Cooperative Multiplexing in Wireless Relay Networks

by

Vinayak Nagpal

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Electrical Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Borivoje Nikolić, Chair
Professor David Tse
Professor Geoffrey Bower

Spring 2012

Cooperative Multiplexing in Wireless Relay Networks

Abstract

Cooperative Multiplexing in Wireless Relay Networks

by

Vinayak Nagpal

Doctor of Philosophy in Electrical Engineering

University of California, Berkeley

Professor Borivoje Nikolić, Chair

Wireless networks are experiencing an explosive growth in the number of users and the demand for data capacity. One of the methods to improve capacity is to use tighter cooperation between terminals. In order to design a cooperative wireless link, several theoretical as well as practical challenges need to be addressed. In this dissertation we develop tools for the design of practical cooperative links that perform very close to fundamental limits.

Using the tools of information theory, we begin by showing that cooperative relaying provides additional degrees-of-freedom for communication. For a simple network with a single-antenna source, single-antenna half-duplex relay and a two antenna destination, we show that cooperation allows the link throughput to increase approximately by a factor of 2.

This gain is achievable using the recently introduced quantize-map-and-forward (QMF) cooperation scheme. However, QMF requires joint decoding of multiple information streams at the destination. The computational complexity of joint decoding is prohibitive for practical implementation. We address this problem by developing a low-complexity practical coding and system design framework for QMF relaying. The framework presents several pragmatic design choices to achieve cooperative degree-of-freedom gains in practice. The framework uses a combination of LDPC and LDGM codes decoded jointly over a low complexity factor graph. Signal processing requirements at all terminals are shown to have linear time complexity. Density evolution tools are developed for the design of specialized linear codes and mapping functions. Based on these tools, we demonstrate the design of cooperative links that perform within 0.5-1.0dB of information-theoretic limits.

To my family.

Contents

List of Figures

List of Tables

List of Abbreviations

AF   Amplify and Forward

APP   A Posteriori Probability

AWGN Additive White Gaussian Noise

BEC   Binary Erasure Channel

BER   Bit Error Rate

BIAWGN Binary Input Additive White Gaussian Noise

BICM Bit Interleaved Coded Modulation

BMS   Binary Memoryless Symmetric

BSC   Binary Symmetric Channel

CF   Compress and Forward

D-BLAST Diagonal Bell-Labs Space Time Architecture

DDF   Dynamic Decode and Forward

DF   Decode and Forward

DMT   Diversity Multiplexing Tradeoff

GABPT Gaussian Approximation Belief Propagation Threshold

i.i.d   Independent and identically distributed

LAN   Local Area Network

LDGM Low Density Generator Matrix

LDPC Low Density Parity Check

LLR   Log Likelihood Ratio

LTE   3G Partnership Project Long Term Evolution

MAP   Maximum A Posteriori

MIMO   Multiple Input Multiple Output

MISO   Multiple Input Single Output

ML   Maximum Likelihood

MMSE   Minimum Mean Square Error

NAF   Non-Orthogonal Amplify and Forward

PBICM   Parallel Bit Interleaved Coded Modulation

PDDF   Partial Dynamic Decode and Forward

PDF   Probability Density Function

QMF   Quantize-Map-and-Forward

SIC   Successive Interference Cancellation

SISO   Single Input Single Output

V-BLAST   Vertical Bell-Labs Space Time Architecture

Acknowledgments

First and foremost, I would like to express my deepest gratitude towards my research advisor Prof. Borivoje Nikolić for his guidance and support throughout my stay at Berkeley. I am indebted to him for helping me to learn the art of collaborative teamwork. His keen interest and ability to bring together expertise from diverse specializations has made it possible for me to complete this work. I want to thank him for always looking out for my best interests and offering timely career advice. I also want to thank Prof. David Tse for supporting this work with his deep expertise and valuable insights. Through him, I have learnt the importance of clear and effective communication in scientific research. His style of expressing complex ideas in a simple and intuitive form is a source of great inspiration. I would like to thank Prof. Jan Rabaey and late Prof. Donald Backer for reviewing my research proposal. I also would like to thank Prof. Geoffrey Bower for reviewing this dissertation. Special thanks to my colleagues, I-Hsiang Wang, Milos Jorgovanovic and Sameer Pawar for their support, hard work and friendship leading to a very successful collaboration. I am also grateful to Artemy Baxansky, Sameet Ramakrishnan, Kathy Sun and Akin Olugbade for their valuable contributions in this project.

My research was funded by the Center for Circuits and Systems (C2S2) focus center. Prototyping experiments were performed using equipment donations from National Instruments. Thanks to Hugo Andrade, Trung Tran and Sadia Malik from NI for always being prompt with support requests and providing long-term guidance.

I feel privileged to have enjoyed the excellent facilities, culture of open collaboration, dialogue with industry and spirit of camaraderie at BWRC. I am grateful to all BWRC faculty directors and staff for making BWRC such a pleasant place to work. Special thanks to Gary Kelson, Tom Boot, Brian Richards, Kevin Zimmerman, Leslie Nishiyama, Olivia Nolan and Bira Coelho.

My friends and colleagues have made this journey especially enjoyable. I would like to thank all members and alumni of the COMIC research group. Special thanks to Dusan Stepanovic, Ji-Hoon Park, Prof. Zhengya Zhang, Renaldi Winoto, Farhana Sheikh, Matthew Weiner, Katerina Papadoupoulou, Sharon Xiao, Jaehwa Kwak, Ruzica Jevtic, Olivier Thomas, Charles Wu, Brian Zimmer, Seng Oon Toh and Zheng Guo. I would also like to thank my colleagues outside of the COMIC group including Melvyn Wright, Dan Werthimer, Prof. Aaron Parsons, Mubaraq Mishra, Rahul Tandra, David Chen, Changho Suh, Debopriyo Chowdhury, Terry Filiba, Alex Krasnov, Mark Wagner, Peter McMahon and Jason Manley.

Finally I would like to thank my loving wife Bhavna for standing by me and patiently encouraging me through this long journey. I am also grateful to my parents Dr. M.K. Nagpal, Mrs. Uma Nagpal and brother Varun Nagpal for encouraging me towards higher studies and always being pillars of support.

Chapter 1

Introduction

1.1 Motivation

In the past two decades, wireless communications have had a profound impact on human productivity. The explosive growth in popularity of smartphones, cloud based services and mobile entertainment has led to the emergence of a wide array of innovative models for business and social interaction.

Wireless services have burgeoned, both in terms of global reach and capacity for data throughput. Subscriptions[1] have already surpassed 5 billion, which corresponds to 75% of the world population. Since their inception, wide area cellular networks have increased traffic capacity by several orders of magnitude. 3GPP Long Term Evolution (LTE) networks currently in deployment are expected to allow a peak download rate of approximately 300Mbps at end user terminals. Notwithstanding such accelerated network upgrades, capacity is lagging far behind traffic demand. The research community is faced with the challenge of meeting this growing demand with aggressive advances in communication technology.

Most challenges in traditional networks were attributed to two fundamental properties of the wireless channel namely, (a) multi-path propagation and (b) broadcast transmission over a superposition medium. As per the modern outlook, the same properties present the most opportunities for addressing the capacity challenge.

a) *Multipath Propagation:* Wireless transmissions reach their intended destination through several disparate paths. This phenomenon is a major contributor to the complexity of current network standards. However, recent advances have transformed this apparent impediment into an opportunity.

Antenna arrays are used at either end of a link to send multiple transmissions simultaneously in different directions. Such multiple input multiple output (MIMO) systems utilize the rich scattering environment for *spatial multiplexing*. Early proposals for

---

[1]Source: iSuppli 2010.

MIMO communication emerged in 1993, followed by several years of development. Today, MIMO techniques are widely deployed e.g. 802.11n local area networks (LAN), WiMAX and LTE.

Spatial multiplexing allows linear scaling of link capacity with the number of antennas in the array. While large arrays can be used at base station terminals, antenna arrays at mobiles are limited due to constraints on size, weight and power. As a result, spatial multiplexing is not fully realized in current networks.

b) *Broadcast Transmissions over a Superposition Medium:* Wireless transmissions from one user interfere with others in the vicinity. Interference management is one of the most important considerations in the design and deployment of practical networks. Nonetheless, this broadcast nature of the wireless medium also presents an opportunity to improve capacity.

Wireless terminals can hear transmissions from each other, providing a natural opportunity for *cooperation.* Terminals can potentially form cooperative antenna arrays, thereby allowing improved spatial multiplexing.

Over the last decade, cooperative communication has received tremendous interest from the research community. Several influential results in information theory have shown that cooperation can address the capacity challenges of future networks. The paradigm has also gained acceptance in studies for upcoming network standards like 3GPP LTE-Advanced.

Despite promising indications, several theoretical as well as practical challenges need to be addressed before cooperative spatial multiplexing techniques can be deployed in practical networks.

1.1.1 Related Work

The earliest information-theoretic study of cooperative networks is due to Cover and El Gamal [16](1963). They consider a network called the *relay channel* with one transmitter, one receiver and one cooperating relay. It represents the foundational building block for the study of cooperative communication. Over the past 40 years, several capacity bounds and cooperation strategies have been proposed for the relay channel. Notwithstanding these extended efforts, the fundamental capacity of a cooperative link remains an open problem even for this simple network. Several alternate and tractable metrics have been proposed to gain insights into performance limits of cooperation. One of the most popular metrics is the diversity multiplexing tradeoff (DMT). The fundamental DMT was characterized for specific configurations of the relay channel in 2009. However, it remains open for a general multiple-antenna relay channel.

In spite of the challenges, there is a widespread consensus that cooperative techniques will play a key role in addressing the capacity challenges of the future. It was shown in

2007 by Ozgur et.al.[51] that cooperative MIMO can allow the capacity of an adhoc network to scale linearly with the number of terminals. In 2009, a deterministic approach proposed by Avestimehr et.al.[6] led to results providing approximate capacity for several cooperative networks. While the general characterization of fundamental limits has remained elusive, results like these represent significant advances in understanding.

Another critical step in the evolution of cooperative communication is to bridge the divide between theory and practice. Coding schemes and decoding algorithms proposed by information theory are often too complex for the design of practical systems. Moreover, cooperation introduces unique performance-complexity tradeoffs in the system. Traditional system building blocks e.g. modulation, error correcting codes, synchronization and access control need to be reexamined for the cooperative communication paradigm. So far, these challenges have recieved relatively little attention.

This dissertation presents an attempt to contribute towards the vision outlined above. For a relatively simple class of networks, a framework of ideas is proposed to aid the design of low complexity cooperative wireless links that perform close to fundamental limits.

1.2 Scope and Organization

The focus of this dissertation is cooperative spatial multiplexing i.e. increasing the throughput of a wireless link using cooperation. In Chapter 2, a full characterization of the fundamental DMT for a representative relay channel is presented. If the relay is relatively close to the transmitter, DMT analysis shows that link throughput can be increased by almost a factor of 2 even with the help of a single half-duplex relay.

It is identified that the quantize-map-and-forward (QMF) cooperation scheme is approximately optimal at high signal-to-noise ratio for networks with an arbitrary number of cooperating relays. However, it presents a unique challenge in the design of computationally efficient error control coding schemes. In Chapter 3, this challenge is addressed with a low complexity coding and system design framework for QMF relaying. Central to the framework is a belief propagation algorithm over a sparse factor graph.

In Chapter 4, density evolution techniques are extended to enable systematic analysis and design of degree profiles for the proposed factor graph. Based on these tools, the design of code profiles for the relay channel that perform within $0.5-1.5$dB of information theoretic limits is demonstrated.

Finally, Chapter 5 presents a summary of results and directions for future work.

Chapter 2

Cooperative Multiplexing: Fundamental Limits

This chapter presents information theoretic tools and results to demonstrate the potential of cooperative techniques towards improving the performance of wireless networks. It primarily focuses on the *diversity multiplexing tradeoff* (DMT) framework for cooperative communication systems in the delay-constrained high SNR regime. It outlines previous work on DMT limits for *cooperative diversity* and presents new results that provide insights into opportunities for *cooperative spatial multiplexing*.

The chapter also introduces an enhancement of the DMT framework that allows capturing relative distances between terminals. It presents a characterization of the fundamental DMT for a network having a single antenna half-duplex relay between a single-antenna source and two-antenna destination. The results show that cooperation can achieve a multiplexing gain greater than that of the direct link and that it is a function of the relative distance between source and relay compared to the destination.

2.1 Outage Probability

Relay terminals create *virtual* antenna arrays for communication. Due to this, the characterization of MIMO channels is relevant for the study of cooperative communication. This section presents an overview of the outage probability formulation for MIMO channels.

Wireless environments are affected by multi-path propagation resulting in random channel fading. In most systems, the transmitter is not aware of the *instantaneous* realization of fading state. It encodes information based on knowledge of *average* channel statistics. This is called *non-coherent* transmission. Note that the receiver can measure instantaneous fading state by utilizing sounding symbols sent by the transmitter i.e. reception is usually *coherent*. Error correcting codes ensure that even if few symbols are corrupted, the receiver can still successfully decode the message. However, there are some instances when the receiver is

unable to successfully decode. These result in transmission errors. One of the major goals of wireless system design is to minimize transmission errors while retaining a high information throughput.

Fundamental limits for system performance are studied in the context of *ergodic* Shannon capacity i.e. the maximum information throughput for which channel coding can ensure arbitrarily small probability of error. Ergodic capacity represents the maximum achievable throughput averaged across all fading conditions. Multi-path fading properties of a channel change over relatively longer timescales (compared to receiver noise) and the fading state stays fairly constant over an interval called the *channel coherence interval*. In typical outdoor settings, the channel coherence interval is of the order of several milliseconds. In order to use an information rate that is close to the channel's ergodic capacity, every message block must be able to average across a sufficiently large number of coherence intervals. This requires long message blocks and increases the communication delay/latency. Most systems in practice have a constraint on the maximum tolerable latency which limits the length of message blocks and hence the number of fading realizations. In situations where the system delay constraint is comparable to channel coherence time, ergodic capacity does not provide a meaningful performance limit. These are known as *delay-constrained* or *slow fading* [67] systems.

Another method to study delay-constrained systems is to consider the probability of decoding error in relation with the information rate. If channel conditions are too severe to allow reliable communication at a given rate, the channel is said to be in *outage*. The probability of decoding error in terms of the outage event is given as follows:

$$P[\text{error}] = P[\text{error}|\text{outage}]P[\text{outage}] + P[\text{error}|\text{no outage}]P[\text{no outage}]$$

- If the channel is in outage, the probability of decoding error will be high irrespective of coding scheme i.e. $P[\text{error}|\text{outage}] \rightarrow 1$.

- If the channel is not in outage, the transmitter uses a coding scheme that allows decoding error to be made very small i.e. $P[\text{error}|\text{no outage}] \rightarrow 0$.

If the transmitter uses the best possible non-coherent coding scheme, the probability of decoding error is dominated by the probability of outage. This formulation gives a useful handle on the information carrying capacity of delay-constrained fading channels.

2.1.1 SISO Channel

As an example, consider a single input single output (SISO) wireless channel having complex Gaussian noise, unit input power constraint and average signal-to-noise ratio without fading, represented as SNR. For simplicity, assume quasi-static flat fading using a single coefficient
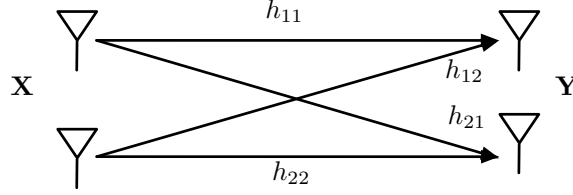
Figure 2.1: 2×2 MIMO Channel.

h. The maximum instantaneous mutual information I for this channel is given as follows:

$$I = \log_2 \left(1 + |h|^2 \mathsf{SNR}\right)$$

If the transmitter communicates using an information rate $\mathcal{R}$, the outage event is the set of channel realizations where $I < \mathcal{R}$. This condition represents the *outage region*:

$$|h|^2 < \frac{2^{\mathcal{R}} - 1}{\mathsf{SNR}}$$

For Rayleigh fading, $|h|^2$ is exponentially distributed with parameter λ, the outage probability $P_{\text{out}}^{\text{SISO}}$ is given as follows [37]:

$$P_{\text{out}}^{\text{SISO}}(\mathsf{SNR}, \mathcal{R}) := P\left[|h|^2 < \frac{2^{\mathcal{R}} - 1}{\mathsf{SNR}}\right]$$

$$\approx \lambda \frac{2^{\mathcal{R}} - 1}{\mathsf{SNR}}$$

This represents a fundamental limit for the SISO channel. It relates the maximum rate $\mathcal{R}$ for a given SNR with the minimum achievable P_{error}.

2.1.2 MIMO Channels

It is worth noting that for a SISO channel, the transmitter can only code across multiple symbols in time. For MIMO channels, coding can additionally be performed across multiple antennas. This offers several options for *space time* coding strategies. Outage probability provides a natural benchmark to evaluate performance of various space time coding schemes. For illustration, a few space-time coding schemes are described below with corresponding expressions for $P_{\text{out}}(\mathsf{SNR}, \mathcal{R})$. Derivations are excluded and can be found in [67].

Consider the 2×2 MIMO channel shown in Fig 2.1. For simplicity, assume quasi-static flat fading represented by matrix

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

Two antennas transmit a sequence of symbols $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots]$, $\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \end{bmatrix}$, $E[|x_{i,j}|^2] = 1$.

The received signal sequence is $\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ldots]$, $\mathbf{y} = \begin{bmatrix} y_{i,1} \\ y_{i,2} \end{bmatrix}$. Every receive symbol $\mathbf{y}$ experiences independent and identically distributed (i.i.d) Gaussian noise $\mathbf{Z} = [\mathbf{z}_1 \ \mathbf{z}_2 \ldots]$, $\mathbf{z}_i = \begin{bmatrix} z_{i,1} \\ z_{i,2} \end{bmatrix}$, $z_{i,j} \sim \mathcal{CN}(0, \sigma)$. The channel is parametrized by the average SNR, $\mathsf{SNR} = 1/\sigma^2$. The channel output is expressed as follows:

$$\mathbf{Y} = \mathbf{HX} + \mathbf{Z}$$

An *inner* space-time code is used for symbols across blocks of N symbol times i.e. $[\mathbf{x}_1 \ldots \mathbf{x}_N]$. Each symbol $x_{i,j}$ in the inner code is encoded using a capacity-achieving *outer* code. If a communication scheme with rate $\mathcal{R}$ is used, the fundamental outage probability is given as follows:

$$P_{\text{out}}(\mathsf{SNR}, \mathcal{R}) = P\left[\log \det\left(\mathbf{I} + \mathsf{SNR} \ \mathbf{HH}^\dagger\right) < \mathcal{R}\right] \tag{2.1}$$

Here the $\dagger$ superscript represents conjugate transpose. Outage probability expressions for representative space-time coding schemes are shown below.

Repetition Scheme

In this scheme, the same symbol is repeated across both antennas one-by-one while the other antenna remains quiet. $\mathbf{X}$ for repetition coding is given as follows:

$$\begin{bmatrix} 0 & x[1] & 0 & \ldots \\ x[1] & 0 & x[2] & \ldots \end{bmatrix} \tag{2.2}$$

Here $\{x[i]\}_{i=1}^{N/2}$ belong to a capacity-achieving outer code having rate $\mathcal{R}$. The repetition scheme encodes messages across time as well as antennas (space). However, the scheme sends only one symbol over two symbol times and is inefficient in terms of throughput. Assuming the channel remains constant for N symbol times (slow fading), the outage probability is given as:

$$P_{\text{out}}^{rep}(\mathsf{SNR}, \mathcal{R}) = P\left[\frac{1}{2}\log\left(1 + \mathsf{SNR}(|h_{11}|^2 + |h_{22}|^2 + |h_{21}|^2 + |h_{12}|^2)\right) < \mathcal{R}\right] \tag{2.3}$$

Alamouti Scheme

The Alamouti scheme is more efficient in use of space-time channel slots. For this scheme, $\mathbf{X}$ is given as follows:

$$\begin{bmatrix} x[1] & -x^*[2] & x[3] & \ldots \\ x[2] & x^*[1] & x[4] & \ldots \end{bmatrix} \tag{2.4}$$

Symbols $\{x[i]\}_{i=1}^{N}$ belong to a capacity-achieving outer code having rate $\mathcal{R}$. The structure of $\mathbf{X}$ ensures that every transmit symbol can be averaged across all 4 fading coefficients and also appears orthogonal to other symbols at the receiver. For slow fading, the outage probability is given as:

$$P_{\text{out}}^{Ala}(\mathsf{SNR}, \mathcal{R}) = P\left[\log\left(1 + \mathsf{SNR}(|h_{11}|^2 + |h_{22}|^2 + |h_{21}|^2 + |h_{12}|^2)\right) < \mathcal{R}\right] \tag{2.5}$$

V-BLAST

The above two schemes ensure that coding is performed across time as well as antennas. The vertical Bell Labs space time architecture [22] (V-BLAST) attempts to maximize system throughput at the cost of averaging across fewer fading coefficients. Under V-BLAST, the message is split across two coded streams. $\mathbf{X}$ is represented as follows:

$$\begin{bmatrix} x_1[1] & x_1[2] & x_1[3] & \dots \\ x_2[1] & x_2[2] & x_2[3] & \dots \end{bmatrix} \tag{2.6}$$

Here $\{x_1[i]\}_{i=1}^{N}$ and $\{x_2[i]\}_{i=1}^{N}$ belong to two independent capacity achieving codes having rates $\mathcal{R}_1$ and $\mathcal{R}_2$ respectively. For each of the two streams, the transmit antenna remains fixed and coding is performed only across time.

In the Alamouti scheme, two information symbols appear orthogonal at the receiver due to the special construction of $\mathbf{X}$. In V-BLAST, such a construction is not used and the receiver estimates $x_1[i]$ and $x_2[i]$ from its received signal using signal processing techniques. The decorrelator, MMSE (minimum mean squared error) and ML (maximum likelihood) are among various possible decoding techniques listed in order of increasing implementation complexity and performance. The 2×2 MIMO channel with V-BLAST decomposes into two SISO parallel sub-channels having complex gains $p_1(\mathbf{H})$ and $p_2(\mathbf{H})$ respectively.

The joint probability distribution of gains $p_1(\mathbf{H}), p_2(\mathbf{H})$ depends on the decoding architecture used. The outage probability for the scheme is as follows:

$$P_{\text{out}}^{VBLAST}(\mathsf{SNR}, \mathcal{R}_1, \mathcal{R}_2) = P\left[\bigcup_{i=1}^{2}\left(\log\left(1 + \mathsf{SNR}|p_i(\mathbf{H})|^2\right) < \mathcal{R}_i\right)\right] \tag{2.7}$$

D-BLAST

The diagonal Bell Labs space time scheme (D-BLAST)[22] uses a diagonal structure and performs coding across antennas as well as space. The sequence $\mathbf{x}' = \{x[i]\}_{i=1}^{N}$ forms one codeword belonging to a capacity achieving outer code having rate $\mathcal{R}$. Each codeword is split into two blocks of symbols $\mathbf{x}'_A$ and $\mathbf{x}'_B$, each having $N/2$ symbols. The transmitter

sends several codewords indexed by j and $\mathbf{X}$ is given as follows:

$$\begin{bmatrix} \cdots & \mathbf{x}'_B[j] & \mathbf{x}'_B[j+1] & \cdots \\ \cdots & \mathbf{x}'_A[j+1] & \mathbf{x}'_A[j+2] & \cdots \end{bmatrix}$$

In order to initialize the diagonal block structure shown above, one initial block of $N/2$ symbols i.e. $\mathbf{x}'_B[0]$ is set to $\mathbf{0}$. If one considers the information rate for D-BLAST over a sufficiently large number of codewords, this one-time overhead can be ignored. As discussed previously, the decorrelator, MMSE and ML receiver structures can be used for D-BLAST. The diagonal structure also allows successive interference cancellation (SIC) [67]. A detailed discussion of D-BLAST with MMSE-SIC is presented in Sec 2.5.3. Using D-BLAST, the MIMO channel decomposes into two parallel SISO sub-channels having gains $g_1(\mathbf{H})$ and $g_2(\mathbf{H})$. The outage probability for D-BLAST is given as follows:

$$P_{\text{out}}^{DBLAST}(\mathsf{SNR}, \mathcal{R}) = P\left[\sum_{k=1}^{2} \log\left(1 + \mathsf{SNR}|g_k(\mathbf{H})|^2\right) < \mathcal{R}\right] \tag{2.8}$$

The joint distribution of $g_1(\mathbf{H})$ and $g_2(\mathbf{H})$ is determined by the receiver architecture.

2.2 Diversity Multiplexing Tradeoff

It is interesting to compare the outage conditions for various space-time coding schemes shown in equations (2.3)(2.5)(2.7)(2.8) with the fundamental channel outage condition given in Eq. (2.1). A scheme is said to be *outage optimal* for a given region in $\{\mathsf{SNR}, \mathcal{R}\}$, when the scheme is in outage if and only if the channel is also in outage for all fading realizations $\mathbf{H}$. An important goal in studying MIMO channels is to understand these outage optimality characteristics.

This goal translates into understanding how various coding schemes can tradeoff information throughput ($\mathcal{R}$) for reliability P_{out}. There are many ways in which the MIMO channel can be poorly conditioned to cause outage, making the analysis of outage probability expressions like (2.1) etc. intractable. In order to overcome this challenge, the diversity multiplexing tradeoff (DMT) framework was introduced by Zheng and Tse in 2003 [80]. DMT is widely recognized for its ability to distill outage performance for space time coded systems in a compact form. Most systems in practice today operate in the moderate to high SNR regime i.e. transmit power is not the major limiting factor for system capacity. At high SNR, it becomes necessary to utilize advanced techniques like MIMO and cooperative relaying to get significant improvement in spectral efficiency. The DMT formulation takes advantage of this fact and simplifies outage analysis by focusing on the high SNR regime.

The ergodic capacity $\mathfrak{C}(\mathsf{SNR})$ for a MIMO channel scales logarithmically with SNR. For

M transmit and N receive antennas it is given as follows[66]:

$$\mathfrak{C}(\mathsf{SNR}) = E\left[\log\det\left(\mathbf{I} + \mathsf{SNR}\,\mathbf{H}\mathbf{H}^{\dagger}\right)\right]$$
$$\to \min(M, N)\log(\mathsf{SNR}) \text{ as } \mathsf{SNR} \to \infty$$

The outage probability $P_{\text{out}}(\mathsf{SNR}, \mathcal{R})$ approaches 0 as $\mathsf{SNR} \to \infty$ for any fixed rate $\mathcal{R}$. However, this fixed $\mathcal{R}$ is only a fraction of the high SNR ergodic capacity. If $\mathcal{R}$ is increased in a logarithmic relation to SNR as $\mathsf{SNR} \to \infty$, then the outage probability can be quantified.

As per the DMT formulation [80], for a coding scheme of interest a family of codes indexed by SNR is considered. If the rate $\mathcal{R}$ for every code in this family is scaled as $r\log(\mathsf{SNR})$, then r represents the normalized data rate and is called the *multiplexing gain*. A high multiplexing gain r represents scenarios where $\mathcal{R}$ is aggressively increased as SNR increases.

$$r = \lim_{\mathsf{SNR}\to\infty} \frac{\mathcal{R}}{\log(\mathsf{SNR})} \tag{2.9}$$

For this family of codes with multiplexing gain r, the outage probability is calculated as a function of SNR. The slope of P_{out} versus SNR on a log-log plot is defined as the *diversity gain $d(r)$* of the scheme.

$$d(r) = \lim_{\mathsf{SNR}\to\infty} \frac{\log(P_{\text{out}})}{\log(\mathsf{SNR})} \tag{2.10}$$

This can be expressed in terms of exponential equality as $P_{\text{out}} \doteq \mathsf{SNR}^{-d(r)}$. The diversity gain provides a measure of link reliability. The DMT is usually represented as a plot of maximum diversity $d(r)$ gain versus multiplexing gain r. The fundamental tradeoff for a given channel is given by the maximum achievable DMT over all possible coding schemes and is represented by $d^*(r)$.

2.2.1 2×2 MIMO DMT

To illustrate the general DMT analysis procedure, the fundamental DMT for the 2×2 MIMO channel is derived below using the system model in Sec 2.1.2 [67]. The outage probability for a given multiplexing gain r is given as follows.

$$P_{\text{out}}^{2\times2}(\mathsf{SNR}, \mathcal{R}) = P\left[\log\det\left(\mathbf{I}_2 + \mathsf{SNR}\,\mathbf{H}\mathbf{H}^{\dagger}\right) < r\log(\mathsf{SNR})\right]$$
$$= P\left[\prod_{i=1}^{2}(1 + \mathsf{SNR}\,\lambda_i) < \mathsf{SNR}^r\right]$$

Here $\lambda_1 > \lambda_2$ are the ordered eigenvalues of $\mathbf{HH}^\dagger$ which are expressed in terms of SNR as $\lambda_i = \mathsf{SNR}^{-\beta_i}$

$$P_{\text{out}}^{2\times2}(\mathsf{SNR}, \mathcal{R}) = P\left[\prod_{i=1}^{2}(1 + \mathsf{SNR}^{1-\beta_i}) < \mathsf{SNR}^r\right]$$

At high SNR, only terms with positive powers of SNR remain in the above expression i.e. for values $0 \leq \beta_1, \beta_2 \leq 1$.

$$P_{\text{out}}^{2\times2}(\mathsf{SNR}, \mathcal{R}) \doteq P\left[2 - \beta_1 - \beta_2 < r\right]$$

The set of β_1, β_2 values for which the outage condition holds is defined as the outage region i.e.

$$\mathcal{O}(r) = \{(\beta_1, \beta_2) : 2 - \beta_1 - \beta_2 < r; \ 0 \leq \beta_1, \beta_2 \leq 1\} \quad (2.11)$$

P_{out} is written as follows:

$$P_{\text{out}}^{2\times2}(\mathsf{SNR}, \mathcal{R}) = \int_{(\beta_1,\beta_2)\in\mathcal{O}(r)} f_{\beta_1,\beta_2}(\beta_1, \beta_2) \ d\beta_1 \ d\beta_2$$

Here $f_{\beta_1,\beta_2}(\beta_1, \beta_2)$ is the joint distribution of β_1, β_2. For Rayleigh fading i.e. entries of $\mathbf{H}$ are i.i.d. $\mathcal{CN}(0, 1)$ it is shown in [80] that $f_{\beta_1,\beta_2}(\beta_1, \beta_2) \doteq \mathsf{SNR}^{-(\beta_1+3\beta_2)}$

$$P_{\text{out}}^{2\times2}(\mathsf{SNR}, \mathcal{R}) \doteq \int_{(\beta_1,\beta_2)\in\mathcal{O}(r)} \mathsf{SNR}^{-(\beta_1+3\beta_2)} \ d\beta_1 \ d\beta_2$$

In the limit $\mathsf{SNR} \to \infty$, this integral can be approximated by using Laplace's method giving the diversity as:

$$d_{2\times2}^*(r) = \inf_{(\beta_1,\beta_2)\in\mathcal{O}(r)} \beta_1 + 3\beta_2$$

For outage region given in (2.11) and $r > 1$, the infimum is achieved by $\beta_2 = 0$; $\beta_1 = 2 - r$. Similarly for for $r < 1$, $\beta_1 = 1$; $\beta_2 = 1 - r$. This gives the fundamental DMT for the 2×2 MIMO channel as follows:

$$d_{2\times2}^*(r) = \begin{cases} 4 - 3r & 0 \leq r \leq 1 \\ 2 - r & 1 \leq r \leq 2 \end{cases} \quad (2.12)$$

Fig 2.2 shows that the 2×2 MIMO channel provides a maximum diversity of 4 at $r = 0$. This means that a coding scheme can prioritize link reliability by averaging every transmit symbol over all 4 fading coefficients (Fig. 2.1) $h_{11}, h_{12}, h_{21}, h_{22}$ to get maximum diversity gain of 4. Space time coding schemes can achieve this diversity gain by keeping $\mathcal{R}$ constant as $\mathsf{SNR} \to \infty$. The channel offers a maximum multiplexing gain of 2 $(d_{2\times2}^*(2) = 0)$. This implies
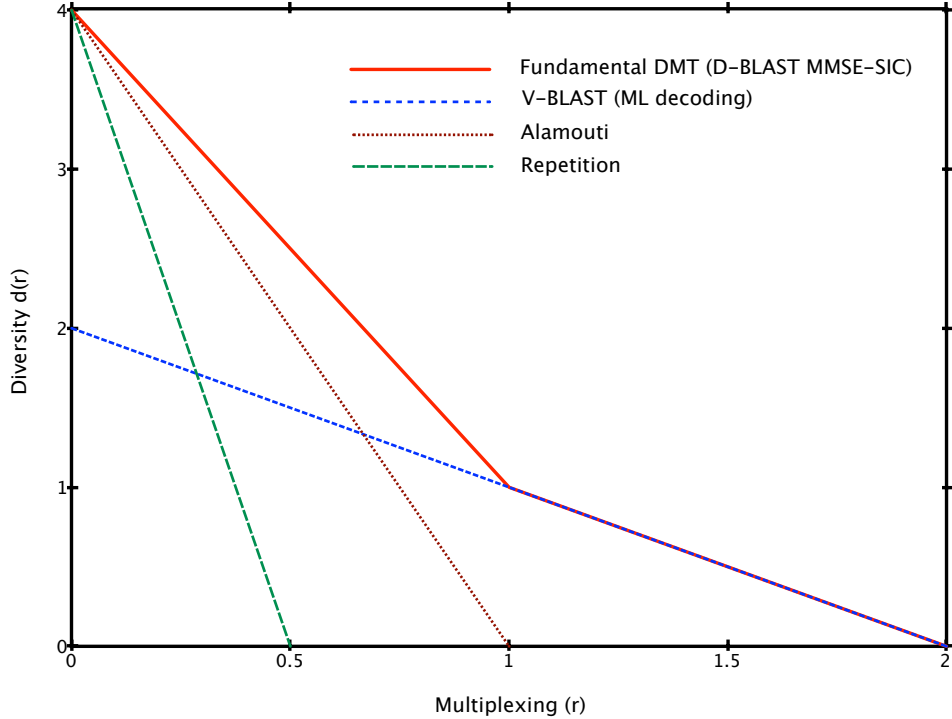
Figure 2.2: Diversity Multiplexing Tradeoff for the 2×2 MIMO Channel.

that at high SNR, information throughput for the 2×2 MIMO channel can be $2\log(\mathsf{SNR})$ as opposed to $\log(\mathsf{SNR})$ for the SISO channel. The maximum multiplexing gain is achieved by a scheme that maintains a constant P_{error} as $\mathsf{SNR} \to \infty$.

Similarly, maximum tradeoffs for various schemes are calculated using respective outage probability expressions and shown in Fig 2.2. The Alamouti scheme performs better than repetition for all $r > 0$. Both schemes achieve the maximum diversity offered by the channel and neither achieve the maximum multiplexing gain. V-BLAST outperforms the Alamouti scheme and achieves the fundamental tradeoff only at high multiplexing gains ($r > 1$). However, it can only achieve maximum diversity of 2. It is shown in [71] that the D-BLAST scheme [22] with MMSE-SIC receiver achieves the fundamental tradeoff $d^*(r)$ for all r.

The DMT framework discussed above has been widely used to study the performance limits of cooperative channels.

2.3 Relay Channel

A wireless transmission can be received by various terminals in the vicinity due to the broadcast nature of the wireless medium. This presents a natural opportunity for cooperation. Fig. 2.3 shows a canonical cooperative network with a single relay (R) to help transmission
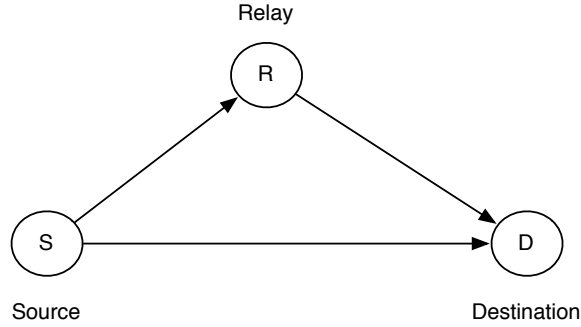
Figure 2.3: Relay Channel.

from S to D. This is also known as the *relay channel*.

The earliest generalized information theoretic study of the relay channel was presented by Cover and El Gamal in [16]. Two fundamental relaying schemes were presented in this paper namely, *decode-and-forward* (DF) and *compress-and-forward* (CF) with corresponding theorems for channel capacity. Since this seminal work, performance limits for relay networks have received a lot of interest.

The relay channel can be studied in context of either half-duplex or full-duplex relays. A *full-duplex* relay transmits and receives in the same frequency at the same time. On the other hand, a *half-duplex* relay uses transmit and receive modes that are non-overlapping e.g. using time division, frequency division or code division. Practical design of wireless radios is currently limited to half-duplex operation. Therefore, this dissertation focusses solely on half-duplex cooperation.

