RICE UNIVERSITY

Distributed Cooperative Communications in Wireless Networks

by

Mohammad Ali Khojastepour

A Thesis Submitted
in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Behnaam Aazhang, Chair
J.S. Abercrombie Professor
Electrical and Computer Engineering

Richard G. Baraniuk
Professor
Electrical and Computer Engineering

Don H. Johnson
J.S. Abercrombie Professor
Electrical and Computer Engineering

Ashutosh Sabharwal
Research Professor
Electrical and Computer Engineering

Richard A. Tapia
N. Harding Professor
Department of Computational and Applied Mathematics

Houston, Texas
Abstract

Distributed Cooperative Communications in Wireless Networks

by

Mohammad Ali Khojastepour

The primary challenge in communication over wireless networks, unlike wireline networks, is the existence of interference and channel variations (fading). Having more users at higher data rates means that current point-to-point networks will not scale. To engineer a scalable network, we introduce a new paradigm that exploits different network characteristics. We show that cooperation between users in the network, network coding, not only reduces existing (destructive) interferences from other users but it can also generate constructive interference, transforming the destructive interference into useful information.

In this thesis, we explore the problem of source and channel coding over wireless networks, ranging from information theoretical analysis to code design and practical implementation issues. We show that significant gains in throughput can be achieved through network coding. Despite the importance of the problem and the work done on wireless networks, little is known about network coding and the effective use of
the relaying function and cooperative strategy at the intermediate nodes. A notable example is the lack of an optimal coding scheme over the relay channel, the simplest form of a network, which has remained an outstanding open question for the last three decades.

We propose new approaches to network coding that improve upon the best known coding schemes by many decibels. Specifically, we develop two main coding techniques, one for the multi-state relay channel and the other for the multiple access with generalized feedback (MAC-GF). We show that by using the new coding techniques, higher transmission rates than those previously known are achievable. The first technique achieves the ultimate transmission rate (capacity) for both half-duplex and the original relay channel under certain conditions. These improved capacity results for the relay channel are the only known results since Cover’s in 1979 and El Gamal’s in 1982. The second coding technique improves the best known achievable transmission rate for the MAC-GF by Willems in 1983. This latter result also improves the achievable transmission rate for the Gaussian relay channel over all other known schemes for some channel conditions.

We also present a practical code design technique for the relay channel. The design gains more than 4dB over direct transmission and closes the gap to the relay channel Shannon limit to less than 1dB with a code length of only $2 \times 10^4$ bits. The new coding techniques and transmission strategies developed in this thesis provide important steps toward overcoming the challenges of wireless network coding.
Acknowledgments

I would like to express my profound gratitude to the individuals who have helped me during my graduate study at Rice University to develop this thesis. My words here are insufficient to capture all of my appreciation.

First and foremost, I would like to thank my advisor Dr. Behnaam Aazhang. He has helped me reach several milestones as an academic and prepared me to reach many more. I want to thank him for being a great source of enthusiasm, encouragement, insight, and inspiration. I have benefited extensively from his countless efforts and dedication to the intellectual and personal growth of his graduate students. He made my graduate-student experience enjoyable and memorable.

I would also like to thank my committee members, Dr. Ashutosh Sabharwal, Dr. Richard G. Baraniuk, Dr. Don H. Johnson, and Dr. Richard A. Tapia for many useful interactions and for their comments and suggestions on my work. I am particularly grateful to Dr. Ashutosh Sabharwal for the productive collaboration we have had during the past few years and for his invaluable insights and comments on my work. I am also grateful to all the professors at Rice from whom I have learned a lot not only during the courses but also during numerous research discussions. In particular, I am thankful to Dr. Joseph Cavallaro and Mike Orchard for the helpful discussions in the joint group meeting and for their support.

I would also like to thank all my friends and members of the communication and
DSP groups. I feel lucky to have so many incredible friends around. I thank each and every one of them for their trust and friendship.

As always, I would like to thank my family for their love and constant support throughout my life and for inspiring me to pursue an academic career: Mom and Dad for giving the very first lectures I attended; and my sisters, Nazi, Mitra, Fariba, Roya, and Sara, for their invaluable love, encouragement, support, and sacrifices. I thank them, and the rest of my extended family. Their understanding throughout these many years has meant more than I can ever express.

Last, my most tender and sincere thanks go to my loving wife Vida who has been impacted by this dissertation and the years of my graduate studies at Rice more than anyone else has. I want to thank her for her love, patience and support.
# Contents

Abstract ii  
Acknowledgments iv  
List of Illustrations xiv

1 Introduction 1  
1.1 Known Results in Network Information Theory . . . . . . . . . . . . 4  
1.2 Multi-Terminal Network Information Theory . . . . . . . . . . . . . 6  
1.3 Information Theory of the wireless Networks . . . . . . . . . . . . . 9  
1.4 Challenges of Communication in Wireless Networks . . . . . . . . . . 10  
1.5 Organization of the Thesis . . . . . . . . . . . . . . . . . . . . . . . 13

2 Multi-State Networks 17  
2.1 General Cut-set Bounds for the Multi-state Networks . . . . . . . . . 18  
  2.1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19  
  2.1.2 Problem Formulation . . . . . . . . . . . . . . . . . . . . . . . 21  
  2.1.3 Multi-State Cut-Set Bounds . . . . . . . . . . . . . . . . . . . . 24  
    2.1.3.1 Single Cut Bound . . . . . . . . . . . . . . . . . . . . . 25  
    2.1.3.2 Max-Flow Min-Cut Bound . . . . . . . . . . . . . . . . 31  
  2.1.4 Applications of the New Bounds . . . . . . . . . . . . . . . . . 34  
    2.1.4.1 Compound Channel . . . . . . . . . . . . . . . . . . . . . 35
2.1.4.2 ‘Cheap’ Relay Channel ........................................ 37
2.1.5 Concluding Remarks .............................................. 43
2.2 General Achievable rate for the Multi-state Networks ............. 44
  2.2.1 Introduction .................................................. 44
  2.2.2 General Results ............................................... 45
  2.2.3 Gaussian Cheap Relay Channel ................................. 49
  2.2.4 Multi-hop Gaussian Cheap Relay Channel ...................... 53
  2.2.5 Concluding Remarks ........................................... 56

3 Gaussian Relay Channel ............................................... 57
  3.1 Summary of existing bounds on the Capacity of the Relay channel . 58
  3.2 Preliminaries ................................................... 60
  3.3 Upper Bound on the Capacity .................................... 61
  3.4 Lower Bounds on the Capacity .................................. 63
  3.5 New Achievable rate for the Gaussian relay channel ............... 69
  3.6 Concluding remarks ............................................. 74

4 Cheap Relay Channel .................................................. 76
  4.1 Discrete Memoryless Cheap Relay Channel ........................ 77
    4.1.1 Introduction ............................................... 77
    4.1.2 Cheap Relay Channel ...................................... 78
4.1.2.1 Achievability of $C$ in Theorem 4.1.1 . . . . . . . . . 82
4.1.2.2 Encoding and Decoding . . . . . . . . . . . . . . . . 84
4.1.2.3 Analysis of the error probability . . . . . . . . . . . 86
4.1.3 Cascaded Network with Cheap Nodes . . . . . . . . . . . . . 97
4.1.3.1 Capacity of the Cheap Cascaded Network . . . . . . 98
4.1.3.2 Sketch of the Proof for Theorem 4.1.2 . . . . . . . . 99
4.1.4 Concluding remarks . . . . . . . . . . . . . . . . . . . . . . . . 102
4.1.5 Bounds on Non-typical Event Probabilities . . . . . . . . . . 103
4.2 The Gaussian Cheap Relay Channel . . . . . . . . . . . . . . . . . . 105
4.2.1 Upper bound on the Capacity . . . . . . . . . . . . . . . . . . 106
   4.2.1.1 Bound on $I(X_{1i};Y_i,Y_{1i}|m_1)$ . . . . . . . . . . . . . 107
   4.2.1.2 Bound on $I(X_{1i};Y_i|X_{2i},m_2)$ . . . . . . . . . . . . . 108
   4.2.1.3 Bound on $I(X_{1i};Y_i|m_1)$ . . . . . . . . . . . . . . . . 110
   4.2.1.4 Bound on $I(X_{1i},X_{2i};Y_i|m_2)$ . . . . . . . . . . . . . 111
4.2.2 The Gaussian Degraded Cheap relay Channel . . . . . . . . 113
4.2.3 Achievability of $C$ in Theorem 4.2.2 . . . . . . . . . . . . 114
4.2.4 Concluding remarks . . . . . . . . . . . . . . . . . . . . . . . . 117

5 New Capacity Results for the Relay Channel 119
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
5.2 Expensive Relay Channel . . . . . . . . . . . . . . . . . . . . . . . . 122
5.3 ‘Cheap’ Relay Channel ....................................................... 123
   5.3.1 Time Division Relay Channels .................................... 124
   5.3.2 Frequency Division Relay Channels .............................. 125
5.4 Unifying Approach to Time and Frequency Division Relaying .... 126
   5.4.1 Bounds on the Capacity of Cheap Relay Channel ............. 127
   5.4.2 New Capacity results for discrete memoryless relay channel .. 131
5.5 Concluding remarks ...................................................... 134

6 Fading Relay Channel 136

   6.1 Introduction ............................................................... 137
   6.2 Preliminaries ............................................................. 138
   6.3 Complex Gaussian Relay Channel ..................................... 139
      6.3.1 Upper Bound on the Capacity of the Complex Gaussian Relay Channel ..................................................... 139
      6.3.2 Lower Bound on the Capacity of the Complex Gaussian Relay Channel ..................................................... 145
   6.4 Fading Relay Channel .................................................... 149
      6.4.1 Upper Bound on the Capacity ..................................... 149
      6.4.2 Lower Bounds on the Capacity .................................... 151
      6.4.3 Ergodic Capacity of the Fading Relay Channel ............... 154
   6.5 Numerical Results ....................................................... 157
7 Fading Channel: Peak and Average Power Constraints 162

7.1 Introduction .......................................................... 162
7.2 Problem Formulation and Notations ................................. 165
7.3 Capacity Analysis ....................................................... 167
7.4 Coding Theorem ......................................................... 171
7.5 Converse ................................................................. 176
7.6 Numerical Results ...................................................... 180
7.7 Conclusion ............................................................... 183

8 Joint Source-Channel Coding in Wireless Networks:

Application to Sensor Networks 185

8.1 Introduction .......................................................... 185
8.2 Prior Works on Joint Source-Channel Coding of Correlated Sources . 188
8.3 Channel and Source Models ......................................... 190
8.4 Rate Region for User Cooperation Channel with Correlated Sources . 194
  8.4.1 Main Result ........................................................ 194
  8.4.2 Special Cases ....................................................... 196
  8.4.3 Examples and Discussion ......................................... 199
8.5 Proof of Theorem 8.4.1 ............................................... 201
9 Code Design For the Relay Channel

9.1 Introduction .................................................. 211
9.2 Preliminaries .................................................. 214
9.3 Capacity of the Gaussian relay channel with finite alphabets .... 215
  9.3.1 Upper and lower bounds on the capacity .................. 217
  9.3.2 Analysis of the minimum energy per bit required for reliable communications .................. 221
9.4 General Code Design Approach for the Relay Channel .......... 226
  9.4.1 Protocol .................................................. 226
    9.4.1.1 A simple protocol ................................. 227
    9.4.1.2 DF-protocol ..................................... 228
  9.4.2 Constituent codes .................................... 228
  9.4.3 Power allocation .................................... 229
9.5 Relay Channel Code Design Based on LDPC Constituent Codes ... 232
  9.5.1 Factor Graph Representation of a Block Code ............. 233
  9.5.2 Relay Node Operation .................................. 235
    9.5.2.1 Relay operation in the simple protocol with LDPC constituent codes ............ 235
9.5.2.2 Relay operation in the DF-protocol with LDPC constituent codes .................................. 238

9.5.3 Decoding at the Destination Node .................................................................................. 239
  9.5.3.1 Forward decoding algorithm .................................................................................... 241
  9.5.3.2 Backward decoding algorithm ................................................................................ 242
  9.5.3.3 Solving the partial factor graphs ............................................................................. 243

9.6 Performance ....................................................................................................................... 244

9.7 Conclusion ......................................................................................................................... 247

10 Cross-Layer Problems in Wireless Networks  .................................................................... 248

10.1 Introduction ...................................................................................................................... 249

10.2 Guaranteed Maximum Delay Scheduler .......................................................................... 251
  10.2.1 Problem Formulation ............................................................................................... 251
  10.2.2 Set of Feasible Schedulers ....................................................................................... 253

10.3 Feasible Schedulers are Low-Pass Filters ....................................................................... 255

10.4 Optimal Time-invariant Scheduling ................................................................................ 258

10.5 Bounds on Scheduler Performance .................................................................................. 261

10.6 Optimal Robust Scheduling ............................................................................................. 267
  10.6.1 Optimal Time-varying Robust Scheduler .................................................................. 268
    10.6.1.1 Different delay constraints for each packet ....................................................... 268
    10.6.1.2 The same delay constraints for all packets ....................................................... 274
10.6.2 Performance Analysis ................................. 277
10.7 Related Work ............................................. 282
10.8 Conclusion ................................................. 285

11 Conclusions ................................. 286

Bibliography ................................................. 289
<table>
<thead>
<tr>
<th>Illustrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 A general multi-terminal network. Two arbitrary cut-sets $C_1$ and $C_2$ are shown, each of which have the set of source nodes, $S_1$, on one side and the set of destination nodes, $S_2$, on the other side.</td>
</tr>
<tr>
<td>2.2 ‘Cheap’ Relay Channel: Representation of the cut-sets. Notes: 1-The relay can either transmit or receive at any given time. 2-Both of the cut-sets exist in either “Broadcast” or “Multiple access states”.</td>
</tr>
<tr>
<td>2.3 ‘Cheap’ relay channel: The representation of two states of operation as a “broadcast state” ($m_1$) and a “multiple access state” ($m_2$).</td>
</tr>
<tr>
<td>2.4 Inner and outer bounds (in bits/channel use) on the capacity of the cheap relay channel of Figure 2.2 versus cross over probability of the direct link, $p$, where each of the links $S \rightarrow D$, $S \rightarrow R$, and $R \rightarrow D$ are binary symmetric channels with cross over probabilities $p$, $q$, and $r$ respectively. The $q$ and $r$ take values from the interval $[0, \frac{1}{4}]$ as a function of $p$, and are symmetric with respect to $p = \frac{1}{2}$, i.e., $q(p) = q(1 - p)$ and $r(p) = r(1 - p)$.</td>
</tr>
<tr>
<td>2.5 Gaussian ‘Cheap’ Relay Channel</td>
</tr>
<tr>
<td>2.6 Multi-hop Gaussian Relay Channel</td>
</tr>
<tr>
<td>3.1 Gaussian Relay Channel</td>
</tr>
</tbody>
</table>
3.2 Relaying region for which each of the different coding schemes
outperforms other schemes. \((\gamma_1 = 1, P_1 = P_2 = -10\text{dB})\). ............. 72

4.1 Discrete Memoryless Relay Channel ................................. 79

4.2 Cascaded Network ....................................................... 97

5.1 (a) Expensive relay channel, (b) Frequency division relay channel -
Model I. The solid and dashed lines represent two orthogonal channels 123

5.2 (a) Frequency division relay channel - Model II, (b) Frequency
division relay channel - Model III. The solid and dashed lines
represent two orthogonal channels ........................................ 126

6.1 Fading relay channel ..................................................... 139

6.2 Relaying region for which the capacity of the Rayleigh fading relay
channel is known where the relay node lies in this region ............. 159

6.3 Upper and lower bounds on the capacity of the fading relay channel,
where \(d\) denotes the source-relay and \(D\) denotes the
source-destination distance ............................................... 161

7.1 System model for communication over the fading channel with
transmitter side information. .............................................. 166

7.2 Optimal power adaption for the fading channel with both peak and
average power constraint. ............................................... 180
7.3 Capacity of the Rayleigh fading channel with different power constraints

7.4 Capacity of the Rayleigh fading channel for a fixed peak to average power ratio.

8.1 User cooperation channel with correlated sources

9.1 Gaussian relay channel.

9.2 A physical model Gaussian relay channel with the source, relay and destination located along a line of unit length.

9.3 Upper and lower bound on the required $E_b/N_0$ for the given transmission rate.

9.4 Effect of the power allocation (splitting) on the rate, for a given $E_b/N_0$. The optimal value of the power splitting factor is denoted by $\kappa^*$ which corresponds to the $(E_b/N_0)_{\text{min}} = -3.5 dB$ required to achieve rate $R = 0.5$ with the decode-and-forward scheme.

9.5 (a) Successive decoding procedure at the destination node. The partial factor graphs are indicated by boxes and the direction of decoding is indicated by the arrows (forward decoding in this example), (b) Solution of a partial factor graph. In this figure, $r_2$ is a multiple access code based on two codewords, while $r_1$ is a received signal based solely on the code indicated by part 1.
9.6 Factor graph representation of the code. (a) Representation of a
(3, 6) regular LDPC code. (b) shorthand (symbolic) representation of
a factor graph with a vector variable node (the circle), a vector check
node (the square), and the parity check matrix $H$ which represents
the connection between these to vectors. (c) shorthand representation
of direct connections which corresponds to $H = I$. 234

9.7 Operation of the relay node in a simple relay protocol (Section 9.4.1). 237

9.8 Decoding at the destination node. The factor graph representation of
the proposed code for the relay channel coding, denoting both the
constituent codes and the protocol that shows the connection of the
constituent codes and received vectors. The optimal decoder at the
destination node should solve this factor graph to decode all $B$
transmitted block jointly. 240

9.9 Performance of proposed relay channel code with simple protocol
defined in Section 9.4.1. Performance of a single user code using the
same constituent parity check matrix and Shannon limit for any
single user codes are shown for comparison. 244
9.10 Performance of the proposed relay channel code with

decode-and-forward relay protocol and optimized LPDC codes.

Constituent code length $N = 20,000$ with $B = 40$ Blocks. Depicted

curve is for the position of relay at $d=0.25$ in figure 9.2. Required

$(E_b/N_0)_{\text{min}}$ for this position of relay is shown for comparison. . . . . .  246

10.1 Conceptual depiction of a queuing system. . . . . . . . . . . . . . . .  251

10.2 Performance of the optimal time-varying robust scheduler and the

bounds on performance. Note that performance of the optimal

time-invariant robust scheduler matches the upper bound. . . . . . .  266

10.3 Representation of the successive water-filling at time $t$ for the $i^{th}$ queue.  269

10.4 Water-filling solution for the time slot $t - 1$ and subsequent time slot $t$.  274
Chapter 1

Introduction

Explosive growth in the use of the Internet and wireless communications in recent years has transformed consumer technology and posed new problems for wireless communication service providers. Not only do people view cell phones, wireless text-messaging devices, wireless-enabled handheld computers, pocket PCs, laptops, and cameras as essential consumer devices, they also expect more reliable and faster services. The proliferation of these wireless applications sets the data rate and quality of service requirements for the next generation of wireless devices an order of magnitude higher in order to meet the increasing demands for high-speed wireless communications. Not only should current data rates be improved by introducing an intelligent network coding but future efforts should also provide increased battery life while supporting the required quality of service over the available bandwidth, which translates to power optimal systems with high spectral efficiency.

The communication problem in wireless networks poses a new set of challenges that do not exist in either point-to-point wireless communication or wired networks. Unlike wired communication, in which the media can be easily duplicated (several wires between two points), communication in the wireless environment is contingent on very scarce and limited resources. The most important of all are frequency and
power. Frequency is scarce because once it is used for communication between two points it cannot be duplicated. The available power of a wireless device is also limited because of the mobility of wireless devices that cannot be attached to power outlets like wired communication systems.

Not only do wireless communication systems have to deal with the scarce and limited system resources, but they also suffer from power attenuation due to large pathloss and channel variation in fading environments which makes the communication problem in a wireless environment more challenging. To overcome these problems, communication can occur over multiple hops in a network of communicating nodes instead of point-to-point communication between a transmitter and a receiver. As a result, the transmission range can be reduced by using multiple hops and the communication message signal can go through multiple paths to increase the reliability.

Multiple active transmitters, however, introduce a unique challenge in wireless networks because simultaneous transmission from different nodes generates interference. In fact, without accounting for interference, the communication problem in wireless networks can be modelled as that of a wired network with time-varying point-to-point links where the fading is represented with time variation. However, the existence of interference can be turned to our favor. One of the primary goals of this thesis is to use network coding to exchange the role of the existing destructive interferences into constructive interference, a process we refer to as cooperation. Cooperation is
concurrent joint channel coding at two or more terminals in the network in order to generate a constructive interference which enables us to increase the achievable rate(s) of transmission and decrease the transmission power.

In addition to using cooperation to resolve the problems of channel coding and communication between the nodes, the same idea of cooperation lies behind a solution to the problem of source coding or joint source-channel coding. We refer to such cooperation between different sources of the network as dialogue. The dialogue between the encoding terminals allows us to match the sources to the channel, while cooperation solely tries to maximize the possible rate of transmission for two independent or uncorrelated sources. Therefore, dialogue becomes necessary when the sources at the transmitting terminals are correlated.

In this introduction, we first give a brief history of the known capacity results in network information theory, starting with the introduction of various spacial multiple user channels. Then, we consider more general results on multi-terminal network information theory. After that, a short overview of the recent works characterizing the capacity of wireless networks is given. Finally, after discussing the challenges of communication over wireless networks (mostly from the physical layer perspective), we present the organization of the rest of the thesis and briefly summarize the major achievements in our work.
1.1 Known Results in Network Information Theory

Network information theory in general addresses the problem of communication and information transfers between intended subsets of the nodes in a multi-terminal network of senders and receivers. Current results in network information theory are mainly a collection of results for special network topologies or channel models, including two-way channels, interference channels, multiple access channels, broadcast channels, and relay channels. The development of multiple user information theory started with the definition of various types of multiple user channels and attempts to find the corresponding capacity regions. Among the investigated channels there are a few channels for which the exact capacity has been derived, but usually the capacity is not known. Finding the capacity of a communication channel is important not only to know the limit of the communication, but also to show how this ultimate communication limit can be achieved. In other words, finding the capacity shows what kind of coding and transmission protocol achieves the highest possible rate of transmission in a given communication system.

The definition of the two-way channel by Shannon, given in one of his pioneering papers in 1961 [1], is often considered the first work on multiple user information theory. It is quite interesting that in different parts of his paper, Shannon also considered two other types of multiple user channels that almost a decade later developed into the multiple access channel and the interference channel.

Shannon did not find the capacity result for the general two-way channel (and it
is still not known), but in his paper [1] Shannon informally defined a channel (which has later been called a multiple access channel) and mentioned that “a complete and simple solution of the capacity region” for this channel has been found. He never published his result on the multiple access channel, and almost a decade later, in 1971, Ahlswede presented a simple characterization of the multiple access channel [2].

Shannon also considered a “restricted” two-way channel where sender and receiver points at each end are in different places with no direct cross communication. Carleial later defined this channel in terms of two sender-receiver pairs where the transmission of information from each sender to its corresponding receiver interferes with the communications between other sender-receiver pairs [3, 4].

During the time Shannon and Ahlswede were working, Cover first introduced the broadcast channel in 1972. In a broadcast channel, a single source sends different pieces of information to several destinations [5]. This channel can be also viewed as a dual to a multiple access channel where multiple sources trying to communicate with a single destination node.

The relay channel was first defined and studied by Van der Mullen [6] for the case of three terminals (or nodes), and it was later considered by Cover [7]. Cover’s results on the relay channel are still the most comprehensive in terms of achievability and capacity bounds (converse). Although in the relay channel coding problem there is just one sender and receiver, it is considered in the domain of network information theory due to the presence of a relay node (or nodes). In fact, the problem of relay
channel coding and finding the best relaying function in a simple three terminal network, consisting of the source, the relay, and the destination, captures the essence of the challenging problem of network coding.

The aforementioned works were first steps in finding better achievable rates or capacities. Although not many capacity results have been derived, these efforts usually have succeeded in developing some important ideas, including deriving better achievable rates, considering some interesting special cases and special constraint, investigating the corresponding channel with continuous alphabet (usually as AWGN channel), and exploring the effect of feedback.

1.2 Multi-Terminal Network Information Theory

In the aforementioned multiple user channels and in general in network information theory, new elements such as interference, cooperation, and feedback make the problem of reliable communications more challenging. Because it is difficult to derive the capacity of different communication channels, capacity results are only known for very limited cases with special constraints. In most of the cases where the capacity results are known, they coincide with the cut-set bound on the achievable rates in a network [8]. The bounds have nice and simple max-flow min-cut representation: The rate of information flow across any boundary is less than or at most equal to the conditional mutual information between the inputs on one side of the boundary (senders’ side) and the output on the other side (receivers’ side) given the inputs on
the receiver’s side.

The cut-set theorem is one of the most general results in network information theory. Although the converse (if it exists) to the coding theorem for different channels has been proven independently, in most of the cases it corresponds to a simple representation of the cut-set bounds. For example, the capacity region of the multiple access channel is known because the discovered achievable region for this channel is large enough to coincide with the cut-set bound. Another example is the relay channel for which an achievable rate was derived by Cover [7]. This achievable rate simply coincides with the cut-set bound if the relay channel is degraded. Therefore, the capacity of degraded relay channel is known, although the bound of non-degraded relay channel is higher and is still not shown to be achievable.

It should be pointed out that the cut-set bounds are not always achievable. In fact, there are some cases where an upper bound on the capacity has been derived which is tighter than the bound obtained by cut-set theorem. For example, the capacity region of the general broadcast channel is still not known because either the outer bounds are not tight enough or the achievable rate region is not large enough. But recently, the sum capacity of the Gaussian vector broadcast channel that is lower than the corresponding suggested cut-set bound on the sum rate (GVBC) has been considered and derived [9, 10, 11, 12]. The outer bound on this sum rate capacity is derived from the work of Sato [13], and the achievability part is based on Costa’s precoding [14] or on Marton’s achievable region for the general broadcast channel [15]. More
recently, Hanan Weingarten, Yossef Steinberg, and Shlomo Shamai showed that the achievable rate obtained earlier from Costa’s precoding is in fact the capacity region of the Gaussian vector broadcast (MIMO broadcast) channel [16]. The proof is based on establishing a new converse by showing the existence of a test channel for any point on the capacity region. This test channel is degraded and its capacity region is known. However, the capacity region of this test channel provides an upper bound on the capacity (at the considered point of interest) of the original Gaussian vector broadcast channel.

Some recent works on network information theory have considered the problem of multi-casting in the networks, which they regard as network information flow. Ahlswede et al. [17, 18] introduced a new class of multi-user source coding problems inspired by network communications and proposed using coding in the network for the purpose of multi-casting. This work can be regarded as the max-flow min-cut theorem for the network information flow. Meanwhile, this work reveals that in general coding at the nodes (network coding) is required to achieve the best rates and that the information should not be treated as a ‘fluid’ that can simply be routed and replicated at the nodes. Li and Yeung [19] presented a linear code for multi-casting from a single-source, which is an example of such a network code. Borade applied the cut-set theorem [8] to the problem of the network information flow and derived an information theoretic upper bound on the information flow in discrete memoryless networks [20]. These works offer the best insights into the importance of
network coding for wired networks and the limitations on our knowledgeable use of network coding to solve the same problems in wireless network with the existence of interference and channel variation that this thesis investigates.

1.3 Information Theory of the wireless Networks

Because information theoretical analysis of capacity and achievable rates in the network is complicated, this analysis might not explain the construction and limits of practical communication systems. Gupta and Kumar’s work on the capacity of large networks that characterized the achievable throughput of the nodes [21] presented a new analysis approach to this problem. The assumption of point-to-point coding in their work precludes any kind of cooperation or network coding, which is possible for example in broadcast or multiple access channels.

With the new interest in sensor networks in the past few years researchers began to understand the importance of network coding. The new effort in better understanding of network information theory has in turn sparked new research into the relay channel coding problem. Subsequent work by Gupta and Kumar [21] on an extension of the relay channel shows that using more sophisticated multi-user coding schemes can provide sizable gains in terms of transport capacity. Also, a follow-up paper by Xie and Kumar [22] established an explicit achievable rate expression for the degraded Gaussian channel with multiple relays and characterized the bounds on transport capacity. Reznik et al. [23] considered Gaussian physically degraded relay
channels and extended the results for multiple relay stages with a total average power constraint.

Gasper and Vetterli [24] derived a lower bound on the capacity of the relay channel by allowing arbitrarily complex network coding. Also, they considered upper bounds from the cut-set theorem [8] and showed that these upper and lower bounds meet asymptotically as the number of relay nodes goes to infinity. This result again shows that network coding is essential.

1.4 Challenges of Communication in Wireless Networks

Interference, feedback, and cooperation pose significant challenges in the communication problem in wireless network. Feedback and cooperation have received most of the attention so far as researchers try to derive better achievable rates or increase the network capacity in different scenarios. However, the effect of interference can be quite capacity limiting, and it has not received enough attention. In addition, most of the current radios in sensor nodes operate in TDD mode when the transmitting and receiving frequencies are the same, thus sensor nodes cannot transmit and receive at the same time. As a result, the achievable rate in the network is lower than a system without TDD limitation.

This practical constraint will force us to use the nodes of the network as either a sender or receiver at any given time, which means that there is more than one mode of operation in the network. Each mode of operation in the network corresponds to the
valid partitioning of the nodes into two disjoint subsets of sender and receiver nodes such that there is no node in the sender nodes set which is going to communicate with another node in that subset.

One of the major motivations of our work on multi-state network information theory in this thesis comes from the aforementioned practical constraint. In fact, the limitation of using only half-duplex radios induces a multi-state characteristic on the network. Because the presence of multiple states in a network might be the effect of channel variation as well, we consider the most general case of the problem where the network has a finite number of states, and we derive upper and lower bounds on the achievable information rate in such a network. The derived bounds coincide with the known cut-set bound in network information theory [8] when the network has just one state. The bounds hold for the network with the mentioned practical constraint if modes are considered as states and if the number of nodes in the network are finite. Later, we will give two applications of the derived bound to provide a tight upper bound where the cut-set bound [8] is unable to provide such bounds. In these two specific examples, we show that the derived bounds are tight enough to characterize the capacity.

The hostile nature of fading in a wireless environment is another challenge in communication over wireless networks. Usually, the transmission strategy and power are adapted to combat the effect of fading. But the limiting property of fading is more dominant in wireless networks in comparison to point-to-point communication
because the number of unknown parameters is larger. Moreover, estimation of these parameters and feeding back the estimated channel characteristics create a much harder problem in the network. Therefore part of this thesis is dedicated to the analysis of efficient strategies for communication over the network.

Cross-layer design issues are an extremely important part of the effort to engineer an efficient communication system over wireless networks. Therefore the last part of the thesis investigates the fundamental problem of packet scheduling with a required quality of service constraint.

In this thesis, one of our goals has been to evaluate the practical issues and constraints in mathematical modelling, analysis, and design of the communication systems. In particular, we have considered three practical constraints:

(i) Capacity analysis with the assumption of ‘cheap’ nodes in the network. A ‘cheap’ node defined as a node in the network which cannot receive information while it transmits information.

(ii) The capacity of the fading channel subject to a peak power constraint.

(iii) The effect of finite delay in communication between two points from the scheduling perspective.

(iv) A finite constellation size (alphabet size at the transmitter)
1.5 Organization of the Thesis

In this section we briefly explain the organization of the thesis and summarize the major contributions we have made.

(a) In Chapter 2, we consider a multi-state network and derive the upper bound on the achievable rate of any arbitrary network coding scheme. This bound is an extension of the known max-flow min cut-set bound for a network without state. It will be shown in Chapter 5 that the new bound is not only applicable to more general scenarios but it also in fact provides a tight bound for the capacity of a relay channel under specific conditions which generates a new capacity result for the general relay channel. This bound can be applied to any multi-state network, whether the multiple state in the network is caused by channel variations (e.g., fading) or by the half-duplex function of the network nodes.

In the second part of Chapter 2, we present a general coding theorem for a network with arbitrary topology. Although for some network topology this coding theorem is not as strong as a coding strategy which is carefully tailored for the specific scenario, our theorem is relatively easy to evaluate and in many cases provides a realistic and good suboptimal strategy.

(b) In Chapter 3, we consider a general relay channel (single state) first introduced by Van der Mullen [6]. The relay channel capacity has been an open problem since
then. We provide a very simple coding strategy for the Gaussian relay channel for which its achievable rates exceed all the achievable rates of all previously known schemes under specific channel conditions. In particular, it exceeds the achievable rate of all the coding schemes of Cover and El Gamal [7], which were the best known results since 1979.

(c) In Chapter 4, we look back at the problem of the relay channel with a half-duplex relay. We provide a coding strategy that proves to be optimal (capacity achieving) for some channel conditions.

(d) In Chapter 5, we show that the coding theorem developed in the preceding chapter for a cheap relay channel is even capable of solving some new cases of the general (non-cheap) relay channel for which the capacity had not been derived earlier. We present some cases of the general relay channel problem for which the capacity is derived based on two major improvements: (i) the achievable rate of the coding theorem developed in Chapter 4 exceeds the achievable rate of the Cover and El Gamal [7] coding theorems and (ii) the new cut-set bound of Chapter 2 provides a smaller bound (and in fact tighter bound) than the previously known cut-set bound.

(e) In Chapter 6, we derive the capacity of the fading relay channel where the relay node is positioned in a relatively large (and practically important) “apple”-shaped area around the source. This derivation, based on an alternative coding
technique to Cover and El Gamal’s coding strategy, shows the same achievable rate without the knowledge of the fading states at the transmitters (source and the relay nodes).

(f) The capacity analysis of the fading channel with peak and average power constraints is derived in Chapter 7. We use the derived multi-state cut-set bound of Chapter 2 to derive the capacity of this channel. This example shows another aspect of the usability of the multi-state cut-set bound, for which the multiple state for the channel is caused by the channel variation.

(g) In Chapter 8, we turn to the problem of joint source and channel coding, which has a strong impact on the current treatment of the sensor networks. Most of the research conducted on the distributed source coding is aimed at the sensor network. However, the existence of the wireless inter-link between the nodes is ignored for the purpose of source coding. The result of this chapter shows not only that the nodes should cooperate for the purpose of channel coding in the sensor networks, but also that they should have a ‘dialogue’ for the purpose of joint source-channel coding.

(h) Chapter 9 is dedicated to a very important problem of a practical code design for the relay channel. Because of the challenging nature of the relaying function, there has been almost no good codes for the relay channel. We present a simple and very effective code design for the relay channel that can achieve a rate very
close to the capacity of the channel.

(i) Lastly, in Chapter 10, we consider a cross-layer design problem. We present the necessary and sufficient condition for a scheduler that meets a strict delay constraint.
Chapter 2
Multi-State Networks

In this chapter, we consider a general multi-terminal communication network with finite number of states. First, we derive upper bounds on the information flow across the network. The first upper bound is a single cut bound which applies to any arbitrary cut of the network and provides a bound on the sum of information flow from one side of the cut to the other side. The second bound is more general and is obtained through repeated application of the single cut bound over multiple cuts. This bound is particularly useful to derive an upper bound for a single flow between two points of the network positioned multiple hops away from each other. Later in this chapter, we present a general approach to derive an achievable rate for a single flow across the network with multiple states. Although this approach does not provide an optimal coding scheme, it is still very useful to obtain a sensible simple lower bound on the capacity. It should be pointed out that in many cases it is hard to derive a better lower bound. An example of such an improved lower bound is explained later in Chapter 4 through the use of a sophisticated coding scheme.
2.1 General Cut-set Bounds for the Multi-state Networks

In this section, we derive new cut-set bounds on the achievable rates in a general multi-terminal network with a finite number of states. Multiple states are commonly encountered in communication networks, in the form of multiple channels and nodes’ states. Our results are broadly applicable and provide much tighter upper bounds than the known single state min-cut max-flow theorem, and hence our results form an important new tool to bound the performance of multi-node networks.

Two diverse examples presented here to illustrate the tightness and the utility of the new bounds. In each of the example applications, the known single-state max-flow min-cut theorem provides a bound strictly looser than the new cut-set bounds. The first illustrative example is single-user compound channels, in which both the transmitter and receiver have channel state information. The example of compound channel represents the smallest possible network with only two nodes, but this channel still has multiple states due to channel variations. The upper bound derived using the proposed bounds turns out to be the capacity of the compound channel, which implies that the bound is tight in this case. The second example is from a contemporary network problem. We demonstrate the application of new bounds to characterize the limits on the rate of information transfer in ‘cheap’ relay networks, in which the relay nodes can either transmit or receive but not do both simultaneously. In this case, each constituent channel has a single state, but relay nodes can be in one of two states, either transmit or receive mode, giving rise to multiple network states. Here,
again, the upper bound coincides with the capacity of the channel if the relay channel is degraded.

2.1.1 Introduction

Network information theory studies the problems related to reliable communication between subsets of nodes in a given network of senders and receivers. Finding limits on the reliable communication in a multi-terminal network often proves to be very challenging, primarily due to the simultaneous existence of interference, cooperation, and feedback between the nodes. The challenge facing network information theory is evident from the fact that the exact capacity is known for only a handful of problems. Hence, we often rely on lower and upper bounds to understand the limits of communication efficiency. One of the most powerful and versatile bounds is known as min-cut max-flow theorem [8]. Interestingly, in most cases where the capacity of a network is known, the min-cut max-flow theorem [8] coincides with the achievable rates, making it both a useful bound and a method to prove converses to many capacity theorems.

Knowing the capacity of communication systems and the general network of communication nodes is crucial to understand the capabilities of such systems and is the first step in engineering the network to achieve these capabilities. The well-known min-cut max-flow theorem unifies the proof of converse to the coding theorem for many known networks. Recently, there has been increased attention paid to the difficult problem of network coding. It is commonly believed that the problem of finding
the capacity of an arbitrary network will remain unsolved. The major reasons for current pessimism can be attributed to limited success in the last fifty years. In light of the above challenges, the general upper bound in [8] has received more attention as researchers strive to understand network capacity limits. For example, cut-set bounds for single-state networks were used in [24] to study the asymptotic behavior in large-scale networks when there is only one active flow. The cut-set bounds were also used in [25] to study a variation of the multiple relay problem. However, when the min-cut max-flow theorem is applied directly to networks with multiple states, the resulting bound is loose to the extent that it is not very useful for either understanding the limits of the network or engineering the nodes.

In this section, we derive cut-set theorems for multi-terminal networks with possible multiple states. The multi-state min-cut max-flow theorem reduces to the well-known min-cut max-flow theorem [8] when it applies to single-state networks. It has been shown that if the network has multiple states, the new theorem provides a tighter upper bound than the bound in [8].

A network can have multiple states for many possible reasons. Commonly studied possibilities include channels and sources with multiple states, of which we discuss the example of a compound channel capacity in a single-user system. Our result also applies to the cases where the transmission and reception of nodes is restricted because of the hardware limitations of each node. As an example, we establish the capacity of a ‘cheap’ relay channel [26, 27, 28, 29], in which the relay node can either
transmit or receive but not do both simultaneously.

The results in this section are derived for the case of discrete memoryless networks with a finite number of states. Although the assumption of a memoryless network is fundamental to derive new bounds, the assumptions of discrete alphabet and a finite number of states can be relaxed in many cases of interest. For example, the second application of a ‘cheap’ relay channel was extended to Gaussian channels by the authors in [28], which is a continuous channel. An example extending to an infinite number of states are the bounds on the capacity of fading channels derived in [30, 31].

The rest of this section is organized as follows: We formally define the problem and notations in Section 2.1.2. The main result which is the cut-set bounds for the multi-state network, is derived in Section 2.1.3. In Section 2.1.4, we show some of the applications of the derived bounds. Finally, we conclude in Section 2.1.5.

### 2.1.2 Problem Formulation

We consider a general multi-terminal network of senders and receivers, shown in Figure 2.1, in which the network has a finite number of states. The network can be considered as a directed graph where each node of the network represents a potential sender or receiver, and each link represents the existence of a one-way channel between two nodes. We define a network channel to denote both the full network and a subset of nodes in the network, and the distinguish it from a channel which exists between
any two nodes. Alternately, the network channel can be understood as a collection of individual channels between the nodes in a network. In accordance with the term channel use in a single channel, we define network use which corresponds to the simultaneous one-time use of the network channel in a given state.

Assume that there are $N$ nodes and for every node $i$, the input random variable $X^{(i)}$ is the transmitted signal, and the output random variable $Y^{(i)}$ is the received signal. We assume that all intended messages $W^{(ij)}$ from node $i$ to node $j$, which are going to be transmitted in $n$ network uses, are independent and uniformly distributed over their respective ranges $\{1, 2, 3, ..., 2^{nR^{(ij)}}\}$, where $R^{(ij)}$ represents the rate in which node $i$ transmits information to node $j$.

Considering a network with a finite number of states and following the notation
of [32], the channel is represented by the collection of channel transition functions or channel probability functions (c.p.f.) defined* as: \( p(y^{(1)}, y^{(2)}, \ldots, y^{(N)}| x^{(1)}, x^{(2)}, \ldots, x^{(N)}| m) \) where \( m \) is the state of the network which takes its values from a set of possible states, \( \mathcal{M} \), with finite cardinality \( M = |\mathcal{M}| \).

The input symbol \( X_{k}^{(i)} \), which is the transmitted signal from node \( i \) in time \( k \) (or \( k^{th} \) network use), depends on \( W^{(ij)} \) \( \forall j \in \{1, 2, \ldots, N\} \) and also on the past values of the received signal \( Y^{(i)} \), i.e., \( Y_{1}^{(i)}, Y_{2}^{(i)}, \ldots, Y_{k-1}^{(i)} \). Also, node \( j \) decodes the intended message \( W^{(ij)} \) for this node from node \( i \), \( \forall i \in \{1, 2, \ldots, N\} \) after the transmission of the whole block of length \( n \). Therefore, decoding this message is based on all the received signals \( Y_{1}^{(j)}, Y_{2}^{(j)}, \ldots, Y_{n}^{(j)} \) and the transmitted message from this node, i.e., \( W^{(jl)} \) \( \forall l \in \{1, 2, \ldots, N\} \). Thus, the encoding and decoding functions of block length \( n \) code for node \( i \) have the following structure:

**Encoder:** \( X_{k}^{(i)}(W_{1}^{(ii)}, W_{2}^{(ii)}, \ldots, W_{N}^{(ii)}, Y_{1}^{(i)}, Y_{2}^{(i)}, \ldots, Y_{k-1}^{(i)}) \) for any node \( i \in \{1, 2, \ldots, n\} \) and any network use \( k \in \{1, 2, \ldots, N\} \).

**Decoder:** \( \hat{W}^{(ij)}(Y_{1}^{(j)}, Y_{2}^{(j)}, \ldots, Y_{n}^{(j)}, W_{1}^{(j)}, W_{2}^{(j)}, \ldots, W_{N}^{(j)}) \) for all values of \( i, j \in \{1, 2, \ldots, N\} \) which is an estimate of the receiver of node \( i \) at node \( j \) based on the received signals at node \( j \) for the whole block of \( n \) transmissions and its own transmitted information for the other nodes.

*The notation \( p(\cdot|\cdot|\cdot) \) is unconventional, but it was used in [32] to represent channel probability functions as a function of the channel state.*
The probability of error for each decoder is defined as: \( P_{e}^{(n)(ij)} = \text{Prob}(\hat{W}^{(ij)} \neq W^{(ij)}) \). The definition is based on the assumption that the messages are independent and uniformly distributed over their respective ranges.

A set of rates \( \{R^{(ij)}\} \) is said to be achievable if there exist an encoding and decoding function with block length \( n \) such that \( P_{e}^{(n)(ij)} \to 0 \) when \( n \to \infty \) for all \( j, i \in \{1, 2, \ldots, N\} \).

We denote the state of the channel in the \( k^{th} \) network use as \( m_k \). For any state \( m \) define \( n_m(k) \) as the number of times that the network is used in state \( m \) in the first \( k \) network uses. Let

\[
t_m = \lim_{k \to \infty} \frac{n_m(k)}{k}
\]

(2.1)

define the portion of the time that the network has been used in state \( m \) as the total number of network uses goes to infinity.

### 2.1.3 Multi-State Cut-Set Bounds

In this section, we first derive a bound on the sum rate of the information transfer across one cut-set in the multi-state network (Section 2.1.3.1). This forms a basis for our main result: the max-flow min-cut theorem for the multi-state network (Section 2.1.3.2). The results of this section are derived assuming that the inputs and outputs are discrete and that the network channel is memoryless. Though the memoryless property of the channel is vital in our derivation, it is relatively easy to extend the same results for continuous input alphabets [28]. It should be noted that in the
continuous alphabet case, typically additional constraints on the input signals (such as limited average power, or peak power) are needed in order to have finite capacity, which requires suitable modifications to subsequent analysis.

2.1.3.1 Single Cut Bound

We partition the given network into two subsets, set $S \subset \{1,2,\ldots,N\}$, and its complement, $S^c$. We will derive two bounds on the rate of information flow from senders in Set $S$ to receivers in Set $S^c$. When the sequence of the network states $(m_k)_{k=1}^\infty = (m_1,m_2,m_3,\ldots)$ are known a priori, Lemma 2.1.1 provides the first upper bound on the achievable rates of transmission across a given cut-set. And in the second bound (Theorem 2.1.2), we consider the case where the sequence of the network states can be chosen arbitrarily.

**Lemma 2.1.1** (Single Cut Bound). Consider a general network with finite number of states, $M$, for which the sequence $(m_k)_{k=1}^\infty$ of the network states is given and known a priori to all nodes. If the information rates $\{R^{(ij)}\}$ are achievable then there exists some joint probability distributions $p(x^{(1)},x^{(2)},\ldots,x^{(N)}|m)$, $m = 1,2,\ldots,M$, such that for any $S \subset \{1,2,\ldots,N\}$,

$$
\sum_{i \in S, j \in S^c} R^{(ij)} \leq \sum_{m=1}^M t_m I(X^{S}_{(m)};Y^{S^c}_{(m)}|X^{S^c}_{(m)})
$$

(2.2)

where $S^c$ denotes the complement of the set $S$, $I(X^{S}_{(m)};Y^{S^c}_{(m)}|X^{S^c}_{(m)})$ is the conditional mutual information between $X^{S}_{(m)}$ and $Y^{S^c}_{(m)}$ given $X^{S^c}_{(m)}$, and $\{t_m\}_{m=1}^M$ represents the...
asymptotic proportions of the states.

**Proof:** Proof of Lemma 2.1.1: Let \( T = \{(i, j) : i \in S, j \in S^c\} \) be the set of all links that cross from node set \( S \) to node set \( S^c \), and let \( T^c \) be all the other links in the network. Since \( W^{(ij)} \) are independent and uniformly distributed over their range and based on the definitions and conditions of Section 2.1.2, we have

\[
\sum_{i \in S, j \in S^c} R^{(ij)}
= \sum_{i \in S, j \in S^c} H(W^{(ij)}) \tag{2.3}
= H(W^T) \tag{2.4}
= H(W^T | W^{T^c}) \tag{2.5}
= I(W^T; Y_1^{(S^c)}, Y_2^{(S^c)}, \ldots, Y_N^{(S^c)} | W^{(T^c)})
+ H(W^T | Y_1^{(S^c)}, Y_2^{(S^c)}, \ldots, Y_N^{(S^c)}, W^{(T^c)}) \tag{2.6}
\leq I(W^T; Y_1^{(S^c)}, Y_2^{(S^c)}, \ldots, Y_N^{(S^c)} | W^{(T^c)}) + n\epsilon_n \tag{2.7}
= \sum_{k=1}^{N} I(W^T; Y_k^{(S^c)} | Y_1^{(S^c)}, Y_2^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)})
+n\epsilon_n \tag{2.8}
= \sum_{k=1}^{N} \left\{ H(Y_k^{(S^c)} | Y_1^{(S^c)}, Y_2^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)})
- H(Y_k^{(S^c)} | Y_1^{(S^c)}, Y_2^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}, W^T) \right\}
+n\epsilon_n \tag{2.9}
\]
\[
\leq \sum_{k=1}^{N} \left\{ H\left( Y^{(S)c}_k | Y^{(S)c}_1, Y^{(S)c}_2, \ldots, Y^{(S)c}_{k-1}, W^{(T)c}, W^{T}, X^{(S)c}_k \right) - H\left( Y^{(S)c}_k | Y^{(S)c}_1, Y^{(S)c}_2, \ldots, Y^{(S)c}_{k-1}, \right. \right. \\
+ n\epsilon_n \right. \right. \} + n\epsilon_n \quad (2.10)
\]

\[
\leq \sum_{k=1}^{N} \left\{ H\left( Y^{(S)c}_k | X^{(S)c}_k \right) - H\left( Y^{(S)c}_k | X^{(S)c}_k, X^{(S)}_k \right) \right. \right. \\
+ n\epsilon_n \quad (2.11)
\]

\[
= \sum_{k=1}^{N} I\left( X^{(S)c}_k ; Y^{(S)c}_k \mid X^{(S)c}_k \right) + n\epsilon_n \quad (2.12)
\]

\[
= \sum_{k=1}^{N} I\left( X^{(S)c}_{Q_m(k)} ; Y^{(S)c}_{Q_m(k)} \mid X^{(S)c}_{Q_m(k)}, Q_m(k) = k \right) \\
+ n\epsilon_n \quad (2.13)
\]

\[
= \sum_{m=1}^{M} n_m(n) I\left( X^{(S)c}_{Q_m} ; Y^{(S)c}_{Q_m} \mid X^{(S)c}_{Q_m}, Q_m \right) + n\epsilon_n \quad (2.14)
\]

\[
= \sum_{m=1}^{M} n_m(n) \left\{ H\left( Y^{(S)c}_{Q_m} \mid X^{(S)c}_{Q_m}, Q_m \right) \\
- H\left( Y^{(S)c}_{Q_m} \mid X^{(S)c}_{Q_m}, X^{(S)c}_{Q_m}, Q_m \right) \right. \right. \\
+ n\epsilon_n \right. \right. \} + n\epsilon_n \quad (2.15)
\]

\[
\leq \sum_{m=1}^{M} n_m(n) \left\{ H\left( Y^{(S)c}_{Q_m} \mid X^{(S)c}_{Q_m} \right) \\
- H\left( Y^{(S)c}_{Q_m} \mid X^{(S)c}_{Q_m}, X^{(S)c}_{Q_m} \right) \right. \right. \\
+ n\epsilon_n \right. \right. \} + n\epsilon_n \quad (2.16)
\]

\[
= \sum_{m=1}^{M} n_m(n) I\left( X^{(S)c}_{Q_m} ; Y^{(S)c}_{Q_m} \mid X^{(S)c}_{Q_m} \right) + n\epsilon_n \quad (2.17)
\]

where

(2.3) follows from the assumption that \( W^{(ij)} \)s are distributed uniformly over their respective ranges \( \{ 1, 2, 3, \ldots, 2^{nR^{(ij)}} \} \).

(2.4) follows from the assumption that \( W^{(ij)} \)s are independent and also the definition
of \( W^{(T)} = \{ W^{(ij)} : i \in S, j \in S^c \} \),

(2.5) follows from the independence of \( W^{(T)}, W^{(T^c)} \),

(2.6) follows from the definition of the mutual information,

(2.7) follows from the Fano’s inequality, because message \( W^{(T)} \) can be decoded from \( Y^{(S)} \) and \( W^{(T^c)} \). (Since we have assumed that the set of rates \( \{ R^{(ij)} \} \) are achievable; and also note that \( \epsilon_n \rightarrow 0 \) as \( n \rightarrow \infty \)),

(2.8) follows from the chain rule,

(2.9) follows from the definition of the mutual information,

(2.10) follows from the fact that the first term has been changed due to the definition of \( X^{(S^c)}_k \), which is a function of the past received symbols \( Y^{(S^c)} \) and the message \( W^{(T^c)} \). Also the second term has not been increased because conditioning can only reduce the entropy,

(2.11) follows from the fact that \( Y^{(S^c)}_k \) only depends on the current input symbols \( X^{(S^c)}_k, X^{(S)}_k \). (Note that although \( Y^{(S^c)}_k \) depends on the state \( m_k \), it is deterministic and predefined before transmission for all the nodes),

(2.12) follows from the definition of the mutual information,

(2.13) follows from introducing time sharing random variables \( Q_1, Q_2, \ldots, Q_M \) for each state of the network, where each state \( m, Q_m \) is uniformly distributed over all values of the time index \( k, k \in \{1, 2, 3, ..., N\} \), for which \( m_k = m \),

(2.14) follows from rearranging the summation and using the definition of the average mutual information,
(2.15) follows from the definition of the mutual information,

(2.16) follows from the fact that conditioning cannot increase the entropy for the first term and the fact that the second term has not been changed because $Y_{Q_m}^{(S^c)}$ only depends on the current input symbols $X_{Q_m}^{(S^c)}, X_{Q_m}^{(S)}$ and conditionally is independent of $Q_m$,

(2.17) follows from the definition of the mutual information.

