfering neighbors (those in $C_i(1)$ and $C_i(2)$) such that the effective channel utilization of the bottleneck AP is minimized, resulting in a least congested network. Being heuristic-based, the algorithm does not
guarantee to give optimal solutions, especially in a large WLAN with several tens of APs. The algorithm can be applied after the initial locations of APs are determined.

### 12.6.3 MinMax II Approach

Based on the static model in [26], the strategy [27] attempts to assign channels in an adaptive manner to a set of APs such that the maximum channel utilization at the most overloaded AP is minimized, for a given traffic load and an interfering set of APs. An interfering set for an AP is defined as the co-channel neighbors of this AP, whose transmissions can cause enough interference for this AP to sense its channel busy. Unlike [26], the channel utilization in [27] was formulated based on a dynamic MAC model found in [28], where the estimated number of active clients (i.e., including those clients associated with the AP of interest and those under the neighboring co-channel APs) were taken into account. The channel assignment algorithm is similar to the one in [26], but it is dynamic in the sense that the estimation of the number of active clients is done in real time for each channel adjustment (assignment) period, and as a result of new channel assignment, a check on the network performance against some predefined QoS threshold is invoked at the end of the algorithm. To ensure quick convergence and to avoid infinite looping, an optimal channel switching is derived based on a Markov state transition diagram. Each AP optimizes its channel assignment simultaneously and independently without relying on inter-AP communication. The optimal solution exists for a small-scale network. For a larger network, the proposed strategy may not be scalable. Another approach similar to [27] was proposed in [29].

### 12.6.4 Hminmax/Hsum: Weighted Coloring Approach

In [30], frequency assignment was modelled as a minimum-sum weighted vertex coloring problem in which different weights are put on interference edges. Looking at interference from the clients’ perspective, this approach attempts to minimize the maximum interference as seen by clients in all common interfering regions. This interference is captured through two functions: interference factor and weight functions. While the interference-factor accounts for the amount of channel overlapping between two interfering APs, the weight can represent the number of clients confined in a common region of these two interfering APs. Mathematically, the objective is to minimize the maximum product of the interference-factor and weight function, which is referred to as the I-value.

Based on this objective, two algorithms were proposed: $H_{\text{minmax}}$ and $H_{\text{sum}}$. Requiring no coordination among APs, the $H_{\text{minmax}}$ algorithm attempts to minimize the maximum interference weight among those edges connecting directly to a vertex (AP) of interest. Each AP executes the algorithm locally and independently in a periodic manner. Due to this distributed nature, $H_{\text{minmax}}$ is most suited for c-existing WLANs managed by different administrators. In $H_{\text{sum}}$, APs are required to transmit their interference metrics to their AP peers. In this way, each AP can have a global view of the network topology. The maximum weighted interference is then minimized by $H_{\text{sum}}$ in a global
sense. Due to the IAPP requirements, \( H_{sum} \) is suitable only for WLANs supervised by the same administrator. In terms of net throughput, \( H_{sum} \) was shown to outperform \( H_{minmax} \).

### 12.6.5 Pick-Rand and Pick-First Approach

In [31], the objective is to assign overlapping channels to APs in such a way that minimizes the total weighted interference as seen by each AP. Such interference is weighted by the so-called overlapping channel interference factor, the AP transmit power, and the path-loss between two interfering APs. The overlapping channel interference factor indicates how much two interfering channels are overlapped in frequency. Each AP independently runs the algorithm that minimizes such interference, either in a periodic manner or whenever the interference rises above a specified level. Two versions of the algorithm were proposed for breaking ties between channels that give the same interference: Pick-Rand and Pick-First. While Pick-Rand randomly picks a channel for assignment, Pick-First simply picks the first channel of the ascendingly ordered channel list.

### 12.6.6 Pick-Rand and Pick-First II Approach

As an extension to [31], the work in [32] incorporates the load-balancing problem into the channel assignment framework. Wireless clients are initially characterized by their data rate requirement, hence signal strength. The AP then can decide to disassociate (break a connection with) its clients through the reduction of its transmit power of beacon packets whenever it becomes overly congested. Those clients whose data rates cannot be supported will eventually turn away and reassociate to a new AP that can accommodate their traffic requirements. A congestion indicator in this scheme is defined as the ratio of aggregate data rates, required by all currently associated clients, to the AP's available bandwidth. Mathematically, this problem is modeled as an ILP whose objective is to minimize the maximum congestion at the most congested AP. Once the optimal power levels and user associations are obtained, the channel assignment problem as formulated in [31] is invoked and directly solved without resorting to a heuristic. A similar approach can be found in [33].

### 12.6.7 Channel Hopping Approach

In [2], a distributed channel assignment algorithm based on the concept of channel hopping was proposed specifically for an uncoordinated WLAN. In particular, each AP is assigned a unique sequence of channels and hops through this sequence over time so as to average out the throughputs of all APs in a long run. This is illustrated by an example in Figure 12.7. Each AP is within transmission ranges of three other APs. The sequence assigned to each AP is also shown in Figure 12.7. Each AP hops to the next channel at the end of each time slot. Suppose that only three non-overlapping channels, namely, 1, 6, and 11, are available for assignment, and that each AP always has data to
transmit. The goal is to average out the throughputs of all APs in the long run. In the first time slot, AP3 and AP4 are assigned the same channel. Due to MAC contention, their normalized throughputs are therefore 0.5 each, while the throughputs of AP1 and AP2 are 1 each. In the second time slot, AP2 and AP3 are interfering. Their resulting throughputs are 0.5 each, and those for AP1 and AP4 are 1 each. Continuing in this manner, every AP will have the equal throughput of 0.75 at the end of the sixth time slot. Long-term fairness is then captured by this approach. The main requirement of this approach is the common notion of time among all APs. In other words, every AP must synchronize to the same clock so that channel hopping is performed by each AP at the same instant.

12.6.8 Measurement-Based No-Coord

This is another variation of Measurement based Local-Coord proposed to avoid the requirement of inter-cell coordination. The formulation is similar to that of Measurement based Local-Coord, but the algorithm itself is greedy in the sense that channel switching is based only on its own weighted interference. That is, each cell switches to operate on another channel if and only if the new weighted interference calculated based on the new channel is lower than the previous one.

12.7 Comparison among Various Channel Assignment Schemes

A summary of the various channel assignment schemes is provided in Table 12.1. In this comparison, the following aspects are considered: (1) how often channel assignment is triggered (static or adaptive), (2) to which type of deployment a channel assignment is applicable (uncoordinated or centrally managed), (3) the type of frequency channels used (overlapping or non-overlapping channels), (4) the procedure in obtaining channel assignment solutions (heuristic or integer linear programming), (5) from which perspective interference is modeled (from AP’s or clients’ point of view), (6) whether Inter-AP communication is needed as part of channel assignment, and (7) whether a scheme is scalable (i.e., whether a solution exists for relatively large networks). According to Table 12.1, all the static schemes rely on a centralized control, while the adaptive schemes reside on both centrally managed and uncoordinated deployments. The solutions for the static schemes are usually executed only once, so that their complexity
Table 12.1 Summary of Channel Assignment Schemes in 802.11 WLAN

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[15]</td>
</tr>
<tr>
<td>Priority-Map</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[18]</td>
</tr>
<tr>
<td>MinMax</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[26]</td>
</tr>
<tr>
<td>ILP-based</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[17]</td>
</tr>
<tr>
<td>Patching</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[19]</td>
</tr>
<tr>
<td>Cov.-oriented</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[21]</td>
</tr>
<tr>
<td>Local-Coord</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>no</td>
<td>[13]</td>
</tr>
<tr>
<td>Global-Coord</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>no</td>
<td>[13]</td>
</tr>
<tr>
<td>No-Coord</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[13]</td>
</tr>
<tr>
<td>MinMax II</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>no</td>
<td>[27, 29]</td>
</tr>
<tr>
<td>Conflict Set</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>no</td>
<td>[12]</td>
</tr>
<tr>
<td>LCCS</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>yes</td>
<td>[25]</td>
</tr>
<tr>
<td>DSATUR</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>yes</td>
<td>[11]</td>
</tr>
<tr>
<td>Hsum</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>yes</td>
<td>[30]</td>
</tr>
<tr>
<td>Hminmax</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>yes</td>
<td>[30]</td>
</tr>
<tr>
<td>Pick-Rand/Pick-1st</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>yes</td>
<td>yes</td>
<td>[31]</td>
</tr>
<tr>
<td>Pick-Rand/Pick-1st II</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>yes</td>
<td>[32]</td>
</tr>
<tr>
<td>Channel Hopping</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>no</td>
<td>yes</td>
<td>[2]</td>
</tr>
</tbody>
</table>

is of little concern. When the planning is done, the channels are established for long-term use. Most schemes consider non-overlapping channels while the new trend tries to utilize both non-overlapping channels and overlapping ones [34]. Due to high computational complexity, most schemes resort to heuristic solutions. While the early schemes view interference only from the AP’s point of view, recent schemes try to model interference also from clients’ perspectives. As can be expected, all the channel assignment schemes in the uncoordinated environments are scalable due to their distributed execution, whereas those in the centrally managed environments are not because of their privileged centralized control.

In addition, the comparison among relevant schemes in terms of performance is highlighted as follows. It was reported that the fully distributed Channel Hopping approach [2] outperforms LCCS [25] in terms of fairness (measured by Jain’s index) and throughput by 42% and 30%, respectively. Its performance is, however, comparable to the fully centralized CFAssign-Rac approach [12]. Further, CFRacAssign [12] is shown to outperform both the weighted coloring (Hminmax/Hsum) [30] and LCCS [25] approaches in terms of application-level throughput and percentage of MAC collisions. The weighted coloring (Hminmax/Hsum) [30] in turn outperforms LCCS [25] in
terms of throughput. It is also reported that the measurement-based *Local-Coord* algorithm [13] gives higher average throughput than the CFRacAssign [12], with the price in signaling overhead to pay. Compared to the classical cellular results, the optimal channel assignment obtained from the *MinMax* approach [26] is also shown to match that for the cellular network with a frequency re-use factor of three for small networks. However, only suboptimal assignment is obtained when the proposed scheme is applied to large networks (≥ 111 APs). Similar results are also reported in [27, 29], where the overall network throughput in this case is also shown to improve over that of the random channel assignment by 65%. Compared to the random channel assignment scheme, the *Vertex Coloring* approach [11, 23, 24] is shown to give higher UDP/TCP aggregate network throughput, especially when the number of APs increases. The *Coverage-Oriented* approach [21] is also shown to outperform a random channel assignment scheme and sequential network planning approaches such as [15, 16], in terms of radio coverage and average throughput. In [31], the *Pick-Rand and Pick-First* approach is shown to outperform the worst-case channel assignment scheme, in which all APs are assigned the same channel, in terms of total interference by a factor of four. In [19], the *Patching* algorithm slightly improves overall throughput over both the ILP-based [17] and traditional [16] approaches in terms of fairness (Jain’s index) and throughput. Further, it has been shown that [32] improves over [17] in terms of load distribution and interference reduction (as much as 4%).

### 12.8 Current Practice in Channel Assignment

Current practice in channel assignment for infrastructure-based 802.11 networks is still far behind current research described through several schemes in Sections 12.5 and 12.6. Most Cisco AP products, for example, still employ LCCS in randomly searching for least congested channels at power-up or when the other parameters of radio interfaces are changed [35]. (The operations of LCCS are described in Section 12.6.1.) In LCCS, human intervention is still required in specifying which channels to ignore or search on. In other AP products, such automatic channel searching is not even available, and it is left to users or network administrators to manually assign channels to APs at the network planning phase or at later stages. Although products such as Cisco WLAN Controllers claim to use a more sophisticated method [36] in assigning channels to APs, which takes into account a variety of network dynamics such as AP received energy, noise, interference, channel utilization, and client load, the centralized operations of the method, however, limit its applicability only to centrally managed networks. Detailed operations and implementation also are not available to public. Another sophisticated method described in Section 12.5.2 and [37] has also been illustrated to work well in practice, but, to the best of our knowledge, has not yet been implemented in real products. This method also suffers from its centralized operations. Being centralized, both of these methods are not applicable to uncoordinated networks, which keep populating in recent years in places like residential neighborhoods, public hotspots, adjacent offices, etc. The recent trend in research then tends to focus on channel assignment schemes
that work distributively across networks overseen by different network administrators and that require less human intervention.

12.9 Open Research Issues

Channel assignment is one mechanism to improve the performance of WLANs. In this survey we discuss several existing channel assignment schemes applicable to either centrally managed or uncoordinated environments. Several possible future research directions and open issues with regard to channel assignment in WLANs are outlined below:

- The research direction tends to shift toward adaptive channel assignment in uncoordinated environments, in which network dynamics is incorporated into the problem formulation. The following system parameters need to be considered: client locations, building layouts and AP locations, time fluctuation of traffic demand of wireless clients at various locations, and application QoS requirements. The challenge would be how to capture the network dynamics as much as possible while maintaining the complexity of implementation of channel assignment algorithm at a practical level. Furthermore, when WLANs are deployed in an uncoordinated fashion by different network administrators, the scalability of the implementation of channel assignment algorithms becomes even a more important issue. In such scenarios, a channel assignment scheme of choice should be cooperative and scalable enough to orchestrate channel switching across the entire network without creating significant interference to the neighbors. Being aware of the neighboring networks located in different administrative domains, the scheme should also be able to interact and exchange necessary information (network topology, channels in use, the number of clients, etc.) with its neighbors in order to allocate appropriate channels to the APs.

- Continually monitoring the network dynamics, say, on a daily basis, at a particular location may lead to a discovery of traffic pattern. Channel assignment can then be performed at a particular location during a particular period based on the prediction (or self-learning experience) as well as the application requirements.

- The schemes discussed in this chapter assume either uplink or downlink traffic. To be more realistic, traffic in both directions should be considered. This is reasonable as peer-to-peer communications become more popular.