2.3.1 System Model

A *discrete memoryless* Gaussian relay channel system model is used. It is indicated by $P(\mathbf{y}, \mathbf{y}_R | \mathbf{x}_S, \mathbf{x}_R)$ as shown in Fig. 2.4. It has three half-duplex terminals: source (S), relay (R) and destination (D) each of which have m,n and k antennas respectively with AWGN channels. It is called the (m, n, k) relay channel. $\mathbf{X}_S \in \mathbb{C}^{m \times N_S}$ and $\mathbf{X_R} \in \mathbb{C}^{n \times N_R}$ are signal blocks transmitted by the antennas at S and R respectively. Each block is a sequence of complex symbols having lengths N_S and N_R respectively. Here bold-face lower case alphabet denotes a column vector of symbols.

$$\mathbf{X}_S = [\mathbf{x}_{S,1}\mathbf{x}_{S,2} \ldots \mathbf{x}_{S,N_S}] \tag{2.13}$$
$$\mathbf{X}_R = [\mathbf{x}_{R,1}\mathbf{x}_{R,2} \ldots \mathbf{x}_{R,N_R}] \tag{2.14}$$

Without loss of generality, it is assumed that R uses time division for half-duplex operation. R listens for fraction $f \in [0, 1]$ of the signal block and transmits for fraction $(1 - f)$ as shown in Figure 2.5. The block lengths for transmit sequences at S and R satisfy the half-duplex
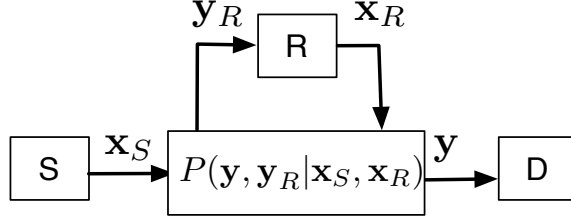
Figure 2.4: Discrete Memoryless Relay Channel

constraint:

$$N_R = (1 - f)N_S$$

Per-node constraints on average symbol power are considered at the terminals. See Remark 1.

$$E||\mathbf{x}_{S,i}||^2 \leq P_S$$
$$E||\mathbf{x}_{R,i}||^2 \leq P_R$$

The received signal sequences $\mathbf{Y} \in \mathbb{C}^{k \times N_S}$ and $\mathbf{Y}_R \in \mathbb{C}^{n \times fN_S}$ are:

$$\mathbf{Y} = [\mathbf{y}_1 \mathbf{y}_2 \ldots \mathbf{y}_{N_S}] \tag{2.15}$$
$$\mathbf{Y}_R = [\mathbf{y}_{R,1} \mathbf{y}_{R,2} \ldots \mathbf{y}_{R,fN_S}] \tag{2.16}$$

The received symbols are given as follows:

$$\mathbf{y}_i = \mathbf{H}_1 \mathbf{x}_{S,i} + \mathbf{H}_2 \mathbf{x}_{R,i} + \mathbf{z}_i$$
$$\mathbf{y}_{R,j} = \mathbf{H}_R \mathbf{x}_{S,j} + \mathbf{z}_{R,j}$$

Here $\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_R$ denote the $S \rightarrow D$, $R \rightarrow D$ and $S \rightarrow R$ channel matrices respectively. $\mathbf{H}_2 = \mathbf{0}$ for $i \in \{1, 2, \ldots, fN_S\}$ when R is listening. $\mathbf{z_i}$ and $\mathbf{z}_{R,j}$ are i.i.d. zero-mean Gaussian noise vectors with covariance matrices $\sigma^2 \mathbf{I}$ and $\sigma_R^2 \mathbf{I}$ respectively. For Rayleigh fading, all elements of the channel matrices are i.i.d $\mathcal{CN}(0, 1)$. $\mathbf{X}_S$ (2.13) and $\mathbf{Y}$ (2.15) are partitioned into $\mathbf{X}_S^f, \mathbf{X}_S^{(1-f)}$ and $\mathbf{Y}^f, \mathbf{Y}^{(1-f)}$ using superscript to represent relay's state e.g. $\mathbf{X}_S^f = [\mathbf{x}_{S,1} \mathbf{x}_{S,2} \ldots \mathbf{x}_{S,fN_S}]$. This channel is characterized using the following average SNR parameters:

$$\mathsf{SNR}_{SR} = \frac{P_S}{\sigma_R^2}, \ \mathsf{SNR}_{SD} = \frac{P_S}{\sigma^2}, \ \mathsf{SNR}_{RD} = \frac{P_R}{\sigma^2}$$

Remark 1. The system model above, uses *per node* constraints on transmit power without an additional *network* power constraint (sum of average transmit power at S and R). If
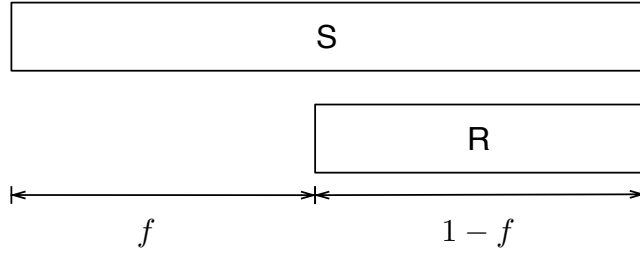
Figure 2.5: Relay listens for f fraction of time and transmits in remaining fraction.

the network power constraint is dominant, power allocation between terminals needs to be considered [24]. However, in the high SNR regime, per-node power constraints tend to be dominant and power allocation does not improve system performance significantly therefore the network constraint is omitted in our analysis.

2.3.2 Capacity Bounds

While the capacity of the relay channel is not known in general, the following bounds act as guidelines for information theoretic limits.

Cut-Set Bound

The cut-set bound (max flow min cut theorem) [21, 18] is a useful upper-bound for the capacity of multi-terminal networks. The network in Figure 2.3 has two cuts namely $\mathcal{C}_S = \{S\}, \{R, D\}$ and $\mathcal{C}_R = \{S, R\}, \{D\}$. The mutual information across these cuts for a given listening fraction f is given as follows:

$$I_{\mathcal{C}_S}(f) = fI(\mathbf{X}_S^f; \mathbf{Y}_R, \mathbf{Y}^f) + (1-f)I(\mathbf{X}_S^{(1-f)}; \mathbf{Y}^{(1-f)}|\mathbf{X}_R) \tag{2.17}$$

$$I_{\mathcal{C}_D}(f) = fI(\mathbf{X}_S^f; \mathbf{Y}^f) + (1-f)I(\mathbf{X}_S^{(1-f)}, \mathbf{X}_R; \mathbf{Y}^{(1-f)}) \tag{2.18}$$

Here $I(;)$ represents mutual information and conditional mutual information as defined in [17]. The maximum achievable rate beyond which reliable communication is not possible is bounded using the cut-set bound as follows [17]:

$$\mathcal{R}^{(cutset)} \leq \max_{P(\mathbf{X}_S, \mathbf{X}_R)} \min(I_{\mathcal{C}_S}(f), I_{\mathcal{C}_D}(f)) \tag{2.19}$$

Decode and Forward

Decode and forward is one of the fundamental schemes for the relay channel introduced in [16] and studied extensively in literature. Under the DF scheme, R decodes its received

message and forwards it to D after re-encoding. During the cooperation phase (f) when R is listening, S must make sure that R can decode i.e. it must use a rate $\leq I(\mathbf{X}_S^f; \mathbf{Y}_R)$. The maximum achievable rate under DF for a given listening fraction f is given as follows:

$$\mathcal{R}^{(DF)} \leq \max_{P(\mathbf{X}_S, \mathbf{X}_R)} \min\{ fI(\mathbf{X}_S^f; \mathbf{Y}_R) + (1-f)I(\mathbf{X}_S^{(1-f)}; \mathbf{Y}^{(1-f)} | \mathbf{X}_R),$$
$$fI(\mathbf{X}_S^f; \mathbf{Y}^f) + (1-f)I(\mathbf{X}_S^{(1-f)}, \mathbf{X}_R; \mathbf{Y}^{(1-f)}) \}$$

Compress and Forward

In CF, the transmission from R depends on a *soft* estimate of $\mathbf{X}_S^f$, as opposed to DF which uses a *hard* estimate. Soft estimate forwarding relaxes the constraint for decoding at the relay. R compresses its observation $\mathbf{Y}_R$ to form $\hat{\mathbf{Y}}_R$ and maps it to a transmit codeword $\mathbf{X}_R$. The destination D first decodes $\mathbf{X}_R$ from $\mathbf{Y}^{(1-f)}$ and recovers $\hat{\mathbf{Y}}_R$ by treating the transmission from S as noise. Since the transmission from S acts as side information at D, R can use Wyner-Ziv [74] type compression having rate $I(\hat{\mathbf{Y}}_R; \mathbf{Y}_R | \mathbf{Y}^f)$. In order to ensure that D can decode $\mathbf{X}_R$ this rate cannot be greater than the capacity of the $R \rightarrow D$ link. The resulting constraint is expressed as follows:

$$fI(\hat{\mathbf{Y}}_R; \mathbf{Y}_R | \mathbf{Y}^f) \leq (1-f)I(\mathbf{X}_R; \mathbf{Y}^{(1-f)}) \tag{2.20}$$

D decodes $\mathbf{X}_S$ using $\hat{\mathbf{Y}}_R$ and $\mathbf{Y}$. The maximum achievable rate under CF for a given listening fraction f is given by

$$\mathcal{R}^{CF} \leq \max_{P(\mathbf{X}_S, \mathbf{X}_R, \hat{\mathbf{Y}}_R, \mathbf{Y})} fI(\mathbf{X}_S; \hat{\mathbf{Y}}_R, \mathbf{Y}^f) + (1-f)I(\mathbf{X}_R; \mathbf{Y}^{(1-f)}) \tag{2.21}$$

subject to (2.20).

Quantize Map and Forward (QMF)

QMF is a modified version of the CF scheme proposed in [6]. Under QMF, R compresses its observation $\mathbf{Y}_R$ to $\hat{\mathbf{Y}}_R$ and maps it to transmit codeword $\mathbf{X}_R$ without performing Wyner-Ziv compression. D decodes $\mathbf{X}_S$ from $\mathbf{Y}^f$ and $\mathbf{Y}^{(1-f)}$ without explicitly recovering $\hat{\mathbf{Y}}_R$. For a single relay, QMF has the same maximum achievable rate as CF (2.21). However, unlike CF, QMF does not require the relay to have any knowledge of the $R \rightarrow D$ channel strength. Therefore, $\hat{\mathbf{Y}}_R$ does not need to satisfy (2.20). For networks with multiple relays it has been shown [6] that QMF outperforms CF.

2.4 Cooperative Diversity

In the high SNR regime, MIMO and cooperative techniques play a key role in improving system spectral efficiency. Therefore, the high SNR DMT discussed in Sec 2.2 is a natural fit to analyze the relay channel. The DMT limits of MIMO channels act as upper-bounds for corresponding cooperative configurations. Various relaying schemes are compared with each other and corresponding MIMO bounds to draw insights into their performance.

A relay channel with antenna configuration having form $(m, n, k = m)$ is referred to as a *cooperative diversity* configuration. The maximum achievable multiplexing gain without cooperation (no relay) i.e. $\min\{m, m\} = m$ is the same as the multiplexing gain with unlimited cooperation i.e. $\min\{m + n, m\} = m$. In these configurations, the range of achievable multiplexing gains r remains the same, but cooperation allows higher diversity for any given r.

2.4.1 Single Antenna Case

The $(1, 1, 1)$ relay channel is the simplest cooperative diversity configuration. The fundamental DMT of the 1×1 SISO ("no cooperation") and 2×1 MISO ("unlimited cooperation") channels act as DMT benchmarks for this channel. Several relaying schemes are proposed for the $(1, 1, 1)$ configuration in literature. These include Amplify-and-Forward (AF) [38], wherein the relay simply forwards its received signal without decoding or compression. AF is traditionally defined as an "orthogonal" scheme, i.e. during the $(1 - f)$ relay-transmit phase, only relay transmits and the source remains silent. A non-orthogonal version of the scheme (NAF) is proposed in [45] which allows S and R to transmit simultaneously. Another prominent scheme is *dynamic* decode-and-forward (DDF). In DDF [7] the relay listening fraction f is chosen dynamically based on instantaneous channel realizations to minimize outage probability. DDF outperforms AF, NAF and traditional DF. The DMT for DDF as shown in Fig 2.6 achieves the 2×1 MISO bound for multiplexing gains $r \leq 0.5$. At high multiplexing gains ($r > 0.5$), the DDF relay spends a large fraction of time listening and thereby reducing the throughput. To address this drawback, another scheme is proposed in [55]. It is called partial-dynamic-decode-and-forward (PDDF), where the transmission from source is split into two parts. The relay is required to decode only one of those parts using the DDF protocol. The relay does not participate in the second part, thereby partially relaxing the constraint on listening time at the relay. While PDDF has a better DMT performance compared to DDF it still does not achieve the 2×1 MISO bound. Finally, it is shown in [75][52] that using soft decision forwarding schemes like CF or QMF the 2×1 MIMO DMT bound is achievable with half duplex cooperation. As opposed to CF, QMF achieves the DMT upper bound with reduced channel knowledge at relay. The scheme in [75][52] uses a fixed listening fraction of $f = 0.5$. These results highlight the power of half-duplex cooperation in terms of system performance at high SNR.

Fig 2.6 shows DMT performance results for the various schemes described above over the
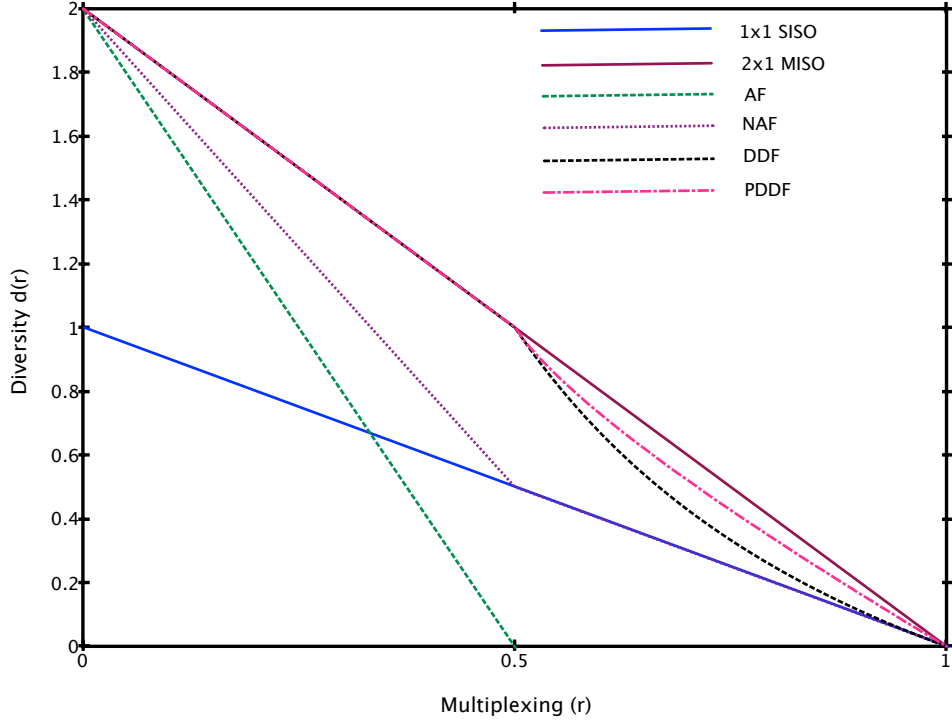
Figure 2.6: Diversity Multiplexing Tradeoff for the $(1, 1, 1)$ Half Duplex Relay Channel.

$(1, 1, 1)$ relay channel. Also shown are corresponding DMT bounds for 1×1 SISO and 2×1 MISO.

2.4.2 Multiple Antennas

Unlike the single antenna case, the half duplex relay channel with multiple antennas cannot achieve corresponding MIMO DMT bounds. For the $(m = 1, n, k = 1)$ relay channel, it is shown in [75] that the fundamental DMT is given as follows:

$$d^*_{(1,n,1)}(r) = \begin{cases} 1 + n - 2nr & 0 \leq r \leq 0.5 \\ 2(1 - r) & 0.5 \leq r \leq 1 \end{cases}$$

The above tradeoff is achievable using soft-decision forwarding schemes e.g. CF and QMF and a static listening schedule with $f = 0.5$. Fig 2.7 shows the tradeoff for $n = 2$ in comparison with the 3×1 MISO bound.

A general form of the above result is reported in [39] for all cooperative diversity configurations $(m, n, k = m)$. The fundamental tradeoff is piecewise linear and can be achieved using a static schedule with $f = 0.5$ and CF/QMF relaying. The corner points of the tradeoff
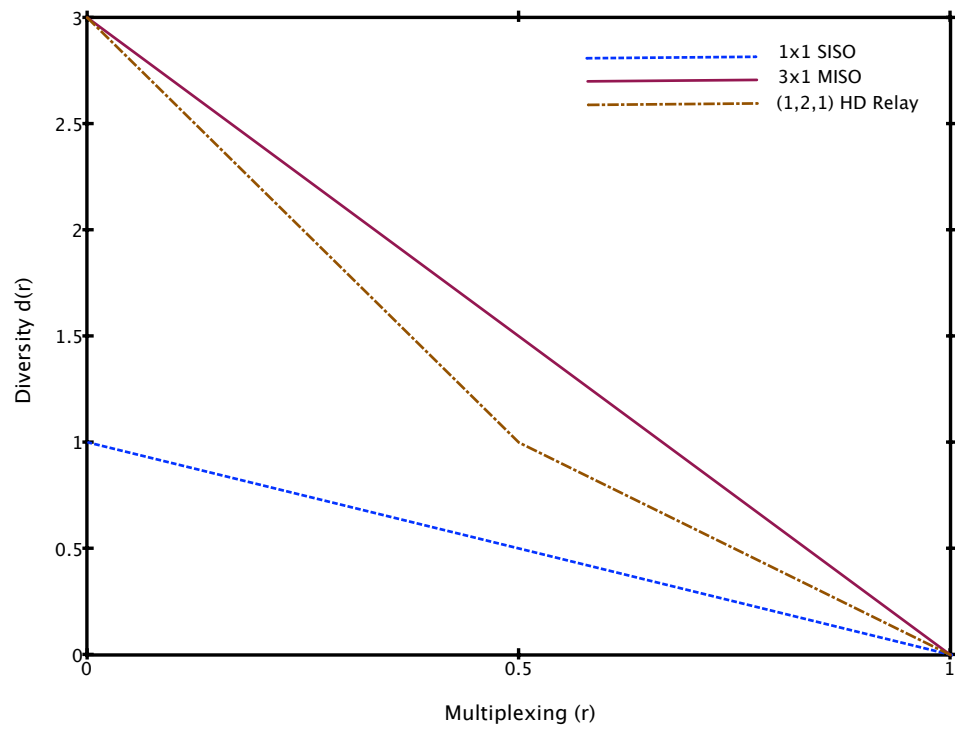
Figure 2.7: Diversity Multiplexing Tradeoff for the $(1, 2, 1)$ Half Duplex Relay Channel.
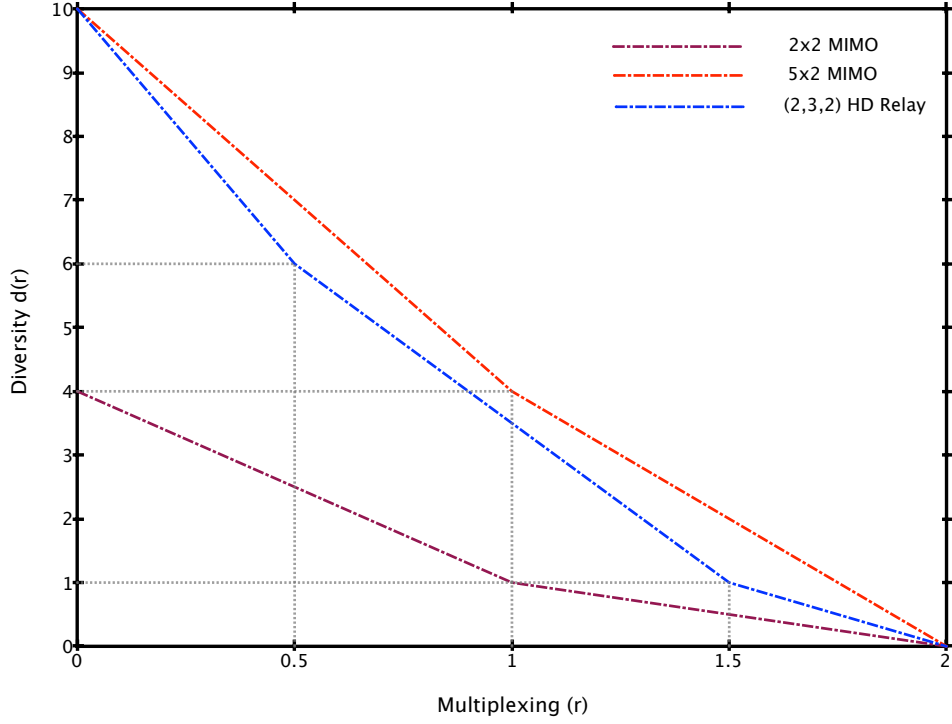
Figure 2.8: Diversity Multiplexing Tradeoff for the $(2, 3, 2)$ Half Duplex Relay Channel.

are given as follows. $l_0 = \min\{m, \lfloor \frac{n+1}{3} \rfloor\}$.

$$d^*_{(m,n,m)}(l/2) = m^2 + (n-l)(m-l)$$
$$\text{for } 0 \le r \le l_0/2 \text{ and } l \in \{0, 1, \ldots, l_0\}$$
$$d^*_{(m,n,m)}(l_0/2 + l) = l^2 + (n+m-l)(m-l_0-l)$$
$$\text{for } l_0/2 \le r \le m - l_0/2 \text{ and } l \in \{0, 1, \ldots, m-l_0\}$$
$$d^*_{(m,n,m)}(m-l/2) = l^2$$
$$\text{for } m - l_0/2 \le r \le m \text{ and } l \in \{0, 1, \ldots, l_0\}$$

Fig. 2.8 shows the DMT for $(2, 3, 2)$ configuration with associated "no cooperation" and "unlimited cooperation" bounds.

2.5 Cooperative Multiplexing

Cooperative diversity refers to the additional diversity gain (over direct link) offered by cooperation. Similarly, *cooperative multiplexing* refers to situations where a relay provides additional multiplexing gain (compared to direct link) [20]. This section presents our findings

on relay channel configurations that allow cooperative multiplexing i.e. those having form (m, n, k) with $m \neq k$.

In this dissertation, only the simplest such configuration i.e. $(1, 1, 2)$ is considered. The fundamental DMT is fully characterized for this case. The configuration is of special interest for current cellular and local area networks, wherein base stations commonly have multiple antennas and mobiles are limited to just one antenna (as discussed in Chapter 1). It is shown that if the source and the relay are relatively close to each other, cooperative multiplexing gain is achievable even with half duplex relaying. The full-duplex case is studied in [20]. It is worth noting that for multiplexing gains less than 1, if source-relay SNR (measured in dB) is at least two times the source-destination SNR a simple scheme with the relay listening $\frac{1}{3}$ of the time and transmitting $\frac{2}{3}$ of the time achieves the 2×2 MIMO DMT. These results are reported in [46].

These results lend insight into the fundamental limits of cooperative multiplexing in the half-duplex relay channel. Two key techniques enable our results:

1. *Distance between terminals:* In most results, it is seen that relative distances between source, relay and destination do not affect DMT performance of the relay channel. Since DMT is calculated at high SNR the path loss and therefore distances are not easily captured in results. This apparent limitation is overcome by scaling the average SNR's of the various links differently.

   The proposed approach enriches the DMT framework by adding insights about network geometry. Sec 2.5.1 describes the model in detail.

2. *MIMO with half duplex antenna:* The cut-set capacity bound (2.19) is widely used [75] to calculate an upper bound for DMT performance. Notice in Figure 2.9 that the $\{S, R\}, \{D\}$ cut corresponds to a 2×2 MIMO system with one source antenna that remains active only for a fraction of total communication time (R is half duplex). It is noted in [75] that an upper bound for mutual information across such a cut is hard to compute. As a result, in prior work fundamental DMT results are only known for the special case of diversity configurations ($m = k$).

   In Sec 2.5.3 a simple channel decomposition is proposed, allowing us to compute the cut-set DMT bound for the $(1, 1, 2)$ configuration. The technique can be applied towards computing DMT for general m, n and k. In Sec 2.5.4 it is shown that the cut-set DMT can be achieved using a dynamic QMF scheme.

2.5.1 Asymmetric System Model

Consider the system in Figure 2.9 with source S, relay R and destination D having $1, 1$ and 2 antennas respectively. Let $D_j, j \in \{1, 2\}$ denote the jth antenna at D. The channel gain for $S \rightarrow R$ is h_{sr}, gains for $S \rightarrow D_j$ are h_{sj} and $R \rightarrow D_j$ are h_{rj}. All the channel gains are
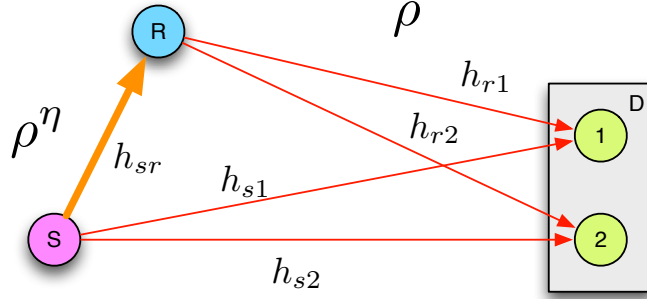
Figure 2.9: Relay channel with 2 antennas at destination and $S \to R$ proximity gain η.

assumed to be flat fading having i.i.d. $\mathcal{CN}(0,1)$ distribution. Quasi-static fading is assumed, i.e. once realized, channel gains remain unchanged for the duration of the codeword and they change independently between codewords. Noise at all receivers is additive i.i.d. $\mathcal{CN}(0,1)$ and independent of all other variables in the system. Transmit power at S and R is limited by an average power constraint specified by the average SNR. For simplicity, it is assumed that transmissions at S and R are synchronous at symbol level.

Consider an asymmetrical network geometry. S and R are modeled to be close to each other as compared to $\{S, R\}$ and D. $S \to D$ and $R \to D$ are assumed to have the same average SNR denoted by ρ. $S \to R$ on the other hand is modeled to have SNR higher than ρ by a factor η on dB scale, i.e. the $S \to R$ average SNR is ρ^η. Thus, the $S \to R$ channel (cooperation link) has $\eta - 1$ more degrees of freedom than other channels in the network. We call η the proximity gain and assume $\eta \geq 1$.

Non-coherent transmission and coherent reception is assumed i.e. only average channel statistics ρ, η are known at S, at D all channel realizations h_{sr}, h_{sj}, h_{rj} are completely known.

Three models for relaying strategy are identified.

- *Global:* The relay uses knowledge of all instantaneous channel realizations to optimize its strategy.

- *Local:* The relay can measure h_{sr} and uses only this (local) information.

- *Blind:* The relay only uses average channel statistics.

The *global* strategy is discussed in Sec 2.5.3 while *local* and *blind* are discussed in Sec 2.5.5.

2.5.2 Results Summary

Theorem 2.5.1. *The maximum achievable DMT for network described in Sec 2.5.1 is given by,*

$$d^*_{(1,1,2)}(r) = \begin{cases} \min\{\eta + 2, 4\} - 3r & 0 \leq r \leq 1, \eta \geq 1 \\ (2\eta - \eta r - 1)/(\eta - r) & 1 \leq r \leq 2 - \frac{1}{\eta}, \eta \geq 2 \\ \eta - \frac{1}{2-r} & 1 \leq r \leq 2 - \frac{1}{\eta}, 1 \leq \eta \leq 2 \end{cases} \qquad (2.22)$$

Corollary 2.5.2. *For system model described in Sec 2.5.1 the maximum achievable multiplexing gain $r^* = \inf_{r \geq 0}\{r|d^*_{(1,1,2)}(r) = 0\}$ is,*

$$r^* = 2 - \frac{1}{\eta} \qquad (2.23)$$

For a symmetrical geometry with all channels having the same degrees of freedom ($\eta = 1$), this implies $r^* = 1$ i.e. cooperation doesn't provide additional multiplexing gain. To enable higher multiplexing gain, the $S \rightarrow R$ channel (cooperation link) needs to have more degrees of freedom than the $S \rightarrow D$ channel (communication link).

Let $d^*_{2\times 2}(r)$ represent the DMT of the 2×2 MIMO channel. For finite η, note that $d^*_{(1,1,2)}(r) \leq d^*_{2\times 2}(r)$ with strict inequality over a non-empty region of r. This suggests that for distributed antennas the finite capacity of the cooperation channel ($S \rightarrow R$) poses a fundamental limitation on the achievable DMT performance. Note that:

$$\lim_{\eta \to \infty} d^*_{(1,1,2)}(r) \rightarrow d^*_{2\times 2}(r)$$

Figure 2.10 shows $d^*_{(1,1,2)}(r)$ for several values of η.

Theorem 2.5.1 is proven in two steps. In Sec 2.5.3 it is shown that the cut-set DMT upper bound for network in Sec 2.5.1 is given by (2.22). In Sec 2.5.4 it is shown that this bound is achievable.

2.5.3 Cut-Set DMT upper bound

Let $0 \leq f \leq 1$ denote a listen-transmit schedule for the half duplex relay. R listens for a fraction f (listening phase) of total communication time and transmits for fraction $(1 - f)$ (cooperation phase). The two cuts of the network $\mathcal{C}_D = \{S, R\}, \{D\}$ and $\mathcal{C}_S = \{S\}\{R, D\}$ are shown in Figure 2.11 for these two phases. The instantaneous mutual information across the two cuts is given in equations (2.17) and (2.18).

To maximize these mutual information expressions, zero-mean complex Gaussian distributions are chosen for $\mathbf{X}_S^f, \mathbf{X}_S^{(1-f)}$ and $\mathbf{X}_R$. The distributions have covariance matrices which satisfy respective average power constraints. Using these, mutual information upper bounds $I'_{\mathcal{C}_S}$ and $I'_{\mathcal{C}_D}$ for $I_{\mathcal{C}_S}$ and $I_{\mathcal{C}_D}$ are computed respectively.
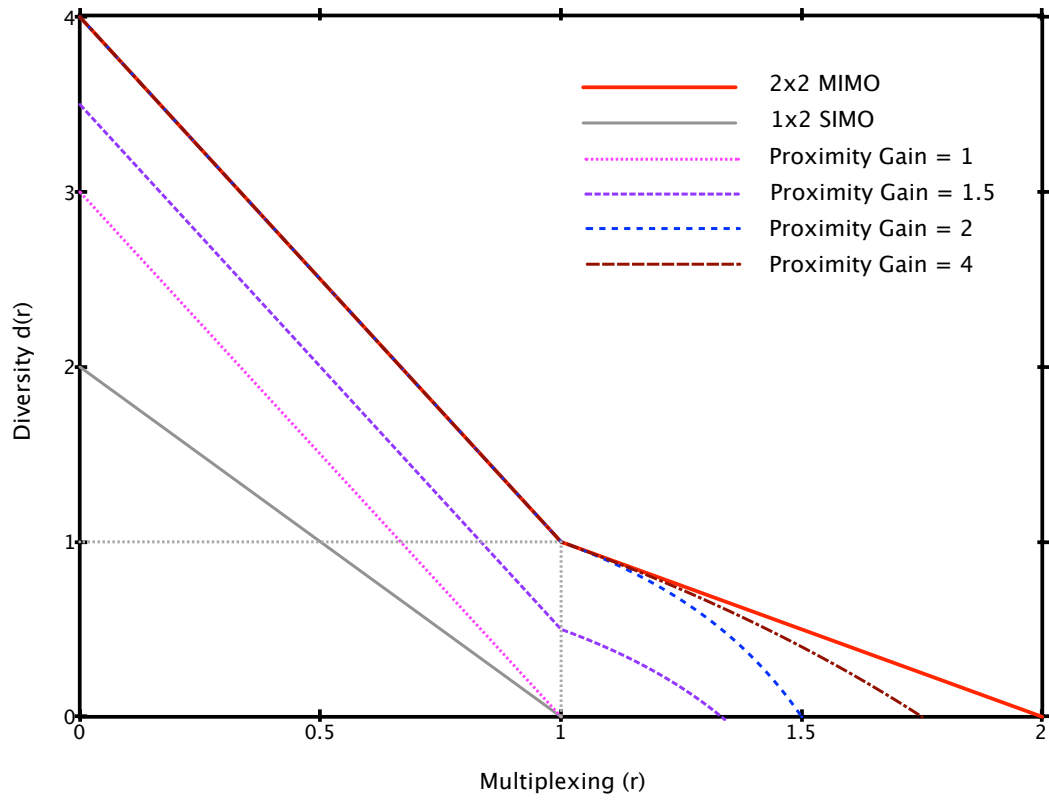
Figure 2.10: $d^*_{(1,1,2)}(r)$ for various values of proximity gain η.

	$\mathcal{C}_S = \{S\}, \{R, D\}$	$\mathcal{C}_D = \{S, R\}, \{D\}$
f		
$1 - f$		
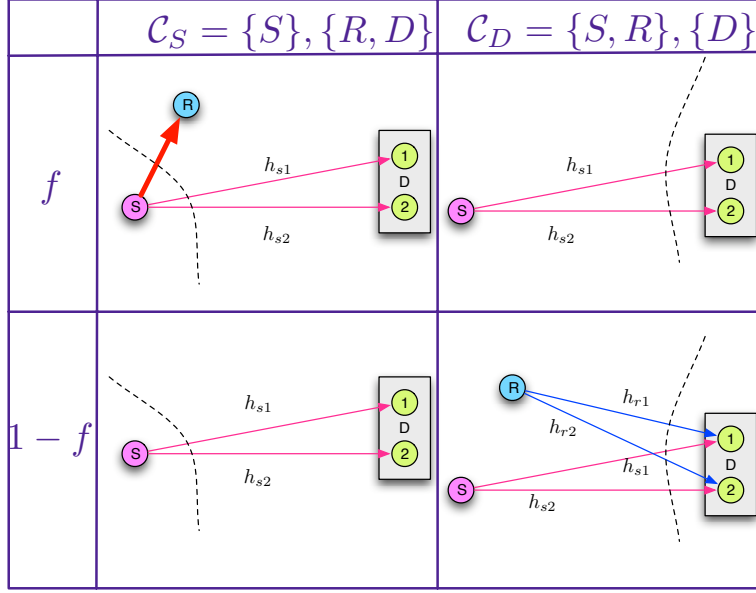
Figure 2.11: Two cuts of network during listen (f) and cooperation phase.$(1 - f)$.

$$
\begin{aligned}
I_{\mathcal{C}_S} \leq I'_{\mathcal{C}_S} &= f \log(1 + \rho^\eta |h_{sr}|^2 + \rho ||\mathbf{h}_s||^2) \\
&\quad + (1 - f) \log(1 + \rho ||\mathbf{h}_s||^2) \\
&\approx f \max\{\log(1 + \rho^\eta |h_{sr}|^2), \log(1 + \rho ||\mathbf{h}_s||^2)\} \\
&\quad + (1 - f) \log(1 + \rho ||\mathbf{h}_s||^2) \\
I_{\mathcal{C}_D} \leq I'_{\mathcal{C}_D} &= f \log(1 + \rho ||\mathbf{h}_s||^2) \\
&\quad + (1 - f) \log \det(\mathbf{I} + \rho \mathbf{H} \mathbf{H}^\dagger)
\end{aligned}
$$

Here $\mathbf{h}_s = \begin{bmatrix} h_{s1} \\ h_{s2} \end{bmatrix}, \mathbf{h}_r = \begin{bmatrix} h_{r1} \\ h_{r2} \end{bmatrix}$ and $\mathbf{H} = [\ \mathbf{h}_s \quad \mathbf{h}_r\ ]$. The approximation is tight to within one bit.