Thus, by dividing both sides of inequality (2.17) by $n$ and finding the limit as $n \to \infty$ we have

$$\sum_{i \in S, j \in S^{(c)}} R^{(ij)} \leq \sum_{m=1}^{M} \frac{n_m(n)}{n} I\left(X_{Q_m}^{(S)}, Y_{Q_m}^{(S^c)} \mid X_{Q_m}^{(S^c)}\right) + \epsilon_n.$$ 

As $n \to \infty$ we have:

$$\sum_{i \in S, j \in S^{(c)}} R^{(ij)} \leq \sum_{m=1}^{M} t_m I\left(X_{Q_m}^{(S)}, Y_{Q_m}^{(S^c)} \mid X_{Q_m}^{(S^c)}\right)$$

and it completes the proof.  

It should be noted that this bound is not a function of the choice of the sequence of the network states, $(m_k)_{k=1}^{\infty}$, directly, rather it depends only on the asymptotic values $t_1, t_2, \ldots, t_M$. In fact this bound holds for the best choice of the deterministic function† $m_k$, as well as the case where it changes randomly from one network use

†Section 2.1.4 shows an interesting case of a deterministic function where this new cut-set bound is tight.
to the other while the asymptotic proportions of the sequence of the states, i.e., $t_i, i \in \{1, 2, \ldots, M\}$, remain fixed. Thus, no information can be embedded in the order of states $m_k$ if their asymptotic proportions are held constant.

If the sequence of the network states can be chosen arbitrarily, the bound across each cut-set would be the supremum of the achievable bounds in Lemma 2.1.1 for all choices of the function $m_k$. On the other hand, since the bound is not a function of the individual sequence of the network states, the supremum would be taken over the possible values of $t_i, i \in \{1, 2, \ldots, M\}$ and specifically we have the following theorem.

**Theorem 2.1.2.** Consider a general network with a finite number of states, $M$. Maximum achievable information rates $\{R^{(i,j)}\}$ across the cut-set $S \subset \{1, 2, \ldots, N\}$ for any choice of network state sequence $(m_k)_{k=1}^\infty$ is bounded by:

$$\sum_{i \in S, j \in S^c} R^{(i,j)} \leq \sup_{t_m} \sum_{m=1}^M t_m I(X_{(m)}^S; Y_{(m)}^{S^c}|X_{(m)}^{S^c}) = \sup_m I(X_{(m)}^S; Y_{(m)}^{S^c}|X_{(m)}^{S^c})$$ (2.18)

for some joint probability distributions $p(x^{(1)}, x^{(2)}, \ldots, x^{(N)}|m), m = 1, 2, \ldots, M$, where $\sum_{i=1}^M t_m = 1$.

Thus, in order to maximize the sum rate of the information transfer across a cut-set, the above result suggests that using the network just in one fixed state $m$ allows for the maximum mutual information $I(X_{(m)}^S; Y_{(m)}^{S^c}|X_{(m)}^{S^c})$. Although Theorem 2.1.2 is
a bound on actual capacity, intuitively it appears that the actual capacity will exhibit
similar behavior using only the most beneficial state of the network. However, in the
case that the sequence of states cannot be chosen, the average mutual information is
the bound given by Lemma 2.1.1. In Section 2.1.4, we show the connection of the
above bound to the results on compound channel capacity.

2.1.3.2 Max-Flow Min-Cut Bound

In many cases, the aim is to maximize the information rate between two disjoint sub-
sets of nodes in the network regardless of any possible simultaneous communication
rates between other nodes in the network. For example, this situation happens in the
relay problem when a source node (or set of source nodes) transmits information to a
destination node (or set of destination nodes) by the help of some intermediate relay
node (or relay nodes). In this case, the aim is to maximize the rate of information
transfer from source to destination. Therefore, instead of a bound on a single cut-set
in the network, we would like to find an upper bound for the rate of information
transfer between two disjoint subsets of nodes $S_1, S_2 \subset \{1, 2, \ldots, N\}$ of the network.
For any cut-set that partitions the set of nodes in the network to $\{S, S^{(c)}\}$ and that
has either one of the sets $S_1$ and $S_2$ in just one side of the cut (e.g., in Figure 2.1
consider sets $S_1$ and $S_2$ for either cuts $C_1$ or $C_2$), we can use the single cut bounds
of Section 2.1.3.1. Therefore, the rate of information transfer between $S_1$ and $S_2$ is
bounded by the minimum of all such single cut bounds.
A simple observation suggests that the sum of information rates from the set $S_1$ to the set $S_2$ is bounded by the sum of the rates across all the cut-sets $S$ such that $S \subset \{1, 2, \ldots, N\}$, $S \cap S_1 = S_1$, $S \cap S_2 = \phi$. This observation is given by Theorem 2.1.2. Thus, we have:

$$\sum_{i \in S_1, j \in S_2} R_{ij}^{(i)} \leq \min_S \sup_m I(X^S_{(m)} ; Y^{S_c}_{(m)} | X^{S_c}_{(m)})$$

(2.19)

for some joint probability distribution $p(x(1), x(2), \ldots, x(N)|m)$.

The bound of (2.19) is a naive bound, because it does not account for the fact that the sequence of network states that gives the maximum achievable sum rate across a given cut-set does not necessarily lead to the maximum achievable rate across other cut-sets in the network. Therefore, since the maximum achievable rates across different cut-sets depend on the sequence of the network states, we use Lemma 2.1.1 directly to find a better upper bound for the achievable rate of the information flow between the two disjoint subsets of the nodes $\{S_1, S_2\}$.

**Lemma 2.1.3.** Consider a general network with a finite number of states, $M$, for which the sequence $(m_k)_{k=1}^{\infty}$ of the states of the network is given and known to all nodes. If the information rates $\{R_{ij}^{(i)}\}$ are achievable, then the sum rate of information transfer from a node set $S_1$ to a disjoint node set $S_2$ where $S_1, S_2 \subset \{1, 2, \ldots, N\}$, is bounded by:
\[
\sum_{i \in S_1, j \in S_2} R^{(ij)} \leq \min_S \sum_{m=1}^{M} t_m I(X_{(m)}^S; Y_{(m)}^{S^c}|X_{(m)}^{S^c}),
\] (2.20)

for some joint probability distributions \( p(x^{(1)}, x^{(2)}, \ldots, x^{(N)}|m), m = 1, 2, \ldots, M \), where the minimization is taken over all set \( S \subset \{1, 2, \ldots, N\} \) subject to \( S \cap S_1 = S_1, S \cap S_2 = \phi \).

**Proof:** Direct application of the Lemma 2.1.1 over multiple cut-sets, and considering the fact that the sequence of network states is given, results in (2.20). ■

Again, considering all possible sequences of network states, the bound of the achievable rate of information transfer from a node set \( S_1 \) to a disjoint node set \( S_2 \), where \( S_1, S_2 \subset \{1, 2, \ldots, N\} \), is given by the supremum of the achievable bounds in Lemma 2.1.3 for all choices of network state sequence function \( m_k \). It again translates to taking supremum over all possible asymptotic values, \( t_i, i \in \{1, 2, \ldots, M\} \).

**Theorem 2.1.4.** Consider a general network with a finite number of states, \( M \). If the information rates \( \{R^{(ij)}\} \) are achievable, then the sum rate of information transfer from a node set \( S_1 \) to a disjoint node set \( S_2 \), where \( S_1, S_2 \subset \{1, 2, \ldots, N\} \) and for any choice of network state sequence \( (m_k)_{k=1}^{\infty} \), is bounded by:

\[
\sum_{i \in S_1, j \in S_2} R^{(ij)} \leq \sup_{S} \min_{t_m} \sum_{m=1}^{M} t_m I(X_{(m)}^S; Y_{(m)}^{S^c}|X_{(m)}^{S^c})
\] (2.21)
for some joint probability distributions $p(x^{(1)}, x^{(2)}, \ldots, x^{(N)}|m)$, $m = 1, 2, \ldots, M$, when the minimization is taken over all set $S \subset \{1, 2, \ldots, N\}$ subject to $S \cap S_1 = S_1$, $S \cap S_2 = \phi$ and the supremum is over all the non-negative $t_m$ subject to $\sum_{i=1}^{M} t_m = 1$.

In the next section we will consider some specific examples of the networks and applications of this bound on the achievable rates of communication in such networks.

### 2.1.4 Applications of the New Bounds

When there is only one state of operation in the network, i.e., $(M = 1)$, the bound of Lemma 2.1.1 reduces to the known cut-set theorem in network information theory [Theorem 14.10.1 in [8]], and Theorem 2.1.4 can be interpreted as max-flow min-cut theorem for a single state network with $(M = 1, S_1 = S, S_2 = S^{(c)})$. Although these cut-set bounds are not tight in general, they are tight enough to find the capacity region or capacity in some cases such as multiple access channel, degraded relay channel, or general relay channel with feedback [7].

In the sequel, we will discuss two applications of the new multi-state cut-set bounds. First, we will consider the application of Lemma 2.1.1 to derive an upper bound on the capacity of compound channels (Section 2.1.4.1). For this example, the network is trivial and has only one channel, i.e., between the sender and receiver nodes. Multiple states in the network result from changes in channel probability function from one channel use to another. In this case, the upper bound from Lemma 2.1.1 turns out to be the capacity of compound channels.
The second application is an upper bound for the capacity of ‘cheap’ relay channels (Section 2.1.4.2), for which previously known cut-set bound does not provide a tight bound. However, the derived bounds in this section not only provide a better bound but they also provide a tight bound for the case of the degraded ‘cheap’ relay channel.

2.1.4.1 Compound Channel

Channels with a finite number of states were considered first in the context of compound channels [32] for which the capacity for discrete memoryless channels with different cases of channel state information at the receiver and the transmitter have been derived. Extension of the results in [32] for the fading channel, which is a multi-state channel with an infinite number of states, can also be found in [30, 31]. The derived cut-set bounds for the multi-state networks unify some of the converses for the known multi-state channels. For example, Lemma 2.1.1 provides a tight bound for the capacity of the compound channel with the known channel state information both at the transmitter and receiver (Theorem 4.6.1 in [32]).

A compound channel is defined as a channel for which the channel probability function governs the transmission of signal changes with time, but it does so independently of all previously transmitted and received signals, with a known probability distribution $P_{\text{state}}(m)$. Lemma 2.1.1 provides a very simple upper bound for the capacity of the compound channel, $C_{\text{compound}}$, where noncausal channel state information
is available to both transmitter and receiver. We have

$$C_{\text{compound}} \leq \sum_{m=1}^{M} t_m I(X(m); Y(m)) = \sum_{m=1}^{M} t_m C(m) \tag{2.22}$$

where $C(m)$ represents the channel capacity of a uni-state channel in state $m$. In fact, it has been shown that this upper bound is the exact capacity of the compound channel (Theorem 4.6.1 in [32]), thus the bound of Lemma 2.1.1 is tight in this case.

It can also be shown directly that the average mutual information bound of Lemma 2.1.1 is an upper bound on the capacity of the channel with noncausal side information at the transmitter introduced in [33], although in this case the bound is loose in general. We have

$$\max_{p(u,x|v)} \{I(U; Y) - I(U; V)\} \leq \sum_{m=1}^{M} t_m I(X(m); Y(m)) \tag{2.23}$$

where $V$ represents the channel state which is available to the transmitter for the whole block of transmission a priori, and $U$ is an auxiliary random variable with a finite alphabet size.

Two notes are in order here. First, the bounds in this section are easily extended for channels with continuous input and output alphabets. For example, Lemma 2.1.1 can also serve to find a bound on the result known as “writing on dirty paper”, which can be considered as an extended version of the channel with noncausal side information at the transmitter [33] with a continuous input alphabet.

Second, although the bounds in this section are obtained with the assumption of a finite number of states in the network, they can also be applied to some channels with
an infinite number of states. For example, Lemma 2.1.1 provides a tight upper bound in (i) the fading channel with side information and average power constraints [30], and (ii) the fading channel with side information and both peak and average power constraints [31]. The fading channel with side information [30, 31] can be considered as an extended version of the compound channel [32] with a continuous alphabet and an infinite number of states.

2.1.4.2 ‘Cheap’ Relay Channel

In chapter 4, the problem of a ‘cheap’ relay channel will be discussed in detail, where we develop the coding strategies for this channel. However, here we briefly discuss the application of the multi-state cut-set theorem to derive an upper bound on the capacity of this channel.

A cheap node is defined as a node which is either a sender or receiver for any channel use. Each state of operation in a network with cheap nodes corresponds to the valid partitioning of the nodes into two disjoint subsets of sender and receiver nodes. It can be deduced that there are a finite number of states for a network with a finite number of nodes.

The ‘cheap’ relay channel, shown in Figure 2.3, is a variation of the relay channel where the relay node functions as either a transmitter or a receiver for any given channel use (or equivalently any given time). The ‘cheap’ relay channel is introduced in [26], and an achievable rate for the defined channel is derived.
Figure 2.2: ‘Cheap’ Relay Channel: Representation of the cut-sets. Notes: 1-The relay can either transmit or receive at any given time. 2-Both of the cut-sets exist in either “Broadcast” or “Multiple access states”.

Figure 2.3: ‘Cheap’ relay channel: The representation of two states of operation as a “broadcast state” ($m_1$) and a “multiple access state” ($m_2$)
Let \( p(y, y_1|x_1, x_2) \) denote the channel probability function of the original relay channel. Because the relay node, \( R \), functions either a transmitter or a receiver there are two possible states of operation in the network. In the state \( m_1 \) ("broadcast state" in Figure 2.3), the relay node \( R \) acts as a receiver and thus the channel probability function is given by \( p(y_1|x_1|m_1) = p(y, y_1|x_1, x_2 = 0) \). While in the state \( m_2 \) ("Multiple access state" in Figure 2.3), the relay node functions as a transmitter and the channel probability function is given by \( p(y|x_1, x_2|m_2) = \sum_{y_1} p(y, y_1|x_1, x_2) \). We use \( x_2 = 0 \) to denote the case that there is no input from relay node \( R \) to the channel in state \( m_1 \). It should be pointed out that we use the terms "broadcast state" and "multiple access state" because of the resemblance of the network in these states to broadcast and multiple access channels.

However, to achieve the full capacity of the channel, the coding scheme cannot be divided into two separate coding schemes, one for each individual state. For example, consider the case that the received signal at the destination is a degraded version of the received signal at the relay in broadcast state. In this situation, the destination cannot decode the whole message sent in this state, but it is decodable at the relay node. This means that there would be an ambiguity about the whole or at least part of the transmitted message from the source, which would remain to be resolved by the next transmission in the multiple access state. In this case, resolving the ambiguity is easier than retransmitting the whole message in the multiple access state. In other words, although the received signal at the destination node in the broadcast state
is not enough to decode the whole transmitted message, it would provide partial information that would be resolved later in multiple access state.

From Theorem 2.1.4, an upper bound for the information transfer rate $R$ from source node $S$ to the destination node $D$ would be

$$R \leq \sup_{t, 0 \leq t \leq 1} \min \{ t I(X_1; Y, Y_1|m_1) + (1 - t) I(X_1; Y|X_2, m_2), t I(X_1; Y|m_1) + (1 - t) I(X_1, X_2; Y|m_2) \},$$

(2.24)

where a minimum of two terms is involved: the first term corresponds to the bound on broadcast cut-set $C_1$ and the second term corresponds to the multiple access cut-set $C_2$ in Figure 2.2. Either of the above cut-set bounds over $C_1$ and $C_2$ have two conditional mutual information terms. The first conditional mutual information term corresponds to the broadcast state $m_1$, and the second conditional mutual information corresponds to the multiple access state.

On the other hand, it has been shown [26] that for every $t$, $0 \leq t \leq 1$, the rate $R^*$ is achievable where $R^*$ is given by

$$R^* \triangleq \min \{ t I(X_1; Y_1|m_1) + (1 - t) I(X_1; Y|X_2, m_2), t I(X_1; Y|m_1) + (1 - t) I(X_1, X_2; Y|m_2) \}.$$  

(2.25)
Thus, if in the state $m_1$ the received signal $y$ at the destination node $D$ is a degraded form of the received signal $y_1$ at the relay node, then the bound of (2.24) would coincide with this achievable rate for some value of $t$. Hence, the bound derived in Theorem 2.1.4 provides the converse for the capacity theorem of the degraded cheap relay channel, and the capacity is given by $C = R^*$ defined in (2.25).

It should be pointed out that the achievable rate of information transmission in the cheap relay channel is lower than that of the relay channel without the extra condition on the relay node (Figure 2.4). Thus, the known cut-set theorem of network information theory is unable to provide tight enough bounds. However, the derived bounds in this section prove to be effective in this case, and they capture the effect of the practical constraint imposed on the network nodes to the extent that we can actually derive the capacity of the degraded cheap relay channel.

Although having more than one state of operation in the network seems to increase degrees of freedom and thus the achievable rates in the network, it is not always true. For example, in the problem of communication in a network with cheap nodes [26], having more than one state is a result of imposing the condition that each node cannot transmit and receive at the same time. It is somehow trivial that posing this condition will not increase any sets of achievable rates in the network in comparison to the same network without this constraint (which has just one state of operation but all of the links or channels between the nodes can be used at the same time).
Figure 2.4: Inner and outer bounds (in bits/channel use) on the capacity of the cheap relay channel of Figure 2.2 versus cross over probability of the direct link, $p$, where each of the links $S \rightarrow D$, $S \rightarrow R$, and $R \rightarrow D$ are binary symmetric channels with cross over probabilities $p$, $q$, and $r$ respectively. The $q$ and $r$ take values from the interval $[0, \frac{1}{4}]$ as a function of $p$, and are symmetric with respect to $p = \frac{1}{2}$, i.e., $q(p) = q(1 - p)$ and $r(p) = r(1 - p)$. 

"Cheap relay upper bound from Theorem 2"

"Cover’s upper bound for expensive relay"

"Direct link capacity"

"Cheap relay (Achievable rate)"

"Expensive relay (Cover’s achievable Rate)"

Figure 2.4 : Inner and outer bounds (in bits/channel use) on the capacity of the cheap relay channel of Figure 2.2 versus cross over probability of the direct link, $p$, where each of the links $S \rightarrow D$, $S \rightarrow R$, and $R \rightarrow D$ are binary symmetric channels with cross over probabilities $p$, $q$, and $r$ respectively. The $q$ and $r$ take values from the interval $[0, \frac{1}{4}]$ as a function of $p$, and are symmetric with respect to $p = \frac{1}{2}$, i.e., $q(p) = q(1 - p)$ and $r(p) = r(1 - p)$. 

"Cheap relay upper bound from Theorem 2"

"Cover’s upper bound for expensive relay"

"Direct link capacity"

"Cheap relay (Achievable rate)"

"Expensive relay (Cover’s achievable Rate)"
2.1.5 Concluding Remarks

Considering the fact that cut-set bounds are tight in many important cases (multiple access, relay channels, etc.), they have received relatively little attention in the literature. The reason may partly lie in the fact that point-to-point (two node network) analysis received far greater attention than other problems, especially those in network information theory. We envision that the situation will gradually change as the focus of the community shifts to analyzing capacity of ad hoc, sensor and distributed wireless networks.

We feel that the new bounds in this section is only a step towards capturing the complexities of actual networks. Two big challenges are as follows. First, the cut-set bounds inherently assume a collaborative reception by the nodes in the receiving subset (e.g., set $S_2$ in Figure 2.1). This is precisely the reason why min-cut max-flow cut-set theorem does not yield the capacity of broadcast channels, where the receiver decodes their signals with no collaborative communication with other receivers.

Second, a more serious deficiency of known cut-set bounds is that they provide no useful bound if there are loops in the network; a classic example of this situation is the two-way channel. Feedback via shared bandwidth is an important aspect of all networks, and information theoretic methods have had limited success in incorporating in its analysis [34].
2.2 General Achievable rate for the Multi-state Networks

In this section, we derive achievable rates for arbitrary network topologies consisting of ‘cheap’ nodes. A node is labeled ‘cheap’ if its radio can only operate in TDD mode when transmitting and receiving in the same frequency band. Two main results are shown here. The first result provides an achievable rate for the channel with either a continuous alphabet or discrete channel. The second result provides the achievable rate for the Gaussian channel with average power constraint. The two results are applied to the case of the Gaussian relay channel and the concatenated channel with cheap nodes that corresponds to a multi-hop channel where each intermediate node functions as either a transmitter or a receiver.

2.2.1 Introduction

In this section, we first derive achievable rates (lower bound on the capacity) for the networks when the channel probability function is known (Theorem 2.2.1) and also for the Gaussian channel networks with a power constraint (Theorem 2.2.2). Then, we present applications of the results for two cases, the Gaussian cheap relay channel and the multi-hop network with cheap nodes.

The relay channel was first defined and studied by Van der Mullen [6] for the case of three terminals (or nodes) and later considered by Cover [7, 8]. Several authors have recently considered a more complex version of the relay channel [24, 23, 22], but all of the results assume the use of ‘expensive’ radios, which can send and receive
at the same time in the same frequency band. This assumption is the key difference between the previous works and the results presented in this section.

The rest of this section is organized as follows: In section 2.2.2, we derive achievable rate for the networks with arbitrary topologies. In section 2.2.3, we present the application of both Theorems 2.2.1 and 2.2.2 for the Gaussian cheap relay channel considering two different cases for the power constraints. In section 2.2.4 we show an application of the results in a more involved case of the multi-hop relay network with cheap nodes. Finally, we conclude with some remarks in section 2.2.5.

2.2.2 General Results

Following the notation of Section 2.1.2, we consider a general multi-terminal network of senders and receivers. The network can be considered as a directed graph where each node of the network represents a potential sender or receiver (or both) and each link represents a (one-way) channel between two nodes. In this section, we establish a lower bound on the achievable rate between any two nodes in a multi-state network with an arbitrary topology.

Theorem 2.2.1. Consider a general network with a finite number of states, $M$, for which the sequence $m_k$ of the states of the network is known a priori to all nodes. If the set of the information rates $\{R_{ij}^m, i, j \in \{1,2,\ldots,N\}\}$ from node $i$ to node $j$ is achievable in state $m$ given the previous sequence of the states, then the rate $R$ given
by

\[ R \leq \sum_{j=2}^{N} \sum_{m=1}^{M} \delta_{ij}^m R_{ij}^m \]  

(2.26)

is achievable from node 1 to node N, provided that

\[ \sum_{j=1, (j \neq i)}^{N} \sum_{m=1}^{M} t_m \delta_{ji}^m R_{ji}^m \leq \sum_{j=1, (j \neq i)}^{N} \sum_{m=1}^{M} t_m \delta_{ij}^m R_{ij}^m \]  

(2.27)

for all the nodes \( i \in \{2, 3, \ldots, N - 1\} \), where \( \delta_{ij}^m = 1 \) iff link from node \( i \) to node \( j \) is active in the state \( m \), otherwise \( \delta_{ij}^m = 0 \).

**Proof** See [29].

Theorem 2.2.1 provides a general framework to find an achievable rate between a single source-destination pair of interest in an arbitrary cheap relay network. It can be used wherever the channel transition probability of the network is fixed given the state of the network as defined earlier. Thus, it is applicable to the network with a continuous input alphabet (e.g. Gaussian) as well as the network with a discrete input alphabet as long as the channel is memoryless and the channel transition probability is fixed for each state. In other words, it is applicable in the case of the Gaussian channel network or a block fading channel network with perfect CSIT and CSIR when the average power constraint is given for each node per state. The following theorem extends the result to the case where the average power constraint is given for the whole block of transmission, which is of more interest for the Gaussian channel network.
**Theorem 2.2.2.** Consider a Gaussian channel network with an arbitrary topology and a finite number of states, \( M \), for which the sequence \( m_k \) of the states of the network is known a priori to all nodes. Suppose that the average power constraint per node is given as \( P_0, P_1, \ldots, P_{N-1} \) for the nodes \( 1, 2, \ldots, N - 1 \). If the set of the information rates \( \{ R_{ij}^m, i, j \in \{1, 2, \ldots, N\} \} \) from node \( i \) to node \( j \) is achievable in state \( m \) given the previous sequence of the states, then the following rate is achievable from node 1 to node \( N \)

\[
R \leq \sum_{j=2}^{N} \sum_{m=1}^{M} \delta_{ij}^m R_{ij}^m \tag{2.28}
\]

provided that

\[
\sum_{j=1,(j\neq i)}^{N} \sum_{m=1}^{M} t_m \delta_{ji}^m R_{ji}^m \leq \sum_{j=1,(j\neq i)}^{N} \sum_{m=1}^{M} t_m \delta_{ij}^m R_{ij}^m \tag{2.29}
\]

for all the nodes \( i \in \{2, 3, \ldots, N - 1\} \), where \( \delta_{ij}^m = 1 \) iff link from node \( i \) to node \( j \) is active in the state \( m \), and

\[
P_i = \sum_{j=1,(j\neq i)}^{N} \sum_{m=1}^{M} t_m \delta_{ij}^m P_{ij}^m \tag{2.30}
\]

for all the nodes \( i \in \{1, 2, \ldots, N - 1\} \), where \( P_{ij}^m = \sum_{j=1,(j\neq i)}^{N} P_{ij}^m \) is the average power constraint of node \( i \) in the state \( m \).

**Proof** See [29].

The maximal achievable rate of the Theorem 2.2.1 is achieved using the optimum sequence of the network states and the corresponding values of the time sharing parameters \( t_1, t_2, \ldots, t_M \). However, the maximal achievable rate in the Theorem 2.2.2
is achieved using joint optimization of the same parameters and power allocations $P_{P^m_{ij}}$ for all possible values of $i$, $j$, and $m$.

Although results of the Theorems 2.2.1 and 2.2.2 do not provide the best possible achievable rate in general and it is sometimes possible to find a better achievable rate for some specific network topology, these results provide a lower bound for the capacity of the single source-destination pair in the network. It is worth mentioning that for some special cases, the derived lower bound is the exact capacity of the network. In [26], the capacity of the cascaded single link channels (multi-hop) has been derived. Although it is not shown in this section, it can be easily verified that the achievable rate of the Theorem 2.2.1 is in fact the capacity of cascaded channels (See Section 4.1.3).

Apart from a very few specific cases, the problem of finding the capacity or even a better achievable rate is intractable, and the derived achievable rate of the Theorems 2.2.1 and 2.2.2 are the best known results. In the following sections we discuss two cases: (i) the Gaussian cheap relay channel, where a better achievable rate has been derived by means of very sophisticated coding techniques which have allowed us to fully characterize the corresponding capacity and (ii) the Multi-hop cheap relay network, where the capacity is hard to find, although Theorem 2.2.2 can give a reasonably good achievable rate in this case.

It is interesting to note that the assumption of the fluid model for the information flow in the network is not right. The linear combination of the rates as stated in
Theorems 2.2.1 and 2.2.2 may give the impression of treating the information as a fluid, but close examination of the achievable rates in each state reveals that the partial shared knowledge of the information between the nodes due to the previous transmission provides the basis of the cooperation between the nodes in the next states. In other words, the cooperation among the relay nodes results in better rates across some cut-set of the network which contradicts the fluid model for information flow. This point will become more clear in the next section where we give more specific examples.

### 2.2.3 Gaussian Cheap Relay Channel

In this section we apply the derived general lower bound to the Gaussian cheap relay channel. The cheap relay channel defined in Section 2.1.4.2 will be further investigated in Chapter 4.

We consider the Gaussian version of the cheap relay channel (Section 2.1.4.2), shown in Figure 2.5, consists of an input $x_1$, a relay output $y_1$, a relay sender $x_2$ (which depends only on the past values of $y_1$), and a channel output $y$. The channel is assumed to be memoryless. Assuming the relay node is cheap, there are two possible states of operation for the cheap relay channel. The dependency of the outputs on the inputs are as follows: In state $m_1$, relay node R acts as a receiver and thus the channel output is given by $y = h_1 x_1 + z$, and the relay output is given by $y_1 = h_0 x_1 + z_1$. In state $m_2$, the relay node functions as a transmitter and the
channel output is given by $y = h_1x_1 + h_2x_2 + z$, where $h_0$, $h_1$, and $h_2$ are channel losses and assumed to be constant, and $z \sim N(0, N)$ and $z_1 \sim N(0, N_1)$ are independent Gaussian noises. Suppose that we can arbitrarily choose between the states $m_1$ and $m_2$ in order to maximize the rate of information transfer between the source node S and destination node D. Suppose that in a total of $n$ channel uses, the channel is used $k$ times in the state $m_1$ and $n - k$ times in the state $m_2$. We assume the following three power constraints: (i) $\frac{1}{k} \sum_{i=1}^{k} x_{1i|m_1}^2 \leq P_0$, (ii) $\frac{1}{n-k} \sum_{i=1}^{n-k} x_{1i|m_2}^2 \leq P_1$, and (iii) $\frac{1}{n-k} \sum_{i=1}^{n-k} x_{2i|m_2}^2 \leq P_2$, where $x_{ji|m_k}$ is the $i$’th transmitted signal from node $j$ while the channel is used in state $m_k$. The problem is to find the capacity of the channel between the sender S and receiver D.

Theorem 2.2.1 gives the following achievable rate for the channel if the channel is degraded:
\[ R = \max_{0 \leq t, \alpha, \beta \leq 1} t(R_{11} + R_{01}) + (1 - t)R_{12} \quad (2.31) \]

where \((1 - t)R_{22} \leq tR_{01}\), and

\[
R_{11} = \frac{1}{2} \log(1 + \frac{\gamma_1(1 - \alpha)P_0}{1 + \gamma_1 \alpha P_0})
\]

\[
R_{01} = \frac{1}{2} \log(1 + \gamma_0 \alpha P_0)
\]

\[
R_{12} \leq \frac{1}{2} \log(1 + \gamma_1(1 - \beta)P_1)
\]

\[
R_{12} + R_{22} \leq \frac{1}{2} \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\sqrt{\gamma_1 \gamma_2 \beta P_1 P_2})
\]

and \(\gamma_i\)'s, \(i = 0, 1, 2\) are defined as: \(\gamma_0 = \frac{|h_0|^2}{N_1}\), \(\gamma_1 = \frac{|h_1|^2}{N}\), and \(\gamma_2 = \frac{|h_2|^2}{N}\). The rates \(R_{01}\) and \(R_{11}\) constitute the achievable rate region of the broadcast channel in state \(m_1\) and the rate \(R_{12}\) and \(R_{22}\) are the achievable rate region of the multiple access channel in state \(m_2\).

The multiple access channel has been extensively studied, and the achievable rate regions for the case of the independent sources, the correlated sources [35], and the hierarchical sources [36] have been derived. Also, the converse has been established for the independent message sources which results in the derivation of the capacity in this case. A result which may be of independent interest is that by proving the converse, we have shown that the rates \(R_{12}\) and \(R_{22}\) are the exact capacity region of the multiple access channel where the node S knows the information of the node R (hierarchical case).
It is also well known that for the Gaussian degraded channel the rates $R_{01}$ and $R_{11}$ form the exact capacity region. However, the capacity of the degraded Gaussian cheap relay channel is not given by the Equation (2.31), although the capacity region is known for both of the states. In fact, by using a more sophisticated coding technique we have shown that a higher rate than (2.31) is achievable for the Gaussian cheap relay channel [28] and are thus able to fully determine the capacity of the degraded Gaussian cheap relay channel. Using the same notation and defined variables we have

**Theorem 2.2.3.** [28] Capacity of the Gaussian degraded cheap relay channel is given by

$$C = \max_{t, \rho, 0 \leq t, \rho \leq 1} \min \{ t \log(1 + \gamma_0 P_0) + (1 - t) \log(1 + (1 - \rho^2)\gamma_1 P_1), t \log(1 + \gamma_1 P_0) + (1 - t) \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \}. \quad (2.32)$$

The reason for achieving the higher rate in Theorem 2.2.3 is that the received signal in the state $m_1$ contains some extra information which can be deduced using joint decoding of all received signals.

In the case of the total average power constraint per nodes as $E[x_1^2] \leq P_1'$ and $E[x_2^2] \leq P_2'$ we can use Theorem 2.2.2 and solve explicitly for $t$, which gives the achievable rate as:

$$R = \max_{0 \leq \alpha, \beta \leq 1} \frac{R_{11}R_{22} + R_{01}R_{22} + R_{12}R_{01}}{R_{22} + R_{01}} \quad (2.33)$$
where the maximization is now subject to these power constraints:

\[ P'_1 = \frac{R_{22}P_0 + R_{01}P_1}{R_{22} + R_{01}} \quad (2.34) \]
\[ P'_2 = \frac{R_{01}P_2}{R_{22} + R_{01}}. \quad (2.35) \]

**2.2.4 Multi-hop Gaussian Cheap Relay Channel**

Consider the multi-hop Gaussian relay channel of Figure 2.6, in which source node S transmitting information to the destination node D with the help of two cheap relay nodes P and Q. It consists of an input \( x_0 \), two relay outputs \( y_1 \) and \( y_2 \), two relay senders \( x_1 \) and \( x_2 \) (which depends only on the past values of \( y_1 \) and \( y_2 \)), and a channel output \( y \). The channel is assumed to be memoryless. Since the relay nodes are cheap, there are three possible states of operation for this network. The dependency of the outputs on the inputs are as follows: In state \( m_1 \), both relay nodes P and Q act as receivers; and thus the output at node P is given by \( y_1 = h_0x_0 + z_1 \), and the relay Q output is given by \( y_2 = h_1x_0 + z_2 \). In state \( m_2 \), relay node P functions as a transmitter, and the other relay works as a receiver with the received signal given by \( y_2 = h_1x_0 + h_2x_1 + z_2 \). In state \( m_3 \), relay node Q functions as a transmitter and the other relay works as a receiver. Then the received signal at node P is given by \( y_1 = h_0x_0 + z_1 \). The channel output \( y \) is given by \( y = h_3x_2 + z \), where \( h_0, h_1, h_2, \) and \( h_3 \) are channel losses and assumed to be constant and where \( z \sim N(0, N) \), \( z_1 \sim N(0, N_1) \) and \( z_2 \sim N(0, N_2) \) are independent Gaussian noises. Again, we suppose that we can arbitrarily choose between the states \( m_1, m_2, \) and \( m_3 \) in order to maximize the rate.
of information transfer between the source node S and destination node D. Let $t_1$, $t_2$, and $t_3$ represent the time proportion spent in the states $m_1$, $m_2$, and $m_3$, respectively.

In the case of the total average power constraint per nodes as $E[x_0^2] \leq P_1$, $E[x_1^2] \leq P_1$, and $E[x_2^2] \leq P_2$, we can use Theorem 2.2.2 to derive the following achievable rate for the channel, if $N_2 \leq N_1$:

$$R = \max_{0 \leq t_1, t_2, t_3 \leq 1} t_1(R_{11} + R_{01}) + t_2R_{12} + t_3R_{03}. \quad (2.36)$$

Subject to three sets of constraints:

(i)

$$t_3R_{33} \leq t_1R_{11} + t_2(R_{12} + R_{22})$$

$$t_2R_{22} \leq t_1R_{01} + t_3R_{03}$$
(ii)

\[ R_{11} = \frac{1}{2} \log \left(1 + \frac{\gamma_1 P_{11}}{1 + \gamma_1 P_{01}}\right) \]

\[ R_{01} = \frac{1}{2} \log \left(1 + \gamma_0 P_{01}\right) \]

\[ R_{12} \leq \frac{1}{2} \log \left(1 + \gamma_1 (1 - \beta) P_{12}\right) \]

\[ R_{12} + R_{22} \leq \frac{1}{2} \log \left(1 + \gamma_1 P_{12} + \gamma_2 P_{22} + 2\sqrt{\gamma_1 \gamma_2 \beta P_{12} P_{22}}\right) \]

\[ R_{03} = \frac{1}{2} \log \left(1 + \gamma_0 P_{03}\right) \]

\[ R_{33} = \frac{1}{2} \log \left(1 + \gamma_3 P_{33}\right) \]

(iii)

\[ P_1 = t_1 (P_{01} + P_{11}) + t_2 P_{12} + t_3 P_{03} \]

\[ P_2 = t_2 P_{22} \]

\[ P_3 = t_3 P_{33} \]

where \( \gamma_i \)'s, \( i = 0, 1, 2 \) are defined as: \( \gamma_0 = \frac{|h_0|^2}{N_1} \), \( \gamma_1 = \frac{|h_1|^2}{N} \), and \( \gamma_2 = \frac{|h_2|^2}{N} \).

The rates \( R_{01} \) and \( R_{11} \) form the achievable rate region of the broadcast channel in state \( m_1 \), the rate \( R_{12} \) and \( R_{22} \) are the achievable rate region of the multiple access channel in state \( m_2 \) with hierarchical source structure, and the rate \( R_{03} \) and \( R_{33} \) are the corresponding single link capacities.
2.2.5 Concluding Remarks

We presented achievable rates for arbitrary network topologies with cheap nodes. Since the actual capacity is known for very limited sets of network topologies, the results presented in this section are important to obtain useful lower bounds to actual capacity for many cases of practical interest. Application of the new results were presented in the context of relay and multi-hop relay channels. It is also possible to show that the achievable rate in Theorem 2.2.1 subsequently finds the capacity of multi-hop network with cheap nodes [26, 29].
Chapter 3

Gaussian Relay Channel

It has been shown recently that the overall throughput of dense wireless networks can increase significantly when wireless nodes collaborate while transmitting different packets. In this section, we focus on a simple network known as the relay channel, where there is a single relaying node in the system assisting a sender-receiver pair. We present a new achievable rate for the Gaussian relay channels, which outperforms the best-known schemes in many cases of interest. The proposed scheme is a variation of the amplify and forward method and is labeled as a scale and forward scheme, signifying that the relay does not use all its power amplifying the received signal. By controlling the scaling and time correlation of the relay signal, we show that the proposed scheme can outperform all known schemes for several channel gains, especially those where the best known methods provide no improvement over direct transmission. Numerical results are presented here to quantify in which regimes the proposed scheme is the best-known coding scheme for the Gaussian relay channel.
3.1 Summary of existing bounds on the Capacity of the Relay channel

Cooperative coding, where nodes collaborate with each other to improve their data rates, holds the potential of much higher network wide throughputs. The challenge in network coding comes from the fact that none of the current schemes truly extract the potential benefits of cooperating nodes in many cases. The above fact is evident even in the simplest network with one transmitter-receiver pair being helped by one relay node, commonly known as the relay channel [7]. The achievable rate of the best-known coding scheme for the relay channel is limited by the channel SNR of the transmitter-relay link (assuming that it is the better channel) and does not scale with the number of relaying nodes [21]. In this section, we provide the first improvements in the achievable rates of the Gaussian relay channel since [7] for several network instances.

In this section, we derive new lower bounds for the Gaussian relay channel, improving upon the best-known lower bound [7] for several combinations of channel loss parameters. A key case where our proposed achievable rates outperform the decode-and-forward scheme proposed in [7] is the case when the SNR of the transmitter-relay channel is no better than the SNR of the transmitter-receiver channel. The proposed coding scheme also enjoys an additional benefit that it does not require a doubly infinite growth in code lengths, like the Markovian coding scheme of [7]. In fact, the code
lengths required to exploit the capacity benefits of cooperation are of the same order as a point-to-point (no relay) transmission, making the proposed scheme amenable to delay-bounded traffic transmission.

In the proposed coding scheme, the relay does not attempt to decode the data like the Markovian scheme of [7]. Instead, the relay relies on a statistical estimate of the transmitted signal and sends out a scaled version of the input signal. In particular, the relay controls two parameters; the amount of scaling performed in each code block and the correlation introduced across time to enable enhanced signal reception at the receiver.

We note that the proposed scheme is not the best-known scheme for all cases of network parameters, which immediately implies that it is not able to achieve the capacity. But it brings forth an important point that the capacity achieving scheme probably (should not) attempt to decode the transmitted codewords completely in many cases. Instead, we conjecture that it should perform a combination of partial decoding and scale-and-forward operations.

The rest of this section is organized as follows. After some preliminaries in the Section 3.2, we briefly discuss an upper bound for the general Gaussian relay channel in Section 3.3, which is based on the best-known upper bounds for the capacity of the general discrete memoryless relay channels derived in [7]. In Section 3.4, we consider a series of lower bounds on the capacity of the Gaussian relay channel. The main result of this section is a new achievable rate given in Section 3.4. Finally, we conclude
with some remarks in Section 3.6.

### 3.2 Preliminaries

Consider the Gaussian relay channel of Figure 3.1, in which the source node S intends to transmit information to the destination node D by using the direct link between the node pair (S, D) as well as the help of another relay node R (if it improves the achievable rate of transmission) by using link pairs (S, R) and (R, D).

A relay channel, shown in Figure 3.1, consists of an input $x_1$, a relay output $y_1$, a relay sender $x_2$ (which depends only on the past values of $y_1$), and a channel output $y$. The channel is assumed to be memoryless. The dependency of the outputs on the inputs are as follows: the channel output is $y = h_1 x_1 + h_2 x_2 + z$, and the relay output is given by $y_1 = h_0 x_1 + z_1$. The constants $h_0$, $h_1$, and $h_2$ are channel losses and assumed to be constant, and $z \sim N(0, N)$ and $z_1 \sim N(0, N_1)$ are independent Gaussian noises. The input power constraints are given by $E[X_1^2] \leq P_1$ and $E[X_2^2] \leq P_2$. The problem is to find the capacity of the channel between the sender S and receiver D.

An $(M, n)$ code for a Gaussian memoryless relay channel consists of a set of integers $\mathcal{M} = \{1, 2, \ldots, M\} \triangleq [1, M]$; a set of encoding functions $x_1^n : \mathcal{M} \rightarrow \mathcal{R}^n$, where $x_1^n$ denotes a $n$-tuple $(x_{11}, x_{12}, \ldots, x_{1n})$; a set of relay functions $\{f_i\}_{i=1}^n$ such that $x_{2i} = f_i(Y_{11}, Y_{12}, \ldots, Y_{i-1})$ for $1 \leq i \leq n$, where $Y_{ii}$ is the received signal at the relay and $x_{2i}$ is the transmitted signal from the relay at time $i$; and a decoding function $g : \mathcal{Y}^n \rightarrow \mathcal{M}$. For generality, the encoding functions $x_1(\cdot)$, $f_i(\cdot)$ and decoding function
Figure 3.1 : Gaussian Relay Channel

g(.) are allowed to be stochastic functions. At the source node, encoding is only based on the input message. However, the relay has no access to the input message, and because of the non-anticipatory relay condition, relay signal $x_2$ is allowed to depend only on the past $y_1^{(i-1)} = (y_{11}, y_{12}, \ldots, y_{1(i-1)})$ values of the received signals [7].

### 3.3 Upper Bound on the Capacity

In [7], an upper bound for the information transfer rate $R$ in the discrete memoryless relay channel is derived as follows

$$C \leq \sup_{P(x_1,x_2)} \min\{I(X_1;Y,Y_1|X_2), I(X_1,X_2;Y)\} \quad (3.1)$$

This bound has a simple interpretation based on the min-cut max-flow theorem (Theorem 14.9.1 in [8]). For the Gaussian relay channel, this upper bound can be expressed
in terms of the channel parameters and the power constraints as

\[
R \leq \frac{1}{2} \max_{0 \leq \rho \leq 1} \min \{ \log(1 + (1 - \rho^2)(\gamma_0 + \gamma_1)P_1), \\
\log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \} \tag{3.2}
\]

where \( \gamma_0 \triangleq \frac{|h_0|^2}{N_1} \), \( \gamma_1 \triangleq \frac{|h_1|^2}{N} \) and \( \gamma_2 \triangleq \frac{|h_2|^2}{N} \). Here, we wish to emphasize the role of parameter \( \rho \) which corresponds to the correlation factor between the channel input \( X_1 \) and relay signal \( X_2 \). While increasing \( \rho \) would increase the mutual information term \( I(X_1, X_2; Y) \) by helping the transmission in the multiple access cut of the relay channel, increasing \( \rho \) limits the information transfer rate in a broadcast cut of the channel. Therefore for different channel parameters \( h_0, h_1, h_2, N_1, \text{ and } N \), there are different values of the correlation factor \( \rho \) which optimizes the mentioned upper bound. In fact, the correlation between the relay signal and the channel input can be introduced by using two factors: (i) the knowledge of received signal at the relay and (ii) the transmission strategy, e.g., repetition coding at the input of the channel. Clearly, introducing correlation between the channel input and relay signal increases the information transfer rate in the multiple access cut of the relay channel. However, it has its own drawback: it means less information transfers in the broadcast cut of the relay channel which can be interpreted as having prior knowledge of some part of the transmitted message from the source at the relay node (because of the correlation).

The above consideration about the correlation between the source input and relay signal would help in understanding the idea behind the achievable rates of the scale-
and-forward schemes for the relay channel introduced by Propositions 3.5.1, 3.5.2, 3.5.3.

3.4 Lower Bounds on the Capacity

The best-known achievable rate for Gaussian relay channels is the Markovian scheme [7] of Cover and El Gamal presented in 1979. In this scheme transmission occurs in several blocks of long codewords. In each block some information is solely encoded for the reception at the relay, and the codeword length is long enough to allow almost error-free decoding by the relay. Therefore, the source and relay nodes cooperate to resolve, at the destination node, the ambiguity about the message which is sent in the previous block by using the information now shared between the source and relay nodes. Although this clever scheme is capacity achieving for some special cases, it is restricted by decoding and re-encoding at the relay node, and hence its achievable rate is no more than the capacity of the source-relay channel. This scheme is also known as the decode-and-forward scheme because of the aforementioned reason. Since the essential part of coding for the relay channel is the function of the relay, we label different schemes based on their relaying function.

Proposition 3.4.1 (from Theorem 5 [7]). The achievable rate of the decode-and-forward scheme for the Gaussian relay channel is given by

\[
R_{DF} = \frac{1}{2} \max_{0 \leq \rho \leq 1} \min \left\{ \log(1 + (1 - \rho^2)\gamma_0 P_1), \\
\log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \right\}
\]

(3.3)
It can be easily verified that if $\gamma_1 > \gamma_0$, then direct transmission without using the above Markovian scheme would achieve a higher rate. Trivially, we have the following.

**Proposition 3.4.2.** The achievable rate of the direct transmission for the Gaussian relay channel is given by

$$R_{Direct} = \frac{1}{2} \log(1 + \gamma_1 P_1)$$

(3.4)

The above propositions imply that if the relay is in a good situation in terms of the received SNR with respect to the destination received SNR, i.e. $\gamma_0 > \gamma_1$, then using the relay is helpful and improves the achievable rate of the direct transmission. Now, the question is, what if a relay node exists that wants to assist the communication but its received SNR is not as good as the destination received SNR ($\gamma_1 > \gamma_0$). The difficulty with Markovian coding becomes clearer if we consider the situation that the available power of the relay is very large. Unfortunately, even for a very large available power at the relay node, there is no gain over direct transmission by using the decode-and-forward scheme [7] where $\gamma_1 > \gamma_0$. However, instead of decoding at the relay we can use the estimate [7] of the received signal at the relay, i.e., $\hat{Y}_1$. Again, in this scheme encoding is performed in several blocks of large codeword length. The transmission of the estimate of the relay received signal $\hat{Y}_1$, involves using the idea of source coding (at the relay node) and decoding with side information (at the destination node) [37] by considering a test channel with given pdf, $p(\hat{y}_1|x_2,y_1)$. By using statistical transmission of the estimate of the relay received signal [38, 7] the
following rate (3.5) is achievable.

**Proposition 3.4.3** (from [38]). The achievable rate of the estimate-and-forward scheme for the Gaussian relay channel is given by

\[
R_{EF} = \frac{1}{2} \log \left( 1 + \gamma_1 P_1 + \frac{\gamma_0 P_1 \gamma_2 P_2}{1 + \gamma_0 P_1 + \gamma_1 P_1 + \gamma_2 P_2} \right) \tag{3.5}
\]

By comparing (3.5) and (3.4), it is quite clear that the achievable rate of estimate-and-forward, \( R_{EF} \), is always greater than the rate of direct transmission, \( R_{Direct} \). However, depending on the channel conditions the decode-and-forward scheme may achieve a superior rate than the estimate-and-forward scheme.

It should be noted that in both of the estimate-and-forward and the decode-and-forward schemes there are two encoding parameters, the number of blocks and the codeword length in each block, that should become infinitely large. Therefore, the required length of the codeword for these schemes are *doubly infinite* in comparison to that of direct transmission in which only the codeword length grows infinite. In the following, we explore a coding scheme in which the codeword length need not be doubly infinite; yet, it will be shown that their rate can even exceed both the rate of estimate-and-forward and decode-and-forward depending on channel parameters.

As we noted, it is possible to use the received signal and to transmit its estimate with the help of a relay. To overcome the required doubly infinite codeword structure of the estimate-and-forward scheme, we note that it is possible to quantize the received relay signals and to transmit the quantized version through the relay-destination
channel. We can show that the following rate is achievable for the quantize-and-forward scheme.

**Proposition 3.4.4.** The achievable rate of the quantize-and-forward scheme for the Gaussian relay channel is given by

\[ R_{QF} = \max I(X_1; Y; \hat{Y}_1) \]  

(3.6) 

Subject to

\[ H(\hat{Y}_1) < \frac{1}{2} \log \left(1 + \frac{\gamma_2 P_2}{1 + \gamma_1 P_1}\right) \]  

(3.7) 

where maximization is taken over all input distributions \( p(x_1) \) and the quantization function at the relay which generates \( \hat{Y}_1 \) from the received signal \( Y_1 \).

With this approach, the Gaussian input is no longer the optimal input distribution. This approach also implies that a closed form expression for the capacity in terms of the channel parameters is also hard to find [39].

An easier approach compared to the above is simply forwarding a scaled version of the received signal at the relay node. The idea of amplify-and-forward has been used in [40, 41], but it has never been shown to achieve a rate higher than the achievable rate of the decode-and-forward scheme. Here, we show how efficient use of this idea not only can improve upon the decode-and-forward scheme, but it can also achieve a rate higher than that of the estimate-and-forward scheme. It should be pointed out that the amplify-and-forward idea has been also used in [25, 42, 24] for different network topologies. The parallel relaying model of [25] assumes that the
transmission occurs through two relays in two different paths and that it lacks a direct transmission link, which makes the problem more tractable. Therefore, this model trivially allows for the coherent combining of the signals from the relay nodes at the destination without needing any synchronization with direct transmissions. However, the network topology of [42, 24] considers multiple relays and a direct link. A simple but very nice observation suggests that if a source can be transmitted over the channel (a relay channel in this case) for a given distortion measure with an average distortion $\Delta$, then the capacity of the channel is not less than $R(\Delta)$, where $R(.)$ denotes the rate distortion function of the source [24]. By considering the transmission of a Gaussian source and picking the mean square error as our distortion measure, it has been shown that the following average distortion is achievable for the transmission over the Gaussian relay channel,

$$D_1 = \frac{P_1 N}{N + h_2^2 P_1 + \frac{<\alpha,\beta>}{1+||\beta||^2} P_1} \quad (3.8)$$

where for the case of the single relay node $\tilde{\alpha} = h_0$ and $\beta = h_2 \sqrt{\frac{P_2}{h_0^2 P_1 + N_1}}$. Therefore, the achievable rate of this scheme, i.e., uncoded transmission, is given by the following.