- Implementation issues with regard to channel switching should be considered. With current hardware, the time required to switch a channel at an AP ranges from 100 µs to 20 ms [2]. As an AP switches to operate on another channel, its associated clients should then be notified and/or handed over properly to another AP within this time limit. Ongoing communications, especially delay-sensitive applications, should be interrupted as little as possible. The concepts of hard and soft handoffs as employed in cellular networks could be developed. Also, for feasible implementation, installing any new channel assignment strategy in a network should result only in a driver update at APs, but not major hardware changes.
Additionally, all the schemes discussed thus far assume fixed transmit power. In practice, this may not be the case since transmit power may vary with channel conditions and application QoS requirements (e.g., data rates). Joint optimization of channel assignment and power control may lead to more efficient utilization of radio channels. With the growing demand for multimedia contents in WLANs, channel assignment that takes power control into consideration is an interesting avenue for research.

With the emerging concept of spectrum aggregation in cellular networks across both licensed and unlicensed bands [38] as well as cellular technologies such as LTE-Unlicensed (LTE-U) [39] which intend to use the unlicensed 5-GHz band, dynamic channel assignments for WLANs may also need to consider the usage of the spectrum bands by cellular networks.

References


Part V

Cross-Layer Modeling for
Resource Allocation in
Wireless Networks
13 Joint PHY/RLC Design in Cellular Wireless Networks

13.1 Introduction

Design and analysis of ARQ and HARQ protocols for error control in the wireless environment has been a very active research topic over several decades. While the ARQ protocols simply aim at retransmitting erroneous packets to achieve the desired reliability, HARQ protocols jointly exploit physical layer techniques including diversity combining, error control coding, and retransmission mechanisms of traditional ARQ protocols to achieve better performance. In particular, HARQ protocols are usually more throughput and delay efficient than ARQ protocols since HARQ protocols adaptively adjust the redundancy level on the fly to achieve reliable data transmission. Moreover, HARQ protocols are able to exploit useful information accumulated over different transmission attempts to efficiently decode data, which are typically overlooked by ARQ protocols.

Much research has been conducted to develop efficient ARQ and HARQ protocols for wireless channels with different characteristics such as large round trip delay, high loss rate, and long bursts of errors. For HARQ protocols, devising a suitable coding strategy to be employed with the underlying retransmission and decoding mechanisms has been of great interest. Moreover, performance analysis of ARQ and HARQ protocols over different communication channels such as channels with independent and identically distributed (i.i.d.) and correlated error patterns has resulted in large body of literature. More recently, there has been interest in jointly designing ARQ/HARQ protocols with advanced physical layer techniques such as adaptive modulation and coding.

This chapter covers fundamental design and analysis aspects of ARQ and HARQ protocols. Different types of ARQ and HARQ protocols with their own characteristics are described and highlighted. In addition, we will discuss important models for performance analysis and cross-layer design of different ARQ and HARQ protocols over diverse communication channels. While Chapter 2 mainly described the throughput of basic ARQ protocols in communication environments with simple error patterns, the analysis presented in this chapter considers error-prone wireless channels with different characteristics.

13.2 Radio Link Control (RLC) Protocols: ARQ and HARQ

An ARQ protocol is typically employed at the RLC layer to eliminate residual errors over an error-prone communication channel and to alleviate the costly use of a strong
error correction code at the physical layer [1–4]. Among the three basic ARQ protocols (namely, stop-and-wait (SW-ARQ), go-back-N (GBN-ARQ), and selective repeat (SR-ARQ)), SR-ARQ is the most efficient. GBN-ARQ is less efficient than SR-ARQ, but its implementation is simpler than SR-ARQ because packets are always received in order in the receiving buffer.

While most wireless data applications are somewhat delay-tolerant, which creates opportunities and flexibility for system design, we usually want to limit the RLC protocol delay to provide QoS support for users. Therefore, the number of retransmissions is usually limited by a maximum value, and corresponding protocols are referred to as truncated ARQ protocols [5]. By setting the maximum number of retransmissions for the truncated ARQ protocols appropriately, one can achieve desirable tradeoff between the end-to-end delay and protocol performance.

For practical implementation of ARQ protocols, data information must be encoded by an error-detecting (ED) code such as a cyclic redundancy check (CRC) code, and the resulting packets comprising both information bits and redundant bits of the ED code are transmitted over the communication channel. The ED code is employed to verify if the packet is received correctly or not at the receiver side based on which, either an ACK or NAK, message is fed back to the transmitter.

In practice, data information is also encoded by a forward error correction (FEC) code, and the combination of FEC and ARQ protocol results in a corresponding HARQ protocol [6–11]. There are two main types of HARQ protocols, which are called Type I and Type II HARQ. Moreover, detailed design and implementation of HARQ protocols can be very different, and depend on the specific choice of FEC codes, methods to form data packets to be transmitted over different transmission rounds, and decoding techniques employed at the receiver.

In Type I HARQ, information data are encoded by both ED and FEC codes to form a transmitted packet. Then the transmitted packet is transmitted repeatedly using one of three standard ARQ protocols. Specifically, in each transmission attempt, the receiver decodes the received packet and corrects errors using the FEC code. If the transmission errors are correctable, the receiver can obtain the correct packet. Otherwise, the receiver will reject the received packet and send back a NAK message to request the retransmission.

For Type II HARQ, incorrectly received packets at different transmission attempts are stored at the receiver rather than discarded. Moreover, the receiver combines the received messages before performing decoding in each transmission round. Therefore, the decoding succeeds with higher probability over consequent attempts. There are two main methods for packet combining: Chase combining and incremental redundancy. For Chase combining, every re-transmission contains the same information including data and parity bits, and the receiver employs the maximum-ratio combining or other combining methods to obtain the combined baseband signal to be used in the decoding process [6].

The HARQ protocol with incremental redundancy transmits different sets of coded bits over different transmission attempts, which are typically obtained by puncturing the
13.3 Link Adaptation with Adaptive Modulation and Coding (AMC)

Achieving high-speed transmission and provisioning of QoS for emerging data-oriented wireless applications through intelligent and flexible radio resource management are among key research challenges for future-generation wireless networks. Adaptive modulation and coding (AMC) is an important technology that has been adopted in many recent wireless standards. The key idea behind AMC is to adaptively change the joint modulation and coding scheme according to the time-varying channel condition [12–15]. AMC is a link adaptation technique since it enables to adapt the communication rate to the varying wireless channel quality.

In fact, most 2.5/3G and beyond wireless systems predetermine a set of modulation and coding schemes (MCS), each of which corresponds to unique modulation scheme and coding strategy with the specific code rate [16–18]. For example, in the LTE wireless cellular system, the data to be transmitted are encoded by the turbo code with different code rates and modulated using one of the following schemes: quadrature-phase shift keying (QPSK), 16-QAM, or 64-QAM, followed by OFDM modulation.

For practical implementation, the receiver estimates the channel quality and transmits this channel state information (CSI) to the transmitter to choose the suitable MCS. In other words, the link adaptation technique via AMC adapts the transmission rate over the wireless channel with the signal to interference plus noise ratio (SINR). To perform link adaptation, the received SINR \( x \) is partitioned into a finite number of intervals. Let \( X_0 (= 0) < X_1 < X_2 < \cdots < X_{K+1} (= \infty) \) be the thresholds of the received SNR for these different intervals, each of which corresponds to one predetermined MCS. The channel is said to be in state \( k \) if \( X_k \leq x < X_{k+1} \) \( (k = 0, 1, 2, \ldots, K) \). For example, in channel state \( k \), the modulation scheme \( 2^k\)-QAM, \( k = 1, 2, \ldots, K \) can be chosen. In state 0, no transmission is allowed to avoid high probability of transmission error. Each MSC is also referred to as a transmission mode, which is mapped to one unique channel state.

We now discuss the transmission errors and one particular method to choose the SINR partition points \( X_k \), which are useful for performance analysis of RLC protocols. For coded systems, the closed-form derivation for the packet error rate (PER) is not easy, so some suitable approximation should be proposed for design and analysis purposes.
One such approximation is proposed in [5] as follows:

\[
\text{PER}_k(x) \approx \begin{cases} 
1, & \text{if } 0 < x < X_{pk} \\
 a_k \exp(-g_k x), & \text{if } x \geq X_{pk}
\end{cases}
\]  

(13.1)

where \(a_k, g_k, \text{ and } X_{pk}\) are obtained by curve fitting [5]. In fact, the similar approximations for bit error rate (BER) has been widely used in the literature [14], [15].

The average PER for each mode \(k\) can be then calculated by using the probability density function (PDF) of the received SINR. Under the general Nakagami-\(m\) fading channel, the PDF of the received SINR \(x\) can be written as follows [15]:

\[
p_X(x) = \frac{m^m x^{m-1}}{\bar{x}^m \Gamma(m)} \exp\left(-\frac{mx}{\bar{x}}\right)
\]  

(13.2)

where \(\bar{x} = E\{x\}\) is the average SNR, \(\Gamma(m) = \int_0^\infty t^{m-1} \exp(-t)dt\) is the Gamma function, and \(m\) is the Nakagami fading parameter \((m \geq 1/2)\). The Nakagami fading model is very general and includes the Rayleigh fading channel as a special case when \(m = 1\), and the Ricean fading can also be approximated by the Nakagami model [19].

Using this PDF of the received SINR \(x\), the average PER for each mode \(k\) can be calculated as

\[
\text{PER}_k = \int_{X_k}^{X_k+1} \frac{a_k}{\text{Pr}(k)} \left(\frac{m}{\bar{x}}\right)^m \frac{\Gamma(m, b_k X_k) - \Gamma(m, b_k X_{k+1})}{(b_k)^m} \exp\left(-\frac{mx}{\bar{x}}\right) dx
\]  

(13.3)

where \(k = 1, \ldots, K\), \(b_k = m/\bar{x} + g_k\), and \(\text{Pr}(k)\) is the probability of channel state \(k\), which can be calculated as [15]

\[
\text{Pr}(k) = \int_{X_k}^{X_k+1} p_X(x) dx = \frac{\Gamma(m, m \bar{x}/x) - \Gamma(m, m \bar{x}_{k+1}/x)}{\Gamma(m)}
\]  

(13.4)

where \(\Gamma(m, x) = \int_x^\infty t^{m-1} \exp(-t)dt\) is the complementary incomplete Gamma function. In practice, the communication performance degrades due to feedback delay and errors. The impacts of feedback errors and delay on the bit error rate performance of AMC schemes have been investigated in the literature [12], [13].

The packet error rate calculation presented above has completely modeled the physical layer of a multi-rate wireless network. One important issue is to determine the SINR thresholds \(X_k\) to realize specific design goals. For example, it was proposed in [4] and [20] that one can calculate these thresholds such that all channel states have the same probability. In [21], a search algorithm was proposed to find the SINR thresholds while constraining \(\text{PER}_k = P_0\). The SINR thresholds chosen by this algorithm allow to guarantee the PER at the physical layer. The algorithm can be described as follows:

**Algorithm 13.1: SINR threshold search algorithm**

1. Set \(k = K\) and \(X_{K+1} = \infty\).
2. For each \(k\), search for \(X_k \in [0, X_{k+1}]\) such that \(\text{PER}_k = P_0\).
3. If \(k > 1\), go to 2, otherwise go to 4.
4. Set \(X_0 = 0\).
Here the design parameter is $P_0$ which can be optimized to achieve good performance for some desired performance measures and also allows cross-layer design and optimization.

## 13.4 Channel Modeling

Performance analysis and design of ARQ/HARQ protocols are typically conducted for some specific channel model, which must balance between accuracy and tractability. We describe some important wireless channel models, which will be employed later in the analytical models of different ARQ/HARQ protocols. In general, there is a correlation in the channel status over consecutive packet transmissions, and the level of such a correlation depends on how fast the channel varies over time. If the channel correlation over two consecutive packet transmissions is sufficiently small, then the independent and identically distributed (i.i.d.) channel model can be adopted. Otherwise, Markov channels can be employed for the highly correlated wireless channels. In the following, we describe both i.i.d. and Markov channel models in more detail where channel evolutions over fixed-size transmission time slots are considered. We adopt an assumption that the channel state remains static in one time slot but may change in consecutive time slots where one packet and multiple packets can be transmitted in one time slot depending on the physical layer specifics.

### 13.4.1 I.I.D. Channel Models

In the i.i.d. channel model, the channel state in each time slot is one of the possible channel states, and it is independent with channel states in the previous time slots. The number of channel states depends on specific physical layer design. In particular, if AMC is employed at the physical layer, then each channel state $k \geq 1$ corresponds to one SINR interval where a specific modulation and coding scheme is employed. The channel state probability and average PER can be calculated as in (13.4) and (13.3), respectively. In contrast, if a single fixed modulation and coding scheme is employed at the physical layer, then the wireless channel can be assumed to induce certain packet error probability $p_e$ independently in each time slot. In addition, the error probability $p_e$ can be calculated based on the employed MCS and underlying wireless fading channel (e.g., Rayleigh, Nakagami-m fading channels).

### 13.4.2 Two-State Markov Channel Model

To capture channel correlation with bursty errors, the Markov channel with a good and bad channel state was first proposed by Gilbert and Elliott in [22] and [23]. This channel is mostly useful to model the correlated transmission outcomes (success or failure) for the constant rate communication scenario where one packet is transmitted in each fixed-length time slot. More recently, Zorzi et al. have validated this channel
model and have shown how to calculate its parameters [3]. Under this two-state Markov channel, the wireless channel is assumed to be in either the good or bad state, and the channel state evolves according to a two-state Markov chain.