Note that the expression for $I'_{\mathcal{C}_D}$ is a linear combination of the capacities of two Raleigh fading Gaussian channels having correlated channel matrices $\mathbf{H}_1 = [\ \mathbf{h}_s \quad 0\ ]$ and $\mathbf{H} = [\ \mathbf{h}_s \quad \mathbf{h}_r\ ]$. Outage probability for the MIMO channel is calculated in [80] using eigenvalue decomposition of the channel matrix. If this same technique is followed, it requires computing the joint eigenvalue distributions for two correlated hermitian matrices, $\mathbf{H}_1 \mathbf{H}_1^\dagger$ and $\mathbf{H} \mathbf{H}^\dagger$. It is noted in [75] that this is hard to compute. An easier decomposition can solve this problem. The second term in $I'_{\mathcal{C}_D}$ is the capacity of a 2×2 MIMO channel which can be represented by two parallel Gaussian channels having gains g_1 and g_2 shown in Figure 2.12. The channel is given as follows:
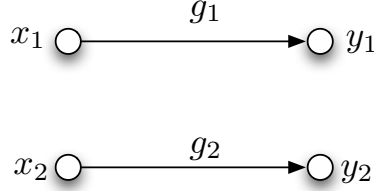
Figure 2.12: Parallel Channel Model for 2×2 MIMO.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \sqrt{\rho} \begin{bmatrix} g_1 & 0 \\ 0 & g_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

Here $E[|x_i|^2] = 1$, $w_1, w_2 \sim \mathcal{CN}(0,1)$. The capacity for $x_i \to y_i$ is given by $\log(1 + \rho g_i^2)$. It is shown in [71][80] that a D-BLAST transmission scheme with MMSE-SIC receiver achieves the mutual information of the MIMO channel. For this scheme g_1^2 and g_2^2 are calculated as follows:

$$\begin{aligned} g_2^2 &= ||\mathbf{h}_{r\perp s}||^2 + \frac{||\mathbf{h}_{r\|s}||^2}{1 + \rho||\mathbf{h}_s||^2} \\ g_1^2 &= ||\mathbf{h}_s||^2 \end{aligned}$$

where $\mathbf{h}_{r\perp s}$ and $\mathbf{h}_{r\|s}$ respectively denote the perpendicular and parallel components of $\mathbf{h}_r$ with respect to $\mathbf{h}_s$.

Note that while g_1^2 and g_2^2 are correlated, $\mathbf{h}_s, \mathbf{h}_{r\perp s}$ and $\mathbf{h}_{r\|s}$ are mutually independent. Therefore the correlation between g_1^2 and g_2^2 is analyzable. The destination decodes $\mathbf{X}_R$ in the presence of interference from $\mathbf{X}_S^{(1-f)}$. It cancels $\mathbf{X}_R$ from its received signal before decoding $\mathbf{X}_S^{(1-f)}$. Therefore, $S \to D$ is effectively an interference free channel (with gain g_1) during both the listen and the cooperation phases.

$$\begin{aligned} I'_{\mathcal{C}_D} &= f \log(1 + \rho g_1^2) \\ &\quad + (1-f)[\log(1 + \rho g_1^2) + \log(1 + \rho g_2^2)] \\ &= \log(1 + \rho g_1^2) + (1-f)\log(1 + \rho g_2^2) \end{aligned}$$

Let α_{sr}, α_1 and α_2 represent channel realizations via the following variable transforma-

tions.

$$\begin{aligned}
\alpha_{sr} &= \lim_{\rho \to \infty} \frac{\log(1 + \rho^\eta |h_{sr}|^2)}{\log \rho} \\
\alpha_1 &= \lim_{\rho \to \infty} \frac{\log(1 + \rho g_1^2)}{\log \rho} \\
\alpha_2 &= \lim_{\rho \to \infty} \frac{\log(1 + \rho g_2^2)}{\log \rho}
\end{aligned}$$

This gives simplified expressions for mutual information upper bounds.

$$\frac{I'_{\mathcal{C}_S}}{\log \rho} = \alpha_1 + f(\alpha_{sr} - \alpha_1)^+ \tag{2.24}$$

$$\frac{I'_{\mathcal{C}_D}}{\log \rho} = \alpha_1 + (1 - f)\alpha_2 \tag{2.25}$$

To achieve desired multiplexing gain r at high SNR ($\rho \to \infty$) the network must achieve a rate $\mathcal{R} = r \log \rho$. The network is in outage if $\min\{I'_{\mathcal{C}_S}, I'_{\mathcal{C}_D}\} \le r \log \rho$. For a given r and schedule f, the outage region $\mathcal{O}(r, f)$ is defined over channel realizations $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_{sr})$.

$$\mathcal{O}(r, f) = \{\vec{\alpha} | \frac{\min\{I'_{\mathcal{C}_S}, I'_{\mathcal{C}_D}\}}{\log \rho} \le r\} \tag{2.26}$$

The outage probability P_{out} is,

$$P_{\text{out}} = \int_{\vec{\alpha} \in \mathcal{O}(r, f)} f_{\vec{\alpha}}(\alpha_1, \alpha_2, \alpha_{sr})$$

Here $f_{\vec{\alpha}_1}(\alpha_1, \alpha_2, \alpha_{sr})$ is the joint distribution of $(\alpha_1, \alpha_2, \alpha_{sr})$.

Lemma 2.5.3. *Proof see Sec 2.6*

$$f_{\vec{\alpha}}(\alpha_1, \alpha_2, \alpha_{sr}) \doteq \rho^{-s(\vec{\alpha})}$$

where $0 \le \alpha_1, \alpha_2 \le 1$, $0 \le \alpha_{sr} \le \eta$ *and*

$$s(\vec{\alpha}) = \begin{cases} \eta + 4 - 3\alpha_1 - 2\alpha_2 - \alpha_{sr} & \alpha_1 + \alpha_2 \le 1 \\ \eta + 3 - 2\alpha_1 - \alpha_2 - \alpha_{sr} & \alpha_1 + \alpha_2 > 1 \end{cases} \tag{2.27}$$

For a given listen-transmit schedule f, the cut-set DMT bound is given by,

$$d^*_{(1,1,2)}(r, f) = \inf_{\vec{\alpha} \in \mathcal{O}(r,f)} s(\vec{\alpha}) \tag{2.28}$$

Optimizing over all listen-transmit schedules, gives the DMT upper bound.

$$d^*_{(1,1,2)}(r) = \inf_{\vec{\alpha} \in \mathcal{O}(r)} \sup_{f} s(\vec{\alpha}) \tag{2.29}$$

Note that this optimization is performed on a per realization basis, i.e. the optimal f depends on all channel realizations α_{sr}, α_1 and α_2. Therefore this corresponds to the *global* strategy discussed in Sec 2.5.1.

Note that $I'_{\mathcal{C}_S}$ (2.24) is monotonically increasing in f and $I'_{\mathcal{C}_D}$ (2.25) is monotonically decreasing in f. This indicates that the optimal listen-transmit schedule f_{glob} is one which sets $I'_{\mathcal{C}_S} = I'_{\mathcal{C}_D}$.

$$f_{glob} = \frac{\alpha_2}{(\alpha_{sr} - \alpha_1)^+ + \alpha_2}$$

This leads to the solution for $d^*_{(1,1,2)}(r)$ given in (2.22).

2.5.4   Achievability: Relaying Scheme

For the same system model described in Sec 2.5.1 but with a single antenna at destination it is shown in [52] that the QMF relaying scheme (proposed in [5] and [3]) is DMT optimal. It is shown below that QMF adapts naturally to the case with 2 antennas at destination and with some modification, achieves the cut-set DMT bound derived in Theorem 2.5.1.

It is interesting to note that if knowledge of all channel gains was available at the source and at the relay (coherent transmission) then a decode-and-forward relaying scheme would also achieve the cut-set DMT bound. In this section, the QMF relaying scheme is considered because it achieves the cut-set DMT bound even without any knowledge of instantaneous channel gains at S and R. A description of the QMF relaying scheme is provided below.

Description of scheme

Source S has a sequence of messages $w_n \in \{1, 2, \ldots, 2^{T\mathcal{R}}\}$, $n = 1, 2, \ldots$ to be transmitted. Both source S and relay R create random Gaussian codebooks. S randomly maps each message to one of its Gaussian codewords and transmits it using T symbol times giving an overall transmission rate of $\mathcal{R}$. Relay listens to the first fT time symbols of each block i.e. $\mathbf{X}_S^f$. It quantizes $\mathbf{Y}_R$ to $\hat{\mathbf{Y}}_R$ and then randomly maps it into a Gaussian codeword $\mathbf{X}_R$ using a random mapping function $f_R(\hat{\mathbf{Y}}_R)$. It transmits this codeword during the next $(1 - f)T$ symbol times. Given the knowledge of all the encoding functions and signals received, D attempts to decode the message sent by S.

DMT of Quantize-Map

By Theorem 7.4.1 in [3], for any fixed listen-transmit schedule f, the quantize-map relaying scheme, uniformly over all channel realizations achieves a rate within a constant gap to the cut-set upper bound $\min\{I'_{\mathcal{C}_S}, I'_{\mathcal{C}_D}\}$ for that particular f. The random Gaussian code-book generated at source is independent of f. Also the code-book generated at relay depends on f only to determine the length of each codeword $(1 - f)T$.

The relay can generate a larger code-book with each codeword of length T. If the relay chooses a listen-transmit schedule f, it can use the first $(1 - f)T$ symbols of the codeword to compose $\mathbf{X}_R$. The destination always knows the schedule f, hence can adapt its decoder accordingly. This construction allows the claim that QMF achieves a rate within a constant gap of $\min\{I'_{\mathcal{C}_S}, I'_{\mathcal{C}_D}\}$ uniformly for each dynamic choice of f i.e.

$$\min\{I'_{\mathcal{C}_S}, I'_{\mathcal{C}_D}\} - \kappa \leq R_{\text{quantize-map}}(h_{sr}, \mathbf{h}_s, \mathbf{h}_r, \rho, f) \tag{2.30}$$

In the above equation, the constant κ does not depend on the channel gains and SNR. At the order of DMT, which assumes high SNR ($\rho \to \infty$) the effect of κ becomes negligible. This leads to the following theorem for achievability.

Theorem 2.5.4. *For dynamic listen-transmit schedules, the modified QMF relaying scheme as described above achieves the diversity multiplexing tradeoff of $\min\{I'_{\mathcal{C}_S}, I'_{\mathcal{C}_D}\}$, where $I'_{\mathcal{C}_S}$ and $I'_{\mathcal{C}_D}$ are given by (2.24)(2.25).*

2.5.5 Achievability: Listen-Transmit Schedule

In Sec 2.5.3 the cut-set DMT upper bound was calculated for the *globally* optimal listen-transmit schedule f_{glob}. However, in a practical communication scenario global knowledge of instantaneous channel realizations may not be available at the relay. To account for this, the *local* and *blind* relaying strategies are identified in Sec 2.5.1. In this section, Theorem 2.5.1 is refined to calculate DMT bounds for *local* and *blind* schedules.

Blind Scheduling

Theorem 2.5.5. *For low multiplexing gain region i.e. $r \leq 1$, the* blind *scheduling strategy is DMT optimal. Additionally for $\eta \geq 2$ the* blind *strategy achieves the 2×2 MIMO DMT bound for $r \leq 1$. The optimal blind schedule for this region is $f_{blind} = \frac{1}{3}$*

$$d^{blind}_{(1,1,2)}(r) = \min\{\eta + 2, 4\} - 3r \quad r \leq 1 \tag{2.31}$$

From (2.28), the DMT bound for blind scheduling can be written as follows:

$$d^{blind}_{(1,1,2)}(r) = \max_f \min_{\vec{\alpha} \in \mathcal{O}(r,f)} s(\vec{\alpha}) \tag{2.32}$$
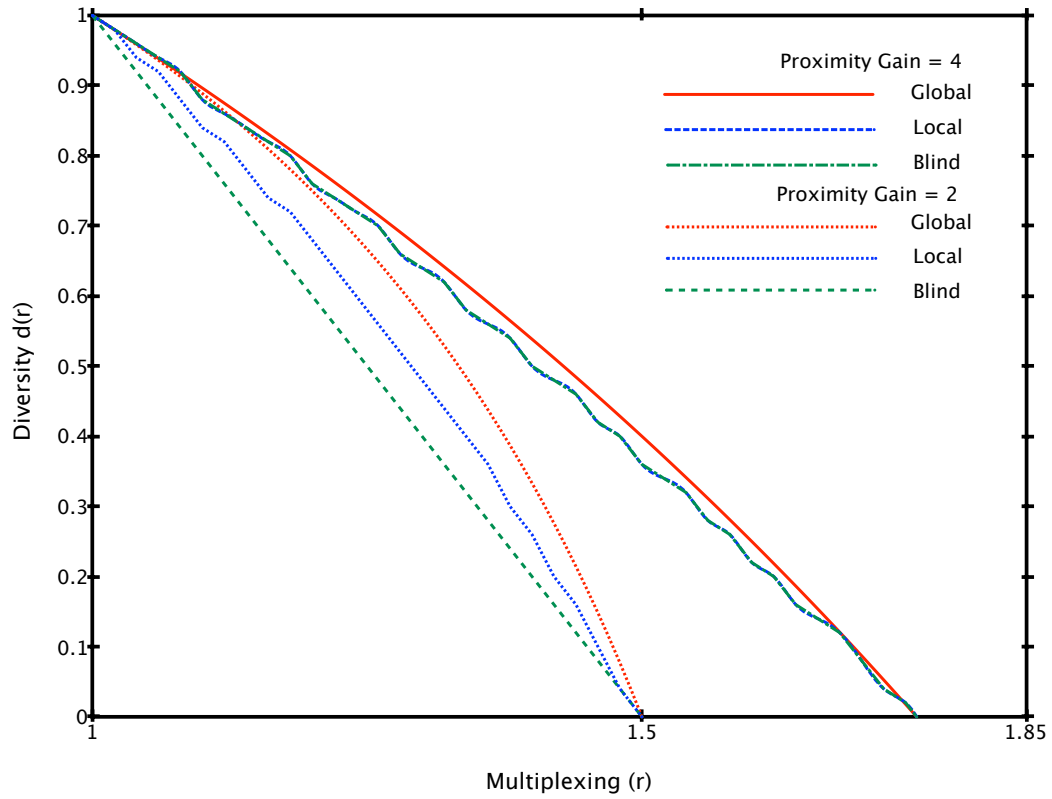
Figure 2.13: $d^*_{(1,1,2)}(r)$ $d^{local}_{(1,1,2)}(r)$ and $d^{blind}_{(1,1,2)}(r)$ comparison for $r \geq 1$.

i.e. f is optimized without knowledge of channel realizations $\vec{\alpha}$. Solving this optimization for $r \leq 1$ yields Theorem 2.5.5. This suggests that as long as cooperative multiplexing is not necessary i.e. desired rate $\mathcal{R} = r \log(\rho)$ is such that $r \leq 1$, static scheduling at relay is sufficient to achieve the DMT upper bound. $f_{blind} = \frac{1}{3}$ turns out to be the optimal listen-transmit schedule for this region.

For the high rate region ($r > 1$), the analytical solution for (2.32) is tedious to obtain. The optimization is convex and has numerical solutions. Figure 2.13 shows a comparison between $d^*_{(1,1,2)}(r)$ and $d^{blind}_{(1,1,2)}(r)$ for $r \geq 1$. Note that for cooperative multiplexing ($r \geq 1$) static scheduling is insufficient to achieve DMT upper bound.

Local Scheduling

Similarly, the DMT bound for local scheduling is expressed as an optimization problem from (2.28).

$$d^{local}_{(1,1,2)}(r) = \min_{\alpha_{sr}} \max_{f} \min_{\alpha_1, \alpha_2 \in \mathcal{O}(r,f)} s(\vec{\alpha}) \tag{2.33}$$

f_{local} is optimized using knowledge of α_{sr} only. The DMT performance of local scheduling must be at-least as good as blind scheduling, therefore by Theorem 2.5.5 for $r \leq 1$ $d^{local}_{(1,1,2)}(r) = d^{blind}_{(1,1,2)}(r) = d(r)$.

Numerical solution to (2.33) for the high rate $r \geq 1$ region is shown in Figure 2.13. Note that local scheduling performs better than blind, but for higher η this advantage diminishes.

2.6 Proof of Lemma 2.5.3

2.6.1 Marginal Distribution of α_{sr}

$f_{\alpha_{sr}}(\alpha)$ is calculated as,

$$\begin{aligned}
\mathcal{P}[\alpha_{sr} < \alpha] &= \lim_{\rho \to \infty} \mathcal{P}[|h_{sr}|^2 < \rho^{\alpha_{sr} - \eta}] \\
f_{\alpha_{sr}}(\alpha) &\doteq \rho^{\eta - \alpha}(0 \leq \alpha_{sr} \leq \eta)
\end{aligned}$$

2.6.2 Joint Distribution of α_1 and α_2

Note that g_1^2 has a χ_4^2 distribution, the marginal distribution of α_1 is given by,

$$\begin{aligned}
f_{\alpha_1}(\alpha_1) &= f_{g_1^2}(\rho^{-(1-\alpha_1)}) \frac{dg_1^2}{d\alpha_1} \\
&\doteq \rho^{-2(1-\alpha_1)}(0 \leq \alpha_1 \leq 1)
\end{aligned}$$

Now for $(0 \leq \alpha_1, \alpha_2 \leq 1)$ their joint CDF can be written as,

$$
\begin{aligned}
F_{\alpha_1,\alpha_2}(\alpha_1, \alpha_2) &= \mathcal{P}[g_1^2 \leq \rho^{\alpha_1-1}, g_2^2 \leq \rho^{\alpha_2-1}] \\
&= \int_{x=0}^{\alpha_1} \mathcal{P}[g_1^2 = \rho^{x-1}, g_2^2 \leq \rho^{\alpha_2-1}]dx \\
&\doteq \int_{x=0}^{\alpha_1} \rho^{-2(1-x)} \\
&\qquad \mathcal{P}(|h_{s\perp r}|^2 + \frac{|h_{s\|r}|^2}{1+\rho^x} \leq \rho^{\alpha_2-1})dx \\
&\doteq \int_{x=0}^{\alpha_1} \rho^{-2(1-x)} \\
&\qquad \int_{y=0}^{\alpha_2-1} \rho^y \mathcal{P}(|h_{s\|r}|^2 \dot{\leq} \rho^{x+\alpha_2-1})\ dy\ dx
\end{aligned}
$$

Case $\alpha_1 + \alpha_2 \leq 1$

$$
\begin{aligned}
F_{\alpha_1,\alpha_2}(\alpha_1, \alpha_2) &\doteq \int_{x=0}^{\alpha_1} \rho^{-2(1-x)} \int_{y=0}^{\alpha_2-1} \rho^{x+y+\alpha_2-1}\ dy\ dx \\
&\doteq \int_{x=0}^{\alpha_1} \rho^{3x+2\alpha_2-4}\ dx \\
f_{\alpha_1,\alpha_2}(\alpha_1, \alpha_2) &\doteq \rho^{3\alpha_1+2\alpha_2-4}
\end{aligned}
$$

Case $\alpha_1 + \alpha_2 > 1$

$$
\begin{aligned}
F_{\alpha_1,\alpha_2}(\alpha_1, \alpha_2) &\doteq \int_{x=0}^{1-\alpha_2} \rho^{3x+2\alpha_2-4}\ dx \\
&\qquad + \int_{x=1-\alpha_2}^{\alpha_1} \rho^{2x+\alpha_2-3}\ dx \\
f_{\alpha_1,\alpha_2} &\doteq \rho^{2\alpha_1+\alpha_2-3}
\end{aligned}
$$

Since α_{sr} is independent of α_1, α_2 we get Lemma 2.5.3.

2.7   Summary

In this chapter information theoretic limits of cooperative multiplexing are derived. These results show that half-duplex cooperation is a promising technique for improving link throughput.

Chapter 3

Coding for Quantize-Map-and-Forward

In the previous chapter, information theoretic results that demonstrate the potential of cooperative relaying were presented. Information theory only proves the existence of error correcting codes and relay mapping functions. To bring these ideas closer to practice, explicit construction of codes that perform close to information theoretic limits and have efficient encoding/decoding algorithms is required. In this chapter progress is reported towards this goal by developing coding schemes for QMF relaying.

In the past, code design for cooperative systems has focussed mainly on the DF and CF relaying schemes. The QMF [6] relaying scheme has several properties that makes it suitable for practical implementation. Additionally, QMF has been shown to achieve the capacity of arbitrary Gaussian relay networks to within a constant gap. As discussed in Sec 2.3.2, the soft decisions (side information) forwarded by relays are not decoded explicitly under QMF. Instead the message is decoded *jointly* with the side information. This presents a unique challenge in the design of efficient coding schemes.

In this chapter, a practical coding framework is developed to addresses this challenge. The initial focus is on the *binary* input Gaussian relay channel. Two code configurations based on combinations of binary low density parity check (LDPC) and low density generator matrix (LDGM) codes are proposed. It is shown that the schemes have linear-time complexity for encoding, relay-mapping and *joint* decoding operations. Simulation results are given using off-the-shelf codes to demonstrate that the schemes provide cooperation gains that are close to those predicted by information theory. Finally, the *binary* coding framework is extended for use with high order signal constellations using bit-interleaved coded modulation (BICM).

3.1 Error Control Coding

In this section, an overview of error control coding for traditional binary input point-to-point channels is presented. A *message* is represented by a sequence of k bits $\mathbf{a} \in \{0, 1\}^k$. It is encoded into $n > k$ bits using a code C to form a transmit *codeword* $\mathbf{b} \in \{0, 1\}^n$. The codeword is modulated to construct symbols $\mathbf{x} \in \mathbb{C}^n$ that are transmitted over the channel. The decoder receives a sequence of symbols $\mathbf{y} \in \mathbb{C}^n$ and attempts to estimate the transmitted message.

Error correcting codes are divided into two structural classes namely, *convolutional* and *block* codes. Block codes encode every message sequence $\mathbf{a}$ to codeword $\mathbf{b}$ *independently*. On the other hand, convolutional codes (e.g. Turbo codes) exhibit *memory* i.e. codewords depend not only on one message but several previous messages. In this dissertation, only schemes based on block coding are considered.

The code rate $\mathcal{R}_{\mathsf{C}} = \frac{k}{n} \leq 1$ represents the number of information/message bits per transmit symbol. The code C is a k-dimensional subspace of the vector space V_n of all binary sequences of length n. If the mapping is linear, then C is said to be a *linear* block code. Linearity is a desirable property to reduce the implementation complexity of encoding. Linear blocks codes can be characterized in terms of a *generator matrix* $\mathbf{G}$ having k rows and n columns ($k \times n$). The encoding operation can be written as:

$$\mathbf{b} = \mathbf{a}\mathbf{G}$$

Here $\mathbf{b}$ and $\mathbf{a}$ are row vectors. An alternate representation of a linear block code is given by its *parity check matrix* $\mathbf{H}$ having dimensions $(n - k) \times n$. A bit-sequence $\mathbf{b}$ is a codeword in C generated by $\mathbf{G}$ if and only if $\mathbf{b}\mathbf{H}^T = \mathbf{0}$. The rows of the parity check matrix represent $(n - k)$ parity constraints that must be satisfied by any $\mathbf{b} \in \{0, 1\}^n$ to be in C.

3.1.1 Low Density Parity Check Codes

LDPC codes were introduced by Gallager in [23]. For point-to-point communication over Gaussian channels, LDPC codes have the best known performance. They have been shown to perform within 0.004dB of the Shannon limit [14]. In addition to their high performance, they can be encoded using a linear-time algorithm and decoded using an intrinsically parallel, linear time *belief propagation* algorithm.

LDPC codes use a *sparse* parity check matrix $\mathbf{H}$. Sparsity ensures that the number of non-zero entries in $\mathbf{H}$ grows only linearly with block length. They can be visualized using a bipartite graph having adjacency matrix given by $\mathbf{H}$. This is called a Tanner graph [73]. In Fig 3.1, a Tanner graph and parity check matrix is shown for an example code having rate $\mathcal{R}_{\mathsf{C}} = \frac{2}{6} = \frac{1}{3}$ and block length $n = 6$. Circles represent the n bits, which are also called *variable nodes*. Squares represent the $(n - k) = 4$ parity check constraints, which are also called *check* or *function* nodes.

$$\mathbf{bh}_0^T = \mathbf{0} \quad \mathbf{bh}_1^T = \mathbf{0} \quad \mathbf{bh}_2^T = \mathbf{0} \quad \mathbf{bh}_3^T = \mathbf{0}$$

$$\mathbf{H} \;=\; \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{h}_0 \\ \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \end{bmatrix}$$
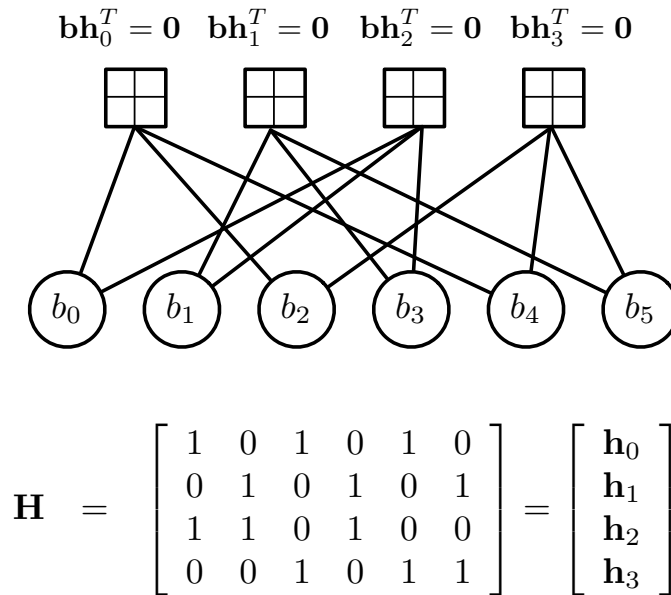
Figure 3.1: Parity check matrix and Tanner graph for example LDPC code.

Decoding is based on an iterative belief propagation algorithm [54] over the Tanner graph. Prior beliefs are calculated for every bit in $\mathbf{b}$ from channel observations $\mathbf{y}$. The variable nodes and check nodes update these beliefs by iteratively passing messages along the edges in the graph. For an outgoing message along any given edge, the nodes compute updated beliefs using *extrinsic* information i.e. based on incoming messages from every other edge. As iterations progress, every variable node receives updates based on information from distant nodes in the graph. The algorithm is terminated after a few (usually tens of) iterations, and final beliefs are read out from the variable nodes. Based on final beliefs, a *hard* decision is made about the message.

In 1963, the bipartite graph representation and associated belief propagation algorithm were first introduced for LDPC codes by Gallager [23]. Later in 1981, Tanner [65] showed that these graph-theoretic models can be generalized for other codes as well. In 1996, Wiberg et. al [73] further generalized the formulation to include *hidden* variables that can represent decoder state in addition to codeword symbols. In 1997, a formulation called *factor graphs* was developed using graphical models to compute marginal product of functions. It was reported jointly in two landmark papers [36] by Kshischang, Frey, Loeliger and [1] by Aji, McEliece. Several important algorithms in signal processing, communications and artificial intelligence are special cases of message passing over factor graphs. These include belief propagation decoding of LDPC codes, the Viterbi algorithm, the Kalman filter and certain fast Fourier transform (FFT) algorithms.

3.2 Factor graphs

The coding schemes for QMF relaying presented in this chapter are derived using the factor graph formulation. In this section, a brief introduction of the model [36, 1] and its illustration for LDPC decoding is presented.

Factor graphs were developed based on the insight that many complex algorithms involve functions of multiple variables and attempt to compute *marginal* functions. Consider a collection of variables z_1, z_2, z_3, z_4 where each z_i ($i \in \{1, 2, 3, 4\}$) can assume values in a finite set S. Consider a real-valued function $g(z_1, z_2, z_3, z_4)$ of these variables. It is associated with 4 *marginal* functions $g_i(z_i)$ ($i \in \{1, 2, 3, 4\}$). A marginal $g_1(z_1)$ is computed by the summation of $g(z_1, z_2, z_3, z_4)$ over all possible values of z_2, z_3, z_4. The *not sum* or *summary* notation ($\sim$) was introduced in [36] to express these. Instead of indicating variables being summed over, this notation indicates variables not being summed over.

$$\sum_{\sim\{z_1\}} g(z_1, z_2, z_3, z_4) := \sum_{z_2 \in S} \sum_{z_3 \in S} \sum_{z_4 \in S} g(z_1, z_2, z_3, z_4)$$

$$g_1(z_1) := \sum_{\sim\{z_1\}} g(z_1, z_2, z_3, z_4)$$

Efficient algorithms to compute marginal functions can be developed by factorizing the *global* function $g(z_1, z_2, z_3, z_4)$ into a product of *local* functions, each of which depend only on a subset of the variables. Computational efficiency is achieved by using two key techniques:

- Exploit the factorization by using the distributive law to simplify summations.

- Reuse intermediate partial sums.

Assume that the function $g(z_1, z_2, z_3, z_4)$ can be factorized into four local functions as follows:

$$g(z_1, z_2, z_3, z_4) = f_A(z_1) f_B(z_1, z_2, z_3) f_C(z_3, z_4) f_D(z_4) \tag{3.1}$$

This can be represented as a bipartite graph called a factor graph. Each variable z_i is represented as a variable node and each function $f_j(\cdot)$ as a *function node*. An edge connects variable z_i to function $f_j(\cdot)$ if and only if z_i is an argument of $f_j(\cdot)$. The factorization (3.1) is shown graphically in Fig. 3.2.

3.2.1 Sum-Product Algorithm

The sum-product algorithm [36, 1] is a generalized message passing algorithm over a factor graph that efficiently computes all the marginal functions by appropriately exploiting the factorization and reusing partial sums.

Nodes exchange messages over the edges using extrinsic information. For any node ν, let $n(\nu)$ denote the set of its neighbors in the factor graph. Let $\mu_{x \to f}(x)$ represent the message
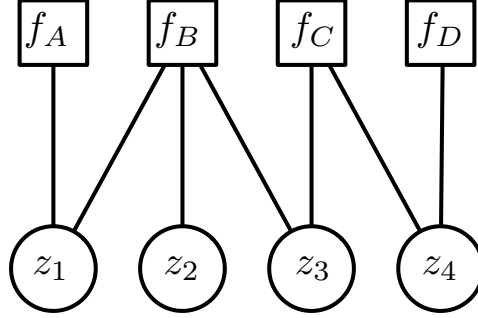
Figure 3.2: Factor graph for function factorization in Eq. (3.1).

sent from variable node x to function node f and $\mu_{f\rightarrow x}(x)$ the message in reverse direction. The generalized message update rules can be written as follows.

Variable to Function:

$$\mu_{x\rightarrow f}(x) = \prod_{h\in n(x)\backslash\{f\}} \mu_{h\rightarrow x}(x) \tag{3.2}$$

Function to Variable:

$$\mu_{f\rightarrow x}(x) = \sum_{\sim\{x\}} \left(f(X) \prod_{y\in n(f)\backslash\{x\}} \mu_{y\rightarrow f}(y) \right) \tag{3.3}$$

Here $X = n(f)$ is the set of arguments for function f. Message passing initiates at leaf nodes i.e. nodes with degree 1. A node must compute an outgoing message for each of its edges once incoming messages have arrived from all other edges. The algorithm terminates once messages have been exchanged in either direction on every edge. After termination, the product of all incoming messages give the respective marginal functions at variable nodes. Here *sum* and *product* refer to the point wise product or sum of functions.

As an example, let us consider the factor graph in Fig 3.2 and examine sum-product updates as defined in (3.2)-(3.3). The algorithm is initiated by nodes z_2, f_A and f_D since they have degree 1. As the algorithm progresses, all marginal functions are computed. For illustration, a sequence of message updates leading to the computation of $g_4(z_4)$ is shown below.

$$\mu_{f_A \to z_1} = f_A(z_1)$$

$$\mu_{z_2 \to f_B} = 1$$

$$\mu_{z_1 \to f_B} = f_A(z_1)$$

$$\mu_{f_B \to z_3} = \sum_{\sim\{z_3\}} f_B(z_1, z_2, z_3) f_A(z_1)$$

$$\mu_{z_3 \to f_C} = \sum_{\sim\{z_3\}} f_B(z_1, z_2, z_3) f_A(z_1)$$

$$\mu_{f_C \to z_4} = \sum_{\sim\{z_4\}} \left(f_C(z_3, z_4) \times \left(\sum_{\sim\{z_3\}} f_B(z_1, z_2, z_3) f_A(z_1) \right) \right)$$

$$\mu_{f_D \to z_4} = f_D(z_4)$$

$$g_4(z_4) = \mu_{f_D \to z_4} \times \mu_{f_C \to z_4}$$

$$= f_D(z_4) \times \left(\sum_{\sim\{z_4\}} \left(f_C(z_3, z_4) \times \left(\sum_{\sim\{z_3\}} f_B(z_1, z_2, z_3) f_A(z_1) \right) \right) \right)$$

$$= \sum_{\sim\{z_4\}} g(z_1, z_2, z_3, z_4)$$

Note that the sequence of computations exploits the distributive law and reuses all partial sums. Also note that degree 1 nodes simply send constant messages without performing any computation. Similarly, degree 2 variable nodes forward messages from one edge to another without computation. This is a consequence of the extrinsic information transfer property of the sum-product algorithm.

3.2.2   Graphs with cycles

The sum product algorithm described above can only be used with graphs that have no cycles. If the graph has cycles, the algorithm can get "stuck". There may be no nodes with degree 1, or with degree n that have received at least $n - 1$ incoming messages. For graphs with cycles, an iterative version of the algorithm has been defined. The iterative algorithm yields approximate solutions for marginal functions. Nonetheless, extensive simulation results show that for large graphs which do not have short cycles, the approximation yields sufficiently accurate results.

The iterative algorithm has no natural start or end conditions and multiple messages are exchanged over every edge. Initialization does not rely on degree 1 nodes. Instead, all the edges are initialized with the identity function. A node is said to have a message *pending* at an edge e, if it has received new messages on any edge other than e since the

previous outgoing message on e. Under the iterative algorithm, only pending messages are transmitted because only pending messages can be different from previously transmitted ones. Different message passing *schedules* are possible for iterations, each having different convergence characteristics.

3.2.3   LDPC Decoding

The above formulation applies naturally to error control coding. As per the coding model described in Sec. 3.1, decoding involves searching for codeword $\mathbf{b} \in \mathsf{C}$ that maximizes the *a posteriori* probability (APP) i.e. maximum *a posteriori* (MAP) decoding rule. The APP is given by $P[\mathbf{b}|\mathbf{y}]$, where $\mathbf{y} \in \mathbb{C}^n$ are observations from the channel. The above search can be made efficient by computing *bitwise* APPs:

$$P[b_i|\mathbf{y}] = \sum_{\sim\{b_i\}} P[\mathbf{b}|\mathbf{y}]$$

for $i \in \{1, 2, 3, \ldots, n\}$. These are nothing but marginals of the APP function. A factor graph can be constructed by factorizing the APP function as follows:

$$P[\mathbf{b}|\mathbf{y}] = \frac{f_{\mathbf{y}|\mathbf{b}}(\mathbf{y}|\mathbf{b})P(\mathbf{b})}{f_{\mathbf{y}}(\mathbf{y})}$$

Here $f_{\mathbf{y}}(\mathbf{y})$ represents the probability density function (PDF) of the observations. Since $\mathbf{y}$ remains constant for a particular decoding instance, the APP can be considered a function only of the variables $\mathbf{b} = [b_1 \; b_2 \; \ldots \; b_n]$.

$$P[\mathbf{b}|\mathbf{y}] \propto f_{\mathbf{y}}(\mathbf{y}|\mathbf{b})P(\mathbf{b})$$

If the channel is memoryless, $f_{\mathbf{y}|\mathbf{b}}(\mathbf{y}|\mathbf{b})$ can be further factorized into symbol-wise functions. If the *a priori* distribution of $\mathbf{b}$ is uniform over all codewords, then $P[\mathbf{b}] = \mathbf{1}_{\mathsf{C}}(\mathbf{b})/|\mathsf{C}|$ where $\mathbf{1}_{\mathsf{C}}(\mathbf{b})$ represents the indicator function that $\mathbf{b} \in \mathsf{C}$.

$$f_{\mathbf{y}}(\mathbf{y}|\mathbf{b})P(\mathbf{b}) = \mathbf{1}_{\mathsf{C}}(\mathbf{b})\prod_{j=1}^{n} f_{y_j}(y_j|b_j)$$

$$b_0 \oplus b_2 \oplus b_4 = 0 \quad b_1 \oplus b_3 \oplus b_5 = 0 \quad b_0 \oplus b_1 \oplus b_3 = 0 \quad b_2 \oplus b_4 \oplus b_5 = 0$$
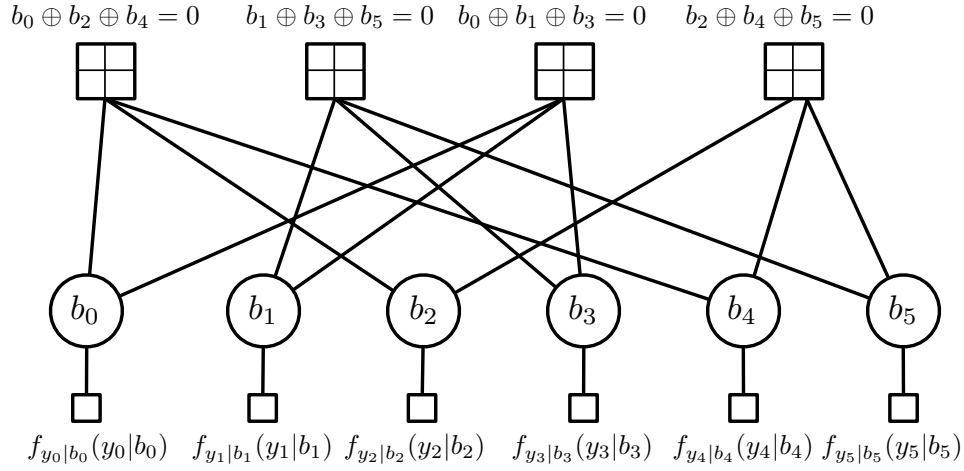
Figure 3.3: Factor graph for LDPC code in Fig. 3.1.