**Proposition 3.4.5** (from Theorem 3 [24]). *The achievable rate of the uncoded-transmission scheme of [24] for the Gaussian relay channel is given by*

$$R_{UT} = \frac{1}{4} \log \left( \frac{P_1}{D_1} \right) \quad (3.9)$$
which can be simplified as

\[ R_{UT} = \frac{1}{4} \log \left( 1 + \frac{\gamma_0 P_1 \gamma_2 P_2}{1 + \gamma_0 P_1 + \gamma_2 P_2} \right) \]  \hspace{1cm} (3.10)

It should be pointed out that the network topology of [25] is different and that it lacks the direct transmission link, therefore this result does not apply to our topology. However, we restate the achievable rate of the devised amplify-and-forward schemes of [40] and [41] just for the sake of comparison.

**Proposition 3.4.6** (from [40]). The achievable rate of the amplify-and-forward scheme of [40] for the Gaussian relay channel is given by

\[ R_{AF_1} = \frac{1}{4} \log \left( 1 + \gamma_1 P_1 + \frac{\gamma_0 P_1 \gamma_2 P_2}{1 + \gamma_0 P_1 + \gamma_2 P_2} \right) \]  \hspace{1cm} (3.11)

**Proposition 3.4.7** (from [41]). The achievable rate of the amplify-and-forward scheme of [41] for the Gaussian relay channel is given by

\[ R_{AF_2} = \frac{1}{2} \max_{\alpha, 0 \leq \alpha \leq 1} \log \left( 1 + \frac{(1 - \alpha) \gamma_1 P_1 \alpha \gamma_1 P_1}{1 + \gamma_0 P_1 + 2 \sqrt{\gamma_0 \gamma_1 \gamma_2 P_1} \sqrt{\frac{(1 - \alpha) \gamma_2 P_2}{1 + \gamma_0 P_1}} + \gamma_2 P_2} \right) \]  \hspace{1cm} (3.12)

It can be easily seen that the achievable rate of the amplify-and-forward scheme of [43] (Proposition 3.11) is equal to that of the uncoded-transmission scheme of [24] (Proposition 3.10). It should also be pointed out that the coding scheme which is associated with Proposition 3.4.6 [40] and also the coding schemes of [25, 42] do not need doubly infinite codeword length. However, the coding scheme of [41] also needs a doubly infinite codeword length as that of the decode-and-forward scheme.
3.5 New Achievable rate for the Gaussian relay channel

We name our coding scheme scale-and-forward algorithm primarily in order to differentiate our technique the present amplify-and-forward techniques in the literature. However, scaling is required to control the transmitted power form the relay node and to keep it below the power constraint. Therefore, it does not always corresponds to the amplification of the signal (that literally means an increase in magnitude).

Assume that the transmission of the message occurs in a block of length $ML$ where $M$ is a fixed integer and $L$ is chosen large enough, i.e., $L \to \infty$. We divide the transmission into $L$ consecutive sub-blocks of length $M$. Furthermore, assume that the relay builds its input signal based on all previously received signals in each sub-block. Consider the simple case of $M = 2$.

**Proposition 3.5.1.** The optimal achievable rate of the scale-and-forward scheme with the sub-block length $M = 2$ for the Gaussian relay channel is given by

$$R_{SF(M=2)} = \frac{1}{4} \max_{\beta, Q_1, Q_2} \log \left( 1 + \frac{h_1^2Q_1}{N} + \frac{h_2^2Q_2 + \beta h_0^2h_2^2Q_1}{N + \beta h_2^2N_1} + \frac{h_1^2Q_1Q_2 + \beta h_0^2h_2^2N^2}{N(N + \beta h_2^2N_1)} \right)$$

where $Q_1 + Q_2 = 2P_1$ and $\beta \leq \frac{2P_2}{h_0^2Q_1 + N_1}$

A few notes about these latter achievable rates are in order. First, the proposed scale-and-forward encoding scheme is very simple to implement and is comparable with that of a single-link channel without relay in terms of encoding and decoding.
complexity. Second, the scale-and-forward scheme can be easily adapted to apply to the cheap relay channel [26, 44, 28] discussed in Section 2.1.4.2 (Please refer to Chapter 4 for detailed discussions on cheap relay channel). By definition, a cheap relay is a relay which can either transmit or receive (not both simultaneously) at any given time. This constraint is due to the restriction of the current wireless radios that cannot transmit and receive at the same time in the same frequency band. For example, consider the simple case of the scale-and-forward scheme for the block length $M = 2$. We note that in each sub-block, the relay node only receives during the first transmission and only transmits during the second transmission. Therefore, the scale-and-forward transmission scheme in this case already satisfies the cheap relaying condition. Third, this achievable rate exceeds all the previously mentioned achievable rates, in particular that of the decode-and-forward and the estimate-and-forward schemes, for some channel conditions (see figure 3.2). It also improves over the rates of amplify-and-forward of [40] and [41]. In fact, neither of the amplify-and-forward rates of [40] and [41] ever exceed the rate of the decode-and-forward scheme [7]. Fourth, the optimal value of parameter $\beta$, say $\beta^*$, can be found both analytically and through simulation, and it can be shown that this value is not always equal to $\beta_{\text{max}} = \frac{2P_2}{h_0^2Q_1 + N_1}$. In fact, the ratio $\beta^*/\beta_{\text{max}}$ can be anywhere between zero and one. This is a particularly important observation, because all previous works have considered using full power of the relay for amplification and hence have fixed the value of the scaling factor $\beta$ to the fixed value $\beta_{\text{max}}$. Fifth, in each sub-block the
average power, which is allocated to the transmission in different channel uses, is
different, and controlling this power is important. That is the basic reason why the
amplify-and-forward scheme of [40] has poor performance. Sixth, not only the power
should be optimally allocated, but also the correlation between the transmitted signals
in each sub-block should also be controlled. A specific example in the case of the sub-
block length \( M = 2 \) is that although input signals are chosen independently from one
sub-block to another sub-block, there is an optimum correlation factor between the
input signal in the first and the second transmissions in the sub-blocks. However,
there is a slight degradation in the performance if this optimal correlation factor is
not considered, and the following rate is then achievable.

**Proposition 3.5.2.** The achievable rate of the scale-and-forward scheme with a fully
correlated input signal in the sub-block of length \( M = 2 \) for the Gaussian relay channel
is given by

\[
R'_{SF(M=2)} = \frac{1}{4} \max_{\beta, \alpha} \log \left( 1 + \frac{2 \alpha h_1^2 P_1}{N} + \frac{(h_1 \sqrt{2(1 - \alpha)} P_1 + h_0 h_2 \sqrt{2 \alpha P_1 \beta})^2}{N + \beta h_2^2 N_1} \right) \quad (3.14)
\]

where \( \beta \leq \frac{2 P_2}{2 \alpha P_1 h_0^2 + N_1} \)

It should be pointed out that although this achievable rate \( R'_{SF(M=2)} \) is always less
than the achievable rate \( R_{SF(M=2)} \) in (3.13), it still improves upon both the decode-
and-forward and the estimate-and-forward achievable rates under some values of the
channel parameters.
Figure 3.2: Relaying region for which each of the different coding schemes outperforms other schemes. ($\gamma_1 = 1$, $P_1 = P_2 = -10$dB).
Figure 3.2 shows three different regions based on the different values of the SNR of the source-relay link and relay-destination link where the SNR of the direct link $\gamma_1 = 1$ is assumed to be fixed. Each region shows the corresponding values of the channel parameters for which the marked scheme achieves superior rate than that of all other schemes. The three different regions correspond to decode-and-forward, estimate-and-forward, and a specific example of the scale-and-forward (for sub-block length $M = 2$) schemes. Although finding analytical expression for the achievable rate of the scale-and-forward is hard, the optimal transmission strategy for a specific channel condition can be readily obtained through simulation. Once the optimal input power allocation, input correlation factors, and relay scaling constants have been found, the implementation of the scheme is extremely simple. It has been observed that by increasing the sub-block length of the scale-and-forward scheme, the corresponding region for which this scheme is superior grows larger.

We conclude this section with the following result which characterizes the capacity of the relay channel under the fixed scheme of scale-and-forward at the relay. It should be noted that if we restrict the form of relaying function to be solely the scaled version of the previously received signals, then the following is not just an achievable rate; it is the capacity of the channel under this condition [45]. Furthermore, because in the following result $M \rightarrow \infty$, there is no need to consider a block of length $L$ anymore, i.e. $L = 1$ and the codeword consists of only one block.

**Proposition 3.5.3.** Let $X_1^M$, $X_2^M$, $Y_1^M$, and $Y_2^M$ denote column vectors the channel
input, the relay signal, the received relay signal and the channel output, respectively, for the $M$ consecutive use of the channel. The capacity of the Gaussian relay channel with fixed relaying function of the scale-and-forward scheme is given by

$$C_{SF} = \lim_{M \to \infty} \frac{1}{2M} \sup_{K_x, B} \log \left( \frac{|(N + N_1 h_0^2 BB^T)|}{|N \ldots + (h_1 + h_0 h_2 B)K_x (h_1 + h_0 h_2 B)^T + N_1 h_0^2 BB^T|} \right)$$

Subject to

$$\frac{1}{M} \text{tr}(K_x) \leq P_1,$$

$$\frac{1}{M} \text{tr}(h_0^2 B K_x B^T + NBB^T) \leq P_2,$$

where optimization is over a positive definite matrix $K_x = \mathbb{E}[X_1^M (X_1^M)^T]$, which is the input covariance matrix, and a strictly lower triangular matrix $B$, which denotes the scaling function of the relay defined such that $X_2^M = BY_1^M$.

### 3.6 Concluding remarks

Deriving the capacity of the relay channel has been an open problem since its introduction [6] by Van Der Mullen in 1971. The capacity analysis of the Gaussian relay channel is especially of interest due to its close connection with fading relay channels. To this end, by using a combination of the results of this section, we have derived the exact ergodic capacity of the Rayleigh fading relay channel and also a coding protocol which is robust to the variation of the channel parameters [46]. Moreover, we have shown that a hybrid scheme can perform very effectively for the block fading relay
channel and can approach the universal lower bound on the outage probability [47].

In fact, the lower bound is either achieved in most of the cases or the difference is
negligible for practical purposes.

Based on our intuition, we believe that the gap between the upper bound and
lower bound should be closed by finding better achievable rates and hence improving
the lower bound. Therefore, our goals have been finding (i) better coding schemes and
(ii) better error exponent, i.e., minimizing the required length of the codewords which
can perform close to the capacity of the channel. In particular, both the Markovian
coding scheme (decode-and-forward scheme) [7] and the estimate-and-forward scheme
need doubly infinite long codewords to perform close to the capacity. However, the
proposed scale-and-forward or quantize-and-forward schemes are easily applicable in
practice because their required codeword length is in the same order of the codeword
length of a simple channel with one source and one destination without a relay.
Chapter 4
Cheap Relay Channel

In this chapter, we consider the communication problem in a multi-hop relay network where we assume that the intermediate relay nodes cannot transmit and receive at the same time. The motivation for this assumption comes from the fact that current radios operate in TDD mode when the transmitting and receiving frequencies are the same. We label such a node radio as a *cheap* radio and the corresponding node of the network as a *cheap* node.

We briefly introduced the cheap relay channel [26, 27, 28, 29, 48] in Section 2.1.4.2 and mentioned it again in earlier sections (Section 2.2.4 and 2.2.3). In this section, we derive an efficient coding technique to determine the capacities of the degraded cheap relay channel and the multi-hop network with cheap nodes. The proof of the achievability parts in coding theorems are presented based on the jointly typical sequences, while the proof of the converses are derived from the direct application of the upper bounds shown in Chapter 2 [44]. The coding technique of this chapter is a powerful coding scheme based on partial decoding. The power of this coding technique becomes more evident later in Chapter 5 where the coding technique of this chapter ill be further used to derive some new capacity results for the original (non-cheap) relay channel problem.
4.1 Discrete Memoryless Cheap Relay Channel

4.1.1 Introduction

In wireless networks, most radios operate in half-duplex (TDD) mode when the transmitting and receiving frequencies are the same. Although it might be possible to build the RF radios that are able to send and receive in the same frequency band, design of such radios needs precise and expensive components. We consider the information theoretical approach to quantify the capacity of the network with this practical constraint. Thus, we assume that the nodes of the network can either send or receive at a given time or a given use of the network. We label such a node radio as a *cheap* radio and the corresponding node of the network as a *cheap* node. In this section we find the capacity of the cheap relay channel, Figure 4.1, and the multi-hop network, Figure 4.2, where the intermediate relay nodes are cheap nodes.

The relay channel was first defined and studied by Van der Mullen [6] for the case of three terminals (or nodes) and later considered by Cover [7]. Cover’s results on the relay channel are still the most comprehensive results in terms of achievability and capacity bounds (converses) for the discrete memoryless relay channel. Recent interest in sensor and ad-hoc networks has sparkled new research into the relay channel. Gupta and Kumar [?] presented an extension of the relay channel and showed that using more sophisticated multiuser coding schemes can provide sizable gains in terms of *transport capacity*. Also a follow-up paper by Xie and Kumar [22] established an
explicit achievable rate expression for the degraded Gaussian channel with multiple relays and characterized the bounds on transport capacity. Reznik et al. [23] also considered Gaussian physically degraded relay channels and extended the results for multiple relay stages with a total average power constraint. Gasper and Vetterli [24] derived a lower bound on the capacity of the relay channel by allowing arbitrarily complex network coding. Also, they considered upper bounds from the cut-set theorem [8] and showed that these upper and lower bounds meet asymptotically as the number of relay nodes goes to infinity. In all of the previous work it has been assumed that transmission and reception can be performed at the time in the same frequency band, which is the key difference between the previous works on the relay channel and the results of this section.

The rest of this section is organized in two sections: In section 4.1.2, we consider the problem of the cheap relay channel. In the section 4.1.3, the problem of the multi-hop network with cheap relays is considered.

4.1.2 Cheap Relay Channel

Consider the discrete memoryless relay channel of Figure 4.1, in which source node S is willing to transmit information to destination node D by using a direct link between the node pair (S, D) as well as the help of another relay node R (if it improves the achievable rate of transmission) by using link pairs (S, R) and (R, D). Furthermore assume that relay node R is a cheap node and thus it cannot transmit and receive at
A cheap relay channel, Figure 4.1, consists of an input $x'_1$, a relay output $y'_1$, a
relay sender $x'_2$ (which depends only on the past values of $y'_1$), and a channel output
$y'$. The channel is assumed to be memoryless. With the assumption of a cheap relay
node, there are two possible states of operation for the cheap relay channel. In state
$m_1$, relay node R acts as a receiver and thus the channel probability function is given
by $p(y', y'_1|x'_1|m_1)$, while in the state $m_2$ the relay node functions as a transmitter
and the channel probability function is given by $p(y'|x'_1, x'_2|m_2)$.

Remark: (i) Throughout this section we shall drop the subscripts of probability mass functions,
for example: $p(u|v) = p_{U|V}(u|v)$, since it can be inferred by inspection of the arguments of the
function. We shall also use the notation $X \sim p(x)$ to indicate that the $p(x)$ is the probability
distribution of random variable $X$, or random variable $X$ is drawn according to the probability mass
function $p(x)$.

(ii) The notation $p(y_1, y_2, \ldots, y_N|x_1, x_2, \ldots, x_N|m)$ follows from the notation of [32], and [44] for
the channel probability function (c.p.f.) of multi-state channels.
Let \( X_0 \triangleq X'_1|m_1 \), \( X_1 \triangleq X'_1|m_2 \), \( X_2 \triangleq X'_2|m_2 \), \( Y_0 \triangleq Y'|m_1 \), \( Y_1 \triangleq Y'|m_1 \), \( Y \triangleq Y'|m_2 \). Using these new random variables, the discrete memoryless cheap relay channel can be denoted by two 3-tuple \((\mathcal{X}_0, p(y_0 | y_1, x_0), \mathcal{Y}_0 \times \mathcal{Y}_1), (\mathcal{X}_1 \times \mathcal{X}_2, p(y | x_1, x_2), \mathcal{Y})\), in modes \( m_1 \) and \( m_2 \) respectively, which consists of six finite sets \( \mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}, \mathcal{Y}_0, \mathcal{Y}_1 \) and a collection of probability distributions \( p(., | x_0), p(., | x_1, x_2) \) on \( \mathcal{Y}_0 \times \mathcal{Y}_1, \mathcal{Y} \), one for each \( x_0, (x_1, x_2) \) respectively. The interpretation is that there are two modes \( m_1 \) and \( m_2 \). In mode \( m_1 \), \( x_0 \) is the input to the channel, \( y_0 \) is the output at destination node D and \( y_1 \) is the output at relay node R. In mode \( m_2 \), \( x_2 \) is the input symbol chosen by the relay node R and \( x_1 \) is the source input. Suppose that we can arbitrarily choose the modes \( m_1 \) and \( m_2 \) in order to maximize the rate of information transfer between source node S and destination node D. The problem is to find the capacity of the channel between the sender S and receiver D.

Let \( m_1 = 0, m_2 = 1, \mathcal{S} = \{m_1, m_2\} \). A \((M, n)\) code for a discrete memoryless cheap relay channel consists of a set of integers:

\[
\mathcal{M} = \{1, 2, \ldots, M\} \triangleq [1, M];
\]

a set of encoding functions\(^\dagger\)

\[
s^n : \mathcal{M} \rightarrow \mathcal{S}^n,
\]

\[
x^n_i : \mathcal{M} \rightarrow (\mathcal{X}_{s_1} \times \mathcal{X}_{s_2} \times \ldots \times \mathcal{X}_{s_n})
\]

where \( s_i \) is the state of the network at the \( i \)’th network use, a set of relay functions \( \{f_i\}_{i=1}^n \) such that

\(^\dagger\)We shall use the superscript notation \( v^n \) to indicate a \( n \)-tuple \( (v_1, v_2, \ldots, v_n) \).
\[ x'_{2i} = f_i(Y'_{11}, Y'_{12}, \ldots, Y'_{i(i-1)}) \text{ for } 1 \leq i \leq n; \]

and a decoding function
\[ g : (Y^n_1, S^n) \rightarrow M. \]

For generality, the encoding functions \( x'_i(\cdot), f_i(\cdot) \) and decoding function \( g(\cdot, \cdot) \) are allowed to be stochastic functions. Note that the structure of both the encoding functions at the source and the relay nodes are actually from the definition of the cheap relay channel. At the source node, encoding is based on choosing the mode of operation of the network and then choosing the symbols in the corresponding mode. Also, because of the non-anticipatory relay condition, relay input \( x'_{2i} \) is allowed to depend only on the past \( y'(i-1) = (y'_{11}, y'_{12}, \ldots, y'_{i(i-1)}) \) [7].

In [44], an upper bound for the information transfer rate \( R \) from source node S to the destination node D is shown to be:

\[
R \leq \sup_{t, 0 \leq t \leq 1} \min \left\{ t I(X_1; Y, Y_1|m_1) + (1 - t) I(X_1; Y|X_2, m_2), \right.
\]

\[ m_2), \quad t I(X_1; Y|m_1) + (1 - t) I(X_1, X_2; Y|m_2) \} \tag{4.1} \]

with the new definition of input and output random variables we have:

\[
R \leq \sup_{t, 0 \leq t \leq 1} \min \{ t I(X_0; Y_0, Y_1) + (1 - t) I(X_1; Y|X_2), \right.
\]

\[ t I(X_0; Y_0) + (1 - t) I(X_1, X_2; Y) \} \tag{4.2} \]

In Section 4.1.2.1 it will be shown that for every \( t, 0 \leq t \leq 1 \), the rate \( R^* \) is achievable where \( R^* \) is given by:
Thus, if in the state $m_1$ the received signal $y$ at the destination node D is a degraded form of the received signal $y_1$ at the relay node, then the bound of (4.2) would coincide with this achievable rate for some value of $t$. Hence the capacity of the degraded cheap relay channel would be given by $C = \mathcal{R}^*$ defined in (4.3). Thus we have:

**Theorem 4.1.1.** The capacity of the degraded cheap relay channel is given by

$$
C = \sup_{t, \ p(x_0, x_1, x_2)} \min \{ t I(X_0; Y_1) + (1 - t) I(X_1; Y | X_2),
\quad t I(X_0; Y_0) + (1 - t) I(X_1, X_2; Y) \} 
$$  

(4.4)

where the supremum is taken over $t, 0 \leq t \leq 1$ and all the joint distributions $p(x_0, x_1, x_2)$ on $\mathcal{X}_0 \times \mathcal{X}_1 \times \mathcal{X}_2$.

**4.1.2.1 Achievability of $C$ in Theorem 4.1.1**

In this section, we will show the achievability of the rate $R^*$ in (4.3) for any $t, \ 0 \leq t \leq 1$. We first begin with a brief outline of the proof. We consider a source $U = (W, V)$ which consists of two independent message sources $W$ and $V$ to be transmitted from the source node S to the destination node D. First, the network will be used $t_1$ times in
mode \( m_1 \), in which the source node S transmits \( v \) from the message \( u = (w, v) \) to the relay node R. We assume that this rate is possibly too high for the destination node D to allow reliable decoding. Then, the network will be used \( t_2 \) times in mode \( m_2 \), in which both source node S and relay node R cooperate to resolve the uncertainty of message \( v \) at the receiver of the destination node D. Also, the source node S will transmit the other part of the message \( u = (w, v) \), i.e. message \( w \in W \), to the destination during the use of the channel in this mode. In other words we consider the message source pair \((U, V)\), where the input of the relay node R will be determined by the message source \( V \) and the input of the source node S will be determined by message source \( U \). The destination node D will collect all of the received information in the \( t_1 + t_2 \) uses of the network about the message \( u = (w, v) \) and then will perform the decoding.

Throughout this section we will use the notion of \( \epsilon \)-typical sequences and the asymptotic equipartition property as defined in [35] and [8]. For simplicity of our notation we use the following definitions:

**Definition 1:** (\( n \)'th extension of the memoryless source) The \( n \)'th extension of a random variable \( X \) with distribution \( X \sim p(x) \) is a sequence of independent and identically distributed (i.i.d.) random variables \( X_i, 1 \leq i \leq n \) with the same distribution \( p(x_i) = p(x) \) and is denoted by \( X^n \).

**Definition 2:** (\( n \)'th extension of discrete memoryless channel) The \( n \)'th extension of a discrete memoryless channel with channel probability function \( p(y_1, y_2, \ldots, y_{N_1} | x_1 \)
A block code for the channel consists of three integers \( t_1, t_2, m \) and three encoding functions. In mode \( m_1 \) the encoding function \( x^{t_1m}_1 : V^m \to X^{t_1m}_1 \) assigns codewords to the source outputs. In mode \( m_2 \), the encoding functions \( x^{t_2m}_1 : U^m \to X^{t_2m}_1 \) and \( x^{t_2m}_2 : V^m \to X^{t_2m}_2 \) assign codewords to the source outputs, and finally decoding function \( d^{(t_1+t_2)m} : Y^{(t_1+t_2)m} \to U^m \) decodes the transmitted message.

Let \( \tilde{X}_0 = X^{t_1}_0 \), \( \tilde{Y}_0 = Y^{t_1}_0 \), and \( \tilde{Y}_1 = Y^{t_1}_1 \) be the \( t_1 \)'th extension of the random variables \( X_0, Y_0, Y_1 \), and let \( \tilde{X}_1 = X^{t_2}_1, \tilde{X}_2 = X^{t_2}_2 \), and \( \tilde{Y} = Y^{t_2} \) be the \( t_2 \)'th extension of the random variables \( X_1, X_2, \) and \( Y \) respectively. Also consider the \( t_1 \)'th extension of the channel in mode \( m_1 \) as a channel with c.p.f. \( p(\tilde{y}_0, \tilde{y}_1 | \tilde{x}_0) = p(y^{t_1}_0, y^{t_1}_1 | x^{t_1}_0) \) and similarly the \( t_2 \)'th extension of the channel in mode \( m_2 \) as a channel with c.p.f. \( p(\tilde{y} | \tilde{x}_1, \tilde{x}_2) = p(y^{t_2} | x^{t_2}_1, x^{t_2}_2) \).

In the following subsections we consider \( m \)-sequences of the random variables \( u, v, w, \tilde{x}_0, \tilde{x}_1, \tilde{x}_2, \tilde{y}, \tilde{y}_0, \tilde{y}_1 \) denoted as \( u, v, w, \tilde{x}_0, \tilde{x}_1, \tilde{x}_2, \tilde{y}, \tilde{y}_0, \tilde{y}_1 \).

4.1.2.2 Encoding and Decoding

Generating Random Codes: Fix the conditional probability mass functions \( p(x_0 | v) \), \( p(x_2 | v) \), and \( p(x_1 | u, x_2) \) where \( p(u, v, x_0, x_1, x_2, y, y_0, y_1) = p(u)p(v)p(x_0 | v)p(x_2 | v)p(x_1 | u, x_2)p(y | x_1, x_2)p(y_0, y_1 | x_0) \).
(i) For each $v \in \mathcal{V}^m$ generate one $\tilde{x}_2$ sequence and one $\tilde{x}_0$ sequence drawn according to the distribution $\prod_{i=1}^{m} p(x_{2i}|v_i)$ and $\prod_{i=1}^{m} p(x_{0i}|v_i)$ and label them as $\tilde{x}_2(v)$ and $\tilde{x}_0(v)$ respectively.

(ii) For each $u \in \mathcal{U}^m$ generate one $\tilde{x}_1$ sequence drawn according to the distribution $\prod_{i=1}^{m} p(x_{1i}|x_{2i}, v_i)$ and label it as $\tilde{x}_1(u|\tilde{x}_2)$.

**Encoding:** Source node $S$ uses the network in mode $m_1$ and transmits $\tilde{x}_0(v)$. Relay node $R$ performs decoding upon receiving $\tilde{y}_1$ using jointly $\epsilon$-typical sequences. Provided that the transmitted $v$ in the mode $m_1$ is decodable at the relay node $R$ (but not necessarily at the destination node $D$), the source node $S$ transmits $\tilde{x}_1(u|\tilde{x}_2)$ and the relay node $R$ transmits $\tilde{x}_2(v)$ in mode $m_2$.

**Decoding at the relay node:** The relay node $R$, performs decoding upon receiving $\tilde{y}_1$ using jointly $\epsilon$-typical sequences. In order to find the transmitted message $v$, the decoder at the relay node $R$ finds the only message $v$ such that $(v, \tilde{x}_0(v), \tilde{y}_1) \in A_\epsilon$, where $A_\epsilon$ is the appropriate set of jointly $\epsilon$-typical sequences. If there is no message $v$ or if there exists more than one such message, then the decoder declares error. Thus, if the source has transmitted the message $v_0$ then the error occurs if

(i) $(v_0, \tilde{x}_0(v_0), \tilde{y}_1) \notin A_\epsilon$, or

(ii) there exists some $v \neq v_0$ such that $(v, \tilde{x}_0(v), \tilde{y}_1) \in A_\epsilon$
Decoding at the destination node: The destination node, D, performs decoding upon receiving $\tilde{y}_0$, $\tilde{y}$ in modes $m_1$ and $m_2$ by using jointly $\epsilon$-typical sequences. In order to find the transmitted message $w$, the decoder at the destination node D finds the only pair $(u, v)$ such that $(u, v, \tilde{x}_0(v), \tilde{x}_1(u|\tilde{x}_2), \tilde{x}_2(v), \tilde{y}, \tilde{y}_0) \in A_\epsilon$, where $A_\epsilon$ is the appropriate set of jointly $\epsilon$-typical sequences. If there is no such pairs or if there exists more than one such pair, then the decoder declares error. Thus, if the source has transmitted the pair $(u_0, v_0)$ then the error occurs if

(i) $(u_0, v_0, \tilde{x}_0(v_0), \tilde{x}_1(u_0|\tilde{x}_2), \tilde{x}_2(v_0), \tilde{y}, \tilde{y}_0) \notin A_\epsilon$, or

(ii) there exists some $(u, v) \neq (u_0, v_0)$ such that $(u, v, \tilde{x}_0(v), \tilde{x}_1(u|\tilde{x}_2), 
\tilde{x}_2(v), \tilde{y}, \tilde{y}_0) \in A_\epsilon$

4.1.2.3 Analysis of the error probability

Analysis of the probability of error of the decoder at the destination node: With the mentioned definition of error declaration at the decoder of destination node D, the average probability of error $\bar{P}_{m,D}$ can be written as:

$$\bar{P}_{m,D} = \sum_{(u,v) \in \mathcal{U} \times \mathcal{V}} p(u,v)P\{\text{error made at the decoder of node D} \mid (u,v) \text{ is the transmitted message from S}\}$$  (4.5)
It can be bounded from above by:

\[
\bar{P}_{m,D} \leq \sum_{(u,v) \in A} p(u,v) P\{ \text{error made at the decoder of node D | (u,v) is the transmitted message from S} \} 
+ \sum_{(u,v) \notin A} p(u,v)
\]

(4.6)

From the asymptotic equipartition property (AEP), for the sufficiently large \( n \),

\[
\sum_{(u,v) \notin A} p(u,v) \leq \epsilon, \text{ thus } \bar{P}_{m,D} \leq \sum_{(u,v) \in A} p(u,v) P\{ \text{error made at the decoder of node D | (u,v) is the transmitted message from S} \} + \epsilon
\]

(4.7)

Now, we consider the terms in the summation and find an upper bound which is independent of the \((u,v)\). To show this we assume that \((u_0, v_0) \in A\) is the transmitted message from the source node S, and we let \( \mathcal{F} \) denote such an event. Thus, we are interested in an upper bound for \( P\{ \text{error made at the decoder | } \mathcal{F} \} \)

The error event at the decoder, \( E = E_1 \cup E_2 \), is a union of two events \( E_1 \) and \( E_2 \), where

\( E_1 \): the event that \((u_0, v_0, \tilde{X}_0(v_0), \tilde{X}_1(u_0|\tilde{X}_2), \tilde{X}_2(v_0), \tilde{Y}, \tilde{Y}_0) \notin A \), and

\( E_2 \): the event that there exist some \((u,v) \neq (u_0, v_0)\) such that \((u,v, \tilde{X}_0(v), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v), \tilde{Y}, \tilde{Y}_0) \in A \)
Note that since we have generated the code randomly and we are averaging the probability of error over all coding schemes generated this way, the only random variables in the event $E$ are $\tilde{X}_0$, $\tilde{X}_1$, $\tilde{X}_2$, $\tilde{Y}$, and $\tilde{Y}_0$. Following from the asymptotic equipartition property (AEP), it is always possible to choose $m$ large enough to make the probability of the error event $E_1|\mathcal{F}$ as small as possible,

$$P\{E_1|\mathcal{F}\} \leq \epsilon \quad (4.8)$$

The event $E_2 = E_{21} \cup E_{22}$ itself can be decomposed into a union of two events $E_{21}$ and $E_{22}$, where

$E_{21}$ : the event that there exist a $u \neq u_0$ such that $(u, v_0, \tilde{X}_0(v_0), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v_0), \tilde{Y}, \tilde{Y}_0) \in A_\epsilon$

$E_{22}$ : the event that there exist a $u \neq u_0$, and a $v \neq v_0$ such that $(u, v, \tilde{X}_0(v), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v), \tilde{Y}, \tilde{Y}_0) \in A_\epsilon$

Now we find the bounds on $P\{E_{2i}|\mathcal{F}\}$ for $i = 1, 2$

**Bound for $P\{E_{21}|\mathcal{F}\}$ :** We have

$$P\{E_{21}|\mathcal{F}\} = Pr\{\exists u \neq u_0 : (u, v_0, \tilde{X}_0(v_0), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v_0), \tilde{Y}, \tilde{Y}_0) \in A_\epsilon|\mathcal{F}\} \quad (4.9)$$
Therefore,
\[
P\{E_{21}|\mathcal{F}\} = \sum_{u \neq u_0, (u,v_0) \in A_\varepsilon} P\{(u, v_0, \tilde{X}_0(v_0), \tilde{X}_1(u|\tilde{X}_2), 
\tilde{X}_2(v_0), \tilde{Y}, \tilde{Y}_0) \in A_\varepsilon|\mathcal{F}\}
\]
\[
(4.10)
\]

From the properties of jointly typical sequences, it can be shown that (see Section 4.1.5)
\[
P\{(u, v_0, \tilde{X}_0(v_0), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v_0), \tilde{Y}, \tilde{Y}_0) \in A_\varepsilon|\mathcal{F}\}
\leq 2^{-m[I(\tilde{X}_1;\tilde{Y}|\tilde{X}_2) - 8\varepsilon]}
\]
\[
(4.11)
\]
for \((u, v_0) \in A_\varepsilon\). Note that this bound is independent of \(u\), hence by substituting (4.11) in (4.10) we have
\[
P\{E_{21}|\mathcal{F}\} \leq \sum_{u \neq u_0, (u,v_0) \in A_\varepsilon} 2^{-m[I(\tilde{X}_1;\tilde{Y}|\tilde{X}_2) - 8\varepsilon]}
\]
\[
(4.12)
\]
Thus
\[
P\{E_{21}|\mathcal{F}\} \leq \|\{u : (u, v_0) \in A_\varepsilon\}\| 2^{-m[I(\tilde{X}_1;\tilde{Y}|\tilde{X}_2) - 8\varepsilon]}
\]
\[
(4.13)
\]
and from the joint typicality of \((u, v)\) we have
\[
\|\{u : (u, v_0) \in A_\varepsilon\}\| \leq 2^{m[H(U|V) + 2\varepsilon]}
\]
\[
(4.14)
\]
Combining Equations (4.13) and (4.14) and using the facts that $u = (w, v)$, and also that $w$ and $v$ are independent, we have

$$P\{E_{21}|F\} \leq m[H(W) - I(\tilde{X}_1; \tilde{Y}|\tilde{X}_2) + 10\epsilon]$$  \hspace{1cm} (4.15)

Thus, if

$$H(W) \leq I(\tilde{X}_1; \tilde{Y}|\tilde{X}_2) - 10\epsilon$$  \hspace{1cm} (4.16)

Then for large enough $m$ we have

$$P\{E_{21}|F\} \leq \epsilon$$  \hspace{1cm} (4.17)

*Bound for $P\{E_{22}|F\}$* : We have

$$P\{E_{22}|F\} = Pr\{\exists u \neq u_0, \exists v \neq v_0 : (u, v, \tilde{X}_0(v), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v), \tilde{Y}, \tilde{Y}_0) \in A_\epsilon|F\}$$  \hspace{1cm} (4.18)

for which we have,

$$P\{E_{22}|F\} = \sum_{u \neq u_0, v \neq v_0, (u, v) \in A_\epsilon} P\{(u, v, \tilde{X}_0(v), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v), \tilde{Y}, \tilde{Y}_0) \in A_\epsilon|F\}$$  \hspace{1cm} (4.19)

From the properties of jointly typical sequences, it can be shown that (see Sec-
\[ P\{(u, v, \tilde{X}_0(v), \tilde{X}_1(u|\tilde{X}_2), \tilde{X}_2(v), \tilde{Y}, \tilde{Y}_0) \in A_\epsilon | \mathcal{F} \} \leq 2^{-m[I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) - 8\epsilon]} \] (4.20)

for \((u, v_0) \in A_\epsilon\). By substituting the bound of (4.20) into (4.19) and noting that this bound is independent of \((u, v)\), we have

\[ P\{E_{22}|\mathcal{F}\} \leq \sum_{u \neq u_0, v \neq v_0, (u, v) \in A_\epsilon} 2^{-m[I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) - 8\epsilon]} \] (4.21)

or

\[ P\{E_{22}|\mathcal{F}\} \leq ||\{(u, v) : (u, v) \in A_\epsilon\}|| \times 2^{-m[I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) - 8\epsilon]} \] (4.22)

and from the joint typicality of \((u, v)\) we have

\[ ||\{(u, v) : (u, v) \in A_\epsilon\}|| \leq 2^{m[H(U, V) + 2\epsilon]} \] (4.23)

Combining Equations (4.22) and (4.23) and using the fact that \(u = (w, v)\), we have

\[ P\{E_{22}|\mathcal{F}\} \leq m[H(U) - I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) + 10\epsilon] \] (4.24)
Thus, if
\[ H(U) \leq I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) - 10\epsilon \] (4.25)

Then for large enough \( m \) we have
\[ P\{E_{22}|\mathcal{F}\} \leq \epsilon \] (4.26)

Combining the bounds on \( P\{E_i|\mathcal{F}\} \) for \( i = 1, 2 \) from the Equations (4.17), (4.26) and using the union bound on \( P\{E_2|\mathcal{F}\} \) will result in
\[ P\{E_2|\mathcal{F}\} \leq P\{E_{21}|\mathcal{F}\} + P\{E_{22}|\mathcal{F}\} \leq 2\epsilon \] (4.27)

Considering the union bound for \( P\{E\} \) and using the bounds on \( P\{E_1|\mathcal{F}\}, P\{E_2|\mathcal{F}\} \) in the Equations (4.8) and (4.27), we have
\[ P\{E|\mathcal{F}\} \leq P\{E_1|\mathcal{F}\} + P\{E_2|\mathcal{F}\} \leq 3\epsilon \] (4.28)

Using the definition of the event \( E \) (which is bounded by (4.28)) in the error probability \( \bar{P}_{m,D} \) and substituting in the Equation (4.7), we have
\[ \bar{P}_{m,D} \leq P\{E|\mathcal{F}\} + \epsilon \leq 4\epsilon \] (4.29)

Thus, the average probability of error at the decoder of destination node D can be made arbitrarily small (as of Equation (4.29)) if we use the mentioned encoding scheme for large enough \( m \) and if the conditions of the Equations (4.16) and (4.25) are satisfied.
Analysis of the probability of error of the decoder at the relay node: Similar to what we had for the decoder at the destination node D, the average probability of error can be made arbitrarily small for large enough m. The relay node R performs decoding upon receiving $\tilde{y}_1$ by finding the only message $v$ such that $(v, \tilde{x}_0(v), \tilde{y}_1) \in A_\epsilon$, where $A_\epsilon$ is the appropriate set of jointly $\epsilon$-typical sequences. Thus, the average probability of error at the decoder of relay node R, $\bar{P}_{m,R}$, can be written as

$$\bar{P}_{m,R} = \sum_{v \in V_m} p(v) P\{\text{error made at the decoder of node R} | v \text{ is the transmitted message from S}\}$$ (4.30)

We use the same line of proof similar to what we used to bound the average error probability at the decoder of the destination node. The average probability of error at the decoder of relay node R, $\tilde{P}_{m,R}$, can be bounded from above by:

$$\tilde{P}_{m,R} \leq \sum_{v \in A_\epsilon} p(v) P\{\text{error made at the decoder of node R} | v \text{ is the transmitted message from S}\} + \sum_{v \notin A_\epsilon} p(v)$$ (4.31)

From the asymptotic equipartition property (AEP), for the sufficiently large $n$, we have $\sum_{v \notin A_\epsilon} p(v) \leq \epsilon$, thus

$$\tilde{P}_{m,R} \leq \sum_{v \in A_\epsilon} p(v) P\{\text{error made at the decoder of node R} | v \text{ is the transmitted message from S}\} + \epsilon$$ (4.32)

Considering the terms in the summation, we find an upper bound which is independent of the $v$. To show this we assume that $v_0 \in A_\epsilon$ is the transmitted message
from the source node $S$, and we let $G$ denote such an event. We are interested in an upper bound for $P\{ \text{error made at the decoder} \mid G \}$.

The event or error at the decoder $E$ is a union of two events $E_1$ and $E_2$,

$$E = E_1 \cup E_2$$ \hfill (4.33)

where

$E_1$ : the event that $(v_0, \tilde{X}_0(v_0), \tilde{Y}_1) \notin A_\epsilon$ and

$E_2$ : the event that there exist some $v \neq v_0$ such that $(v, \tilde{X}_0(v), \tilde{Y}_1) \in A_\epsilon$

Note that in this case the only random variables in the event $E$ are $\tilde{X}_0$ and $\tilde{Y}_1$. Following from the asymptotic equipartition property (AEP), it is always possible to choose $m$ large enough to make the probability of the error event $E_1 \mid G$ as small as possible,

$$P\{E_1 \mid G\} \leq \epsilon.$$ \hfill (4.34)

Also, for the event $P\{E_2 \mid G\}$ we have

$$P\{E_2 \mid G\} = Pr\{\exists v \neq v_0 : (v, \tilde{X}_0(v), \tilde{Y}_1) \in A_\epsilon \mid G\}. \hfill (4.35)$$

Thus,

$$P\{E_2 \mid G\} = \sum_{v \neq v_0, v \in A_\epsilon} P\{(v, \tilde{X}_0(v), \tilde{Y}_1) \in A_\epsilon \mid G\}. \hfill (4.36)$$
Using properties of typical sequences it is fairly easy to see that [8]

\[ P\{(v, \tilde{X}_0(v), \tilde{Y}_1) \in A_e | G \} \leq 2^{-m|I(\tilde{X}_0; \tilde{Y}_1) - 3\epsilon} \]  

(4.37)

Therefore, by substituting (4.64) in (4.36) we have

\[ P\{E_2 | G \} \leq \sum_{v \neq v_0, v \in A_e} 2^{-m|I(\tilde{X}_0; \tilde{Y}_1) - 3\epsilon} \]  

(4.38)

or

\[ P\{E_2 | G \} \leq ||\{v : v \in A_e\}|| \cdot 2^{-m|I(\tilde{X}_0; \tilde{Y}_1) - 3\epsilon} \]  

(4.39)

and from the typicality of \( v \) we have

\[ ||\{v : v \in A_e\}|| \leq 2^{m[H(V) + 2\epsilon]} \]  

(4.40)

Combining equations (4.39) and (4.40), we have

\[ P\{E_2 | G \} \leq m[H(V) - I(\tilde{X}_0; \tilde{Y}_1) + 5\epsilon]. \]  

(4.41)

Thus, if

\[ H(V) \leq I(\tilde{X}_0; \tilde{Y}_1) - 5\epsilon \]  

(4.42)

then for large enough \( m \) we have

\[ P\{E_2 | G \} \leq \epsilon. \]  

(4.43)
Therefore from the equations (4.32), (4.34), and (4.43) we have

\[ \bar{P}_{m,R} \leq 3\epsilon. \]  (4.44)

Thus, from the above analysis of the probability of error and from the Equations (4.16), (4.42), and (4.25), the source \( U = (W, V) \) can be transmitted with an arbitrarily small probability of error for large enough \( m \) provided that the following conditions hold

\[
H(W) \leq I(\tilde{X}_1; \tilde{Y}|\tilde{X}_2) \quad (4.45)
\]
\[
H(V) \leq I(\tilde{X}_0; \tilde{Y}_1) \quad (4.46)
\]
\[
H(U) \leq I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) \quad (4.47)
\]

Because the message source \( U = (W, V) \) is composed of two independent message sources \( W \) and \( V \), by combining the Equations (4.45) and (4.46) we have

\[
H(U) = H(W) + H(V) \leq I(\tilde{X}_1; \tilde{Y}|\tilde{X}_2) + I(\tilde{X}_0; \tilde{Y}_1) \quad (4.48)
\]

If for the message source \( U \), the conditions of the Equations (4.47) and (4.48) are satisfied, it can be decomposed into two independent message sources \( W \) and \( V \) such that the rates of the message sources \( W \) and \( V \) satisfy the conditions of the Equations (4.45) and (4.46). Thus, the source \( U \) can be reliably transmitted if

\[
H(U) = \min \{I(\tilde{X}_1; \tilde{Y}|\tilde{X}_2) + I(\tilde{X}_0; \tilde{Y}_1),
\]
\[
I(\tilde{X}_1, \tilde{X}_2; \tilde{Y}) + I(\tilde{X}_0; \tilde{Y}_0) \} \quad (4.49)
\]
Now, Assume that the average rate of source \( U \) per channel use is \( R \), then \( H(U) = (t_1 + t_2)R \). Thus, by using the definitions of \( \hat{X}_0, \hat{X}_1, \hat{X}_2, \hat{Y}, \hat{Y}_0, \) and \( \hat{Y}_1 \), we have

\[
(t_1 + t_2)R = \min\{t_1 I(X_0; Y_1) + t_2 I(X_1; Y|X_2),
\]

\[
t_1 I(X_0; Y_0) + t_2 I(X_1, X_2; Y)\}
\]

(4.50)

By defining \( t \triangleq \frac{t_1}{t_1 + t_2} \), we conclude the achievability of the rate \( R \) for any \( t, 0 \leq t \leq 1 \) defined below which is the same as the rate \( R^* \) in the Equation (4.3).

\[
R = \min\{t I(X_0; Y_1) + (1 - t) I(X_1; Y|X_2),
\]

\[
t I(X_0; Y_0) + (1 - t) I(X_1, X_2; Y)\}
\]

(4.51)

### 4.1.3 Cascaded Network with Cheap Nodes

Consider \( L \) discrete memoryless channel in the cascaded network shown in Figure 4.2. We index each channel from left to right as \( i = 1, 2, \ldots, L \), and we index each node from left to right as \( 0, 1, 2, \ldots, L \). We are interested in transmitting information from node 0 to node \( L \). Thus, node \( i \) receives \( Y_i \) which is the output of the channel \( i \) and transmit its information through signal \( X_{i+1} \) via channel \( i + 1 \), which is the
input to the next channel. Because we have assumed that channels are cascaded and that there are no other connections between the nodes other than stated, the channel output \( Y_i \) just depends on the input \( X_i \) and not on other transmitted signals. For each channel \( i \), where \( i \in \{1, 2, \ldots, L\} \), define capacity of each individual link as \( R_i \triangleq \max_{p(x)} I(X_i; Y_i) \) where maximization is over all possible distributions of \( X_i \).

### 4.1.3.1 Capacity of the Cheap Cascaded Network

It has been shown [8] that the capacity of such a cascaded system without the mentioned practical limitation on transmission and reception at the same time is the minimum of the individual rates of the channels \( C_1 = \min\{R_1, R_2, \ldots, R_L\} \). Because each channel can transmit information at least at the rate of \( C_1 \) without any restriction on receiving data from previous nodes, achievability of this minimum rate is immediate. Also, the known cut-set bound [8] of network information theory states that no higher rate is achievable.

However, imposing the mentioned practical limitation will decrease the achievable rate in this cascaded channel and the mentioned known cut-set bound is no longer tight. In [44], an upper bound for the information transfer rate \( R \) from the node source 0 to the destination node \( L \) is shown to be

\[
R \triangleq R^{(0L)} \leq \sup_{t_m} \min_i \left\{ \sum_{m=1}^{M} t_m \delta_{im} R_i \right\}
\]

(4.52)

when the minimization is taken over \( i, i \in \{1, 2, \ldots, L\} \) and the supremum is over all
the non-negative \( t_m \) subject to \( \sum_{i=1}^{M} t_m = 1 \). In the above expression \( \delta_{im} = 1 \) iff link \( i \) is used in state \( m \) of the network, otherwise \( \delta_{im} = 0 \).

It is also possible to prove that the above bound is actually achievable. The proof is based on the fact that for any given sets of \( \{t_1, t_2, \ldots, t_M\} \) associated with the states 1, 2, \ldots, \( M \) satisfying \( \sum_{i=1}^{M} t_m = 1 \), the rate of \( \min_i \{(\sum_{m=1}^{M} t_m \delta_{im}) R_i\} \) is achievable with an arbitrarily small probability of error. Thus, the above rate

\[
\sup_{t_m} \min_i \{(\sum_{m=1}^{M} t_m \delta_{im}) R_i\}
\]

is the capacity of the multi-hop network with cheap nodes, and in section 4.1.3.2.2 we will prove the following capacity expression.

**Theorem 4.1.2.** The capacity of the Cascaded Network of the figure 4.2 is given by

\[
C = \min \left\{ \frac{R_1 R_2}{R_1 + R_2}, \frac{R_2 R_3}{R_2 + R_3}, \ldots, \frac{R_{L-1} R_L}{R_{L-1} + R_L} \right\}
\]  

(4.53)

where \( R_i := \max_{p(x)} I(X_i; Y_i) \) is defined as the capacity of link \( i \).

**4.1.3.2 Sketch of the Proof for Theorem 4.1.2**

Consider the cascaded network of Figure 4.2. First we prove that the rate \( R \) defined in the Equation (4.52) cannot be higher than \( C \) defined in the Equation (4.53). To show this, we consider two cut-set \( C_i, C_{i+1} \) for each \( i \in \{1, 2, \ldots, L - 1\} \). Let \( T_i \triangleq \sum_{m=1}^{M} t_m \delta_{im} \). Because for each \( i \) and \( m \) we have \( \delta_{im} \delta_{(i+1)m} = 0 \), it can easily be verified that \( T_i + T_{i+1} \leq 1 \). Thus,

\[
R \leq \min \{T_i R_i, T_{i+1} R_{i+1}\} \leq \frac{R_i R_{i+1}}{R_i + R_{i+1}}
\]  

(4.54)
Therefore, if we consider the Equation (4.54) for all values of $i \in \{1, 2, \ldots, L - 1\}$, then we conclude that $R \leq C$.

It is possible to show that states 1, 2, \ldots, $M$ and their associated set of $\{t_1, t_2, \ldots, t_M\}$ exist such that the rate $C$ defined in the Equation (4.53) is achievable. Assume that in an $L$-hop network the minimum of

$$\{ \frac{R_1 R_2}{R_1 + R_2}, \frac{R_2 R_3}{R_2 + R_3}, \ldots, \frac{R_{L-1} R_L}{R_{L-1} + R_L} \}$$

occurs at node $i$, i.e.

$$C = \frac{R_i R_{i+1}}{R_i + R_{i+1}} = \min \left\{ \frac{R_1 R_2}{R_1 + R_2}, \frac{R_2 R_3}{R_2 + R_3}, \ldots, \frac{R_{L-1} R_L}{R_{L-1} + R_L} \right\}$$

(4.55)

then it is easy to see that $T_i = \frac{R_{i+1}}{R_i + R_{i+1}}$ and $T_{i+1} = \frac{R_i}{R_i + R_{i+1}}$. We use strong induction on the number of hops to prove that $C$ in the Equation (4.53) is achievable. For $L = 2$ it is fairly easy to see that using two states 1, 2 with the associated timing set of $\{t_1, t_2\}$ are enough when, $T_1 = t_1 = \frac{R_2}{R_2 + R_1}$ and $T_2 = t_2 = \frac{R_1}{R_1 + R_2}$. Now suppose that for all values of $L$, $L \in \{2, 3, \ldots, K - 1\}$ it is possible to find the states 1, 2, \ldots, $M$ and their associated timing set of $\{t_1, t_2, \ldots, t_M\}$ for any $L$-hop network, such that the rate $C$ defined in the Equation (4.53) is achievable. We will show that the rate $C$ is achievable for any K-hop network as well, and we will show how to choose the corresponding states and the associated set of timing for each state. Assume that for the K-hop network the min of the expression

$$\{ \frac{R_1 R_2}{R_1 + R_2}, \frac{R_2 R_3}{R_2 + R_3}, \ldots, \frac{R_{K-1} R_K}{R_{K-1} + R_K} \}$$
occurs at node $i$. If $i \in \{2, \ldots, K - 2\}$, then we can consider the left and right networks defined by the sets of nodes $\{0, 1, 2, \ldots, i+1\}$ and $\{i-1, i, i+1, \ldots, K\}$, respectively. From the assumption of induction, it is possible to find two different sets of modes and their associated timing sets for the left and right network such that rate $\frac{R_i R_{i+1}}{R_i + R_{i+1}}$ is achievable in both networks. Because for both the left and right networks the minimum occurs at node $i$, for both of the mentioned sets of modes, we have $T_i = \frac{R_{i+1}}{R_i + R_{i+1}}$ and $T_{i+1} = \frac{R_i}{R_i + R_{i+1}}$. Therefore, it is possible to compose the set of modes and timing sets for the left and right network in order to find a set of modes and the associated timing set for the K-hop network.