Let $X(t) \in \{0, 1\}$ denote the channel state in time slot $t$ where $X(t) = 0$ means the channel is in the good state and the packet transmission is successful, and $X(t) = 1$ implies the transmission error in the bad channel state. The transition probability matrix of the two-state Markov chain describing the wireless channel can be written as

$$ T = \begin{bmatrix} p & q \\ r & s \end{bmatrix}. $$

(13.5)

Here we assume that states 1 and 2 are good and bad states, respectively. Therefore, $p$ and $r$ represent the probabilities that the transmission in a particular time slot is successful given the transmission in the previous time slot is successful and unsuccessful, respectively. Parameters $q$ and $s$ can be explained similarly. Moreover, we have $q = 1 - p$ and $s = 1 - r$ according to the standard property of the transition probability matrix. From the stationary analysis of this Markov chain, we can calculate the probability of transmission error (i.e., the stationary probability of state two) as

$$ p_e = 1 - \frac{r}{1 - p + r}. $$

(13.6)

Note also that $1/r$ represents the average length of a burst of errors.

To completely characterize the two-state Markov channel, we need to determine two transition probabilities among the four probabilities $p, q, r, s$. Equivalently, we can calculate one transition probability $p$ or $r$ and the probability of transmission error $p_e$, and then by using the relationship in (13.6) we can determine the remaining transition probabilities. We describe how to complete such calculation in the following.

In the wireless channel, the fading coefficient $\alpha$ at one particular sampling instant, which is a complex number, is a random variable. The amplitude of such fading coefficient, i.e., $v = |\alpha|$, plays an important role in the transmission error process. Let $P_e(v)$ denote the conditional probability that the packet transmission is in error for a given $v$, which is typically a function of the employed modulation scheme and forward error correction (FEC) code. Then we can calculate the average probability transmission error as

$$ p_e = \Pr(X = 1) = \Pr[1] = \int_0^\infty P_e(x)f_v(x)\,dx $$

(13.7)

where $f_v(x)$ represents the probability density function (PDF) of the fading envelope $v$, and we have denoted the probability of the bad state as $\Pr(X = 1) = \Pr[1]$ for brevity.

Moreover, the probability that the two successive packet transmissions are both in error is given by

$$ \Pr[1, 1] = \int_0^\infty \int_0^\infty P_e(x_1)P_e(x_2)f_{v_1,v_2}(x_1, x_2)\,dx_1\,dx_2 $$

(13.8)
where \( f_{v_1v_2}(x_1,x_2) \) is the joint PDF of the fading envelops sampled at the two consecutive packet transmission time slots. Then we have

\[
r = 1 - \Pr[1|1] = 1 - \frac{\Pr[1,1]}{\Pr[1]} = 1 - \frac{\Pr[1,1]}{p_e}.
\]

(13.9)

Therefore, by using the results in (13.7) and (13.8), we can calculate the transition probability \( r \), and using this result and \( p_e \) given in (13.7) allows us to determine all parameters of the two-state Markov channel.

For performance evaluation of ARQ protocols, it is mostly sufficient to employ a threshold model to describe the packet transmission outcome [3]. Specifically, the binary error process \( X \) can be modeled as

\[
X = \begin{cases} 
0, & \text{if } v^2 > b \\
1, & \text{if } v^2 \leq b
\end{cases}
\]

(13.10)

where \( b \) denotes the power threshold and “0” and “1” stand for a packet transmission success and error, respectively. The related quantity \( F = 1/b \) is referred to as fading margin, which describes the maximum fading attenuation that the system can tolerate [3]. The parameter \( b \) or \( F \) in the threshold error model must be determined for a given channel with a certain average signal to noise ratio so that the average packet error probability resulting from this threshold model is equal to that calculated by using the original conditional packet error probability function \( P_e(v) \).

If \( F_0(x) \) represents the cumulative density function (cdf) of the fading envelop \( v \) and \( F_{u_1v_2}(x,y) \) represents the joint cdf of fading envelops in two consecutive transmission time slots, then the average transmission error probability \( p_e \) and \( P[1,1] \) under the threshold error model (13.10) can be calculated as

\[
p_e = F_0(\sqrt{b}) \tag{13.11}
\]

\[
P[1,1] = F_{u_1v_2}(\sqrt{b}, \sqrt{b}). \tag{13.12}
\]

For the case of Rayleigh fading channel, the PDF of the fading envelope for the case \( \mathbb{E}(v^2) = 1 \) is

\[
f_v(x) = 2xe^{-x^2} \tag{13.13}
\]

and the joint PDF is

\[
f_{u_1v_2}(x,y) = \frac{xy}{1 - \rho^2}e^{-(x^2+y^2)/2(1-\rho^2)} I_0 \left( \frac{\rho xy}{1 - \rho^2} \right) \tag{13.14}
\]

where \( \rho = J_0(2\pi f_d T_s) \) is the correlation coefficient of the fading envelopes in two consecutive time slots, \( f_d \) is the Doppler shift, \( T_s \) is the time slot interval, and \( J_0(.) \) and \( I_0(.) \) denote the Bessel function and the modified Bessel function of the first kind, respectively, both of zero-th order. Using these results, we have

\[
p_e = 1 - e^{-b} \tag{13.15}
\]

\[
r = \frac{Q(\theta, \rho \theta) - Q(\rho \theta, \theta)}{e^b - 1} \tag{13.16}
\]
where
\[ \theta = \sqrt{\frac{2b}{1 - \rho^2}} \] (13.17)
and \( Q(.) \) is the Marcum \( Q \) function.

Moreover, once the threshold parameter \( b \) has been determined, the four transition probabilities of the two-state Markov channel can also be obtained readily via computer simulation. This simulation approach has the advantages of being simple and applicable to any fading environment where the distribution of the fading envelope may be unknown. One can accomplish this by implementing a fading envelope generator, and the generated values are compared to the threshold value \( b \) as described in (13.10) to obtain the error process.

### 13.4.3 Finite-State Markov Channel Model

The more general model of the above two-state Markov channel is the Finite-State Markov Channel (FSMC) model where the channel state is in a finite set of states. Moreover, the channel state evolves according to a finite-state Markov chain [24]. This FSMC model is especially useful for performance analysis and design of wireless systems employing adaptive modulation and coding since wireless channel states can be naturally mapped to corresponding transmission modes (i.e., difference MCSs) [25]. Moreover, the FSMC model can also be employed for performance analysis of ARQ protocols in the fixed-rate communications system if high accuracy is required [26]. This FSMC model [20, 21, 27, 28] was described in detail in Chapter 1 (see Section 1.1.6).

### 13.5 ARQ Protocols with I.I.D. Errors

We first discuss the throughput analysis for three ARQ protocols under the fixed rate and i.i.d. error scenario. This is the direct extension of the analysis in communication environments with simple error patterns in Chapter 2. Throughput is defined as the average number of bits transmitted successfully per second. Let \( L_p \) and \( L_b \) denote the number of total encoded bits and the number of information bits per RLC packet, respectively. The packet transmission error probability is \( p_e \), and errors are assumed to be i.i.d. We further assume that the feedback channel is fully reliable, i.e., the ACK or NAK message is always received successfully at the transmitter. Moreover, suppose that the round trip delay is \( T_{RTT} \) and the time slot interval is \( T_s \) seconds. Finally, we consider both cases in which the maximum number of transmissions per packet is \( N_{\text{max}} \) (truncated ARQ) or unbounded (infinitely persistent ARQ). Most results in this section are taken from the analysis presented in [29].
Then the average number of transmissions per packet for truncated ARQ protocols can be written as [5]

\[ E_T = 1 + p_e + p_e^2 + \cdots + p_e^{N_{\text{max}}-1} = \frac{1 - p_e^{N_{\text{max}}}}{1 - p_e}. \]  

(13.18)

Hence, the average number of transmissions per packet for infinitely persistent ARQ protocols becomes

\[ E_\infty = \frac{1}{1 - p_e} \]  

(13.19)

which can be obtained from (13.18) by setting \( N_{\text{max}} \to \infty \). Using this result, the throughput of truncated and infinitely persistent Stop-and-Wait ARQ (SW-ARQ) can be calculated respectively as [29]

\[ \eta_{\text{SW-ARQ}}^{\text{tr}} = \frac{L_b}{(L_p + T_{\text{RTT}}R)E_T} \]  

(13.20)

\[ = \frac{L_b/L_p}{(1 + T_{\text{RTT}}R/L_p)E_T} \]  

(13.21)

\[ \eta_{\text{SW-ARQ}}^{\infty} = \frac{L_b}{(L_p + T_{\text{RTT}}R)E_\infty} \]  

(13.22)

\[ = \frac{L_b/L_p(1 - p_e)}{(1 + T_{\text{RTT}}R/L_p)} \]  

(13.23)

where \( R \) denotes the communication rate. Let us define \( n = 1 + T_{\text{RTT}}R/L_p \); then the throughput of truncated and infinitely persistent SW-ARQ protocols becomes

\[ \eta_{\text{SW-ARQ}}^{\text{tr}} = \frac{L_b/L_p}{nE_T} \]  

(13.24)

\[ \eta_{\text{SW-ARQ}}^{\infty} = \frac{L_b/L_p(1 - p_e)}{n}. \]  

(13.25)

For simplicity, we will assume that \( n \) is an integer number throughout this chapter since \( n \) can be approximated by \([1 + T_{\text{RTT}}R/L_p]\). This is a reasonable approximation for large round trip delay (i.e., \( n \gg 1 \)).

For GBN-ARQ, all packets after each erroneous packet need to be retransmitted. Therefore, the throughput of truncated and infinitely persistent GBN-ARQ protocols can be calculated, respectively, as

\[ \eta_{\text{GBN-ARQ}}^{\text{tr}} = \frac{L_b}{(L_p + p_eT_{\text{RTT}}R)E_T} \]  

(13.26)

\[ \eta_{\text{GBN-ARQ}}^{\infty} = \frac{L_b}{(L_p + p_eT_{\text{RTT}}R)E_\infty}. \]  

(13.27)

GBN-ARQ protocols achieve higher throughput than the SW-ARQ counterparts since the effective round trip time of GBN-ARQ protocols is only \( p_eT_{\text{RTT}} \). The throughput of
truncated and infinitely persistent GBN-ARQ protocols can be rewritten as

\[
\eta_{\text{GBN-ARQ}}^\text{tr} = \frac{L_b/L_p}{(1 + p_e T_{\text{RTT}} R/L_p) E_\text{tr}} 
\]

(13.28)

\[
= \frac{L_b/L_p}{(1 - p_e + p_e n) E_\text{tr}} 
\]

(13.29)

\[
\eta_{\text{GBN-ARQ}}^\infty = \frac{L_b/L_p}{(1 + p_e T_{\text{RTT}} R/L_p) E_\infty} 
\]

(13.30)

\[
= \frac{L_b/L_p(1 - p_e)}{(1 - p_e + p_e n)}. 
\]

(13.31)

For the SR-ARQ protocol, the transmitter transmits packets continuously and only retransmits an erroneous packet upon receiving its NAK message. Therefore, the throughput of truncated and infinitely persistent SR-ARQ protocols can be expressed, respectively, as

\[
\eta_{\text{SR-ARQ}}^\text{tr} = \frac{L_b}{L_p E_\text{tr}} 
\]

(13.32)

\[
\eta_{\text{SR-ARQ}}^\infty = \frac{L_b}{L_p E_\infty} = \frac{L_b}{L_p(1 - p_e)}. 
\]

(13.33)

It can be verified that the throughput of SR-ARQ protocol is larger than the throughput of GBN and SW-ARQ protocol.

13.6 ARQ Protocols in Two-State Markov Channel

We now study the throughput analysis of ARQ protocols over the two-state Markov channel. Channel memory captured by the Markov channel has significant impacts on the throughput efficiency of the ARQ protocols. Moreover, the throughput analysis under this two-state Markov channel is more complicated than that for the i.i.d. channel studied previously since we have to track both protocol operations and the evolutions of the channel state. We are interested only in the GBN-ARQ and SR-ARQ protocols since they are more efficient than the SW-ARQ protocol. Both scenarios where the feedback channel is fully reliable and erroneous will be considered. It is again assumed that each packet transmission is equal to one time slot interval \( T_s \).

The analysis for both GBN-ARQ and SR-ARQ protocols is performed by assuming that the round trip delay is equal to \( n - 1 \) time slots, i.e., the rounded ratio between the round trip time \( T_{\text{RTT}} \) and the time slot interval \( T_s \) satisfies \( n = 1 + \lceil T_{\text{RTT}}/T_s \rceil \). The throughput \( \eta \) is described via the throughput efficiency measure \( \eta^e \), which is defined as the ratio of the number of correctly received packets and the number of transmitted packets over a long period of time. Then the throughput \( \eta_X \) of protocol \( X \) is given as

\[
\eta_X = \frac{L_p}{L_p} \eta^e_X 
\]

(13.34)
where we recall that $L_p$ and $L_b$ denote the total number of encoded bits and the number of information bits per RLC packet, respectively, and $X$ stands for GBN-ARQ or SR-ARQ protocol.

In general, for channels with unreliable feedback, each successful ACK/NAK message can be employed to carry communication status for all previously transmitted packets. Moreover, each packet can be associated with a timer so that it can be transmitted at timeout if the transmitter has not received its status [30, 31]. While these enhanced features can potentially improve the throughput performance, we consider the simpler version of ARQ protocols for which the transmitter simply retransmits a packet if its feedback message is lost [32, 33].

13.6.1 GBN-ARQ Protocol in Two-State Markov Channel

From the throughput analysis perspective, packet transmissions under the GBN-ARQ protocol can be grouped into cycles where each cycle commences with an erroneous packet followed by $n - 1$ packets whose transmission outcomes are not important for the throughput analysis since they will be retransmitted by the transmitter according to the GBN-ARQ protocol. Then the cycle is terminated by the first erroneous transmission after the $n$-th packet. It is also noted that if the cycle length is $k \geq n$ packets, then only $k - n$ packets after the $n$-th packet in the cycle are accepted by the receiver.