The indicator function for the code can be further factorized into local functions i.e. parity constraints that must be satisfied in order for $\mathbf{b} \in \mathsf{C}$. As an example, consider the example code shown in Fig. 3.1. The factorization results in four local functions.

$$
\begin{aligned}
\mathbf{1}_{\mathsf{C}}(\mathbf{b}) = & \mathbf{1}[b_0 \oplus b_2 \oplus b_4 = 0]\mathbf{1}[b_1 \oplus b_3 \oplus b_5 = 0] \\
& \mathbf{1}[b_0 \oplus b_1 \oplus b_3 = 0]\mathbf{1}[b_1 \oplus b_3 \oplus b_5 = 0]
\end{aligned}
$$

The above factorization is expressed as a factor graph in Fig 3.3. It is similar to the Tanner graph in Fig. 3.1, except for the presence of dangling *observation* nodes from every variable. These represent prior beliefs for each of the bits. As discussed in Sec 3.2.1, such degree 1 nodes simply forward the prior beliefs and do not play an active role in the decoding algorithm. Note that this factor graph has cycles and therefore requires the use of iterative sum product decoding. All the variables in this graph are binary i.e. over Galois field GF(2) and all function nodes (with degree > 1) have identical form (parity check constraints). Therefore, specialized sum-product update rules can be derived from the general expressions given in (3.2) and (3.3). These are discussed later in Sec. 3.3.8.

3.3 Coding for cooperative channels

In this section, the factor graph formulation is used to design efficient coding schemes for relay cooperation. As discussed previously, cooperation using multi-antenna terminals in the high SNR regime is of special interest for practical system design. A cooperative link typically consists of an information source, a destination and one or more cooperating *half*

duplex relays. Coding schemes for such links are closely tied to the relaying scheme used at relays.

Various strategies for relaying have been proposed in literature. Among these amplify-and-forward (AF), decode-and-forward (DF) and compress-and-forward (CF) [16, 37] have received the most attention. Under DF, relays decode the source's message and forward a hard estimate of it, whereas under AF and CF they forward soft estimates without explicitly decoding it. Furthermore, DF and CF relays forward coded information, whereas AF relays forward it in uncoded form. For the simple one relay configuration with additive white Gaussian noise, it is known that CF performs better than DF and AF in the high SNR regime [75]. In fact, CF can perform within one bit/sec/Hz to the information theoretic capacity [53, 6] for single-antenna terminals.

In the conventional CF scheme, relays compress their observed signal and encode it before forwarding. The encoding rate is chosen such that the destination can use *successive decoding* i.e. first decode compressed signals from relays and then use this as side information to decode the message from source. Hence, CF requires that every relay know the quality of its forward channel. For complex networks, this estimation overhead becomes large and rate adaptation schemes become increasingly difficult. Moreover, for configurations that involve multiple relays CF is no longer within bounded gap from the information theoretic capacity.

Recent work [6] has shown that QMF achieves performance within bounded gap of capacity for Gaussian networks with arbitrary number of relays. In addition to having close-to-optimal performance QMF offers several advantages that make it suitable for practical system design. In the QMF scheme, a relay quantizes its received signal at noise level, maps it to a codeword and forwards it. QMF does not require forward channel knowledge at relays minimizing the channel estimation overhead and simplifying rate adaptation protocols. Not having to perform any decoding reduces processing at relays and allows freedom to independently optimize the listening-time fraction. QMF has played a key role in several recent information-theoretic results on cooperative networks [40, 53, 46, 72]. The focus of this dissertation is on various system components for QMF relaying.

QMF presents a unique challenge in the design of practical coding schemes. Since the mapping at relays is performed without any knowledge of forward channel strength, the side information from relays cannot be decoded at the destination independently. Therefore, QMF uses *joint decoding* of the message (from source) and side information (from relays). Joint decoding usually requires higher complexity and makes it harder to design practical codes. The key result developed in this chapter is a low complexity coding framework for QMF that performs close to information theoretic bounds.

3.3.1   Related Work

Channel coding schemes for cooperative communication were first proposed around 2003 by Hunter et.al, Janani et.al. and Zhao et.al. [79, 26, 31]. Majority of previous work has focused on the DF scheme. DF coding schemes involve computing and forwarding additional parity

constraints for the source's codeword. Turbo code designs with full-duplex and half-duplex DF relays were shown in [78] and [77] respectively, where the proposed schemes performed $\approx$ 1dB away from the DF information theoretic threshold. LDPC profiles were first developed for DF in [13]. A bilayer LDPC structure [57] and the protograph method [49] were used to get LDPC designs $\leq$ 0.5dB from DF threshold. The bilayer structure is extended for use at high SNR using BICM in [56]. Since DF does not perform within bounded gap to capacity, schemes like CF and QMF always out-perform DF at high SNR.

A coding scheme for CF relaying using a combination of LDPC and irregular repeat accumulate (IRA) codes is presented in [70]. Rateless coding schemes for both CF and DF are developed in [69].

In independent work [50], a coding technique for QMF has been proposed based on lattice strategies. The scheme in [50] reduces the complexity of mapping at the relay to polynomial-time while the joint decoding complexity remains exponential-time.

3.3.2 Key Ideas

In this work, a coding framework for QMF relaying with linear complexity encoding at the source, mapping at the relays and joint decoding at the destination is developed. A low complexity factor graph structure and decoding algorithm are proposed. The framework is validated using channel simulation experiments. In Chapter 4, the framework is further enriched by the development of codes specialized for cooperative relaying. The key techniques used in the framework are summarized as follows:

1. *BICM: Binary* channel codes with standard higher order signal constellations are considered, based on the widely used BICM technique [12].

2. *LDPC-LDPC and LDPC-LDGM:* The proposed scheme uses LDPC codes at the source for channel coding and either a LDGM, or LPDC code at the relay for mapping.

3. *Joint Factor Graph:* The joint decoding procedure at the destination is formulated as a belief propagation algorithm over a factor graph. This graph contains the original channel code (LDPC) and relay mapping functions (LDPC/LDGM) as subgraphs connected via probabilistic constraints that model the QMF relay operations.

4. *Practical Decoding Algorithm:* Using a suitable space-time architecture, scalar quantization procedure at relays and specific choice of component codes, the resulting factor graph can be greatly simplified making it suitable for practical decoder implementation.

5. *Code Design:* Density evolution analysis tools [58, 59] are used for code design over joint LDPC-LDGM graphs. This is covered in Chapter 4.

3.3.3 System Model

Initially, the *binary memoryless symmetric* (BMS) relay channel is considered as described below. In Sec 3.4, this model is extended to the high SNR regime.

The BMS Gaussian relay channel has three half-duplex terminals: source (S), relay (R) and destination (D) with binary input additive white Gaussian (BIAWGN) channels between them (Fig. 3.4). The model is similar to the one described in Sec. 2.3.1, the only exception is that channel inputs are binary symbols. The focus of the current discussion is limited to cooperative multiplexing relay channel antenna configurations having form $(1, 1, M)$. The transmit codewords are $\mathbf{b}_S \in \{0, 1\}^{N_S}$ and $\mathbf{b}_R \in \{0, 1\}^{N_R}$. The corresponding transmit signals are $\mathbf{x}_S \in \{\pm\sqrt{P_S}\}^{N_S}$ and $\mathbf{x}_R \in \{\pm\sqrt{P_R}\}^{N_R}$. P_S, P_R and SNR parameters $\mathsf{SNR}_{SD}, \mathsf{SNR}_{SR}, \mathsf{SNR}_{RD}$ are as defined in Sec. 2.3.1.
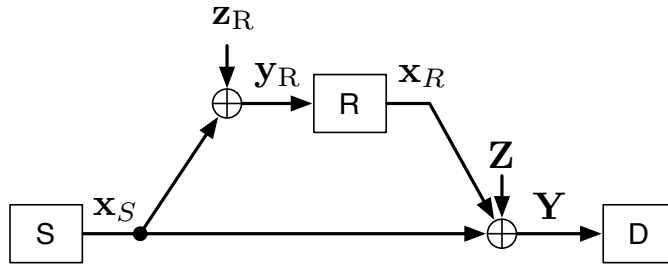
Figure 3.4: Half-Duplex Binary Input Gaussian Relay Channel

For the above model, quantize-map-and-forward relaying [6] is summarized as follows: S transmits a sequence of messages $m_k \in \{1, \ldots, 2^{N_S \mathcal{R}}\}$ ($k = 1, 2, \ldots$). At both S and R, binary linear block codes C_S^b and C_R^b are created. S maps each message to one of its codewords and transmits it using N_S symbol times. This results in an overall transmission rate of $\mathcal{R}$. Relay listens to the first $f N_S$ time symbols of each block. It quantizes its observation and maps it to a codeword in C_R^b having length $(1 - f)N_S$. In order to construct an efficient practical scheme, the mapping at relay C_R^b is assumed to be linear and can be represented by a $f N_S \times (1 - f)N_S$ matrix $\mathbf{G}_R$ with entries in GF(2). The mapping C_R^b can be thought of as a linear block code having generator matrix $\mathbf{G}_R$. The linear mapping has a corresponding parity check matrix $\mathbf{H}_R$. Given the knowledge of all encoding functions and signals received ($\mathbf{Y}$), destination D attempts to decode the message sent by S.

3.3.4 QMF Joint Decoding Factor Graph

The proposed scheme is based entirely on the use of *sparse graph* codes. It is known that encoding operations for such codes have linear complexity [41, 60]. Therefore, it is sufficient to focus on the *joint decoding* operation at destination which is central to QMF relaying. As discussed previously, linear binary block codes C_S^b and C_R^b can be represented as bipartite Tanner graphs. Our scheme expresses the joint decoding problem as a factor graph that has

Tanner graphs of component codes $(\mathsf{C}_S^b, \mathsf{C}_R^b)$ as its sub-graphs. The sub-graphs are connected via probabilistic constraints. These constraints represent the QMF relaying operation.

Joint decoding involves searching for the codeword $\mathbf{b}_S \in \mathsf{C}_S^b$ that maximizes the APP $P(\mathbf{b}_S|\mathbf{Y})$. Here $\mathbf{Y}$ represents all the channel observations at D. Consider the bitwise MAP decoder, where the bitwise APPs are computed.

$$P(b_{S,i}|\mathbf{Y}) = \sum_{\sim b_{S,i}} P(\mathbf{b}_S|\mathbf{Y}) \quad i \in \{1, 2, \dots N_S\}$$

This is achieved by marginalizing the APP using the following factorization:

$$P(\mathbf{b}_S|\mathbf{Y}) = \sum_{\mathbf{b}_R} \frac{f(\mathbf{Y}|\mathbf{b}_S, \mathbf{b}_R) P(\mathbf{b}_S, \mathbf{b}_R)}{f(\mathbf{Y})}$$

$$\propto \sum_{\mathbf{b}_R} f(\mathbf{Y}|\mathbf{b}_S, \mathbf{b}_R) P(\mathbf{b}_S, \mathbf{b}_R).$$

For the first fN_S bits, R is listening and D observes an interference-free signal from S. During the remaining transmissions, D observes a superposition of signals from S and R. As a result, the term $f(\mathbf{Y}|\mathbf{b}_S, \mathbf{b}_R)$ can be factorized as follows:

$$f(\mathbf{Y}|\mathbf{b}_S, \mathbf{b}_R) = \prod_{i=1}^{fN_S} f(\mathbf{y}_i|b_{S,i}) \prod_{j=1}^{N_R} f(\mathbf{y}_{(fN_S+j)}|b_{S,(fN_S+j)}, b_{R,j})$$

The codes C_S^b and C_R^b have characteristic functions $\mathbf{1}(\mathbf{b}_S \in \mathsf{C}_S^b)$ and $\mathbf{1}(\mathbf{b}_R \in \mathsf{C}_R^b)$ respectively.

$$P(\mathbf{b}_S, \mathbf{b}_R) = P(\mathbf{b}_S) P(\mathbf{b}_R|\mathbf{b}_S)$$

$$\propto \mathbf{1}(\mathbf{b}_S \in \mathsf{C}_S^b) P(\mathbf{b}_R|\mathbf{b}_S)$$

$$\stackrel{(a)}{=} \mathbf{1}[\mathbf{b}_S \in \mathsf{C}_S^b]\mathbf{1}[\mathbf{b}_R \in \mathsf{C}_R^b]P(\mathbf{b}_R|\mathbf{b}_S).$$

(a) is due to the fact that $\mathbf{b}_R$ must be a codeword in C_R^b. As discussed previously, if C_S^b and C_R^b are sparse, then their indicator functions can be factorized into local parity checks.

The resulting factor graph is shown in Fig. 3.5. In addition to nodes representing channel observations i.e. $f(\mathbf{y}_i|b_{S,i}, b_{R,j})$, the subgraphs $\mathbf{1}(\mathbf{b}_S \in \mathsf{C}_S^b)$ and $\mathbf{1}(\mathbf{b}_R \in \mathsf{C}_R^b)$ are connected by $P(\mathbf{b}_R|\mathbf{b}_S)$. These connections represent the quantization operation at R.

The complexity of a sum-product algorithm over a factor graph is proportional to the number of edges in the graph. If the component codes C_S^b and C_R^b are sparse, the overall factor graph is also sparse. This implies that decoding complexity is linear in the length of component codes.

Note that the sum-product update rules at the function node $P(\mathbf{b}_R|\mathbf{b}_S)$ may get very complex due to its high degree $(N_S + N_R)$. Moreover, nodes like this introduce very short
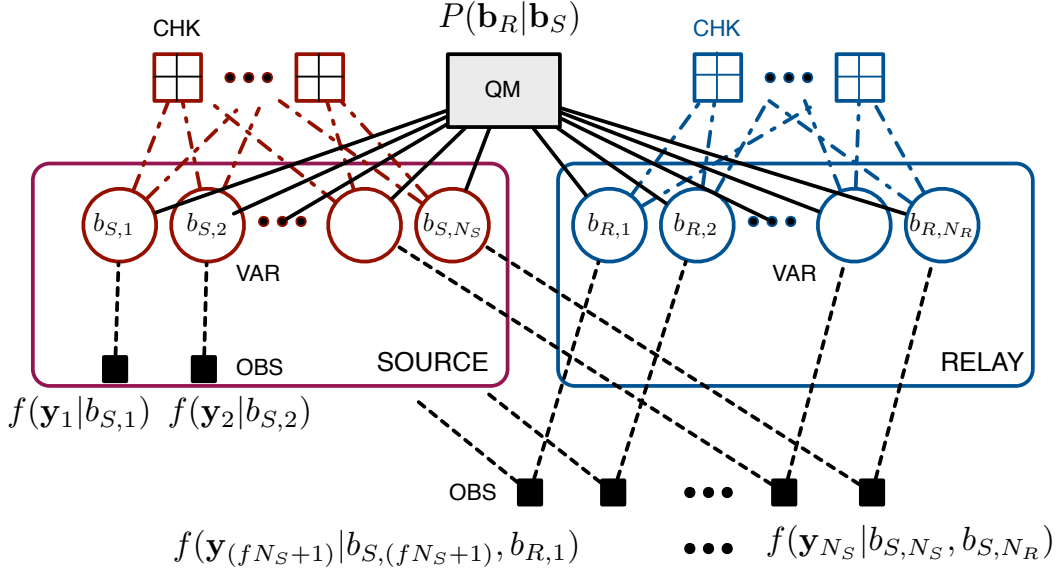
Figure 3.5: Joint Decoding Factor Graph.

cycles in the graph. In this case, sum-product decoding does not approach the performance of MAP decoding. In order to approach MAP performance and low decoding complexity, the $P\left(\mathbf{b}_R|\mathbf{b}_S\right)$ nodes need further factorization. In the following subsections, specific techniques are discussed for the same.

3.3.5   LDPC-LDGM Configuration

The factor graph shown in Fig. 3.5 corresponds to an LDPC-LDPC configuration i.e. source and relay both use LDPC codes. LDGM codes are another class of sparse graph codes that are commonly used for lossy data compression [44]. As their name suggests, they utilize a sparse generator matrix, which can be represented as a Tanner graph. Note that the parity check matrix of an LDGM code is not guaranteed to be sparse. In an LDGM Tanner graph, variable nodes represent n input bits, check nodes represent k parities that form the output. For an LDGM code, if $k \leq n$, then it performs data compression, else $(k > n)$ it introduces redundancy in the output. Estimates for the input bits can be recovered from observations of the output by using sum-product decoding over the corresponding Tanner graph.

Two alternative factor graph configurations are studied in this chapter. The first one uses LDPC codes at both S and R as discussed previously. Henceforth, it is referred to as the LDPC-LDPC configuration. The second choice is to use an LDPC code at S and an LDGM code at R. This is referred to as the LDPC-LDGM configuration. In order to construct LDPC-LDGM factor graphs, auxiliary variable nodes $\mathbf{b}_Q = \{b_{Q,i}\}_{i=1}^{K_R}$ are introduced. $\mathbf{b}_Q$ represents the K_R bits after quantization at R. $\mathbf{b}_Q$ is mapped to the codeword $\mathbf{b_R}$ of length

N_R, after GF(2) multiplication with low density generator matrix $\mathbf{G}_R$. The LDGM code has rate N_R/K_R and characteristic function $\mathbf{1}[\mathbf{b}_R \in \mathsf{C}_R^b]$. Since $\mathbf{b}_R$ is a deterministic function of $\mathbf{b}_Q$, $P(\mathbf{b}_R|\mathbf{b}_S)$ is factorized as follows (Fig 3.6):

$$\begin{aligned}
P(\mathbf{b}_R|\mathbf{b}_S) &= P(\mathbf{b}_R, \mathbf{b}_Q|\mathbf{b}_S) \\
&= P(\mathbf{b}_R|\mathbf{b}_Q, \mathbf{b}_S)P(\mathbf{b}_Q|\mathbf{b}_S) \\
&= \mathbf{1}[\mathbf{b}_R \in \mathsf{C}_R^b]P(\mathbf{b}_Q|\mathbf{b}_S)
\end{aligned}$$

The $\mathbf{b_R}$ variable nodes always have degree 2 and simply perform forwarding of messages under the sum product algorithm.

3.3.6   Scalar Quantizer

It is possible to perform vector quantization at R. However, it is shown [6] that QMF performs within bounded gap of capacity even with a scalar quantizer. Under scalar quantization, observations at R are independently quantized in a symbol-by-symbol fashion. For the LDPC-LDGM configuration, if each $y_{R,i}$ is quantized into $b_Q[A_i]$ ($i \in \{1, 2, \ldots fN_S\}$), then the $P(\mathbf{b}_Q|\mathbf{b}_S)$ function node factorizes into fN_S separate nodes. Each of them represent a scalar quantization operation.

$$P(\mathbf{b}_Q|\mathbf{b}_S) = \prod_{i=1}^{fN_S} P(b_Q[A_i]|b_{S,i}), \text{ where}$$

$$\bigcup_{i=1}^{fN_S} A_i = \{1, 2, \ldots, K_R\}, \ A_i \cap A_j = \emptyset \ \forall i \neq j$$

Here A_i denotes the subset of indices in $\mathbf{b}_Q$ that symbol $y_{R,i}$ is quantized into.

A similar factorization is possible for the LDPC-LDPC configuration as well [47]. The factorized nodes represent $P(b_Q[A_i]|b_{S,i})$ relations for LDPC-LDGM and $P(b_R[A_i]|b_{S,i})$ relations for the LDPC-LDPC configurations respectively. Henceforth, these nodes are called *quantize* (Q) nodes, because they represent quantization at the relay. Example factor graphs showing the LDPC-LDGM and LDPC-LDPC constructions are illustrated in Fig 3.6 and Fig. 3.7. In the examples, each symbol observation $y_{R,i}$ is quantized into one bit (i.e. $K_R = fN_S$ and $A[i] = \{i\}$) . As shown, there are four kinds of nodes in the resulting factor graphs: Observation (OBS) nodes, Variable (VAR) nodes, Check (CHK) nodes, and Quantize (Q) nodes. Some VAR nodes in the C_S^b subgraph share OBS nodes with VAR nodes in the C_R^b subgraph. This is a consequence of multiple access at D i.e. both S and R transmit simultaneously for $(1-f)N_S$ symbols and their observations interfere at D.
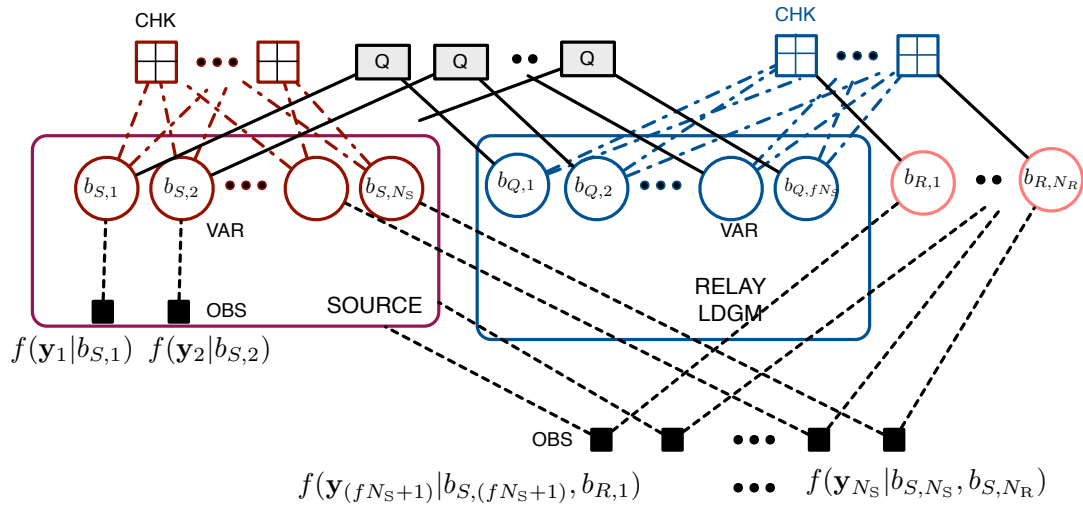
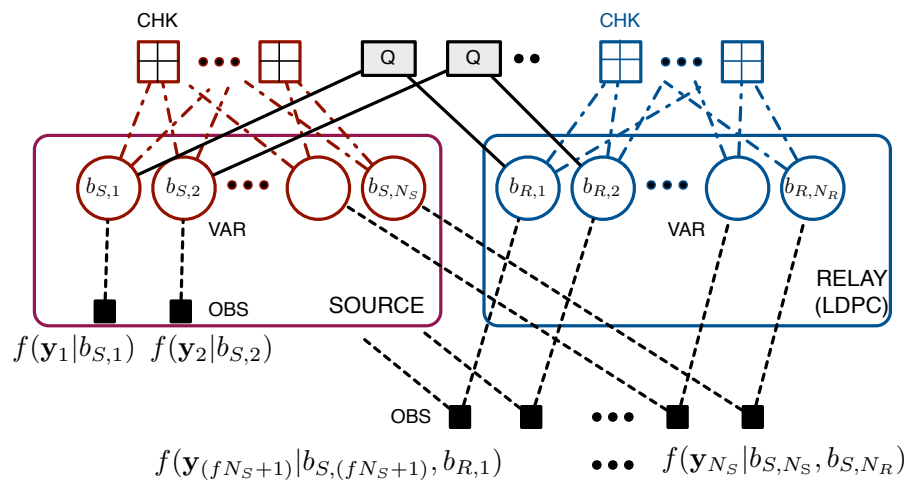Figure 3.6: LDPC-LDGM factor graph with 1 bit scalar quantizer.

Figure 3.7: LDPC-LDPC factor graph with 1 bit scalar quantizer.
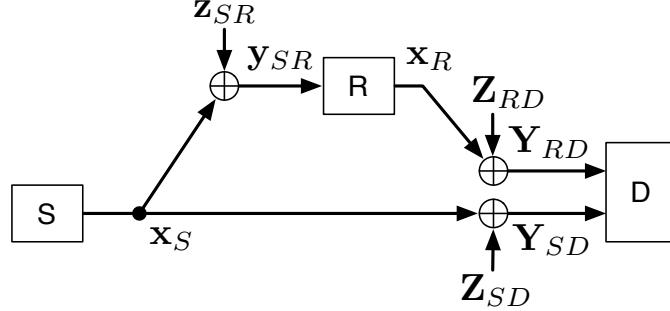
3.3.7 Space Time Architecture

Figure 3.8: Channel model using a space time coding scheme like DBLAST.

The factor graphs shown in Figs. 3.6-3.7 are further simplified by the use of a suitable space-time architecture. Several space-time coding schemes for MIMO channels are discussed previously in Sec. 2.1.2. Similar schemes are applicable to the relay channel. Repetition and Alamouti schemes (2.2)-(2.4) require that both transmit antennas have knowledge of all codeword bits. This is not achievable with QMF relaying, because R only has noisy estimates of the first fN_S codeword bits.

The V-BLAST architecture (2.6) requires both antennas to use independent codewords. S and R transmissions belong to independent codebooks but represent the same underlying message. This leads to correlation between $\mathbf{x}_S$ and $\mathbf{x}_R$. In spite of this, an architecture *similar* to V-BLAST can be used for QMF relaying, if correlation between codewords is appropriately handled at the receiver D.

Correlation between $\mathbf{x}_S$ and $\mathbf{x}_R$ is a form of coding across transmit antennas. The D-BLAST space time architecture is natural fit for this scenario. D-BLAST for the relay channel is also proposed in [77, 78, 22, 34]. In a MIMO channel, coding across antennas is *deterministic* i.e. introduced by the inner code. In the relay channel, this correlation is probabilistic because $\mathbf{x}_R$ is derived from a noisy observation of $\mathbf{x}_S$.

The MMSE-SIC decoding scheme discussed in Sec 2.5.3 can be used over the relay channel. It involves introducing a delay of one block at the relay. At the k-th block, the destination receives the superposition of the following:

- signal from the source containing message sent at block p, namely, $\mathbf{x}_S(m_p)$

- signal from the relay, containing side information about source's message at block $p-1$ i.e. $\mathbf{x}_R(m_{k-1})$

Messages from the source are independent across blocks. At the p-th block, the destination jointly decodes block $p - 1$, i.e. $(\mathbf{x}_S(m_{p-1}))$ and side information $(\mathbf{x}_R(m_{p-1}))$ by treating $\mathbf{x}_S(m_p)$ as Gaussian noise. The receiver subtracts relay's codeword $\mathbf{x}_R(m_{p-1})$ from its received signal $\mathbf{Y}[p]$ and retains residual $\widetilde{\mathbf{Y}}[p]$ for decoding the next block. This architecture

permits the use of a simplified equivalent channel model shown in Fig. 3.8. The destination observes two orthogonal sets of observations. This leads to factorization of the degree 2 OBS nodes (representing multiple access). The factorization is shown in equation (3.4) where $\underline{\mathbf{Y}} := [\mathbf{Y}_{SD} \ \mathbf{Y}_{RD}]$.

$$
\begin{aligned}
f(\underline{\mathbf{Y}}|\mathbf{b}_S, \mathbf{b}_Q) &= \prod_{i=1}^{fN_S} f(\mathbf{y}_{SD,i}|b_{S,i}) \prod_{j=1}^{N_R} f(\mathbf{y}_{SD,(fN_S+j)}, \mathbf{y}_{RD,(fN_S+j)}|b_{Q,j}, b_{S,(fN_S+j)}) \\
&= \prod_{i=1}^{fN_S} f(\mathbf{y}_{SD,i}|b_{S,i}) \prod_{j=1}^{N_R} f(\mathbf{y}_{SD,(fN_S+j)}|b_{S,(fN_S+j)}) f(\mathbf{y}_{RD,(fN_S+j)}|b_{Q,j}) \\
&= \prod_{i=1}^{N_S} f(\mathbf{y}_{SD,i}|b_{S,i}) \prod_{j=1}^{N_R} f(\mathbf{y}_{RD,(fN_S+j)}|b_{Q,j}),
\end{aligned}
\tag{3.4}
$$

Examples of simplified LDPC-LDPC and LDPC-LDGM factor graphs are shown in Figs. 3.9-3.10. Note that VAR nodes of the S and R Tanner graphs are connected *only* through Q nodes. Since $\mathbf{y}_{RD,i} = \mathbf{0}$ for $i \in \{1, \ldots, fN_S\}$, we rename $\mathbf{y}_{RD,(fN_S+j)} \equiv \mathbf{y}_{RD,j}$, for $j = 1, \ldots, N_R$.

The structure of resulting graphs is very similar to Tanner graphs with the exception of Q constraints.

Remark 2. Consider the original channel model (without D-BLAST) in Sec. 3.3.3 and the DBLAST-equivalent channel. Note that the capacities of these two channels are within two bits of each other. This conclusion is based on the following observations:

1. *The cut-set capacity upper bound for both channels is within a one bit gap (for any listening fraction $f \in [0, 1]$).*
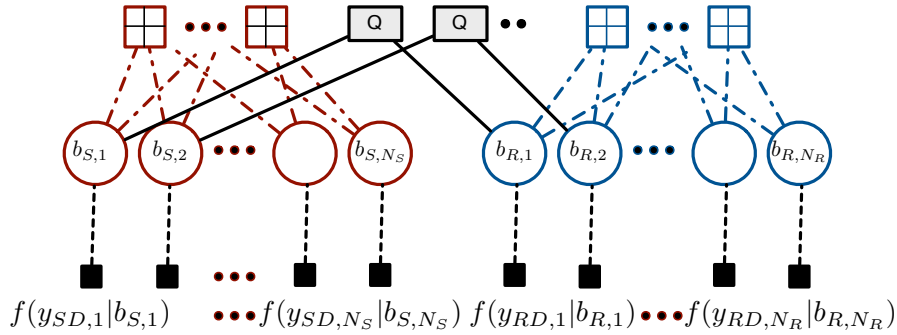
Figure 3.9: Simplified LDPC-LDPC factor graph with one-bit scalar quantizer and DBLAST.
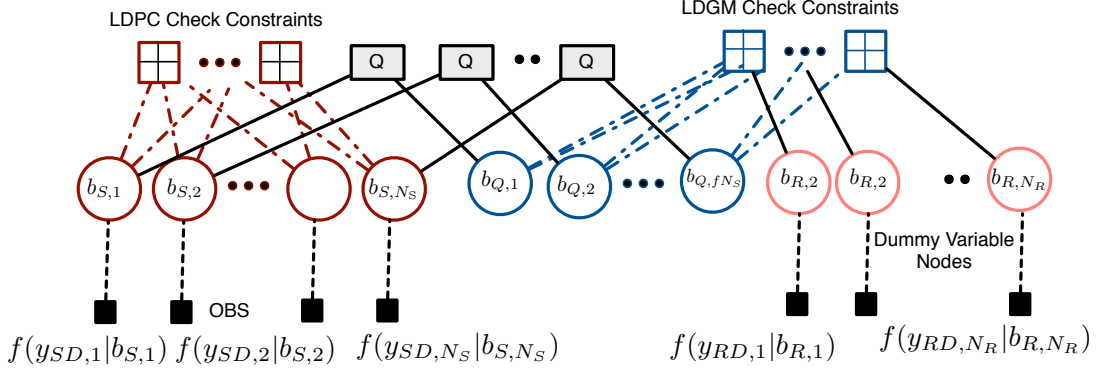
Figure 3.10: Simplified LDPC-LDGM factor graph with one-bit scalar quantizer and DBLAST.

The mutual information across cut $\{S\}, \{R, D\}$ is the same for the two channels. Consider the mutual information across the cut $\{S, R\}, \{D\}$. It is known that SIC achieves the sum capacity of multiple-access channels. In the original channel (Fig. 3.4), S and R are allowed unlimited cooperation. This is not the case in the D-BLAST equivalent channel. As a result, the cut-set bound for DBLAST incurs a power-gain loss of at most $(1 - f)$ bits.

2. *QMF relaying scheme achieves the cut-set upper bound to within one bit for both channels.*

   Constant gap to cut-set bound is shown in the original work on QMF [6]. It is also shown that for a single relay network the gap is within one bit [6].

3.3.8   Decoding Algorithm

Sum product decoding computes the MAP estimate precisely only if the factor graph has no cycles. Otherwise, it computes the approximate APP for each bit. For the factor graph presented above, messages passed on the edges and update rules at the VAR/CHK nodes remain unchanged from the case of LDPC decoding. The only new ingredient in the mix are Q nodes introduced by the joint factor graph framework.

   For completeness, the messages and update rules at all kinds of nodes are described in this section. Let the subscripts V, K, and Q denote VAR nodes, CHK nodes, and Q nodes respectively. For $F \in \{K, Q\}$, let $\omega_{VF}^{(l)}$ denote the message sent from variable node V to function node F in the lth iteration. Every edge in the graph is connected to exactly one variable node and a message on the edge represents the APP for the respective variable. Messages can be represented as *log likelihood ratios* (LLR), but for the sake of simplicity,

consider messages represented by a two-dimensional vector[1].

$$\omega_{\mathsf{VF}} := [p_0 \ p_1]$$

Here $\omega_{\mathsf{VF}}(1) = p_0 \in [0, 1]$ represents the probability that bit V is 0 and $\omega_{\mathsf{VF}}(2) = p_1 \in [0, 1]$ represents the probability that bit is 1 ($p_0 + p_1 = 1$).

As per (3.2), the message sent from V to node $\mathsf{F} \in \{\mathsf{K}, \mathsf{Q}\}$ is the normalized product of all incoming messages into node V except the message from F. Normalization ensures that $p_0 + p_1 = 1$ for the outgoing message.

Variable nodes having degree 2 do not play an active role in computing updates, they merely forward messages from one edge to another. For a node V having degree 3, the message sent on any edge is a function of the incoming messages on the 2 other edges. Consider incoming messages, $\omega_{\mathsf{F}_1\mathsf{V}} = [p_0 \ p_1]$ and $\omega_{\mathsf{F}_2\mathsf{V}} = [q_0 \ q_1]$, the update rule is written as a function $\mathtt{VAR3}(\omega_{\mathsf{F}_1\mathsf{V}}, \omega_{\mathsf{F}_2\mathsf{V}})$. This is computed from (3.2) as follows.

$$\omega_{\mathsf{VF}_3} = \mathtt{VAR3}(\omega_{\mathsf{F}_1\mathsf{V}}, \omega_{\mathsf{F}_2\mathsf{V}}) = \left[\frac{p_0 q_0}{p_0 q_0 + p_1 q_1} \ \frac{p_1 q_1}{p_0 q_0 + p_1 q_1} \right] \tag{3.5}$$

Updates for variable nodes having degree > 3 are computed recursively from (3.5).

$$\omega_{\mathsf{VF}_4} = \mathtt{VAR4}(\omega_{\mathsf{F}_1\mathsf{V}}, \omega_{\mathsf{F}_2\mathsf{V}}, \omega_{\mathsf{F}_3\mathsf{V}}) = \mathtt{VAR3}(\omega_{\mathsf{F}_1\mathsf{V}}, \mathtt{VAR3}(\omega_{\mathsf{F}_2\mathsf{V}}, \omega_{\mathsf{F}_3\mathsf{V}}))$$

The message sent from check node K to node V' is the indicator function that the check is satisfied, marginalized on the bit represented by V'. A degree 2 CHK node represents the function $\mathbf{1}[b_1 \oplus b_2 = 0]$ which is equal to $\mathbf{1}[b_1 = b_2]$. The update rule for degree 2 CHK nodes therefore is given as

$$\omega_{\mathsf{KV}_2} = \mathtt{CHK2}(\omega_{\mathsf{V}_1\mathsf{K}}) = \omega_{\mathsf{V}_1\mathsf{K}}$$

Update rule for a degree 3 check node is computed from (3.3) in terms of incoming messages $\omega_{\mathsf{V}_1\mathsf{K}} = [p_0 \ p_1]$ and $\omega_{\mathsf{V}_2\mathsf{K}} = [q_0 \ q_1]$.

$$\omega_{\mathsf{KV}_3} = \mathtt{CHK3}(\omega_{\mathsf{V}_1\mathsf{K}}, \omega_{\mathsf{V}_2\mathsf{K}}) = [p_0 q_0 + p_1 q_1 \ p_0 q_1 + p_1 q_0] \tag{3.6}$$

Updates for check nodes of higher degree are computed recursively from (3.6).