Now we consider the case that the minimum occurs at node $i = 1$, i.e. $C = \frac{R_1 R_2}{R_1 + R_2}$ (the other case when the minimum occurs at the node $i = K - 1$ is similar). In this case we choose another $R'_3$ such that

$$C = \frac{R_1 R_2}{R_1 + R_2} = \frac{R'_3 R_2}{R'_3 + R_2}$$

if $R_2 \leq R_4$, or

$$C = \frac{R_1 R_2}{R_1 + R_2} = \frac{R'_3 R_4}{R'_3 + R_4}$$

if $R_4 \leq R_2$, and we define a new K-hop network in which the capacity of link 3 is given by $R'_3$ instead of $R_3$. Note that with this definition the minimum still occurs at Node 1 as well as Node 2 if $R_2 \leq R_4$, or at Node 3 if $R_2 \leq R_4$.

In the first case when $R_2 \leq R_4$, we decompose the network into the left and right networks defined by the sets of the nodes $\{0, 1, 2, 3\}$ and $\{1, 2, \ldots, K\}$ respectively.
From the assumption of induction it is possible to find a set of modes and their associated timing sets for the right network such that rate $\frac{R'_3 R_2}{R'_3 + R_2}$ is achievable in this network. Since in this case $R'_3 = R_1$, it is sufficient to assign all the modes in which the Link 2 is "off" to the Link 1.

In the other case when $R_4 \leq R_2$, we decompose the network into the left and right networks defined by the sets of the nodes $\{0, 1, \ldots, 4\}$ and $\{2, 3, \ldots, K\}$ respectively. From the assumption of induction it is possible to find a set of modes and their associated timing sets for both the left and right networks such that rate $\frac{R'_3 R_4}{R'_3 + R_4}$ is achievable in both of the networks. Note that for both of the above cases retrieving the original $R_3$ will not decrease the achievable rates because by our definition $R'_3$ it is always greater than or equal to the $R_3$. It completes the inductive proof.

4.1.4 Concluding remarks

We derived the capacity of the degraded cheap relay channel and the multi-hop network with the cheap nodes as stated in Theorems 4.1.1 and 4.1.2. The assumption of cheap relay nodes in the network is important for design of the practical systems, and the results of this section characterize the limits of information transfer in such a networks.
4.1.5 Bounds on Non-typical Event Probabilities

In this Section we find the bounds on $P\{(u, v, X_0(v), X_1(u|X_2), X_2(v), Y, Y_0) \in A_\epsilon|\mathcal{F}\}$ in Equations (4.11) and (4.20) by using the following lemma[35].

**Lemma 4.1.3.** Let $(Z_1, Z_2, Z_3, Z_4, Z_5)$ be random variables with distribution $P(z_1, z_2, z_3, z_4, z_5)$. Fix $(z_1, z_2) \in A_\epsilon$, and let $Z_3, Z_4, Z_5$ be m-sequences drawn according to

$$P(Z_3 = z_3, Z_4 = z_4, Z_5 = z_5|z_1, z_2) = \prod_{i=1}^{m} P(z_{3i}|z_{1i}, z_{2i})P(z_{4i}|z_{3i}, z_{2i})P(z_{5i}|z_{3i}, z_{1i})$$

(4.56)

In other words, $Z_3$ depends only on $Z_1$ and $Z_2$, $Z_4$ depends only on $Z_3$ and $Z_2$, and finally $Z_5$ depends only on $Z_3$ and $Z_1$. Then

$$P\{(z_1, z_2, Z_3, Z_4, Z_5) \in A_\epsilon\} \leq 2^{-m[I(Z_1;Z_4|Z_2,Z_3)+I(Z_5;Z_2,Z_4|Z_1,Z_3) - 8\epsilon]}$$

(4.57)

In order to find the bound on $P\{(u, v_0, X_0(v_0), X_1(u|X_2), X_2(v_0), Y, Y_0) \in A_\epsilon|\mathcal{F}\}$ in Equation (4.11), we use the Lemma with $z_1 = v_0$, $z_2 = u$, $Z_3 = X_2(u)$, $Z_4 = X_1(u|X_2)$, $Z_5 = Y$. Clearly, the conditions of the Lemma on the conditional distribution of $Z_3, Z_4, Z_5$ given $z_1, z_2$ are held. Using the fact that $X_1$ and $(X_2, U)$ are conditionally independent given $V$ and also that $Y$ and $(U, V)$ are conditionally independent given $(X_1, X_2)$, it is possible to show that

$$I(Z_1; Z_4|Z_2, Z_3) = I(X_1; V|X_2, U) = 0,$$

(4.58)
\[ I(Z_5; Z_2, Z_4 | Z_1, Z_3) = I(Y; U, X_1 | X_2, V) = I(Y; X_1 | X_2) \] (4.59)

Thus

\[ P\{(u, v_0, X_0(v_0), X_1(u | X_2), X_2(v_0), Y, Y_0) \in A_\epsilon | \mathcal{F}\} \leq 2^{-m[I(X_1; X_2) - 8\epsilon]} \] (4.60)

Also, to find the bound on \( P\{(u, v, X_0(v), X_1(u | X_2), X_2(v), Y, Y_0) \in A_\epsilon | \mathcal{F}\} \) in Equation (4.20), we use the Lemma with \( z_1 = \emptyset, z_2 = (u, v), Z_3 = \emptyset, Z_4 = (X_1, X_2), Z_5 = (Y, Y_0) \). It is easy to verify that the conditions of the lemma on the conditional distribution of \( Z_3, Z_4, Z_5 \) given \( z_1, z_2 \) are held. In this case, we use the following facts that (i) \( X_1 \) and \( (X_2, U) \) are conditionally independent given \( V \), (ii) \( Y \) and \( (U, V) \) are conditionally independent given \( (X_1, X_2) \), and (iii) \( Y_0 \) and \( (U, V) \) are conditionally independent given \( X_0 \). It is possible to verify that

\[ I(Z_1; Z_4 | Z_2, Z_3) = 0 \quad \text{(Note that } Z_1 = \emptyset) \] (4.61)

\[ I(Z_5; Z_2, Z_4 | Z_1, Z_3) = I(X_1, X_2; Y) + I(X_0; Y_0) \] (4.62)

Thus

\[ P\{(u, v, X_0(v), X_1(u | X_2), X_2(v), Y, Y_0) \in A_\epsilon | \mathcal{F}\} \leq 2^{-m[I(X_1, X_2; Y) + I(X_0; Y_0) - 8\epsilon]} \] (4.63)

Also, it is fairly easy to show that if the rate of source \( V \) is less than \( I(X_0; Y_1) \), then

\[ P\{(v, X_0(v), Y_1) \in A_\epsilon | \mathcal{G}\} \leq 2^{-m[I(X_0; Y_1) - 7\epsilon]} \] (4.64)
4.2 The Gaussian Cheap Relay Channel

In this section, we consider the Gaussian version of the cheap relay channel considered in the previous sections. First, we derive an upper bound on the capacity of the Gaussian cheap relay channel. Then, we derive the capacity of the Gaussian degraded cheap relay channel (Theorem 4.2.2). The proof of achievability in Theorem 4.2.2 is derived based on the same coding scheme in [26]. Because we do not use the assumption of degradedness in the proof of the coding theorem, this rate can be interpreted as a lower bound on the capacity of the Gaussian cheap relay channel in general. By applying the assumption of degradedness, the converse is proved with slight changes from the proof of the derived upper bound on the capacity of the Gaussian cheap relay channel. Our results show that even in cheap relay channels, there is a sizable gain in capacity if we use the relay.

The general problem setup for the Gaussian cheap relay channel is similar to that of the discrete memoryless cheap relay channel defined in Section 4.1.2. However, for a detailed definition of Gaussian relay channel, its specific input power constraints, notations, and parameters, we refer to Section 2.2.3. In the next section (Section 4.2.1), we drive an upper bound on the capacity of the cheap relay channel for a given power constraint. In the subsequent sections, Section 4.2.2 and 4.2.3, we derive the capacity of the Gaussian degraded cheap relay channel which follows the detailed derivation of Section 4.2.1 for the lower bound on the capacity.
4.2.1 Upper bound on the Capacity

In [44], an upper bound for the information transfer rate $R$ in the discrete memoryless cheap relay channel is derived as follows

$$R \leq \sup_{t, 0 \leq t \leq 1} \min \{ t I(X_1; Y, Y_1 | m_1) + (1 - t) I(X_1; Y | X_2, m_2), \ t I(X_1; Y | m_1) + (1 - t) I(X_1, X_2; Y | m_2) \}. \quad (4.65)$$

Although the above bound has been obtained for the discrete memoryless cheap relay channel, the same proof is valid for any discrete time channel even with continuous input and output alphabets. To express the bound in terms of the power constraints in a Gaussian cheap relay channel, we use the following bound from [44]:

$$n R \leq \max_{k, 0 \leq k \leq n} \min \left\{ \sum_{i=1}^{k} I(X_{1i}; Y_i, Y_{1i} | m_1) + \sum_{i=1}^{n-k} I(X_{1i}; Y_i | X_{2i}, m_2), \right.$$  
$$\left. \sum_{i=1}^{k} I(X_{1i}; Y_i | m_1) + \sum_{i=1}^{n-k} I(X_{1i}, X_{2i}; Y_i | m_2) \right\} + n \epsilon_n. \quad (4.66)$$

Now, we find an upper bound for each of the mutual information terms in the above expression in terms of the power constraints and the channel parameters.‡

‡For simplicity we shall drop the state indices in the entropy functions, and the choice will be clear by examining the state index in the associated mutual information.
4.2.1.1 Bound on $I(X_{1i};Y_i, Y_{1i}|m_1)$

$$
\sum_{i=1}^{k} I(X_{1i};Y_i, Y_{1i}|m_1) \\
\leq \sum_{i=1}^{k} [h(Y_i) + h(Y_{1i}|Y_i) - h(Y_i|X_{1i}) - h(Y_{1i}|X_{1i}, Y_i)] \\
\leq \sum_{i=1}^{k} [h(Y_i) + h(Y_{1i}|Y_i) - \frac{1}{2} \log(2\pi eN) - \frac{1}{2} \log(2\pi eN_1)].
$$

(4.67)

Now, for any $i$,

$$
h(Y_{1i}|Y_i) = E[h(Y_{1i}|y_i)] \\
\leq E \left[ \frac{1}{2} \log(2\pi e \text{Var}(Y_{1i}|y_i)) \right] \\
\leq \frac{1}{2} \log(2\pi e E[\text{Var}(Y_{1i}|y_i)]) \\
\leq \frac{1}{2} \log(2\pi e (E[Y_{1i}^2] - E[E^2[Y_{1i}|Y_i]])) \\
\leq \frac{1}{2} \log \left( 2\pi e \left( E[Y_{1i}^2] - \frac{E^2[Y_{1i}Y_i]}{E[Y_i^2]} \right) \right) \\
\leq \frac{1}{2} \log \left( 2\pi e \frac{(h_0^2 N + |h_1|^2 N_1 E[X_{1i}^2] + NN_1)}{|h_1|^2 E[X_{1i}^2] + N} \right).
$$

(4.68)

Also

$$
h(Y_i) \leq E \left[ \frac{1}{2} \log(2\pi e \text{Var}(Y_i)) \right] \\
\leq \frac{1}{2} \log(2\pi e(|h_1|^2 E[X_{1i}^2] + N)).
$$

(4.69)
From (4.67), (4.68), and (4.69) and using Jensen’s inequality we have

$$\sum_{i=1}^{k} I(X_{1i}; Y_{i}, Y_{1i}|m_1)$$

$$\leq \sum_{i=1}^{k} \left[ \frac{1}{2} \log(2\pi e (|h_1|^2 E[X_{1i}^2] + N))
+ \frac{1}{2} \log \left( 2\pi e \frac{|h_0|^2 N + |h_1|^2 N_1 E[X_{1i}^2] + NN_1}{|h_1|^2 E[X_{1i}^2] + N} \right)
- \frac{1}{2} \log((2\pi e)^2 NN_1) \right]$$

$$\leq \sum_{i=1}^{k} \frac{1}{2} \log \left( 1 + \frac{|h_0|^2 E[X_{1i}^2]}{N_1} + \frac{|h_1|^2 E[X_{1i}^2]}{N} \right)$$

$$\leq \frac{k}{2} \log \left( 1 + \left( \frac{|h_0|^2}{N_1} + \frac{|h_1|^2}{N} \right) \frac{1}{k} \sum_{i=1}^{k} E[X_{1i}^2] \right).$$  \hspace{1cm} (4.70)

Thus, by using the power constraints we have

$$\sum_{i=1}^{k} I(X_{1i}; Y_{i}, Y_{1i}|m_1) \leq \frac{k}{2} \log \left( 1 + \frac{|h_0|^2 P_0}{N_1} + \frac{|h_1|^2 P_0}{N} \right).$$  \hspace{1cm} (4.71)

### 4.2.1.2 Bound on $I(X_{1i}; Y_{i}|X_{2i}, m_2)$

Similarly, for the second mutual information term in (4.66), we have

$$\sum_{i=1}^{n-k} I(X_{1i}; Y_{i}|X_{2i}, m_2)$$

$$\leq \sum_{i=1}^{n-k} [h(Y_{i}|X_{2i}) - h(Y_{i}|X_{1i}, X_{2i})]$$

$$\leq \sum_{i=1}^{n-k} \left[ h(Y_{i}|X_{2i}) - \frac{1}{2} \log(2\pi e N) \right].$$  \hspace{1cm} (4.72)
Now, for any $i$, 

$$h(Y_i|X_{2i}) = E[h(Y_i|x_{2i})]$$

$$\leq E \left[ \frac{1}{2} \log(2\pi e \text{Var}(Y_i|x_{2i})) \right]$$

$$\leq \frac{1}{2} \log(2\pi e E[\text{Var}(Y_i|X_{2i})])$$

$$\leq \frac{1}{2} \log(2\pi e (E[Y_i^2] - E^2[Y_i|X_{2i}]))$$

$$\leq \frac{1}{2} \log \left( 2\pi e \left( \frac{E[Y_i^2] - E^2[Y_i|X_{2i}]}{E[X_{2i}^2]} \right) \right)$$

$$\leq \frac{1}{2} \log \left( 2\pi e \left( |h_1|^2 E[X_{1i}^2] + N \right) \right)$$

$$- \frac{|h_1|^2 E^2[X_{1i},X_{2i}]}{E[X_{2i}^2]} \right) \right). \quad (4.73)$$

Define

$$\rho^2 = \sum_{i=1}^{n-k} \frac{1}{(n - k)} \frac{E^2[X_{1i},X_{2i}]}{E[X_{1i}^2]E[X_{2i}^2]} \quad (4.74)$$

From (4.72) and (4.73), and using Jensen’s inequality we have

$$\sum_{i=1}^{n-k} I(X_{1i}; Y_i|X_{2i}, m_2)$$

$$\leq \sum_{i=1}^{n-k} \left[ \frac{1}{2} \log \left( 2\pi e \left( |h_1|^2 E[X_{1i}^2] + N \right) \right) - \frac{1}{2} \log(2\pi e N) \right]$$

$$\leq \sum_{i=1}^{n-k} \frac{1}{2} \log \left( 1 + \frac{(1 - \rho^2)|h_1|^2}{N} E[X_{1i}^2] \right)$$

$$\leq \frac{n - k}{2} \log \left( 1 + \frac{(1 - \rho^2)|h_1|^2}{N} \frac{1}{n - k} \sum_{i=1}^{n-k} E[X_{1i}^2] \right). \quad (4.75)$$
Thus, by using the power constraints

\[
\sum_{i=1}^{n-k} I(X_{1i}; Y_i | X_{2i}, m_2) \leq \frac{n-k}{2} \log \left( 1 + \frac{(1-\rho^2)|h_1|^2 P_1}{N} \right).
\] (4.76)

### 4.2.1.3 Bound on \( I(X_{1i}; Y_i | m_1) \):

Now consider the third mutual information term in (4.66). We have

\[
\sum_{i=1}^{k} I(X_{1i}; Y_i | m_1) \leq \sum_{i=1}^{k} [h(Y_i) - h(Y_i | X_{1i})]
\]

\[
\leq \sum_{i=1}^{k} \left[ h(Y_i) - \frac{1}{2} \log(2\pi e) \right].
\] (4.77)

Now, for any \( i \),

\[
h(Y_i) \leq E \left[ \frac{1}{2} \log(2\pi e \text{Var}(Y_i)) \right]
\]

\[
\leq \frac{1}{2} \log(2\pi e (|h_1|^2 E[X_{1i}^2] + N)).
\] (4.78)

From (4.77) and (4.78) and using Jensen’s inequality we have

\[
\sum_{i=1}^{k} I(X_{1i}; Y_i | m_1)
\]

\[
\leq \sum_{i=1}^{k} \left[ \frac{1}{2} \log(2\pi e (|h_1|^2 E[X_{1i}^2] + N)) - \frac{1}{2} \log(2\pi e) \right]
\]

\[
\leq \sum_{i=1}^{k} \frac{1}{2} \log \left( 1 + \frac{|h_1|^2 E[X_{1i}^2]}{N} \right)
\]

\[
\leq \frac{k}{2} \log \left( 1 + \frac{|h_1|^2}{N} \frac{1}{k} \sum_{i=1}^{k} E[X_{1i}^2] \right).
\] (4.79)
Thus, by using the power constraints
\[
\sum_{i=1}^{k} I(X_{1i}; Y_{i|m_1}) \leq k \log\left( 1 + \frac{|h_1|^2 P_0}{N} \right).
\] (4.80)

### 4.2.1.4 Bound on \( I(X_{1i}, X_{2i}; Y_{i|m_2}) \)

Finally, the last mutual term in (4.66) can be bounded as
\[
\sum_{i=1}^{n-k} I(X_{1i}, X_{2i}; Y_{i|m_2}) \leq \sum_{i=1}^{n-k} [h(Y_i) - h(Y_i|X_{1i}, X_{2i})]
\]
\[
\leq \sum_{i=1}^{n-k} \left[ h(Y_i) - \frac{1}{2} \log(2\pi e N) \right].
\] (4.81)

For any \( i \),
\[
h(Y_i) = h(h_1 X_{1i} + h_2 X_{2i} + Z_i)
\]
\[
\leq E \left[ \frac{1}{2} \log(2\pi e (E[(h_1 X_{1i} + h_2 X_{2i})^2] + N)) \right]
\]
\[
\leq \frac{1}{2} \log(2\pi e(|h_1|^2 E[X_{1i}^2] + |h_2|^2 E[X_{2i}^2]
\]
\[
+ 2|h_1 h_2| E[X_{1i} X_{2i}]))\]
\[
\leq \frac{1}{2} \log(2\pi e(|h_1|^2 E[X_{1i}^2] + |h_2|^2 E[X_{2i}^2]
\]
\[
+ 2|h_1 h_2| E[X_{2i} E[X_{1i} X_{2i}]]).\] (4.82)
Recall the definition of $\rho$ in Equation (4.74). From (4.81) and (4.82) and using Jensen’s inequality we have

$$
\sum_{i=1}^{n-k} I(X_{1i}, X_{2i}; Y_i|m_2)
\leq \sum_{i=1}^{n-k} \left[ \frac{1}{2} \log(2\pi e (|h_1|^2 E[X_{1i}^2] + |h_2|^2 E[X_{2i}^2]) + 2|h_1 h_2| E[X_{2i} E[X_{1i} | X_{2i}]]) - \frac{1}{2} \log(2\pi e N) \right]
\leq \sum_{i=1}^{n-k} \frac{1}{2} \log \left( 1 + \left( \frac{|h_1|^2}{N} E[X_{1i}^2] + \frac{|h_2|^2}{N} E[X_{2i}^2] + \frac{2|h_1 h_2|}{N} E[X_{2i} E[X_{1i} | X_{2i}]] \right) \right)
\leq \frac{n-k}{2} \log \left( 1 + \left( \frac{|h_1|^2}{N} + \frac{|h_2|^2}{N} \sum_{i=1}^{n-k} E[X_{1i}^2] + \frac{2|h_1 h_2|}{N} \sum_{i=1}^{n-k} E[X_{2i} E[X_{1i} | X_{2i}]] \right) \right). 
$$

Thus, by using Cauchy-Schwartz inequality, and substituting the values of $\rho$ and the power constraints, we have

$$
\sum_{i=1}^{n-k} I(X_{1i}, X_{2i}; Y_i|m_2)
\leq \frac{n-k}{2} \log \left( 1 + \left( \frac{|h_1|^2 P_1}{N} + \frac{|h_2|^2 P_2}{N} + \frac{2|h_1 h_2|}{N} \sum_{i=1}^{n-k} \sqrt{E[X_{2i}^2] E[E^2[X_{1i} | X_{2i}]]} \right) \right)
\leq \frac{n-k}{2} \log \left( 1 + \left( \frac{|h_1|^2 P_1}{N} + \frac{|h_2|^2 P_2}{N} + \frac{2|h_1 h_2| \rho \sqrt{P_1 P_2}}{N} \right) \right). 
$$
Now, by substituting the calculated bounds for the mutual information in Inequalities (4.71), (4.76), (4.80), and (4.84) into Inequality (4.66) we have

\[
n R \leq \max_{0 \leq k \leq n} \min \left\{ \frac{k}{2} \log \left( 1 + \frac{|h_0|^2 P_0}{N_1} + \frac{|h_1|^2 P_0}{N} \right) + \frac{n-k}{2} \log \left( 1 + \frac{(1-\rho^2)|h_1|^2 P_1}{N} \right) , \right. \\
\frac{k}{2} \log \left( 1 + \frac{|h_1|^2 P_0}{N} \right) + \frac{n-k}{2} \log \left( 1 + \frac{|h_1|^2 P_1}{N} \right) + \frac{|h_2|^2 P_2}{N} + 2 |h_1 h_2| \rho \sqrt{P_1 P_2} \left) \right\} + n \epsilon_n. \tag{4.85}
\]

Define \( \gamma_0 \triangleq \frac{|h_0|^2 P_0}{N_1} \), \( \gamma_1 \triangleq \frac{|h_1|^2 P_0}{N} \) and \( \gamma_2 \triangleq \frac{|h_2|^2 P_2}{N} \). By dividing both sides of the above bound (4.85) by \( n \) and letting \( n \to \infty \), we have the following theorem.

**Theorem 4.2.1.** An upper bound for the capacity of the Gaussian cheap relay channel with the input-output relations and power constraints defined in Section 2.2.3 is given by

\[
R \leq \frac{1}{2} \max_{0 \leq t, \rho \leq 1} \min \{ t \log(1 + (\gamma_0 + \gamma_1) P_0) + (1-t) \log(1 + (1-\rho^2) \gamma_1 P_1), \ t \log(1 + \gamma_1 P_0) + (1-t) \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2 \rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \}. \tag{4.86}
\]

### 4.2.2 The Gaussian Degraded Cheap relay Channel

Suppose that the transmitter at the node S with some power constraint sends a signal intended for the receiver of D. However, this signal is also received by the relay node R which is perhaps physically closer to node S. This is an example of a scenario where
we can assume that the received signal at the node D is a degraded version of the
received signal at the node R. We derive the capacity of the Gaussian degraded cheap
relay channel in the following theorem.

**Theorem 4.2.2.** Capacity of the Gaussian degraded cheap relay channel with the
input-output relations and power constraints defined in section 2.2.3 is given by

\[
C = \frac{1}{2} \max_{t, \rho, 0 \leq t, \rho \leq 1} \min \{ t \log (1 + \gamma_0 P_0) + \\
(1 - t) \log (1 + (1 - \rho^2) \gamma_1 P_1), t \log (1 + \gamma_1 P_0) + \\
(1 - t) \log (1 + \gamma_1 P_1 + \gamma_2 P_2 + 2 \rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \} \tag{4.87}
\]

In Section 4.2.3, we prove the achievability of any rate \( R \leq C \) in (4.87) whether
the channel is degraded or not. Thus, this achievable rate can be considered as a
lower bound to the capacity of the Gaussian cheap relay channel in general. On the
other hand, with the assumption of a degraded relay channel the upper bound of the
Theorem 4.2.1 can be made smaller. In fact \( I(X_{1i}; Y_i, Y_{1i}|m_1) = I(X_{1i}; Y_{1i}|m_1) \) and
with the analysis similar to the one in the Section 4.2.1.3, the bound of Theorem 4.2.1
reduces to the expression of \( C \) in (4.87).

### 4.2.3 Achievability of \( C \) in Theorem 4.2.2

In this section, we will show the achievability of the rate \( R \leq C \) in the Theorem 4.2.2
for any \( t, 0 \leq t \leq 1 \). The achievability proof is based on a coding scheme similar to
the coding scheme in Theorem 1 of [26], which is restated below:
Theorem 4.2.3. [26] The capacity of the degraded cheap relay channel is given by

\[ C = \sup_{t, p(\cdot)} \min \{ t I(X_1; Y_1|m_1) + (1 - t) I(X_1; Y|X_2,m_2), \\
 t I(X_1; Y_0|m_1) + (1 - t) I(X_1, X_2; Y|m_2) \} \]  

(4.88)

where the supremum is taken over \( t, 0 \leq t \leq 1 \) and all the joint distributions \( p(x_0, x_1, x_2) \) on \( X_0 \times X_1 \times X_2 \).

We first begin with a brief sketch of the coding scheme and then give the random code that achieves it. We consider a source \( U = (W, V) \) which consists of two independent message sources \( W, V \) to be transmitted from the source node S to the destination node D. First, the network will be used \( t_1 \) times in the mode \( m_1 \), in which the source node S transmits \( v \) from the message \( u = (w, v) \) to the relay node R. We assume that this rate is too high for the destination node D to allow reliable decoding. Then, the network will be used \( t_2 \) times in mode \( m_2 \), in which both the source node S and relay node R cooperate to resolve the uncertainty of message \( v \) at the receiver of the destination node D. Also, the source node S will transmit the other part of the message \( u = (w, v) \), i.e. message \( w \in W \), to the destination during the use of the channel in this mode. In other words we consider the message source pair \( (U, V) \), where the input of the relay node R will be determined by the message source \( V \) and the input of the source node S will be determined by message source \( U \). The destination node D will collect all of the received information in the \( t_1 + t_2 \) uses of the network about the message \( u = (w, v) \) and then perform the decoding.
For some $\rho, 0 \leq \rho \leq 1$, let $X_U \sim N(0, (1 - \rho^2)P_0)$, $X_V \sim N(0, P_2)$, with $U$ and $V$ independent. Also, let $X_1|m_1 = \sqrt{P_0/P_2}X_V$, $X_1|m_2 = \sqrt{\rho^2P_1/P_2}X_V + X_U$, and $X_2|m_2 = X_V$. Referring to the Theorem 4.2.3, we evaluate
\[
I(X_1; Y_1|m_1) = \frac{1}{2} \log(1 + \gamma_0 P_0)
\]
\[
I(X_1; Y|X_2, m_2) = \frac{1}{2} \log(1 + (1 - \rho^2)\gamma_1 P_1)
\]
\[
I(X_1; Y_0|m_1) = \frac{1}{2} \log(1 + \gamma_1 P_0)
\]
\[
I(X_1, X_2; Y|m_2) = \frac{1}{2} \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\rho\sqrt{\gamma_1\gamma_2 P_1 P_2})
\]
The assertion that this distribution for $X_1|m_1$, $X_1|m_2$, and $X_2|m_2$ actually achieve the capacity $C$ in Equation (4.88) follows from the proof of the converse. The random codebook associated with this distribution is then given by the random choice of
\[
\begin{align*}
X_U(u) \text{ i.i.d. } &\sim N_n(0, (1 - \rho^2)P_0 I) & u \in [1, 2^{nH(U)}] \\
X_V(v) \text{ i.i.d. } &\sim N_n(0, P_2 I) & v \in [1, 2^{nH(V)}]
\end{align*}
\]
(4.89)
where $N_n(0, C_n)$ denotes the $n$-variate normal distribution with zero mean and covariance matrix $C_n$. The codebook is given by
\[
\begin{align*}
X_1|m_1(v) &= \sqrt{\frac{P_0}{P_2}} X_V(v) \\
X_1|m_2(u|v) &= \sqrt{\frac{\rho^2P_1}{P_2}} X_V(v) + X_U(u) \\
X_2|m_2(v) &= X_V(v)
\end{align*}
\]
(4.90)
The generated codewords would satisfy the power constraints with high probability $(1 - \epsilon)$, and thus the overall average probability of error can be shown to be small.
4.2.4 Concluding remarks

We derived the capacity of the Gaussian degraded cheap relay channel as stated in Theorem 4.2.2. Because in the proof of the achievability we do not use the degraded assumption, this capacity can be considered as a lower bound for the capacity of Gaussian cheap relay channels in general. We also derived an upper bound for the capacity of the Gaussian cheap relay channel (Theorem 4.2.1). The assumption of cheap relay nodes in the network is important for the design of the practical systems, and the results of this section characterize the limits of information transfer in such networks.

An interpretation for the states of the cheap relay channel is that mode $m_1$ resembles a broadcast channel from source node S to the relay node R and Destination node D and that mode $m_2$ resembles a multiple access channel from source node S and relay node R to the destination node D. This interpretation might be considered in the capacity formulation of the Theorem 4.2.2 as being $t$ portion of the times in the broadcast mode and $(1 - t)$ portion of the time in the multiple access mode. The latter interpretation is misleading, because while the multiple access part of this capacity expression coincide with the capacity of the multiple access channel with partially known inputs [35], the broadcast part does not agree with the capacity of the broadcast channel. In fact the relay network coding of Section 4.2.3 allows us to achieve a higher rate, which is equal to the upper bound for the capacity in the case
of a degraded cheap relay channel.

Although the capacity of a cheap relay channel is less than the capacity of the conventional relay channel (where the relay node can send and receive at the same time), the lower bound on this capacity shows that it can provide a higher capacity with respect to the direct link. Thus, the conclusion is that even though the relay nodes in the network are cheap, we can achieve a higher transmission rate by exploiting network coding over the transmitter and the relay nodes.
Chapter 5

New Capacity Results for the Relay Channel

In this section, we revisit the problem of communication between a single source-destination pair using a cheap relay that cannot send and receive in the same frequency band simultaneously. Since its introduction, the cheap relay channel has been considered as a channel either with time division relaying or frequency division relaying. In this section, we show that the treatment of a cheap relay channel in [26] as a channel with multiple states provides a unifying theory for the thereafter works on the capacity of the relay channel under cheap relay constraint, specifically under frequency division and time division relaying. Moreover, we present three non-degraded cases of the general relay channels for which the coding scheme [26] developed in Chapter 4 is used to characterize the channel capacity. These capacity results cannot be deduced from Cover and El Gamal’s work on the general relay channel [7] and are the only capacity results known to date for the non-degraded discrete memoryless relay channel besides that of the semi-deterministic relay channel [49].

5.1 Introduction

As previously discussed in Chapter 4, in current wireless networks, most radios operate in time division mode (TDD) when the transmitting and receiving frequencies
are the same. This constraint results because building RF radios that are capable of receiving and transmitting simultaneously in the same frequency band require precise and expensive components and elegant design. This practical limitation differentiates an expensive relay, which can simultaneously send and receive in the same frequency band, from a cheap relay, which is not able to do so. Therefore, the coding technique and theoretical limits of the expensive relay channel [6, 7] are not applicable or insightful in the cheap relay channel problem.

The contribution of this section is twofold. First, based on the result of Chapter 4, we provide a unifying theory for driving the capacity of the relay channel under the practical constraint of cheap relaying, specifically frequency division and time division relaying. Second, we present two cases of general relay channel for which the channel is not degraded and the capacity is not yet known. We derive the capacity of these two cases based on the coding scheme of [26], which directly improves upon the best known results for the relay channel given by Cover and El Gamal [7] for these cases. These new results are the only known capacity results for the non-degraded discrete memoryless relay channel besides that of the semi-deterministic relay channel [49]. It is worth mentioning that our capacity derivation is based on enlarging the achievable rate enough for some special cases to coincide with the cut-set upper bound [7, 44]. However, still no general upper bound better than the cut-set upper bound [8] for the expensive relay channel and that of [44] for the cheap relay channel is known to the authors.
Recently, the problem of cheap relaying has drawn a great deal of attention [44, 50, 51, 52]. Zahedi and El Gamal have considered two different cases of frequency-division Gaussian relay channel and derived lower and upper bounds on the capacity, which in turn translates to upper and lower bounds on the minimum required energy per bit for the reliable transmission [51]. For the same model, Zahedi and El Gamal have looked at the problem of linear relaying [53] similar to the linear relaying approach of [54]. Upper and lower bounds for a different model of frequency division relaying have also been derived by Liang and Veeravalli [52]. The problem of time division relaying has been considered by Host-Madsen and Zhang [50, 55] and by the authors in [44, 26, 27, 28]. An extension of the problem of time division relaying to multiple antenna systems can also be found in [56].

Besides the capacity analysis of the relay channel with time and frequency division relaying, we have extended the concept of a network with a cheap node to other problems and different types of a network. The existence of cheap relays in a network can be easily modelled as multiple states in the network, where in each state the received signal at each node of the network only depends on the transmitted signals in the same state. Further treatment of a general network with multiple states due to cheap relaying can be found in [44, 26, 27]. An example of such a network is the broadcast channel with a cooperative decoder considered by Liang and Veeravalli [57].

The rest of this section is organized as follows. After some preliminary definitions in Section 5.2, we formally define the cheap relay channel in Section 5.3. The main
result of the section, which is a unifying approach to the capacity analysis of time and frequency division relaying, is presented in Section 5.4. We also present some new capacity results for the general discrete memoryless relay channel. Finally, we conclude in Section 5.5.

5.2 Expensive Relay Channel

The relay channel (with expensive relaying) was first introduced and studied by Van der Mullen [6] and later considered by Cover and El Gamal [7]. A relay channel, shown in Figure 5.1(a), is a three terminal channel consisting of an input $x_1$ that is transmitted from the source, a signal $y_1$ received with the relay, a signal $x_2$ (which depends only on the past values of $y_1$) transmitted by the relay, and a channel output $y$ that is received by the destination. The goal in a relay channel is to communicate a message from the source node to the destination node by the help of the relay node. The channel is assumed to be memoryless and the received signals $y$ and $y_1$ are obtained from the inputs $x_1$ and $x_2$ through channel transition function $p(y,y_1|x_1,x_2)$. It should be pointed out that the transmitted or received signals can be vectors, which allows us to treat the most general definition of the relay channel including multiple antenna relay channel and frequency division relay channel. For example, consider a channel for which the destination received vector $y$ and the source transmitted vector $x_1$ are given by $y = (y_b, y_m)$ and $x_1 = (x_{1b}, x_{1m})$, respectively, where the channel transition function is given by
Figure 5.1: (a) Expensive relay channel, (b) Frequency division relay channel - Model I. The solid and dashed lines represent two orthogonal channels.

\[ p(y, y_1 | x_1, x_2) = p(y_b, y_m | x_{1b}, x_{1m}, x_{2m}) = p(y_b, y_{1b} | x_{1b})p(y_m | x_{1m}, x_{2m}). \]

Note that \( y_b, y_m, x_{1b}, \) and \( x_{1m} \) can also be vectors themselves. This channel can be realized as it is depicted in Figure 5.1(b) and can be interpreted as a discrete memoryless counterpart of a frequency division relay channel.

### 5.3 ‘Cheap’ Relay Channel

In its introduction in [44, 26], a *cheap* node is defined as a node which is either a sender or a receiver for any channel use. Also, a state of operation in a network with cheap nodes corresponds to the valid partitioning of the nodes into two disjoint subsets of sender and receiver nodes. However, the above notion of state in the network can be easily generalized as follows.

**Definition 5.3.1.** A set of states in a network is a disjoint partitioning of the input space (transmitted signal space) where the channel outputs (received signals) at each state depend only on the channel inputs in the associated state.

Clearly, Definition 5.3.1 covers the aforementioned notion of state in the network.
due to the time division relaying. Moreover, Definition 5.3.1 also enables us to define states for some channels where the previous definition is not applicable. Examples of such a channel is given in Section 5.3.2.

5.3.1 Time Division Relay Channels

In this section we consider a general model of a discrete memoryless relay channel for which the relay node operates in time division mode to transmit and receive information. This general model enables us to easily derive the corresponding results for the case of the Gaussian channel or multiple antenna systems.

Let \( p(y, y_1|x_1, x_2) \) denote the channel transition function of the original relay channel. For a time division relaying, the relay node is either a transmitter or a receiver. Therefore, there are two possible states of operation in the network. In the state \( b \) ("broadcast state"), the relay node acts as a receiver, and thus the channel transition function is given by \( p(y, y_1|x_1, s = b) = p(y, y_1|x_1, x_2 = 0) \), where \( s \) denotes the state variable. Also, in the state \( m \) ("Multiple access state"), the relay node functions as a transmitter, and the channel probability function is given by \( p(y|x_1, x_2, s = m) = \sum_{y_1} p(y, y_1|x_1, x_2) \). We use \( x_2 = 0 \) to denote the case that there is no input from the relay node to the channel in state \( b \). This notation is specifically natural considering the fact that \( x_2 = 0 \) corresponds to no transmitted signal in Gaussian channels. It should be pointed out that we use the terms "broadcast state" and "multiple access state" because of the resemblance of the network in these states
to broadcast and multiple access channels. However, to achieve the full capacity of
the channel, the coding schemes may not be divided into two separate coding scheme
one for each individual state.

5.3.2 Frequency Division Relay Channels

Similar to Section 5.3.1, we consider a general model of a discrete memoryless relay
channel for which the relay node operates in two different parallel channels to transmit
and receive information. This general model in fact enables us to easily derive the
corresponding results for the frequency division relaying case of the Gaussian channel.
Admitting the fact that frequency cannot be defined for the discrete memoryless
channels, by an abuse of name convention, we call the channel model of this section
as frequency division relay channel model. In fact, a frequency division relay channel
can be realized in three different ways (Figure 5.1(b), Figure 5.2(a), and Figure 5.2(b))
which are all special cases of the model presented in this section.

Let the channel transition function of the original channel be given as
\[ p(y, y_1|x_1, x_2) = p(y_b, y_m|x_{1_b}, x_{1_m}, x_{2_m}) = p(y_b, y_{1_b}|x_{1_b})p(y_m|x_{1_m}, x_{2_m}). \] The interpretation is
that there are two parallel channels which define the channel transition function of
this relay channel. One channel is a broadcast channel from the source node to the
destination and relay nodes denoted by channel transition function \( p(y_b, y_{1_b}|x_{1_b}) \). The
other channel is a multiple access channel from the source and relay nodes to the des-
tination given by channel transition function \( p(y_m|x_{1_m}, x_{2_m}) \). We keep the vector
representation of the transmitted and received signals to allow for easy extension of the results to multiple antenna systems.

It should be noted that the frequency division relay channel has been considered in three different models by various authors as depicted in Figure 5.1(b) and Figure 5.2. However, models II and III in Figure 5.2 are special cases of the channel model I represented in Figure 5.1(b).

5.4 Unifying Approach to Time and Frequency Division Relaying

While the constraint of cheap relaying in general reduces the capacity of the relay channel, the coding schemes for this channel have contributed to better achievable rates even for the original relay channel with an expensive relay. A simple example of such a contribution is the linear relaying strategy developed in [26], where the achievable rate of a cheap relay scheme, namely the scale-and-forward scheme, surpasses the
rates of all the known schemes for some channel conditions. In this section, we pro-
vide three more cases (Cases 5 through 7 in Section 5.4.2) where a coding technique
developed originally for the cheap relay channel [26] achieves a rate higher than the
achievable rate of all the other known schemes [7] even though they are not subject
to the cheap relaying restriction.

In this Section, we first revisit the problem of the cheap relay channel and provide
upper and lower bounds on the capacity (Section 5.4.1). We then show how this
result subsumes all the cases of the frequency division relay channel [51, 53, 52] and
the time division relay channel [26, 50, 55]. Later, we present new capacity results
for the relay channel in Section 5.4.2.

5.4.1 Bounds on the Capacity of Cheap Relay Channel

The new definition of state in Section 5.3 clarifies that the outputs of the channel
that has the inputs in the same states are independent of the channel inputs in the
other states. Therefore, a cheap relay channel in general can be denoted as two
possible states $b$ and $m$ governed by two channel transition functions $p(y_b, y_{1b}|x_{1b})$
and $p(y_m|x_{1m}, x_{2m})$, respectively. However, a cheap relay channel is not a compound
channel, thus, we may allow the two states to be used in parallel fashion.

For example, let $t_b$ and $t_m$ denote the fraction of time that the states $b$ and $m$
can be used. Therefore, we must have $t_b + t_m = 1$ for the time division relay channel.
However, with the same channel transition function model as above, we can have
\( t_b = t_m = 1 \) for the frequency division relay channel because the states can be used in parallel fashion. Fortunately, neither the coding theorem of [26] nor the multi-state cut-set bound of [44] relies on the values of \( t_b \) and \( t_m \), and they only exploit the fact that the outputs in each state are only stochastic functions of the inputs in the same state. Therefore, we have the following upper (from [44]) and lower (from [26]) bounds on the capacity of the cheap relay channel.

**Theorem 5.4.1.** Consider a cheap relay channel defined by two possible states \( b \) and \( m \) governed by two channel transition functions \( p(y_b, y_{1b}|x_{1b}) \) and \( p(y_m|x_{1m}, x_{2m}) \), respectively. A lower bound on the capacity is given by

\[
C_l \triangleq \sup_{(t_b, t_m) \in \mathcal{T}, \ p(x_{1b})p(x_{1m}, x_{2m})} \min\{t_b I(X_{1b}; Y_{1b}) + t_m I(X_{1m}; Y_m|X_{2m}),
\]

\[
t_b I(X_{1b}; Y_b) + t_m I(X_{1m}, X_{2m}; Y_m)\},
\]

and an upper bound on the capacity is given by

\[
C_u \triangleq \sup_{(t_b, t_m) \in \mathcal{T}, \ p(x_{1b})p(x_{1m}, x_{2m})} \min\{t_b I(X_{1b}; Y_{1b}) + t_m I(X_{1m}; Y_m|X_{2m}),
\]

\[
t_b I(X_{1b}; Y_b) + t_m I(X_{1m}, X_{2m}; Y_m)\},
\]

where the supremum has taken over all joint distributions \( p(x_{1b})p(x_{1m}, x_{2m}) \) on \( \mathcal{X}_{1b} \times \mathcal{X}_{1m} \times \mathcal{X}_{2m} \), and over all possible time fractions \((t_b, t_m) \in \mathcal{T}\), where the set \( \mathcal{T} \subset [0, 1] \times [0, 1] \) is defined by the channel model.

Based on the notion of cheapness presented in this section, the bound of Theorem 5.4.1 can be used to provide respective bounds for the time division and various
cases of a frequency division relay channel. We state the following corollary for the sake of completeness and easy comparison between different cases of a cheap relay channel, i.e., frequency division and time division relay channels.

**Corollary 5.4.2.** Consider a cheap relay channel defined by two possible states $b$ and $m$ governed by two channel transition functions $p(y_b, y_{1b}|x_{1b})$ and $p(y_m|x_{1m}, x_{2m})$, respectively. Lower and upper bounds on the capacity for different cases of time division (TDD) and frequency division (FDD) relay channels are given as follows.

For time division relaying [26]:

$$C^\text{TDD}_l \triangleq \sup_{0 \leq t \leq 1} \min \{ t \, I(X_{1b}; Y_{1b}) + (1-t) \, I(X_{1m}; Y_m | X_{2m}), \\ t \, I(X_{1b}; Y_b) + (1-t) \, I(X_{1m}, X_{2m}; Y_m) \}, \quad (5.3)$$

$$C^\text{TDD}_u \triangleq \sup_{0 \leq t \leq 1} \min \{ t \, I(X_{1b}; Y_{1b}) + (1-t) \, I(X_{1m}; Y_m | X_{2m}), \\ t \, I(X_{1b}; Y_b) + (1-t) \, I(X_{1m}, X_{2m}; Y_m) \}, \quad (5.4)$$

For frequency division relaying depicted in Figure 5.1(b):

$$C^\text{FDD}_l \triangleq \sup_{p(x_{1b})p(x_{1m}, x_{2m})} \min \{ I(X_{1b}; Y_{1b}) + I(X_{1m}; Y_m | X_{2m}), \\ I(X_{1b}; Y_b) + I(X_{1m}, X_{2m}; Y_m) \}, \quad (5.5)$$

$$C^\text{FDD}_u \triangleq \sup_{p(x_{1b})p(x_{1m}, x_{2m})} \min \{ I(X_{1b}; Y_{1b}) + I(X_{1m}; Y_m | X_{2m}), \\ I(X_{1b}; Y_b) + I(X_{1m}, X_{2m}; Y_m) \}, \quad (5.6)$$
where the supremum has taken over all joint distributions $p(x_{1b})p(x_{1m}, x_{2m})$ on $\mathcal{X}_{1b} \times \mathcal{X}_{1m} \times \mathcal{X}_{2m}$.

The bounds of Corollary 5.4.2 can be easily computed for the Gaussian channels acknowledging the fact that the same coding scheme can be simply extended for the case of the Gaussian channel with continuous input signals. These bounds for the time division relaying are also derived in [28, 50, 55]. However, the bounds for the frequency division relaying obtained in Corollary 5.4.2 directly improve upon the corresponding bounds derived in [52]. This improvement is not surprising, because the heart of the result obtained in [52] is Cover and El Gamal’s bounds (Theorem 1 in [7]).

However, the coding scheme presented in Chapter 4 is tailored for the cheap relay channel, and it also provides absolutely higher achievable rate for some channel conditions where both coding schemes are applicable. The relay channel cases for which the rate achieved by the coding scheme of Chapter 4 surpasses the achievable rate of Cover and El Gamal coding scheme in [7] will be discussed in greater details in Section 5.4.2.

Figure 5.2(a) and (b) show more restricted cases of frequency division relaying, where either the multiple access state or broadcast state is solely used as a single link channel between relay-destination or source-relay, respectively. Zahedi and El Gamal [51] and also Liang and Veeravalli [52] have found upper and lower bounds for the Gaussian case of the frequency division relay channel models depicted in Figure 5.2(a) and (b) which coincide with the bounds of Corollary 5.4.2 evaluated for
the Gaussian channel. However, since transmission from the source in each of the
frequency bands (or channel states) can always be heard by the destination, it should
be possible to improve the achievable rate of the transmission. It should be pointed
out that if the destination node uses all the received signals in both frequency bands
as depicted in Figure 5.1(b), the bounds of Corollary 5.4.2 directly improve upon all
the existing results.

5.4.2 New Capacity results for discrete memoryless relay channel

Deriving the capacity of the relay channel has been a challenge for the last 30 years.
However, except for some special cases, no capacity result has been derived for the
general relay channel. Current capacity results for the general discrete memoryless
relay channel are restricted to the following three results.

Case 1: (Degraded relay channel [7]) If the channel transition function can
be written in the form of \( p(y, y_1|x_1, x_2) = p(y_1|x_1, x_2)p(y|y_1, x_2) \), then the capacity of
the channel is given by

\[
C_1 = \sup_{p(x_1, x_2)} \min\{I(X_1; Y_1 | X_2), I(X_1, X_2; Y)\}. \tag{5.7}
\]

Case 2: For a general \( p(y, y_1|x_1, x_2) \), if \( I(X_1; Y_1 | X_2) \geq I(X_1, X_2; Y) \) for all
\( p(x_1, x_2) \), then the capacity of the channel is given by [7]

\[
C_2 = \sup_{p(x_1, x_2)} I(X_1, X_2; Y). \tag{5.8}
\]
Case 3: (Reversely Degraded relay channel [7]) If the channel transition function can be written in the form of $p(y, y_1|x_1, x_2) = p(y|x_1, x_2)p(y_1|y, x_2)$ then the capacity of the channel is given by

$$C_3 = \sup_{x_2 \in X_2, p(x_1)} I(X_1; Y_1|x_2). \quad (5.9)$$

Case 4: (Semi-deterministic relay channel [49]) If the relay received signal is a deterministic function of the transmitted signals as $y_1 = f(x_1, x_2)$, i.e., the channel transition function can be written in the form of $p(y, y_1|x_1, x_2) = p(y|x_1, x_2), I(y_1 = f(x_1, x_2))$, then the capacity of the channel is given by

$$C_4 = \sup_{p(x_1, x_2)} \min \{H(Y_1|X_2) + I(X_1; Y|X_2, Y_1), I(X_1, X_2; Y)\}. \quad (5.10)$$

In this section, we present three more cases for which the capacity can be exactly characterized based on the coding scheme presented in [26]. These results are obtained directly from Theorem 5.4.1, where the upper and lower bounds coincide.

Case 5: If the channel transition function can be written in the form of $p(y, y_1|x_1, x_2) = p(y, y_1|(x_{1b}, x_{1m}), x_2) = p(y_1|x_{1b})p(y|x_{1m}, x_2)$, then the capacity of the channel is given by

$$C_5 = \sup_{p(x_{1b})p(x_{1m}, x_2)} \min \{I(X_{1b}; Y_1) + I(X_{1m}; Y|X_2), I(X_{1m}, X_2; Y)\}. \quad (5.11)$$

Case 6: If the channel transition function can be written in the form of $p(y, y_1|x_1, x_2) = p((y_b, y_m), y_{1b}|(x_{1b}, x_{1m}), x_{2m}) = p(y_b|y_{1b})p(y_{1b}|x_{1b})p(y_m|x_{1m}, x_{2m}),$ then the
capacity of the channel is given by

$$C_6 = \sup_{p(x_{1b})p(x_{1m},x_{2m})} \min \{ I(X_{1b};Y_{1b}) + I(X_{1m};Y_m|X_{2m}),$$

$$I(X_{1b};Y_b) + I(X_{1m},X_{2m};Y_m) \}.$$  \(5.12\)

**Case 7:** If the channel transition function can be written in the form of

$$p(y, y_1| x_1, x_2) = p(y_b, y_m|x_{1b}, x_{1m}, x_{2m}) = p(y_b, y_{1b}|x_{1b})p(y_m|x_{1m}, x_{2m}),$$

then the capacity of the channel is given by

$$C_7 = \sup_{p(x_{1b})p(x_{1m}, x_{2m})} I(X_{1b};Y_b) + I(X_{1m}, X_{2m};Y_m), \quad (5.13)$$

provided that $I(X_{1b};Y_{1b}) + I(X_{1m};Y_m|X_{2m}) \geq I(X_{1b};Y_b) + I(X_{1m}, X_{2m};Y_m)$ for all input distributions of $p(x_{1b})p(x_{1m}, x_{2m})$.

Here, a few notes are in order. First, the channel in Figure 5.2(b) can be shown to satisfy the condition of Case 5. Also, Cases 6 and 7 can be realized as special cases of Figure 5.1(b). Second, we emphasize the fact that the channel defined in any of Cases 5 through 7 can be considered in the context of the original expensive relay channel without any specific treatment. Therefore, the coding scheme of the original relay channel in [7] is also directly applicable to these channels without any specific treatments. However, the achievable rate obtained from the coding scheme of [7] is less than the capacity of these channels. For example the achievable rate of Cover and El Gamal’s [7] scheme for the relay channel in Case 5 is given by

$$R_5 = \sup_{p(x_{1b})p(x_{1m}, x_{2})} \min \{ I(X_{1b};Y_1), I(X_{1m}, X_2;Y) \}. \quad (5.14)$$
Third, the coding scheme developed in [26] not only is tailored for the cheap relay channel, but it also can contribute to the expensive relay channel problem [7] by providing a better rate than all the other known schemes. In fact, Cases 5 through 7 show that the coding scheme of [26] is even optimal for some channel conditions. Fourth, it should be noted that none of Cases 5 through 7 are degraded channels. Specifically, we note that the received signal at the destination for Case 6 is directly a function of the source input, even if the received signal at the relay is given. Therefore, the channel is not a degraded channel. By looking at the channel conditions, it is clear that the other two cases are also not degraded channels. Fifth, it can be seen that Case 5 can be easily derived from Case 6. However, Case 6 is more general than Case 5 and it cannot be realized in either of the frequency division relaying models of Figure 5.2. Sixth, the capacity result of Case 5 for the Gaussian channel is also indirectly reported in [51] by Zahedi and El Gamal, where they have used the capacity expression to obtain the minimum energy per bit required for reliable transmission.