The throughput efficiency can be derived by studying the limiting behavior of the system considering relevant quantities in a single cycle [32]. Let $\bar{A}$ denote the average number of accepted packets during a single cycle. Then the throughput efficiency of the GBN-ARQ protocol can be expressed as

$$\eta_{\text{GBN-ARQ}} = \frac{\bar{A}}{n + \bar{A}}.$$ (13.35)

Therefore, one must determine $\bar{A}$ to complete the analysis. In the following, we consider the analysis for two different scenarios. In the first scenario, we assume that the error process in the forward channel (i.e., from the transmitter to the receiver) follows the two-state Markov channel while the backward channel (i.e., from the receiver to the transmitter) is fully reliable (i.e., no feedback error). In the second scenario, forward and backward error processes follow two independent two-state Markov channels. The transition probability matrix of the forward Markov channel is given as

$$T_f = \begin{bmatrix} p & q \\ r & s \end{bmatrix}. \quad (13.36)$$

To analyze the throughput in the first scenario, we define $\varphi = p + s$, which captures the clustering of packets of the same type (success or failure). In addition, we need to employ the $n$-step transition probability matrix of the error process, which gives transition probabilities of two transmissions $n$ slots apart. This $n$-step transition probability matrix $T_f(n)$ can be written as

$$T_f(n) = \begin{bmatrix} p(n) & q(n) \\ r(n) & s(n) \end{bmatrix}. \quad (13.37)$$
This \( n \)-step transition probability matrix can be derived as [32]
\[
T_f(n) = T_f^n = \left[ \begin{array}{cc} r & q \\ q & r \end{array} \right] + \left[ \begin{array}{cc} (1 - r - q)^n/(r + q) \\ -r \\ r \end{array} \right] \left[ \begin{array}{c} q \\ r \end{array} \right].
\] (13.38)

Then we can calculate \( A \) for the first scenario as
\[
\overline{A}_1 = qr(n) \sum_{k=1}^{\infty} k p^{k-1} = r(n)/q.
\] (13.39)

This expression accounts for the probability that the system evolves from error-free state to erroneous state in one slot with probability \( q \), comes back to the error-free state after \( n \) slots with probability \( r(n) \), then stays in the error-free state for \( k - 1 \) more slots. Using this result and the fact that \( r + q = 2 - \varphi \), the throughput efficiency of the GBN-ARQ protocol in this scenario is given as
\[
\eta_{\text{GBN-ARQ}}^{0,1} = \frac{r[1 - (\varphi - 1)^n]}{r[1 - (\varphi - 1)^n] + m(2 - \varphi)(2 - \varphi - r)}.
\] (13.40)

We now study the throughput for the second scenario, which considers both forward and backward errors. Suppose that the transition matrix of the two-state Markov channel of the backward channel is given as
\[
T_b = \left[ \begin{array}{cc} a & b \\ c & d \end{array} \right].
\] (13.41)

In order to derive the throughput efficiency, we consider the transition probability matrix of the combined forward and backward channel states \((X, Y)\) where \( X \in \{0, 1\} \) and \( Y \in \{0, 1\} \) represent the states of the forward and backward channels, respectively. Since the two channels are independent, the combined transition probability matrix can be written as
\[
T_c = \left[ \begin{array}{cccc} pa & pb & qa & qb \\ pc & pd & qc & qd \\ ra & rb & sa & sb \\ rc & rd & sc & sd \end{array} \right] = \left[ \begin{array}{cccc} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{array} \right].
\] (13.42)

In addition, we also need the \( n \)-step transition probability matrix of the combined channels \( T_c(n) = T_c^n \), which is given as
\[
T_c(n) = \left[ \begin{array}{cccc} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{21} & e_{22} & e_{23} & e_{24} \\ e_{31} & e_{32} & e_{33} & e_{34} \\ e_{41} & e_{42} & e_{43} & e_{44} \end{array} \right].
\] (13.44)
Following the same argument as in the first scenario, the parameter $\overline{A}$ in the second scenario is given as [33]

$$\overline{A}_2 = (\alpha_{12}e_{21} + \alpha_{13}e_{31} + \alpha_{14}e_{41}) \sum_{k=1}^{\infty} kp^{k-1}$$  \hfill (13.45)

$$= (\alpha_{12}e_{21} + \alpha_{13}e_{31} + \alpha_{14}e_{41})/(1 - \alpha_{11})^2$$  \hfill (13.46)

where the interpretation of this expression is similar to that for (13.39); however, error events can be due to errors in the forward or backward or both forward and backward channels. Substituting this result for $\overline{A}_2$ into (13.35), we can obtain the throughput efficiency of the GBN-ARQ protocol in this scenario $\eta_{\text{GBN-ARQ}}^e.2$.

In the special case with reliable feedback, we have $a = 1$ and $b = c = d = 0$, then (13.47) reduced to

$$\overline{A} = e_{31}/\alpha_{13}$$  \hfill (13.47)

which is equal to $\overline{A}_1$ given in (13.39).

### 13.6.2 SR-ARQ Protocol in Two-State Markov Channel

For simplicity, we consider only the scenario where the error process on the forward channel follows a two-state Markov channel with transition probability given in (13.36) and the backward channel has random error with probability $p_{e,b}$ in the analysis of the SR-ARQ protocol. The operations of the SR-ARQ protocol implies that its throughput efficiency can be calculated as

$$\eta_{\text{SR-ARQ}}^e = \frac{1}{E_{MC}}$$  \hfill (13.48)

where $E_{MC}$ denotes the average number of time slots required to successfully deliver one packet.

Suppose that the initial state of the forward channel is “0” (i.e., good state). In order to derive $E_{MC}$, we have to determine the probability $\Pr(X_k, Y_k)$ that the combined channel state $(X, Y) = (0, 0)$ is terminated after the $k$-th transmission of the same packet. We have

$$\Pr(X_1, Y_1) = p(n)p_{e,b}$$  \hfill (13.49)

$$\Pr(X_k, Y_k) = p(n)p_{e,b}^{k-1}p_{e,b} + q(n)r(n)p_{e,b} \sum_{i=0}^{k-2} p_{e,b}^{k-2-i}s(n)^i, \quad k \geq 2$$  \hfill (13.50)

where $p_{e,b} = 1 - p_{e,b}$, and we recall the transition probabilities of the $n$-step probability transition matrix $T_f(n)$ of the forward channel given in (13.37).

Using these results, $E_{MC}$ can be calculated as

$$E_{MC} = \sum_{j=1}^{\infty} j\Pr(X_j, Y_j) = \frac{1}{p_{e,b}} + \frac{q(n)}{r(n)}.$$  \hfill (13.51)
Table 13.1 Transmission Modes with Convolutionally Coded Modulation in HYPERLAN/2 Standard

<table>
<thead>
<tr>
<th>Mode</th>
<th>Modulation</th>
<th>Coding rate</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BPSK</td>
<td>1/2</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>QPSK</td>
<td>1/2</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>QPSK</td>
<td>3/4</td>
<td>1.50</td>
</tr>
<tr>
<td>4</td>
<td>16-QAM</td>
<td>9/16</td>
<td>2.25</td>
</tr>
<tr>
<td>5</td>
<td>16-QAM</td>
<td>3/4</td>
<td>3.00</td>
</tr>
<tr>
<td>6</td>
<td>64-QAM</td>
<td>3/4</td>
<td>4.50</td>
</tr>
</tbody>
</table>

From (13.48) and (13.51), the throughput efficiency of the SR-ARQ protocol is given as

$$
\eta_{\text{SR-ARQ}}^e = \frac{p_{e,b}r(n)}{r(n) + p_{e,b}q(n)}. \tag{13.52}
$$

13.7 Truncated ARQ Protocol with Link Adaptation under I.I.D. Channels

The throughput analysis in this chapter up to this point has been limited to wireless systems with a fixed transmission rate. Most of today’s wireless systems, however, implement the link adaptation to achieve higher system throughput (e.g., Table 13.1). Motivated by this fact, we study the throughput performance of the truncated ARQ protocol in wireless systems using link adaptation. For simplicity, we assume that the feedback delay is non-negligible, i.e., the transmitter knows the transmission outcomes immediately after transmission. Under this assumption, three ARQ protocols become the same.

Since different transmission modes have different transmission rates, we analyze the throughput with respect to the average spectral efficiency [5]. Suppose that there are $K$ transmission modes, each of which corresponds to one particular modulation scheme with specific spectral efficiency (transmitted bits per symbol period) and forward error correction (FEC) code with specific coding rate. We denote the spectral efficiency of mode $k$ as $R_k$, which is calculated as the multiplication of spectral efficiency of the modulation scheme and the coding rate of the FEC code. We simply refer to $R_k$ as the rate of mode $k$ for brevity. For illustration, we present the transmission modes of the HYPERLAN/2 standard in Table 13.2 [5].

Then the average packet error rate (PER) can be calculated using the PERs of all transmission modes as follows [5]:

$$
pe = \frac{\sum_{k=1}^{K} R_k \Pr(k) \overline{\text{PER}}_k}{\sum_{k=1}^{K} R_k \Pr(k)} \tag{13.53}
$$

where the average PER of mode $k$, denoted as $\overline{\text{PER}}_k$, is given in (13.3), and the probability that mode $k$ is activated $\Pr(k)$ is given in (13.4). If the SINR thresholds $X_k$ are chosen so that the average PER of any mode is equal to a target value $P_0$ as in Algorithm 13.1, then we have $\overline{\text{PER}}_k = pe = P_0, \forall k$. Moreover, the average number of transmissions per packet for truncated and infinitely persistent ARQ protocols can be calculated as in (13.18) and (13.19), respectively.
If one does not employ the ARQ protocol, then the average rate (spectral efficiency) achieved by AMC at the physical layer is

\[
\bar{R}_{\text{PHY}} = \sum_{k=1}^{K} R_k \Pr(k).
\] (13.54)

And this physical layer rate is achieved with the achieved average PER \(p_e\) given in (13.3).

For systems with the ARQ protocol, the average rates (spectral efficiency) of the truncated and infinitely persistent ARQ protocols can be calculated, respectively, as [5]

\[
\bar{R}_{\text{ARQ}}^\infty = \sum_{k=1}^{K} R_k \Pr(k) = \frac{\bar{R}_{\text{PHY}}}{E_{\infty}} = \bar{R}_{\text{PHY}} (1 - p_e). \tag{13.56}
\]

Note that the average rate of the truncated ARQ is higher than that of the infinitely persistent ARQ protocol since \(E_{\text{tr}} < E_{\infty}\). However, this throughput gain is achieved at the cost of having average PER of \(p_e^{N_{\text{max}}}\) while the infinitely persistent ARQ protocol is fully reliable. It can be verified that \(\bar{R}_{\text{ARQ}}^\infty < \bar{R}_{\text{ARQ}}^\text{tr} < \bar{R}_{\text{PHY}}\) and the corresponding reliability levels of no-transmission, truncated ARQ, and infinitely persistent ARQ protocols increase in this order. Therefore, the truncated ARQ protocol allows us to achieve the desirable tradeoff between throughput efficiency and reliability. This can be achieved by choosing a suitable value of \(N_{\text{max}}\) (i.e., the number of transmissions per packet).

### 13.8 Delay Analysis of GBN-ARQ Protocol with Link Adaptation under FSMC

We have studied the throughput analysis of ARQ protocols over different wireless channels. This section discusses the delay analysis, which is indeed more challenging. The throughput analysis has been presented for the saturated-buffer scenario where the transmitter always has packets to transmit over the wireless channel. In practice, traffic arriving to the link layer buffer from the higher layer can be bursty, which implies that the buffer can be empty or low-loaded from time to time.

In general, the end-to-end delay experienced by an RLC packet comprises queuing delay at the transmitter’s buffer and transmission delay over the wireless channel. In fact, there is also re-sequencing delay at the receiver’s buffer for the SR-ARQ protocol since packets accepted at the receiver can be out of order. Analytical models for delay analysis of different ARQ protocols over wireless channels can be quite different. In addition, one can derive either average delay or delay distribution even though delay distribution provides more detailed information for system design and QoS provisioning.

Existing research on delay analysis of ARQ protocol has mainly conducted for wireless systems with a fixed transmission rate and either i.i.d. or Markovian error models [34–38]. We will derive the delay distribution for the GBN-ARQ protocol for the multi-rate wireless system employing the link adaptation technique by using the
matrix geometric method (MGM) [25, 39]. Delay analysis of the SR-ARQ protocol for the same system is available in [25].

### 13.8.1 System and Protocol Description

We consider a transmitter node using adaptive modulation and coding at the physical layer and GBN-ARQ protocol for RLC error recovery to communicate with a receiver node over a wireless channel. Transmissions occur in fixed-size time slots where the number of packets transmitted during each time slot depends on the chosen transmission mode. The receiver decodes the received packets and sends back an ACK/NAK message to the transmitter depending on the decoding outcomes. In case of transmission failure of one or more packets transmitted during a time slot, an error recovery based on the GBN-ARQ protocol is initiated.

For the GBN-ARQ protocol, the transmitter continuously transmits packets from the buffer in sequence until it detects a transmission error through the NAK message. In case of transmission failure(s), the GBN-ARQ protocol retransmits all the packets starting from the first erroneous packet. We assume that the feedback message (i.e., the ACK/NAK) arrives at the transmitter node \(n\) slots after the beginning of the corresponding transmission slot. In addition, an error-free feedback channel is assumed.

In addition to the ACK/NAK information, the feedback channel carries the channel state information (CSI) or the selected transmission mode to be used for dynamic link adaptation. We assume that CSI is available at the transmitter without delay, which is reasonable in slow fading channels. The maximum number of retransmissions allowed for a packet is assumed to be unbounded (i.e., the infinitely persistent GBN-ARQ protocol).

**Example 69** The operations of the GBN-ARQ protocol are illustrated in Figure 13.1 for \(n = 3\), where the transmission batch size is determined by the rate defined in terms of packets/slot. In this figure, packet 2 in time slot 1 and packets 8 and 9 in time slot 3 are assumed to be in error. In the GBN-ARQ protocol, the NAK message for packet 2 arrives...
at the transmitter side at the end of time slot 3, and retransmission of all packets starting from packet 2 begins at time slot 4. Note that, packets 3, 4, . . . , 7 are retransmitted even though they were correctly received before.