At node Q, the message sent from Q to V is the marginalization of the function $p\left(b_Q[A_i] | b_{S,i}\right)$ on the bit represented by V. b_Q is computed from a noisy observation of $b_{S,i}$. Since the quantization is scalar, $\forall \mathbf{u} \in \{0, 1\}^{|A_i|}$ and $v \in \{0, 1\}$,

$$g(\mathbf{u}, v) := p\left(b_R[A_i] = \mathbf{u} | b_{S,i} = v\right)$$

The above function is fully represented by a lookup table with $2^{|A_i|+1}$ values. As an example,

---

[1] Later in Chap. 4, ω is replaced by w, the commonly used message $\log \frac{p_0}{p_1}$ (LLR) in belief propagation.
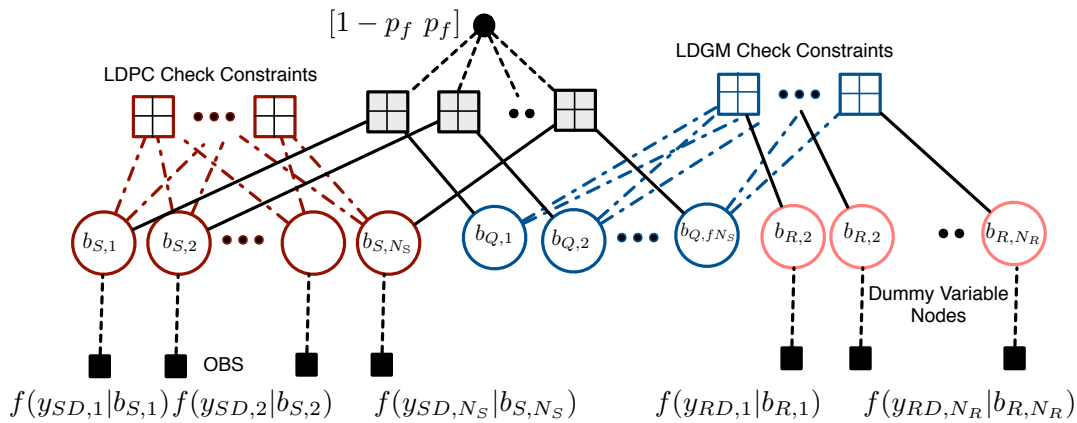
Figure 3.11: Equivalent LDPC-LDGM factor graph of that in Fig. 3.10. The Q constraints are decomposed into check constraints and a dummy variable node.
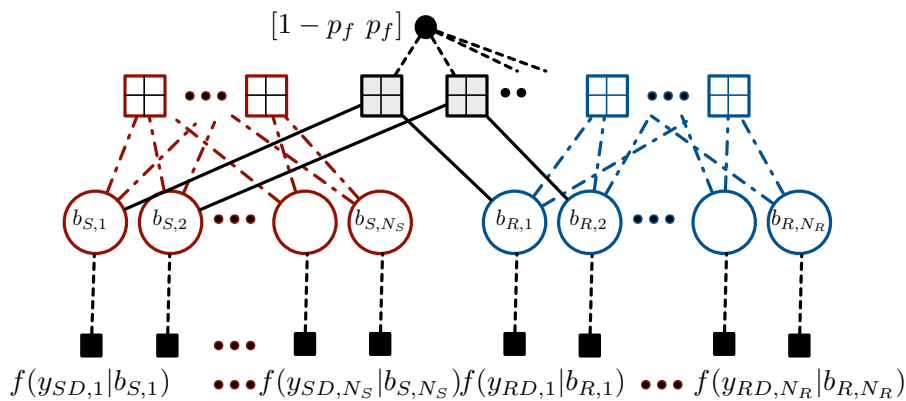

Figure 3.12: Equivalent LDPC-LDPC factor graph of that in Fig. 3.9. The Q constraints are decomposed into check constraints and a dummy variable node.

consider a one-bit scalar quantizer at the relay. Note that $A_i = \{i\}$, $i = 1, 2, \ldots, fN_S$, and $K_R = fN_S$. $\forall u \in \{0, 1\}$ and $v \in \{0, 1\}$,

$$\begin{aligned}
g(u, v) :=&p\left(b_{R,i} = u | b_{S,i} = v\right) \\
=&(1 - p_f)\mathbf{1}\{u = v\} + p_f\mathbf{1}\{u \neq v\}
\end{aligned}$$

where $p_f := \frac{1}{2}\operatorname{erfc}\sqrt{\frac{\mathsf{SNR}_{SR}}{2}}$ denotes the probability of bit error for scalar one-bit quantization over a BIAWGN channel. Since the function g is symmetric in u and v, assume that the VAR node belongs to the source and the marginalization is on v. Let the other VAR node be V'.

$$\begin{aligned}
\omega_{\mathsf{QV}}(1) = (1 - p_f)\omega_{\mathsf{V'Q}}(1) + p_f\omega_{\mathsf{V'Q}}(2) \\
\omega_{\mathsf{QV}}(2) = (1 - p_f)\omega_{\mathsf{V'Q}}(2) + p_f\omega_{\mathsf{V'Q}}(1),
\end{aligned}$$

The update has the same form as a degree 3 CHK node (3.6) with incoming messages $\omega_{\mathsf{V'Q}}$ and $[1 - p_f\ p_f]$. In this example, the Q node specializes to a CHK node with additional dummy VAR node sending constant message $[1 - p_f\ p_f]$ that depends on SNR_{SR}. Resulting LDPC-LDGM and LDPC-LDPC factor graphs are depicted in Figs. 3.11-3.12.

3.4 Bit Interleaved Coded Modulation

Thus far, only the binary input Gaussian relay channel ($\mathbf{x}_S \in \{\pm\sqrt{P_S}\}^{N_S}$ and $\mathbf{x}_R \in \{\pm\sqrt{P_R}\}^{N_R}$) has been considered. The capacity advantage of cooperative relaying and effectiveness of QMF is most significant in the high SNR regime. In this regime, general complex input symbols $\mathbf{x}_S \in \mathbb{C}^{N_S}$ and $\mathbf{x}_R \in \mathbb{C}^{N_R}$ must be considered. Information theory suggests the use of input symbols drawn randomly from a Gaussian distribution in order to achieve channel capacity. This is impractical for implementation because of two reasons:

- modulators capable of generating arbitrary symbols are required

- decoding requires an exhaustive search over all possible symbol sequences

For efficient implementation, it is desirable to have input symbols drawn from a fixed *finite alphabet* and consider design of codes over sequences of such elementary waveforms e.g. $\mathbf{x}_S \in \mathcal{A}^{N_S}$ and $\mathbf{x}_R \in \mathcal{A}^{N_R}$ where $\mathcal{A}$ represents a finite set of constellation points. The goal is to have schemes that use preferably simple, structured signals and yet offer a reasonably large space of possible codes to choose from. Examples of signal constellations (modulation schemes) like pulse amplitude modulation (PAM), quadrature amplitude modulation (QAM) etc. with representations in I-Q (inphase-quadratute) signal space are shown in Fig. 3.13.

Various *coded modulation* methods have been proposed to address the above goal. Schemes like trellis coded modulation (TCM) [68] and multilevel coded modulation (MLC) [29] pro-
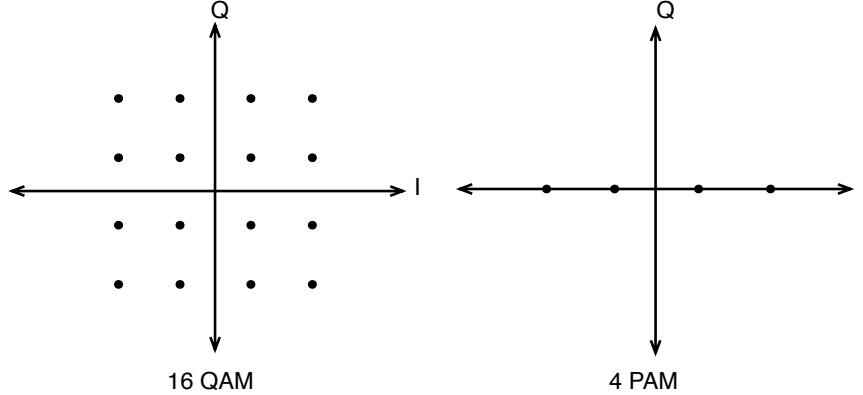
Figure 3.13: I-Q signal space representation for popular modulation constellations.

pose a unified approach for designing coding and modulation schemes jointly. BICM, introduced by Zehavi [76] in 1992, uses a radically different approach. It logically separates the coding and modulation problems by using an interleaving permutation. BICM is inherently sub-optimal i.e. involves loss of information between demodulation and decoding operations. It proposes soft-output demodulation followed by soft-input decoding to minimize information loss. A comprehensive analysis of BICM due to Caire et. al. [12] in 1998, shows that the loss incurred by BICM is very small. The ease of design and system flexibility outweigh this minor performance loss. For practical wireless systems, BICM is considered the *de facto* standard and is widely used in both cellular and local area networks. BICM has also been proposed for various cooperative communication systems [34][33][8][56]. In this section, a brief overview of BICM is provided, followed by its extension to the factor graph based coding framework developed above.

3.4.1 Classical and Parallel BICM

The classical BICM [12] scheme for a point to point channel consists of a binary code C, a bit interleaver represented by permutation π and a binary mapping μ. The modulation order 2^L represents the cardinality of symbol constellation $\mathcal{A}$. The mapping $\mu : \{0,1\}^L \to \mathcal{A}$ corresponds to a binary labeling for every symbol in the constellation. As an example, consider a 4-PAM constellation with average symbol energy P as shown in Fig. 3.14. 4-PAM has a cardinality of 4 i.e. $\mathcal{A} = \{-\frac{3\sqrt{P}}{2}, -\frac{\sqrt{P}}{2}, \frac{\sqrt{P}}{2}, \frac{3\sqrt{P}}{2}\}$. Consider a sequence of binary codewords $\mathbf{c_0}, \mathbf{c_1}, \mathbf{c_2}, \ldots$ each having blocklength N. Assume that the interleaver π has a depth of ξ codewords i.e. π is a random permutation with $N\xi$ bits at input arranged as a $\xi N/L$ sequence of L-tuples at output. For $L = 2$, the tuples are simply ordered pairs of form $\{b_A, b_B\}$. Each tuple is mapped to a symbol in $\mathcal{A}$ using the mapping μ, used to form transmit symbols $\mathbf{x} \in \mathcal{A}^{N\xi/2}$. In the 4-PAM example, consider μ to be a Gray labeling as shown in Fig. 3.14.
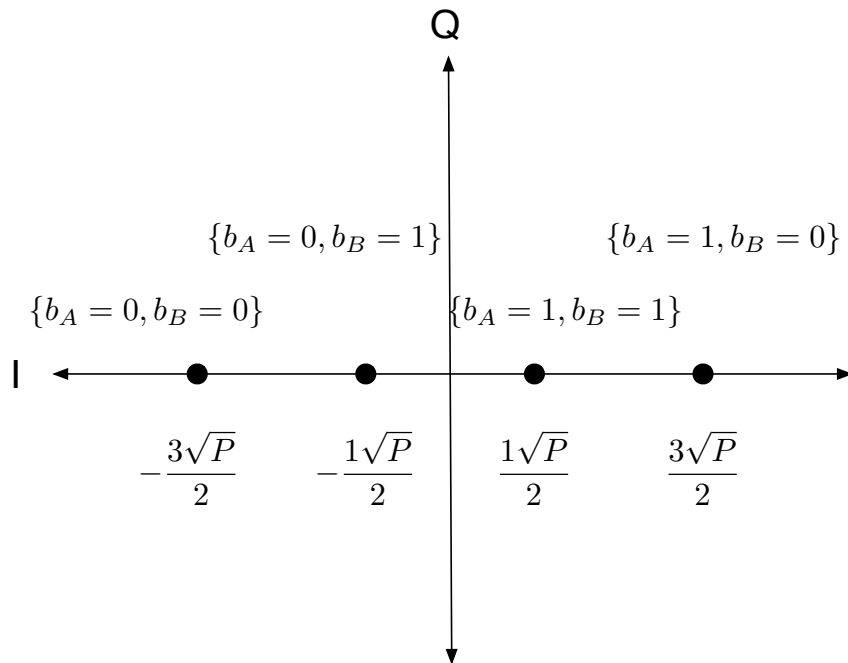
Figure 3.14: 4-PAM with Gray labeling.

At the receiver, consider the inverse mapping of μ i.e. $b_j : \mathcal{A} \to \{0,1\}$. $b_j(x)$ is used to denote the j bit in the labeling of symbol $x \in \mathcal{A}$ e.g. $b_A(\frac{\sqrt{P}}{2}) = 1$. The following subsets of $\mathcal{A}$ can be defined:

$$\mathcal{A}_{j,b} := \{x \in \mathcal{A} : b_j(x) = b\}$$

For the 4-PAM example, these are given below:

$$\mathcal{A}_{A,0} = \left\{ -\frac{3\sqrt{P}}{2}, -\frac{\sqrt{P}}{2} \right\}$$

$$\mathcal{A}_{B,0} = \left\{ -\frac{3\sqrt{P}}{2}, \frac{3\sqrt{P}}{2} \right\}$$

$$\mathcal{A}_{A,1} = \left\{ \frac{\sqrt{P}}{2}, \frac{3\sqrt{P}}{2} \right\}$$

$$\mathcal{A}_{B,1} = \left\{ -\frac{\sqrt{P}}{2}, \frac{\sqrt{P}}{2} \right\}$$

The BICM demodulator treats each bit in the symbol independently and computes a *bit*

wise APP metric from channel observations $\mathbf{y} \in \mathbb{C}^{N\xi/2}$.

$$P_{B_j|Y}[b_j = b|\mathbf{y}] = \frac{1}{|\mathcal{A}_{j,b}|} \sum_{x' \in \mathcal{A}_{j,b}} P_{X|Y}(x'|\mathbf{y})$$

For the 4-PAM example, $P_{B_j|Y}(b_B = 0|\mathbf{y})$ can be computed as:

$$P_{B_j|Y}(b_B = 0|\mathbf{y}) = \frac{1}{2} P_{X|Y}\left(x = -\frac{3\sqrt{P}}{2} \mid \mathbf{y}\right) + \frac{1}{2} P_{X|Y}\left(x = \frac{3\sqrt{P}}{2} \mid \mathbf{y}\right)$$

These bitwise soft estimates are rearranged by deinterleaving i.e. using reverse permutation π^{-1} before respective codewords are decoded independently.

A sufficiently deep interleaver $\xi \to \infty$ ensures that for any codeword $\mathbf{c}$, each bit is mapped to an independent transmit symbol with high probability. For a memoryless channel, this also implies that there is no correlation between bits (of the same codeword). Assume that transmit codewords $\mathbf{c_0}, \mathbf{c_1}, \ldots$ are i.i.d and uniformly distributed over the space of all possible codewords. This means that for b_j from a given symbol x, the other bits $b_{k \neq j}$ in the symbol are Bernoulli$(1/2)$ random variables. Since every codeword is decoded separately, correlation between bits of different codewords is explicitly ignored leading to a minor loss of information.

Caire et. al. [12] proposed a useful model to visualize BICM, wherein the original memoryless channel is decomposed into *parallel independent memoryless* "sub-channels". Every "sub-channel" $P_{Y|B,S}(y|b,s)$ has binary inputs $b \in \{0,1\}$. It depends on state $s \in \{1, 2, \ldots, L\}$ which is chosen uniformly and known to both the terminals. At the receiver, LLR for a bit that was mapped to state s is calculated from observation $y \in \mathbb{C}$.

$$LLR(y, s) = \log \frac{P_{B|Y,S}(b = 0|y, s)}{P_{B|Y,S}(b = 1|y, s)}$$

$LLR(y, s) \in \mathbb{R}$ serves as the output for each logical binary "sub-channel".

These binary "sub-channels" are not guaranteed to be output-symmetric i.e. the detection error probability for a bit is not independent of its value. Let $f_\Lambda(\lambda)$ represent the PDF of $LLR(y, s)$, the channel is output symmetric if the following condition is true:

$$f_{\Lambda|B}(\lambda|b = 0) = f_{\Lambda|B}(-\lambda|b = 1)$$

In the 4-PAM example, the demodulation rule is to select the constellation point in $\mathcal{A}$ that is nearest to the channel observation y. If $b_B = 0$ is transmitted, the error event at the demodulator is :

$$\mathcal{E}_{0 \to 1}(y) := \mathbf{1}[\Re(y) \in (-\sqrt{P} \ \sqrt{P})]$$
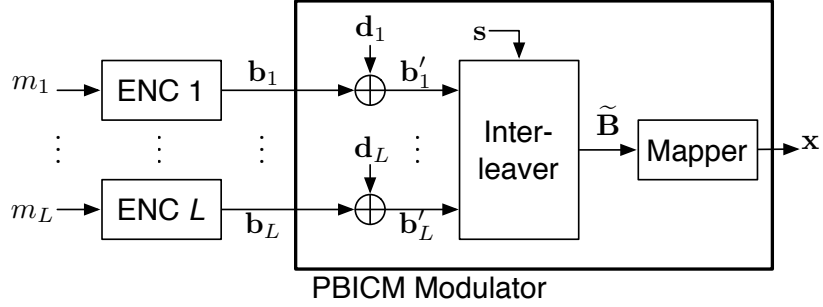
(a) PBICM Transmitter.
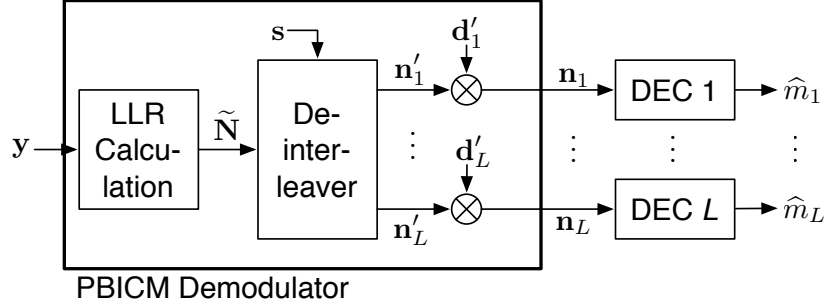
(b) PBICM Receiver.

Figure 3.15: PBICM Architecture: $\{\mathbf{d}_i\}_{i=1}^{L}$ are dithers, and $\mathbf{d}_i' = 1 - 2\mathbf{d}_i$.

Instead, if $b_B = 1$ is transmitted the error event is:

$$\mathcal{E}_{1 \to 0}(y) := \mathbf{1}[\Re(y) \in (-\infty, -\sqrt{P}) \cup \Re(y) \in (\sqrt{P}, \infty)]$$

As seen from Fig. 3.14, for Gaussian y:

$$P[\mathcal{E}_{0 \to 1}(y)] \neq P[\mathcal{E}_{1 \to 0}(y)]$$

Conventional methods for analysis and design of linear coding schemes cannot be used with asymmetric channels. This issue can be resolved by adding random dithers at every bit to make the channel output-symmetric as proposed in [12, 25, 30]. Dithers are i.i.d. Bernoulli $\left(\frac{1}{2}\right)$ variables known to both the transmitter and receiver. For a dither $d \in \{0, 1\}$, the channel $p_{Y|B,S,D}(y|b, s, d)$ is BMS. Its output is given by:

$$LLR(y, s, d) = (-1)^d LLR(y, s)$$

This method is called parallel BICM (PBICM) in [30]. Fig. 3.15 shows the architecture

for a PBICM point to point link having L states. $\{m_i\}_{i=1}^L$ represent messages. $\mathbf{b}_i$ and $\mathbf{b}'_i$ represent the transmit codewords before and after dithering. The equivalent BMS channel is characterized by the number of states L, the SNR of the underlying AWGN channel and the symbol mapping in modulation. It is known that Gray labeling is close to optimal in several BICM scenarios. Therefore, Gray labeling is used for all examples in the rest of this dissertation. For the 4-PAM example considered above, b'_B is the transmitted bit after applying Bernoulli$(\frac{1}{2})$ random dither to b_B. The detection error probability $P_{\text{error}}[b'_B]$ for b'_B is independent of its value as shown below:

$$P_{\text{error}}[b'_B] = \frac{1}{2}P[\mathcal{E}_{0\to1}(y)] + \frac{1}{2}P[\mathcal{E}_{1\to0}(y)]$$

3.4.2 BICM for QMF relaying

In order to use PBICM with the relay channel, a definition of the quantize-and-map operation under PBICM is required. With a PBICM modulator at source S, the observations at relay R ($\mathbf{y}_R$) represent L interleaved codewords. If R performs quantization at the *symbol* level, then the PBICM decomposition into independent binary sub-channels is lost. As an alternative, it is proposed that R perform quantization at the *bit* level.

S and R use PBICM modulator blocks with constellation size 2^L having state and dither vectors given by $\mathbf{s}_S, \mathbf{s}_R, \mathbf{D}_S$ and $\mathbf{D}_R$ respectively. The QMF operation at R is described below (depicted in Fig. 3.16):

1. For observed symbol sequence $\mathbf{y}_{SR} := \{y_{SR,j}\}_{j=1}^{fN_S}$ perform PBICM demodulation. The output is represented as $\{\mathbf{n}_{SR,i}\}_{i=1}^L$ where each $\mathbf{n}_{SR,i} := \{n_{SR,i,j}\}_{j=1}^{fN_S}$ represents LLRs for the ith codeword.

2. Quantize every LLR in $\{\mathbf{n}_{SR,i}\}_{i=1}^L$. As an example, for a one bit scalar quantizer this simply involves observing the sign of LLRs.

3. Encode the quantizer output $\{m_{R,i}\}_{i=1}^L$ using an LDPC/LDGM code.

4. Transmit the resultant codewords $\{\mathbf{b}_{R,1}\}_{i=1}^L$ using a PBICM modulator.

Using this definition of QMF, the Gaussian relay channel is decomposed into L parallel BMS relay channels. It is characterized by constellation size and SNRs of the underlying AWGN links i.e. $\mathsf{SNR}_{SR}, \mathsf{SNR}_{SD}, \mathsf{SNR}_{RD}$.

3.5 Simulation Experiments

In this section, results from simulation experiments are shown. Factor graphs shown in Figs. 3.11-3.12 i.e. using a scalar quantizer at the relay and D-BLAST channel decomposition
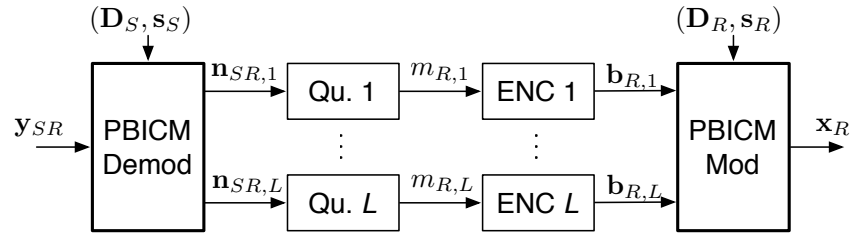
Figure 3.16: QMF Relaying with PBICM.

(a) Equivalent Relay Channel.

(b) Equivalent Point-to-Point Channel.
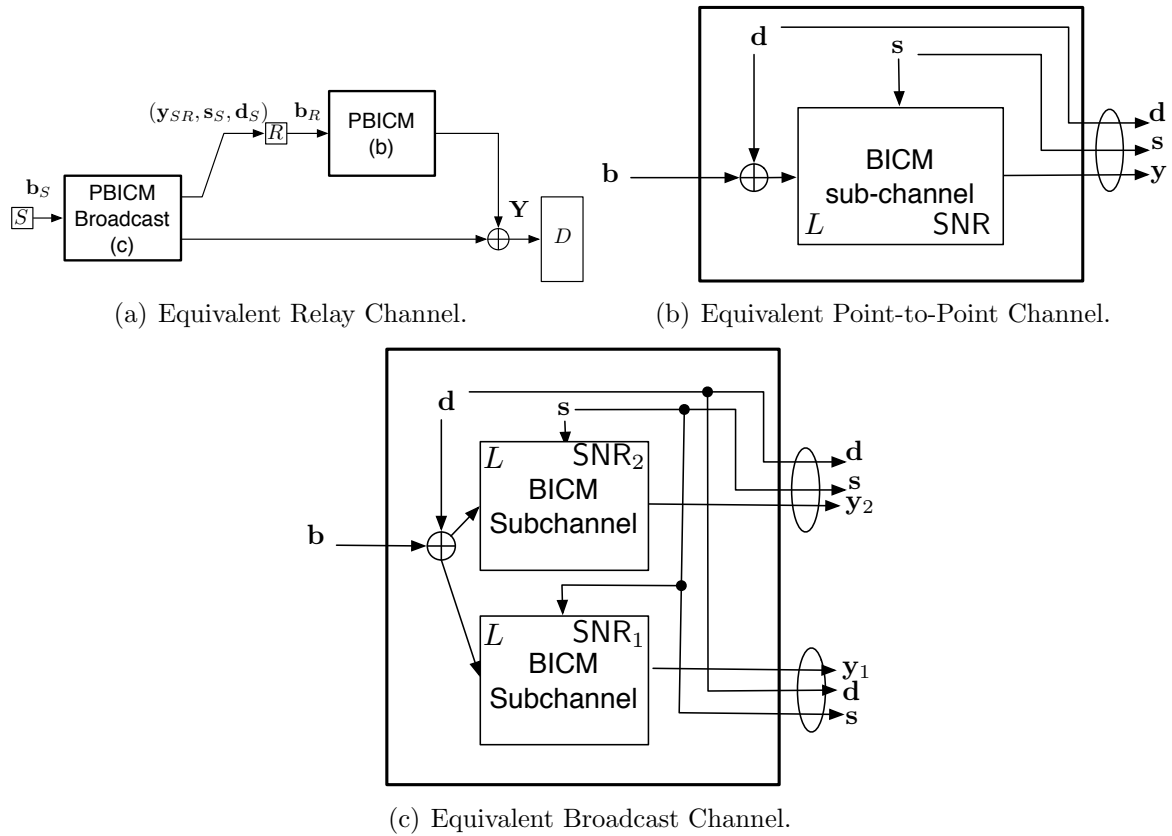
(c) Equivalent Broadcast Channel.

Figure 3.17: Equivalent binary-input system.

are used. For the experiments, bit-error-rate (BER) performance is computed as a function of SNR. For construction of the joint-factor graph, component sparse graph code matrices using off-the-shelf designs are used. Sparse graph code matrices are broadly classified into *regular* and *irregular* codes. In a (d_v, d_c) regular code, every variable node has degree d_v and every check node has degree d_c. Irregular codes allow nodes to have arbitrary degree distributions subject to maximum constraints $d_v^{\max}$ and $d_c^{\max}$. Since decoding is performed on the *joint* factor graph, performance of the coding scheme depends on the *combination* of component codes chosen for the experiment. In this section, only results from selected code combinations that were found to perform well are discussed.

Both LDPC-LDPC and LDPC-LDGM graphs structures are considered. Off-the-shelf LDPC code matrices designed in [58] are used. For the LDPC-LPDC scheme, both component LDPC codes C_S^b and C_R^b are chosen from an irregular code profile designed in [58] having $d_v^{\max} = 8$ and $d_c^{\max} = 7$. For the LDPC-LDGM scheme, the same profile is used for C_S^b with a regular $(10, 5)$ LDGM code C_R^b at relay. The design rate for the codes are as follows:

$$r_{\mathrm{LDPC}} = 0.5 \quad r_{\mathrm{LDGM}} = 2$$

To satisfy the half-duplex constraint with these rates, a listening fraction at the relay $f = 1/3$ is used. Practical wireless systems use channel codes with blocklengths $\approx 10^3 - 10^4$. In our simulations $N_S = 20400$ is used. $N_R = 13600$ is calculated from the listening fraction f and N_S.

The SNRs of the channels are as described in Sec. 3.3.7. Network geometry is modeled by considering SNR_{SD} as reference and using linear scaling factors q_{SR} and q_{RD} for other SNR quantities.

$$\mathsf{SNR}_{\mathrm{SR}} = q_{\mathrm{SR}}\mathsf{SNR}_{\mathrm{SD}}$$
$$\mathsf{SNR}_{\mathrm{RD}} = q_{\mathrm{RD}}\mathsf{SNR}_{\mathrm{SD}}$$

To provide a baseline for performance comparisons, an information theoretic *upper* bound is used on maximal achievable rate using QMF relaying with continuous Gaussian inputs and a vector Gaussian quantizer at the noise level.

$$\mathcal{R}^{QMF} = \min \left\{ \begin{array}{l} (1 - f)\mathfrak{C}_{\mathrm{G}}\left(\mathsf{SNR}_{SD}\right) + f\mathfrak{C}_{\mathrm{G}}\left(\frac{\mathsf{SNR}_{SR}}{2} + \mathsf{SNR}_{SD}\right), \\ (1 - f)\mathfrak{C}_{\mathrm{G}}\left(\mathsf{SNR}_{RD}\right) + \mathfrak{C}_{\mathrm{G}}\left(\mathsf{SNR}_{SD}\right) - f \end{array} \right\} \tag{3.7}$$

where $\mathfrak{C}_{\mathrm{G}}\left(z\right) := \log\left(1 + z\right)$ is the AWGN point-to-point capacity at signal-to-noise ratio z. If inputs are constrained to structured constellations like 16-QAM, 64-QAM etc, the achievable QMF rate with BICM is given as follows:

$$\mathcal{R}_L^{QMF} = \min \left\{ \begin{array}{l} (1 - f)\mathfrak{C}_L\left(\mathsf{SNR}_{SD}\right) + f\mathfrak{C}_L\left(\frac{\mathsf{SNR}_{SR}}{2} + \mathsf{SNR}_{SD}\right), \\ (1 - f)\mathfrak{C}_L\left(\mathsf{SNR}_{RD}\right) + \mathfrak{C}_L\left(\mathsf{SNR}_{SD}\right) - f \end{array} \right\} \tag{3.8}$$

Here $\mathfrak{C}_L(z)$ denotes the point-to-point capacity at SNR ratio z with QAM structured signal constellations and BICM. L denotes the constellation order i.e. $|\mathcal{A}| = 2^L$. As a reference, the cut-set bound (defined in Sec. 2.3.2) under BICM ($\mathcal{R}_L^{(cutset)}$) is also computed. $\mathfrak{C}_L(z)$ and simulation performance of C_S^b over a point-to-point channel (no cooperation) are used as comparison benchmarks.

Since the chosen binary codes have rate 0.5, the overall throughput after modulation is fixed to $0.5 \times L$. The maximum achievable rate quantities described above are expressed in terms of the minimum SNR threshold at which the rate $0.5 \times L$ becomes achievable[2]. For 2-PAM ($L = 1$) and 16-QAM ($L = 4$) calculated SNR thresholds are shown in Table. 3.1-3.2. Corresponding BER simulations for the selected codes and above parameters plotted w.r.t. SNR_{SD} are shown in Fig. 3.18-3.21.

Table 3.1: Information theoretic quantities in terms of SNR_{SD} (decibels) for 2-PAM (L=1) and $f = 1/3$

Point-to-Point	Relay: $q_{RD} = 1, q_{SR} = 2$		Relay: $q_{RD} = 1, q_{SR} = 10$	
$\mathfrak{C}_L$	$\mathcal{R}_L^{(cutset)}$	$\mathcal{R}_L^{QMF}$	$\mathcal{R}_L^{(cutset)}$	$\mathcal{R}_L^{QMF}$
0.19	-1.51	-1.11	-2.83	-2.56

Table 3.2: Information theoretic quantities in terms of SNR_{SD} (decibels) for 16-QAM (L=4) and $f = 1/3$

Point-to-Point	Relay: $q_{RD} = 1, q_{SR} = 2$		Relay: $q_{RD} = 1, q_{SR} = 10$	
$\mathfrak{C}_L$	$\mathcal{R}_L^{(cutset)}$	$\mathcal{R}_L^{QMF}$	$\mathcal{R}_L^{(cutset)}$	$\mathcal{R}_L^{QMF}$
5.11	3.44	3.95	1.79	2.72

3.5.1 Discussion

In Fig. 3.18, note that the BER performance over the relay channel with the LDPC-LDPC coding scheme is ≈ 0.9dB better than the same code over a point-to-point channel (*cooperation gain*). The plot also shows the corresponding information-theoretic thresholds $\mathfrak{C}_L$ and $\mathcal{R}_L^{QMF}$ for chosen parameters f, q_{SR}, q_{RD}. The gap between these two thresholds is ≈ 1.3dB. This can be considered as the theoretical maximum *cooperation gain* using a 1-bit quantizer and QMF relaying. The observed performance gains are therefore quite comparable to those predicted by information theory. Moreover, the threshold corresponding to $\mathcal{R}_L^{QMF}$ is only

---

[2]A rate is said to be achievable at a given SNR if there exists a coding scheme that has arbitrarily small probability of decoding error at that SNR.