5.5 Concluding remarks

We presented a unifying approach to time and frequency division relaying, where we derived lower and upper bounds based on the coding scheme developed in [26] for the cheap relay channel and also the cut-set bound for multi-state networks developed in [44]. These bounds improve upon the existing bounds for the frequency division relaying for many channel conditions. A notable result of this section is obtained
by using the derived upper and lower bounds to characterize the capacity for a non-degraded relay channel (Case 6 in Section 5.4.2) where no prior coding scheme is known that can achieve the capacity of such a channel.
Chapter 6
Fading Relay Channel

Deriving the capacity of the relay channel has been a challenge for the last 30 years. Except for some trivial or degenerate cases no capacity result has been derived for the relay channel. In this section, we consider a very practical case, the Rayleigh fading relay channel. The contributions of this section are twofold. First, we derive the exact ergodic capacity of the Rayleigh fading relay channel for a range of channel conditions which include the most interesting and practical relaying cases. An illustrative example is considered where the average fade is inversely proportional to the distance. It is shown that if the relay lies in a relatively large region surrounding the source node, the exact capacity is known. Second, we show that the estimate-and-forward coding scheme is more robust to the variation of fading parameters. Furthermore, this scheme performs relatively close to the upper bound on the capacity even for the cases where we do not know the exact capacity. Numerical results are presented to quantify the capacity gain when using the relay over direct transmission. Also, it is shown for which regimes the proposed scheme is the best-known coding scheme for the Rayleigh fading relay channel.
6.1 Introduction

Recently, there has been a great interest in network coding and its potential to increase the overall throughput of a network. Cooperative coding, where nodes collaborate with each other to improve their data rates, has shown its potential [58] and attracted the main research focus in network coding. The challenge in network coding comes from the fact that none of the current schemes truly extract the potential benefits of cooperating nodes. The above fact is evident even in the simplest network with one transmitter-receiver pair being helped by one relay node (Figure 6.1), commonly known as the relay channel [7]. Surprisingly, aside from some trivial or degenerate cases [7, 59, 60], no capacity result has been derived even for this simplest form of the networks.

In this section, we derive the capacity of the Rayleigh fading relay channel which is especially of interest because of its practical importance in wireless networks. The contributions of this section are twofold. First, we derive the exact ergodic capacity of the Rayleigh fading relay channel for a range of channel conditions which include the most interesting and practical relaying cases. We consider an illustrative example scenario where the average fades of the source-relay, source-destination, and relay-destination links are inversely proportional to their mutual distances, respectively. We also take into account the path loss exponent which ranges between 2 to 4 for practical wireless communications. Our results demonstrate the exact capacity for a relatively large region surrounding the source node. Second, we show that the
estimate-and-forward [54, 38] coding scheme is more robust to the changes in average fades of the links and performs relatively close to the upper bound on the capacity even for the cases that we do not know the exact capacity. Therefore, aside from the analytical beauty of the first coding scheme which achieves the exact capacity of the Rayleigh fading relay channel, the latter coding technique seems to be of more practical interest because of its robustness.

The rest of the section is organized as follows. After some preliminaries in Section 6.2, we briefly discuss an upper bound in Section 6.4.1. In Section 6.4.2, we consider two lower bounds on the capacity of the fading relay channel. The main result of this section, the capacity of the fading relay channel, is derived in Section 6.4.3. Finally, after some numerical results in Section 6.5, we conclude in Section 6.6.

6.2 Preliminaries

Consider the fading relay channel of Figure 6.1, in which the source node S intends to transmit information to the destination node D by using the direct link between the node pair (S, D) as well as the help of another relay node R (if it improves the achievable rate of transmission) by using link pairs (S, R) and (R, D). A fading relay channel, shown in Figure 6.1, consists of a channel input \( x_1 \), a relay output \( y_1 \), a relay signal \( x_2 \) (which depends only on the past values of \( y_1 \)), and a channel output \( y \). The channel is assumed to be memoryless. The dependency of the outputs on the inputs are as follows: the channel output is \( y = h_1 x_1 + h_2 x_2 + z \), and the relay output is given
by $y_1 = h_0 x_1 + z_1$. The fading coefficients $h_0$, $h_1$, and $h_2$ are assumed to be independent random variables each with zero mean complex Gaussian distribution with variance values $\lambda_0$, $\lambda_1$, and $\lambda_2$, respectively. The additive white Gaussian noises $z \sim N(0, N)$ and $z_1 \sim N(0, N_1)$ are assumed to be mutually independent, and also independent of the fading process. The input power constraints are given by $\mathbb{E}[X_1^2] \leq P_1$ and $\mathbb{E}[X_2^2] \leq P_2$. The problem is to find the ergodic capacity of the channel between the sender $S$ and receiver $D$.

6.3 Complex Gaussian Relay Channel

6.3.1 Upper Bound on the Capacity of the Complex Gaussian Relay Channel

**Lemma 6.3.1.** Consider a complex Gaussian relay channel, which is a special case of fading relay channel in Figure 6.1 where the channel state $\mathbf{h} = (h_0, h_1, h_2)$ is constant.
An upper bound on the capacity is then given by

\[ R \leq \max_{\rho, 0 \leq |\rho| \leq 1} \min \{ \log(1 + (1 - |\rho|^2)(\gamma_0 + \gamma_1)P_1), \]

\[ \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2Re(\rho e^{j\angle h_1-j\angle h_2})\sqrt{\frac{\gamma_1 \gamma_2 P_1 P_2}}) \} \tag{6.1} \]

where \( \gamma_0 \triangleq \frac{|h_0|^2}{N_1}, \gamma_1 \triangleq \frac{|h_1|^2}{N} \) and \( \gamma_2 \triangleq \frac{|h_2|^2}{N} \). Furthermore, the constant \( \rho \) in general is a complex value and a function of the channel state \( \mathbf{h} = (h_0, h_1, h_2) \) at the transmitters, i.e., \( \rho = \rho(\mathbf{h}) \). However, if the value of the channel state \( \mathbf{h} = (h_0, h_1, h_2) \) is not known at the transmitters, then transmitters cannot align their received signal at the destination and therefore the optimal value of \( \rho \) cannot be found at the transmitters. Therefore, without knowledge of \( \mathbf{h} = (h_0, h_1, h_2) \) at the transmitter, the upper bound is obtained for a fixed value of \( \rho \) which is given by the coding scheme.

**Proof:** Before we give the details of the proof, we note that if the channel state \( \mathbf{h} = (h_0, h_1, h_2) \) is known to all nodes, then transmitters can perfectly align their signals at the receiver of the destination node. More specifically, we can completely ignore all the phase’s values of the channel coefficients \( h_1, h_2, h_0 \) by multiplying the transmitted signals \( X_1, X_2 \) by the values \( e^{-j\angle h_1}, e^{-j\angle h_2} \) and by multiplying the relay received signal \( Y_1 \) by the value \( e^{j\angle h_1-j\angle h_0} \). Also, \( \rho = \rho(\mathbf{h}) \) can be optimally found, therefore the upper bound is given by

\[ R \leq \max_{\rho, 0 \leq |\rho| \leq 1} \min \{ \log(1 + (1 - |\rho|^2)(\gamma_0 + \gamma_1)P_1), \]

\[ \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \} \tag{6.2} \]
where $\rho$ is a real number.

To prove the upper bound of Lemma 6.3.1 we use the cut-set lemma in [8]. It can be easily shown from Theorem 14.10.1 in [8] that the maximum achievable rate for the complex Gaussian relay channel is bounded by the bounds across two cut-set as follows.

\begin{equation}
C \leq I(X_1; Y_1(h), Y(h)|X_2), \quad \text{Broadcast cut-set} \quad (6.3)
\end{equation}

and

\begin{equation}
C \leq I(X_1, X_2; Y(h)). \quad \text{Multiple-access cut-set} \quad (6.4)
\end{equation}

Consider $n$ consecutive channel uses and let index $i = 1, 2, \ldots, n$ denote the time index of the channel uses. For the multiple access bound of (6.4), i.e., $I(X_1, X_2; Y(h))$, we have

\begin{equation}
C \leq \frac{1}{n} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; Y_i(h)) \quad (6.5)
\end{equation}

\begin{align*}
&\leq \frac{1}{n} \sum_{i=1}^{n} \{H(Y(h)) - H(Y(h)|X_1, X_2)\} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \{H(Y(h)) - \log(2\pi eN)\}
\end{align*}
where

\[ H(\mathbf{Y}(h)) \quad (6.6) \]

\[ \leq E[\log((2\pi e) \text{Var}(\mathbf{Y}(h)))] \]
\[ \leq \log((2\pi e) E[\text{Var}(\mathbf{Y}(h))]) \]
\[ \leq \log((2\pi e) E[\mathbf{Y}(h)\mathbf{Y}^*(h)]) \]
\[ \leq \log \left( (2\pi e) \left[ E[|X_1|^2] + E[|X_2|^2] + 2\text{Re}(h_1 h_2^* E[X_1 X_2^*]) + N \right] \right) \]

Combining Equations (6.5) and (6.6) and using Jensen’s inequality we obtain

\[ C \leq \frac{1}{n} \sum_{i=1}^{n} \{ H(\mathbf{Y}(h)) \} - \log(2\pi eN) \quad (6.7) \]

\[ C \leq \frac{1}{n} \sum_{i=1}^{n} \log \left( 1 + \frac{|h_1|^2}{N} E[|X_1|^2] + \frac{|h_2|^2}{N} E[|X_2|^2] + 2\text{Re}(\frac{h_1 h_2^*}{N} E[X_1 X_2^*]) \right) \]
\[ C \leq \log \left( 1 + \frac{|h_1|^2}{N} \frac{1}{n} \sum_{i=1}^{n} E[|X_1|^2] + \frac{|h_2|^2}{N} \frac{1}{n} \sum_{i=1}^{n} E[|X_2|^2] + 2\text{Re}(\frac{h_1 h_2^*}{N} \frac{1}{n} \sum_{i=1}^{n} E[X_1 X_2^*]) \right) \]

Therefore, by using the definition of channel quality parameters \( \gamma_1 \) and \( \gamma_2 \), and the input power constraint \( P_1 \) and \( P_2 \), we have

\[ C \leq \log \left( 1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\text{Re}(\sqrt{\gamma_1 \gamma_2} e^{j\delta_1} - j\delta_2 \rho \sqrt{P_1 P_2}) \right) \quad (6.8) \]

where \( \rho \) is defined as the correlation factor between the input signals as

\[ \rho \triangleq \frac{1}{n} \sum_{i=1}^{n} \frac{E[X_1 X_2^*]}{\sqrt{P_1 P_2}} \quad (6.9) \]

For the broadcast bound of (6.3), i.e., \( I(X_1; Y_1(h), Y(h)|X_2) \), again we consider
n consecutive channel uses. Let index \( i = 1, 2, \ldots, n \) denote the time index of the channel uses, we have

\[
C \leq \frac{1}{n} \sum_{i=1}^{n} I(X_{ii}; Y_{i(h)}, Y_i(h) | X_{2i})
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \{ \mathcal{H}(Y(h), Y_1(h) | X_2) - \mathcal{H}(Y(h) | X_1, X_2) - \mathcal{H}(Y_1(h) | Y(h), X_1, X_2) \}
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \{ \mathcal{H}(Y(h) | X_2) - \log(2\pi eN) - \log(2\pi eN_1) \}
\]

where \( \mathcal{H}(X) \) denotes the differential entropy of \( X \), and

\[
Y(h)|X_2 \triangleq \begin{bmatrix} Y(h)|X_2 \\ Y_1(h)|X_2 \end{bmatrix}
\]

is defined as the vector of received signals at the destination and relay nodes. Let

\[
\text{Var}(X) \triangleq |E[XX^*] - E[X]E^*[X]|
\]

in general denote the determinant of the covariance matrix of \( X \) where \( X \) is a vector that is the variance of \( X \) where \( X \) is a scalar. For \( \mathcal{H}(Y(h)|X_2) \) we have
\[ H(Y(h)|X_2) = E[H(Y(h)|x_2)] \]  
\[ \leq E \left[ \log((2\pi e)^2 \text{Var}(Y(h)|x_2)) \right] \]  
\[ \leq \log((2\pi e)^2 E[\text{Var}(Y(h)|X_2)]) \]  
\[ \leq \log((2\pi e)^2 (E[\text{Var} \left( \begin{array}{c} Y(h|X_2) \\ Y_1(h|X_2) \end{array} \right)]) \]  
\[ \leq \log((2\pi e)^2 (E[\text{Var} \left( \begin{array}{c} h_1(X_1|X_2) + Z \\ h_0(X_1|X_2) + Z_1 \end{array} \right)]) \]  
\[ \leq \log((2\pi e)^2 \left| \begin{array}{cc} |h_1|^2 E[\text{Var}(X_1|X_2)] + N & h_1h_0^* E[\text{Var}(X_1|X_2)] \\ h_1^*h_0 E[\text{Var}(X_1|X_2)] & |h_0|^2 E[\text{Var}(X_1|X_2)] + N_1 \end{array} \right| \]  
\[ \leq \log \left( (2\pi e)^2 ((N_1|h_1|^2 + N|h_0|^2)E[\text{Var}(X_1|X_2)] + NN_1) \right) \]  
\[ \leq \log \left( (2\pi e)^2 ((N_1|h_1|^2 + N|h_0|^2)(E[|X_1|^2] - E[E[|X_1|X_2|^2]]) + NN_1) \right) \]  
\[ \leq \log \left( (2\pi e)^2 \left( (N_1|h_1|^2 + N|h_0|^2)(E[|X_1|^2] - \frac{E^2[|X_1X_2^*|]}{E[|X_2|^2]}) + NN_1 \right) \right) \]

Combining Equations (6.10) and (6.11) and using Jensen’s inequality we obtain

\[ C \leq \frac{1}{n} \sum_{i=1}^{n} \{H(Y(h)|X_2) - \log(2\pi eN) - \log(2\pi eN_1) \} \]  
\[ C \leq \frac{1}{n} \sum_{i=1}^{n} \{\log \left( 1 + \left( \frac{|h_1|^2}{N} + \frac{|h_0|^2}{N_1} \right) \left( E[|X_1|^2] - \frac{E^2[|X_1X_2^*|]}{E[|X_2|^2]} \right) \right) \} \]  
\[ C \leq \log \left( 1 + \left( \frac{|h_1|^2}{N} + \frac{|h_0|^2}{N_1} \right) \left( \frac{1}{n} \sum_{i=1}^{n} E[|X_1|^2] - \frac{1}{n} \sum_{i=1}^{n} \frac{E^2[|X_1X_2^*|]}{E[|X_2|^2]} \right) \right) \]

Therefore, by using the definition of channel quality parameters $\gamma_1$ and $\gamma_0$, the input
power constraint $P_1$, and the correlation factor $\rho$, we have

$$C \leq \log \left( 1 + (\gamma_1 + \gamma_0) \left( 1 - |\rho|^2 \right) P_1 \right)$$  \hfill (6.13)

Because both bounds of (6.13) and (6.8) should be satisfied, the minimum of these two is the upper bound which is given by

$$R \leq \max_{\rho, 0 \leq |\rho| \leq 1} \min \{ \log(1 + (1 - |\rho|^2)(\gamma_0 + \gamma_1)P_1),$$

$$\log(1 + \gamma_1 P_1 + \gamma_2 P_2 +$$

$$2\text{Re}(\rho e^{j\angle h_1 - j\angle h_2})\sqrt{\gamma_1 \gamma_2 P_1 P_2} \}$$ \hfill (6.14)

where one can optimize the bound by finding the appropriate value of $\rho$ if and only if knowledge of channel state $\mathbf{h} = (h_0, h_1, h_2)$ is available at the transmitter. Therefore, where the knowledge of the channel state $\mathbf{h} = (h_0, h_1, h_2)$ is not available at the transmitter, the upper bound on the achievable rate of any coding scheme is given by (6.1) for a constant $\rho$ defined in (6.9) which only depends on the correlation between the input signals. It should be noted that the critical information at the transmitter is only the difference between the phase of $h_1$ and $h_2$. This completes the proof of Lemma 6.3.1.

6.3.2 Lower Bound on the Capacity of the Complex Gaussian Relay Channel

Lemma 6.3.2. Consider a complex Gaussian relay channel, which is a special case of fading relay channel in Figure 6.1 where the channel state $\mathbf{h} = (h_0, h_1, h_2)$ is constant.
A lower bound on the capacity is then given by

\[
R \leq \max_{\rho, \, 0 \leq |\rho| \leq 1} \min \{ \log(1 + (1 - |\rho|^2)\gamma_0 P_1), \\
\log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\Re(\rho e^{j\angle h_1 - j\angle h_2}) \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \} 
\]

(6.15)

where \( \gamma_0 \triangleq \frac{|h_0|^2}{N}, \gamma_1 \triangleq \frac{|h_1|^2}{N}, \) and \( \gamma_2 \triangleq \frac{|h_2|^2}{N}. \) Furthermore, the constant \( \rho \) in general is a complex value and a function of the channel state \( \bar{h} = (h_0, h_1, h_2) \) at the transmitters, i.e., \( \rho = \rho(\bar{h}). \) However, if the value of the channel state \( \bar{h} = (h_0, h_1, h_2) \) is not known at the transmitters, then transmitters cannot align their received signal at the destination and therefore the optimal value of \( \rho \) cannot be found at the transmitters. Therefore, without knowledge of \( \bar{h} = (h_0, h_1, h_2) \) at the transmitter, the lower bound is obtained for a fixed value of \( \rho \) which is given by the coding scheme.

**Proof:** With the same argument stated in the proof of Lemma 6.3.1, we note that if the channel state \( \bar{h} = (h_0, h_1, h_2) \) is known to all nodes, then transmitters can perfectly align their signals at the receiver of the destination node. More specifically, we can completely ignore all the phase’s values of the channel coefficients \( h_1, h_2, h_0 \) by multiplying the transmitted signals \( X_1, X_2 \) by the values \( e^{-j\angle h_1}, e^{-j\angle h_2} \) and by multiplying the relay received signal \( Y_1 \) by the value \( e^{j\angle h_1 - j\angle h_0} \). Also, \( \rho = \rho(\bar{h}) \) can be optimally found, therefore the lower bound is given by

\[
R \leq \max_{\rho, \, 0 \leq |\rho| \leq 1} \min \{ \log(1 + (1 - \rho^2)\gamma_0 P_1), \\
\log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\rho \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \} 
\]

(6.16)
where $\rho$ is a real number.

To prove the lower bound of Lemma 6.3.2 we use the achievable rate of the decode-and-forward algorithm [7, 54]. It can be easily shown from Theorem 1 in [7] that an achievable rate for the complex Gaussian relay channel is the minimum of $R_1$ and $R_2$ as given by

$$R \leq R_1 = I(X_1; Y_1(h) | X_2), \quad (6.17)$$

and

$$R \leq R_2 = I(X_1, X_2; Y(h)). \quad (6.18)$$

Consider $n$ consecutive channel uses and let index $i = 1, 2, \ldots, n$ denote the time index of the channel uses. The bound of (6.18), i.e., $I(X_1, X_2; Y(h))$, is the multiple access cut bound derived in Lemma 6.3.1. However, the bound of (6.17) is slightly different from the broadcast bound of (6.3), i.e., $I(X_1; Y_1(h), Y(h) | X_2)$, again we consider $n$ consecutive channel uses. Let index $i = 1, 2, \ldots, n$ denote the time index of the channel uses, we have

$$R \leq \frac{1}{n} \sum_{i=1}^{n} I(X_{1i}; Y_{1i}(h) | X_{2i}) \quad (6.19)$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \{ \mathcal{H}(Y_1(h) | X_2) - \mathcal{H}(Y_1(h) | X_1, X_2) \}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \{ \mathcal{H}(Y_1(h) | X_2) - \log(2\pi e N_1) \}$$

where $\mathcal{H}(X)$ denotes the differential entropy of $X$. Following the proof of Lemma 6.3.1
and especially the derivation of the broadcast bound in (6.11), we have

$$\mathcal{H}(Y_1(h)|X_2) = E[\mathcal{H}(Y_1(h)|x_2)]$$

$$\leq \log \left( (2\pi e)^2 \left( |h_0|^2 (E[|X_1|^2] - \frac{E^2[|X_1X_2^*|]}{E[|X_2|^2]}) + N_1 \right) \right)$$

Combining Equations (6.19) and (6.20) and using Jensen’s inequality we obtain

$$C \leq \frac{1}{n} \sum_{i=1}^{n} \{\mathcal{H}(Y_1(h)|X_2) - \log(2\pi e N_1)\}$$

$$C \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ \log \left( 1 + \frac{|h_0|^2}{N_1} \left( E[|X_1|^2] - \frac{E^2[|X_1X_2^*|]}{E[|X_2|^2]} \right) \right) \right\}$$

$$C \leq \log \left( 1 + \frac{|h_0|^2}{N_1} \left( \frac{1}{n} \sum_{i=1}^{n} E[|X_1|^2] - \frac{1}{n} \sum_{i=1}^{n} \frac{E^2[|X_1X_2^*|]}{E[|X_2|^2]} \right) \right)$$

Therefore, by using the definition of channel quality parameters $\gamma_1$ and $\gamma_0$, the input power constraint $P_1$, and the correlation factor $\rho$, we have

$$C \leq \log (1 + \gamma_0 (1 - |\rho|^2) P_1)$$

(6.22)

Because both bounds of (6.22) and (6.8) should be satisfied, the minimum of these two is the lower bound which is given by

$$R \leq \max_{\rho, 0 \leq |\rho| \leq 1} \min \{ \log(1 + (1 - |\rho|^2)(\gamma_0 + \gamma_1) P_1), \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2 \text{Re}(\rho e^{j\angle h_1 - j\angle h_2}) \sqrt{\gamma_1 \gamma_2 P_1 P_2}) \}$$

(6.23)

where one can optimize the bound by finding the appropriate value of $\rho$ if and only if knowledge of channel state $h = (h_0, h_1, h_2)$ is available at the transmitter. Therefore,
where the knowledge of the channel state $h = (h_0, h_1, h_2)$ is not available at the
transmitter, the lower bound on the achievable rate of any coding scheme is given by
(6.15) for a constant $\rho$ defined in (6.9) which only depends on the correlation between
the input signals. It should be noted that the critical information at the transmitter
is only the difference between the phase of $h_1$ and $h_2$. This completes the proof of
Lemma 6.3.2.

\section*{6.4 Fading Relay Channel}

\subsection*{6.4.1 Upper Bound on the Capacity}

In [7], an upper bound for the information transfer rate $R$ in the discrete memoryless
relay channel is derived as follows

$$
C \leq \sup_{P(x_1, x_2)} \min \{I(X_1; Y, Y_1|X_2), I(X_1, X_2; Y)\} \tag{6.24}
$$

This bound has a simple interpretation based on the min-cut max-flow theorem (Theo-
rem 14.9.1 in [8]). For the fading relay channel, this upper bound can be expressed
in terms of the channel parameters and the power constraints as follows.

\textbf{Proposition 6.4.1} (based on cut-set theorem [8, 44]). \textit{The achievable rate of any
coding scheme for the fading relay channel in Figure 6.1 without side information at}
the transmitters is bounded above by

\[ R \leq \max_{\rho, 0 \leq \rho \leq 1} \min \left\{ \mathbb{E} \left[ \log(1 + (1 - |\rho|^2)(\gamma_0 + \gamma_1 P_1)) \right], \right. \]
\[ \left. \mathbb{E} \left[ \log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2 \Re \left( \rho e^{j\angle h_1 - j\angle h_2} \sqrt{\gamma_1 \gamma_2 P_1 P_2} \right) \right] \right\} \]  \tag{6.25}

where \( \gamma_0 \triangleq \frac{|h_0|^2}{N_1}, \gamma_1 \triangleq \frac{|h_1|^2}{N} \) and \( \gamma_2 \triangleq \frac{|h_2|^2}{N} \) and where the expectations have taken over the fading coefficients \( h_0, h_1, \) and \( h_2. \) Furthermore, while \( \rho \) in general can be a function of channel state information at the transmitters (i.e., if the perfect channel state information is available at the transmitters then \( \rho = \rho(h_0, h_1, h_2) \)), without the side information at the transmitters \( \rho \) is a constant.

Here, we wish to emphasize the role of parameter \( \rho \) which corresponds to the correlation factor between the channel input \( X_1 \) and relay signal \( X_2. \) While increasing \( \rho \) would increase the mutual information term \( I(X_1, X_2; Y) \) by helping the transmission in multiple access cut of the relay channel, it limits the information transfer rate in the broadcast cut of the channel. Therefore for different channel parameters \( h_0, h_1, h_2, N_1, \) and \( N, \) there are different values of the correlation factor \( \rho \) which optimizes the mentioned upper bound. In fact, the correlation between the relay input and channel input can be introduced by using two factors: (i) the knowledge of the received signal at the relay and (ii) the transmission strategy, e.g., repetition coding at the input of the channel. Clearly, introducing the correlation between the channel input and relay input increases the information transfer rate in the multiple access.
cut of the relay channel. However, it has its own drawback: it means less information transfer in the broadcast cut of the relay channel which can be interpreted as having prior knowledge of some part of the transmitted message from the source at the relay node (because of the correlation).

6.4.2 Lower Bounds on the Capacity

The best-known achievable rate for Gaussian relay channels is the Markovian scheme [7] of Cover and El Gamal presented in 1979. In this scheme, transmission occurs in several blocks of long codewords. In each block some information is solely encoded for the reception at the relay, and the codeword length is long enough to allow almost error-free decoding by the relay. Therefore the source and relay nodes cooperate to resolve the ambiguity at the destination node about the message which is sent in the previous block by using the information which is now shared between the source and relay nodes. Although this clever scheme is capacity achieving for some special cases, it is restricted to decoding and re-encoding at the relay node, and hence its achievable rate is no more than the capacity of the source-relay channel. This scheme is also known as the decode-and-forward scheme. Therefore, an achievable rate of the decode-and-forward coding scheme for the ergodic fading relay channel can be found as follows.

Proposition 6.4.2 (based on Theorem 5 [7]). The achievable rate of the decode-and-forward scheme for the fading relay channel in Figure 6.1 without side information
at the transmitters is given by

\[
R_{DF} = \max_{\rho, 0 \leq \rho \leq 1} \min \{\mathbb{E}[\log(1 + (1 - |\rho|^2)\gamma_0 P_1)],
\mathbb{E}[\log(1 + \gamma_1 P_1 + \gamma_2 P_2 + 2\text{Re}(\rho e^{j\angle h_1 - j\angle h_2})\sqrt{\gamma_1 \gamma_2 P_1 P_2})]\} \quad (6.26)
\]

where \(\gamma_0 \triangleq \frac{|h_0|^2}{N}, \gamma_1 \triangleq \frac{|h_1|^2}{N}\) and \(\gamma_2 \triangleq \frac{|h_2|^2}{N}\) and where the expectations have taken over the fading coefficients \(h_0, h_1,\) and \(h_2\). Furthermore, while \(\rho\) in general can be a function of channel state information at the transmitters (i.e., if the perfect channel state information is available at the transmitters then \(\rho = \rho(h_0, h_1, h_2)\)), without the side information at the transmitters \(\rho\) is a constant.

It can be easily verified that if \(\gamma_1 > \gamma_0\) then direct transmission without using the above Markovian scheme would achieve a higher rate. Thus, we have the following proposition.

**Proposition 6.4.3** (from [30]). The achievable rate of the direct transmission for the fading relay channel in Figure 6.1 without side information at the transmitters is given by

\[
R_{Direct} = \mathbb{E}[\log(1 + \gamma_1 P_1)] \quad (6.27)
\]

where \(\gamma_1 \triangleq \frac{|h_1|^2}{N}\) and where the expectation has taken over the fading coefficient \(h_1\).

The above propositions imply that if the relay has better received SNR than the destination, i.e. \(\gamma_0 > \gamma_1\), then using the relay is helpful and improves the achievable
rate of the direct transmission. Now, the question is, what if a relay node exists that wants to assist the communication but its received SNR is not as good as the destination received SNR, that is, $\gamma_1 > \gamma_0$? The difficulty with Markovian coding (Proposition 6.4.2) becomes clearer if we consider the situation where the available power of the relay is very large. Unfortunately, even for a very large available power at the relay node, there is no gain over direct transmission by using the decode-and-forward scheme [7] where $\gamma_1 > \gamma_0$. However, instead of decoding at the relay we can use the estimate [7] of the received signal at the relay, i.e., $\hat{Y}_1$ [54, 38]. Again, in this scheme encoding is performed in several blocks of large codeword length. Transmitting the estimate of the relay received signal, $\hat{Y}_1$, involves using the idea of source coding (at the relay node) and decoding with side information (at the destination node) [37] by considering a test channel with given pdf, $p(\hat{y}_1|x_2,y_1)$. Basically, a random code for this test channel is generated at the relay node based on its received signal in the previous block; and the generated codeword is then transmitted during the current block. The destination also knows the probability density function (pdf) of the test channel, $p(\hat{y}_1|x_2,y_1)$, and by using its received signal, the destination finds the estimate of the relay received signal, $\tilde{y}_1$. By using statistical transmission of the estimate of the relay received signal [38, 7] the following rate (6.28) is achievable.

**Proposition 6.4.4** (based on Proposition 3 [54]). *The achievable rate of the estimate-and-forward scheme for the fading relay channel in Figure 6.1 without side informa-
tion at the transmitters is given by

\[ R_{EF} = \mathbb{E} \left[ \log \left( 1 + \gamma_1 P_1 + \frac{\gamma_0 P_1 \gamma_2 P_2}{1 + \gamma_0 P_1 + \gamma_1 P_1 + \gamma_2 P_2} \right) \right] \] (6.28)

where \( \gamma_0 \triangleq \frac{|h_0|^2}{N_1} \), \( \gamma_1 \triangleq \frac{|h_1|^2}{N} \), \( \gamma_2 \triangleq \frac{|h_2|^2}{N} \) and where the expectations have taken over the fading coefficients \( h_0, h_1, \) and \( h_2 \).

By comparing (6.28) and (6.27), it is quite clear that the achievable rate of estimate-and-forward, \( R_{EF} \), is always greater than the rate of direct transmission, \( R_{Direct} \). On the other hand, depending on the channel conditions, the decode-and-forward scheme may achieve a higher rate than the estimate-and-forward scheme.

6.4.3 Ergodic Capacity of the Fading Relay Channel

In this section we use the upper and lower bounds of Section 6.4.1 and 6.4.2 to derive the ergodic capacity of the Rayleigh fading relay channel without side information at the transmitters, i.e., the value of \( h_0 \) is only known to the relay node, and \( h_1 \) and \( h_2 \) are only known to the destination node. First, we find the optimal value of the parameter \( \rho \) in Propositions 6.4.2 and 6.4.1, which are different in general. We show that for many cases of interest the value of this parameter is zero which immediately results in the exact characterization of the capacity for these cases. It should be pointed out that the performance of the decode-and-forward scheme (Proposition 6.4.2) degrades severely when the channel parameters do not fall into the favorable region. However, it can be shown that the estimate-and-forward scheme (Proposition 6.4.4) is not very
sensitive to the changes in the long-term channel parameters, and in almost all cases, this scheme performs reasonably close to the upper bound on the capacity. Therefore, from a practical standpoint, the decode-and-forward protocol, which corresponds to a multi-hopping strategy [60], might not be a good choice. Nevertheless, even though the estimate-and-forward protocol is not the optimal coding scheme for cases in which the capacity is achieved with the decode-and-forward scheme, it may still be considered the best strategy because of its robustness. It should be pointed out that for the cases where the decode-and-forward strategy is superior to the estimate-and-forward scheme the difference between their achievable rates is usually tolerable. The same, however, is not true when estimate-and-forward is the better scheme. In this case, the achievable rate of the estimate-and-forward scheme might be even an order of magnitude higher than that of the decode-and-forward scheme.

**Theorem 6.4.5.** The capacity of the Rayleigh fading relay channel (Figure 6.1) without side information at the transmitters - where the fading coefficients of the source-destination, source-relay, and relay-destination links are complex Gaussian, mutually independent, and also independent of the additive noise at the source and the relay - is given by the following:

\[
C = \mathbb{E}\left[\log(1 + \gamma_1 P_1 + \gamma_2 P_2)\right]
\]

Provided that

\[
\mathbb{E}\left[\log(1 + \gamma_1 P_1 + \gamma_2 P_2)\right] \leq \mathbb{E}\left[\log(1 + \gamma_0 P_1)\right]
\] (6.29)
Where the expectation has taken over the random variables $\gamma_0 \triangleq \frac{|h_0|^2}{N_1}$, $\gamma_1 \triangleq \frac{|h_1|^2}{N}$ and $\gamma_2 \triangleq \frac{|h_2|^2}{N}$, which are mutually independent and exponentially distributed with the means $\frac{\lambda_0}{N_1}$, $\frac{\lambda_1}{N}$ and $\frac{\lambda_2}{N}$, respectively.

Proof: First, we note that with the assumption of Rayleigh fading, the distribution of random fading coefficients $h_0$, $h_1$, and $h_2$ are complex Gaussian, and they are mutually independent. Therefore, the channel quality parameters $\gamma_0 \triangleq \frac{|h_0|^2}{N_1}$, $\gamma_1 \triangleq \frac{|h_1|^2}{N}$ and $\gamma_2 \triangleq \frac{|h_2|^2}{N}$ are mutually independent and exponentially distributed with the means $\frac{\lambda_0}{N_1}$, $\frac{\lambda_1}{N}$ and $\frac{\lambda_2}{N}$, respectively. Furthermore the phase of the fading coefficients $h_0$, $h_1$, and $h_2$ are uniformly distributed in $[0, 2\pi)$. The phases $\theta_0 \triangleq \angle h_0$, $\theta_1 \triangleq \angle h_1$, and $\theta_2 \triangleq \angle h_2$ are mutually independent and also independent of the amplitudes $|h_0|$, $|h_1|$, and $|h_2|$.

Consider the achievable rate of Proposition 6.4.2 as a function of the correlation factor $\rho$, i.e., $R_{DF}(\rho)$. Based on the above discussion, we note that the second term of Equation 6.26 is independent of the phase of the correlation parameter $\rho$, because $\theta_1 = \angle h_1$ and $\theta_2 = \angle h_2$ are uniformly distributed over $[0, 2\pi)$. On the other hand, it is clear that the first term of Equation 6.26 is also independent of the phase of the correlation parameter $\rho$. Therefore, we have $R_{DF}(\rho) = R_{DF}(-\rho)$. Thus, based on Jensen’s inequality we have

$$\forall \rho : R_{DF}(\rho) = \left( \frac{R_{DF}(\rho) + R_{DF}(-\rho)}{2} \right) \leq R_{DF}(0)$$ (6.30)

Therefore, $\rho = 0$ maximizes the decode-and-forward achievable rate expression. The
same analysis is true for the upper bound of Proposition 6.4.1, i.e., \( \rho = 0 \) maximizes the upper bound expression of (6.25). It is now very simple to see that if condition (6.29) is satisfied then the upper bound of Proposition 6.4.1 and the lower bound of Proposition 6.4.2 coincide, and we have

\[
C = \mathbb{E}_{\gamma_1, \gamma_2}[\log(1 + \gamma_1 P_1 + \gamma_2 P_2)]
\]

This completes the proof.

The fact that \( \rho = 0 \) maximizes the decode-and-forward achievable rate expression means that in the Rayleigh fading environment, the best strategy for the decode-and-forward protocol is to avoid any correlation between the input signal, \( X_1 \), and the relay signal, \( X_2 \). Since the upper bound is also maximized without any correlation between the input signals, that is \( \rho = 0 \), it seems that this property holds the key to the exact characterization of the capacity for the remaining cases.

### 6.5 Numerical Results

As an illustrative example we consider a case (Figure 6.2) in which the source node is located at \((0, 0)\) and the destination node is located in the \((0, d_1)\). We assume that the mean values \( \lambda_0 \), \( \lambda_1 \), and \( \lambda_2 \) of the norm squared of the fading coefficients are proportional to the distance between the source-relay \((d_1)\), source-destination \((d_2)\), and relay-destination \((d_3)\), respectively. Specifically, let \( \lambda_i \propto \frac{1}{d_i^\alpha} \), for \( i = 0, 1, 2 \), where \( \alpha \) denotes the path-loss exponent which is some constant value between 2 and 4 in
practice. As shown in Figure 6.2, if the relay lies in the “apple”-like region around the source, condition (6.29) is satisfied. The size of the region for which condition (6.29) is satisfied depends on the available average powers $P_1$ and $P_2$ and on the pathloss exponent $\alpha$. However, this region usually contains a circle around the source with the radius $\frac{d_1}{2}$. Therefore, if the source node finds a relay node (in any direction) which is not further than half of the distance between source and relay, the exact capacity is known. In fact, this region corresponds to the most useful region, because if the source node wants to find a relay, it is very likely to choose the relay which sees the best SNR from the source, i.e., A relay which is relatively close to the source node.

Figure 6.3 shows the upper bound on the capacity as well as the achievable rates of the different mentioned schemes for different locations of the relay node between the source and destination nodes. As shown in Figure 6.3, the increase in capacity by the use of relay is usually remarkable. For example, in Figure 6.3 if the relay is located in $0.62d_1$ then the capacity by using the relay is about 150% of the capacity without using the relay. It should be noted that the achievable rate of the estimate-and-forward protocol is always higher than that of the direct transmission (relay off), but this is not true for the decode-and-forward protocol. In fact, the achievable rate of the decode-and-forward protocol is limited to the capacity of the source-relay link which might be even less than the capacity of the direct link. However, for the range of the relay position plotted in Figure 6.3, the achievable rate of decode-and-forward is also superior to that of direct transmission.
Figure 6.2: Relaying region for which the capacity of the Rayleigh fading relay channel is known where the relay node lies in this region.
However, if we use direct transmission with the aggregate powers of the source and relay, i.e., $P'_1 = P_1 + P_2$, then the achievable rate of direct transmission might be higher than that of the decode-and-forward scheme as it can be seen in the small portion of Figure 6.3 where the relay node is located close to the destination node.

### 6.6 Conclusion

The capacity of the fading relay channel with Gaussian noise is especially of interest due to its practical applications in wireless communication systems and networks. We have derived the capacity of the fading relay channel for many cases of interest, and we have shown that the decode-and-forward scheme achieves it. However, the decode-and-forward protocol is very sensitive to the prior assumption about the channel distribution and location of the relay, and it suffers a huge loss in performance when the relay is not located in the “apple”-like relaying region of Figure 6.2.

Despite the sensitivity of the decode-and-forward scheme to the relay position, the estimate-and-forward protocol [54, 38] is a very robust coding scheme that always performs better than direct transmission, and in most of the cases it closely follows the upper bound on the capacity. In short, the estimate-and-forward protocol seems to be a better choice from the practical point of view, especially for ad-hoc networks, where the location of the relay is not known and where the relay may move outside the “apple”-like region of Figure 6.2.
Figure 6.3: Upper and lower bounds on the capacity of the fading relay channel, where $d$ denotes the source-relay and $D$ denotes the source-destination distance.
Chapter 7

Fading Channel: Peak and Average Power Constraints

We derive the ergodic capacity of the discrete-time fading channel with additive Gaussian noise subject to both peak and average power constraints. The average power can be interpreted as the cost that we incur to achieve a certain rate. However, the motivation of this analysis results from a peak power limitation in a practical communication system. It is been shown that the optimal power adaption is no longer water-filling or a constant power adaption, which is the case when there is no limitation on the peak power. The numerical results show that the importance of peak power constraints become negligible for relatively low available average power, while it is limiting the capacity to be finite even as available average power goes to infinity.

7.1 Introduction

The capacity of the various single-user memoryless channels with different constraints on the channel input and channel characteristics has been extensively studied since the early age of information theory [61, 62, 63, 30]. When the cardinality of the input alphabet is not finite, a constraint on the input is needed in order to have finite
channel capacity. The most commonly used constraint is one on the average power, for which the associated capacity of the Gaussian channel is derived by Shannon [61]. Also, the capacity of the fading channel subject to average power constraint with perfect side information about the fading state both at the transmitter and the receiver is determined in [30], and the corresponding capacity without side information at the transmitter is found in [64]. However, there has not been much attention to the capacity of the channels subject to both average and peak power constraints, although its importance has been known for decades [61], because it better embraces the practical limitation of communication systems [65, 66, 67].

The Gaussian channel subject to an input peak power constraint is the first channel for which a constraint other than the average power is considered. The behavior of the Gaussian channels subject to peak and average power constraints is first studied in the original work of Shannon [61] for asymptotically low and high SNR in the context of a bandlimited continuous-time Gaussian channel. The capacity of the Gaussian channel subject to both peak and average power constraints was later studied by Smith [62], who showed that the capacity achieving distribution is discrete. The generalization of this result for the quadrature additive Gaussian channel, along with lower and upper bounds on the capacity of this channel, can also be found in [63].

In this section, we consider the capacity of the fading channel subject to both peak and average power constraints under the assumption of perfect side information about the fading state both at the transmitter and the receiver. We show that a multiplexed
Gaussian codebook with the rate and power adaption based on the fading states is sufficient to achieve the capacity. We prove a coding scheme which is more general to some extent and which subsumes the achievability of the rate for the specific case of peak and average power constraints over the fading channel. The converse part of the proof is derived from a newly derived cut-set theorem for the multi-state network [44, 68]. We show that as the peak power constraint approaches the average power, the capacity of fading channel under both peak and average power constraints approaches the capacity without side information at the receiver. However, as peak power goes to infinity, i.e., the peak power constraint relaxes, the optimal power adaption becomes pure water-filling, and the corresponding capacity approaches the capacity of the fading channel with perfect side information at the transmitter and the receiver.

The rest of this section is organized as follows: We define some terminologies and notations and also formulate the problem in Section 7.2. Our main result which is the capacity of the fading channel with both peak and average power constraints is presented in Section 7.3. The coding theorem and proof of the achievability is given in Section 7.4, and the converse is derived in Section 7.5. In Section 7.6, we present the power adaption technique and the capacity results for the Rayleigh fading channel. Finally, we conclude in Section 7.7.
7.2 Problem Formulation and Notations

We consider a discrete-time fading channel where in each time index $i$ the channel is characterized by two time-varying parameters: the channel gain $\sqrt{g[i]}$ and the additive Gaussian noise $z[i]$. The channel power gain $g[i]$ is also known as the fading state of the channel. The random processes $g[i]$ and $z[i]$ are assumed to be ergodic and independent from each other and from the channel input. Let $N_0$ denote the noise density and $B$ denote the received signal bandwidth. The additive noises $\{z[i]\}_{i=1}^{\infty}$ are assumed to be zero mean i.i.d. Gaussian random variables with the distribution $N(0,N_0B)$. The channel power gains $\{g[i]\}_{i=1}^{\infty}$ are also assumed to be i.i.d. and non-negative real random variables with a known distribution such that the expected value of the channel power gain is unity. If $P_{in}[i]$ denotes the input power at time $i$ and $\bar{P}$ denotes the average input power, then the instantaneous received signal-to-noise ratio without power control is given by $\gamma[i] = \bar{P}g[i]/(N_0B)$, and its expected value is equal to $\bar{P}/(N_0B)$.* Because the distribution of the channel gain is assumed to be known, the distribution of $\gamma$ is known and will be denoted by $f(\gamma)$.

The channel model is illustrated in Figure 7.1. The message $w$ is encoded into the codeword $x$ of length $n$, which is transmitted over the time-varying channel as $x[i]$ at

*Remark: Throughout this section we shall refer to $\gamma$ as “Received SNR” for short, but we imply “Received SNR without power control” which only shows the variation of the fading state and not the power adaption. Clearly, with power control the actual value of the received SNR depends on both the fading state and the average power used in this state.
Figure 7.1: System model for communication over the fading channel with transmitter side information.

time \( i \). The channel gain \( g[i] \) varies from time to time, and the coding scheme, which depends on both the message \( w \) and the codeword length \( n \), achieves the capacity of the channel if it is long enough to see all possible realizations of the channel. This will happen by letting the codeword length \( n \to \infty \) assuming ergodicity. We assume that the instantaneous value of the channel gain \( g[i] \) is known to the receiver at each time \( i \) through a perfect channel estimator. Therefore the receiver can keep track of the channel state sequence \( \{g[i]\}_{i=1}^{\infty} \) and it always has perfect channel state information.

If a perfect and delay-less feedback channel is available, then the transmitter also has perfect channel state information. Thus, we consider that receiver always has perfect CSI, but we distinguish between the cases where the transmitter has perfect CSI or no CSI.

For any channel with continuous input alphabet, the capacity of the channel is infinite unless a constraint on the input power is considered. Usually, this constraint is in the form of an upper bound on the average input power \( E_{\gamma}[P(\gamma)] \leq P \), which
is an indication of the cost that we incur to achieve a given rate. Here, $P(\gamma)$ is the power adaption based on the state of the channel $\gamma$. Therefore, the instantaneous power adaption at time $i$ is given by $P_m[i] = P(\gamma[i])$. However, in many practical systems, there is a limitation on the peak power [65, 66] due to the non-linear function of the amplifiers and also due to compliance with other standards (e.g. to limit the interference on the other communications). In this section, we consider the case that both of the above mentioned constraints are present as

$$E_\gamma[P(\gamma)] \leq \overline{P} \tag{7.1}$$

$$\forall \gamma : P(\gamma) \leq P_{\text{max}} \tag{7.2}$$

### 7.3 Capacity Analysis

The ergodic capacity of fading channels with average power constraint is studied in [64, 30]. If the channel state information is not available at the transmitter, the capacity of the channel can be achieved by the Gaussian codebook with constant power $\overline{P}$ and the capacity is given by average capacity formula [64] over all possible channel states

$$C_{\text{RCSI}}(\overline{P}) = E_\gamma[B \log(1 + \gamma)]. \tag{7.3}$$

It is also possible to interpret this average capacity formula by using the result of [69] on the capacity of the channels with side information at the receiver only. In fact, if the same input distribution maximizes the mutual information for all the channel
states and if the sequence of the channel states are i.i.d., then it has been shown that
the capacity is given by the average capacity formula over all channel states [69].

However, if the perfect channel state information is available both at the transmitter and at the receiver, then capacity is again given by an average capacity formula with variable power adaption over all channel states [30]

\[ C_{T\&RC} = E_{\gamma} \left[ B \log \left( 1 + \frac{P(\gamma)\gamma}{\bar{P}} \right) \right], \quad (7.4) \]

where power adaption over the channel states, \( P(\gamma) \), is in the form of the water-filling solution given by

\[
\frac{P(\gamma)}{\bar{P}} = \begin{cases} 
\frac{1}{\gamma_0} - \frac{1}{\gamma}, & \gamma \geq \gamma_0 \\
0, & \gamma < \gamma_0 
\end{cases} , \quad (7.5)
\]

such that it satisfies the average power constraint

\[ E_{\gamma}[P(\gamma)] = \bar{P}, \quad (7.6) \]

If the input signal is subject to a peak power constraint of the form (7.2) in addition to the average power constraint of (7.1), then we show that still a multiplexed Gaussian codebook with variable power would achieve the capacity of this channel. We derive an achievable rate by proving the existence of a coding scheme that satisfies the above constraints, and by proving the converse we show that this achievable rate is the capacity of the channel. We also derive explicitly the form of the optimal power adaption which is needed to achieve the capacity in this case. It will be shown that the power adaption is divided into three regions based on the channel state. In the
first region, we do not use the channel for the values of $\gamma < \gamma_0$. In other words, for the values of $\gamma$ less than a threshold $\gamma_0$, the input power is set to zero. In the second region, defined as the values of $\gamma_0 \leq \gamma < \gamma_1$, the power adaption takes the water-filling form. Finally, for the third region, which corresponds to $\gamma_1 \leq \gamma$, a constant power equal to $P_{\text{max}}$ will be used. The threshold values of the received SNR, $\gamma_0$ and $\gamma_1$, are determined such that the average power constraint is satisfied. We show that for the fixed average power constraint by relaxing the peak-power constraint, the optimal power solution is in the form of the pure water-filling power control, which coincides with the results of [30]. However, if the peak-power constraint remains unchanged, relaxing the average power constraint allows the optimal power constraint to take the form of constant power adaption. In fact, in this case there would be no power adaption and capacity would coincide with the capacity of the channel without side information at the transmitter. Specifically, we have the following theorem.

**Theorem 7.3.1.** The capacity of the fading channel with perfect channel state information both at the transmitter and the receiver and the input signals subject to both peak and average power constraints of (7.1) and (7.2), respectively, is given by

$$C_{T\&RCSI}(\mathcal{P}, P_{\text{max}}) = E_\gamma \left[ B \log \left( 1 + \frac{P(\gamma)\gamma}{P} \right) \right],$$

(7.7)
where the power adaption over the channel states, $P(\gamma)$, is in the form of

$$
P(\gamma) = \begin{cases} 
0, & \gamma < \gamma_0 \\
\frac{P - P^*}{\gamma_0 - \gamma}, & \gamma_0 \leq \gamma < \gamma_1 \\
\max, & \gamma_1 \leq \gamma 
\end{cases} \quad (7.8)$$

such that the threshold values $\gamma_0$ and $\gamma_1$ are determined from the following equations

$$
\frac{P}{\gamma_0} - \frac{P}{\gamma_1} = \max, \quad (7.9)
$$

$$
E_{\gamma}[P(\gamma)] = \overline{P}. \quad (7.10)
$$

In order to prove the above theorem, it is necessary to show both the achievability of the rate $C_{TKRCSI}(P, \max)$ in (7.7) and the converse for the coding theorem for any coding schemes which satisfy both of the average and peak power constraints of (7.1) and (7.2), respectively. In Section 7.4, the achievability is proved for a broader class of power allocation schemes defined in Section 7.4 that easily subsumes the class of $F_1$ defined as

$$
F_1 = \left\{ P(\gamma) : E_{\gamma}[P(\gamma)] \leq \overline{P}, \; \forall \gamma : P(\gamma)_{a.s.} \leq \max \right\} \quad (7.11)
$$

which is the class of power allocation schemes that satisfies both the average and peak power constraints of (7.1) and (7.2). The converse part is proved based on the newly derived max-flow min-cut theorem for multi-state networks in Chapter 2 [44, 68] and it will be presented in Section 7.5 where we prove that the achievable rate of any possible coding schemes, for which the power control belongs to the class of $F_1$, is not greater than the rate given by Theorem 7.3.1.
7.4 Coding Theorem

Consider a power allocation policy $P(\gamma)$ (policy in general might be stochastic or deterministic), which allocates the power $P(\gamma)$ for the received SNR value of $\gamma$. We define a class of power allocation policies as a given subset of all possible power allocation policies. A class of power allocation policies $\mathcal{P}$ may be defined as all power allocation policies that satisfy a given set of conditions making them distinguishable from other possible policies. For example, $\mathcal{F}_1$ in (7.11) identifies the class of power allocation policies for which the power constraints of (7.1) and (7.2) are fulfilled.

The following theorem shows a coding scheme which gives an achievable rate for the power control policy that belongs to a given class. Therefore, supremum of the achievable rates in Theorem 7.4.1 over all possible power allocation policies of the same class is also achievable. In the next section, instead of taking the supremum which is hard, we pick a given policy in the class of $\mathcal{F}_1$ and show that this is actually the supremum of the achievable rate in this class by establishing the proof of the converse.

**Theorem 7.4.1.** Let $\mathcal{P}$ denote the class of power adaption policies defined as

$$\mathcal{P} = \left\{ P(\gamma) : E_{\gamma} \left[ B \log \left( 1 + \frac{P(\gamma)\gamma}{\mathcal{P}} \right) \right] < \infty \right\}. \quad (7.12)$$

Consider a power control policy $P(\gamma) \in \mathcal{P}$ and let

$$\mathcal{D}(P) = \{ \gamma : P(\gamma) \neq 0 \} \quad (7.13)$$
Denote the domain of $P(\gamma)$. There exists a coding scheme which achieves the rate $R$ defined as
\begin{equation}
R = \int_{D(P)} B \log \left( 1 + \frac{P(\gamma) \gamma}{P} \right) f(\gamma) d\gamma \quad (7.14)
= E_\gamma \left[ B \log \left( 1 + \frac{P(\gamma) \gamma}{P} \right) \right],
\end{equation}
with vanishing probability of error as the codeword length becomes infinitely large over the fading channel with perfect side information both at the transmitter and the receiver.