The channel is modeled as an FSMC with $K + 1$ states (0, 1, . . . , $K$). When the channel is in state $k$ (1, 2, . . . , $K$), $h_k$ packets are transmitted in one time slot. In fact, each channel state corresponds to one transmission mode of the AMC technique. We further assume that the transmitter does not transmit in channel state 0 to avoid high probability of transmission error.

Each transmission mode corresponds to a unique modulation and coding scheme, and the number of packets transmitted in mode $k$ (equal to $h_k$) is proportional to the spectral efficiency of mode $k$. For example, if the spectral efficiencies of five transmission modes in a particular system are 0.5, 1, 1.5, 2.5 and 3.5 (bits/s/Hz) and one packet can be transmitted in mode 1 (with spectral efficiency of 0.5), the number of packets transmitted in one time slot using the other modes is 2, 3, 5, 7, respectively. The maximum number of packets that can be transmitted in one time slot (in channel state $K$) is denoted as $L$.

The radio link level queueing for the GBN-ARQ protocol is modeled in discrete time with one time interval equal to one time slot, and the system states are observed at the beginning of each time slot. The buffer size at the transmitter is assumed to be infinite. Packet arrivals to the transmitter’s buffer follow a Bernoulli process with arrival probability $\lambda$. We assume that a packet arriving during time slot $t−1$ cannot be transmitted until time slot $t$ at the earliest.

### 13.8.2 Queuing Model

Since the result of the decoding process for each packet reaches the transmitter $n$ slots only after the beginning of the transmission slot, if a transmitted packet is in error in time slot $t$, all transmissions from time slot $t + 1$ to $t + n − 1$ will be discarded. Therefore, we need to keep track not only of the channel state, which determines how many packets can be transmitted in one time slot, but also the useful time slot, which is defined as the slot where the transmitted packets, if successfully decoded, will be accepted by the receiver.

Let $x(t) \geq 0$ represent the number of packets in the queue, including packets that will be retransmitted but whose NAKs have not yet been received, $0 \leq u(t) \leq n − 1$ track the useful time slot, and $0 \leq w(t) \leq K$ represent the channel state. We assign the value for $u(t)$ as follows. If a transmission failure occurs in a useful time slot, $u(t)$ will be equal to $n − 1$ in the next time slot. Then it will be decreased during the subsequent slots until $u(t) = 0$, where a useful transmission period starts (an example is shown in Figure 13.2). As is evident, the number of useless slots following transmission error(s) in a useful slot is $n − 1$. It can be shown that the random process $Z(t) = \{x(t), u(t), w(t)\}$ forms a discrete-time Markov chain (MC). For brevity, we will omit time index $t$ in the related variables if it does not cause confusion.

In order to calculate the steady state probability for the underlying MC, it is important to put its transition probability matrix in a nice form where its specific transition
Example: For ease of exposition, we consider a very simple case where there are two channel states (states 0, 1) and feedback delay is two slots (i.e., \( n = 2 \)). The matrix block \( D_{i,l} \) can be expanded as in (13.57) where element \( D_{i,l}(1,0)(1,0) \), for example, represents the probability of the state transition \( (i, 1, 1) \rightarrow (i + 1 - l, 0, 0) \).

\[
D_{i,l} = \begin{bmatrix}
D_{i,l}(0,0) & D_{i,l}(0,1) \\
D_{i,l}(1,0) & D_{i,l}(1,1)
\end{bmatrix}
\]

The resulting transition matrix for the MC \( Z(t) \) is written in (13.57) for \( L = 3 \). Recall that \( L \) is the maximum number of packets that can be transmitted in one time slot (i.e., equal to \( h_k \)). In this probability transition matrix, \( D_{i,l} \) contains the probabilities of system transitions where \( x = i \) before the transitions. All transitions captured by \( D_{i,l} \) for different \( l \) will be called transitions in level \( i \) of the transition matrix in the sequel.

Figure 13.2 Modeling of GBN-ARQ for \( n = 4 \).

structure can be exploited. Now, let \( (i, j, k) \) be the generic system state (i.e., \( x = i \), \( u = j \), and \( w = k \)) and \( (i, j, k) \rightarrow (i', j', k') \) denote the system transition from state \( (i, j, k) \) to state \( (i', j', k') \). For fixed \( i \), the probabilities corresponding to system state transitions \( (i, *, *) \rightarrow (i + 1 - l, *, *) \) can be written in a matrix block \( D_{i,l} \). We further put the probabilities of state transitions \( (i, j, *) \rightarrow (i + 1 - l, j', *) \) into a sub-matrix \( D_{i,l}(j, j') \) of \( D_{i,l} \). Also, the probability of transition \( (i, j, k) \rightarrow (i + 1 - l, j', k') \) is denoted by \( D_{i,l}(j, j')(k, k') \), which is an element of \( D_{i,l}(j, j') \). In fact, the probabilities corresponding to transitions from state \( (i, j, k) \) to any other state will be in the \((i(K+1)n+j(K+1)+k)\)-th row of the probability transition matrix, and they are elements of \( D_{i,l} \) for some value of \( l \) depending on the destination state. The following example clarifies further how the system state transition probabilities are ordered in the matrix form.
that, in the generic system state \((i, j, k)\), \(j\) can have \(n\) possibilities and \(k\) can have \(K + 1\) possibilities (i.e., \(K + 1\) channel states). Thus, the order of \(\mathbf{D}_{i,l}\) is \(n(K + 1) \times n(K + 1)\).

The derivations of matrix blocks \(\mathbf{D}_{i,l}\) and \(\mathbf{D}_l\) are detailed later in Section 13.8.3. As can be seen from these derivations, for \(i \geq L\), \(\mathbf{D}_{i,l}\) is independent of the level index \(i\); therefore, for brevity we denote \(\mathbf{D}_{i,l}\) by \(\mathbf{D}_l\) in (13.57).

\[
\mathbf{P} = \begin{bmatrix}
\mathbf{D}_{0,0} & \mathbf{D}_{0,1} & \cdots & \mathbf{D}_{0,L-1} \\
\mathbf{D}_{1,0} & \mathbf{D}_{1,1} & \cdots & \mathbf{D}_{1,L-1} \\
\mathbf{D}_{2,0} & \mathbf{D}_{2,1} & \cdots & \mathbf{D}_{2,L-1} \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{D}_{L-1,0} & \mathbf{D}_{L-1,1} & \cdots & \mathbf{D}_{L-1,L-1}
\end{bmatrix}.
\] (13.57)

Note that, in (13.57), there is at most one arriving packet and at most \(L = 3\) packets successfully transmitted in one time slot. Therefore, for level \(i \geq 3\), the transitions can go up at most one level (represented by \(\mathbf{D}_0\)) and go down at most three levels (represented by \(\mathbf{D}_4\)). The probability transition matrix in (13.57) describes a GI/M/1 Markov chain, where the solution can be found by the well-established method proposed by Neuts [39]. In fact, the steady-state probability \(\mathbf{x} = [x_0 \ x_1 \ x_2 \ \cdots]\) satisfies

\[
\mathbf{x}\mathbf{P} = \mathbf{x}
\]

\[
\sum_{i=0}^{\infty} x_ie = 1
\]

where \(e\) is a column vector of all ones with the same dimension as \(x_i\) which is \(n(K + 1)\).

We can find \(x_0, x_1, \ldots, x_L\) using the boundary and the normalization conditions. Other values of \(x_i\) \((i > L)\) can be calculated from \(x_L\) by using a non-negative matrix \(\mathbf{R}\) as follows: \(x_i = x_L \mathbf{R}^{i-L}\) [39]. Here the order of matrix \(\mathbf{R}\) is \(n(K + 1) \times n(K + 1)\).

### 13.8.3 Derivations of Matrix Blocks in (13.57)

We derive the matrix blocks for the transition matrix (13.57) in this section. The number of packets transmitted in time slot \(t\) is the minimum of the number of packets available in the queue and the transmission capability (i.e., equal to \(\min\{x(t), h_{u(t)}\}\)). Let \(a(t)\) be the number of arriving packets during slot \(t\), and \(d(t)\) be the number of packets that will not be retransmitted due to transmissions in slot \(t\) (in fact, \(d(t) \leq \min\{x(t), h_{u(t)}\}\)), then we have \(x(t + 1) = x(t) + a(t) - d(t)\).

To derive the matrix blocks \(\mathbf{D}_{i,l}\) and \(\mathbf{D}_l\) in (13.57), we consider the following cases that may occur in each time slot. First, the slot is not useful (i.e., \(u(t) \neq 0\)), and therefore, no packet can depart. We need to keep track of the channel state evolution only for this case. Second, the slot is useful (i.e., \(u(t) = 0\)), and all transmitted packets are received correctly. In this case, the number of packets in the queue changes according to the number of successfully transmitted packets and the number of arriving packets in that slot. Also, the next time slot will be a useful one (i.e., \(u(t + 1) = 0\)). Third, the slot is useful (i.e., \(u(t) = 0\)), and there exists at least one packet among those transmitted in
error. The number of packets in the queue at the end of the slot depends on the error pattern and the number of arriving packets in that slot. However, the next time slot will not be useful (in fact, \( u(t + 1) = n - 1 \)).

Now let us define the following matrices:

- \( T_k \) (\( k = 0, 1, \ldots, K \)) are constructed by keeping only the \((k + 1)\)-th row of the channel transition probability matrix \( T \) and setting all other rows to 0. These matrices capture the case the channel is in state \( k \) at the beginning of a particular time slot.
- \( \Psi_{ij}^{(0)} \) are matrices of order \((K + 1) \times (K + 1)\) in which element \( \Psi_{ij}^{(0)}(k, k') \) is the probability that all \( i \) transmitted packets are received correctly given that there were \( j \) packets in the queue before transmission (i.e., \( i = \min\{x(t), h_{u(t)}\} = \min\{j, h_{u(t)}\} \)), the channel changes from state \( k \) to state \( k' \) in the transmission slot.
- \( \Psi_{ij}^{(1)} \) are matrices of order \((K + 1) \times (K + 1)\) whose element \( \Psi_{ij}^{(1)}(k, k') \) is the probability that \( i \) transmitted packets are received correctly given that there were \( j \) packets in the queue before transmission, there is at least one erroneous packet (i.e., \( i < \min\{x(t), h_{u(t)}\} = \min\{j, h_{u(t)}\} \)), and the channel changes from state \( k \) to state \( k' \) in the transmission slot.
- \( C_{ij}^{(k)} \) (\( k = 1, 2, 3 \)) are matrices of order \( n(K + 1) \times n(K + 1) \) representing the aforementioned three cases, respectively, whose element \( C_{ij}^{(k)}(l, l')(h, h') \) (\( 0 \leq l, l' \leq n - 1, 0 \leq h, h' \leq K \)) represents the probability of system transition \((j, l, h) \rightarrow (j - i, l', h')\) (i.e., \( i \) packets are successfully transmitted given there were \( j \) in the queue before transmissions).

From foregoing definitions, \( C_{ij}^{(k)} \) can be written as follows:

\[
C_{ij}^{(1)} = \begin{cases} 0 & \text{if } i \neq 0 \\
0 & \text{else}
\end{cases}
\]

\[
C_{0,j}^{(1)} = \begin{bmatrix}
0 & 0 & \cdots & 0 & 0 \\
T & 0 & \cdots & 0 & 0 \\
0 & T & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & T & 0
\end{bmatrix}
\]

\[
C_{i,j}^{(2)} = \begin{bmatrix}
\Psi_{ij}^{(0)} & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & 0 & 0
\end{bmatrix}
\]

\[
C_{i,j}^{(3)} = \begin{bmatrix}
0 & 0 & \cdots & 0 & \Psi_{ij}^{(1)} \\
0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & 0 & 0
\end{bmatrix}
\]
Equation (13.61) simply captures the channel state transitions where \( u(t) \neq 0 \). Note that \( u(t) \) decreases by one in each time slot, which explains the structure of \( C_{0,i}^{(1)} \) (i.e., \( T \) is in positions \((u, u - 1)\)). For \( u(t) \neq 0 \), no packet can depart; therefore, we have \( C_{i,i}^{(1)} = 0 \) for \( i \neq 0 \). In (13.62), \( C_{i,j}^{(2)} \) contains the probabilities of transition between two useful slots (i.e., \( u(t) = u(t + 1) = 0 \)), and \( C_{i,j}^{(3)} \) contains the probabilities of transition from a useful time slot to a useless one (i.e., \( u(t) = 0 \) and \( u(t + 1) = n - 1 \)), where at least one transmission error must have occurred. This explains the position of \( \Psi_{i,j}^{(0)} \) and \( \Psi_{i,j}^{(1)} \) in the matrices \( C_{i,j}^{(2)} \) and \( C_{i,j}^{(3)} \), respectively.

Before we derive the matrix blocks in (13.57), let \( C_{i,j} = \sum_{k=1}^{3} C_{i,j}^{(k)} \), which contains the probabilities that \( i \) packets are successfully transmitted given that there were \( j \) packets in the queue before transmission without distinguishing the aforementioned three cases. As we discussed above, if \( d(t) \) is the number of packets that will not be retransmitted due to transmissions in slot \( t \), we have \( x(t + 1) = x(t) + a(t) - d(t) \). It can be easily seen that \( D_{i,l} \) and \( D_{l} \) contains the probabilities of system state transition where \( x(t + 1) = x(t) + 1 - l \). Thus, we have \( a(t) - d(t) = 1 - l \). As a result, \( d(t) = l - 1 \) if \( a(t) = 0 \) (i.e., no arrival) and \( d(t) = l \) if \( a(t) = 1 \) (i.e., one arrival). Therefore, \( D_{i,l} \) can be calculated as follows:

\[
D_{i,l} = (1 - \lambda)C_{l-1,i} + \lambda C_{i,i}. 
\]

Similarly, \( D_{l} \) can be calculated as

\[
D_{l} = (1 - \lambda)C_{l-1,L} + \lambda C_{l,l}. 
\]

It can be easily observed that \( C_{i,j} = C_{l,l} \) for \( j \geq L \).