Figure 3.18: 2-PAM simulation results for $q_{\mathrm{SR}} = 2$, $q_{\mathrm{RD}} = 1$, $N_{\mathrm{S}} = 20400$, $N_{\mathrm{R}} = 13600$, $f = \frac{1}{3}$

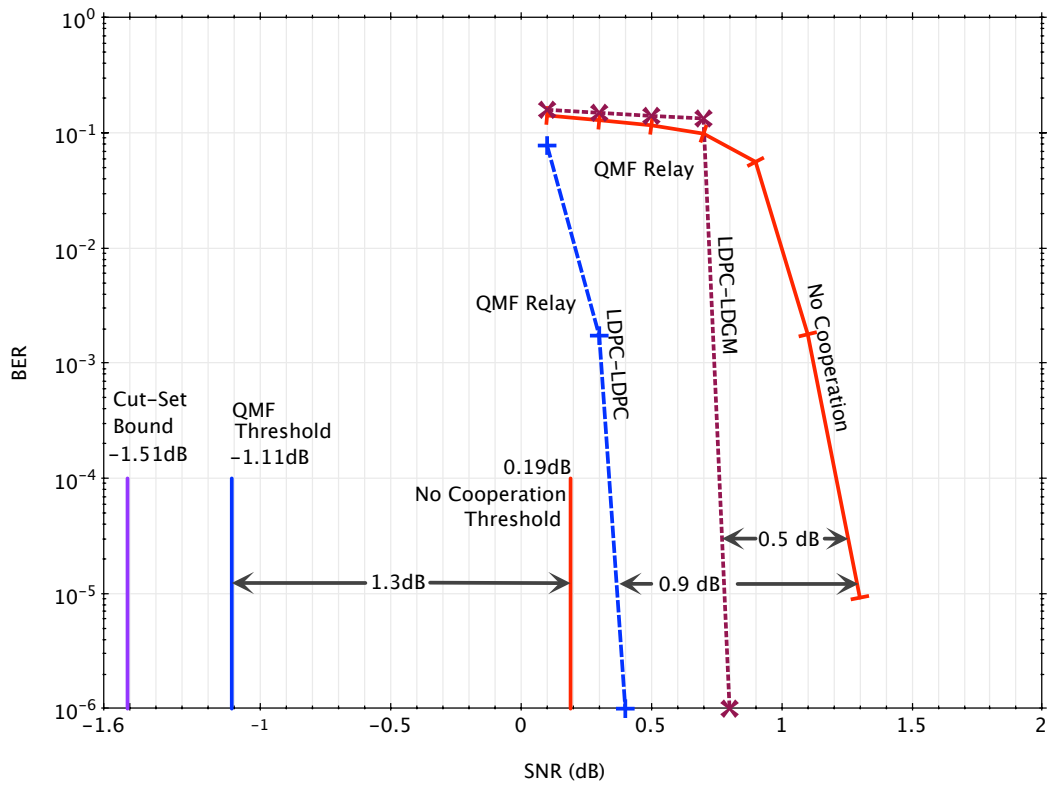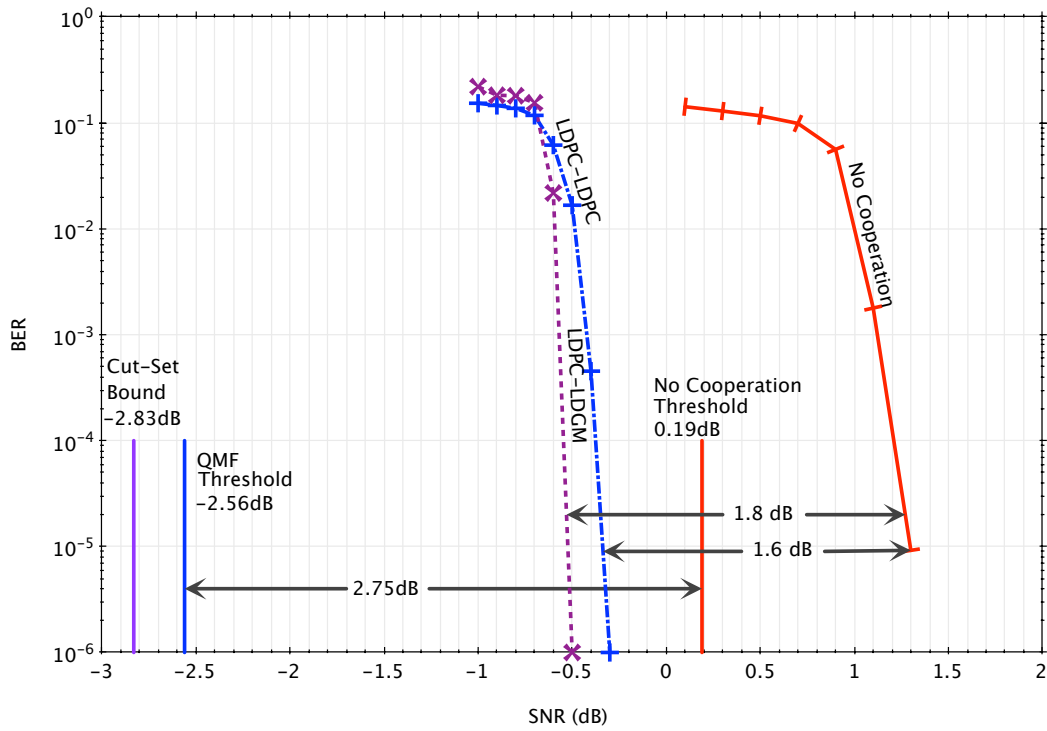Figure 3.19: 2-PAM simulation results for $q_{SR} = 10$, $q_{RD} = 1$, $N_S = 20400$, $N_R = 13600$, $f = \frac{1}{3}$
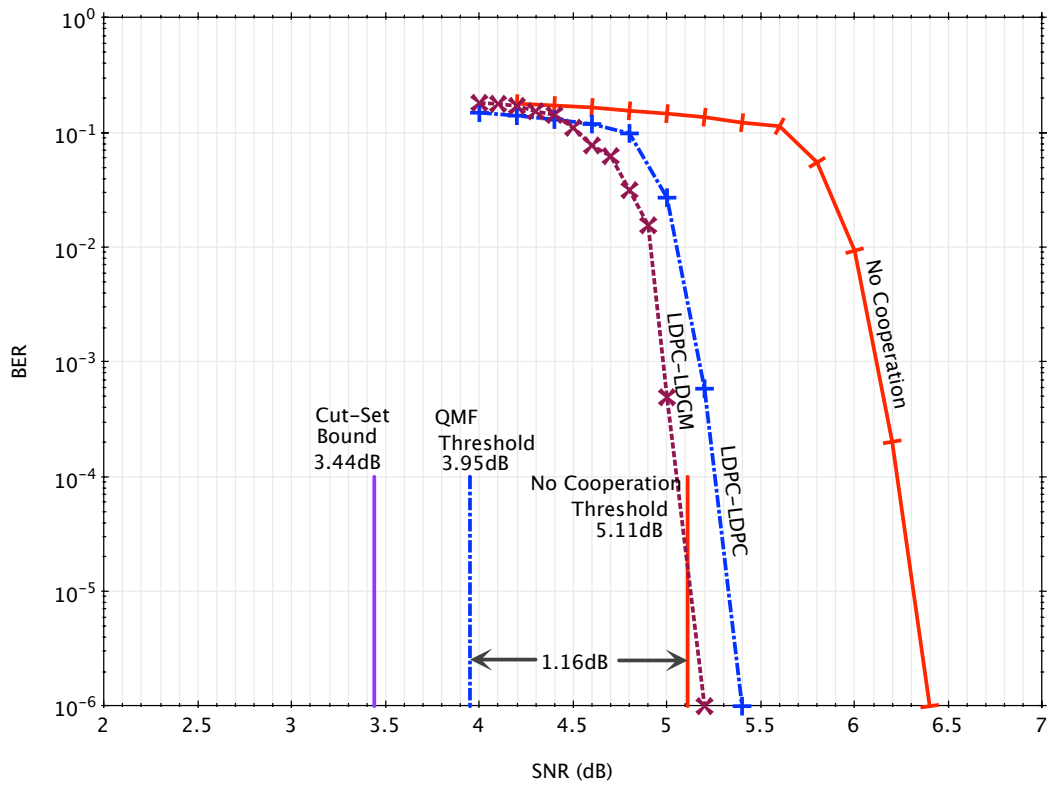
Figure 3.20: 16-QAM simulation results for $q_{\mathrm{SR}} = 2$, $q_{\mathrm{RD}} = 1$, $N_{\mathrm{S}} = 20400$, $N_{\mathrm{R}} = 13600$, $f = \frac{1}{3}$
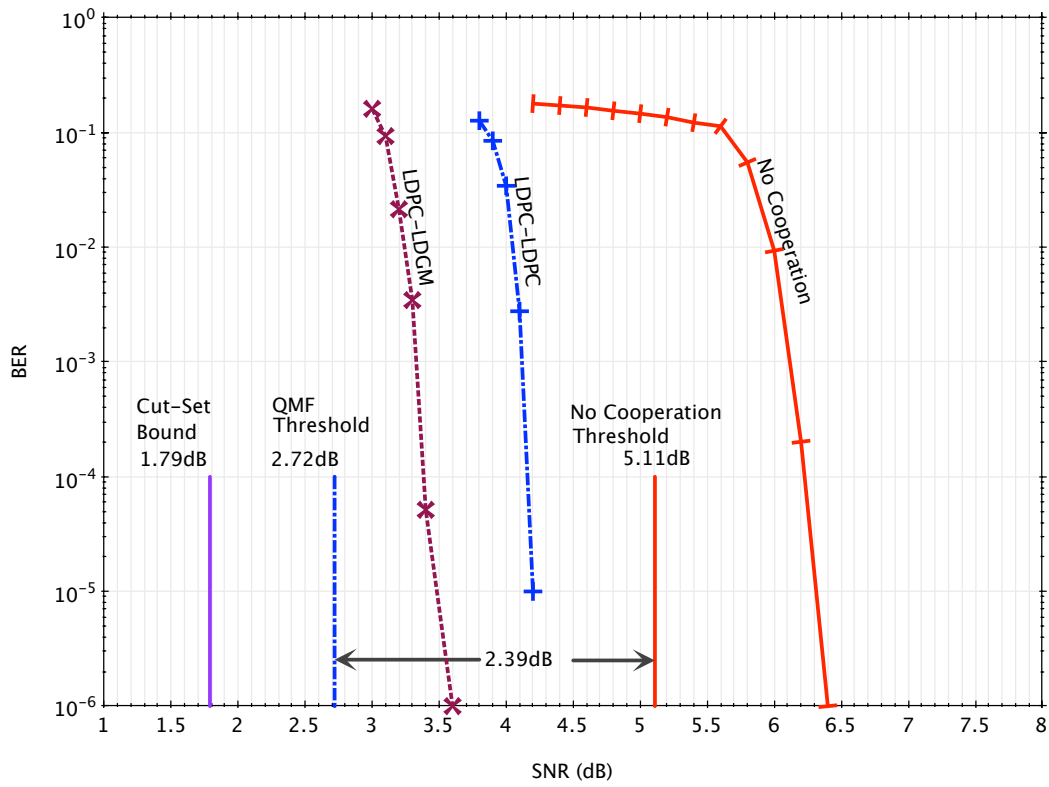
Figure 3.21: 16-QAM simulation results for $q_{SR} = 10$, $q_{RD} = 1$, $N_S = 20400$, $N_R = 13600$, $f = \frac{1}{3}$

0.4dB away from the cut-set upper bound, indicating that even with a 1-bit scalar quantizer, QMF relaying allows performance very close to capacity.

BER performance of a scheme depends both on the choice of code structures and the blocklengths. Large blocklengths offer better BER performance but are not very suitable for delay-constrained systems. Irregular point-to-point LDPC code designs that perform within 0.1dB of the Shannon limit are shown in [58] using a block length of 10^6. However, even the best point-to-point codes from [58] perform ≈ 0.8dB away from the Shannon limit at a blocklength of 10^4. Consider the BER performance for system parameters shown in Fig. 3.21 using block-lengths of the same order (10^4). Note that the performance of the LDPC-LDGM scheme is just ≈ 1dB away from the QMF information theoretic threshold. This is an encouraging indication that the proposed cooperative coding scheme can perform very close to fundamental limits in some scenarios.

However, a general comparison between the LDPC-LDPC and LDPC-LDGM configurations is not possible using above results. Furthermore, for 3 of the 4 cases shown, BER performance of both LDPC-LDPC and LDPC-LDGM schemes are still far (> 1.5dB) from $\mathcal{R}_L^{QMF}$.

Choice of component codes for the above results is based on an empirical "trail-and-error" process. Such an *ad hoc* design approach prevents exploration of the full design space. In the next chapter, progress towards addressing this limitation is presented.

Chapter 4

Code Profiles

The previous chapter outlines a coding and system design framework for the relay channel using QMF. It is based on using sparse binary codes, BICM and iterative sum-product decoding over a low complexity factor graph. By using simulation experiments, it is shown to achieve gains similar to those predicted by information theory. The scheme uses a *pair* of codes, one at the source and the other for mapping at the relay. The overall performance depends on the choice of suitable code *pairs*.

In Chapter 3, off-the-shelf code profiles are picked using a trail-and-error procedure for finding good pairs. Each trail involves time consuming simulation experiments. For practical systems that operate over a wide range of channel conditions, it is desirable to have a more systematic approach to study code pairs.

This chapter presents progress towards the above goal. Density evolution techniques are typically used for analysis and design of LDPC codes. This chapter presents methods to adapt these tools for analysis of QMF factor graphs. In Chapter 3, two configurations are considered namely, LDPC-LDPC and LDPC-LDGM. It turns out that density evolution tools do not adapt readily to LDPC-LDPC graphs. Nonetheless, they are suitable for analysis and design of codes in the LDPC-LDGM configuration. Based on these tools, examples of LDPC-LDGM code designs that perform within $0.5 - 1.5$dB of information theoretic limits are demonstrated for the relay channel.

4.1   Analysis of Sparse Graph Codes

This section presents an overview of the density evolution framework developed by Richardson and Urbanke [59]. A sketch of the theory and associated historical context is provided followed by a detailed discussion of selected components. In Sec. 4.2, these tools are applied towards analysis of QMF factor graphs.

A "good" code over a given channel is one that achieves a small probability of decoding error at a rate that is close to the channel's information theoretic threshold. Constructing

specific codes that are provably good is difficult. Most constructions are based on the *probabilistic method.* An *ensemble of codes* is constructed using a random process and it is shown that good codes occur in this ensemble with high probability. Several information theoretic results are based on codes drawn from loosely constrained random ensembles. For example, any linear code picked uniformly at random is "good" with high probability over a binary symmetric channel (BSC). However, the ML decoder for an arbitrary linear code is NP-Complete [10] i.e. decoding is exponential in block length. This remains true even if the destination has full knowledge of the code and unlimited pre-processing time to find an efficient decoder [11].

Design of practical codes with efficient encoding and decoding algorithms requires the use of constrained and structured ensembles. This work focuses on ensembles of sparse bipartite graphs in conjuction with belief propagation algorithms, which are known to have low complexity (Ch. 3).

4.1.1   Overview of Density Evolution

The earliest approach to design and analyze LDPC codes for a BSC is due to Gallager [23] (1963). This method considers a sequence of (d_v, d_c)-regular bipartite graphs with increasing block length n containing no cycles shorter than $2\ell(n)$ where:

$$\ell(n) := \frac{\ln n - \ln \frac{d_v d_c - d_v - d_c}{2d_c}}{\ln[(d_c - 1)(d_v - 1)]}$$

Consider belief propagation decoding over this sequence (of graphs). Messages passed along graph edges may be *correct* or *incorrect.* For an incorrect message, the associated hard decision does not agree with the true value of the associated codeword bit. The average fraction of incorrect messages passed at the ℓth iteration is expressed as a system of coupled recursive functions. The functions depend on the channel parameter (i.e. BSC crossover probability δ) and the degree profile of the graph sequence (d_v, d_c). Based on these recursions the following condition is tested: "Does average probability of error approach zero as number of iterations are increased?" The worst channel parameter (e.g. highest δ) for which the condition holds true is called the *belief propagation threshold* for the graph sequence.

Luby, Mitzenmacher, Shokrollahi and Spielman [42, 43] (1998) propose a refinement of this approach. They observe that codes with irregular degree profiles have better performance than regular codes.

For an intuitive explanation of this phenomenon [42, 43], consider a game to design a fixed rate code. The variable and check nodes are players in the game and each player wants to determine its degree. One constraint is that the sum of degrees on the variable node side must equal the sum of degrees on the check node side. In order to get information from as many check nodes as possible and thereby improve their chances of being decoded correctly, variable nodes tend to high degrees. Check nodes want to have a low degree so as

to transmit more valuable information to their respective neighbors. Irregular graphs offer greater flexibility to balance these competing requirements. Some variable nodes can end up with high degrees while others with low degrees. The variable nodes with high degrees tend to correct their value in just a few iterations. In turn, they provide more valuable information to check nodes which pass it on to variable nodes having lower degrees. This leads to a wave effect and improves the overall decoding performance.

Gallager's construction, which uses a sequence of graphs with increasing block-length and tight constraints on the length of cycles, becomes intractable for complex irregular graphs. Therefore, Luby et.al deviate from it and consider graphs that are sampled from random ensembles without having any constraints on cycles. Richardson, Urbanke and Shokrollahi have unified the above ideas into a comprehensive density evolution framework presented in two seminal papers [59, 58]. They have extended the analysis tools beyond the BSC and binary erasure channel (BEC) to cover a broad class of channels and decoding algorithms. For example, their tools are readily applicable to general binary memoryless symmetric (BMS) channels. The BICM equivalent QMF relay channel discussed in Ch. 3 falls under this category.

The theory is built upon three fundamental observations which are summarized below. Rigorous proofs are available in [59, 61].

1. **Concentration around ensemble average:**

   For any random ensemble defined using degree profiles, almost all graphs in it behave increasingly alike with increasing block-length. This convergence is exponentially fast in block-length. Therefore, it suffices to study the average behavior of the ensemble, to characterize any individual graph contained in it. The essence of this observation is that decoding performance inherently depends on the structure of local neighborhoods in the graph. For large block-lengths, all graphs in an ensemble have similar local neighborhoods.

2. **Convergence to cycle-free case:**

   Local neighborhoods in typical graphs appear like trees if the number of iterations is fixed and block-lengths tend infinity. Consequently for large graphs, cycles can be ignored from analysis of *typical structures*. It suffices to simply consider message passing over typical *computation trees*.

3. **Density evolution and threshold determination:**

   The typical computation tree for an ensemble is expressed using normalized degree profiles. It is possible to track the PDF of a typical message as it passes through a typical computation tree. The evolution is expressed as a system of coupled recursive functions that allow computation of the average probability of decoding error at any given iteration. For channels that can be ordered by *physical degradation* e.g. BSC, BEC and BIAWGN, the channel parameter shows a threshold phenomenon. The belief

propagation threshold clearly separates a region where probability of decoding error converges to zero and a region where it converges to a non-zero value.

The framework allows optimizing over the space of degree profiles to find ensembles whose average performance is close to fundamental limits. In [58], degree profile designs for irregular LDPC codes within 0.06dB of Shannon capacity are demonstrated for the BIAWGN channel. Code profile optimization is known to be non-convex and computationally challenging. The density evolution procedure by itself is significantly complex and involves tracking continuous PDF functions over many iterations. To simplify the search for good degree profiles, Chung et.al. propose [15] (2001) a Gaussian approximation to density evolution. If messages are approximated to have a Gaussian distribution, density evolution reduces to the tracking of one scalar quantity. Though approximate, this simplified procedure allows a more comprehensive exploration of the design space and has led to better code designs.

In the remainder of this section, random graph ensembles, density evolution recursions and the Gaussian approximation are discussed in greater detail.

4.1.2   Random Graph Ensembles

A binary linear block code C of rate $\mathcal{R} = \frac{k}{n}$ is represented either by a $k \times n$ generator matrix $\mathbf{G}$ or a $(n - k) \times n$ parity check matrix $\mathbf{H}$. Note that these matrices are not unique. Elementary row transformations on $\mathbf{G}$ and $\mathbf{H}$ do not change the code. This is because, C is formed by the column space of $\mathbf{G}$ and the null space of $\mathbf{H}$. A code has several possible $\mathbf{G}$ and $\mathbf{H}$ matrices. An LDPC code has at least one sparse $\mathbf{H}$ and an LDGM code has at least one sparse $\mathbf{G}$. Sparse matrices are represented using sparse bipartite graphs as discussed in Sec. 3.1.1. An LDPC code is represented by a bipartite graph having n variable nodes and $(n-k)$ check nodes and $\mathbf{H}$ as the graph adjacency matrix. An LDGM code is represented by a bipartite graph having n variable nodes and k check nodes and $\mathbf{G}$ as its adjacency matrix.

Degree Profiles

An ensemble of bipartite graphs is defined by using two polynomials. The number of variable nodes having degree i is Λ_i. The number of check nodes with degree i is P_i. These are written in polynomial form as :

$$\Lambda(x) = \sum_i \Lambda_i x^i$$

$$P(x) = \sum_i P_i x^i$$

Note that each graph in the ensemble has $\Lambda(1)$ variable nodes and $P(1)$ check nodes. The number of edges emanating from variable nodes and check nodes must be equal. Therefore,

the polynomials must satisfy the following condition on their first derivative.

$$\Lambda'(1) = P'(1)$$

Edge Perspective

A random graph ensemble is completely characterized by the polynomials (Λ, P). These are called *degree distributions* from the *node perspective*. The goal of this formulation is to search for random graph ensembles that are likely to contain "good" codes. As discussed in Chapter 3, performance of a code is related to its block-length $n = \Lambda(1)$. Codes with longer block-length perform closer to information theory limits. Therefore, it is useful to consider degree distributions *normalized* by block-length.

A convenient normalized representation is to consider degree distributions $\lambda(x)$ and $\rho(x)$ from the *edge perspective*.

$$\lambda(x) = \sum_i \lambda_i x^{i-1} = \frac{\Lambda'(x)}{\Lambda'(1)}$$

$$\rho(x) = \sum_i \rho_i x^{i-1} = \frac{P'(x)}{P'(1)}$$

λ_i represents the *fraction of edges* that connect to a variable node of degree i. Similarly ρ_i is the fraction of edges that connect to a check node of degree i.

Random Sampling from Ensembles

Consider sampling a random graph from ensemble (Λ, P). This is achieved by the following construction procedure. Consider $\Lambda(1)$ variable nodes and $P(1)$ check nodes. Λ_i variable nodes have degree i and P_i check nodes have degree i. For a node with degree i, consider i *sockets* on the node such that there are $\Lambda'(1) = P'(1)$ sockets on each side of the graph. Label the sockets on each side with the set $[\Lambda'(1)] = \{1, 2, \ldots, \Lambda'(1)\}$. Consider a permutation σ on $[\Lambda'(1)]$. Connect the socket labelled j on the variable side with socket labelled $\sigma(j)$ on the check side. Induce a uniform probability distribution on the set of possible permutations. This implies that the graph is sampled uniformly at random from the ensemble. As discussed previously, a graph does not uniquely identify a code. Therefore, codes are not sampled using a uniform distribution.

The resulting graph may have multiple edges connecting a given pair of variable and check nodes. Consider a $P(1) \times \Lambda(1)$ adjacency matrix $\mathbf{M}$ for the graph with the entry at row i and column j giving the number of edges between check node i and variable node j. The purpose of constructing this graph is to represent a binary linear code with check nodes representing addition in the $GF(2)$ finite field. Multiple edges between two nodes are superfluous from this perspective. A transformation of the graph is defined by removing all

even pairs of multiple edges. This results in a graph, which has at most one edge between any variable node and check node. The resulting adjacency matrix $\mathbf{A}$ has $\{0, 1\}$ entries obtained by interpreting the entries of $\mathbf{M}$ in GF(2) i.e. (mod 2).

The resulting graph and adjacency matrix $\mathbf{A}$, have multiple interpretations. $\mathbf{A}$ can be interpreted as a parity check matrix $\mathbf{H}$ resulting in an LDPC construction. Alternatively $\mathbf{A}$ can be interpreted as a generator matrix $\mathbf{G}$ resulting in an LDGM construction.

Design Rate

The ensemble (Λ, P) is said to have an LDPC *design rate*:

$$r_{LDPC}(\Lambda, P) = \frac{k}{n} = 1 - \frac{n - k}{n} = 1 - \frac{P(1)}{\Lambda(1)}$$

The LDGM *design rate* is equivalently written as:

$$r_{LDGM}(\Lambda, P) = \frac{k}{n} = \frac{P(1)}{\Lambda(1)}$$

The ensemble *design rate* does not give the rate of every code in the ensemble. The random edge construction can possibly lead to parity checks that are not linearly independent. This can result in a code rate that is higher than the ensemble design rate.

For large block lengths, the actual rate of a random element is close to the design rate with high probability [61].

Design rates are also defined in terms of the edge perspective (n, λ, ρ) representation as follows:

$$r_{LDPC}(\lambda, \rho) = 1 - \frac{\int_0^1 \rho(x) \, dx}{\int_0^1 \lambda(x) dx}$$

$$r_{LDGM}(\lambda, \rho) = \frac{\int_0^1 \rho(x) \, dx}{\int_0^1 \lambda(x) dx}$$

For example, a $(3, 6)$ regular graph ensemble with arbitrary block-length is represented as follows:

$$\lambda(x) = x^2 \; ; \; \rho(x) = x^5$$

This gives an ensemble of LDPC codes having design rate $r_{LDPC} = 1/2$. It also gives an ensemble of LDGM codes having design rate $r_{LDGM} = 1/2$. Not all profiles have both LDPC and LDGM interpretations. As noted previously, LDGM codes are used for both data compression $(k < n)$ and channel coding $(k > n)$ scenarios. Consider a regular $(6, 3)$

profile.

$$\lambda(x) = x^5 \; ; \; \rho(x) = x^2$$

This represents an LDGM ensemble with design rate $r_{LDGM} = 2$. Since LDPC codes are only defined for $0 \leq r_{LDPC} \leq 1$, this profile cannot represent a valid LDPC ensemble.

4.1.3   Recursive Density Updates

Density evolution recursion functions for LDPC/LDGM codes are described in terms of ensembles defined above. Consider the iterative sum-product algorithm with LLRs as the messages passed among nodes. Let $w_{\mathsf{KV}}^{(\ell)}$ and $w_{\mathsf{VK}}^{(\ell)}$ denote the message sent from the check node K to the variable node V and vice versa, at the ℓ-th iteration.

The update rules for the algorithm in terms of LLRs are written as follows: ($\mathcal{N}(\cdot)$ here denotes the set of neighboring nodes)

$$w_{\mathsf{VK}}^{(\ell)} = \sum_{\mathsf{K'} \in \mathcal{N}(\mathsf{V}) \backslash \{\mathsf{K}\}} w_{\mathsf{K'V}}^{(\ell)} \tag{4.1}$$

$$w_{\mathsf{KV}}^{(\ell+1)} = 2 \tanh^{-1} \left(\prod_{\mathsf{V'} \in \mathcal{N}(\mathsf{K})} \tanh \left(\frac{1}{2} w_{\mathsf{V'K}}^{(\ell)} \right) \right) \tag{4.2}$$

In this representation, the messages $w_{\{\cdot\}}^{(\ell)}$ are continuous in $(-\infty \; \infty)$. Extreme values $\pm\infty$ represent full confidence for the associated bit. The goal is to track the PDF of a typical message as it propagates through a typical computation tree. Long block-lengths and cycle free graphs are assumed for the purpose of analysis. Due to these assumptions and the extrinsic information transfer property, the messages are treated as independent random variables. As seen in Eq. (4.1), the updates at variable nodes simply involve summation of such independent variables. As a result, PDF of an outgoing message from a variable node is computed as the convolution (denoted as $\otimes$) of the PDF's of incoming messages. For check nodes, the update rule is more complex (4.2). In [58] a variable transformation rule $\Gamma(\cdot)$ and its inverse $\Gamma^{-1}(\cdot)$ are defined. Under this transformation, the check node update rule is also expressed as a summation of incoming messages. If the density of the message $w_{\{\cdot\}}^{(\ell)}$ is denoted by $P_{\{\cdot\}}^{(\ell)}$, the density evolution recursion for a graph ensemble (λ, ρ) is written as

follows:

$$P_{\mathsf{KV}}^{(\ell+1)} = \Gamma^{-1}\left(\sum_j \rho_j \left(\Gamma\left(P_{\mathsf{VK}}^{(\ell)}\right)\right)^{\otimes(j-1)}\right)$$

$$P_{\mathsf{VK}}^{(\ell)} = P_{\mathsf{V}} \otimes \sum_i \lambda_i \left(P_{\mathsf{KV}}^{(\ell)}\right)^{\otimes(i-1)}$$

Here P_{V} represents the conditional PDF of the LLR's at the output of the channel, conditioned on the actual codeword that was transmitted. If the channel is output symmetric, error performance does not depend on the transmitted codeword. Consequently, it suffices to pick any one codeword for conditioning. A natural choice is to use the all-zero codeword. Since all codes in the ensemble are linear, they must all contain it. For the BIAWGN channel having noise variance σ^2, the P_{V} function is given as follows:

$$P_{\mathsf{V}}(y) = \sqrt{\frac{\sigma^2}{8\pi}} \exp\left(-\frac{(y - \frac{2}{\sigma^2})^2}{8}\right)$$

The above recursion on continuous PDF functions is significantly complex for symbolic analysis. Efficient numerical techniques are reported in [58, 15]. The Gaussian approximation approach [15] is widely used to obtain a simplified formulation. $P_{\{\cdot\}}^{(\ell)}$ is approximated by a Gaussian distribution $\mathcal{N}(\mu, \sigma)$. It is shown that, for such messages, $\sigma^2 = 2\mu$ due to the symmetry condition, which is preserved under density evolution [58, 15]. This implies that the recursion can be written in terms of a scalar variable. The recursion for ensemble (λ, ρ) is written in terms of the density mean $\mu^{(\ell)}$ at the ℓ-th iteration as follows:

$$\mu^{(\ell)} = \sum_j \rho_j \phi^{-1}\left(1 - \left[1 - \sum_i \lambda_i \phi\left(\mu_0 + (i-1)\mu^{(\ell-1)}\right)\right]\right) \tag{4.3}$$

Here μ_0 is the mean of P_{V}. $\phi(\cdot)$ is a continuous and monotonically decreasing function on $[0\ \infty)$ defined as follows:

$$\phi(x) := \begin{cases} 1 - \frac{1}{\sqrt{4\pi x}} \int_{\mathbb{R}} \tanh\frac{u}{2} \exp\left(-\frac{(u-x)^2}{4x}\right)\, du & \text{if } x > 0 \\ 1 & \text{if } x = 0 \end{cases}$$

For a BIAWGN channel with noise variance σ^2, the starting point of the recursion is given by $\mu_0 = \frac{2}{\sigma^2}$. Fig. 4.1 shows this evolution for a $(3, 6)$ (design rate $= 0.5$) ensemble over a BIAWGN channel with various noise parameters. Since it is assumed that the all-zero codeword is transmitted, *incorrect* messages correspond to LLR's that are less than zero. If the mean of the message PDF tends to infinity, the fraction of incorrect messages tends to zero. The Gaussian approximation belief propagation threshold (GABPT) is evident in
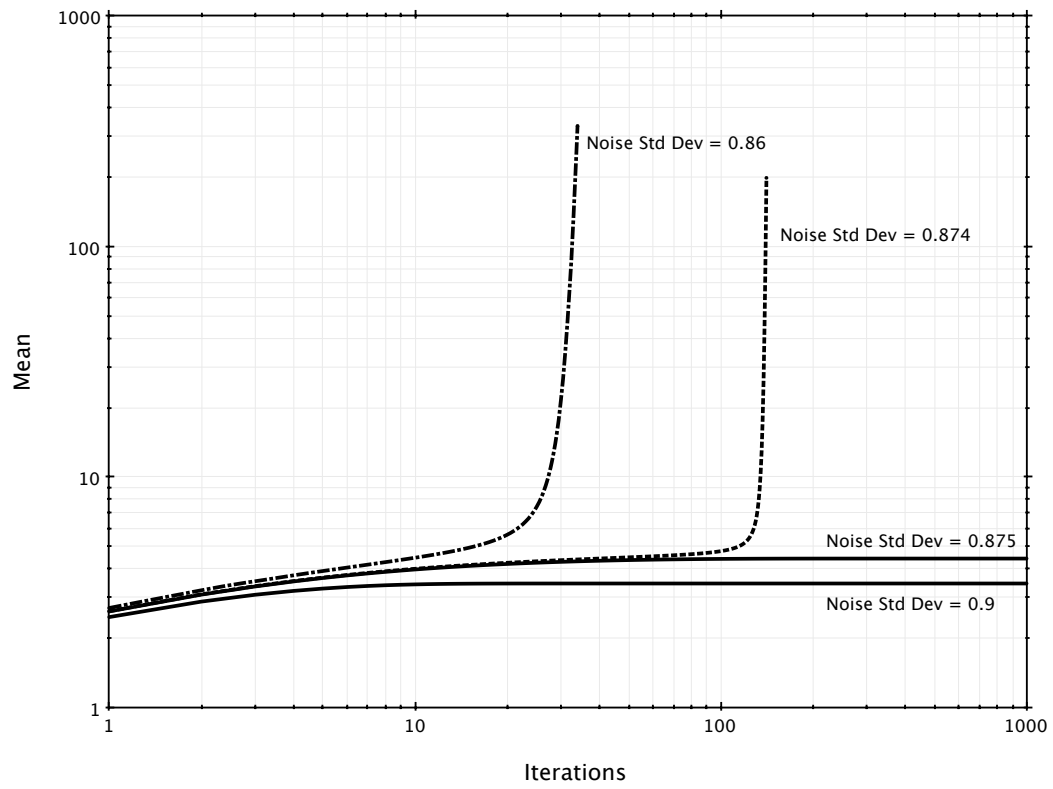
Figure 4.1: Density evolution using Gaussian approximation for $(3, 6)$ regular LDPC ensemble on BIAWGN channel.
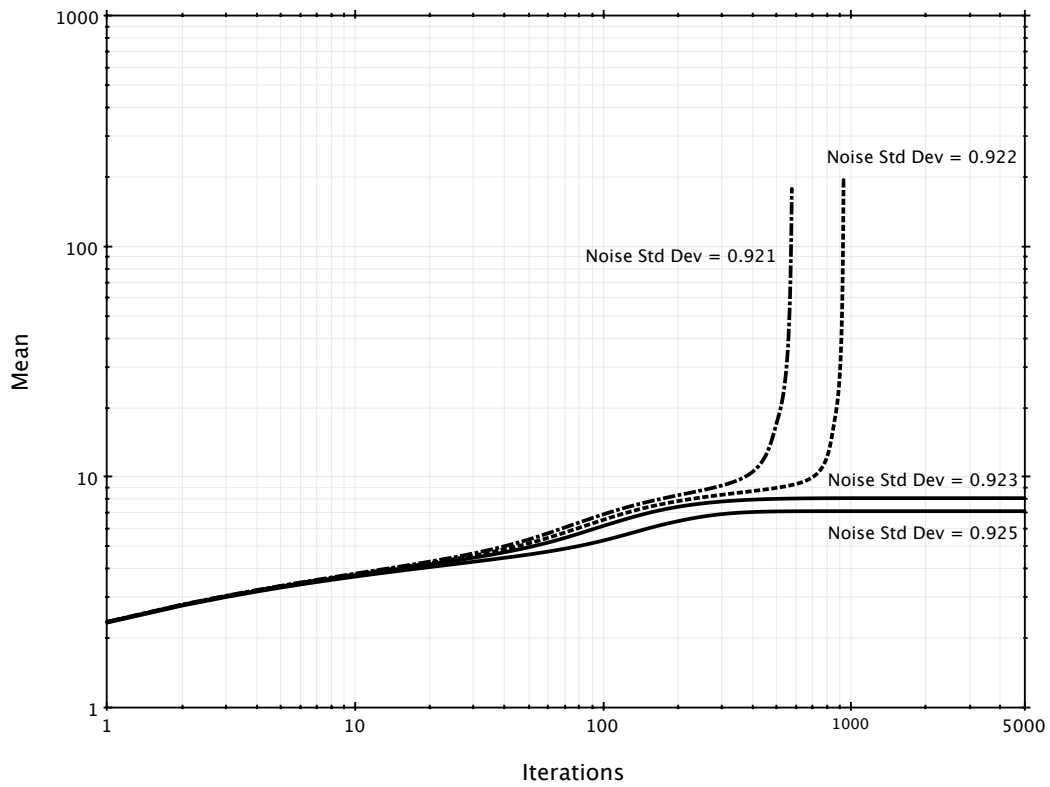
Figure 4.2: Density evolution using Gaussian approximation for a $d_v^{\max} = 8$ and $d_c^{max} = 7$ irregular LDPC ensemble on BIAWGN channel.

Fig. 4.1. For noise standard deviation $\sigma > 0.874$, the mean converges to a finite value whereas for $\sigma \leq 0.874$, it mean tends to ∞ with increasing iterations. For comparison, consider an irregular LDPC ensemble with same design rate from [58, 15] having profile:

$$\lambda(x) = 0.30013x + 0.28395x^2 + 0.41592x^7 \ ; \ \rho(x) = 0.22919x^5 + 0.77081x^6 \qquad (4.4)$$

This irregular profile has $d_v^{\max} = 8$ and $d_c^{\max} = 7$ and is used previously for simulation results shown in Chapter 3. Note that it has a better threshold than the regular $(3, 6)$ profile in Fig 4.2.

4.2 Analysis of QMF Factor Graphs

The above analysis model is applicable to QMF factor graphs. As discussed in Chapter 3, the BICM equivalent relay channel is BMS allowing the construction of an equivalent density evolution framework. Joint decoding for QMF relaying is based on factor graphs shown in Figs. 3.9-3.10 based on either the LDPC-LDPC or LDPC-LDGM configurations respectively. Both these configurations involve Q function nodes that model the QMF operation at relay. The number of Q nodes depends on the relay listening fraction f. The degree of Q nodes depends on the quantization levels. For given channel parameters, f is computed based on the network cut-set bound as discussed in Chapter 2. In Chapter 3, it is shown that a simple one-bit scalar quantizer performs reasonably well and yields a practical decoder. It is desirable to have a model for random QMF factor graphs where these parameters can be designed using the above considerations. This is achieved by defining some *fixed* graph elements i.e. fixed for all graph instances in the ensemble.

4.2.1 QMF Factor Graph Ensembles

In Sec. 4.1.2, random ensembles of Tanner graphs representing LDPC or LDGM codes are considered. A Tanner graph has only two types of nodes (variable V, check K) and one type of edge (VK). However, the corresponding factor graph also has observation nodes O and (VO) type edges. O nodes always have degree one and do not play an active role in the decoding. Including them in the analysis framework will lead to the same density evolution functions and thresholds as before.

For the sake of explanation, consider an ensemble for LDPC/LDGM factor graphs including these additional nodes and edges. Each instance in the ensemble has (VK) type edges drawn using a random permutation between sockets as defined previously. Additionally, the graph has *fixed* O nodes attached to every V node by (VO) type edges. This distinction between *random* edges/nodes and *fixed* edges/nodes allows the construction of complex graph ensembles.

QMF factor graphs have multiple classes of nodes and edges. Consider factor graphs

Variable Nodes				
Node Class	K-Degree	Q-Degree	O-degree	Neighbor Selection
V_S	random	random (0 or 1)	fixed	random (K_S)
V_R	random	random (0 or 1)	fixed	random (K_R)
Function Nodes				
Node Class	V_S-Degree	V_R-Degree	Neighbor Selection	
K_S	random	-	random	
K_R	-	random	random	
Q	fixed (always 1)	fixed (always 1)	random	
O_S	fixed	-	fixed	
O_R	-	fixed	fixed	

Table 4.1: Random LDPC-LDPC graph ensemble construction.

developed in Chapter 3 with one-bit scalar quantizer at relay. The LDPC-LDPC configuration has two classes of variable nodes $\{V_S, V_R\}$ representing the codewords at source $\mathbf{b}_S$ and relay $\mathbf{b}_R$. Similarly, the LDPC-LDGM configuration has variables $\{V_S, V_Q, V_R\}$ representing codeword at source $\mathbf{b}_S$, quantized bits at relay $\mathbf{b}_Q$ and codeword at relay $\mathbf{b}_R$ respectively. Each variable node is connected to multiple classes of function nodes (K, Q, O etc.) thereby having multiple types of degrees e.g. K-degree, Q-degree etc. The structures of these degrees are chosen to be fixed or random depending on the configuration. There are five classes of function nodes $\{K_S, K_R, Q, O_S, O_R\}$ namely, check constraints at source, at relay, quantization constraints and observations. This leads to many classes of edges. Depending on the ensemble configuration i.e. LPDC-LDPC or LDPC-LDGM, some of these classes are considered to be fixed and others to be random. The construction of random edges and selection of random node degrees is based on the same socket permutation procedure described previously.

1. **LDPC-LDPC Graph Ensemble:** Various classes of nodes and edges for the LDPC-LDPC configuration are shown in Fig. 4.3. Ensemble properties for various nodes are listed in Table 4.1. Note that the number and degree of Q nodes is fixed but their neighbors are selected at random. This leads to 4 classes of random edges $\{(V_S K_S), (V_R K_R), (V_S Q), (V_R Q)\}$. All other edge-classes are fixed. Degree profiles are considered only for 4 classes of nodes i.e. $\{V_S, V_R, K_S, K_R\}$, this is because degrees of other nodes are fixed.