**Proof:** Clearly the values of the received SNRs, $\gamma$, are $0 \leq \gamma < \infty$. Let $\rho = \inf_{\gamma \in D(P)} \gamma$ and define the set
\begin{equation}
D_1 = \{ \gamma : \rho \leq \gamma \leq \rho + M \} \quad (7.15)
\end{equation}
for some integer value $M$ (to be defined later). For the sake of simplicity we assume that the power adaption policy $P(\gamma)$ is a deterministic policy and continuous over $\gamma$. Extension to general stochastic policies with possible discontinuity is not hard but also not needed to derive our results in this section. In other words, we will later choose a deterministic and continuous policy to achieve the capacity of the channel over the class of $\mathcal{F}_1$.

Let $m$ be a positive integer and define $\gamma_i = \rho + \frac{i}{m}$ for all $i = 0, 1, \ldots, mM$. Also, let
\begin{equation}
\xi_i = \int_{\gamma_i}^{\gamma_{i+1}} P(\gamma) d\gamma \quad (7.16)
\end{equation}
denote the average power for any interval \( I_i = [\gamma_i, \gamma_{i+1}) \).

Over a given time interval \([0, n]\), let \( N_i \) denote the number of times that the received SNR is in interval \( I_i \). With the stationarity and ergodicity of the channel gains \( g[i] \) we have

\[
\frac{N_i}{n} \to \int_{\gamma_i}^{\gamma_{i+1}} f(\gamma) d\gamma. \tag{7.17}
\]

For a given \( n \), let

\[
n_i = \left\lfloor n \int_{\gamma_i}^{\gamma_{i+1}} f(\gamma) d\gamma \right\rfloor \tag{7.18}
\]

which is equal to \( N_i \) for sufficiently large \( n \). For any \( i = 0, 1, \ldots, mM \) define the rate \( R_i \) as

\[
R_i = B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \tag{7.19}
\]

and the overall rate \( R^{(n)} \) as

\[
R^{(n)} = \sum_{i=0}^{mM} \frac{n_i R_i}{n}. \tag{7.20}
\]

Clearly the rate \( R_i \) defined in (7.19) is achievable over an AWGN channel with the received SNR \( \frac{\xi_i \gamma_i}{P} \). Therefore, there exists a code \((n_i, 2^{n_i R_i})\) which encodes \( 2^{n_i R_i} \) equally likely messages, say \( w_i \), with average input power \( \xi_i \) and error probability \( \epsilon_{n,i} \to 0 \) as the codeword length \( n_i \to \infty \).

Now we define the structure of the codebook and corresponding coding scheme. For a given \( n \), the message index \( w \in \{1, 2, \ldots, 2^{n R^{(n)}}\} \) is mapped to the set of independent messages \( \{w_i\}_{i=1}^{mM} \) by partitioning \( n R^{(n)} \) bits into sets of \( n_i R_i \) bits. Definition of the \( n R^{(n)} \) in (7.20) ensures the existence of such a partitioning. The codebook for
the channel consists of \( mM \) independent Gaussian codebook, each with input average power \( \xi_i \), rate \( R_i \), and code length \( n_i \) for \( i = 1, 2, \ldots, mM \). In order to transmit the message index \( w \), we use the multiplexing strategy to transmit the corresponding message index \( w_i \) defined in the aforementioned mapping over the channel when the received SNR \( \gamma \) is in the interval \( I_i \). By definition, \( N_i \) denotes the number of times that the received SNR is in interval \( I_i \) over the time interval \([0, n]\). Therefore we can achieve the transmission rate of

\[
R^{(n)} = \sum_{i=0}^{mM} \frac{N_i R_i}{n} = \sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \frac{N_i}{n}. \tag{7.21}
\]

Taking the limit of \( R^{(n)} \) in (7.21) as \( n \to \infty \) and using the property (7.17), we have

\[
\lim_{n \to \infty} R^{(n)} = \lim_{n \to \infty} \sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \frac{N_i}{n}
= \sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \lim_{n \to \infty} \frac{N_i}{n}
= \sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \int_{\gamma_i}^{\gamma_i+1} f(\gamma) d\gamma.
\]

Thus, for any \( \epsilon > 0 \) there is a sufficiently large \( n \) such that

\[
R^{(n)} \geq \sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \int_{\gamma_i}^{\gamma_i+1} f(\gamma) d\gamma - \epsilon. \tag{7.23}
\]

Clearly, the value of the received SNR for the codeword \( w_i \) is not less than \( \frac{\xi_i \gamma_i}{P} \) for the codes from the \( i \)th sub-codebook; therefore, the error probability of the multiplexed coding scheme is bounded by the union bound \( \epsilon_n \leq \sum_{i=0}^{mM} \epsilon_{n,i} \) where \( \epsilon_n \to 0 \) as codeword length \( n \) becomes infinitely large because \( n \to \infty \) implies that either
\( n_i \to \infty \) or that there is no power allocated in the interval \( I_i \). In other words, \( I_i \notin D(P) \), which indicates that there is no input for these channel states, and thus no encoding or decoding will be performed.

From the assumption of the theorem, \( P(\gamma) \in \mathcal{P} \), therefore

\[
R = \int_{D(P)} B \log \left( 1 + \frac{P(\gamma)\gamma}{P} \right) f(\gamma) d\gamma < \infty
\]  

(7.24)

thus, for a given \( \epsilon > 0 \) there exists an (integer) value \( M \), such that

\[
\int_{\rho+M}^{\infty} B \log \left( 1 + \frac{P(\gamma)\gamma}{P} \right) f(\gamma) d\gamma < \epsilon.
\]  

(7.25)

However, for a fixed value of \( M \), the monotone convergence theorem [70] implies that

\[
\lim_{m \to \infty} \sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \int_{\gamma_i}^{\gamma_{i+1}} f(\gamma) d\gamma = \int_{\rho}^{\rho+M} B \log \left( 1 + \frac{P(\gamma)\gamma}{P} \right) f(\gamma) d\gamma.
\]  

(7.26)

where the last equation is obtained by using the definition of \( \xi_i \) in (7.16). Thus, for a given \( \epsilon > 0 \) there is a sufficiently large \( m \) such that

\[
\sum_{i=0}^{mM} B \log \left( 1 + \frac{\xi_i \gamma_i}{P} \right) \int_{\gamma_i}^{\gamma_{i+1}} f(\gamma) d\gamma \geq \int_{\rho}^{\rho+M} B \log \left( 1 + \frac{P(\gamma)\gamma}{P} \right) f(\gamma) d\gamma - \epsilon.
\]  

(7.27)
By combining (7.23), (7.25), and (7.27), we conclude that for any \( \epsilon > 0 \) there exist the values of \( n, m, \) and \( M \) sufficiently large such that

\[
R^{(n)} \geq \int_{\rho}^{\infty} B \log \left( 1 + \frac{P(\gamma) \gamma}{P} \right) f(\gamma) d\gamma - 3\epsilon. \tag{7.28}
\]

Thus by letting all three values of \( n, m, \) and \( M \to \infty \), any rate close to \( R \) in (7.14) is achievable, and this completes the proof of the coding theorem.

It is worth noting that the condition on the class of power adaption scheme in Theorem 7.4.1 is satisfied by all the power adaption policies subject to both peak and average power constraints defined as the class \( \mathcal{F}_1 \) in (7.11). Therefore, for any power adaption policy in the class \( \mathcal{F}_1 \), there exists a coding scheme which achieves the rate \( R \) in (7.14) with vanishing error probability as codeword length goes to infinity. Specifically, we have the following corollary.

**Corollary 7.4.2.** The class of power adaption policies \( \mathcal{F}_1 \) defined in (7.11) satisfies the condition of the class \( \mathcal{P} \) in Theorem 7.4.1, i.e., \( \mathcal{F}_1 \subset \mathcal{P} \).

**Proof:** Proof simply follows from Jensen’s Inequality.

### 7.5 Converse

In this section we derive a bound on the achievable rate of any coding scheme for which the input power allocation belongs to a given class of power allocation policies, say \( \mathcal{P} \). To find this bound we assume that both the transmitter and the receiver have noncausal and perfect knowledge of the sequence of the fading states, or equivalently
received SNRs \( \{ \gamma_i \}_{i=1}^\infty \). The proof is based on the newly derived min-cut max-flow theorem for multi-state networks [44, 68]. To be able to use this bound we first partition the range of the received SNRs to a finite number of intervals and then take the limit as the number of intervals grows infinitely.

**Theorem 7.5.1.** The achievable rate, \( R \), of any coding scheme for which the input power distribution, \( P(\gamma) \), belongs to the class of input power adaption policy \( \mathcal{P} \) is bounded above by

\[
R \leq \sup_{P \in \mathcal{P}} \int_{\mathcal{D}(P)} B \log \left( 1 + \frac{P(\gamma)\gamma}{\mathcal{P}} \right) f(\gamma) d\gamma
\]  

(7.29)

where \( \mathcal{D}(P) \) is defined as

\[
\mathcal{D}(P) = \{ \gamma : P(\gamma) \neq 0 \}.
\]  

(7.30)

**Proof:** It has been shown in [44, 68] that for a multi-state single link channel if the sequence of states of the network are non-causally known to both the transmitter and the receiver, then the achievable rate of information transfer over \( n \) uses of the channel is upper-bounded by

\[
R \leq \sum_{m=1}^{M} \frac{N_m}{n} I(X(m); Y(m))
\]  

(7.31)

where \( M \) is the number of states of the channel, \( N_m \) denotes the number of times that state \( m \) is used in past \( n \) uses of the channel, \( X(m) \) denotes the input, and \( Y(m) \) is the output where the channel is used in state \( m \). Let \( M(n) \) denote the possible states of the network in first \( n \) channel uses, and for any \( m = 1, 2, \ldots, M(n), \) we define \( \gamma(m) \)
to be the received SNR in state $m$. Because the value of $\gamma(m)$ is unique for different $m$, the set of $\{\gamma(m)\}_{m=1}^{M(n)}$ can also be used to denote the set of possible state of the channels. Let $N(n, \gamma)$ denotes the number of times that the channel is used in state $\gamma$, and let $D(n, P)$ denote the set of all observed $\gamma(m)$ over the first $n$ channel uses during first $n$ channel uses. Clearly as $n \to \infty$ the number of times that the channel is in the state $\gamma$ for the first $n$ channel uses approaches $nf(\gamma)d\gamma$.

Because for any state $\gamma$ the value of the state variable $\gamma$ is known to both the receiver and transmitter, the channel is equivalent to an AWGN channel with average noise power $\overline{\gamma}$. From the power control policy, $P(\gamma)$ denotes the average power used in state $\gamma$. Therefore, the maximum of the mutual information $I(X_{(m)}; Y_{(m)})$ is given by

$$I(X_{(m)}; Y_{(m)}) \leq B \log \left( 1 + \frac{P(\gamma(m))\gamma(m)}{\overline{\gamma}} \right). \quad (7.32)$$

By combining (7.31) and (7.32) and taking the limit as $n$ goes to infinity, we have

$$R \leq \lim_{n \to \infty} \sum_{m=1}^{M(n)} \frac{N_m}{n} B \log \left( 1 + \frac{P(\gamma(m))\gamma(m)}{\overline{\gamma}} \right) \quad (7.33)$$

$$\leq \lim_{n \to \infty} \int_{D(n, P)} \frac{N(n, \gamma)}{n} B \log \left( 1 + \frac{P(\gamma)\gamma}{\overline{\gamma}} \right)$$

$$\leq \int_{D(P)} f(\gamma) B \log \left( 1 + \frac{P(\gamma)\gamma}{\overline{\gamma}} \right) d\gamma.$$

Because as $n \to \infty$ the proportion of times of being in state $\gamma$ approaches the $f(\gamma)d\gamma$, and also the set $D(n, P) \to D(P)$ since each state $\gamma$ with nonzero $f(\gamma)$ would have nonzero probability of occurrence over $n$ independent channel use for sufficiently large $n$. 
Therefore, the supremum of the above bound over all power control policies $P \in \mathcal{P}$ defined in (7.29) establishes an upper-bound on the achievable rate of all coding schemes for which its input power belongs to the class of policies $\mathcal{P}$, and this completes the proof.

Now we consider the class $\mathcal{F}_1$ of the power control policies defined in (7.11) and find the power adaption policy $P(\gamma) \in \mathcal{F}_1$ which gives the supremum value of the bound of (7.29). It is not hard to see that the problem

\[
\begin{aligned}
\text{Maximize } & \int_{\mathcal{D}(P)} B \log \left(1 + \frac{P(\gamma)\gamma}{P}\right) f(\gamma) d\gamma \\
\text{Subject to } & E_{\gamma}[P(\gamma)] \leq \bar{P} \\
& \forall \gamma : P(\gamma) \leq P_{\text{max}}
\end{aligned}
\tag{7.34}
\]

is a constrained convex optimization problem which can be easily solved by lagrange multiplier techniques [71]. By applying Karush-Kuhn-Tucker (KKT) optimality conditions, the solution of $P(\gamma)$ is obtained. This solution is equivalent to the power adaption policy defined in (7.8), (7.9), and (7.10) in Theorem 7.3.1. Furthermore, it is quite easy to verify that this power adaption policy belongs to class $\mathcal{F}_1$. Thus, the rate of (7.7) is achievable for a code with power distributions belonging to class $\mathcal{F}_1$ through the coding scheme of the Theorem 7.4.1. Therefore, we have shown both the achievablity and the converse for the rate in Theorem 7.3.1 and also proved fact that the power control policy is the solution to the optimization problem defined in (7.34).
Figure 7.2: Optimal power adaption for the fading channel with both peak and average power constraint.

7.6 Numerical Results

In this section, we first demonstrate a visualization of the optimal power adaption for fading channels with both peak and average power constraints. Then, the capacity plots will be presented. It is clear that for a given peak power value, $P_{\text{max}}$, if we relax the average power $\bar{P}$, the capacity would be limited to the capacity with constant power adaption of (7.3) and will not grow by increasing the input average power beyond the peak power, i.e. $\bar{P} \geq P_{\text{max}}$.

The optimal power adaption defined in (7.8)-(7.10) is represented in Figure 7.2.
We recognize three regions which are distinguished by two instantaneous received SNR thresholds $\gamma_0$ and $\gamma_1$. When the received SNR is below $\gamma_0$, no transmission occurs, but, if the received SNR is greater than $\gamma_0$ but less than $\gamma_1$, power adaption is in the water-filling form. Finally, for the values of received SNR greater than $\gamma_1$, constant power equal to the peak power constraint is used.

The capacity of the Rayleigh fading channel with both peak and average power versus normalized average input power, defined as $\frac{P}{N_0 B}$, is depicted in Figure 7.3 for the various peak power constraints. The capacity of the Rayleigh fading channel without peak power constraints [30] for both cases of perfect CSI at the transmitter and receiver $C_{T\&RCSI}$, i.e., water-filling power adaption, and the receiver channel state information only $C_{RCSI}$, i.e. constant power adaption, is also shown as a reference. On the horizontal axis, the normalized average input power $\frac{P}{N_0 B}$ corresponds to average received SNR for the capacity plots of $C_{T\&RCSI}$ and $C_{RCSI}$, so it is labeled accordingly. However, the horizontal axis is in fact the normalized available input average power when we interpret the other capacity plots in this figure.

It can be noted that the capacity with both peak and average power constraints is the same as $C_{T\&RCSI}$ for low SNR, but as the average input power approaches the peak power constraint, it becomes flat, hits the curve $C_{RCSI}$ at the $C_{RCSI}(P_{max})$, and remains constant afterward.

It is a well-known fact that the capacity gain becomes “negligible” [30] for high SNR. Therefore, the capacity plots of Figure 7.3 and Figure 7.4 are shown over the
Figure 7.3: Capacity of the Rayleigh fading channel with different power constraints
range of SNRs where the capacity gain from water-filling adaption over constant power adaption is still considerable. We are expecting to have even lower capacity when the input is more constrained, i.e., where both peak and average power constraints exist. Figure 7.4 shows the capacity for the case that the peak power constraint is in the form of the fixed peak to average power ratio (PAR). Surprisingly, for the range of SNR depicted in Figure 7.4, the loss in capacity is not considerable even for the PAR = 1.7. The capacity with PAR = 1.1 is also depicted in Figure 7.4 which shows the trend of the capacity in this case. The capacity with constant peak to average power ratio is equal to the capacity with receiver side information only, $C_{RCSI}$, for low values of SNR, and this capacity diverges from $C_{RCSI}$ and approaches the capacity $C_{T&RCSI}$ for higher values of SNR.

### 7.7 Conclusion

We have derived the capacity of the fading channel subject to both peak and average power constraints. It has been shown that the capacity is achieved with a variable power multiplex Gaussian codebook. The optimal power adaption based on the fading state of the channel is divided into three separate regions: (i) zero power allocation (no transmissions), (ii) water-filling power allocation, and (iii) constant power allocation.
Figure 7.4: Capacity of the Rayleigh fading channel for a fixed peak to average power ratio.
Chapter 8

Joint Source-Channel Coding in Wireless Networks:
Application to Sensor Networks

In this section, we analyze a new abstraction to characterize the performance of joint source-channel coding in wireless sensor networks. The new abstraction models all wireless channels available in wireless networks and can be viewed as a user cooperation channel over which a pair of nodes convey two correlated sources. The proposed abstraction subsumes most of the models used as a motivation for distributed source coding in sensor networks. We derive the rate region for the proposed model, which matches numerous known results under special conditions. Our new results motivate a reconsideration of common paradigm in distributed source coding literature in which the existing wireless channels between encoding sensor nodes are not used. We contend that a “dialogue” between sensing nodes can lead to significant improvements in joint source-channel performance, and hence should be considered in the design of practical codes.

8.1 Introduction

Our focus in this section is to study a new abstraction for wireless sensor networks collecting spatially correlated data. The new abstraction is based on a simple obser-
vation that in wireless networks (sensor or otherwise), nodes receive radio signals from all neighboring nodes (especially if the nodes use omnidirectional antennas). Hence, locally the networks are fully connected meshes. In dense sensor networks, as the nodes get closer, two things happen simultaneously. First, the correlation between their observed data increases. Second, the channel quality between the node also improves. Thus, it is important to consider both the spatial correlation and available wireless channels to understand the fundamental limits on distributed source-channel coding.

Inspired by the simple observation noted above, we propose to study the transmission of correlated sources over the channel depicted in Figure 8.1. Labelled as the user cooperation channel (UCC) [58, 72], the channel model accounts for the presence of wireless channels between the two observers (or users). Coupled with a correlated source, it is clear that the network model based on UCC subsumes many well-studied problems as special cases; notable examples include relay channels, distributed Slepian-Wolf source coding, and multiple access channels with or without source correlation and with or without feedback.

Our main result is an achievable rate region characterization for a discrete memoryless user cooperation channel with a special class of correlated sources (sources observe a common random variable in addition to an independent random variable, see Figure 8.1). As expected, the rate region for a UCC is strictly greater than a multiple access channel, where the channels between the two users do not exist.
The new result brings us to the main question of this section: should wireless sensor nodes use or ignore the channels which exist between them? Cooperative channel coding has been shown to increase both the capacity and diversity order for improved communication efficiency. Much like performance improvement of channel coding via cooperation, we demonstrate that a “dialogue” between sensing nodes can improve source coding performance. Thus, instead of nodes source coding in a monologue (without any active collaboration with other nodes), they should participate in an active dialogue to exploit all wireless resources. A word of caution: Information theoretical optimal joint source-channel coding methods do not distinguish between source or channel coding but perform them jointly, in an inseparable manner. Our purpose for distinguishing between “cooperation” and “dialogue” is to qualitatively differentiate between channel coding and joint source-channel coding for UCC channels. In fact, the dialogue is not to remove all correlation between the sources but to jointly use it with cooperative channel coding for better communication over noisy channels.

Though the information theoretic answer clearly points to a dialogue, we have chosen to leave the question of “dialogue or monologue?” as an open question. In practice, the utility of dialogue will be affected by many important factors like the complexity and latency of coding. But the authors do believe that the commonly followed paradigm of completely distributed source coding (a la monologue) should be revisited.
The rest of this section is organized as follows. After a brief survey on the related work and on source-channel coding in Section 8.2, we define the problem and some notations and terminologies in Section 8.3. Then, we present the main result of this section in Section 8.4. This section is divided into three parts: The statement of the main result in Section 8.4.1, followed by some special cases which include some of the previously known cases of source-channel coding in Section 8.4.2 as well as some examples and discussion in Section 8.4.3. The proof of Theorem 8.4.1, which is the main result of this section, is given in Section 8.5. Finally, we conclude in Section 8.6.

8.2 Prior Works on Joint Source-Channel Coding of Correlated Sources

An introduction of noiseless coding of correlated information sources by Slepian and Wolf in 1973 [73] started a series of work in the area of distributed source coding. Their work showed a surprising result. In distributed coding of correlated information sources, even if encoders code without the knowledge of the other source outcomes, the sum rate of all encoders need not to exceed the joint entropy of the sources. It is almost equivalent to saying that, in terms of the required sum rate of the encoders, there is no penalty in performing distributed source coding in comparison to the centralized source coding. However, it should be noted that the admissible rate region for the case that encoders are separately distributed is strictly inside the rate
region of the case of centralized source coding (The cases ‘0011’ and ‘1111’ in [73]).

Inspired by [73], recent efforts [74, 75] have considered the design of actual distributed source codes and application to sensor networks. However, the result of [73] loses its charm when the sources are to be encoded and then transmitted through a noisy channel instead of an error-free link. Because removing source correlation completely with Slepian and Wolf source coding [73] followed by channel coding is not optimal in a general network (e.g. see [35] for multiple-access channel example). It should be pointed out that there are still some cases [76] based on the specific network topology and source correlation model for which the source-channel separation principle is valid. Although this is now a well established fact, this point was not made clear until the work of Cover et. al. [35] on the multiple access channel with arbitrarily correlated sources. In 1973, Slepian and Wolf first looked at the problem of transmission of correlated sources over the multiple access channel and derived a coding theorem for the problem [73]. The result of Cover et. al. [35] was more general, and it included some of the past results, including Slepian and Wolf’s source coding theorem [73], the correlated source multiple access channel capacity region of Slepian and Wolf [77], and classical multiple access channel [2]. Despite the generality of the result of Cover et. al. [35], almost a year later, Dueck [78] showed that their coding strategy is not optimal by constructing a counterexample. The optimal strategy in joint source-channel encoding requires a matching condition between the source and the channel in order to exploit the correlation between the sources to achieve the cod-
ing rates that are optimal (capacity achieving) for the channel while those rate are admissible for the sources. The problem of finding the optimal matching condition for the multiple access channel in general is still an open problem, although some minor cases have been solved [76].

The problem of the transmission of correlated sources over other channels have also been studied in the past two decades. The two-user interference channel with correlated sources was first studied in 1980 for a restricted correlation model [79], and revisited for the general correlated source model by Salehi in 1993 [80]. The broadcast channel with arbitrarily correlated sources was also considered by Han and Costa in 1987 [81]. Their results cover the best achievable rate for transmission of uncorrelated sources over the broadcast channel which is also obtained by Marton [15]. Salehi also considered two other interesting channel models, namely the restricted two-way channel and the $K$-out-of-$M$ multiple access channel with correlated sources [82, 83]. In all of the above mentioned cases only a suboptimal matching strategy has been found, illustrating the difficulty of the problem.

### 8.3 Channel and Source Models

We define a discrete memoryless user-cooperation channel (UCC) depicted in Figure 8.1 as a channel with two inputs $x_1$ and $x_2$, the output $y$, and two feedback outputs $y_1$ and $y_2$, taking values from the set of alphabets $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{Y}$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$, respectively. Therefore, the channel is denoted by $(\mathcal{X}_1 \times \mathcal{X}_2, p(y, y_1, y_2|x_1, x_2), \mathcal{Y} \times \mathcal{Y}_1 \times \mathcal{Y}_2)$ where
the channel transition probability is given by $p(y, y_1, y_2|x_1, x_2)$. Therefore, for the block of $N$ consecutive transmission we have

$$p(y^N, y_1^N, y_2^N|x_1^N, x_2^N) = \prod_{n=1}^{N} p(y_n, y_{1n}, y_{2n}|x_{1n}, x_{2n})$$

(8.1)

by the memoryless assumption of the channel, where $y^N \in \mathcal{Y}_N$, $y_1^N \in \mathcal{Y}_1^N$, $y_2^N \in \mathcal{Y}_2^N$, $x_1^N \in \mathcal{X}_1^N$, and $x_2^N \in \mathcal{X}_2^N$.

This channel has been considered by Carleial [84] as a multiple access channel with generalized feedback, for which he first established an achievable rate region. He also showed that the rate region, previously obtained for the multiple access channel with “perfect” feedback to both senders, remains achievable even if the feedback link to one of the sender is eliminated. Later, Willems [85] found an achievable rate region for this channel which improved upon Carleial’s results and, in the process, gave a simpler characterization of the rate region.

It can be noted that this channel has the flavor of the two-way channel [1] (by ignoring the output $Y$), the relay channel [7] (by ignoring output $Y_1$), the multiple access channel [2] (by ignoring both of outputs $Y_1$ and $Y_2$), and also the multiple-access channel with perfect feedback [86] (by letting both of the feedback outputs $Y_1$ and $Y_2$ be equal to $Y$). Sendonaris et al [58, 72] first proposed the idea of increasing diversity gains through cooperation between the users and appropriately named it “user cooperation diversity”. We will thus label the channel user cooperation channel (UCC), to emphasize the fact that the channel allows for a tight cooperation between users.
Assume that we have two information sources $S'$, $T'$ which generate two sequence of $S'_1, S'_2, \ldots$ and $T'_1, T'_2, \ldots$ by independently drawing a pair of discrete random variables* $S'$ and $T'$ from a given bivariate distribution $p(s', t')$. We use the following notion of the common part of two random variables.

**Definition 8.3.1.** [87] The common part $K$ of two random variables $S'$ and $T'$ is defined by the maximum integer $l$ such that there exist functions $f$ and $g$

$$f : S' \rightarrow \{1, 2, \ldots, l\}$$

$$g : T' \rightarrow \{1, 2, \ldots, l\}$$

with $\text{Prob}\{f(S') = i\} > 0$, $\text{Prob}\{g(T') = i\} > 0$, $i = 0, 1, \ldots, l$ such that $f(S') = g(T')$ with probability one. Define the common part $K = f(S')$ (= $g(T')$).

*Remark:* With a little bit of abuse of notation, we use the same symbol to denote a source and also the random variable corresponding to it.
With this definition, it is obvious that the observer of \( S' \) and \( T' \) can agree on the value of \( K \) with probability one. Note that any pair of random variables \( S' \) and \( T' \) have a trivial common part \( K = f(S') = g(T') = 1 \). Therefore, we shall say that two random variables \( S' \) and \( T' \) have a common part if \( l > 2 \) in the definition. It is also shown [87] that the common part of the i.i.d sequence of a pair of random variables \((S'_i, T'_i)\) drawn from a bivariate distribution \( p(s', t') \) is the sequence of the common parts \( K_i \).

In this section, we consider a special case of correlated sources over the user cooperation channel. We assume that the sources \( S' \) and \( T' \) each can be expressed as \((S, K)\) and \((T, K)\) where \( S, T \) and \( K \) are independent random variables (see Fig 8.1). In this case \( K \) is the common part of the random variables \( S' \) and \( T' \), and it models the “correlation” between \( S' \) and \( T' \). This situation can be interpreted as a communication scenario in which a private source \( S \) is connected to one user, and another private source \( T \) is connected to the other user. Meanwhile, a common source \( K \) is available to both users. This notion of a common part between two sources introduces a very strong form of correlation between two random variables that have been always exploited and treated separately from any other kind of correlation in derivations of source-channel coding theorems[35, 81, 77, 79].
8.4 Rate Region for User Cooperation Channel with Correlated Sources

8.4.1 Main Result

The message sources $K$, $S$, and $T$ produce random integers (messages) $W_0 \in \{1, 2, \ldots, M_0\}$, $W_1 \in \{1, 2, \ldots, M_1\}$, and $W_2 \in \{1, 2, \ldots, M_2\}$, respectively, at the beginning of each block of $N$ channel uses. Each triple $(w_0, w_1, w_2)$ occurs with probability $1/M_0M_1M_2$, which agrees with the assumption of independence between the sources $K$, $S$, and $T$.

Each encoder is completely described by a set of $N$ encoding functions. These functions map the message and the sequence of the corresponding outputs in the previous transmission into the next channel input. We have

$$x_{1n} = f_{1n}(w_0, w_1, Y_1^{n-1}), \quad (8.4)$$

$$x_{2n} = f_{2n}(w_0, w_2, Y_2^{n-1}), \quad (8.5)$$

for all $n \in \{1, 2, \ldots, N\}$, where $Y_i^{n-1} = (Y_{i1}, Y_{i2}, \ldots, Y_{in-1})$ denote the sequence of previously received signals at node $i$. The decoder estimates the messages $W_0$, $W_1$, and $W_2$ based on its knowledge of the sequence of the received $N$ channel outputs. That is

$$(\hat{w}_0, \hat{w}_1, \hat{w}_2) = g(Y^N). \quad (8.6)$$

An $(M_0, M_1, M_2, N, P_e)$-code for the d.m. UCC with two private and one common
sources (Figure 8.1), consists of two sets of \( N \) encoding functions (8.4) and (8.5), and a decoding function (8.6) such that

\[
\text{Prob}\{(\hat{w}_0, \hat{w}_1, \hat{w}_2) \neq (w_0, w_1, w_2)\} = P_e. \tag{8.7}
\]

A rate triple \((R_0, R_1, R_2)\) is said to be achievable for this channel, if for any \( \epsilon > 0 \) there exists an \((M_0, M_1, M_2, N, P_e)\)-code for sufficiently large \( N \) such that

\[
0 \leq R_0 \leq \frac{1}{N} \log(M_0), \tag{8.8}
\]

\[
0 \leq R_1 \leq \frac{1}{N} \log(M_1), \tag{8.9}
\]

\[
0 \leq R_2 \leq \frac{1}{N} \log(M_2). \tag{8.10}
\]

with small decoding error probability, i.e. \( P_e \leq \epsilon \).

\( \mathcal{R}_{UCC} \), an achievable rate region for the d.m. UCC with two private and one common sources, is the closure of the set of all achievable triples \((R_0, R_1, R_2)\). We have the following Theorem which gives a “single-letter” characterization of the achievable rate region for this channel.

**Theorem 8.4.1.** For the discrete memoryless user cooperation channel \((X_1 \times X_2, p(y, y_1, y_2|x_1, x_2), \mathcal{Y} \times \mathcal{Y}_1 \times \mathcal{Y}_2)\) with two private sources \( S \) and \( T \), one connected to each user, and with a common source \( K \), connected to both users, an achievable rate region is given by the closure of all rate triples \((R_0, R_1, R_2)\) given by

\[
0 \leq R_1 \leq I(V_1; Y_2|X_2, U, Q) + I(X_1; Y|X_2, V_1, U, Q), \tag{8.11}
\]
\[ 0 \leq R_2 \leq I(V_2; Y_1 | X_1, U, Q) + I(X_2; Y | X_1, V_2, U, Q), \]  
\[ 0 \leq R_1 + R_2 \leq \min \{ I(V_1; Y_2 | X_2, U, Q) + I(V_2; Y_1 | X_1, U, Q) \]  
\[ + I(X_1, X_2; Y | V_1, V_2, U, Q), I(X_1, X_2; Y | Q) \}, \] (8.13)
\[ 0 \leq R_0 + R_1 + R_2 \leq I(X_1, X_2; Y), \] (8.14)

for some probability mass functions \( p(q), p(u|q), p(v_1|u, q), p(v_2|u, q), p(x_1|v_1, u, q), \) \( p(x_2|v_2, u, q), \) and \( p(q, u, v_1, v_2, x_1, x_2, y, y_1, y_2) = p(q)p(u|q)p(v_1|u, q)p(v_2|u, q) p(x_1|v_1, u, q)p(x_2|v_2, u, q)p(y, y_1, y_2|x_1, x_2). \)

### 8.4.2 Special Cases

Theorem 8.4.1 establishes a very general result and subsumes many known results as a specific case. The following are examples for which the rate region of Theorem 8.4.1 coincide with the exact capacity of the channel:

1. The capacity of the discrete memoryless multiple access channel with correlated sources [77],

2. The capacity of the classical multiple access channel [2, 8],

3. The capacity of the multiple access channel with cribbing encoders [88] when the cribbing signals are causal.

Theorem 8.4.1 can also be used to provide an achievable rate (or achievable rate region) for many other channels. In the following examples, the achievable rate region given by Theorem 8.4.1 is the best known achievable rates to date:
1. An achievable rate for a discrete memoryless multiple access channel with partial feedback [86],

2. An achievable rate for the relay channel. This achievable rate coincides with the capacity when the relay channel is degraded. In the case that the channel is not degraded, this achievable rate is usually larger [7],

3. An achievable rate region for the multiple access channel with generalized feedback [85, 84, 89].

We specifically state two corollaries to Theorem 8.4.1. First, Corollary 8.4.2 shows the effectiveness of this Theorem for a channel with correlated sources [77] by showing that the rate region of Theorem 8.4.1 is good enough to characterize the capacity of the channel and to provide the corresponding source matching for the channel. Second, Corollary 8.4.3 provides an achievable rate for the discrete memoryless multiple access channel with generalized feedback [85, 84, 89]. This achievable rate region is the best known region to date. Although the form of the rate region given in Corollary 8.4.3 is different from the one which is found by Willems [85] because it has a smaller number of constraints, it can be shown that they are identical. Comparison of the rate region given by Corollary 8.4.3 with that of Carleial [84] is not easy because of the complexity of the definition of the rate region in [84]. However, we have shown examples in which the achievable rate given by Corollary 8.4.3 is outside the region found by Carleial [84] and also outside the region given by King [89].
Corollary 8.4.2. For the discrete memoryless multiple access channel with correlated sources defined in [77] as two private sources, one for each user, and one common source connected to both users, the capacity region is given by

\[ 0 \leq R_1 \leq I(X_1; Y|X_2, Q), \]  
\[ 0 \leq R_2 \leq I(X_2; Y|X_1, Q), \]  
\[ 0 \leq R_1 + R_2 \leq I(X_1, X_2; Y|Q), \]  
\[ 0 \leq R_0 + R_1 + R_2 \leq I(X_1, X_2; Y), \]

for some probability mass function \( p(q), p(x_1|q), p(x_2|q), \) where \( p(q, x_1, x_2, y) = p(q) p(x_1|q)p(x_2|q)p(y|x_1, x_2). \)

Proof. In the statement of the region \( \mathcal{R} \) in Theorem 8.4.1 by the inequalities (8.11-8.14), simply ignore the feedback signals \( Y_1 \) and \( Y_2 \) by letting \( Y_1 = Y_2 = \phi \). Also, ignore the cooperation signals \( U, V_1, V_2 \) by letting \( U = V_1 = V_2 = \phi \). \[ \square \]

Corollary 8.4.3. For the discrete memoryless multiple access channel with generalized feedback [85], an achievable rate region is given by

\[ 0 \leq R_1 \leq I(V_1; Y_2|X_2, U) + I(X_1; Y|X_2, V_1, U), \]  
\[ 0 \leq R_2 \leq I(V_2; Y_1|X_1, U) + I(X_2; Y|X_1, V_2, U), \]  
\[ 0 \leq R_1 + R_2 \leq \min\{I(V_1; Y_2|X_2, U) + I(V_2; Y_1|X_1, U)\} \]
\[ + I(X_1, X_2; Y|V_1, V_2, U), I(X_1, X_2; Y)\}, \]  
\[ (8.21) \]
for some probability mass function $p(u)$, $p(v_1|u)$, $p(v_2|u)$, $p(x_1|v_1,u)$, $p(x_2|v_2,u)$, where $p(u, v_1, v_2, x_1, x_2, y, y_1, y_2) = p(u) p(v_1|u) p(v_2|u) p(x_1|v_1,u) p(x_2|v_2,u) p(y, y_1, y_2|x_1, x_2)$.

Proof. Let $Q = \phi$ in the region $\mathcal{R}$ defined in Theorem 8.4.1 by the inequalities (8.11-8.14).

8.4.3 Examples and Discussion

It is not hard to show that the source-channel separation principle does not hold for coding correlated sources in a network setting. In fact, source coding followed by channel coding is only optimal for a single channel. The reason is because source coding effectively removes the correlation between the sources and makes the outputs of the source encoders independent. By removing the correlation we lose the opportunity to use the correlation between the sources to establish cooperation for the sake of channel coding and to increase the rate. This cooperation based on the correlation is quite familiar. Examples include, (i) coherent combining of the signals in multiple-transmit or multiple-receive antenna systems that leads to a capacity increase or (ii) cooperation between the relay node and source node in the relay channel [7]. In other words, instead of compressing the sources in the network and destroying the correlation between them, it is always better to exploit the correlation. Therefore, channel code will be designed specifically to the source distributions, i.e. source coding and channel coding are not disjointed. It follows that the optimal joint source-channel
coding scheme will match the source to the channel, which is referred to as “matching condition between the source and the channel” [86, 82, 80].

For example, consider a multiple access channel with two inputs $X_1 \in X_1$ and $X_2 \in X_2$ and an output $Y \in Y$ where $Y = X_1 + X_2$ and where the input and output alphabets are given by $X_1 = X_2 = \{0, 1\}$ and $Y = \{0, 1, 2\}$. Consider the transmission of two correlated sources $(S, T)$ with the joint distribution $p(s, t)$ such that $p(0, 0) = p(0, 1) = p(1, 1) = 1/3$ and $p(1, 0) = 0$. It can be easily seen that $H(S, T) = \log_2(3) = 1.58$ bits. If source coding and channel coding is performed separately, with the best possible coding we need to transfer a sum rate of $H(S, T) = 1.58$ bits, while the maximum sum rate of the multiple access channel with independent distributions of $p(x_1)$ and $p(x_2)$ is

$$\max_{p(x_1)p(x_2)} I(X_1, X_2; Y) = 1.5 \text{ bits.}$$

Therefore, even using Slepian-Wolf data compression on $S$ and $T$ we cannot reliably communicate these sources to the destination. However, it is simple to see that with the choice $X_1 = S$ and $X_2 = T$, error-free transmission of the sources over the channel is possible. This simple example shows that choosing the encoding signals based on the source outputs results in a better achievable rate of the transmission over the channel. Also, it reveals the fact that separating source coding from the channel coding is not optimal.

Another example comes from Corollary 8.4.2 in which two correlated sources $S' = (S, K)$ and $T' = (T, K)$ are to be transmitted over the multiple access channel. If
source coding and channel coding is done separately we have $H(S'|T') = H(S)$, $H(T'|S') = H(T)$, $H(T') = H(T) + H(K)$, and $H(S') = H(S) + H(K)$. Thus, the corner point of the Slepian and Wolf [73] data compression region is given by $(H(S), H(T) + H(K))$ and $(H(S) + H(K), H(T))$. If channel coding follows the source coding, the following conditions are required for the reliable transmission.

$$H(S) + \alpha H(K) \leq I(X_1; Y|X_2), \quad (8.22)$$

$$H(T) + (1 - \alpha) H(K) \leq I(X_2; Y|X_1), \quad (8.23)$$

$$H(S) + H(K) + H(T) \leq I(X_1, X_2; Y), \quad (8.24)$$

for some probability mass function $p(x_1), p(x_2)$, where $p(x_1, x_2, y) = p(x_1)p(x_2)p(y|x_1, x_2)$ and where $0 \leq \alpha \leq 1$ is the time sharing variable. This region is strictly inside the region given by Corollary 8.4.2.

### 8.5 Proof of Theorem 8.4.1

To prove the achievability, we use the notion of $\epsilon$-typical $N$-sequences and random coding argument. We also use the superposition block Markov encoding technique [86] to exploit the feedback signals and to introduce the correlation between the encoding process for the consecutive blocks. Here, we use the idea of backward decoding (or restricted decoding) introduced in [85] which is a very strong technique, often used to prove multi-user information theory results.
Choose a sufficiently small $\epsilon > 0$ and fix the distribution

$$p(q, u, v_1, v_2, x_1, x_2) = p(q)p(u|q)p(v_1|u, q)p(v_2|u, q)p(x_1|v_1, u, q)p(x_2|v_2, u, q).$$

Assume that the messages $W_0 \in \{1, 2, \ldots, M_0\}$, $W_1 \in \{1, 2, \ldots, M_1\}$, and $W_2 \in \{1, 2, \ldots, M_2\}$ will be sent over the channel in $N$ channel uses. In order to use superposition block Markov encoding, we transmit the common message $W_0, b \in \{1, 2, \ldots, M_0\}$, and two private messages $W_1, b \in \{1, 2, \ldots, M_1\}$ and $W_2, b \in \{1, 2, \ldots, M_2\}$ in $B$ blocks of $N$ transmissions. Now, to use backward decoding we divide each message $W_1, b$ and $W_2, b$ into two parts $W_1, b = (W_{11, b}, W_{12, b})$ and $W_2, b = (W_{21, b}, W_{22, b})$ where

$$W_{11, b} \in \{1, 2, \ldots, M_{11} = \exp(NR_{11})\},$$

$$W_{12, b} \in \{1, 2, \ldots, M_{12} = \exp(NR_{12})\},$$

$$W_{21, b} \in \{1, 2, \ldots, M_{21} = \exp(NR_{21})\},$$

$$W_{22, b} \in \{1, 2, \ldots, M_{22} = \exp(NR_{22})\},$$

which are mutually independent of each other and uniformly distributed over their respective ranges. In backward decoding, actual transmission will take place only in the $B - 1$ blocks out of possible $B$ blocks. Therefore there would be a negligible rate loss by factor of $1 - \frac{B - 1}{B} = \frac{1}{B}$ which vanishes as $B \to \infty$. As it will become more clear in the explanation of the encoding process, the messages $W_{11}$ and $W_{22}$ can be interpreted as the messages which are directly sent to the decoder. While,
the messages $W_{12}$ and $W_{21}$ contain new high rate information that is sent to user 2 and user 1, respectively, via the feedback outputs $Y_2$ and $Y_1$. Both of the encoders then cooperate in sending the resolution information to the decoder by means of the decoded message pairs $(\hat{W}_{12}, \hat{W}_{21})$ at the decoders of node 1 and 2 obtained from the transmission in the previous block. The codebook is constructed in the following way.

**Codebook Generation:**

(i) Generate $M_0$ sequences $q = (q_1, q_2, \ldots, q_N)$ each drawn according to $\text{Prob}\{q\} = \prod_{n=1}^{N} P(q_n)$. Label them $q(w_0)$, for $w_0 \in \{1, 2, \ldots, M_0\}$.

(ii) For every $q(w_0)$ generate $M_{12}M_{21}$ sequences $u = (u_1, u_2, \ldots, u_N)$ each drawn according to $\text{Prob}\{u|q(w_0)\} = \prod_{n=1}^{N} P(u_n|q_n(w_0))$. Label them $u(w_0, w_{01}, w_{02})$, for $w_0 \in \{1, 2, \ldots, M_0\}$, $w_{01} \in \{1, 2, \ldots, M_{12}\}$ and $w_{02} \in \{1, 2, \ldots, M_{21}\}$.

(iii) For every pair $q(w_0)$ and $u(w_0, w_{01}, w_{02})$ generate $M_{12}$ sequences $v_1 = (v_{11}, v_{12}, \ldots, v_{1N})$, each drawn according to $\text{Prob}\{v_1|q(w_0), u(w_0, w_{01}, w_{02})\} = \prod_{n=1}^{N} P(v_{1n}|q_n(w_0), u(w_0, w_{01}, w_{02}))$. Label them $v_1(w_0, w_{01}, w_{02}, w_{12})$, for $w_0 \in \{1, 2, \ldots, M_0\}$, $w_{01} \in \{1, 2, \ldots, M_{12}\}$, $w_{02} \in \{1, 2, \ldots, M_{21}\}$, and $w_{12} \in \{1, 2, \ldots, M_{12}\}$. 
(iv) For every pair \( q(w_0) \) and \( u(w_0, w_{01}, w_{02}) \) generate \( M_{21} \) sequences \( v_2 = (v_{21}, v_{22}, \ldots, v_{2N}) \), each drawn according to \( \text{Prob}\{v_2|q(w_0), u(w_0, w_{01}, w_{02})\} = \prod_{n=1}^{N} P(v_{2n}|q_n(w_0), u(w_0, w_{01}, w_{02})) \). Label them \( v_2(w_0, w_{01}, w_{02}, w_{21}) \), for \( w_0 \in \{1, 2, \ldots, M_0\} \), \( w_{01} \in \{1, 2, \ldots, M_{12}\} \), \( w_{02} \in \{1, 2, \ldots, M_{21}\} \), and \( w_{21} \in \{1, 2, \ldots, M_{21}\} \).

(v) For every triple \( q(w_0) \), \( u(w_0, w_{01}, w_{02}) \), and \( v_1(w_0, w_{01}, w_{02}, w_{12}) \) generate \( M_{11} \) sequences \( x_1 = (x_{11}, x_{12}, \ldots, x_{1N}) \), each drawn according to \( \text{Prob}\{x_1|q(w_0), u(w_0, w_{01}, w_{02}), v_1(w_0, w_{01}, w_{02}, w_{12})\} = \prod_{n=1}^{N} P(x_{1n}|q_n(w_0), u(w_0, w_{01}, w_{02}), v_1(w_0, w_{01}, w_{02}, w_{12})) \). Label them \( x_1(w_0, w_{01}, w_{02}, w_{12}, w_{11}) \), for \( w_0 \in \{1, 2, \ldots, M_0\} \), \( w_{01} \in \{1, 2, \ldots, M_{12}\} \), \( w_{02} \in \{1, 2, \ldots, M_{21}\} \), \( w_{12} \in \{1, 2, \ldots, M_{12}\} \), and \( w_{11} \in \{1, 2, \ldots, M_{11}\} \).

(vi) For every triple \( q(w_0) \), \( u(w_0, w_{01}, w_{02}) \), and \( v_2(w_0, w_{01}, w_{02}, w_{21}) \) generate \( M_{22} \) sequences \( x_2 = (x_{21}, x_{22}, \ldots, x_{2N}) \) each drawn according to \( \text{Prob}\{x_2|q(w_0), u(w_0, w_{01}, w_{02}), v_2(w_0, w_{01}, w_{02}, w_{21})\} = \prod_{n=1}^{N} P(x_{2n}|q_n(w_0), u(w_0, w_{01}, w_{02}), v_2(w_0, w_{01}, w_{02}, w_{21})) \). Label them \( x_2(w_0, w_{01}, w_{02}, w_{21}, w_{22}) \), for \( w_0 \in \{1, 2, \ldots, M_0\} \), \( w_{01} \in \{1, 2, \ldots, M_{12}\} \), \( w_{02} \in \{1, 2, \ldots, M_{21}\} \), \( w_{21} \in \{1, 2, \ldots, M_{21}\} \), and \( w_{22} \in \{1, 2, \ldots, M_{22}\} \).
Encoding:

The message $w_0$, $w_1 = (w_{12}, w_{11})$, and $w_2 = (w_{21}, w_{22})$ are now encoded using superposition block Markov encoding in $B$ blocks. Let $w_{0,b}$, $w_{1,b} = (w_{12,b}, w_{11,b})$, and $w_{2,b} = (w_{21,b}, w_{22,b})$ denote the message to be transmitted in block $b$, for $b = 1, 2, \ldots, B$.

(i) In block $b = 1$, we have

\[
\begin{align*}
x_{11} &= x_1(w_{0,1}, 1, 1, w_{12,1}, w_{11,1}) \quad \text{and} \\
x_{21} &= x_2(w_{0,1}, 1, 1, w_{21,1}, w_{22,1})
\end{align*}
\]  

(ii) In block $b$, $b = 2, 3, \ldots, B$, because of the feedback links, encoder of user 1 has the estimate $w'_{21,b-1}$ of the cooperation signal $w_{21,b-1}$ transmitted by user 2 in previous block, $b - 1$, and the encoder of user 2 has the estimate $w'_{12,b-1}$ of the cooperation signal $w_{12,b-1}$ transmitted by user 1 in previous block, $b - 1$. Then in block $b$, $b = 2, 3, \ldots, B - 1$, we have

\[
\begin{align*}
x_{1b} &= x_1(w_{0,b}, w_{12,b-1}, w'_{21,b-1}, w_{12,b}, w_{11,b}) \quad \text{and} \\
x_{2b} &= x_2(w_{0,b}, w'_{12,b-1}, w_{21,b-1}, w_{21,b}, w_{22,b}).
\end{align*}
\]

Finally in block $B$, we have

\[
\begin{align*}
x_{1B} &= x_1(1, w_{12,B-1}, w'_{21,B-1}, 1, 1) \quad \text{and} \\
x_{2B} &= x_2(1, w'_{12,B-1}, w_{21,B-1}, 1, 1).
\end{align*}
\]
Decoding:

We will now explain the decoding process for the decoder of the destination node and two decoders of the users. (a) Decoding at the users is quite straightforward, and it is done in order to obtain the cooperation signal from the other node. This signal is necessary to establish cooperation, and basically it would result in a gain for UCC over the multiple access channel. (b) The main decoding occurs at the destination node, and we will use the idea of backward decoding to find the transmitted information. If in a decoding step there are no message indices to satisfy the decoding rule or if there are more than one (vector or n-tuple), then the index (vector or n-tuple) is chosen at random or an error is declared. For the decoding steps (a) and (b), we have

(a) After each block \( b, b = 1, 2, \ldots, B - 1 \), encoder 1 chooses \( w'_{21,b} \) such that

\[
(q(w_0,b), u(w_0,b, w_{12,b-1}, w'_{21,b-1}), v_1(w_0,b, w_{12,b-1}, w'_{21,b-1}, w_{12,b}), v_2(w_0,b, w_{12,b-1}, w'_{21,b-1}, w'_{21,b})x_1(w_0,b, w_{12,b-1}, w'_{21,b-1}, w_{12,b}, w_{11,b}), y_{1b})
\]

\[\in \mathcal{A}_e(Q, U, V_1, V_2, X_1, Y_1), \quad (8.31)\]

where \( \mathcal{A}_e(\cdot) \) represents the jointly typical set of the arguments. The decoder 2 chooses \( w'_{12,b} \) such that

\[
(q(w_0,b), u(w_0,b, w'_{12,b-1}, w_{21,b-1}), v_1(w_0,b, w'_{12,b-1}, w_{21,b-1}, w'_{12,b}, w_{12,b}), y_{1b})
\]
\begin{equation}
\begin{aligned}
v_2(w_{0,b}, w'_{12,b-1}, w_{21,b-1}, w_{21,b})x_2(w_{0,b}, w'_{12,b-1}, w_{21,b-1}, w_{21,b}, w_{22,b}, y_{2b}) \\
\in \mathcal{A}_c(Q, U, V_1, V_2, X_2, Y_2).
\end{aligned}
\tag{8.32}
\end{equation}

(b) To do the backward decoding at the destination node, the decoder starts from the last block, block \( B \), and looks for a pair \((\hat{w}_{01,B}, \hat{w}_{02,B})\) such that

\begin{equation}
\begin{aligned}
(q(1), u(1, \hat{w}_{01,B}, \hat{w}_{02,B}), v_1(1, \hat{w}_{01,B}, \hat{w}_{02,B}, 1), \\
v_2(1, \hat{w}_{01,B}, \hat{w}_{02,B}, 1)x_1(1, \hat{w}_{01,B}, \hat{w}_{02,B}, 1, 1), y_B) \\
\in \mathcal{A}_c(Q, U, V_1, V_2, X_1, X_2, Y_1).
\end{aligned}
\tag{8.33}
\end{equation}

Now, in block \( B-1 \), the decoder already has an estimate for \( \hat{w}_{12,B-1} \) and \( \hat{w}_{21,B-1} \) as \((\hat{w}_{12,B-1}, \hat{w}_{21,B-1}) = (\hat{w}_{01,B}, \hat{w}_{02,B})\). Then, the decoder looks for the 5-tuple \((\hat{w}_{01,B-1}, \hat{w}_{02,B-1}, \hat{w}_{0,B-1}, \hat{w}_{11,B-1}, \hat{w}_{22,B-1})\) such that

\begin{equation}
\begin{aligned}
(q(\hat{w}_{0,B-1}), u(\hat{w}_{0,B-1}, \hat{w}_{01,B-1}, \hat{w}_{02,B-1}), \\
v_1(\hat{w}_{0,B-1}, \hat{w}_{01,B-1}, \hat{w}_{02,B-1}, \hat{w}_{12,B-1}), v_2(\hat{w}_{0,B-1}, \hat{w}_{01,B-1}, \hat{w}_{02,B-1}, \hat{w}_{21,B-1}), \\
x_1(\hat{w}_{0,B-1}, \hat{w}_{01,B-1}, \hat{w}_{02,B-1}, \hat{w}_{12,B-1}, \hat{w}_{11,B-1}), \\
x_2(\hat{w}_{0,B-1}, \hat{w}_{01,B-1}, \hat{w}_{02,B-1}, \hat{w}_{21,B-1}, \hat{w}_{22,B-1}), y_{B-1}) \\
\in \mathcal{A}_c(Q, U, V_1, V_2, X_1, X_2, Y_1).
\end{aligned}
\tag{8.34}
\end{equation}

Therefore, in this step, the decoder finds the messages \((\hat{w}_{0,B-1}, \hat{w}_{11,B-1}, \hat{w}_{22,B-1})\) of the common sources and both of the private sources for the block \( B-1 \) as
well as the cooperative messages \((\hat{w}_{01,B-1}, \hat{w}_{02,B-1})\). At this point, the decoder has an estimate for \(\hat{w}_{12,B-2}\) and \(\hat{w}_{21,B-2}\) through decoded message \(\hat{w}_{01,B-1}\) and \(\hat{w}_{02,B-1}\) as \((\hat{w}_{12,B-2}, \hat{w}_{21,B-2}) = (\hat{w}_{01,B-1}, \hat{w}_{02,B-1})\) which is necessary to perform decoding of the block \(B - 2\). Continuing this trend, the decoder successively finds the cooperative signals at the current block and uses it to perform decoding of the messages in the previous blocks.