The remaining task is to determine \( \Psi_{i,j}^{(0)} \) and \( \Psi_{i,j}^{(1)} \), which is pursued now. Let \( \theta_k = \text{PER}_k \) be the probability of transmission error when the channel is in state \( k \). Assuming that the transmission outcomes of different packets are independent and let us define

\[
p_{i}^{(k)} = (1 - \theta_k)^i. 
\]

Then, for \( i, j > 0 \), \( \Psi_{i,j}^{(0)} \) can be calculated as

\[
\Psi_{i,j}^{(0)} = \begin{cases} 
  p_{i}^{(k)} T_k, & i < j, \ i = h_k \\
  0, & i < j, \ \text{if } \exists \ k \text{ s.t. } i = h_k \\
  \sum_{h=K}^{K} p_{i}^{(h)} T_h, & i = j, \ k = \min\{w(t) : h_k \geq j\}.
\end{cases} 
\]

For \( i = 0 \), \( \Psi_{0,j}^{(0)} \) can be calculated as

\[
\Psi_{0,j}^{(0)} = T, \quad \Psi_{0,j}^{(0)} = T_0, \quad j > 0.
\]

Also, \( \Psi_{i,j}^{(1)} \) can be calculated as

\[
\Psi_{i,j}^{(1)} = \sum_{m=k}^{K} p_{i}^{(m)} \theta_m T_m 
\]

where \( k = \min\{w(t) : h_k \geq i\} \).
Equations (13.67)–(13.69) can be interpreted as follows. In (13.67), we must have \( i = \min\{j, h_{u(t)}\} \); therefore, \( h_{u(t)} = i \) if \( j > i \) or \( h_{u(t)} \geq i \) if \( i = j \). For \( \Psi^{(0)}_{0,0} \), there is no transmission since there is no packet in the queue at the beginning of the slot. Therefore, the channel can be in any state without introducing any transmission error. For \( \Psi^{(0)}_{0,j} \), there are \( j > 0 \) packets in the queue in this case, and no transmission error occurs only if the channel is in state 0 (since no transmission is allowed in this channel state). The interpretation for (13.69) is similar, but at least one packet among those transmitted must be in error (i.e., \( i < \min\{j, h_{u(t)}\} \)). Note that in this case the first \( i \) packets are received correctly and the \((i + 1)\)-th packet must be in error.

### 13.8.4 Delay Analysis

We now derive the delay distribution of a packet arriving at the queue for the GBN-ARQ protocol. The delay is the time for all packets ahead of the target packet (if any) and itself successfully leaving the queue. In the following calculation, the delay is considered at the transmitter side. Let the arrival slot be numbered as slot zero, and it is not included in the delay calculation.

Now let \( \Phi_{(p,d)} \) be matrices with order \( n(K + 1) \times n(K + 1) \) whose element \( \Phi_{(p,d)}(j, j') \) \((0 \leq j, j' \leq n - 1, 0 \leq k, k' \leq K)\) is the probability of state transition \((p, j, k) \rightarrow (0, j', k')\) in \( d \) time slots. In short, \( \Phi_{(p,d)} \) contains system transition probabilities such that \( p \) packets are successfully transmitted in \( d \) slots. From the definition of \( \Phi_{(p,d)}\), \( \Phi_{(p,d)}(j, j') \) contains the channel state transition probabilities such that \( u \) (in the system state \((x, u, w)\)) evolves from \( j \) to \( j' \). Thus, the order of \( \Phi_{(p,d)}(j, j') \) is \((K + 1) \times (K + 1)\).

We also define \( C_{h,p} \) to be matrices of order \( n(K + 1) \times n(K + 1) \) with the same structure as \( \Phi_{(p,d)} \) whose elements are the system transition probabilities such that \( h \) packets are successfully transmitted in one particular time slot given that there are \( p \) packets in the queue at the beginning of the time slot. We have the following recursive relation:

\[
\Phi_{(p,d)} = \sum_{h=0}^{L} C_{h,p} \Phi_{(p-h,d-1)}
\]

(13.70)

where \( \Phi_{(0,0)} = I_{n(K+1)} \). Equation (13.70) can be interpreted as follows. If there are \( p \) packets that must be delivered in \( d \) time slots (captured by \( \Phi_{(p,d)} \)) and \( h \) packets are successfully transmitted in the first time slot (captured by \( C_{h,p} \)), there are remaining \( p - h \) packets to be delivered in \( d - 1 \) slots (captured by \( \Phi_{(p-h,d-1)} \)). Here \( \Phi_{(0,0)} \) simply captures the end point where the target packet leaves the queue.

To calculate the delay distribution, we need to obtain the steady-state vector seen by an arriving packet to the queue. Note that we have assumed a Bernoulli arrival process so that the ASTA (arrivals see time averages) property holds here. Also, packets arriving to the queue after the tagged packet do not affect the delay experienced by the tagged packet, so they are ignored in the following derivation. Let \( y_i \) be a vector of dimension \( n(K + 1) \) that represents the system state probabilities where an arriving packet sees \( i \)
Table 13.2 Transmission Modes with Uncoded $M_n$-QAM Modulation [21]

<table>
<thead>
<tr>
<th>Mode</th>
<th>Mode 2</th>
<th>Mode 3</th>
<th>Mode 4</th>
<th>Mode 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modulation</td>
<td>BPSK</td>
<td>QPSK</td>
<td>8-QAM</td>
<td>16-QAM</td>
</tr>
<tr>
<td>Rate</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$a_k$</td>
<td>67.7328</td>
<td>73.8279</td>
<td>58.7332</td>
<td>55.9137</td>
</tr>
<tr>
<td>$g_k$</td>
<td>0.9819</td>
<td>0.4945</td>
<td>0.1641</td>
<td>0.0989</td>
</tr>
<tr>
<td>$X_{pk}$ (dB)</td>
<td>6.3281</td>
<td>9.3945</td>
<td>13.9470</td>
<td>16.0938</td>
</tr>
</tbody>
</table>

head-of-line (HOL) packets at the end of its arrival time slot. We have

$$y_i = \sum_{h=0}^{L} x_{i+h} C_{h,i+h}.$$  \hspace{1cm} (13.71)

Equation (13.71) can be interpreted as follows. If there are $i + h$ packets in the queue at the beginning of the arrival time slot (captured by $x_{i+h}$) and $h$ packets successfully leave the queue in this time slot (captured by $C_{h,i+h}$), the arriving packet will see exactly $i$ HOL packets (captured in $y_i$) at the end of this time slot. The probability that the delay is $D$ slots (not including the arrival slot) can therefore be written as follows:

$$P_d(D) = \sum_{h=0}^{DL-1} y_h \Phi_{(h+1,D)} e_{n(K+1)}$$  \hspace{1cm} (13.72)

where $e_{n(K+1)}$ is a column vector of all ones with dimension $n(K+1)$. The sum in (13.72) is limited to $DL - 1$ since at most $L$ packets can be successfully transmitted in one time slot.

13.8.5 Numerical Example

To present some illustrative results, we consider five transmission modes (i.e., $K = 5$). We use the PER fitting values of $a_k$, $g_k$, and $X_{pk}$ summarized in Table 13.2, which was presented in [21], to obtain the SNR thresholds of the FSMC model such that $\text{PER}_k = P_0$ for all the transmission modes, where $P_0$ is a certain target packet error rate. A wireless system using adaptive modulation is considered where $h_k = k$ (i.e., the transmitter transmits $k$ packets/slot in channel state $k$). We assume that the time slot interval $T_s = 1$ ms. The complementary delay distribution is given as $\text{Pr}(\text{delay} > d) = 1 - \sum_{k=1}^{d} P_d(k)$.

An important issue in designing dynamic link adaptation mechanisms is the selection of the mode-switching SNR thresholds for the different transmission modes (i.e., the SNR thresholds $X_k$, $k = 1, 2, \ldots, K$). Specifically, we have obtained the SNR thresholds such that the average packet error rate for all modes is equal to some target packet error rate $P_0$ (i.e., $\text{PER}_d = P_0$). Different values of $P_0$ result in different sets of SNR thresholds for the AMC at the physical layer.

Variations in the complementary cumulative delay distributions for GBN-ARQ with different values of $P_0$ are illustrated in Figure 13.3. The lowest delay is obtained for $P_0 = 0.1$ in this case. Basically, for higher values of $P_0$, the average transmission rate...
increases, but the wireless link becomes less reliable. In other words, we are more likely to use high transmission modes by choosing large \( P_0 \). However, the high probability of transmission errors may require many retransmissions, which may increase the link level delay. Thus, there exists a value of \( P_0 \) for which the link level delay is minimized.

### 13.9 Hybrid ARQ Protocol with Transmission Size Adaptation

We now turn our focus to the analysis and design issues for the HARQ protocol with incremental redundancy. Specifically, we will derive the throughput of the HARQ protocol with incremental redundancy under the fairly general design where the numbers of coded bits transmitted in different transmissions are different. Then we discuss simple approaches to optimize these numbers of coded bits delivered over different transmissions to optimize the achievable throughput [8]. More sophisticated design based on information theoretic metrics for rate allocation and adaptation of the HARQ protocol with incremental redundancy can be found in [41].

Assume that \( L_b \) information bits are encoded into \( L_p \) coded bits with code rate of \( R_C = L_b/L_p \). We assume that a punctured convolutional code is employed to create the coded bits [8], [11]. The HARQ with incremental redundancy and maximum number of transmissions \( N_{\text{max}} \) operates as follows. In each transmission, the transmitter transmits a subset of the coded bits, the receiver attempts to decode the codeword, and the receiver requests the next transmission if the decoding fails. In general, the subsets of coded bits chosen in different transmissions can be optimized together with convolutional code
13.9 Hybrid ARQ Protocol with Transmission Size Adaptation

Let $B_j$ denote the number of coded bits in the $j$-th transmission. In a wireless system employing incremental redundancy, the decoding is performed by jointly combining the soft inputs received from all previous and current transmissions. Once the set of coded bits, which has not been transmitted previously, is exhausted in a particular transmission, one can choose the subset of coded bits randomly from the codeword, and the corresponding soft inputs at the receiver side can be combined with the previously received soft inputs to achieve a combining gain. If the receiver still fails to decode after $N_{\text{max}}$ transmissions, then all accumulated soft inputs for the current packet are discarded, and the system proceeds to transmit the next packet. This description implies that the code rate in the $j$-th transmission is

$$R_j = \frac{L_b}{\sum_{i=1}^{j} B_i}.$$  \hspace{1cm} (13.73)

It can be observed that $R_i > R_j$ for $i < j \leq N_{\text{max}}$. Consequently, the decoding succeeds with higher probability over subsequent transmissions. In general, one can optimize the transmission sizes $B_j$ in different transmissions $j$ to achieve the best throughput performance. This transmission size adaptation will be briefly discussed by the end of this section. To facilitate the throughput analysis, let $p_j$ denote the decoding failure for the $j$-th transmission. Moreover, suppose that the round trip delay is $T_{\text{RTT}}$ and the symbol interval lasts $t_s$ seconds. Then, the throughput of HARQ with incremental redundancy can be expressed as [8]

$$\eta = \frac{L_b}{T_{\text{trans}}} = \frac{L_b (1 - p_{N_{\text{max}}})}{t_s \left( B_1 + \sum_{j=2}^{N_{\text{max}}} B_j p_{j-1} \right)}$$  \hspace{1cm} (13.74)

where $T_{\text{trans}}$ represents the average transmission time until the successful decoding of the current packet. Here $T_{\text{trans}}$ can be calculated by using the fact that the $j$-th transmission is required only if the decoding during the $(j-1)$-th transmission fails (with probability $p_{j-1}$). Moreover, the last decoding succeeds with probability $1 - p_{N_{\text{max}}}$.

The average packet delay, which is counted from the instant a particular packet is transmitted in the first attempt until it is either decoded successfully or discarded, can be calculated as [8]

$$D_{\text{HARQ}} = T_{\text{trans}} + T_{\text{wait}} = t_s \left( \frac{B_1 + \sum_{j=2}^{N_{\text{max}}} B_j p_{j-1}}{1 - p_{N_{\text{max}}}} + \frac{T_{\text{RTT}} \sum_{j=1}^{N_{\text{max}}} p_j}{1 - p_{N_{\text{max}}}} \right)$$  \hspace{1cm} (13.75)

where $T_{\text{wait}}$ denotes the average cumulative waiting time until successful decoding of a packet and $T_{\text{RTT}}$ is the round trip time.