2. **LDPC-LDGM Graph Ensemble:** Node classes for the LDPC-LDGM ensemble are shown in Fig. 4.4 with properties listed in Table 4.2. Similar to the LDPC-LDPC
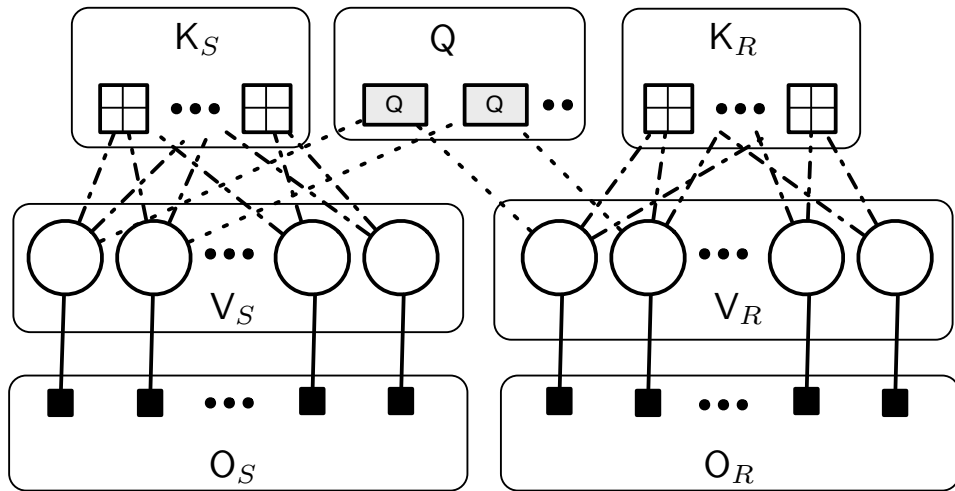
Figure 4.3: Various classes of nodes and edges for LDPC-LDPC configuration. Solid lines depict fixed edges in the random ensemble.
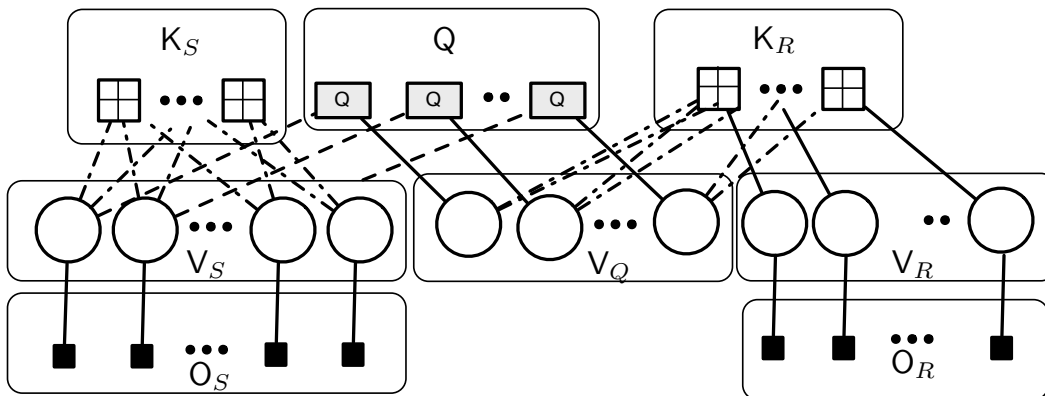

Figure 4.4: Various classes of nodes and edges for LDPC-LDGM configuration. Solid lines depict fixed edges in the random ensemble.

Variable Nodes				
Class	K-Degree	Q-Degree	O-degree	Neighbor Selection
V_S	random	random (0 or 1)	fixed	random (K_S)
V_Q	random	fixed (always 1)	-	random (K_R)
V_R	fixed (always 1)	-	fixed	fixed
Function Nodes				
Class	V_S-Degree	V_Q-Degree	V_R-Degree	Neighbor Selection
K_S	random	-	-	random
K_R	-	random	fixed (always 1)	random (V_Q)
Q	fixed	fixed	-	random for V_S, fixed for V_Q
O_S	fixed	-	-	fixed
O_R	-	-	fixed	fixed

Table 4.2: Random LDPC-LDGM graph ensemble construction.

case, degree profiles for 4 classes of nodes are considered $\{V_S, V_Q, K_S, K_R\}$ along with 4 classes of random edges $\{(V_S, K_S), (V_Q, K_R), (V_S, Q), (V_Q, Q)\}$.

Let $\lambda^K_{S,i,j}$ represent the total fraction of $(V_S K_S)$ edges that are connected to a V_S node having K-degree i and Q-degree j. Similarly, define $\lambda^K_{R,i,j}$ for $V_R(V_Q)$ nodes in the LDPC-LDPC(LDPC-LDGM) configurations respectively. It is possible to marginalize these profiles over the Q-degree and denote them as $(\tilde{\lambda}_S, \tilde{\rho}_S)$. The $(V_S Q)$ edges are distributed equally among V_S nodes of all degrees based on the listening fraction f.

$$\lambda^K_{S,i,0} = f\tilde{\lambda}_{S,i}$$
$$\lambda^K_{S,i,1} = (1-f)\tilde{\lambda}_{S,i}$$

Similarly, consider marginalized profiles for the relay mapping $(\tilde{\lambda}_R, \tilde{\rho}_R)$ between V_R (V_Q) and K_R nodes for the LDPC-LDPC (LDPC-LDGM) configuration.

$$\lambda^K_{R,i,0} = \begin{cases} f\tilde{\lambda}_{R,i} & \text{for LDPC-LDPC} \\ 0 & \text{for LDPC-LDGM} \end{cases}$$
$$\lambda^K_{R,i,1} = \begin{cases} (1-f)\tilde{\lambda}_{R,i} & \text{for LDPC-LDPC} \\ \tilde{\lambda}_{R,i} & \text{for LDPC-LDGM} \end{cases}$$

Check node profiles $\tilde{\rho}_{S,i}$ and $\tilde{\rho}_{R,i}$ have only one degree dimension because all check nodes have only one random neighbor as shown in Tables 4.1-4.2. It is also possible to compute $\lambda^Q_{S,i}$

representing the fraction of $\mathsf{V}_S\mathsf{Q}$ edges connected to a V_S node having K-degree i. (Q-degree is always 1 for $\mathsf{V}_S\mathsf{Q}$ edges)

$$\lambda^{\mathsf{Q}}_{S,i} = \frac{\lambda^{\mathsf{K}}_{S,i,1}}{i\sum_k \frac{\lambda^{\mathsf{K}}_{S,k,1}}{k}}$$

Similarly, $\lambda^{\mathsf{Q}}_{R,i}$ is computed as follows.

$$\lambda^{\mathsf{Q}}_{R,i} = \frac{\lambda^{\mathsf{K}}_{R,i,1}}{i\sum_k \frac{\lambda^{\mathsf{K}}_{R,k,1}}{k}}$$

4.2.2 Recursive Density Updates

Using the above model, recursive density updates for the QMF factor graph are derived. Both configurations have 4 classes of edges that play an active role in the decoding algorithm. Therefore, it is required to track the PDF of 4 messages at each iteration. To simplify the evolution, a Gaussian distribution for all the messages is assumed and 4 message means are tracked. For the LDPC-LDPC configuration, the evolution of $\mu_{\mathsf{K}_S\mathsf{V}_S}, \mu_{\mathsf{QV}_S}, \mu_{\mathsf{K}_R\mathsf{V}_R}, \mu_{\mathsf{QV}_R}$ is derived and shown below:

$$\mu^{(\ell+1)}_{\mathsf{K}_S\mathsf{V}_S} = \sum_k \tilde{\rho}_{S,k}\phi^{-1}\left(1 - \left(1 - \sum_{i,j}\lambda^{\mathsf{K}}_{S,i,j}\phi\left(\mu_{\mathsf{V}_S} + (i-1)\mu^{(\ell)}_{\mathsf{K}_S\mathsf{V}_S} + j\mu^{(\ell)}_{\mathsf{QV}_S}\right)\right)^{k-1}\right)$$

$$\mu^{(\ell+1)}_{\mathsf{QV}_S} = \phi^{-1}\left(1 - \tanh\left(\frac{L}{2}\right)\left(1 - \sum_i \lambda^{\mathsf{Q}}_{R,i}\phi\left(\mu^{(\ell)}_{\mathsf{K}_R\mathsf{V}_R}\right)\right)\right)$$

$$\mu^{(\ell+1)}_{\mathsf{K}_R\mathsf{V}_R} = \sum_k \tilde{\rho}_{R,j}\phi^{-1}\left(1 - \left(1 - \sum_{i,j}\lambda^{\mathsf{K}}_{R,i,j}\phi\left((i-1)\mu^{(\ell)}_{\mathsf{K}_R\mathsf{V}_R} + j\mu_{\mathsf{QV}_R}\right)\right)^{k-1}(1 - \phi(\mu_{\mathsf{V}_R}))\right)$$

$$\mu^{(\ell+1)}_{\mathsf{QV}_R} = \phi^{-1}\left(1 - \tanh\left(\frac{L}{2}\right)\left(1 - \sum_i \lambda^{\mathsf{Q}}_{S,i}\phi\left(\mu_{\mathsf{V}_S} + i\mu^{(\ell)}_{\mathsf{K}_S\mathsf{V}_S}\right)\right)\right)$$

Here μ_{V_S} and μ_{V_R} represent the mean LLR for observations from source and relay respectively. L is defined as follows:

$$L := \log\left(\frac{1 - Q(\sqrt{\mathsf{SNR}_{SR}})}{Q(\sqrt{\mathsf{SNR}_{SR}})}\right)$$

The updates for LDPC-LDGM are obtained by replacing $\mu_{\mathsf{K}_R\mathsf{V}_R}$ and μ_{QV_R} with $\mu_{\mathsf{K}_R\mathsf{V}_Q}$ and μ_{QV_Q} respectively in the above.

4.2.3 Initial Conditions

As in the point-to-point case, initial means i.e. μ_{V_S} and μ_{V_R} are required to compute the GABPT. Without loss of generality, it is assumed that the all-zero codeword is transmitted from the source. However, it cannot be assumed that the relay also transmits the all-zero codeword because the source to relay channel is noisy.

This creates a roadblock in the analysis of LDPC-LDPC graphs preventing further progress. It is required to determine the typical codeword $\mathbf{b}_R$ transmitted by the relay. This depends on (a) statistics of the noisy channel between source and relay and (b) the characteristics of the LDPC encoder at relay. Encoding of linear block codes is based on the generator matrix. Given just a degree profile for an LDPC ensemble it is difficult to draw insights into the structure of generator matrices for the codes in the ensemble.

The LDPC-LDGM configuration does not present a roadblock. LDPC-LDGM ensembles allow characterization of the typical sparse generator matrix used for encoding at relay. Conditioned on $\mathbf{b}_S = \mathbf{0}$ and a given value of SNR_{SR}, there is a *typical* sequence $\mathbf{b}_Q$ that is mapped to a *typical* $\mathbf{b}_R$ based on the LDGM code. The probability of occurrence for atypical codewords vanishes as the block length becomes large and is ignored for the ensemble average performance. A typical $\mathbf{b}_Q$ comprises of $fN_S(1 - p_f)$ 0's and $fN_S p_f$ 1's (p_f is defined in Sec. 3.3.8). For a given marginalized relay degree profile $\left(\tilde{\lambda}_R, \tilde{\rho}_R\right)$, each bit of the typical $\mathbf{b}_R$ is i.i.d. Bernoulli(q), where q depends on the probability of having odd number of 1's in a column of the relay generator matrix (drawn randomly from the LDPC-LDGM QMF ensemble).

$$q = \sum_j \left(\frac{\tilde{\rho}_{R,j}/j}{\sum_i \tilde{\rho}_{R,i}/i} \right) \frac{1 - (1 - 2p_f)^j}{2}$$

μ_{V_R} is the marginal density of the LLR of the relay to destination channel under the marginal law that $\mathbf{b}_R$ is i.i.d. Bernoulli(q). Using this μ_{V_R} the probability of error in decoding of $\mathbf{b}_S$ and associated thresholds for LDPC-LDGM ensembles can be computed.

4.2.4 Simulation Experiments

In this subsection, the analysis framework developed above is validated using channel simulation experiments. For the LDPC code at source, consider marginalized degree profiles $(\tilde{\lambda}_S, \tilde{\rho}_S)$ as given in Eq. (4.4). For the LDGM relay mapping, consider $(6, 3)$ and $(8, 4)$ regular profiles. Both ensembles have matching design rates and correspond to a relay listening fraction of $f = 1/3$. Channel parameters are defined similarly as in Sec. 3.5 with $q_{SR} = 10, q_{RD} = 1$. GABPT (SNR_{SD}) for these LDPC-LDGM ensembles are listed in Table. 4.3. They are calculated based on the condition that mean of messages to V_S nodes tends to ∞ with increasing iterations i.e. assuming $\mathbf{b}_S = \mathbf{0}$. The evolution of means is shown in Fig. 4.5 for one of the ensembles at the GABPT. Note that at around 300 iterations the

LDPC Profile	LDGM Profile	GABPT (dB)
$\tilde{\lambda}_S(x) = 0.30013x + 0.28395x^2 + 0.41592x^7$; $\tilde{\rho}_S(x) = 0.22919x^5 + 0.77081x^6$	$\tilde{\lambda}_R(x) = x^5$; $\tilde{\rho}_R(x) = x^2$	-1.81
	$\tilde{\lambda}_R(x) = x^7$; $\tilde{\rho}_R(x) = x^3$	-2.04

Table 4.3: Thresholds (GABPT) for selected LDPC-LDGM ensembles.
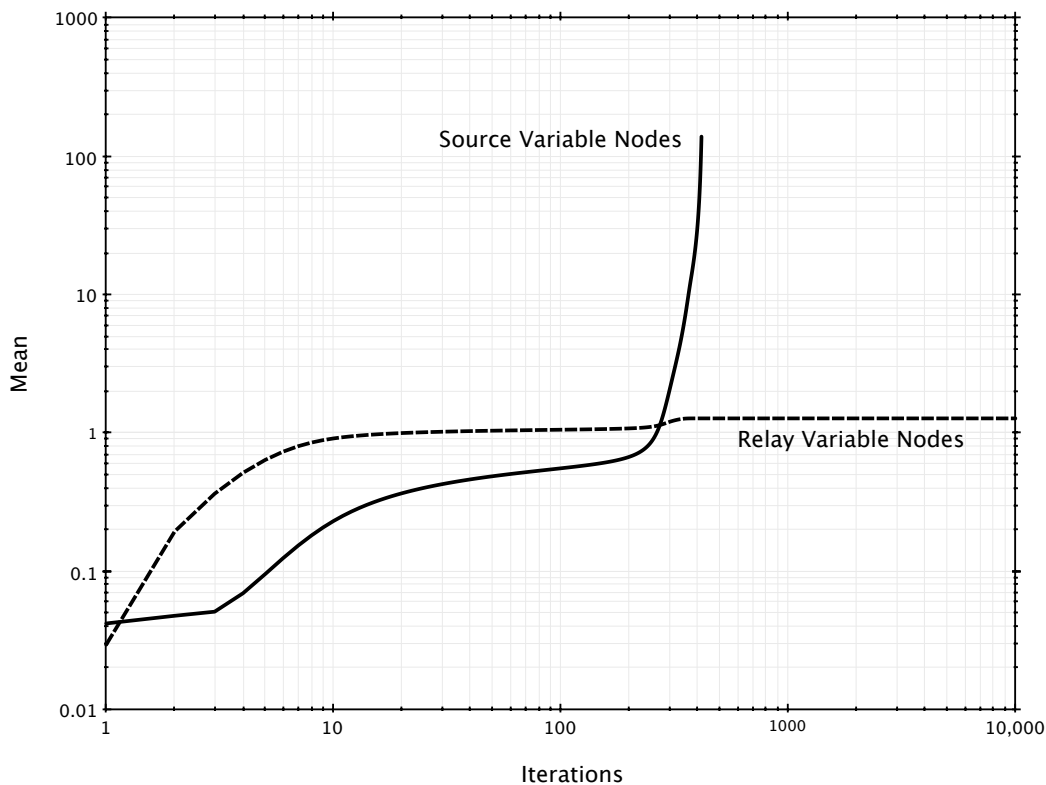

Figure 4.5: Evolution of $\mu_{\mathsf{K}_S \mathsf{V}_S}$ and $\mu_{\mathsf{K}_R \mathsf{V}_Q}$ for LDPC-LDGM ensemble in first row of Table 4.3 at $\mathsf{SNR}_{SD} = -1.81$dB.
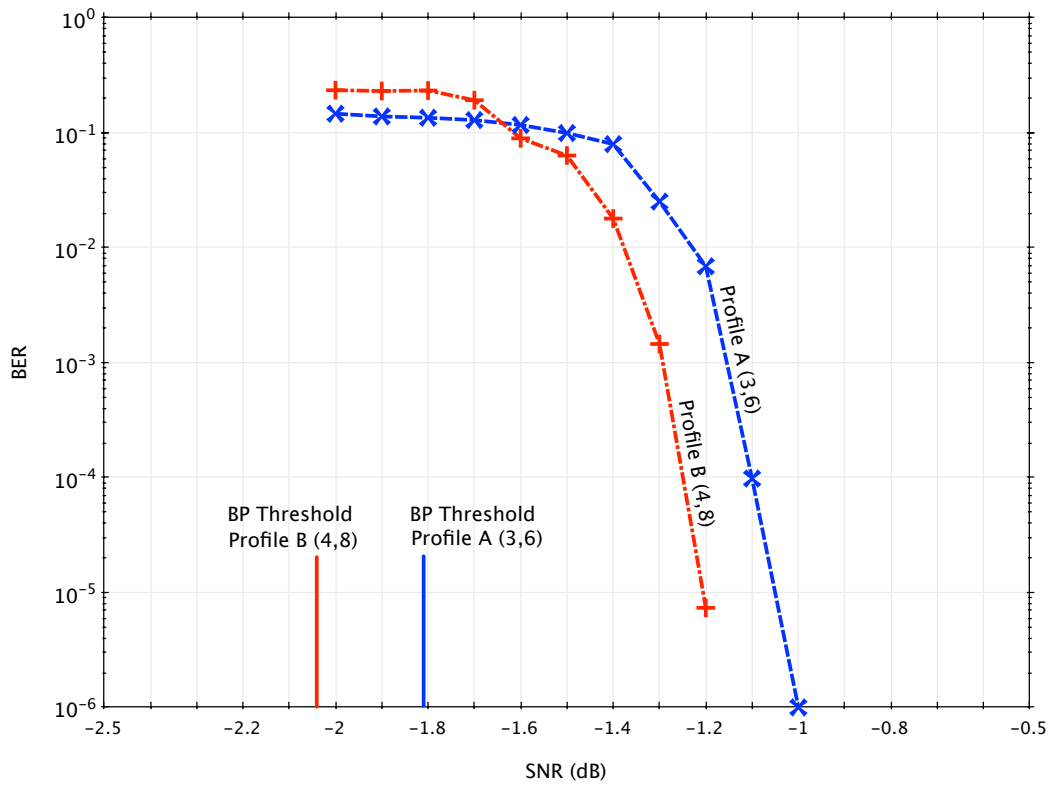
Figure 4.6: GABPT of two LDPC-LDGM ensembles shown in Table 4.3 and BER performance simulation over a BIAWGN channel.
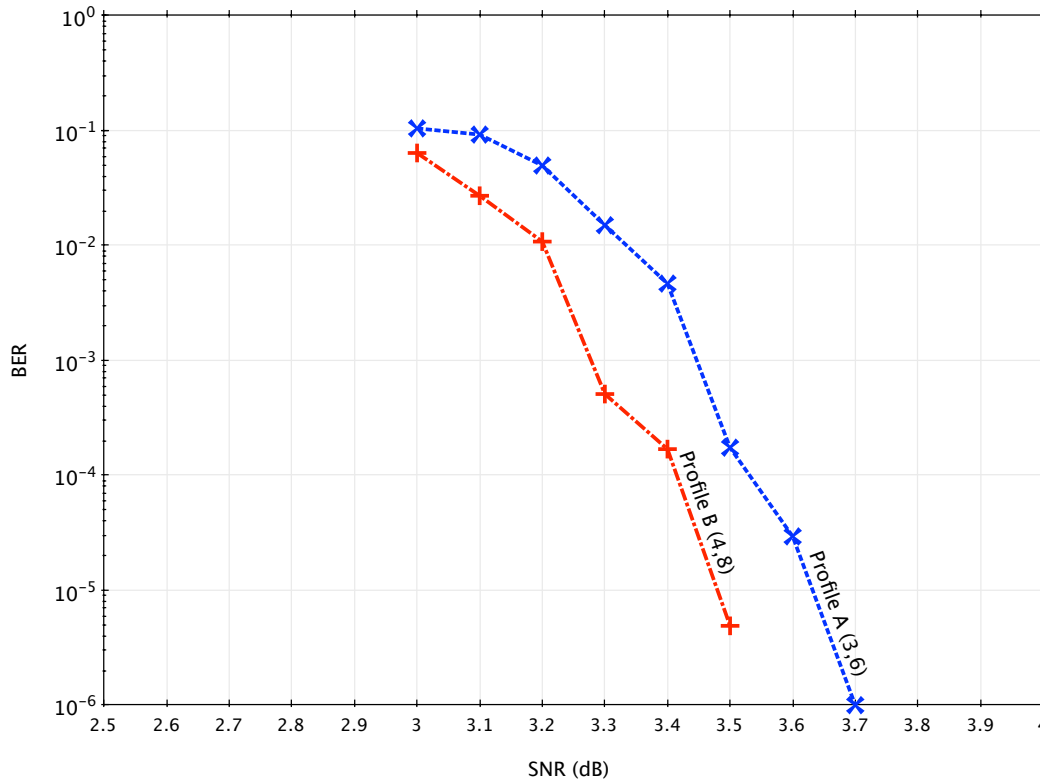
Figure 4.7: BER performance comparison of two LDPC-LDGM ensembles shown in Table 4.3 over an AWGN channel using 16QAM modulation and BICM.

message mean for V_S grows rapidly towards ∞. At the same time message mean for V_Q nodes stabilizes to a finite value. This is because the typical $\mathbf{b}_Q \neq \mathbf{0}$, as discussed previously.

In Fig. 4.6, the BER performance of graph instances from these two ensembles is shown for the BIAWGN relay channel. The block length is chosen to be $N_S = 20400$. Cycles of length 2 and 4 are avoided. The iterative sum product algorithm is run for 50 iterations each time. The calculated thresholds reasonably predict the BER performance of the ensembles.

4.2.5 BICM Equivalent Channel

In the examples above, GABPT is computed for LDPC-LDGM ensembles over the BIAWGN channel. The next step is to consider the BICM equivalent BMS channel (as discussed in Sec. 3.4). The Gaussian approximation to density evolution assumes a Gaussian distribution for all messages. In this computation the channel model plays a role only during the computation of initial conditions. In order to extend the above analysis for BICM equivalent BMS channels, these initial conditions need to be derived. Note that the goal of this analysis is to produce a *comparative ordering* of ensembles by GABPT. Such an ordering allows the

search of profiles that have good thresholds. For this purpose, the absolute value of GABPT is not important. This leads to the following claim.

Claim 1. *Consider an ordering of LDPC-LDGM ensembles based on GABPT calculated over the BICM equivalent BMS relay channel. If equivalent GABPTs are computed over the BIAWGN relay channel, the ordering is retained.*

Consider a point-to-point Gaussian channel and its corresponding BICM equivalent BMS channel. Let the underlying Gaussian channel have signal-to-noise ratio $\mathsf{SNR}_{\mathrm{AWGN}}$. For a fixed modulation order, the BMS channel is also parameterized by $\mathsf{SNR}_{\mathrm{AWGN}}$ and ordered by physical degradation. Note that the BICM BMS channel is not Gaussian. Alternatively consider a BIAWGN channel having parameter $\mathsf{SNR}_{\mathrm{BIAWGN}}$. The parameter is chosen such that this channel has the same mean LLR as the BMS channel above, conditioned on the all-zero codeword. Note that $\mathsf{SNR}_{\mathrm{BIAWGN}}$ varies monotonically with $\mathsf{SNR}_{\mathrm{AWGN}}$.

The above argument also applies to the BICM equivalent BMS relay channel. SNR_{SD} is used to parametrize the relay channel:

$$\mathsf{SNR}_{RD} = q_{RD}\mathsf{SNR}_{SD}, \mathsf{SNR}_{SR} = q_{SR}\mathsf{SNR}_{SD}$$

For given $q_{RD}, q_{SD} > 0$ terms and modulation order, this BMS channel is ordered by physical degradation using SNR_{SD}. An equivalent BIAWGN relay channel is defined with same scaling factors and parameter $\mathsf{SNR}_{SD}^{\mathrm{BIAWGN}}$ which is monotonically increasing in SNR_{SD}. Since the mapping between SNR_{SD} and $\mathsf{SNR}_{SD}^{BIAWGN}$ is monotonic, BIAWGN thresholds can be used to order factor graph ensembles by comparative performance.

Fig. 4.7 shows BER performance of the ensembles in Table 4.3 using BICM and 16QAM modulation. As shown, the ensemble with $(4, 8)$ regular LDGM mapping performs better as predicted by BIAWGN GABPTs.

In the next section, above techniques are used to design example cooperative links. For given relay channel parameters, density evolution (assuming BIAWGN channels) is used to search for degree profiles that have the best GABPT. Procedures to select modulation order and relay listening fraction are also outlined.

4.3   Link Design Examples

Consider a Gaussian (1,1,2) relay channel with system parameters defined previously $q_{RD} = 1, q_{SR} = 10$, i.e.

$$\mathsf{SNR}_{SD} = \mathsf{SNR}_{RD}, \quad \mathsf{SNR}_{SR} = 10 \times \mathsf{SNR}_{SD} \tag{4.5}$$

In previous simulation experiments, modulation order and relay listening fraction f were chosen arbitrarily. In this section, selection of these parameters is performed using information theoretic guidelines. As discussed previously, for QMF the listening-time fraction f at
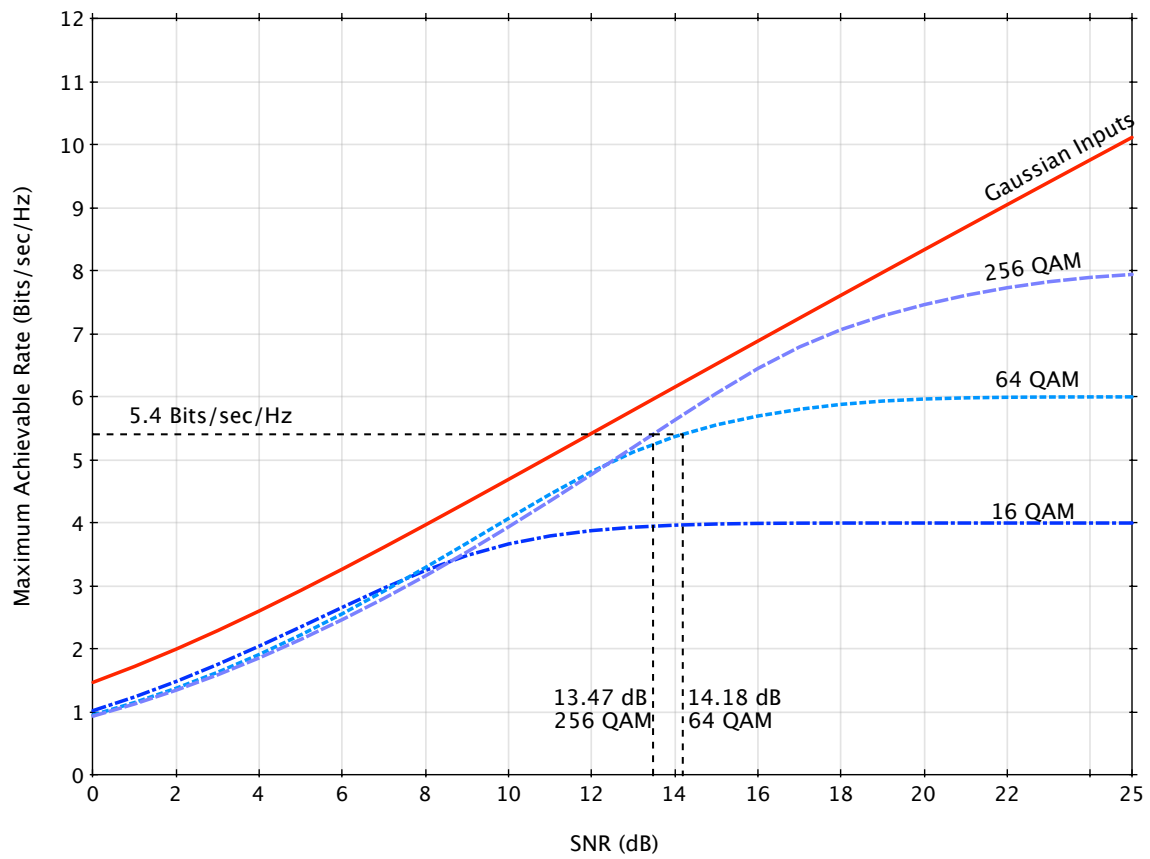
Figure 4.8: Maximum achievable rate for QMF relaying with modulation constraints on channel inputs plotted vs SNR_{SD}. $q_{SR} = 10, q_{RD} = 1$
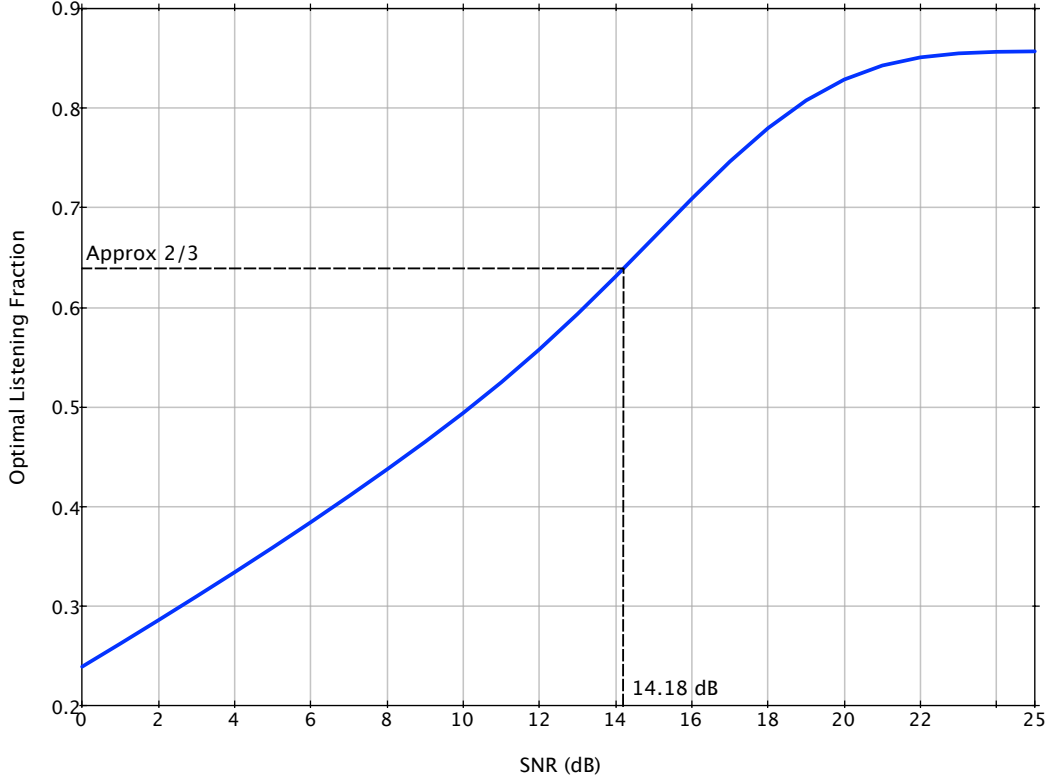
Figure 4.9: Optimal listening fraction plotted with respect to SNR_{SD}. $q_{SR} = 10, q_{RD} = 1$

R is independently optimized to maximize system throughput [75, 46, 53]. The optimal f^* is found by balancing the two terms in the minimization of (3.7):

$$f^* = \frac{\mathfrak{C}_{\mathrm{G}}\left(\mathsf{SNR}_{RD}\right)}{1 + \mathfrak{C}_{\mathrm{G}}\left(\mathsf{SNR}_{RD}\right) + \mathfrak{C}_{\mathrm{G}}\left(\frac{\mathsf{SNR}_{SR}}{2} + \mathsf{SNR}_{SD}\right) - \mathfrak{C}_{\mathrm{G}}\left(\mathsf{SNR}_{SD}\right)}$$

Equivalent optimal fractions are computed when modulation constraints and BICM are considered as in Eq. (3.8). Fig. 4.8 shows a plot of the information theoretic QMF achievable rate upper bound for Gaussian, 16QAM, 64QAM and 256QAM constrained channel inputs (with BICM) as computed in Eqs.(3.7) and (3.8). For each point in the plots, an optimized listening fraction f^* is used to compute the rate.

Consider the design of a cooperative link that achieves an information throughput of 5.4bits/sec/Hz. Practical choices for modulation order using BICM can be made using Fig. 4.8 as shown. Since the required SNR_{SD} to achieve this rate is comparable for 64QAM and 256QAM inputs, let us choose the simpler scheme with 64QAM. This gives a fundamental information theoretic threshold for the link i.e. $\mathsf{SNR}_{SD} = 14.18$dB. A plot of the optimal listening fraction for 64QAM inputs is shown in Fig. 4.9. As shown, the optimal fraction
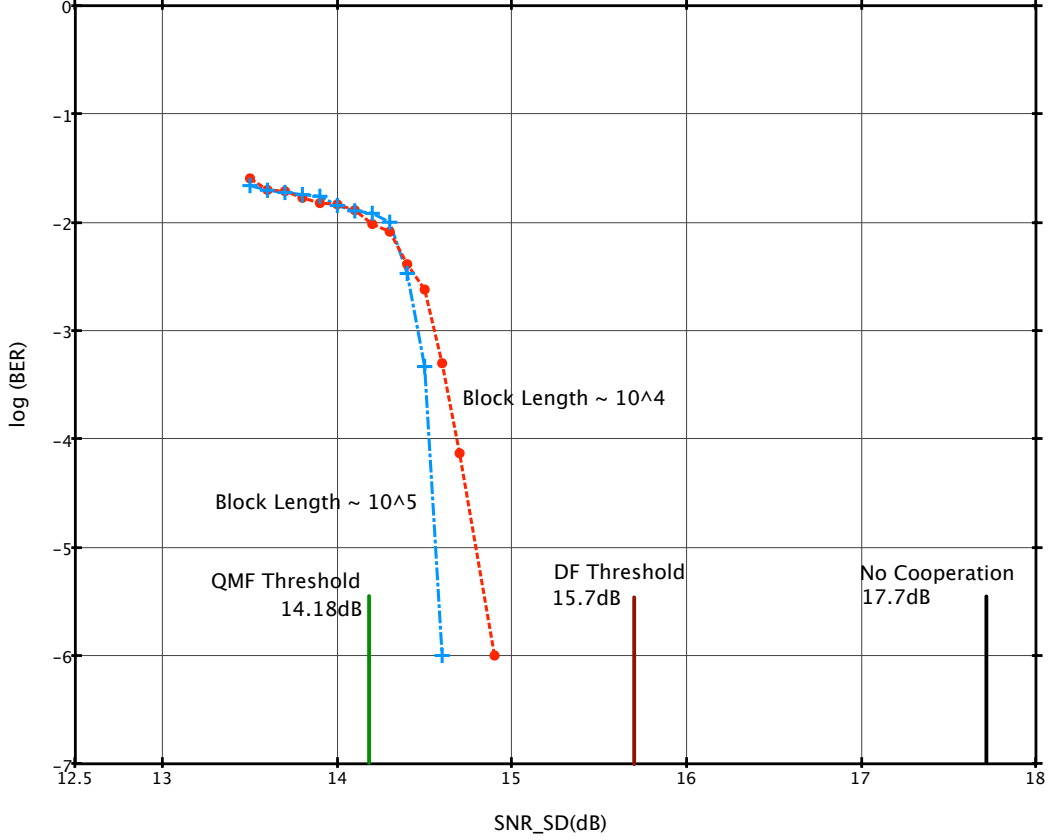
Figure 4.10: BER simulation for $\mathbf{b}_S$ using design rate of 5.4bits/sec/Hz with 64QAM. $q_{SR} = 10, q_{RD} = 1$

corresponding to $\mathsf{SNR}_{SD} = 14.18$dB is $f^* \approx 2/3$.