*Probability of error analysis:*

We are left to prove that the average error probability over the ensemble of \(B - 1\) block of transmission can be made smaller than any positive number \(\epsilon > 0\), if the rate triple \((R_0, R_1, R_2)\) is inside the rate region defined by (8.11)-(8.14). Although writing the expressions for the decoders (a) and (b) is cumbersome, it is quite straightforward and no special treatment is needed. It can be easily verified that the constraints

\[
R_{11} \leq I(X_1; Y|X_2, V_1, U, Q), \tag{8.35}
\]

\[
R_{22} \leq I(X_2; Y|X_1, V_2, U, Q), \tag{8.36}
\]

\[
R_{11} + R_{22} \leq I(X_1, X_2; Y|V_1, V_2, U, Q), \tag{8.37}
\]

\[
R_{11} + R_{22} + R_{12} + R_{21} \leq I(X_1, X_2; Y|Q), \tag{8.38}
\]

\[
R_0 + R_{11} + R_{22} + R_{12} + R_{21} \leq I(X_1, X_2; Y), \tag{8.39}
\]

are required to obtain a small error probability for the decoding operation performed by the decoder of the destination node. It is also clear that each user can reliably
decode the information which is sent by the other user with a small enough probability of error provided that

\[ R_{12} \leq I(V_1; Y_2 | X_2, U, Q), \]  
\[ R_{21} \leq I(V_2; Y_1 | X_1, U, Q). \]

From the definition of \( R_{12}, R_{21}, R_{11}, \) and \( R_{22}, \) we have \( R_1 = R_{12} + R_{11} \) and \( R_2 = R_{21} + R_{22}. \) Now, we need to show that the region \( \mathcal{R} \) defined by (8.11)-(8.14) is exactly equal to the rate region \( \mathcal{R}_1 \) which corresponds to the set of all rate triples \( (R_0, R_1, R_2) \) given by the necessary conditions (8.35)-(8.41) in order to guarantee a small enough decoding error probability. By examining the regions \( \mathcal{R} \) and \( \mathcal{R}_1 \) it is immediate that \( \mathcal{R}_1 \subseteq \mathcal{R}. \) Therefore, we still have to prove that \( \mathcal{R} \subseteq \mathcal{R}_1. \) Consider \( (R_0, R_1, R_2) \in \mathcal{R} \) and let

\[ R_{12} = \min\{R_1, I(V_1; Y_2 | X_2, U, Q)\}, \]  
\[ R_{21} = \min\{R_2, I(V_2; Y_1 | X_1, U, Q)\}. \]

It follows that

\[ 0 \leq (R_1 - R_{12}) \leq I(X_1; Y | X_2, V_1, U, Q), \]
\[ 0 \leq (R_2 - R_{21}) \leq I(X_2; Y | X_1, V_2, U, Q), \]
\[ (R_1 - R_{12}) + (R_2 - R_{21}) \leq I(X_1, X_2; Y | V_1, V_2, U, Q), \]
\[ (R_1 - R_{12}) + (R_2 - R_{21}) + R_{12} + R_{21} \leq I(X_1, X_2; Y | Q), \]
\[ R_0 + (R_1 - R_{12}) + (R_2 - R_{21}) + R_{12} + R_{21} \leq I(X_1, X_2; Y). \]
Now if we set $R_{11} = (R_1 - R_{12})$ and $R_{22} = (R_2 - R_{21})$, it can be easily seen that any triple $(R_0, R_1, R_2)$ in the region $\mathcal{R}$ corresponds to a 5-tuple $(R_0, R_{12}, R_{21}, R_{11}, R_{22})$ in the region $\mathcal{R}$ which establishes that $\mathcal{R} \subseteq \mathcal{R}_1$.

*Remark:* Note that the region described in the Theorem is convex. Therefore, introduction of the time sharing variable or the use of convex hull operation is not needed in characterizing the achievable rate region $\mathcal{R}$.

8.6 Conclusion

In this section, we have analyzed the problem of joint source-channel coding over user cooperation channels with correlated information sources. This channel model includes the relay channel, the two-way channel, and the multiple-access channel with or without feedback as special cases. Therefore, our result reduces to the known result for these channels with correlated sources. Our analysis of a user cooperation channel with correlated sources reveals two facts:

1. First is emphasizing a well established fact: in the presence of the noise in the channel between the transmitter and receiver, Slepian and Wolf source coding followed by channel coding is not optimal.

2. Second is that encoding nodes should talk to each other not only for the sake of channel coding (which is referred to as cooperation), but also for the sake of joint source-channel encoding (that we label as dialogue).
Chapter 9

Code Design For the Relay Channel

Recent information theoretical results have shown a considerable improvement in the performance of communication systems through the use of relaying and cooperation. Despite a collection of strong information theoretical results on the relay channel, there has been almost no attention given to real code design for the relay channel. In this section, we present a powerful modular code design approach for the relay channel and corresponding decoding algorithms based on the factor graph representation of the code. The presented code construction stems from the capacity analysis of the Gaussian relay channel constrained to a finite input alphabet. For most of the relay channel conditions, the constructed code with a low-complexity simple relay protocol outperforms any possible code design for the direct channel by achieving an $E_b/N_0$ below the minimum required $E_b/N_0$ of a single-link transmission. Moreover, the designed codes close the gap to less than 1dB (at a BER of $10^{-6}$) of the Shannon limit for the relay channel with a code length of only $2 \times 10^4$ bits.

9.1 Introduction

There has been an increasing demand for high speed wireless communication in the last few years. Multiple antenna systems have traditionally been used to provide
reliable and fast point-to-point communication over a wireless channel. However, recent results have shown that cooperative relaying strategies obtain substantial gains over multiple antenna systems using the same total power and number of antennas [47, 38, 46]. Nevertheless, up to this point there has been almost no attention to the problem of real code design for the relay channel beyond the information theoretical analysis. In fact, relay channel code design introduces new challenges because of the distributed nature of the coding scheme at the source and relay nodes.

In this work, we consider a code design for the Gaussian relay channel. The contributions of this section are threefold. First, we derive the information theoretic capacity bounds for the Gaussian relay channel subject to a finite alphabet constraint. Based on the derived bounds we present the minimum required energy per bit (per noise variance), $E_b/N_0$, for reliable communication over the relay channel, which gives the appropriate benchmark to evaluate the performance of the designed codes. Second, we present a modular code construction that integrates three main components: (i) protocol design, (ii) constituent codes design, and (iii) the optimal power allocation. Although, it is usually possible to find a code that performs relatively well over a practical range of the channel conditions, this modular structure allows us to adapt the codes to the channel conditions and properties of the transmission media. Third, besides the optimal decoding strategy, we present two iterative decoding techniques, the forward and the backward decoding schemes, with significantly lower complexity. The backward decoding scheme exploits the idea of the information theoretical
decode-and-forward coding scheme [7, 54] and hence has good performance when the relay node is located relatively close to the source node. The forward decoding scheme exploits the idea of the information theoretical estimate-and-forward coding scheme [54, 38] and as a result has good performance when the relay node is located relatively far from the source node, e.g., about half way or more between the source and destination nodes.

Many coding protocols for the relay channel have limited performance because orthogonal transmission between the source-destination and relay-destination links. In particular the coding scheme of [90] assumes a half duplex relay and therefore TDM for the reception and transmission from the relay, which can be thought of as a ‘cheap’ [26, 28] transmission scheme. Moreover, the available bandwidth is divided into two parts and the relay and the source are assumed to transmit in two different frequency bands. Thus, the signals transmitted from the relay and the source would not interfere, and therefore, no joint decoding is necessary at the destination. As a result, there is a rate loss from not exploiting all available dimensions for transmission. Our use of joint decoding on the destination resolve the interference from the source and the relay.

The rest of this section is organized as follows. After some preliminaries in Section 9.2, we derive lower and upper bounds on the capacity of the constrained Gaussian relay channel with binary input alphabets in Section 9.3. The general code design is then presented in Section 9.4. In Section 9.5, we present a specific relay
channel code construction based on the idea of low-density parity check codes. We present the performance of the designed codes in comparison to single-link codes and also to the fundamental limits of the channel in Section 9.6, and finally we conclude in Section 9.7.

9.2 Preliminaries

A relay channel, shown in Figure 9.1, consists of an input $x_1$, a relay-received signal $y_1$, a relay-transmit signal $x_2$ (which depends only on the past values of $y_1$), and a channel output $y$. The channel is assumed to be memoryless. The dependency of the outputs on the inputs are as follows: the channel output is $y = h_1x_1 + h_2x_2 + z$, and the relay received signal is given by $y_1 = h_0x_1 + z_1$. The inter-channel gains $h_0$, $h_1$, and $h_2$ are assumed to be constant, and $z \sim \mathcal{N}(0,N)$ and $z_1 \sim \mathcal{N}(0,N_1)$ are independent Gaussian noises. The input power constraints are given by $E[X_1^2] < P_1$ and $E[X_2^2] < P_2$. In the sequel, we denote $\gamma_0 = \frac{|h_0|^2}{N_1}$, $\gamma_1 = \frac{|h_1|^2}{N}$ and $\gamma_2 = \frac{|h_2|^2}{N}$.

Because in practice the relay position gives a better sense in terms of network
topology than the abstract relay channel parameters, we consider the model shown in Figure 9.2 for evaluation of the proposed relay channel codes. We assume that both receivers have equal noise power spectral density $N = N_1 = N_0$. We normalize all the distances based on the source-destination distance; therefore, the source-destination distance is set to unity in Figure 9.2. To simplify the model, the relay is positioned along a straight line between the source and destination at a distance $d$ from the source. Based on the value of the distance $d$, then $\gamma_0 = \frac{1}{d^\alpha}$, $\gamma_1 = 1$ and $\gamma_2 = \frac{1}{(1-d)^\alpha}$, where $\alpha$ is the pathloss exponent and typically lies in the range $[2, 5)$. The set of channel gains $(\gamma_0, \gamma_1, \gamma_2)$ are assumed to be fixed over time.

### 9.3 Capacity of the Gaussian relay channel with finite alphabets

Although the channel input for the Gaussian or fading channels can be continuous, current practical codes are subject to a finite input alphabet constraint. Thus, in order to evaluate the performance of the proposed code, we analyze the capacity of the relay channel subject to the finite input alphabet constraint. Because this section
presents the design of binary codes, the capacity analysis is accordingly performed for the choice of the binary signaling from the source and the relay.

In this section, we analyze the fundamental limits of information transfer in the relay channel with binary signaling for given power constraints. This limit in turn translates to the minimum required $E_b/N_0$ (Energy per bit / Receiver noise variance) to reliably communicate through the relay channel for a specified rate of transmission.

Because the capacity of the relay channel is not known in general, the required $(E_b/N_0)_{\text{min}}$ cannot precisely be established. However lower and upper bounds on $(E_b/N_0)_{\text{min}}$ are presented in this section, and they are usually very close. El Gamal and Zahedi [51] have also looked at the problem of minimum required $E_b/N_0$ for the frequency division Gaussian relay channel with a Gaussian input alphabet. However, their analysis does not provide the right benchmark on the performance of the binary codes over the relay channel because (i) the input is continuous and (ii) the frequency division relay channel is more constrained than the general relay channel.

In addition to the analysis of the minimum required $E_b/N_0$, the capacity analysis of this section provides great insights into the problem of actual code design. Moreover, the optimal values of the parameters of the designed codes are also a byproduct of the capacity analysis. One of the most important parameters is the power allocation between the relay and the source. This power allocation strongly affects the overall performance of the relay channel code, and it should be carefully designed.
9.3.1 Upper and lower bounds on the capacity

The best known upper bound for the relay channel is obtained through a simple application of the cut-set theorem [7, 8] and is given by

\[ C \leq C_{UB} = \max_{p(x_1, x_2)} \min\{I(X_1; Y, Y_1|X_2), I(X_1, X_2; Y)\} \]  \hspace{1cm} (9.1)

where the first mutual information term is the cut-set bound on the broadcast cut of the relay channel and where the second mutual information term is the cut-set bound on the multiple access cut. It should be noted that the correlation between the channel input signal \(X_1\) and the relay transmit signal \(X_2\) makes the second mutual information term in (9.1) larger, while it makes the first term smaller. For the Gaussian relay channel depicted in Figure 9.1, the cut-set bound leads to the following theorem.

**Theorem 9.3.1.** An upper bound on the capacity of the Gaussian relay channel with binary input signals defined in Section 9.2 is given by

\[ C_{UB} = \max_{\rho} \min\{I_{BC}(\rho), I_{MA}(\rho)\}, \]  \hspace{1cm} (9.2)

where

\[ I_{BC}(\rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sum_{b_2=1}^{2} \frac{1}{2} -f_{Y, Y_1}(y, y_1|X_2 = b_2) \log_2(f_{Y, Y_1}(y, y_1|X_2 = b_2)) dy_1 dy - \]

\[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} -f_{Z, Z_1}(z, z_1) \log_2(f_{Z, Z_1}(z, z_1)) dz_1 dz \]  \hspace{1cm} (9.3)

\[ I_{MA}(\rho) = \int_{-\infty}^{\infty} -f_Y(y) \log_2(f_Y(y)) dy - \int_{-\infty}^{\infty} -f_Z(z) \log_2(f_Z(z)) dz \]  \hspace{1cm} (9.4)
and the density functions are given by

\[ f_{Y_1}(y, y_1 | X_2 = b_2) = \sum_{b_1 = -1, 1} 2f(b_1, b_2)f_Z(y - b_1 \sqrt{P_1 \gamma_1})f_Z(y - b_1 \sqrt{P_1 \gamma_0}), \]

\[ f_Y(y) = \sum_{b_1 = -1, 1} \sum_{b_2 = -1, 1} f(b_1, b_2)f_Z(y + b_1 \sqrt{P_1 \gamma_1} + b_2 \sqrt{P_2 \gamma_2}), \quad (9.5) \]

\[ f_{Z_1}(z, z_1) = f_Z(z)f_Z(z_1), \quad (9.6) \]

\[ f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{\frac{z^2}{2\sigma^2}}. \quad (9.7) \]

Furthermore, the optimal input distribution \( f(b_1, b_2) = \text{Prob}[X_1 = b_1, \text{and} \ X_2 = b_2] \) is such that \( f(1, 1) = f(-1, -1) = \frac{1 + \rho}{4}, \ f(1, -1) = f(-1, 1) = \frac{1 - \rho}{4}. \)

**Proof:** Because the input signals of the source input, \( X_1 \), and the relay signal, \( X_2 \), are both binary, the maximum of the above mutual expression subject to the given power constraint is simply obtained by antipodal signal sets. Moreover, the probability of each of the two possible signals should be equal to \( \frac{1}{2} \). However, there could be a correlation between the source input, \( X_1 \), and the relay signal, \( X_2 \). Let \( \rho \) denote the correlation between \( X_1 \) and \( X_2 \), where (i) \( \rho = 0 \) means that the two signals are uncorrelated and (ii) \( \rho = 1 \) and \( \rho = -1 \) means that the two signals are fully correlated. Moreover, the case \( \rho = 1 \) denotes positive correlation, i.e., if \( X_1 = i \) then \( X_2 = i \) for \( i = -1, 1 \), and the case \( \rho = -1 \) denotes negative correlation, i.e., if \( X_1 = i \) then \( X_2 = -i \) for \( i = -1, 1 \). We find the mutual information terms in (9.1) considering the fact that the input signals are discrete while the output signals are continuous. After some manipulation the bound can be written in the form of (9.2)-(9.7).
However, the above upper bound has not been shown to be achievable in general. Cover and El Gamal [7] presented a strong way of encoding for the relay channel based on the idea of block Markov encoding. In this coding technique, transmission of the information occurs in blocks, where the source node sends some new information at each block. The relay node decodes the transmitted information at each block and sends helper information as a codeword in the next block. The destination node then uses the transmitted information from the source along with the helper information from the relay to correctly decode each blocks. This coding scheme is also known as the decode-and-forward coding scheme [7, 54] because of the need for perfect decoding at the relay. Using the same idea of decode-and-forward, we drive the following lower bound on the capacity of the Gaussian channel subject to binary signals both at the source and relay nodes.

**Theorem 9.3.2.** A lower bound on the capacity of the Gaussian relay channel with binary input signals defined in Section 9.2 is given by

\[ C_{DF} = \max_{\rho} \min \{ I_{SR}(\rho), I_{MA}(\rho) \}, \]  

where

\[ I_{SR}(\rho) = \int_{-\infty}^{\infty} \frac{1}{2} \sum_{b_2 = -1, 1} -f_{Y_1}(y_1|X_2 = b_2) \log_2(f_{Y_1}(y_1|X_2 = b_2)) dy_1 - \int_{-\infty}^{\infty} -f_Z(z_1) \log_2(f_Z(z_1)) dz_1 dz \]  

and

\[ I_{MA}(\rho) = \int_{-\infty}^{\infty} -f_Y(y) \log_2(f_Y(y)) dy - \int_{-\infty}^{\infty} -f_Z(z) \log_2(f_Z(z)) dz \]
and the density functions are given by

\[
f_{Y_1}(y_1|X_2 = b_2) = \sum_{b_1=-1,1} 2f(b_1, b_2)f_Z(y - b_1\sqrt{P_1\gamma_0}), \quad (9.11)
\]

\[
f_Y(y) = \sum_{b_1=-1,1} \sum_{b_2=-1,1} f(b_1, b_2)f_Z(y + b_1\sqrt{P_1\gamma_1} + b_2\sqrt{P_2\gamma_2}), \quad (9.12)
\]

\[
f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{z^2}{2\sigma^2}}. \quad (9.13)
\]

Furthermore, the optimal input distribution \( f(b_1, b_2) = \text{Prob}[X_1 = b_1, \text{and } X_2 = b_2] \) is such that \( f(1, 1) = f(-1, -1) = \frac{1+\rho}{4}, f(1, -1) = f(-1, 1) = \frac{1-\rho}{4} \).

**Proof:** The proof is similar to that of Theorem 9.3.1. However, the lower bound is obtained through block Markov coding technique introduced in [7] for the relay channel as follows

\[
C \leq C_{DF} = \max_{p(x_1, x_2)} \min\{I(X_1; Y|X_2), I(X_1, X_2; Y)\}. \quad (9.14)
\]

This achievable rate is also known as the decode-and-forward scheme [54] which is known to have a good performance when the relay node receives a better signal than the destination node. Again, after some manipulation of the mutual information terms in (9.14) considering that (i) the input signals are discrete and the output signals are continuous, (ii) antipodal signals both at the relay and the source maximizes the mutual information, and (iii) the source and relay may be correlated in general, the bound can be written in the form of (9.8)-(9.13).

Here, two notes are in order about the correlation factor between the source input and the relay signal. First, it should be noted that the optimal value of the correlation
factors $\rho$ in the bounds of Theorem 9.3.1 and 9.3.2 are not usually equal. Second, the
correlation factor can be thought to vary with time. However, it can be shown that
a variable correlation factor would not help to obtain better upper or lower bounds.
In fact, the optimal value of the correlation factor in either of Theorem 9.3.1 or 9.3.2
is constant over time.

9.3.2 Analysis of the minimum energy per bit required for reliable com-
munications

Based on the derived upper and lower bounds on the capacity of the relay channel
with binary inputs (Theorem 9.3.1 and 9.3.2), we can bound the minimum required
$E_b/N_0$ for reliable transmission. The energy per bit for the relay channel can be viewed in two ways. The first case is when the relay node has a separate source of
power and assists the transmission between the source and relay node. Therefore,
the relay power is not counted towards the calculation of $E_b/N_0$. In this case, for all
possible relay channel conditions there is a gain in the capacity of the relay channel
with respect to single-link transmission [54]. Second is the case that the sum power
of the source and relay nodes which contribute towards the transmission of the source
information is counted in the calculation of energy per bit. In this section we focus
only on the sum power case which is clearly more restrictive. Nonetheless, the sum
power scenario is more interesting because it shows the true power of relaying and
cooperative coding more effectively.
The lower bound on \((E_b/N_0)_{\text{min}}\) is obtained through Theorem 9.3.1 and denotes the least required energy per bit per noise spectral density that any possible transmission scheme requires to reliably communicate over a Gaussian relay channel with binary signaling.

**Proposition 9.3.3.** Let \(R_{UB}(P)\) denote the maximum value of the upper bound on capacity, \(C_{UB}\), in Theorem 9.3.1 subject to a sum power constraint \(P_1 + P_2 = P\). A lower bound on the minimum energy per bit for reliable communication over the Gaussian relay channel with binary signaling is given by

\[
(E_b/N_0)_{\text{min}}(R^*) = \frac{R_{UB}^{-1}(R^*)}{N_0 R^*},
\]

(9.15)

where \(R_{UB}^{-1}(\cdot)\) denotes the inverse function of the maximum possible transmission rate \(R_{UB}(\cdot)\).

Although a lower bound on \((E_b/N_0)_{\text{min}}\) is given in (9.15), no coding scheme is known to achieve it. Therefore, we consider the required \((E_b/N_0)_{\text{min}}\) of the decode-and-forward coding scheme [7, 54], which provides a more realistic lower bound for practical code design.

**Proposition 9.3.4.** Let \(R_{DF}(P)\) denote the maximum value of the achievable rate of the decode-and-forward scheme, \(C_{DF}\), in Theorem 9.3.2 subject to a sum power constraint \(P_1 + P_2 = P\). The minimum energy per bit for reliable communication over the relay channel with binary signaling by using the decode-and-forward coding
scheme [7] is given by

\[(E_b/N_0)_{\text{min}}(R^*) = \frac{R_{DF}^{-1}(R^*)}{N_0 R^*}\]  \hspace{1cm} (9.16)

where \(R_{DF}^{-1}(\cdot)\) denotes the inverse function of the maximum possible transmission rate \(R_{DF}(\cdot)\).

Figure 9.3 shows the lower bound of Proposition 9.3.3 on \((E_b/N_0)_{\text{min}}\) as well as the required \((E_b/N_0)_{\text{min}}\) for the decode-and-forward coding scheme in Proposition 9.3.4, which will be used later to evaluate the performance of the proposed codes. It is interesting to note that the decode-and-forward strategy performs very close to optimal for the location of the relay close to the source, say \(d \leq 0.25\), as the corresponding curve almost coincides with the lower bound in Figure 9.3.

One of the most important aspects of the code design for the relay channel with the sum power constraint is the optimal power allocation between the relay and the source nodes. The power allocation defined in terms of \(\kappa \triangleq \frac{P_2}{P_1 + P_2}\) is the ratio of the relay power to the sum power of the source and the relay. The optimal value \(\kappa\) denoted as \(\kappa^*\) can be found by maximizing the rate of transmission in Theorem 9.3.2 subject to a sum power constraint \(P_1 + P_2 = P\), which does not lead to a closed form expression in general. However, for some specific relay protocols, simple analytic expression for the power allocation can be found [for example see (9.17)-(9.19)]. It should be pointed out that a non-optimal choice of power splitting factor \(\kappa\) might lead to a huge performance degradation. For example, using the decode-and-forward
Figure 9.3: Upper and lower bound on the required $E_b/N_0$ for the given transmission rate.
Figure 9.4: Effect of the power allocation (splitting) on the rate, for a given $E_b/N_0$. The optimal value of the power splitting factor is denoted by $\kappa^*$ which corresponds to the $(E_b/N_0)_{\text{min}} = -3.5 dB$ required to achieve rate $R = 0.5$ with the decode-and-forward scheme.

Scheme a rate of 0.5 can be achieved by an $(E_b/N_0)_{\text{min}} = -3.5 dB$, as can be seen in Figure 9.3. Figure 9.4 shows that for the $(E_b/N_0)_{\text{min}} = -3.5 dB$, only using the optimal value of the power splitting factor, $\kappa = \kappa^*$, can achieve the rate of 0.5. It can be seen that by simply using a value of $\kappa = 0.7$, instead of the optimal value of $\kappa^* = 0.33$, more than a 45% rate loss occurs.
9.4 General Code Design Approach for the Relay Channel

The primary difficulty of code construction for the relay channel that makes it inherently different from ordinary single-link code design is because the distributed nature of coding at the source and relay nodes. Therefore, the main challenges are designing the forwarding strategy at the relay, and also the corresponding joint coding between the source and relay nodes. The forwarding strategy shows how to build the relay transmit signal based on the past relay received signals. For this strategy, two codebooks should be generated, one to be used by the encoder at the source node, and one for the encoder at the relay node. In this document, we refer to the specific choice of the relaying function as a protocol, and the codes that are used at the source and relay nodes, we refer to as constituent codes. Because the power splitting between the source and relay significantly affects the performance of the designed code, and should be considered as a part of code construction, the proposed construction has three major elements: protocol, constituent codes, and power assignments.

9.4.1 Protocol

In this section we discuss two protocols: a simple protocol and a DF-protocol. The relay protocols in code design correspond to the same notion of protocols in an information theoretical sense for the relay channel [54]. Although the presented simple protocol in this section is inspired by combining both the decode-and-forward and the estimate-and-forward coding schemes [54], it is much simpler but effective. Our nu-
merical results show that for most of the relay channel conditions, this simple protocol surpasses the performance of any single-link code. The next protocol relies heavily on successful decoding at the relay node and exploits the concept of coherent combining at the destination. Therefore, it is closer to the decode-and-forward coding scheme. However, there is still no rate control in the relay-destination link which make this protocol slightly different from its information theoretical counterpart.

9.4.1.1 A simple protocol

The transmission of the information from the source occurs in B blocks $b = 1, 2, \ldots, B$ of equal length $N$. Every two consecutive blocks use two jointly designed codes called constituent codes. In a simple design, one can choose only two constituent codes which are used alternately in the blocks with an odd or even index.

At each block $b$, the source sends a new codeword $w_b$. At the end of block $b$, the relay node estimates the transmitted codeword $w_b$ from the source by using the relay’s received signal in this block. This estimate, $\hat{w}_b$, is the closest codeword to the received signal which is then sent in block $b + 1$ without the need to re-encode. It should be noted that if the source-relay link is good, $\hat{w}_b$ is most likely decoded correctly, thus, this protocol resembles the decode-and-forward coding scheme $\hat{w}_b$. However, when the source-relay link is not good, $\hat{w}_b$ can be interpreted as the best estimate of the relay received signal; therefore, this protocol resembles the estimate-and-forward coding scheme[54].
9.4.1.2 DF-protocol

The DF-protocol is inspired by the decode-and-forward [54] coding scheme and has a similar structure as that of the simple protocol defined in Section 9.4.1.1. However, the encoding at the source and the decoding and re-encoding at the relay is different for the DF-protocol, as will be outlined next.

Again, the transmission of the information from the source occurs in B blocks $b = 1, 2, \ldots, B$ of equal length $N$. For each block $b$, the source send the superposition of a new codeword $w_b$ and a repetition of the previous codeword $w_{b-1}$ with an appropriate power ratio, except for the first and the last blocks. In the last block there is no transmission of a new codeword $w_B = \phi$ and in the first block there is no prior transmitted codeword $w_0 = \phi$. At the end of block $b$, the relay node decodes the new transmitted codeword $w_b$ from the source by using the relay’s received signal in this block. This codeword, $w_b$ is then transmitted in the next block without re-encoding.

9.4.2 Constituent codes

The optimal design of the constituent codes depends on both the chosen protocol and its given power allocation. However, the main idea of random coding remains a dominant factor in the choice of good constituent codes. Inspired by the very good performance of irregular low density parity-check codes (LDPC) [91, 92, 93, 94], we develop a coding structure based on irregular LDPC constituent codes. It should be pointed out that any set of block codes such as convolutional codes or turbo codes can
also be used as constituent codes. Although, it is yet not clear how to design good joint convolutional or turbo codes for the mentioned coding protocols, LDPC codes can be optimized jointly for the structure of the code depicted in Figure 9.5(b) using two possible approximations of the density evolution method, Gaussian approximation [94] and Erasure channel approximation [95].

9.4.3 Power allocation

To have a fair comparison with single-link communication, in this work we assume that the relay channel codes are subject to a sum power constraint which is equal to the average power of the single-link code. Therefore, if the source and the relay nodes share the available power, there would be an optimal ratio of the total power which achieves the best performance; we refer to this as the spatial power allocation ratio. The spatial power allocation is obtained from the capacity analysis of the finite alphabet Gaussian relay channel. Moreover, there could be a temporal power allocation where the average power across the $B$ blocks of transmission is allocated in order to minimize the possibility of error propagation. While the mentioned temporal power allocation across $B$ blocks of transmissions is important for the case of fading relay channels, it is not as important as spatial power allocation for the Gaussian relay channel.

The spatial power allocation is a function of the chosen protocol and the relay channel condition and is found through the capacity analysis of the underlying
Figure 9.5: (a) Successive decoding procedure at the destination node. The partial factor graphs are indicated by boxes and the direction of decoding is indicated by the arrows (forward decoding in this example), (b) Solution of a partial factor graph. In this figure, $r_2$ is a multiple access code based on two codewords, while $r_1$ is a received signal based solely on the code indicated by part 1.
protocol of the code. For the simple relay protocol of Section 9.4.1.1, the power allocation is the ratio of transmit power from the source and the relay. However, for the DF-protocol three values need to be determined: (i) the source power for the new information, $Q_1$; (ii) the source power for the repeated information from the previous block, $Q_2$; and (iii) the relay power, $Q_3$, such that $Q_1 + Q_2 + Q_3 = P$ where $P$ denotes the total average power.

Because the decoding is essential at the relay for the DF-protocol, the goal is to try to furnish equal effective power for the decoding at the relay and at the destination for the same block of transmission. It should be noted that the effective power for a block which is transmitted simultaneously by the source and by the relay is obtained through coherent combining. Therefore, the received power at the relay signal $\gamma_0 Q_1$ should be equal to the received power at the destination for the same block, i.e., $(\sqrt{\gamma_1 Q_2} + \sqrt{\gamma_2 Q_3})^2$. On the other hand, it can be easily shown that $\gamma_1 Q_3 = \gamma_2 Q_2$ in order to maximize the gain from the coherent combining for a given average sum power $Q_2 + Q_3$. Thus, we have

\[
Q_1 = \frac{\gamma_1 + \gamma_2}{\gamma_0 + \gamma_1 + \gamma_2} P, \quad (9.17)
\]
\[
Q_2 = \frac{\gamma_2}{\gamma_1 + \gamma_2} \frac{\gamma_0}{\gamma_0 + \gamma_1 + \gamma_2} P, \quad (9.18)
\]
\[
Q_3 = \frac{\gamma_1}{\gamma_1 + \gamma_2} \frac{\gamma_0}{\gamma_0 + \gamma_1 + \gamma_2} P. \quad (9.19)
\]

Therefore, the power allocation factor for the relay is given by $\kappa = \frac{Q_3}{P}$, which shows the fraction of total power which is used by the relay node to retransmit a block.
However, in the DF-protocol, the power splitting between the superimposed blocks that are transmitted from the source in any time slot is important and is given by $\frac{Q_1}{Q_2}$.

9.5 Relay Channel Code Design Based on LDPC Constituent Codes

In this section, we present a specific code design for the relay channel based on LDPC constituent codes. The choice of using LDPC codes is based on two reasons: (i) the good performance of irregular LDPC codes [93, 94] and (ii) the possibility of code optimization and more importantly joint code design based on density evolution analysis [94] that can be extended to multi-user channels. Because it is extremely hard to find the optimized LPDC code profile for the factor graph of the whole $B$ blocks of transmission even for the simple relaying protocol of Section 9.4.1, we alternatively consider joint code design for the partial factor graph depicted in Figure 9.5(b). The resulting two set of codes can then be alternately used over $B$ transmitted blocks. This latter strategy enables us to design good codes for the successive decoding scheme discussed in Section 9.5.3. Moreover, because the joint decoding of all $B$ transmitted codewords is tedious in general, successive decoding algorithms are used in practice. Therefore, the designed codes are actually optimally designed for successive decoding strategy.
The encoding procedure for a relay channel code which exploits LDPC constituent codes follows exactly the structure of Section 9.4. However, the operation of the relay node depends on the chosen protocol and involves partial decoding and retransmission. The details of relay operation are presented in Section 9.5.2, after a brief review of some definitions and notations concerning the factor graph representation of the block codes in Section 9.5.1. The factor graph representation of the relay channel code and the corresponding decoding algorithms at the destination node are then presented in Section 9.5.3.

9.5.1 Factor Graph Representation of a Block Code

Any block code can be fully described by its parity check matrix $H$. A factor graph representation of the code (e.g. Figure 9.6) consists of a vector of variable nodes, a vector of check nodes, and some connection between them which is defined by the parity check matrix $H$. Each variable node is denoted by a circle in the graph and corresponds to a column of the parity check matrix $H$. Each check node is denoted by a square and corresponds to a row of the parity check matrix $H$. Finally, there is a connection between a check node and variable node if and only if the parity check matrix $H$ has the element 1 in the corresponding row and column.

For example, a factor graph representation of a rate $\frac{1}{2}$ regular $(3,6)$ LDPC code
Figure 9.6: Factor graph representation of the code. (a) Representation of a (3, 6) regular LDPC code. (b) Shorthand (symbolic) representation of a factor graph with a vector variable node (the circle), a vector check node (the square), and the parity check matrix $H$ which represents the connection between these to vectors. (c) Shorthand representation of direct connections which corresponds to $H = I$. 
with the parity check matrix

$$H = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1
\end{bmatrix}$$ (9.20)

is depicted in Figure 9.6(a), where the shorthand notation is denoted in Figure 9.6(b).

If $H = I$, the factor graph is just a series of parallel connections between the check nodes and the variable nodes for which we use the shorthand notation of Figure 9.6(c).

### 9.5.2 Relay Node Operation

As we discussed earlier in Section 9.4, the choice of the protocol defines the operation of the relay node. Here, we provide more details regarding the decoding at the relay node and retransmission for both the simple protocol and the DF-protocol defined in Section 9.4.1 for the LDPC constituent codes.

#### 9.5.2.1 Relay operation in the simple protocol with LDPC constituent codes

For the simple protocol defined in Section 9.4.1, the factor graph of the code at the receiver of the relay node is represented in Figure 9.7, where $\vec{r}_1', \vec{r}_2', \ldots, \vec{r}_B'$ denote the received vectors in the $B$ consecutive transmission blocks at the relay node. At
each block $b$, the relay node wants to find the best (MAP) estimate of the transmitted codeword from the source based on its received vector in the same block. This process is indeed identical to solving the corresponding factor graph at the relay node, although the goal is necessarily not to decode the transmitted codeword $w_b$ from the source. In fact, for some relaying conditions the transmission rate might be higher than the capacity of the source-relay link, and hence, the resulting codeword would be erroneous. However, in such a situation, the factor graph solver acts as a quantizer which quantizes the relay received vector with a rate that can be reliably transmitted over the relay-destination link. Therefore, the factor graph solver finds the closest codeword to the received signal, and clearly the outcome of the factor graph solver is already a valid codeword. Therefore, the decoded codeword can be transmitted over the relay-destination link without any extra encoding.

As was discussed earlier, the simple relay protocol defined in Section 9.4.1 does not fully follow the decode-and-forward or the estimate-and-forward coding schemes. The above discussion reveals that in the case that we are interested in sending the best estimate of the relay received signal, the relay should find the optimal quantizer which considers the existing side information at the destination node for the current relay received signal. Therefore, the above choice of the quantizer in Figure 9.7 is not optimal. However, the robust nature of the relaying protocol of Figure 9.7 serves as a reasonably good approximation of both the decode-and-forward and estimate-and-forward protocols in different relaying regimes. From a practical point of view, a
Figure 9.7: Operation of the relay node in a simple relay protocol (Section 9.4.1).
low complexity operation in the relay node might be more desirable, especially if the relay node is not only dedicated to a relaying operation but also to its own separate role in the network.

9.5.2.2 Relay operation in the DF-protocol with LDPC constituent codes

The relay can also use the DF-protocol of Section 9.4.1.2. At the end transmission in block $b$, the relay has access to $b$ received signals $r'_1, r'_2, \ldots, r'_b$, and, because of superposition encoding at the source, the structure of the optimal decoder at the relay is similar to Figure 9.8 by setting $B$ to $b$ and $r_i$ to $r'_i$, for $i = 1, 2, \ldots, b$ in this figure. Clearly, solving $B$ complicated factor graphs at the relay nodes for the $B$ blocks of transmission is tedious and also not feasible. Moreover, the decoding at the relay and retransmission should be fast; therefore, we use two simple algorithms for decoding at the relay node. The first decoding algorithm uses the idea of successive cancellation. At the end of block $b$, the relay has already decoded $w_1, \ldots, w_{b-1}$; therefore, it cancels the effect of $w_{b-1}$ from the received signal at block $b$ and decodes $w_b$. The above simple algorithm in fact throws away the useful information about $w_{b-1}$ in the relay received signal $r'_b$ at block $b$. Thus, a better strategy is to jointly decode $w_{b-1}$ and $w_b$ by using the relay received signal and a priori knowledge about $w_{b-1}$ from the previous block. The decoding structure is then similar to that of Figure 9.5(b) (where indices $b-1, b$ correspond to indices 2, 1 in the figure). More discussion about the decoding structure of Figure 9.8 and Figure 9.5(b) can be found
in Section 9.5.3.

### 9.5.3 Decoding at the Destination Node

By using the shorthand notation of Figure 9.6, the factor graph of the designed relay channel code is represented in Figure 9.8, where $r_1, r_2, \ldots, r_B$ denote the received vectors in the consecutive $B$ transmission blocks at the destination node. The parity check matrix of the constituent codes for the consecutive codewords in the $B$ blocks are denoted by $H_1, H_2, \ldots, H_B$, respectively. For both protocols defined in Section 9.4.1 every new codeword which is transmitted from the source in block $b$, $b = 1, 2, \ldots, B$, is either decoded or estimated by the relay node and then retransmitted in the next block $b + 1$. The codeword $w_b$ which is encoded by the code with the parity check matrix $H_b$ affects both the received vectors $r_b$ and $r_{b+1}$ at the destination. Therefore, a priori information about the codeword $w_b$ is obtained through both $r_b$ and $r_{b+1}$ as is indicated in Figure 9.8. The optimal decoding algorithm in general is the joint decoding of all $B$ blocks which is the solution to the factor graph of Figure 9.8 using the MAP algorithm which is extremely difficult. However, if LDPC codes are chosen as the constituent codes, the practical idea of belief propagation can be used in the decoding process. Still, the application of belief propagation for the factor graph of Figure 9.8 is unfeasible, as this method is known to be very sensitive to short cycles in the graph. To overcome this problem, we consider two successive decoding algorithms, the forward and the backward decoding algorithms, both of which have very good
Figure 9.8: Decoding at the destination node. The factor graph representation of the proposed code for the relay channel coding, denoting both the constituent codes and the protocol that shows the connection of the constituent codes and received vectors. The optimal decoder at the destination node should solve this factor graph to decode all $B$ transmitted block jointly.
performance with orders of magnitude lower decoding complexity. It should be noted that if the constituent codes are some other form of block codes such as turbo codes or convolutional codes, the same forward or backward decoding ideas can still be successfully exploited. However, the joint design of such block codes for the coding structure denoted in Figure 9.8 or 9.5(b) remains a challenge. To avoid any confusion, it should be pointed out that the forward-backward or BCJR decoding algorithm for convolutional codes is a completely different decoding strategy than our successive decoding algorithm defined as the forward or the backward decoding algorithms.

### 9.5.3.1 Forward decoding algorithm

In the forward decoding algorithm, the original factor graph of the code denoted in Figure 9.8 is broken into a sequence of smaller factor graphs, called partial factor graphs that are represented in Figure 9.5(a). Decoding starts from the left to decode the first block and successively proceeds forward by removing the interference of the last decoded block from the current block.

The benefit of this algorithm is that the decoding delay is not more than two blocks, because the decoding of block $b$ can be done right after the reception of block $b+1$. However, the forward decoding algorithms is better than the backward decoding algorithm only if the position of the relay is far from the source. In fact, the forward decoding algorithm in conjunction with the presented coding strategy follows the idea of the information theoretical estimate-and-forward coding scheme [54] for the relay
channel. For relay positions far from the source, say \( d = 0.75 \) in Figure 9.2, a simple calculation of the a-priori bit probabilities of the codewords for the first and the last partial factor graph of Figure 9.5(a) also confirms that the a priori information is stronger if decoding starts from the first partial factor graph.

### 9.5.3.2 Backward decoding algorithm

In the backward decoding algorithm, the original factor graph of the code is broken into partial factor graphs. However, unlike the forward decoding algorithm, the decoding starts from the right to decode the last block and successively proceeds backward by removing the interference of the last decoded block from the current block.

Backward decoding is in fact more efficient for positions of the relay node closer to the source, say \( d < 0.5 \) in Figure 9.2. The reason is that the backward decoding algorithm along with the presented coding strategy follows the idea of the information theoretical decode-and-forward coding scheme [54] for the relay channel. Again, it is not hard to see that the a priori probabilities of the first and the last codewords are in favor of decoding the last block first. The backward decoding algorithm might be of more interest because the relay which is positioned close to the source is usually more desirable. However, the backward decoding algorithm suffers from the fact that decoding cannot be started before receiving the entire block of \( B \) transmission. Therefore, this algorithm has a decoding delay of at least \( B \) blocks as opposed to the
forward decoding, which has a decoding delay of only two blocks.

### 9.5.3.3 Solving the partial factor graphs

As we discussed, at each step, the decoding of the current block is performed by successive interference cancellation from the last decoded codeword and by solving the resulting partial factor graphs. It should be noted that the resulting partial factor graphs from the forward or backward decoding algorithms after successive interference cancellation have an identical structure. Therefore, we just consider the partial factor graph of the first block of transmission as depicted in Figure 9.5(b).

After reception of both the received vectors for $r_1$ and $r_2$ for the first and second block of transmission at the destination node, bit probabilities for both $w_1$ and $w_2$ codewords are calculated based on these received vectors. Then, the joint decoding of both transmitted codewords $w_1$ and $w_2$ at the first and second blocks will be performed iteratively by passing messages between the two parts of the factor graph that correspond to the first and second codeword. Part1 and Part2 denote these two parts in Figure 9.5(b), and arrows show the message passing between them through a vector of check nodes. This operation is similar to the decoding of the multiple access codes [96], except that there is another message coming from the received vector $r_1$ which can be considered as a side information about the codeword $w_1$. Therefore, after solving the factor graph of Figure 9.5(b), the estimate of $w_1$ is much better than the estimate of $w_2$ because of the stronger side information.
Figure 9.9: Performance of proposed relay channel code with simple protocol defined in Section 9.4.1. Performance of a single user code using the same constituent parity check matrix and Shannon limit for any single user codes are shown for comparison.

9.6 Performance

Our results show that even the simple protocol defined in Section 9.4.1 can easily outperform any single-link code. Figure 9.9 shows the performance of a relay channel code (code length \( N = 20,000 \), Number of blocks \( B = 40 \)) using the simple relaying protocol for relay positions of \( d = 0.5 \), \( d = 0.25 \), and \( d = 0.75 \). The performance of the single-link code for an optimized irregular LDPC code (the same constituent code is used for the relay channel) with the single-link Shannon limit is also shown for comparison.
It can be seen that the single-link code is approximately 1.3 dB away from the Shannon limit at an error rate of $10^{-5}$. However, for relays at a distance of $d = 0.5$, $d = 0.25$, and $d = 0.75$, the proposed relay channel code with the simple relaying protocol of Section 9.4.1 can achieve an error rate of $10^{-5}$ with the $E_b/N_0$ of approximately $-0.8$ dB, $-0.5$ dB, and $+0.2$ dB, respectively. It shows that not only can the performance of the single-link code be easily surpassed (for example about $2.1$ dB gain for $d = 0.5$) with the proposed simple relay channel code, but it also obtains a performance below the single-link Shannon limit ($0.187$ dB) for most of the relay channel conditions. It means that a simple relaying protocol with a constituent code of length 20,000 easily surpasses any code with any length for the single-link. More interestingly, this is true for most relay positions, say $d < 0.75$. This result highlights the fact that relay coding is beneficial even with the use of a simple relaying protocol.

Figure 9.9 also shows that for the case of the relay positioned close to the destination ($d = 0.75$), the forwarding decoding algorithm outperforms backward decoding, which is expected because it behaves similar to the idea of the estimate-and-forward coding scheme.

However, to come close to the actual capacity (or achievable rate) of the relay channel, we need to exploit more powerful protocols and jointly optimized constituent codes. For example, for $d = 0.25$, the performance of the proposed code design based on the DF-protocol and constituent codes of length $N = 20,000$ and $B = 40$ is shown in Figure 9.10. It can be seen that a BER of $10^{-5}$ can be achieved with an $E_b/N_0$
Figure 9.10 : Performance of the proposed relay channel code with decode-and-forward relay protocol and optimized LPDC codes. Constituent code length $N = 20,000$ with $B = 40$ Blocks. Depicted curve is for the position of relay at $d=0.25$ in figure 9.2. Required $(E_b/N_0)_{min}$ for this position of relay is shown for comparison.

of approximately $-2.6$dB, which is only about $0.9$dB away from the capacity of the relay channel at $d = 0.25$ and corresponds to about $3.9dB$ coding gain over the single-link transmission with the same constituent code. It should also be noted that the information theoretical decode-and-forward coding scheme is in fact optimal for the case of $d = 0.25$ as can be inferred from Figure 9.3. Therefore, it is not surprising that a relay channel code design with the decode-and-forward relaying protocol performs so well for this scenario.
Figure 9.10 also shows the performance of the proposed coding scheme with DF-protocol for the relay at distances of $d = 0.5$ and $d = 0.75$, which shows about a $3.7 dB$ and $2.2 dB$ coding gain over single-link transmission. Moreover, it should be noted that the required $E_b/N_0$ of the proposed codes are lower for the closer source-relay distance which is completely in line with the capacity analysis results for the decode-and-forward coding scheme. However, for the simple relay protocol the required $E_b/N_0$ becomes smaller as the relay moves midway between the source and destination because that this simple protocol exploits both the advantages of decode-and-forward and estimate-and-forward coding schemes.

9.7 Conclusion

We presented a general approach for the relay channel code design. Due to the distributed nature of coding, the proposed code structure consists of three equally important elements: protocol, constituent codes, and power allocation. Based on our results, a careful design of these three elements can be used to design a code that operates very close to the fundamental limits of the channel. Although good single-link codes are not necessarily good constituent codes for the relay channel code design, we have shown that the rich and advanced theory of single-link code design can effectively be used to find jointly optimized codes for the source and the relay channel.
Chapter 10
Cross-Layer Problems in Wireless Networks

In this section, packet scheduling with maximum delay constraints is considered with the objective to minimize average transmit power over Gaussian channels. The main emphasis is on deriving robust schedulers which do not rely on the knowledge of the source arrival process. Toward that end, we first show that all schedulers (robust or otherwise) which guarantee a maximum queuing delay for each packet are equivalent to a time-varying linear filter. Using the connection between filtering and scheduling, we study the design of optimal power minimizing robust schedulers. Two cases, motivated by filtering connection, are studied in detail. First, a time-invariant robust scheduler is presented and its performance is completely characterized. Second, we present the optimal time-varying robust scheduler and show that it has a very intuitive time water-filling structure. We also present upper and lower bounds on the performance of power-minimizing schedulers as a function of delay constraints. The new results form an important step toward understanding the packet time-scale interactions between a physical layer metric of power and a network layer metric of delay.
10.1 Introduction

Most multimedia sources are bursty in nature, a property which can be used to trade queuing delay with the resulting average transmission power [97, 98, 99, 100, 101]. In this section, we study the relation between average transmission power and strict delay constraints.

The work of this section is motivated by the fact that in most practical situations, source and channel probability distributions are seldom known completely. Thus, optimal schedulers built using an assumption on probability distributions can possibly incur significant loss in the presence of distribution mismatch. To avoid this loss we seek robust schedulers by investigating the performance achievable with no prior assumptions about the probability distribution of the source.

Our main contributions are three-fold. First, we establish necessary and sufficient conditions on the service rates of the wireless transmitter in order to meet the delay deadline of every packet in the queue. The conditions are used to show that any scheduler which meets a delay guarantee $D_{\text{max}}$ for every packet over Gaussian channels is a linear time-varying low-pass filter of the order no more than $D_{\text{max}}$. This confirms the intuitive explanation for power reduction due to an additional queuing delay provided in [98].

Second, using the relation between delay bounded scheduling and linear filtering, we construct robust schedulers that do not rely on the knowledge of source statistics. This marks a significant departure from most information-theoretic work on power
efficient scheduling [97, 98]. We construct both the optimal time-varying and time-invariant robust schedulers. The time-invariant robust scheduler turns out to be simply a moving average filter with filter length $D_{\text{max}}$. Similarly, the time-varying scheduler has an elegant structure and water-fills the newly arrived packets over the next $D_{\text{max}}$ time-slots.

Third and lastly, the performance of the optimal scheduler (robust or otherwise) is characterized via an upper and a lower bound. The upper bound is derived using Jensen’s inequality and the lower bound uses Hardy-Littlewood-Polya inequality [102]. The upper bound also characterizes the performance of the optimal time-invariant robust scheduler. Lastly, we also derive the exact performance of the optimal time-varying robust scheduler. To the best of our knowledge, these are the first bounds on average power consumption of delay-constrained scheduling, and they are most useful for studying small delays by providing trends of actual performance.

We note that the results in this section are limited to Gaussian channels, much like the works in [98, 99, 101]. We expect that the connections between filtering and scheduling will continue to hold for time-varying fading channels, albeit the nature of filters will change (possibly no longer low-pass).

The rest of this section is organized as follows. In Section 10.2, we formulate the problem and characterize the set of feasible schedulers. In Section 10.3, the relation between filtering and scheduling is clarified. Motivated by the filtering property of the schedulers, the design of the optimal time-invariant schedulers is presented in
Section 10.4. In Section 10.5, the upper and lower bounds on the performance of optimal scheduler are derived. The design of optimal robust schedulers, designed without the knowledge of source statistics, are presented in Section 10.6. Finally, we briefly survey the related works in Section 10.7 and we conclude in Section 10.8.