In the following, we discuss a simple technique to set the transmission sizes $B_j$ in different transmissions to maximize the HARQ throughput [8]. Toward this end, let $z_j$ denote the cumulative soft inputs over all transmissions up to transmission $j$. Moreover,
a posteriori log-likelihood ratios (LLRs) of the $i$-th information bit in the $j$-th transmission can be written as

$$LLR_j(i) = \log \frac{\Pr[b_i = 0|z_j]}{\Pr[b_i = 1|z_j]}$$

(13.76)

where $b_i$ denotes information bit $i$. Following the $j$-th transmission, the codeword reliability $\mu_j$ can be expressed as

$$\mu_j = E[\log \frac{\Pr[b_i = 0|z_j]}{\Pr[b_i = 1|z_j]}]$$

(13.77)

where the expectation $E[.]$ is taken over noise realization, cumulative set of transmitted bits, and channel fading. Moreover, one can estimate the codeword reliability $\hat{\mu}_j$ following the $j$-th transmission as follows:

$$\hat{\mu}_j = \frac{1}{L_b} \sum_{i=1}^{L_b} |LLR_j(i)|.$$ 

(13.78)

In order to optimize the transmission sizes for the HARQ with incremental redundancy, we assume that the following two mappings can be estimated, namely, the reliability to block error rate (BLER) mapping $BLER = F(\mu)$ and the code rate to reliability mapping $\mu = G(R)$. While exact expressions for these mappings are not easy to derive, bounds for these mappings are given in [8]. Given the knowledge of these mappings, one can optimize the reliability levels to achieve the maximum throughput as follows:

$$[\bar{\mu}_2, \bar{\mu}_3, \ldots, \bar{\mu}_{N_{\text{max}}}] = \arg \max_{\bar{\mu}_2, \bar{\mu}_3, \ldots, \bar{\mu}_{N_{\text{max}}}} \frac{1 - p_{N_{\text{max}}}}{B_1 + \sum_{j=2}^{N_{\text{max}}} B_j p_{j-1}}.$$ 

(13.79)

While closed-form optimal solution for this problem may not be tractable to obtain, one can search the optimal solution satisfying $\bar{\mu}_2 < \bar{\mu}_3 < \cdots < \bar{\mu}_{N_{\text{max}}}$. Note that the transmission size of the first transmission is assumed to be known, which can be determined according to the modulation and coding scheme employed in the first transmission. Based on this optimal solution, the offline non-adaptive transmission sizes can be determined as

$$B_{j+1} = \frac{L_b}{G^{-1}(\bar{\mu}_{j+1})} - \sum_{k=1}^{j} B_k.$$ 

(13.80)

Note that $G^{-1}(\bar{\mu}_{j+1})$ provides the required code rate to achieve the reliability level $\bar{\mu}_{j+1}$. In addition, the adaptive transmission size for the $(j+1)$-th transmission can be determined as

$$B_{j+1} = \frac{L_b}{G^{-1}(\bar{\mu}_{j+1})} - \frac{L_b}{G^{-1}(\hat{\mu}_j)}.$$ 

(13.81)

Here $G^{-1}(\hat{\mu}_j)$ is the estimated code rate achieved with the $j$-th transmission with the estimated LLR value of $\hat{\mu}_j$. In practice, this adaptation can be performed in real time based on the estimated reliability level $\hat{\mu}_j$. 

---

*Available at [link](https://www.cambridge.org/core/terms).*
Table 13.3 Transmission Probability Matrix

<table>
<thead>
<tr>
<th></th>
<th>(0, i)</th>
<th>(1, i)</th>
<th>(0, i + 1)</th>
<th>(1, i + 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, i + 1)</td>
<td>(1 - λ)p_{00}</td>
<td>(1 - λ)p_{01}</td>
<td>λp_{00}</td>
<td>λp_{01}</td>
</tr>
<tr>
<td>(1, i)</td>
<td>(1 - λ)p_{10}</td>
<td>(1 - λ)p_{11}</td>
<td>λp_{10}</td>
<td>λp_{11}</td>
</tr>
</tbody>
</table>

13.10 Exercises

The following exercises are based on the results in several references [35], [40–42].

Exercise 13.1: Consider the ARQ protocol operating in a time slotted manner over the two-state Markov channel comprising the bad and good states, which are denoted as states 0 and 1, respectively. The transition matrix of this Markov channel is given as

\[
T = \begin{bmatrix}
p_{00} & p_{01} \\
p_{10} & p_{11}
\end{bmatrix}
\] (13.82)

Assume that one packet can be successfully transmitted in a good state and no packet can be delivered in a bad state in each time slot. Moreover, we assume that the transmitter knows the transmission outcome right after its transmission, and the feedback channel is instantaneous and error free (i.e., there is no feedback delay and error). Packets arrive at the transmitter’s buffer according to the Bernoulli process where one packet arrives by the end of each time slot with probability λ (i.e., a packet arriving in time slot \(n\) can be transmitted only in time slot \(n + 1\)). To analyze this ARQ-based system, we define the Markov chain \(X(n) = (l(n), i(n))\) with state space \{(l, i), l \in \{0, 1\}, i \geq 0\} where \(l(n)\) denotes the wireless channel state and \(i(n)\) represents the number of packets in the transmitter’s buffer in time slot \(n\).

i. Prove that the transition probabilities of the Markov chain \(X(n)\) are shown in Table 13.3. For example, we have \(\Pr((0, i + 1) \rightarrow (0, i)) = (1 - \lambda)p_{00}\).

ii. Let \(\pi(l, i)\) denote the stationary probability of state \((l, i)\), then show the following:

\[
\pi(1, i) = \beta^i \pi(1, 0), \quad i \geq 1
\]

\[
\pi(0, i) = \frac{\lambda}{1 - \lambda} \beta^{i-1} \pi(1, 0), \quad i \geq 1
\]

\[
\pi(0, 0) = \frac{(1 - \lambda)p_{10} + \lambda p_{00}}{p_{01}} \pi(1, 0)
\]

where

\[
\beta = \frac{(1 - \lambda)p_{11} + \lambda p_{01}}{(1 - \lambda)p_{10} + \lambda p_{00}} \frac{\lambda}{1 - \lambda}
\]

\[
\pi(1, 0) = \frac{[(1 - \lambda)p_{10} - \lambda p_{01}]p_{01}}{(1 - \lambda)(p_{10} + p_{01})[(1 - \lambda)p_{10} + \lambda p_{00}]}
\]

Hint: Study the set of balance equations of the underlying Markov chain.
iii. Show the following relations:

\[
\pi(0, 0) = 1 - \frac{p_e}{1 - \lambda}
\]
\[
\pi(0, i) = \frac{\lambda p_e (1 - \beta) \beta^{i-1}}{1 - \lambda}, \quad i \geq 1
\]
\[
\pi(1, i) = p_e (1 - \beta) \beta^i, \quad i \geq 0
\]

where \( p_e \) represents the transmission error probability, which is equal to

\[
p_e = \frac{p_{01}}{p_{10} + p_{01}}.
\]

**Exercise 13.2:** For the ARQ-based wireless system considered in Exercise 13.1, let \( \Phi_{lj}(k, n) \) denote the probability that there are \( k \) “good” slots in time slots \( \{0, 1, \ldots, n - 1\} \) and the channel is in state \( l \) in slot 0 and in state \( j \) in slot \( n \).

i. Show the following relation:

\[
\Phi_{lj}(k, n) = \Phi_{l0}(k - 1, n - 1) p_{0j} + \Phi_{l1}(k, n - 1) p_{1j} + \delta_{lj} \delta_{k0} \delta_{n0}
\]

where \( \delta_{lj} \) equals 1 if \( l = j \) and equals 0, otherwise.

ii. Prove that the probability that a packet spends more than \( d \) slots in the buffer can be calculated as

\[
P_D(d) = \sum_{i=1}^{\infty} [P_L(0, i) \pi(0, i) + P_L(1, i) \pi(1, i)]
\]

which can be calculated as

\[
P_L(l, i) = \sum_{k=0}^{i} [\Phi_{l0}(k, d) + \Phi_{l1}(k, d)].
\]

**Hint:** Since the packet arrival process is memoryless, an arriving packet “sees” the steady state distribution of the system.

**Exercise 13.3:** Consider a Chase-combining based HARQ protocol with \( L \) transmission rounds. The same data symbols are transmitted over different transmission rounds where ACK/NAK control bits are fed back from the receiver to the transmitter indicating a transmission success/failure in each transmission round. Received signals over transmissions are combined by using the maximal ratio combining (MRC) technique before the receiver performs decoding. If the receiver still cannot correctly decode the data packet after \( L \) transmission rounds, an outage event is declared.

Assume that the transmitted signals \( x_s \) have average unit power and the Gaussian noise power at the receiver side is \( N_0 \). Moreover, let \( P_j \) denote the transmission power in
round $l$ and $h_{sd}$ represent the channel-gain, which is assumed to remain the same over all transmission rounds for each packet and $|h_{sd}|$ follows the Rayleigh distribution with mean value of zero and variance of $\sigma_0^2$. Then the SNR achieved in transmission round $l$ with the MRC can be expressed as

$$\gamma_l = \frac{\sum_{i=1}^{L} P_i |h_{sd}|^2 |x_i|^2}{N_0}.$$  

Since $h_{sd}$ remains the same over transmission rounds, the probability of decoding failure in round $l$, which is assumed to occur if the achieved SNR is below a target value $\gamma_0$, can be written as

$$p_{\text{out}}^l = \Pr\{\gamma_l < \gamma_0\} = 1 - e^{-\frac{\gamma_0 N_0}{\sigma_0^2 \sum_{i=1}^{L} r_i}} \quad \text{(13.85)}$$

where it is assumed that $p_{\text{out}}^0 = 1$.

i. Show that the probability that the HARQ protocol stops successfully its transmission in round $l$ (i.e., the transmission round $l-1$ is not successful but the transmission round $l$ is successful) is equal to $p_{\text{out}}^{l-1} - p_{\text{out}}^l$.

ii. Show that the average power required to transmit one packet (successfully or unsuccessfully) can be calculated as

$$P_{\text{avg}} = P_1 + \sum_{i=2}^{L} P_i p_{\text{out}}^{l-1}. \quad \text{(13.86)}$$

iii. Show that the outage constraint $p_{\text{out}}^L \leq p_0$ where $p_0$ denotes the maximum allowable value of the outage probability is equivalent to

$$\sum_{l=1}^{L} P_l \geq P_0 \quad \text{(13.87)}$$

where $P_0 = \frac{\gamma_0 N_0}{\sigma_0^2 \ln \frac{1}{1-p_0}}$.

**Exercise 13.4:** For the Chase-combining based HARQ system considered in Exercise 13.3, suppose we are interested in performing power allocation to minimize the average transmission power subject to the outage probability constraint. That means we want to solve the following problem:

$$\min_{\mathbf{P}} \quad P_{\text{avg}} \quad \text{s.t.} \quad \sum_{l=1}^{L} P_l \geq P_0 \quad \text{(13.88)}$$

where $\mathbf{P} = [P_1, P_2, \ldots, P_L]$ denotes the transmission power vector.
i. Prove that the optimal power allocation solution should satisfy the following relations:

\[ P_2 = \frac{\sigma_0^2 P_1^2}{\gamma_0 N_0} \]

\[ P_k = \frac{\sigma_0^2 \left( \sum_{l=1}^{k-1} P_l \right)}{\gamma_0 N_0} \left[ 1 - e^{-\frac{\gamma_0 N_0}{\sigma_0^2 \left( \sum_{l=1}^{k-1} P_l \right)}} \right] \left( \sum_{l=1}^{k-2} P_l \right) \].

Hint: Form the Lagrangian by relaxing the power constraint as follows:

\[ L(P, \lambda) = P_1 + \sum_{l=2}^{L} P_l \left[ 1 - e^{-\frac{\gamma_0 N_0 P_l}{\sigma_0^2 \left( \sum_{l=1}^{k-1} P_l \right)}} \right] - \lambda \left[ \sum_{l=1}^{L} P_l - P_0 \right]. \] (13.90)

Then study the Karush-Kuhn-Tucker optimality conditions, namely, \( \frac{\partial L}{\partial P_1} = 0, \frac{\partial L}{\partial P_2} = 0, \ldots, \frac{\partial L}{\partial P_k} = 0 \), for \( k = 3, 4, \ldots, L \).

ii. Prove the inequality constraint (13.89) should be met with equality at optimality:

\[ \sum_{l=1}^{L} P_l = P_0. \] (13.91)

Hint: This statement can be proved by contradiction.

**Exercise 13.5:** Consider the incremental redundancy HARQ protocol operating over a block fading channel where the channel is assumed to be invariant in each transmission round but change independently over different transmission rounds. The maximum allowable number of transmission rounds is equal to \( L \). We assume that each packet is formed by \( L_b \) information bits, which are encoded into \( L_s \) symbols \( x_1, x_2, \ldots, x_{L_s} \) drawn randomly and independently from a complex zero-mean Gaussian distribution. These \( L_s \) symbols are then partitioned into \( L \) sub-codewords to be transmitted over corresponding transmission rounds. Specifically, we assume that a sub-codeword \( x_i \) with \( L_s \) symbols is transmitted in round \( l \) if the decoding process in round \( l-1 \) fails. Moreover, the standard ACK/NAK bits are fed back from the receiver to the transmitter indicating the decoding outcome over an error-free and no-delay feedback channel.

We are interested in optimizing the rates in different transmission rounds to maximize the throughput of this HARQ system where the rate of round \( i \) is equal to \( R_i = L_b/L_{s,i} \) calculated in bits per channel use where each symbol corresponds to one channel use. We also define a related quantity \( \rho_i = 1/R_i = L_{s,i}/L_b \). Let \( c_i \) denote the mutual information achieved in round \( i \) conditioned on the channel-gain \( h_i \), whose amplitude follows the Rayleigh distribution. Then we have \( c_i = \ln(1 + |h_i|^2) \) where the unit noise variance and unit transmission power are assumed for simplicity. The transmission in round \( l \) is assumed to be successful if we have

\[ \sum_{i=1}^{L} c_i L_{s,i} \geq L_b \]
which is equivalent to

\[ S_l = \sum_{i=1}^{l} c_i \rho_i \geq 1. \]

We assume that the transmitter adapts its rate in transmission round \( i \) based on the feedback information \( S_{i-1} \) from the receiver in round \( i - 1 \). Let \( \bar{L}_b \) and \( \bar{L}_s \) denote average number of correctly received bits and the average number of channel uses, respectively.

i. Prove that the throughput can be expressed as

\[
\eta(\rho) = \frac{\bar{L}_b}{\bar{L}_s} = 1 - \frac{f_L(\rho)}{D(\rho)} = \frac{1 - \mathbb{E}_{C_1,\ldots,C_{K-1}}[\mathbb{I}(S_L < 1)]}{\mathbb{E}_{C_1,\ldots,C_{K-1}}\sum_{i=1}^{L}[\rho_i(S_{i-1})]} \tag{13.92}
\]

where \( f_L(\rho) \) denotes the outage probability as a function of \( \rho = [\rho_1, \rho_2, \ldots, \rho_L] \), \( \mathbb{I}(\cdot) \) represents the indication function, and \( \mathbb{E}_{C_1,\ldots,C_{K-1}}[\cdot] \) denotes the expectation over random variables \( C_1, \ldots, C_{K-1} \), which represent the mutual information in different transmission rounds.