64QAM supports at most 6 coded bits per symbol. To achieve an information rate of 5.4bits/sec/Hz, the source should use an LDPC design rate $r_{LDPC} = \frac{5.4}{6} = 0.9$. The optimal listening fraction determines the LDGM design rate

$$r_{LDGM} = \frac{1-f}{f} \approx \frac{1}{2}$$

4.3.1 Profile Design

An LDPC-LDGM ensemble optimized for the above system parameters is designed using density evolution tools. In order to reduce the computational complexity of profile optimization, the Gaussian approximation to density evolution is used as developed in [15]. Additionally, the following heuristics are used to reduce the search space of profiles.
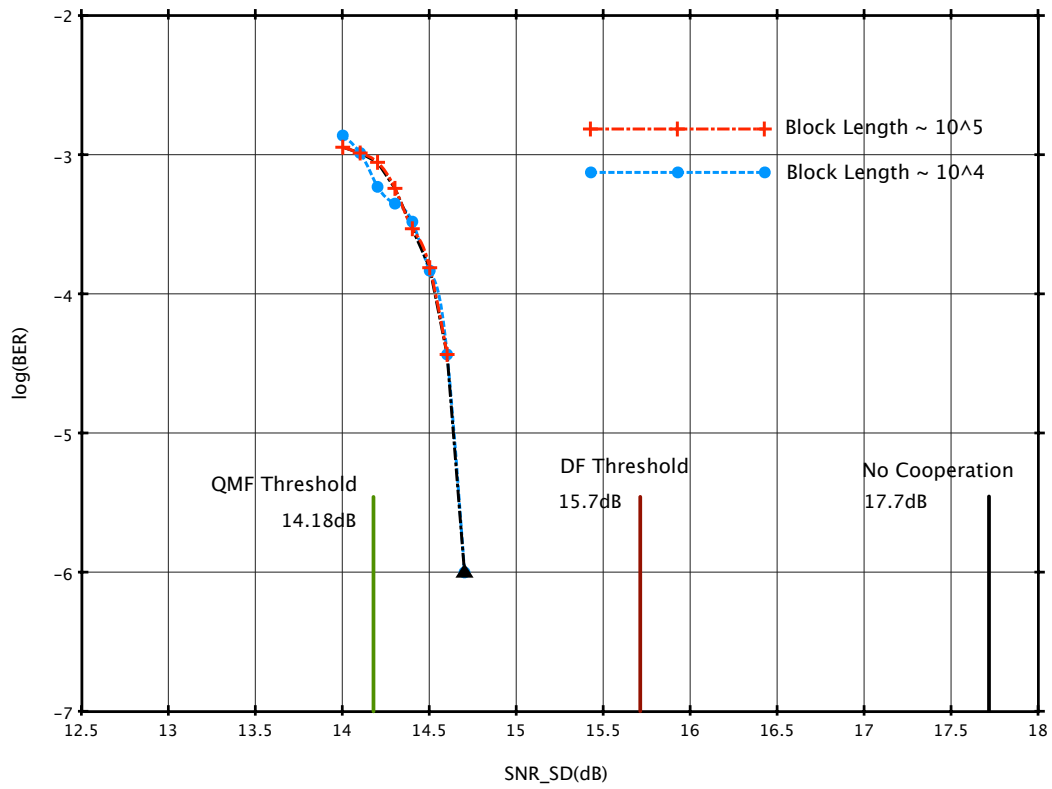
Figure 4.11: BER simulation for $\mathbf{b}_R$ using design rate of 5.4bits/sec/Hz with 64QAM. $q_{SR} = 10, q_{RD} = 1$

1. For the LDPC code, check degree profiles are constrained to be concentrated [15] i.e. all check degrees (from edge perspective) are either k or $k + 1$ for some integer $k \geq 2$.

2. For the LDPC code: variable degree profiles have maximum degree of $d_v^{\max} = 8$.

3. For the relay mapping: search is limited to regular LDGM profiles.

Profile optimization techniques based on linear programming proposed in [14, 58] cannot be readily adapted to the multi-dimensional density evolution of QMF factor graphs. Therefore, a coarse brute force optimization over all possible degree profiles using the above heuristics is used to find the following degree profile:

$$\tilde{\lambda}_S(x) = 0.28x + 0.32x^2 + 0.28x^3 + 0.12x^6 + 0.0009x^7$$
$$\tilde{\rho}_S(x) = 0.04x^{28} + 0.96x^{29}$$
$$\tilde{\lambda}_R(x) = x^4, \tilde{\rho}_R(x) = x^9$$

Simulation results for BER in decoding of $\mathbf{b}_S$ using two graph instances (with block lengths $N_S = 20400$ and $N_S = 204000$ respectively) drawn from above profiles are shown in Fig. 4.10. The simulation models BICM with 64QAM modulation, one bit scalar quantizer and an ideal interleaving. 50 iterations of sum product decoding are used at every step. As shown, the BER performance is within $\leq$ 1dB of the QMF threshold. For a network with one relay, information theoretic thresholds for QMF and CF are identical. But unlike CF, QMF does not require forward channel knowledge at relay and remains within bounded gap to capacity for networks with more relays. As references for comparison, equivalent[1] information theoretic thresholds for DF and the no-cooperation case are shown. For this example, QMF is 1.5dB better than DF and 3.5dB better than no-cooperation.

For the DBLAST architecture, $\mathbf{b}_R$ must also be reliably decoded at or below the target SNR (for successive interference cancellation to work). Fig. 4.11 shows the BER for $\mathbf{b}_R$ which is also within $\leq$ 1dB of the QMF threshold for both of the block-lengths considered.

4.3.2 Log Scaled Model

In Chapter 2, an alternative model is introduced where the channel parameters $\mathsf{SNR}_{SD}, \mathsf{SNR}_{RD}$ and SNR_{SR} are scaled logarithmically using proximity gain η.

$$\mathsf{SNR}_{RD} = \mathsf{SNR}_{SD}$$
$$\mathsf{SNR}_{SR} = (\mathsf{SNR}_{SD})^{\eta}$$

---

[1]DF threshold is computed with BICM and 64QAM inputs. f is optimized so as to maximize throughput for the DF scheme.
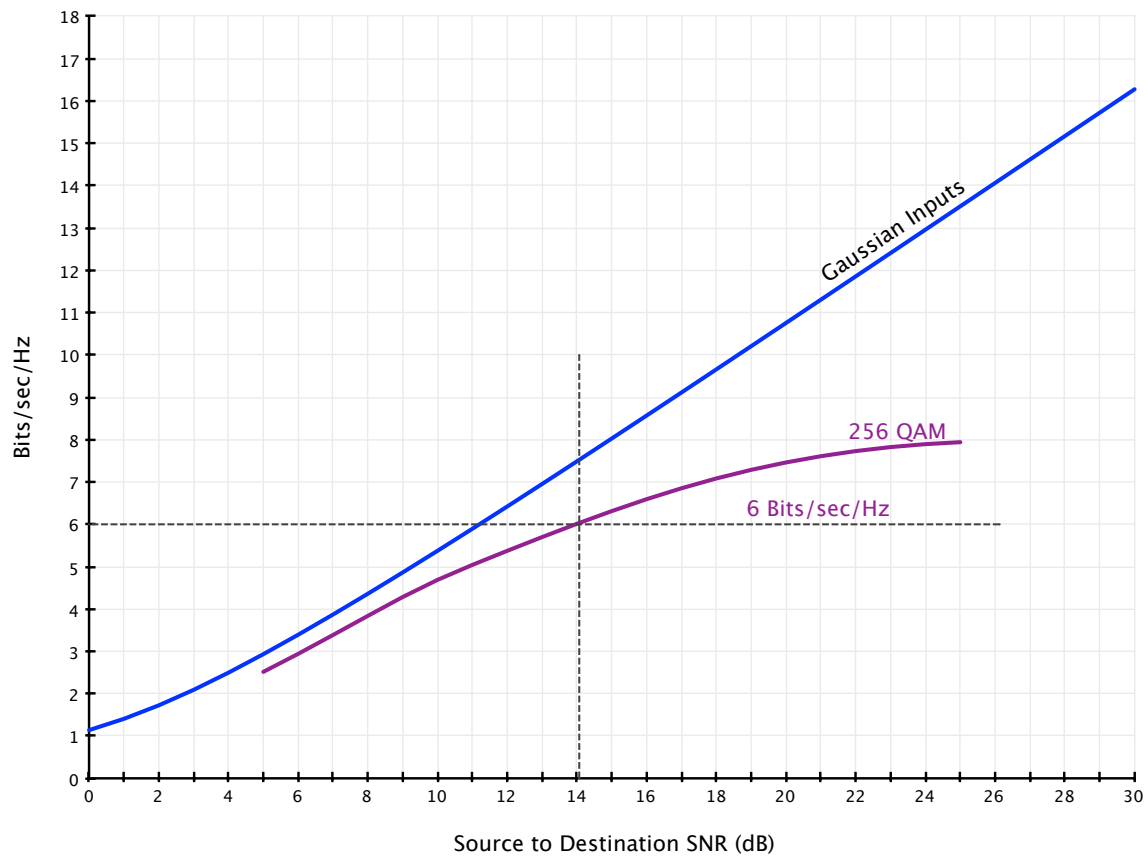
Figure 4.12: Maximum achievable rate for QMF relaying with modulation constraints on channel inputs plotted vs SNR_{SD}. $\eta = 3$
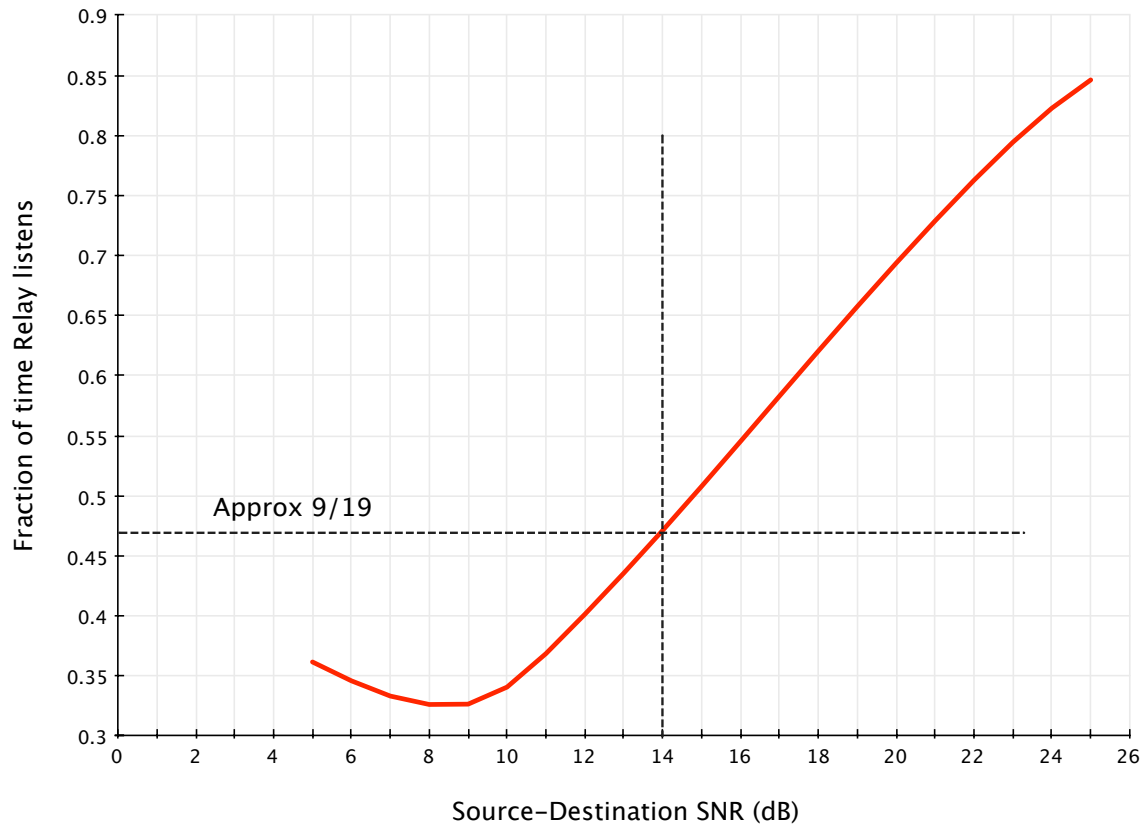
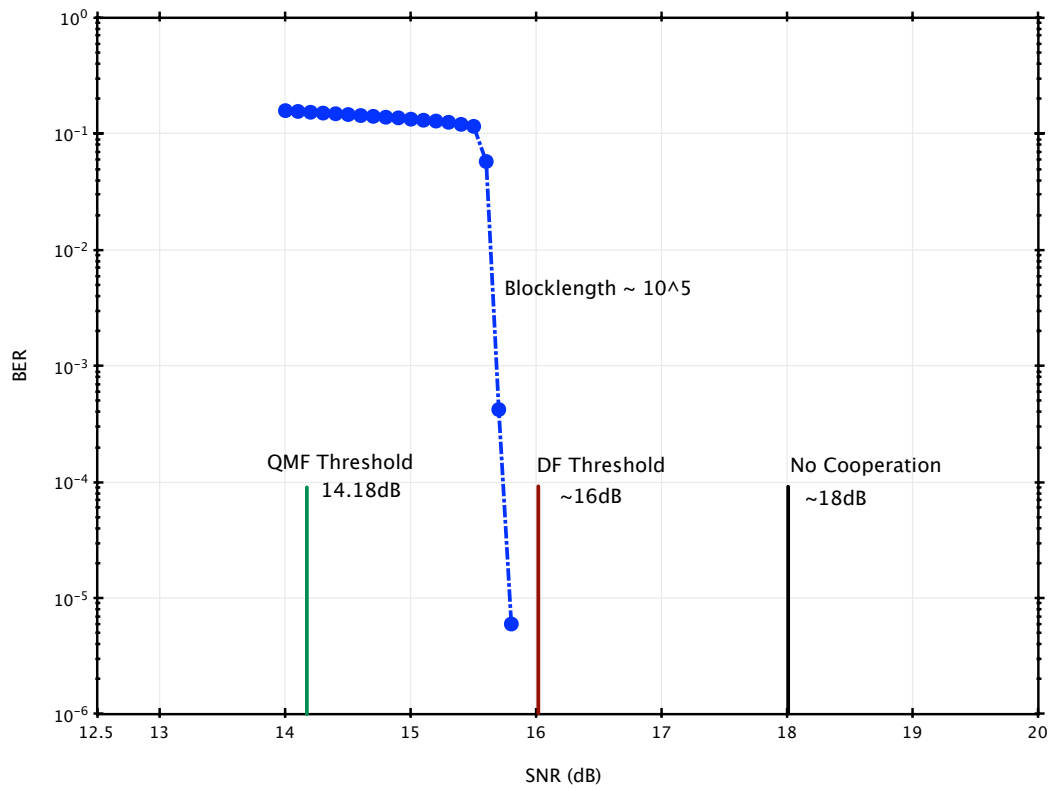Figure 4.13: Optimal listening fraction plotted with respect to SNR_{SD}. $\eta = 3$

Figure 4.14: BER simulation for $\mathbf{b}_S$ using design rate of 6bits/sec/Hz with 256QAM. $\eta = 3$

The previous design example used a linear scaling model:

$$q_{SR} = \frac{\mathsf{SNR}_{SR}}{\mathsf{SNR}_{SD}} = \frac{24.18\text{dB}}{14.18\text{dB}} = 10$$

Let us consider a design example using the logarithmic model with $\eta = 3$. Under this scaling, for $\mathsf{SNR}_{SD} = 14.18$ the channel can support 6bits/sec/Hz using 256QAM as shown in Fig. 4.12. The optimal listening fraction is $f^* = \frac{9}{19}$ as shown in Fig. 4.13. The design rate parameters are calculated as $r_{LDPC} = 0.75$ and $r_{LDGM} = \frac{1-f}{f} = \frac{10}{9} = 1.11$. To simplify design of code profiles, approximate choices are made: $f = 0.5$ and $r_{LDGM} = 1$. The following optimized degree profile is found for these parameters:

$$\tilde{\lambda}_S(x) = 0.34x + 0.25x^2 + 0.15x^3 + 0.23x^5 + 0.0066x^6$$
$$\tilde{\rho}_S(x) = x^{11}$$
$$\tilde{\lambda}_R(x) = x^9, \tilde{\rho}_R(x) = x^9$$

BER simulations for $\mathbf{b}_S$ using $N_S = 209000$ are shown in Fig. 4.14. The performance is within ≈ 1.5dB of the fundamental threshold.

Chapter 5

Conclusion

5.1 Summary of results

In this dissertation, an improved understanding of cooperative relaying for wireless networks is presented. Based on the fundamental DMT of the $(1, 1, 2)$ relay configuration, it is shown that cooperation improves the throughput of wireless links by providing additional degrees of freedom.

The QMF cooperation scheme introduced by Avestimehr et. al. in 2009 promises to achieve the above cooperation gains in a fashion that scales with the number of cooperating terminals in the network. However, it presents unique challenges in terms of designing schemes for efficient coding and decoding. With a focus the single relay network, a coding framework that brings QMF cooperation closer to practical implementation is proposed. It is shown that the framework has linear complexity encoding, mapping and joint decoding operations.

The proposed decoder is based on a sparse factor graph. Using suitable choice of codes, space time coding strategies and quantization procedures it is shown that the decoding algorithm can be reduced to belief propagation over a Tanner graph. Simulation results validate that the proposed scheme can achieve cooperation gains as predicted by information theory.

Finally, tools based on density evolution techniques are presented for systematic analysis and design of factor graphs. The design of profiles that perform within $0.5 - 1.5$dB of fundamental limits is demonstrated.

The maximum achievable information throughput for the $(1, 1, 2)$ relay channel with asymmetric network model (Ch.2) is plotted in Fig. 5.1. For comparison, the point-to-point AWGN Shannon capacity is also shown. Two points are shown corresponding to performance of link designs presented in Chapter 4.

This dissertation focuses on a network with one relay. However, the principles can be applied to networks with multiple relays leading to possibly an order-of-magnitude improve-
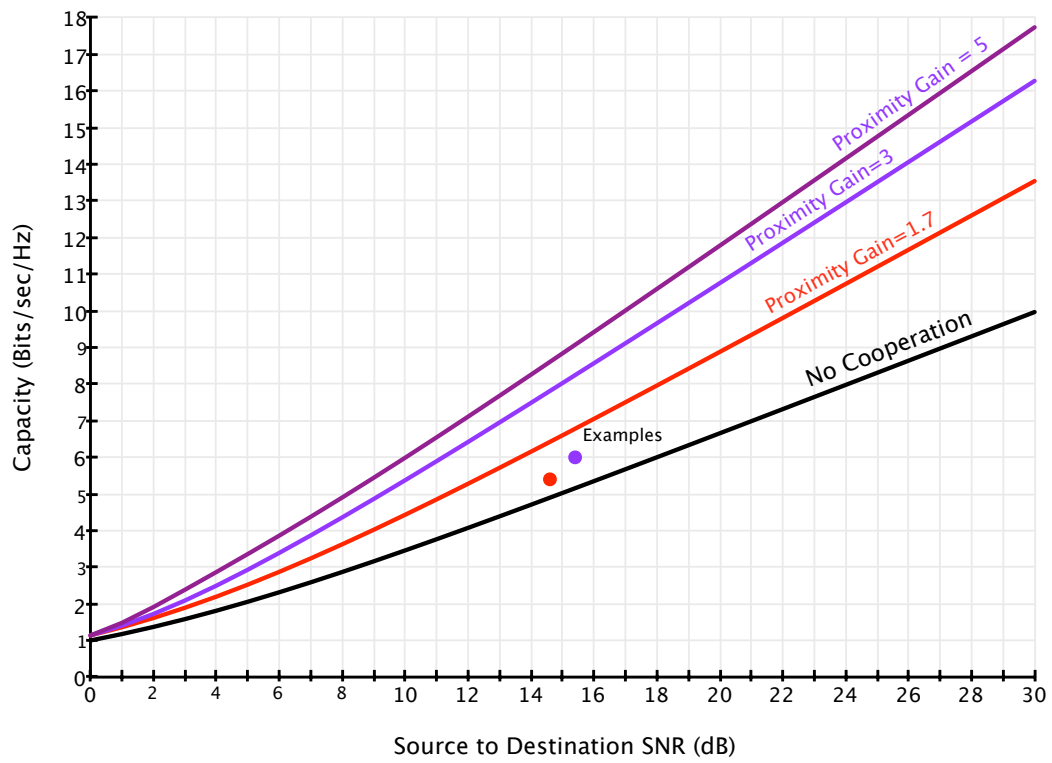
Figure 5.1: $(1, 1, 2)$ Relay channel with QMF relaying. Throughput scaling for various proximity gains.

ment in spectral efficiency allowing future cellular and local area networks to deliver improved user experience.

The principles of cooperative communication studied in this dissertation can also be applied to other scenarios. As an example, NASA's Deep Space Network (DSN) uses large ground based antenna arrays to communicate with deep space probes. Intermediate probes or space stations can be used as relays to improve the range and capacity of these deep space communication links.

5.2   Future Directions

5.2.1   Improved Profiles

Code design examples reported in this work use the Gaussian approximation to density evolution. This approximation is poor for graphs that have function nodes with small degrees [2] like Q nodes. Precise density evolution for BICM equivalent BMS channels can be used to address this.

For link design examples, a brute force optimization technique is used. This procedure requires several simplifying heuristics and does not explore the available design space exhaustively. The study of optimization techniques like differential evolution [64] may lead to better code constructions.

5.2.2   Rate Adapation

Optimizing code design for every instantaneous channel realization is not feasible in practice. Most systems use rate adaptation mechanisms wherein terminals switch between a small set of characteristic codes and constellations as channel conditions vary.

A large set of channel parameters must be considered for cooperative relay networks e.g. a single relay network has 3 SNR parameters. Design of efficient rate adaptation schedules with many channel parameters is a challenging problem.

A possible direction is to develop cooperative rate-less techniques based on schemes such as Fountain or Tornado codes.

Another alternative is to incorporate modern rate adaptation mechanisms like hybrid automatic repeat request (hybrid ARQ) into the proposed framework.

5.2.3   Multiple Relays

As discussed previously, a major advantage of the QMF scheme is that it can perform within bounded gap to fundamental limits for arbitrary relay networks. An important direction for future work is to develop practical schemes for networks with multiple relays.

Consider a network having one source, one destination and N half duplex relays. Several interesting problems arise in the study of this topology. Optimal listening fractions for each relay are determined based on channel conditions at all other relays. This requires a large overhead for sharing channel knowledge among terminals. Practical techniques to minimize such overhead are required. Furthermore, while one relay listens several other terminals could be transmitting. Signal processing techniques to appropriately handle situations like this need to be studied.

5.2.4 Practical System Design Constraints

Communication systems have several components. Frequency and timing offsets between terminals need to be estimated and corrected, channel coefficients must be estimated using pilot symbols etc.

The use of physical layer cooperation increases the complexity of all these operations. It also imposes tighter constraints on several related system parameters. Further study is required to manage this complexity.

5.2.5 Unified Cooperation Framework

It has been shown that cooperation [51] can achieve optimal scaling of capacity in wireless networks. Relaying is just one of the several proposed cooperative communication techniques. Other approaches like cooperative interference alignment and cancellation are equally promising. A long term research vision is to bring all these techniques closer to implementation and design a unified framework for practical cooperative wireless networks.

Bibliography

[1] S.M. Aji and R.J. McEliece. The generalized distributive law. *Information Theory, IEEE Transactions on*, 46(2):325 –343, mar 2000.

[2] M. Ardakani and F.R. Kschischang. A more accurate one-dimensional analysis and design of irregular ldpc codes. *Communications, IEEE Transactions on*, 52(12):2106 – 2114, dec. 2004.

[3] Amir Salman Avestimehr. *Wireless network information flow: a deterministic approach.* PhD thesis, EECS Department, University of California, Berkeley, Oct 2008.

[4] A.S. Avestimehr, S.N. Diggavi, and D. Tse. A deterministic approach to wireless relay networks. In *Proc. Forty-Fifth Allerton Conf. Commun. Contr. Comput.*, Illinois, 2007.

[5] A.S. Avestimehr, S.N. Diggavi, and D. Tse. Approximate capacity of gaussian relay networks. *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 474–478, July 2008.

[6] A.S. Avestimehr, S.N. Diggavi, and D.N.C. Tse. Wireless network information flow: A deterministic approach. *Information Theory, IEEE Transactions on*, 57(4):1872 –1905, april 2011.

[7] K. Azarian, H. El Gamal, and P. Schniter. On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels. *Information Theory, IEEE Transactions on*, 51(12):4152 –4172, dec. 2005.

[8] M. Benjillali and L. Szczecinski. A simple detect-and-forward scheme in fading channels. *Communications Letters, IEEE*, 13(5):309 –311, may 2009.

[9] Amir Bennatan, Shlomo Shamai, and A. Robert Calderbank. In praise of bad codes for multi-terminal communications. *CoRR*, abs/1008.1766, 2010.

[10] E. Berlekamp, R. McEliece, and H. van Tilborg. On the inherent intractability of certain coding problems (corresp.). *Information Theory, IEEE Transactions on*, 24(3):384 – 386, may 1978.

[11] J. Bruck and M. Naor. The hardness of decoding linear codes with preprocessing. *Information Theory, IEEE Transactions on*, 36(2):381–385, mar 1990.

[12] G. Caire, G. Taricco, and E. Biglieri. Bit-interleaved coded modulation. *Information Theory, IEEE Transactions on*, 44(3):927 –946, may 1998.

[13] Arnab Chakrabarti, Alexandre De Baynast, Ashutosh Sabharwal, and Behnaam Aazhang. Low density parity check codes for the relay channel. *Selected Areas in Communications, IEEE Journal on*, 25(2):280 –291, february 2007.

[14] Sae-Young Chung, Jr. Forney, G.D., T.J. Richardson, and R. Urbanke. On the design of low-density parity-check codes within 0.0045 db of the shannon limit. *Communications Letters, IEEE*, 5(2):58 –60, feb 2001.

[15] Sae-Young Chung, T.J. Richardson, and R.L. Urbanke. Analysis of sum-product decoding of low-density parity-check codes using a gaussian approximation. *Information Theory, IEEE Transactions on*, 47(2):657 –670, feb 2001.

[16] T. Cover and A.E. Gamal. Capacity theorems for the relay channel. *Information Theory, IEEE Transactions on*, 25(5):572 – 584, sep 1979.

[17] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John-Wiley and Sons Inc., 2006.

[18] P. Elias, A. Feinstein, and C. Shannon. A note on the maximum flow through a network. *Information Theory, IRE Transactions on*, 2(4):117 –119, december 1956.

[19] Y. Fan, C. Wang, H.V. Poor, and J. S. Thompson. Cooperative multiplexing: Toward higher spectral efficiency in multi-antenna relay networks. *Information Theory, IEEE Transactions on*, submitted.

[20] Yijia Fan, H.V. Poor, and J.S. Thompson. Cooperative multiplexing in full-duplex multi-antenna relay networks. In *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, pages 1 –5, 30 2008-dec. 4 2008.

[21] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. In Ira Gessel and Gian-Carlo Rota, editors, *Classic Papers in Combinatorics*, Modern Birkhäuser Classics, pages 243–248. Birkhäuser Boston, 1987.

[22] Gerard J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal*, 1(2):41–59, 1996.

[23] R.G. Gallager. *Low Density Parity Check Codes*. MIT Press, Cambridge MA, 1963.

[24] A. Host-Madsen and J. Zhang. Capacity bounds and power allocation for wireless relay channels. *Information Theory, IEEE Transactions on*, 51(6):2020 –2040, june 2005.

[25] Jilei Hou, Paul H. Siegel, Laurence B. Milstein, and Henry D. Pfister. Capacity-approaching bandwidth-efficient coded modulation schemes based on low-density parity-check codes. *Information Theory, IEEE Transactions on*, 49(9):2141–2155, 2003.

[26] T.E. Hunter and A. Nosratinia. Cooperation diversity through coding. In *Information Theory, 2002. Proceedings. 2002 IEEE International Symposium on*, page 220, 2002.

[27] T.E. Hunter and A. Nosratinia. Diversity through coded cooperation. *Wireless Communications, IEEE Transactions on*, 5(2):283 – 289, feb. 2006.

[28] H. Imai and S. Hirakawa. Correction to 'a new multilevel coding method using error-correcting codes'. *Information Theory, IEEE Transactions on*, 23(6):784, nov 1977.

[29] H. Imai and S. Hirakawa. A new multilevel coding method using error-correcting codes. *Information Theory, IEEE Transactions on*, 23(3):371 – 377, may 1977.

[30] Amir Ingber and Meir Feder. Parallel bit interleaved coded modulation. *Proceedings of Annual Allerton Conference on Communications, Control, and Computing*, September 2010.

[31] M. Janani, A. Hedayat, T.E. Hunter, and A. Nosratinia. Coded cooperation in wireless communications: space-time transmission and iterative decoding. *Signal Processing, IEEE Transactions on*, 52(2):362 – 371, feb. 2004.

[32] M.A. Khojastepour, N. Ahmed, and B. Aazhang. Code design for the relay channel and factor graph decoding. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on*, volume 2, pages 2000 – 2004 Vol.2, 7-10 2004.

[33] G.M. Kraidy, N. Gresset, and J.J. Boutros. Coding for the non-orthogonal amplify-and-forward cooperative channel. In *Information Theory Workshop, 2007. ITW '07. IEEE*, pages 626 –631, 2-6 2007.

[34] G. Kramer. Distributed and layered codes for relaying. In *Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on*, pages 1752 – 1756, october 2005.

[35] G. Kramer, M. Gastpar, and P. Gupta. Cooperative strategies and capacity theorems for relay networks. *Information Theory, IEEE Transactions on*, 51(9):3037 – 3063, sept. 2005.

[36] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498 –519, feb 2001.

[37] J.N. Laneman, D.N.C. Tse, and G.W. Wornell. Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *Information Theory, IEEE Transactions on*, 50(12):3062 – 3080, dec. 2004.

[38] J.N. Laneman and G.W. Wornell. Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks. *Information Theory, IEEE Transactions on*, 49(10):2415 – 2425, oct. 2003.

[39] O. Lé andvê andque, C. Vignat, and M. Yü andksel. Diversity-multiplexing tradeoff for the mimo static half-duplex relay. *Information Theory, IEEE Transactions on*, 56(7):3356 –3368, july 2010.

[40] Sung Hoon Lim, Young-Han Kim, A. El Gamal, and Sae-Young Chung. Noisy network coding. *Information Theory, IEEE Transactions on*, 57(5):3132 –3152, may 2011.

[41] Jin Lu and J.M.F. Moura. Linear time encoding of ldpc codes. *Information Theory, IEEE Transactions on*, 56(1):233 –249, jan. 2010.

[42] M.G. Luby, M. Amin Shokrolloahi, M. Mizenmacher, and D.A. Spielman. Improved low-density parity-check codes using irregular graphs and belief propagation. In *Information Theory, 1998. Proceedings. 1998 IEEE International Symposium on*, page 117, aug 1998.

[43] M.G. Luby, M. Mitzenmacher, M.A. Shokrollahi, and D.A. Spielman. Improved low-density parity-check codes using irregular graphs. *Information Theory, IEEE Transactions on*, 47(2):585 –598, feb 2001.

[44] Emin Martinian and Jonathan S. Yedidia. Iterative quantization using codes on graphs. *CoRR*, cs.IT/0408008, 2004.

[45] R.U. Nabar, H. Bolcskei, and F.W. Kneubuhler. Fading relay channels: performance limits and space-time signal design. *Selected Areas in Communications, IEEE Journal on*, 22(6):1099 – 1109, aug. 2004.

[46] V. Nagpal, S. Pawar, D. Tse, and B. Nikolic. Cooperative multiplexing in the multiple antenna half duplex relay channel. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 1438 –1442, june 2009.

[47] V. Nagpal, I-Hsiang Wang, M. Jorgovanovic, D. Tse, and B. Nikolić. Quantize-map-and-forward relaying: Coding and system design. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 443 –450, 29 2010-oct. 1 2010.

[48] R. Narasimhan. Finite-snr diversity multiplexing tradeoff for correlated rayleigh and rician mimo channels. *Information Theory, IEEE Transactions on*, 52(9):3965–3979, Sept. 2006.

[49] Thuy Van Nguyen, A. Nosratinia, and D. Divsalar. Bilayer protograph codes for half-duplex relay channels. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 948 –952, june 2010.

[50] A. Ozgur and S. Diggavi. Approximately achieving gaussian relay network capacity with lattice codes. *ArXiv e-prints*, May 2010.

[51] A. Ozgur, O. Leveque, and D.N.C. Tse. Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *Information Theory, IEEE Transactions on*, 53(10):3549 –3572, oct. 2007.

[52] S. Pawar, A.S. Avestimehr, and D. Tse. Diversity-multiplexing tradeoff of the half-duplex relay channel. In *Proc. Forty-Sixth Allerton Conf. Commun. Contr. Comput.*, Illinois, 2008.

[53] S. Pawar, A.S. Avestimehr, and D.N.C. Tse. Diversity-multiplexing tradeoff of the half-duplex relay channel. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 27 –33, 23-26 2008.

[54] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Kauffman, San Francisco CA, 1988.

[55] N. Prasad and M.K. Varanasi. Cth17-3: High performance static and dynamic cooperative communication protocols for the half duplex fading relay channel. In *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE*, pages 1 –5, 27 2006-dec. 1 2006.

[56] P. Razaghi, M. Aleksic, and Wei Yu. Bit-interleaved coded modulation for the relay channel using bilayer ldpc codes. In *Information Theory, 2007. CWIT '07. 10th Canadian Workshop on*, pages 101 –104, 6-8 2007.

[57] P. Razaghi and Wei Yu. Bilayer low-density parity-check codes for decode-and-forward in relay channels. *Information Theory, IEEE Transactions on*, 53(10):3723 –3739, oct. 2007.

[58] T.J. Richardson, M.A. Shokrollahi, and R.L. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *Information Theory, IEEE Transactions on*, 47(2):619 –637, feb 2001.

[59] T.J. Richardson and R.L. Urbanke. The capacity of low-density parity-check codes under message-passing decoding. *Information Theory, IEEE Transactions on*, 47(2):599 –618, feb 2001.

[60] T.J. Richardson and R.L. Urbanke. Efficient encoding of low-density parity-check codes. *Information Theory, IEEE Transactions on*, 47(2):638 –656, feb 2001.

[61] T.J. Richardson and R.L. Urbanke. *Modern Coding Theory*. Cambridge University Press, 2008.

[62] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity. part i. system description. *Communications, IEEE Transactions on*, 51(11):1927 – 1938, nov. 2003.

[63] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity. part ii. implementation aspects and performance analysis. *Communications, IEEE Transactions on*, 51(11):1939 – 1948, nov. 2003.

[64] Rainer Storn and Kenneth Price. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J. of Global Optimization*, 11:341–359, December 1997.

[65] R. Tanner. A recursive approach to low complexity codes. *Information Theory, IEEE Transactions on*, 27(5):533 – 547, sep 1981.

[66] Emre Telatar. Capacity of multi-antenna gaussian channels. *European Transactions on Telecommunications*, 10(6):585–595, 1999.

[67] David N. C. Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[68] G. Ungerboeck. Channel coding with multilevel/phase signals. *Information Theory, IEEE Transactions on*, 28(1):55 – 67, jan 1982.

[69] M. Uppal, Guosen Yue, Xiaodong Wang, and Zixiang Xiong. A rateless coded protocol for half-duplex wireless relay channels. *Signal Processing, IEEE Transactions on*, 59(1):209 –222, jan. 2011.

[70] Momin Uppal, Zhixin Liu, Vladimir Stankovic, and Zixiang Xiong. Compress-forward coding with bpsk modulation for the half-duplex gaussian relay channel. *Trans. Sig. Proc.*, 57(11):4467–4481, 2009.

[71] M.K. Varanasi and T. Guess. Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the gaussian multiple-access channel. *Signals, Systems and Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, 2:1405–1409, Nov 1997.

[72] I-Hsiang Wang and David N. C. Tse. Interference mitigation through limited receiver cooperation. *Submitted to IEEE Transactions on Information Theory*, November 2009, `http://arxiv.org/abs/0911.2053`.

[73] N. Wiberg. *Codes and decoding on general graphs.* PhD thesis, Linköping Univ, Linköping, Sweden, 1996.

[74] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *Information Theory, IEEE Transactions on*, 22(1):1 – 10, jan 1976.

[75] M. Yuksel and E. Erkip. Multiple-antenna cooperative wireless systems: A diversity-multiplexing tradeoff perspective. *Information Theory, IEEE Transactions on*, 53(10):3371 –3393, oct. 2007.

[76] E. Zehavi. 8-psk trellis codes for a rayleigh channel. *Communications, IEEE Transactions on*, 40(5):873 –884, may 1992.

[77] Z. Zhang and T. M. Duman. Capacity approaching turbo coding for half-duplex relaying. *Communications, IEEE Transactions on*, 55(9):1822, sept. 2007.

[78] Zheng Zhang and T.M. Duman. Capacity-approaching turbo coding and iterative decoding for relay channels. *Communications, IEEE Transactions on*, 53(11):1895 – 1905, nov. 2005.

[79] B. Zhao and M.C. Valenti. Distributed turbo coded diversity for relay channel. *Electronics Letters*, 39(10):786 – 787, may 2003.

[80] Lizhong Zheng and D.N.C. Tse. Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *Information Theory, IEEE Transactions on*, 49(5):1073 – 1096, may 2003.