10.2 Guaranteed Maximum Delay Scheduler

10.2.1 Problem Formulation

Consider the system in Figure 10.1, where the number of input packets to the queue at time $t$ is given by a random process $X_t$. The arriving packets are queued in a buffer whose backlog before time $t$ is denoted by $B_t$. We assume that all packets have the same size. At any time $t$, a causal scheduler transmits $R_t$ packets out of the queue by looking at the queue size $B_t$ and by having the knowledge of all the previous input arrivals $\{X_i\}_{i=1}^{t}$. However, a non-causal scheduler uses all arrivals $\{X_i\}_{i=1}^{\infty}$.

In this section, we will consider the case when each packet is required to be delivered within the $D_{\text{max}}$ time-slots of its arrival. The scheduler will be designed
to adapt the transmission rate $R_t$ such that the average transmit power $\mathbb{E}[P_t]$ is minimized while meeting the maximum queuing delay constraint for each packet. We consider a Gaussian channel between the transmitter and receiver and assume that all transmissions occur in packets and that the length of the packet is long enough to allow reliable communication close to the mutual information of the channel. This is motivated by the outage versus capacity analysis in [103] and by other recent works in scheduling problem [97, 98]. Solving the mutual information formula for Gaussian channels shows that the average power is exponentially related to the rate, \textit{i.e.} $P_t \propto 2^{R_t}$. By looking at the average power required for \( N \) time slot the scheduler design based on two different optimization criteria is defined.

**Criterion 10.2.1.**

$$\min \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2^{R_t} \quad (10.1)$$

**Criterion 10.2.2.**

$$\min \mathbb{E} \left[ \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2^{R_t} \right] \quad (10.2)$$

The above optimization problems are not necessarily identical. It can be readily shown that a sufficient condition for the above two minimization problems to be identical is that any given time, \( t \), the scheduler output rate, \( R_t \), is just a function of the past \( M \) input arrivals, \( X_t, X_{t-1}, \ldots, X_{t-M+1} \), for some fixed value of \( M \) which does not depend on \( t \).
10.2.2 Set of Feasible Schedulers

We first characterize the set of all feasible schedulers by deriving the necessary and sufficient conditions for a scheduler to guarantee the maximum delay of $D_{\text{max}}$, in terms of $X_i$ and $R_i$ for $i = 1, 2, \ldots, t$. The results in this section form the basis for all the subsequent results.

**Lemma 10.2.3** (Feasible Schedulers). For any scheduler, the maximum delay of each packet is less than or equal to $D_{\text{max}}$ if and only if for all positive integer values of $t$ and $k$, the output service rates $R_i$ and input arrival rates $X_i$ for $i = t, t+1, t+2, \ldots, t+k+D_{\text{max}}$, and queue backlog $B_t$ satisfy the following inequalities:

$$X_t + X_{t+1} + \ldots + X_{t+k} \leq R_t + R_{t+1} + \ldots + R_{t+k+D_{\text{max}}-1} \quad (10.3)$$

$$R_t + R_{t+1} + \ldots + R_{t+k} \leq B_t + X_t + X_{t+1} + \ldots + X_{t+k}. \quad (10.4)$$

**Proof:** First we prove the necessity of the conditions. Condition (10.3) gives a lower bound for the output service rates out of the queue based on the input arrival rates to the queue. It shows that for any $t$ and $k$ the output rate out of the queue from time $t$ to time $t+k+D_{\text{max}}-1$ should be more than or equal to the input to the queue from the time $t$ to the time $t+k$. Suppose not, thus there exist some values $t$ and $k$ such that the output out of the queue from time $t$ to the time $t+k+D_{\text{max}}-1$ is less than the input to the queue from time $t$ to the time $t+k$. Therefore, even if the queue backlog at time $t$ is zero ($B_t = 0$), the last packet of which has arrived in the time interval of $t$ to $t+k$ would not be serviced until time $t+k+D_{\text{max}}$ which makes
the delay more than $D_{\text{max}}$ for this packet. Condition (10.4) gives an upper bound on the output of the queue based on the input arrivals and previous backlog in the queue. It shows that output in the time interval $t$ to $t+k$ cannot be greater than the sum of the input arrivals in the same time period plus the amount of queue backlog at the time $t$. This is equivalent to the condition that the queue backlog cannot be negative.

Next we prove that the mentioned conditions are sufficient as well. To do this, we need to show that if input arrival rates and output service rates satisfy both of the Conditions (10.3) and (10.4), then the maximum delay for any packet would not exceed $D_{\text{max}}$. Consider a packet which has arrived at arbitrary time $k$. Assume that $X_i$, $i = 1, 2, \ldots, k + D_{\text{max}} - 1$ are arbitrary input arrival rates and the output service rates $R_i$, $i = 1, 2, \ldots, k + D_{\text{max}} - 1$ have been chosen such that the conditions of the lemma are satisfied. For $t = 1$, from the Condition (10.3) we have

$$X_1 + X_2 + \ldots + X_{t+k} \leq R_1 + R_2 + \ldots + R_{t+k+D_{\text{max}}-1}. \quad (10.5)$$

Because there is no arrival to the queue before time $t = 1$ and because the queue initially is empty, the above condition guarantees that all packets which have arrived before time $k+1$ would be out of the queue before time $t + D_{\text{max}}$, which means that the delay for all the packets arrived at time $k$ is at most $D_{\text{max}}$. We still need to show that the above set of output service rates $R_i$, $i = 1, 2, \ldots, k + D_{\text{max}} - 1$ do not violate
the queue rule, i.e. \( B_i \geq 0 \), for all \( i = 1, 2, \ldots, k + D_{\text{max}} - 1 \). By definition

\[
B_i = X_1 + X_2 + \ldots + X_i - (R_1 + R_2 + \ldots + R_i) + B_0 \tag{10.6}
\]

From the condition 10.4, for \( t = 1 \) and \( k = i - 1 \) it is immediate that \( B_i \geq 0 \). ■

### 10.3 Feasible Schedulers are Low-Pass Filters

In this section, we first show that the structure of any scheduler which guarantees the maximum delay of \( D_{\text{max}} \) is equivalent to a linear time-varying filter of size no more than \( D_{\text{max}} \). The characterization of the finite impulse response filter immediately leads to the fact that it is always a low-pass filter.

**Theorem 10.3.1** (Filter Characterization). A scheduler which guarantees the maximum delay of \( D_{\text{max}} \) for any arrived packet is a linear time-variant filter of size \( D_{\text{max}} \), denoted by:

\[
R_t = \alpha_{t0}^i X_t + \alpha_{t1}^i X_{t-1} + \ldots + \alpha_{tD_{\text{max}}-1}^i X_{t-D_{\text{max}}+1},
\tag{10.7}
\]

where the filter coefficients satisfy the following

\[
\sum_{i=0}^{D_{\text{max}}-1} \alpha_{t+i}^i = 1, \forall \ t,
\tag{10.8}
\]

\[
0 \leq \alpha_i^t \leq 1, \forall \ t, i.
\tag{10.9}
\]

Furthermore, any time-variant filter of size \( D_{\text{max}} \) which satisfies the above constraints is a valid scheduler which guarantees the maximum delay of \( D_{\text{max}} \) for any packet.
Proof: First we prove the necessary condition by showing that every scheduler which guarantees the maximum delay of $D_{\text{max}}$ can be expressed as a linear time-variant filter. Consider the output of the queue at time $t$. Because the scheduler guarantees the maximum delay of $D_{\text{max}}$, all of the scheduled packets at time $t$ are packets which have arrived in the past $D_{\text{max}}$ time slots. Assume that the first output packet has the delay of exactly $i$ and that the last output packet has the delay of exactly $j$, where $D_{\text{max}} - 1 \geq i \geq j \geq 0$. Thus, the output rate $R_t$ can be written in the form of the Equation (10.7) where

(i) $\alpha_t^i$ and $\alpha_t^j$ are the portions of packets which have arrived at the time $t - i$ and $t - j$ and are transmitted at time $t$, and

(ii) $\alpha_t^k = 1$ for any $k$, such that $i > k > j$ if it exists.

Because packets which have arrived at time $t$ should be out before the time $t + D_{\text{max}}$ and because $\alpha_t^{i+i}$ is the portion of it which is out at time $t + i$, the summation of the Equation (10.8) should hold.

Now we relax the above conditions (i) and (ii) to that of (10.9) and prove the sufficiency condition which means that any scheduler that has output rate $R_t$ satisfying condition of the Theorem 10.3.1 will have a delay of no more than $D_{\text{max}}$ time unit for all of the packets. The proof follows immediately by verifying that the Inequalities (10.3) and (10.4) of Lemma 10.2.3 hold.

$\blacksquare$
Based on Theorem 10.3.1, whenever the scheduler is treated as a filter, we will use the term *scheduling filter* as an equivalent term for a scheduler. It is worth mentioning that a scheduler does not have a unique representation as a linear time-varying filter. In most cases, there is more than one possible way of representing a fixed scheduler in the form of a linear time-varying filter.

Theorem 10.3.1 turns the design of the guaranteed maximum delay scheduler into the problem of filter design with a ‘linear’ structure. Therefore, we foresee the vast literature on linear filtering theory as a fundamental tool in designing power-efficient schedulers [104, 105]. The following result follows directly from basic filtering theory.

**Corollary 10.3.2.** *Every feasible time-invariant scheduler is a low-pass filter.*

**Proof:** Because all of the coefficients of the filter are chosen in the interval [0, 1], all the zeros of the corresponding $z$-transform of the filter are on the left hand side of the origin which means that the filter is a low-pass filter [106].

**Remark 1:** For power-efficiency, additional delay helps by smoothening the input arrival process via queuing [98]. It is clear that by increasing the number of filter taps (equivalently increasing maximum possible scheduling delay), one can design better low-pass filters, leading to transmit power reduction. The intuition behind the fact that the optimal average power scheduler would have smoother output sequence than the input sequence comes from the convexity of the objective function of the optimization problem in either of Criteria (10.2.1) or (10.2.2).
Remark 2: It can be easily proved that the optimal scheduler with guaranteed maximum delay $D_{\text{max}}$ depends on the distribution of the input arrivals, on the past observation of the input arrival rates $X_{t-i}$, $i = 0, 1, \ldots, D_{\text{max}} - 2$, and on the queue backlog $B_t$ in order to determine the optimal value of output service rate, $R_t$, at time $t$. In other words, it does not depend on the individual values of $X_{t-i}$ for $i \geq D_{\text{max}} - 1$. We will use this property along with the linear filtering property of the optimal scheduler to find the optimal robust scheduler when the distribution of the input arrival rates is not known. The structure of the robust scheduler even extends to the case in which the packets have different QoS requirements (i.e. maximum delay requirements) while the supremum of all of the delays is still bounded by the fixed value $D = D_{\text{max}}$.

10.4 Optimal Time-invariant Scheduling

Motivated by the filtering property of the scheduler, we consider design of the optimal time-invariant scheduler for a given input arrival distribution. Theorem 10.4.1 characterizes the optimal time-invariant scheduler, and it turns out that the optimal solution is independent of the input arrival distribution. In other words, the optimal time-invariant scheduler is robust to the changes of the input arrival distribution as it will be further explored in Section 10.6. Also, for any input arrival distribution, the performance of optimal time invariant scheduler provides an upper bound on the performance of the optimal scheduler (which is time-varying in general) discussed in
Theorem 10.4.1 (Time-invariant Scheduler). Let input arrival rates \( X_t, t = 0, 1, \ldots \) be i.i.d. random variables with distribution \( f_X(x) \). For both cases of the optimization problems posed in Criteria 10.2.1 and 10.2.2, the optimal time invariant scheduling filter (scheduler) has the form

\[
R_t = \frac{X_t + X_{t-1} + \ldots + X_{t-D_{\text{max}}+1}}{D_{\text{max}}}. \tag{10.10}
\]

**Proof**: Assume that the optimal time invariant scheduler for a time index \( k \) is given by 

\[
R_k = \alpha_0 X_k + \alpha_1 X_{k-1} + \ldots + \alpha_{D_{\text{max}}-1} X_{k-D_{\text{max}}+1},
\]

where \( 0 \leq \alpha_i \leq 1 \) and \( \sum_{i=0}^{D_{\text{max}}-1} \alpha_i = 1 \).

First we prove that in the class of time invariant schedulers both of the objective functions are equivalent and equal to the objective function \( J = \mathbb{E}[2^{R_k}] \). The proof is based on the property that for all \( t \), \( R_t \) are identically distributed and also \( R_t \) and \( R_{t+M} \) are independent if \( M \geq D_{\text{max}} \). For the first objective function, we have

\[
J_1 \triangleq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2^{R_t} = \lim_{L \to \infty} \frac{1}{LD_{\text{max}}} \sum_{t=1}^{LD_{\text{max}}} 2^{R_t} \tag{10.11}
\]

\[
J_1 = \frac{1}{D_{\text{max}}} \sum_{i=1}^{D_{\text{max}}} \lim_{L \to \infty} \frac{1}{L} \sum_{t=1}^{L} 2^{R_{t+iD_{\text{max}}}} \tag{10.12}
\]

\[
J_1 = \frac{1}{D_{\text{max}}} \sum_{i=1}^{D_{\text{max}}} \mathbb{E}[2^{R_k}] = \mathbb{E}[2^{R_k}], \tag{10.13}
\]

where Equation (10.13) is derived from the law of large numbers in probability theory.

For the second objective function, we have
\[ J_2 \triangleq \mathbb{E} \left[ \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2^{R_t} \right] \]  
(10.14)

\[ J_2 = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \mathbb{E}[2^{R_k}] = \mathbb{E}[2^{R_k}] . \]  
(10.15)

Therefore, the optimization problem can be written as:

\[
\min J = \min \mathbb{E}[2^{R_k}]
\]
\[
= \min \mathbb{E}[2^{\alpha_0 X_k + \alpha_1 X_{k-1} + \ldots + \alpha_{\text{max}-1} X_{k-\text{max}+1}}]
\]
\[
= \min \mathbb{E}[2^{\alpha_0 X_k}] \mathbb{E}[2^{\alpha_1 X_{k-1}}] \ldots
\]
\[
\ldots \mathbb{E}[2^{\alpha_{\text{max}-1} X_{k-\text{max}+1}}]
\]
\[
= \min \mathbb{E}[2^{\alpha_0 X_k}] \mathbb{E}[2^{\alpha_1 X_k}] \ldots \mathbb{E}[2^{\alpha_{\text{max}-1} X_k}].
\]  
(10.16)

It can be easily shown that \( f(\alpha) = \ln \mathbb{E}[2^{\alpha X_k}] \) is a convex function of \( \alpha \) for any distribution of \( X_k \). The second derivative of \( f(\alpha) \) is given by:

\[
\frac{\partial^2 f(\alpha)}{\partial \alpha^2} = \frac{\mathbb{E}[(X_k \ln 2)^2 2^{\alpha X_k}] \mathbb{E}[2^{\alpha X_k}] - \left( \mathbb{E}[X_k \ln 2] 2^{\alpha X_k} \right)^2}{(\mathbb{E}[2^{\alpha X_k}])^2}.
\]  
(10.17)

The numerator is positive by using the Cauchy-Schwartz inequality in the form \( \mathbb{E}[^2\mathbb{E}[g^2] \geq \mathbb{E}[fg] \), and the denominator is obviously positive. Thus the second
derivative of \( f(\alpha) \) is always positive, and therefore it is a convex function. By applying Jensen’s inequality and considering that \( \sum_{i=0}^{D_{\text{max}}-1} \alpha_i = 1 \), we have

\[
\frac{1}{D_{\text{max}}} \sum_{i=0}^{D_{\text{max}}-1} f(\alpha_i) \geq f \left( \frac{1}{D_{\text{max}}} \sum_{i=0}^{D_{\text{max}}-1} \alpha_i \right)
\]

\[
\sum_{i=0}^{D_{\text{max}}-1} f(\alpha_i) \geq D_{\text{max}} f \left( \frac{1}{D_{\text{max}}} \right)
\]

\[
\sum_{i=0}^{D_{\text{max}}-1} \ln \mathbb{E}[2^{\alpha_i X_{k-1}}] \geq D_{\text{max}} \ln \mathbb{E}[2^{\frac{X_k}{D_{\text{max}}}}]
\]

\[
\prod_{i=0}^{D_{\text{max}}-1} \mathbb{E}[2^{\alpha_i X_{k-1}}] \geq \mathbb{E}^{D_{\text{max}}}[2^{\frac{X_k}{D_{\text{max}}}}].
\]

Thus, \( J \geq \mathbb{E}^{D_{\text{max}}}[2^{\frac{X_k}{D_{\text{max}}}}] \) with equality if and only if \( \alpha_i = 1/D_{\text{max}} \) for all \( i = 0, 1, \ldots, D_{\text{max}} - 1 \).

10.5 Bounds on Scheduler Performance

Upper and lower bounds on the optimal scheduler performance (robust or otherwise) are presented in this section. The upper bound is, in fact, the performance of the robust time-invariant scheduler presented in the previous section. Also, the bounds apply to both optimization Criteria 10.2.1 and 10.2.2.

Corollary 10.5.1 (Upper Bound). \textit{An upper bound on the optimal scheduler guaranteeing maximum delay of \( D_{\text{max}} \) is given by:}

\[
P_{\text{avg}} \leq \mathbb{E}^{D_{\text{max}}}[2^{\frac{X_k}{D_{\text{max}}}}],
\]

(10.19)
where input arrival rates are i.i.d. random variables with distribution \( f_X(x) \).

**Proof**: Follows directly from the existence of the scheduler in Theorem 10.4.1 which requires the power average of \( \mathbb{E}^{D_{\text{max}}}[2^{\frac{X}{D_{\text{max}}}}] \).

It is easy to prove that this upper bound is a decreasing function of \( D_{\text{max}} \) for any distribution of the input process \( X_t, f_X(x) \) (Jensen’s inequality). It again shows that as the maximum delay increases, the number of scheduling filter (scheduler) taps \( D_{\text{max}} \) increases, and the required average transmit power is lower.

By using Jensen’s inequality it is fairly easy to see that \( 2^\mathbb{E}[X] \) is a lower bound on the average power of any scheduler which satisfies conditions of the Lemma 10.2.3. In order to establish a stronger lower bound, we will use Hardy-Littlewood-Polya inequality [102] which captures the effect of permutation and ordering and which fundamentally differs from the class of inequalities (such as Jensen’s inequality) that are based on the property of convex functions.

**Theorem 10.5.2** (Lower Bound). Let input arrival rates \( X_t, t = 0, 1, \ldots \) be i.i.d. random variable with distribution \( f_X(x) \). For both cases of the optimization problems posed in Criteria 10.2.1 and 10.2.2, a lower bound on the required average transmission power for any scheduler which guarantees the maximum delay of \( D_{\text{max}} \) is given by:

\[
P_{\text{ave}} \geq \mathbb{E}^{l+1}\left[2^{\frac{X}{D_{\text{max}}}}\right] \tag{10.20}
\]

for any value of parameter \( l \).
**Proof:** Consider Criterion 10.2.1 of the objective function. We have

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2R_t \geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \left( 2 \frac{R_t}{l+D_{\text{max}}} \right)^{l+D_{\text{max}}}
\]

\[
\geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2 \frac{R_t + R_{t+1} + \ldots + R_{t+l} + D_{\text{max}}}{l+D_{\text{max}}}
\]

\[
\geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2 \frac{X_t + X_{t+1} + \ldots + X_{t+l}}{l+D_{\text{max}}}
\]

where \((10.21)\) is from the application of Hardy-Littlewood-Polya inequality and \((10.22)\) is derived from the law of large numbers in probability theory. Similarly, for the Criterion 10.2.2 of the objective function, we have

\[
\mathbb{E} \left[ \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2R_t \right] \geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \left( 2 \frac{R_t}{l+D_{\text{max}}} \right)^{l+D_{\text{max}}}
\]

\[
\geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2 \frac{R_t + R_{t+1} + \ldots + R_{t+l} + D_{\text{max}}}{l+D_{\text{max}}}
\]

\[
\geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \mathbb{E} \left[ 2 \frac{X_t + X_{t+1} + \ldots + X_{t+l}}{l+D_{\text{max}}} \right]
\]

\[
\geq \mathbb{E}^{l+1} \left[ 2 \frac{X}{l+D_{\text{max}}} \right]
\]

\[
(10.22)
\]
\[
\geq \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \mathbb{E} \left[ 2^{\frac{X_t + X_{t+1} + \ldots + X_{t+1}}{l + D_{\max}}} \right] 
\]  
(10.25)

\[
\geq \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}^{l+1} \left[ 2^{\frac{l+1}{D_{\max}}} \right] 
\]  
(10.26)

\[
\geq \mathbb{E}^{l+1} \left[ 2^{\frac{x}{D_{\max}}} \right] 
\]  
(10.27)

where (10.24) is from the application of Hardy-Littlewood-Polya inequality. ■

**Remark 1:** The bound of the Theorem 10.5.2 gives the trivial lower bound of \(2^{\mathbb{E}[X]}\) as \(l \to \infty\). A special case of this lower bound is given by letting \(l = D_{\max} - 1\) as \(P_{ave} \geq \mathbb{E}^{D_{\max}}[2^{\frac{x}{D_{\max}^{-1}}}]\) which is tight only in the case of \(D_{\max} = 1\). The derived upper and lower bounds are equal in the case of \(D_{\max} = 1\) which corresponds to the transmission of the entire packets in the queue at each time instant. However, it is possible to prove that the upper bound of the Corollary 10.5.1 converges to the lower bound \(2^{\mathbb{E}[X]}\) as \(D_{\max} \to \infty\); thus, the lower and upper bound asymptotically converge together.

**Theorem 10.5.3.** The upper bound \(\mathbb{E}^{D_{\max}}[2^{\frac{x}{D_{\max}}}]\) converges to the lower bound \(2^{\mathbb{E}[X]}\) as \(D_{\max} \to \infty\).

**Proof:** For simplicity we assume that the random variable X is discrete, and we expand the upper bound as:

\[
\mathbb{E}^{D_{\max}}[2^{\frac{x}{D_{\max}}}] = 
\]

\[
(\frac{w_1(2^{x_1})}{D_{\max}} \frac{1}{D_{\max}} + \frac{w_2(2^{x_2})}{D_{\max}} \frac{1}{D_{\max}} + \ldots + \frac{w_n(2^{x_n})}{D_{\max}} \frac{1}{D_{\max}}) D_{\max} 
\]

which is the weighted power mean of \(2^{x_i}\)’s. It can be easily shown that as \(D_{\max} \to \infty\)
this quantity converges to the weighted geometric mean as: [102]

\[
\lim_{D_{\text{max}} \to \infty} E^{D_{\text{max}}}[2^{\frac{X}{D_{\text{max}}}}] = (2^{x_1})^{w_1}(2^{x_2})^{w_2} \ldots (2^{x_n})^{w_n} = 2^{x_1w_1+x_2w_2+\ldots+x_nw_n} = 2^{E[X]} \quad (10.28)
\]

which completes the proof.

Remark 2 : The bounds on the performance of the scheduler is depicted in Figure 10.2 for the Bernoulli input arrivals distribution. Usually, we need to design a scheduler which performs well for low maximum packet delay, while design of scheduler which performs well for relatively high maximum delay constraint is not difficult. In fact, the performance of the derived robust time invariant scheduler asymptotically is optimal for large maximum delay constraint. Although the lower bound of $2^{E[X]}$ is asymptotically tight, it is a loose lower bound for small values of the maximum delay (Figure 10.2), which are of more interest. The previous works on the power optimal packet scheduling [98, 97] also fail to provide a good lower bounds for small values of the average delay. However, the family of the lower bounds given by Theorem 10.5.2 as \( \left\{ E^{l+1} \left[ 2^{\frac{X}{D_{\text{max}}}} \right] \right\}_{l=1}^{\infty} \) provide useful lower bounds for the required average power. For any value of $D_{\text{max}}$, the supremum of this family of bounds is the best lower bound. It is illustrated as the envelope of lower-bound curves.
Figure 10.2: Performance of the optimal time-varying robust scheduler and the bounds on performance. Note that performance of the optimal time-invariant robust scheduler matches the upper bound.
10.6 Optimal Robust Scheduling

In practice, the exact knowledge of the input arrival distribution may not be available or it may change from time to time. The scheduler might also perform poorly when there is a mismatch between the actual source distribution and the assumed distribution by the scheduler. We discuss the design of a Robust scheduler which performs well regardless of the choice of source distribution. In Section 10.4, we derived the optimal time-invariant scheduler and also showed that it is independent of the distribution of the input arrival rates to the queue. Thus, the optimal time-invariant scheduler is in fact the optimal time-invariant Robust scheduler when the distribution of the input arrival rates is not known.

In this section, we present the optimal robust scheduler (which is time-varying in general) for guaranteed maximum delay. The robustness of this scheduler comes from the fact that the scheduled output rate \( R_t \) at time \( t \) is chosen without the knowledge of the future input arrival statistics. The scheduler is based on the previously observed arrivals to the queue and the amount of queue backlog. The optimality of this scheduler guarantees that there is no scheduler which can outperform this scheduler for all possible source distributions.

We also find the exact distribution of the output service rates for the optimal time-varying robust scheduler, which is then used to find the average transmission power. The average power of optimal time-varying scheduler is generally lower than that of the optimal time-invariant robust scheduler which provides a better bound
than the one presented in Corollary 10.5.1.

10.6.1 Optimal Time-varying Robust Scheduler

First, we extend the problem of finding the optimal robust scheduler for guaranteed maximum delay to a more general case, where there are different maximum delays associated to different arrived packets. Then, we consider the case where the maximum delay constraint for all the packets is the same.

10.6.1.1 Different delay constraints for each packet

We find the optimal scheduler for the general delay-constrained scheduling problem when knowledge of input distribution is not available. Solving the general case of the problem would be of more interest in practice where different packets have different maximum delay constraints associated with different QoS requirements. We assume that the supremum of the delay constraints for all arrived packets is equal to $D_{\text{max}}$. In this case we need to make $D_{\text{max}}$ separate FIFO queues, one for each set of packets with a delay requirement of $1, 2, \ldots, D_{\text{max}}$. In the simpler problem where there is only one maximum delay associated with each packet, only one FIFO queue is needed.

Assume that the scheduler does not know the distribution of future input arrivals. In this case, the only information available to the scheduler at time $t$ would be the past observed values of input arrival rates and the delay associated with each packet. Furthermore, assume that the maximum delay bound of all the packets is no more
than $D_{\text{max}}$. We make a circle consisting of $D_{\text{max}}$ queues and label them from 1 to $D_{\text{max}}$ in the clockwise order. At each time $t$, queue $i$ consists of the packets that have delay constraint $i$. So we push all the arrived packets with the same delay constraint $i$ into this queue. After we have serviced all of the scheduled packets out of the queues in time $t$, we go to the time $t + 1$ and we rotate the queue labels clockwise by one position. In other words, the queue which has delay constraint $i$ at the time $t$ would have delay constraint $i - 1$ in the time $t + 1$. Also, the queue 1 would be empty after this queue is serviced at time $t$, and its label will change to $D_{\text{max}}$ to accommodate the newly arrived packets at time $t + 1$ which have an associated delay constraint of $D_{\text{max}}$. 

Figure 10.3: Representation of the successive water-filling at time $t$ for the $i^{\text{th}}$ queue.
Lemma 10.6.1. Consider the constrained optimization problem

\[ J = \min \sum_{i=0}^{D-1} \phi \left( \sum_{j=0}^{D-i} X_j^i \right) \]

subject to the constraints

\[ \forall 0 \leq i, j \leq D - 1 : X_j^i \geq 0 \] (10.29)

\[ \forall 0 \leq j \leq D - 1 : \sum_{i=0}^{D-1-j} X_j^i = A_j \] (10.30)

where \( \phi(.) \) is a convex function and \( A_j \)'s are some nonnegative real values. The successive water-filling defined in Equations (10.31)-(10.33) is the optimal solution.

\[ X_{D-1}^0 = A_{D-1} \] (10.31)

\[ X_{D-2}^0 = (\mu_{D-2} - X_{D-1}^0)^+ \]

\[ X_{D-2}^1 = \mu_{D-2} \]

\[ X_{D-2}^0 + X_{D-2}^1 = A_{D-2} \] (10.32)

and successively water-fill for all \( i = 3, 4, \ldots, D - 1 \) according to,

\[ X_{D-i}^j = (\mu_{D-i} - \sum_{k=j+1}^{i-1} X_{D-k}^j)^+ \forall 0 \leq j \leq (i - 1) \]

\[ X_{D-i}^{i-1} = \mu_{D-i} \]

\[ \sum_{j=1}^{i-1} X_{D-i}^j = A_{D-i} \] (10.33)

Sketch of the proof: Because it is a convex optimization problem, it is easy to find the necessary and sufficient condition for the optimal solution based on Lagrange
multiplier technique. After finding the gradient and applying KKT conditions the solution turns out to be,

\[ X_{D-i}^j = (\mu_{D-i} - \sum_{k=j+1}^{i-1} X_{D-k}^j - \sum_{k=0}^{i-1} X_{k}^j)^+ \]

\[ \forall \ 0 \leq j \leq (i - 1) \]

Because determining \( X_{k}^j \) depends on both prior \((i < k)\) and subsequent \((i > k)\) values of \( X_{i}^j \), the solution of the above simultaneous set of equations is not easy to see. The successive water-filling Equations (10.31)-(10.33) remove the dependency in determining the value of \( X_{k}^j \) to the prior values of \( X_{i}^j \) for the \((i < k)\). Therefore, first we use the successive water-filling Equations (10.31)-(10.33) to find the values of \( X_{i}^j \) for all \( (i, j) : 0 \leq j \leq D - 1, \) and \( 0 \leq i \leq D - 1 - j \), and \( \mu_i \) for all \( 0 \leq i \leq D - 1 \). Then, we prove that there exists some other value of \( \mu'_i \) such that \( \mu'_i \) together with the same values of \( X_{i}^j \) satisfy the KKT conditions of the Equation (10.34).

It should be pointed out that the optimal solution to the above optimization problem is not necessarily unique and that once we get one solution (for example successive water-filling) it is easy to generate other solutions. Also, it is easy to prove the following interesting property for the water-filling solution:
Proposition 10.6.2. The argument of the function $\phi$ in the objective function $J$, 
$$\sum_{j=0}^{D-1-i} X_j^i$$, is a non-increasing function of $i$ for $i = 0, 1, \ldots, D - 1$.

Using Lemma 10.6.1 we can now characterize the optimal time-varying robust scheduler as stated in the following theorem.

Theorem 10.6.3. At any given time $t$, let $A_i$ denote the backlog of the queue $i$ for all $i$, $i = 1, 2, \ldots, D_{\text{max}}$. The optimal output service rate out of the $i^{th}$ queue is given by $X_{D_{\text{max}}-i}^0$, where this value is obtained through successive water-filling of the Equations (10.31)-(10.33) in Lemma 10.6.1 by letting $D = D_{\text{max}}$.

Proof: It is clear that there is no difference between any two different packets in the same queue in terms of our scheduling problem. Thus, at any given time $t$, all the observed information about the input arrival rates would reduce to the size of the backlog in each of the queues, which is assumed to be $A_i^t$ for all $i = 1, 2, \ldots, D_{\text{max}}$.

Let $\phi(x) = 2^x$ which is a convex function of $x$, and $A_i = A_i^t$ for all $i = 1, 2, \ldots, D_{\text{max}}$.

Using Lemma 10.6.1 from the Appendix I, successive water-filling over the next $D_{\text{max}}$ time slot gives the optimal solution for the values of $X_i^j$ for all $\{(i, j) : 0 \leq j \leq D_{\text{max}} - 1, \text{ and } 0 \leq i \leq D_{\text{max}} - 1 - j\}$. Thus, we find the optimal service rates out of the present queue given by $X_{D_{\text{max}}-i}^0$ for the $i^{th}$ queue (Figure 10.3). Successive water-filling means that the water-filling for different queues is performed sequentially and that the water-filling process for the $i^{th}$ queue treats all the other queues yet to be water-filled as empty queues.
Going to the next time $t + 1$ we need to push new packets to the queues by considering their maximum tolerable packet delays. Thus, the new queue backlog would be $A_{i}^{t+1} = A_{i}^{t} - X_{D_{\text{max}}-i}^{0} + N_{i}^{t}$ for all $i = 1, 2, \ldots, D_{\text{max}}$, where $N_{i}$ is the number of arrived packets with maximum delay constraint $i$ at time $t$. Therefore, in the next time $t + 1$ we need to perform successive water-filling of Lemma 10.6.1 again to find the optimal values of service rates out of the queues. Clearly, this process is optimal at each time instant $t$ based on the known information at the time $t$.

Note that the successive water-filling gives a nice, practical way of finding the current transmission rates out of each queue. From Lemma 10.6.1, one would notice that at any time instant $t$, successive water-filling gives the optimal values of the output service rates not only for the present time $t$ but also for the $D_{\text{max}} - 1$ future time instants. But, we would like to emphasize that the entire process of successive water-filling should be redone at each time instant. The reason is that at the new time instant $t + 1$, new input packet arrivals change the backlog of the entire set of queues; thus, the known information has been changed. With this new information, it is not hard to show that the optimal values of the output service rates out of the queues at the time $t + 1$ would not necessarily match the anticipated values at the time $t$ for this time slot (Figure 10.4).
Figure 10.4: Water-filling solution for the time slot $t - 1$ and subsequent time slot $t$.

10.6.1.2 The same delay constraints for all packets

For the case that all packets have the same maximum delay constraint $D_{\text{max}}$, the solution is fairly easy. Basically there are two main differences which make it simpler. First is the fact that a single FIFO queue is sufficient to perform scheduling. It is interesting to note that because of the scheduling process, there are still different packets with different delay constraints remaining in the queue at any given time $t$ and that the water-filling solution involves finding the optimal scheduled service rate for each of them. Still, the sum of the scheduled output service rates is enough to be considered as an optimal service rate out of the single FIFO queue to guarantee the maximum delay constraint. Second, the water-filling solution is much simpler because
at each step we just need to do water-filling for the currently arrived packets on top of the old solution. In other words, instead of making $D_{\text{max}}$ queues, all we need is to keep track of $D_{\text{max}}$ numbers at each time slot that corresponds to the water-filling solution of the previous time slot. We specifically have the following corollary based on Theorem 10.6.3.

**Corollary 10.6.4.** At any given time slot $t$, let $(S^t_0, S^t_1, \ldots, S^t_{D_{\text{max}}-1})$ denote the vector of the scheduled output service rates out of the queue from the previous time slot $t-1$, and let $X_t$ be the new arrival which needs to be scheduled. The optimal output service rate for the current time $t$, $R_t$, and the new vector of the scheduled service rate for the time $t+1$, $(S^{t+1}_0, S^{t+1}_1, \ldots, S^{t+1}_{D_{\text{max}}-1})$, are determined by the water-filling solution as:

\[
R_t = (\mu - S^t_0)^+ \\
S^{t+1}_i = (\mu - S^{t+1}_{i+1})^+ \quad \forall \ 0 \leq i \leq D_{\text{max}} - 2 \\
S^{t+1}_{D_{\text{max}}-1} = 0 \\
R_t + \sum_{i=0}^{D_{\text{max}}-1} S^{t+1}_i = X_t. \quad (10.35)
\]

Therefore, the coefficients of the optimal robust time-variant scheduling filter (scheduler) are given by:

\[
\alpha^t_i = \frac{(\mu - S^{t+1}_{i+1})^+}{X_t} \quad \forall \ 0 \leq i \leq D_{\text{max}} - 1. \quad (10.36)
\]

The above discussion provides an extremely simple solution to the problem of
robust scheduling for delay constrained inputs to a queue. There are some important observations which come out of the above solution. First, although we did not use the filtering property of the scheduler to find the optimal robust scheduler in Theorem 10.6.3 and in Corollary 10.6.4, the optimal solution turns out to be exactly a linear time-variant filter of size $D_{\text{max}}$, which is not surprising given Theorem 10.3.1. Second, the optimal values of the filter coefficient are exactly a function of the past $D_{\text{max}} - 2$ values of the input arrivals and the queue backlog, which is expected as we discussed above. Third, as we mentioned earlier, the optimal scheduler intuitively should try to make the output service rate as smooth as possible, which corresponds to the low-pass property of the scheduler. Indeed, the water filling solution is the best verification of this intuition. Fourth, the water-filling solution and Proposition 10.6.2 reveal that the optimal solution always tries to push the scheduled time of the packets as far as possible to the maximum value of the tolerable delay.

In summary, Theorem 10.4.1 gives the best time-invariant robust scheduler, while Theorem 10.6.3 gives the best time-variant robust scheduler, or simply the optimal robust scheduler. The performance of the optimal time-varying robust scheduler is depicted in Figure 10.2. It can be observed that the performance of the optimal time-varying robust scheduler is much better than the performance of the time invariant scheduler which is depicted as an upper bound in Figure 10.2. In fact, the performance of the optimal robust scheduler follows the trend of the envelope of the lower bound curves.
10.6.2 Performance Analysis

In this section we first derive the distribution of the output service rates, \( f_R(r) \), for the time-varying robust scheduler when the input arrival rates are iid with the known distribution \( f_X(x) \). The derived distribution can be used to evaluate the performance of the robust time-varying scheduler when the input distribution is known. In fact, it is possible to find an analytic expression for the performance of the robust time-varying scheduler which provides a better upper-bound for the performance of the optimal scheduler (which is not necessarily robust). Although this upper bound is lower than that of Corollary 10.5.1, the upper bound of Corollary 10.5.1 is still useful because the simplicity of the expression and the purpose of the numerical calculation.

We assume that the input arrival rate is a stationary and ergodic random process. In this case, the output service rate is an ergodic and stationary random process that can be found by recursive solution. First, we consider the case of \( D_{\text{max}} = 2 \) to show how the recursive solution works to find the distribution of the output arrival rates. Then, we derive the distribution of the output arrival rates for the general case. Although, the solution in this case is more involved, the same approach like the simple case of \( D_{\text{max}} = 2 \) is used.

Suppose that \( f_1(.) \) and \( f_2(.) \) denote the output service rate distributions of the current and the next time intervals, respectively, after the water-filling process is performed. Let \( Y \) be the random variable showing the value of a queue backlog.
and \( y \) be the actual backlog of the queue at time \( i \) before performing water-filling. Because, \( Y \) is in fact the scheduled service rate for the next time slot at time \( i - 1 \), its distribution is given by \( f_2(y) \). Suppose that the input arrival at time \( i \) is given by the random variable \( X \). If \( X < Y \) then the scheduled rate for the next time slot would be the same as random variable \( X \), but if \( X \geq Y \), then the scheduled rate for the next time slot would be equal to \( \frac{X+Y}{2} \). Therefore, at any given time, for the stationary distribution of the next time slot conditioned on the queue backlog \( Y = y \) we have

\[
f_2(z | Y = y) = \begin{cases} 
  f_X(z) & \text{if } z < y \\
  2f_X(2z - y) & \text{if } z > y.
\end{cases}
\] (10.37)

Thus, considering the distribution of the random variable \( Y \), we have

\[
f_2(z) = \int_{0}^{z} 2f_X(2z - y)f_2(y)dy + \int_{z}^{\infty} f_X(z)f_2(y)dy,
\] (10.38)

which can be written in the form of the homogeneous Fredholm integral equation of the second kind

\[
f_2(z) = \int_{0}^{\infty} f_2(y)[f_X(z)U(y - z)
+ 2f_X(2z - y)(1 - U(y - z))]dy.
\] (10.39)

Therefore, for any input distribution of \( f_X(x) \), \( f_2(.) \) can be obtained by solving the above integral equation, where the kernel is a given function of the input distribution.
Also, if \( X < Y \) then the scheduled rate for the current time slot would be the same as random variable \( Y \), but if \( X \geq Y \), then the scheduled rate for the current time slot would be equal to \( \frac{X+Y}{2} \). Therefore, at any given time, for the stationary distribution of the current time slot conditioned on the queue backlog \( Y = y \) we have

\[
f_1(z|Y = y) = \begin{cases} 
0 & \text{if } z < y \\
\int_0^z f_X(x)dx & \text{if } z = y \\
2f_X(2z - y) & \text{if } z > y.
\end{cases}
\] (10.40)

Therefore, considering the distribution of the random variable \( Y \), we have

\[
f_1(z) = \int_0^z f_X(x)dx f_2(z) + \int_z^\infty 2f_X(2z - y)f_2(y)dy.
\] (10.41)

Thus, after some manipulation the distribution of the output service rate, \( f_R(r) \), which is equal to \( f_1(r) \) is given by:

\[
f_R(r) = \int_0^r f_X(\gamma)[f_2(r) + f_2(2r - \gamma)]d\gamma.
\] (10.42)

This result is easily extendable to the case of an arbitrary time delay constraint of \( D_{\text{max}} \) for \( D_{\text{max}} > 2 \). Although the solution is more involved for the case of \( D_{\text{max}} > 2 \), it has the same nature of first finding the solution of a homogenous Fredholm equation of the second kind and then plugging in the answer in a definite multiple integral. We use the same approach to find the distribution of the output service rate in general for an arbitrary \( D_{\text{max}} \). Let \( \bar{Y} \) denote the vector of the random variables
(Y_0, Y_1, \ldots, Y_{D_{\text{max}}-1}) which are the scheduled service rates for the current and next $D_{\text{max}}-1$ time slots. To find the stationary distribution, we use the recursive equation which relates the values of these random variable at time $i$ to the previous time slot $i-1$. It should be pointed out that in the case of $D_{\text{max}} = 2$ we do not need to consider the joint distribution of the random variables $Y_0, Y_1$, but in the case of $D_{\text{max}} \geq 3$ it is essential. Let $(y_0, y_1, \ldots, y_{D_{\text{max}}-1})$ be the actual values of the scheduled service rates for the current and next $D_{\text{max}}-1$ time slots at time $i-1$. If $X$ denotes the input arrival at time $i$, then the vector of the scheduled service rate $\vec{Z}$ would be equal to $(y_1, y_2, \ldots, y_{D_{\text{max}}-k}, z, \ldots, z)$ if and only if $k \in \{1, 2, \ldots, D_{\text{max}}\}$ is the minimum value for which $X + y_{D_{\text{max}}-1} + y_{D_{\text{max}}-2} + \ldots + y_{D_{\text{max}}-k+1} < ky_{D_{\text{max}}-k}$. We define this minimum value $k \triangleq \theta(\vec{y}, X)$ as a function of the value of $X$ and vector $\vec{y}$. Therefore, at any given time $i$, for the stationary distribution of the joint scheduled service times $\vec{Z}$ conditioned on the queue backlog $\vec{Y} = \vec{y}$, we have

$$f_{\vec{y}}(\vec{Z} = (y_1, y_2, \ldots, y_{D_{\text{max}}-k}, z, \ldots, z) | \vec{Y} = \vec{y}) =$$

\[
\begin{cases}
  f_X(z) & \text{if } z < y_{D_{\text{max}}-1} \\
2f_X(2z - y_{D_{\text{max}}-1}) & \text{if } y_{D_{\text{max}}-1} \leq z < y \\
\vdots & \\
D_{\text{max}} f_X(D_{\text{max}} z - \sum_{j=1}^{D_{\text{max}}-1} y_j) & \text{if } y_1 \leq z.
\end{cases}
\] (10.43)
Using the defined function \( k = \theta(\vec{y}, X) \), it can be simplified to

\[
f_{\vec{Y}}(\vec{z} = (y_1, y_2, \ldots, y_{D_{\text{max}}-k}, \ldots, z) | \vec{Y} = \vec{y})
= k f_X \left( kz - \sum_{j=1}^{D_{\text{max}}-1} y_i \right) \text{ if } k = \theta(\vec{y}, X).
\]

(10.44)

Therefore, the problem of finding the distribution of the joint output scheduled rates of the current and next \( D_{\text{max}} - 1 \) time slots reduces to solving the following integral equation.

\[
f_{\vec{Y}}(\vec{z}) = \int_{(\mathcal{R}^+)^{D_{\text{max}}+1}} f_{\vec{Y}}(\vec{z}|\vec{y}) f_{\vec{y}}(\vec{y}) d\vec{y} dx
\]

(10.45)

\[
= \int_{(\mathcal{R}^+)^{D_{\text{max}}+1}} \theta(\vec{y}, X)
\times f_X \left( \theta(\vec{y}, X) z - \sum_{j=1}^{D_{\text{max}}-1} y_i \right) f_{\vec{y}}(\vec{y}) d\vec{y} dx,
\]

where \( \mathcal{R}^+ \) is the set of nonnegative real numbers. Thus, the stationary distribution of the output service rate is given by:

\[
f_R(y_0) = \int_{(\mathcal{R}^+)^{D_{\text{max}}-1}} f_{\vec{y}}(\vec{y}) dy_1 dy_2 \ldots dy_{D_{\text{max}}-1}.
\]

(10.46)

Therefore, the average required power would be

\[
P_{\text{ave}} = \mathbb{E}_r[2^r] = \int_0^{\infty} 2^r f_R(r) dr,
\]

(10.47)

which gives an upper bound obtained with the assumption of a time-invariant scheduler that is lower than that of Corollary 10.5.1. As we mentioned earlier, although
the value of $P_{\text{ave}}$ in Equation (10.47) gives a better upper bound on the $P_{\text{ave}}$ for the optimal scheduler with the maximum delay constraint $D_{\text{max}}$, because of its ease of computation, the simple upper-bound of Corollary 10.5.1 is still more useful in most of the cases.

10.7 Related Work

It is well known that the required power for reliable communication between any two points is exponentially related to the rate of the information which is coded for transmission [98, 97]. Thus, by lowering the transmission power and transmitting over a longer period of time, the energy required to transmit a packet can be significantly reduced [100, 107, 108, 109]. However, the information is usually delay-sensitive; thus, the transmission time cannot be made arbitrarily long. There has been a lot of work to carry out schedulers which minimize the required average power (or transmission energy) with different variations on packet delay constraints. In [100], the off-line energy optimal solution has been found for the case that packets arrive in some time interval $[0, T)$, and they have to be delivered by time $T$. This is a deadline constraint for the arrived packets; thus, the delay corresponding to different packets varies with the arrival time of the packets in the queue. Therefore, the packets arriving at a later time have smaller delay constraints than those that arrived earlier. The off-line optimal scheduler [100] assumes noncausal knowledge of all arrival times of the packets in this interval. Although optimality of the causal on-line scheduler is not
proved, it has been shown through simulation that it is energy efficient much like its off-line counterpart. The extension of this work for the case of multiple users has been studied in [108] and solved by proposing two algorithms: the MoveRight and the MoveRightExpress algorithms corresponding to noncausal and causal schedulers, respectively. In [107], authors combined the energy efficiency by lowering the transmission rate and increasing the transmission time with the recovery of the batteries because of the slow Electro-chemical mechanism. Also, this work considers both the deadline and average delay constraint for the packets. In [98, 97], minimization of the average transmission power with the average delay constraint has been considered. This minimization problem can be turned into a convex optimization problem for which the mentioned works have proposed a dynamic programming formulation to find the optimal solution.

In this work, we consider a strict maximum delay constraint for each arrived packet instead of the average delay [97, 98] or deadline constraint [100, 108, 109] for a group of packets. We also consider a more general case where each packet has a different strict delay constraint. In [100], an off-line scheduler assumes prior knowledge of the packet arrival time (noncausal scheduler), while we consider the causal scheduler that looks at the previous time and number of packet arrivals without prior knowledge of the time or number of packets to arrive later. However, it is not always possible to assume perfect knowledge of the input arrival distribution, because there might be a mismatch between the assumed input distribution and the actual distribution. We
introduce a robust scheduler that performs well regardless of the choice of the input distribution. The off-line scheduler (which is noncausal) in [100] follows a distribution free approach, but the online version of the scheduler (which is causal) was not proved to be optimal. The robust scheduler presented in this section is proved to be optimal in the sense that if the knowledge of the input arrival distribution is not available, then it is the best possible scheduler to minimize the average power based on just the past observation of the input arrivals. The nice property of the derived analytic solution for the optimal robust scheduler is that the solution follows a simple water-filling of the arrived packets in time, which leads to the design of a scheduler with very low-complexity in comparison to the design of schedulers in [97, 98] that require dynamic programming solution.

In addition, we establish the intuitive known connection between the filtering and scheduling [98]. We prove that any scheduler which guarantees the maximum delay constraint $D_{\text{max}}$ over the Gaussian channel is always equivalent to a linear filter of the order no more than $D_{\text{max}}$. We use this filtering connection to establish new lower and upper bounds on the performance of the optimal scheduler and provide the scheduling schemes which achieve the upper bounds. The new lower bounds are especially useful for low values of the maximum delay constraint where no prior good bound is known. The proof technique in deriving the lower bounds in this section might be of independent interest. Finding a lower bound by using inequalities that rely on the property of convex functions usually reduces to Jensen’s lower bound
which is asymptotically optimal but does not provide a good bound for low values of
the delay constraint. We use Hardy-Littlewood-Polya inequality which is inherently
different from the class of inequalities which rely on convex function to capture the
effect of ordering.

10.8 Conclusion

In this section, we presented several fundamental results pertaining to scheduling
with maximum delay constraints for Gaussian channels. We established a connection
between low-pass filtering and scheduling, and we derived both time-invariant and
time-varying optimal schedulers which do not rely on the knowledge of source distri-
bution. New bounds on the average power consumption were presented, which are
particularly useful when we are interested in small delays.
Chapter 11
Conclusions

Communication in wireless networks faces many challenges including the time-varying nature of the wireless environment (that also exists in point-to-point wireless communication) and the effective use of network coding (that also exists in wired networks). However, the problem of communication in a wireless network introduces a unique set of challenges because interference naturally exists when two or more neighboring nodes are transmitting simultaneously. In this thesis we have analyzed the communication problems in wireless networks under some practical constraints (such as using half-duplex or cheap wireless radios, peak power constraint, and strict packet delay), and we evaluated the performance with or without these practical constraints.

In particular, two simple forms of the network, the relay channel and user cooperation channel (known also as multiple access with generalized feedback), have been considered and analyzed in greater detail. Several new coding schemes have been developed to obtain lower and upper bounds on the capacity of these channels under different assumptions. These bounds have also shown to be tight in some cases, and the capacity is derived for the underlying channel. A notable example is the capacity results of Chapter 5 in which two improvements are achieved to derive the capacity.
First, we have derived an upper bound based on the multi-state cut-set theorem of Chapter 2 which is better than the best known upper bound on the capacity of the relay channel since Cover and El Gamal [7]. Second, we have developed a new coding theorem which gives a higher rate than the best known achievable rates for the relay channels obtained by to Cover and El Gamal [7]. Capacity results of Chapter 5 are the only capacity results for the relay channel except that of Cover and El Gamal’s on the degraded relay channel in 1979 [7] and that of El Gamal and Aref’s in 1982 [49].

For the simplest case of the network, the relay channel, we have derived the optimal power allocation strategy for minimizing outage [110]. We also derived the ergodic capacity of the Rayleigh fading relay channel. The relay channel with half-duplex relaying (cheap relaying) has also been studied extensively in this thesis. We have presented a unified approach to both time division and frequency division relaying. Finally, we have developed a practical code design for the relay channel. This code design is of particular importance because the relay channel previously lacked a good code. Also, it verifies and strengthens the information theoretical results on the relay channel by showing how to exploit the anticipated gain in performance through the use of relay.

Despite the knowledge we acquired and developed about the relay channel in the past few years, many problems are still left open about this simplest form of the network, and they remain to be explored. Extending the ideas to more general
network topologies also remains an open and elusive question, and it seems hard to get a good grasp of the problem. Therefore, the problem of network coding for wireless communication networks lies within the scope of our future research efforts.
Bibliography


Notes in Computer Science, LNCS 2634, 2003.


[57] Y. Liang and V. V. Veeravalli, “The impact of relaying on the capacity of broadcast channels,” in *IEEE International Symposium on Information Theory*,
Chicago, IL, June 2004.


[120] ——, “Design of the power optimal robust scheduler with strict delay constraints,” in *Provisional Patent has been filed*, 2003.


[138] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, “Redundancy rates of