**Hint:** Utilize the relation

\[ \bar{L}_s = \sum_{i=1}^{L} \mathbb{E}_{C_1,\ldots,C_{K-1}}[L_{L_s}] = \sum_{i=1}^{L} \mathbb{E}_{S_{i-1}}[L_b] = \sum_{i=1}^{L} \mathbb{E}_{C_1,\ldots,C_{K-1}}[\rho_i(S_{i-1})]. \]

ii. Suppose that we wish to solve the following optimization problem to determine \( \rho \):

\[
\min_\rho D(\rho) \tag{13.93}
\]

s.t. \( f_L(\rho) = \epsilon. \tag{13.94} \]

Note that with outage probability constraint (13.94), minimization of \( D(\rho) \) is equivalent to maximization of the throughput \( \eta(\rho) \). Consider the following Lagrangian obtained from this optimization problem by relaxing the outage constraint:

\[
L(\rho, \lambda) = D(\rho) + \lambda(f_L(\rho) - \epsilon). \tag{13.95}
\]

Show that the optimal \( \rho \) can be determined from the following dynamic programming (DP) based relations:

\[
J_1(S_0) = \min_{\rho_1(S_0)} \{ \rho_1(S_0) + \mathbb{E}_{C_1}[J_2(S_0 + C_1 \rho_1(S_0))] \}
\]

\[
J_2(S_1) = \min_{\rho_2(S_1)} \{ \rho_2(S_1) + \mathbb{E}_{C_2}[J_3(S_1 + C_2 \rho_2(S_1))] \}
\]

\[
\vdots
\]

\[
J_l(S_{l-1}) = \min_{\rho_l(S_{l-1})} \{ \rho_l(S_{l-1}) + \mathbb{E}_{C_l}[J_{l+1}(S_{l-1} + C_l \rho_l(S_{l-1}))] \}
\]

\[
\vdots
\]

\[
J_L(S_{K-1}) = \min_{\rho_L(S_{L-1})} \{ \rho_L(S_{L-1}) + \mathbb{E}_{C_L}[\mathbb{I}(S_{L-1} + C_L \rho_L(S_{L-1}) < 1)] \}
\]

for the optimal value of \( \lambda \) and \( S_0 = 0 \).

**Hint:** The DP based relations can be derived by minimizing the Lagrangian.
iii. To obtain the optimal $\rho$ from the recursive relations given in the previous question, one can solve the last equation for the discretized and finite values of $S_{L-1}$ (e.g., $K$ values of $S_{L-1}$ over the interval $[0, 1]$) to obtain $J_L(S_{K-1})$, which can then be used to optimize $\rho_{L-1}(S_{L-2}), \ldots, \rho_2(S_1), \rho_1(S_0)$. Show that the last equation in question (2) is equivalent to

$$J_L(S_{K-1}) = \min_{\rho_L} \left\{ \rho_L + \lambda F_C \left( \frac{1 - S_{L-1}}{\rho_L} \right) \right\}$$  \hspace{1cm} (13.96)

where $F_C(.)$ denotes the probability distribution function of the random variable $C$ (i.e., the mutual information).

**Exercise 13.6:** Consider an HARQ system operating over block fading channel where the wireless channel remains unchanged in each transmission round but changes independently over different rounds. Both Chase combining HARQ (CC-HARQ) and incremental redundancy HARQ (IR-HARQ) protocols are considered, and the number of symbols transmitted in each round is the same. The maximum number of transmission rounds is $L$. The next transmission round is requested if the transmitter receives an NAK message from the receiver over an error-free and no-delay feedback channel. For the CC-HARQ protocol, the maximal ratio combining (MRC) technique is employed to combine signals over different transmission rounds before decoding in each round. Let $A_i$ and $P_i$ denote the power channel-gain and transmission power of round $i$, respectively.

Assuming the unit noise variance, the outage probability in transmission round $l$ can be written as

$$f_l = \begin{cases} 
\Pr\left\{ \sum_{i=1}^l \ln(1 + P_i A_i) < R \right\}, & \text{IR-HARQ protocol} \\
\Pr\left\{ \ln(1 + \sum_{i=1}^l P_i A_i) < R \right\}, & \text{CC-HARQ protocol} 
\end{cases} \hspace{1cm} (13.97)
$$

$$= \Pr\{ S_l < S_{th} \} = F_{S_l}(S_{th}) = \int_0^{S_{th}} p_{S_l}(x) dx \hspace{1cm} (13.98)$$

where $F_{S_l}(.)$ and $p_{S_l}(.)$ denote the probability distribution function and probability density function of $S_l$, respectively, and

$$S_l = \sum_{i=1}^l c_i, \quad S_{th} = R \quad \text{for IR-HARQ} \hspace{1cm} (13.100)$$

$$S_l = \sum_{i=1}^l P_i A_i, \quad S_{th} = 2^R - 1 \quad \text{for CC-HARQ} \hspace{1cm} (13.101)$$

where $c_i = \ln(1 + P_i A_i)$.

Suppose that we are interested in performing power adaptation in transmission round $l$ based on the ACK/NAK as well as the receiver’s feedback information $S_{l-1}$ to minimize the outage probability subject to the average power constraint. This design problem
can be formally stated as

\[
\min_P f_L \quad \text{(13.102)}
\]

\[
\text{s.t. } \bar{P} \leq \bar{P}_\text{max} \quad \text{(13.103)}
\]

where \( P = [P_1, P_2, \ldots, P_L] \) denotes the power vector. We set \( \bar{P}_\text{max} = 1 \) in the following for simplicity.

i. Show that the average power can be calculated as

\[
\bar{P} = \frac{\sum_{i=1}^{L} E_{S_{i-1}} [P_i(S_{i-1})]}{\sum_{i=0}^{L-1} f_i} \quad \text{(13.104)}
\]

\[
= \frac{\sum_{i=1}^{L} \int_0^{S_{th}} P_i(x) p_{S_{i-1}}(x) dx}{\sum_{i=0}^{L-1} f_i} \quad \text{(13.105)}
\]

where \( f_0 = 1 \). We have explicitly specified \( P_i \) as a function of \( S_{i-1} \), i.e., \( P_i(S_{i-1}) \), and we have \( P_i(S_{i-1}) = 0 \) for \( S_{i-1} > 1 \) (i.e., correctly decoding the packet in transmission round \( i - 1 \)) according to the operations of the considered HARQ protocols.

ii. Show that the optimal power solution \( P \) can be obtained by solving the following recursive relations:

\[
J_1(S_0) = \left\{ \beta P_1 - \beta E_{A_1} \{1(S_1 \leq S_{th})\} + E_{A_1} \{J_2(S_1)\} \right\}
\]

\[
J_2(S_1) = \min_{P_2(S_1)} \left\{ \beta P_2(S_1) - \beta E_{A_2} \{1(S_2 \leq S_{th})\} + E_{A_2} \{J_3(S_2)\} \right\}
\]

\[
\vdots
\]

\[
J_{L}(S_{L-1}) = \min_{P_L(S_{L-1})} \left\{ \beta P_L(S_{L-1}) - \beta E_{A_L} \{1(S_L \leq S_{th})\} + E_{A_L} \{J_{L+1}(S_L)\} \right\}
\]

where

\[
S_I = S_{I-1} + \ln(1 + P_I(S_{I-1}) A_I) \quad \text{for IR-HARQ} \quad \text{(13.106)}
\]

\[
S_I = S_{I-1} + P_I(S_{I-1}) A_I \quad \text{for CC-HARQ} \quad \text{(13.107)}
\]

**Hint:** Study the related optimization problem:

\[
\hat{f}_L(P_1) = \min_{P_2(S_1), P_3(S_2), \ldots , P_L(S_{L-1})} f_L \quad \text{(13.108)}
\]

\[
\text{s.t. } \sum_{i=2}^{L} E_{S_{i-1}} [P_i(S_{i-1})] - \sum_{i=2}^{L-1} f_i \leq 1 + f_1 - P_1. \quad \text{(13.109)}
\]

In this problem, the outage probability \( f_i \) can be calculated as \( f_i = E_{A_i} \{1(S_i \leq S_{th})\} \). Then, the original problem becomes equivalent to

\[
\min_{P_1} \hat{f}_L(P_1). \quad \text{(13.110)}
\]
It can be verified that $J_1(S_0)$ is equal to the minimized Lagrangian of the optimization problem (13.108)–(13.109) obtained by relaxing its constraint where $\beta$ denotes the corresponding Lagrange multiplier.

References


[41] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, “Rate allocation and adaptation for incremental redundancy truncated HARQ,” *IEEE Transactions on Communications*, vol. 61, no. 6, June 2013, pp. 2580–2590.

Index

5G, 39
admission control, 105
area mean power, 20
ARQ, 379
    HARQ, 379
    Type I HARQ, 380
    Type II HARQ, 380
Automatic Repeat Request (ARQ), 93
binary exponential backoff, 79
Bluetooth, 43
Bose-Chaudhuri-Hocquenghem (BCH) code, 95
capture ratio, 83
Carrier Sense Multiple Access (CSMA), 77
cell association, 102, 276
    Cell Range Expansion (CRE), 282
    joint uplink cell association and TPC (UCA-TPC), 279
    reference signal received power (RSRP)-based cell association, 281
    Reference Signal Received Quality (RSRQ)-based cell association, 283
combined delay and loss system, 71
convex function, 117
convex optimization, 123
    duality, 124
    interior point method, 128
    KKT conditions, 125
    projected subgradient method, 131
    subgradient method, 130
convolutional code, 95
cooperative diversity, 309
cooperative relaying, 309
    amplify-and-forward relaying, 309
    decode-and-forward relaying, 310
    incremental relaying, 312
    opportunistic relaying, 311
    two-way relaying, 313
D2D communication, 37
digital modulation, 8, 10
Distributed Coordination Function (DCF), 78
distributed joint admission and power control
    TPC with both temporary and permanent removal (TPC-TPR), 235
    TPC with soft removal (TPC-SR), 239
    TPC with temporary removal, 234
distributed joint power and admission control
    TPC with permanent removal (TPC-PR), 233
diversity combining, 65
diversity transmission, 65
dynamic programming, 143
electromagnetic spectrum, 4f
Erlang delay system, 71
Erlang loss system, 70
Erlang-B formula, 70
Erlang-C formula, 71
Exponential/Proportional Fair (EXP/PF) scheduling, 74
FEC (Forward Error Correction), 93
Finite-State Markov Channel (FSMC) Model, 24
Forward Error Correction Coding (FEC)
    coding gain, 95
    incremental redundancy, 98
game theory, 154
    auction theory, 178
    combinatorial auction, 181
    double auction, 180
    second-price auction, 179
    battle of sexes, 185
    Bayesian game, 166
    Bayesian Nash equilibrium, 167
    Bertrand competition, 182
    coalition game, 174
    continuous game, 159
    cooperative game, 172
    Cournot duopoly, 184
    dynamic game, 162
    evolutionary duopoly, 170
    evolutionary game, 168
    evolutionary stable strategy, 169
Index

game theory (cont.)
gloves game, 186
Nash bargaining solution, 172
Nash equilibrium, 157
non-cooperative game, 159
Pareto efficiency, 172
Pareto optimality, 158
pirate game, 184
Shapley value, 177
Stackelberg equilibrium, 164
static game, 159
ultimatum game, 184
Go-Back-N (GBN) ARQ, 97
gradient descent method, 121
greedy scheduling, 72
handoff management, 104
harmonic mean fairness, 74
heterogeneous network, 37
hybrid ARQ, 98
type-I hybrid ARQ, 98
type-II hybrid ARQ, 98
i.i.d. channel model, 23
IEEE 802.11, 39
IEEE 802.11 WLAN, 353
channel assignment, 356
channelization, 353
IEEE 802.15, 43
IEEE 802.16, 42
integer programming, 132
branch and bound method, 135
cutting plane method, 132
Jain’s fairness index, 73
link adaptation, 62
Adaptive Modulation and Coding (AMC), 381
local mean power, 20
Local Multipoint Distribution Service (LMDS), 5
LTE, 39
LTE-Advanced (LTE-A), 39
max-min fairness, 73
max-SNR scheduling, 72
maximal ratio combining, 66
Maximum-Largest Weighted Delay First (M-LWDF) scheduling, 74
Moment Generating Function (MGF), 61
Multichannel Multipoint Distribution Service (MMDS), 5
multipath fading, 16
Newton method, 122
OFDM, 291
adaptive resource allocation, 293
MIMO-OFDM, 298
OFDM-CDMA, 293
OFDM-TD, 293
OFDMA, 291
packet error rate, 64
Packet Reservation Multiple Access (PRMA), 77
Point Coordination Function (PCF), 78
Poisson point process, 82
power control, 99
closed loop power control, 100
distributed dynamic target-SINR tracking power control (DTPC), 223
distributed opportunistic power control (OPC), 222
distributed target-SINR tracking power control (TPC), 220
inner loop power control, 100
open loop power control, 99
outer loop power control, 100
power-update function, 218
SINR feasibility, 204
standard functions, 217
standard type I functions, 218
standard type II functions, 217
TPC with PU-Protection Algorithm (TPC-PP), 267
proportional fairness, 73
pure ALOHA, 75
Radio Link Control (RLC), 379
Reed-Solomon (RS) code, 95
robust optimization, 141
Selective Repeat (SR) ARQ, 97
shadow fading, 14
signal area coverage/area reliability, 16
signal bandwidth, 8
signal coverage probability, 15
signal outage probability, 15
SINR coverage probability, 23
SINR outage probability, 23
slotted-ALOHA, 75
small cell network, 37
spectral efficiency, 23, 59
stochastic optimization, 140
Stop-and-Wait (SW) ARQ, 96
Taylor’s theorem, 120
transmission capacity, 23
WiMAX, 42
ZigBee, 43