uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Jeffrey Castura

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Rateless Coding over Wireless Channels:
Theory, Design and Applications

TITRE DE LA THÈSE / TITLE OF THESIS

Yongyu Mao

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Shahram Yousefi

Halim Yanikomeroglu

David Falconer

Abbas Yongacoglu

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Rateless Coding over Wireless Channels: Theory, Design and Applications

by

Jeff Castura

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the Ph.D. degree in
Electrical and Computer Engineering

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-50721-6
Our file Notre référence
ISBN: 978-0-494-50721-6

Abstract

In this work we investigate rateless coding and modulation for wireless communication systems. In particular we are interested in channels in which only the transmitter has channel knowledge. For this setting, we show that rateless codes have properties that make them well suited for wireless systems. We advance the theory of rateless codes by proving the existence of rateless codes which are "good" for all possible realizations of fading channels simultaneously.

To complement rateless coding, we introduce a novel modulation scheme called μ-PAM modulation. We show that typical M-ary modulation schemes are not always well suited for the channels considered in this work, and that μ-PAM modulation has benefits relative to standard M-ary PAM. A simple demodulation scheme is presented and its performance is analyzed and simulated.

Applications using rateless codes over wireless channels are considered. We examine the behavior of a streaming application with delay constraints over fading channels, comparing rateless codes with other traditional systems. Benefits of the rateless approach are presented, and theoretical guarantees on probability of outage are developed.

Finally, the use of rateless codes for relay networks are explored and we demonstrate theoretic limits on the rate of communication for the proposed relaying strategy. Simulations validate the performance of the strategy.

Acknowledgements

There are many people who have contributed to the completion of this work, either directly or otherwise.

Everlasting thanks to my wife, your support has only grown during my tenure at the University of Ottawa, and I can never express how much it has helped me.

To my supervisor, Dr. Yongyi Mao, you have been an excellent mentor. From you I have gained appreciation for the worlds of information theory and probability theory. You have been tireless and encouraging, and have put up with my many distractions.

To my colleagues with whom I have had the chance to discuss research ideas and collaborate on projects, thank you for all your assistance and support.

All errors in this work are my own.

Contents

List of Tables

List of Figures

List of Symbols

T_c Coherence time specified as a number of channel uses

h_i Complex channel gain for the fading channel model at discrete time i

$\mathbf{h^n}$ Sequence of channel gains $(h_1, h_2, \ldots, h_n)$

ζ A channel realization representing $\mathbf{h}^\infty$

$C(n;\zeta)$ Realized capacity of a channel given channel realization ζ and n channel uses

γ SNR at the receiver

k/T_c A normalized information rate where k is a number of information bits and T_c is the coherence time

μ-PAM Modulation scheme parameterized by $0 < \mu \leq 1/2$

M Number of constellation symbols in (μ-)PAM modulation

L base-2 logarithm of M, $L = \log_2 M$

D Delay constraint specified in number of channel uses

D/T_c Normalized delay constraint in terms of coherence time

G Gain (in dB) of the SNR between the source to relay channel relative to the source to destination channel in the relay channel

Chapter 1

Introduction

Digital mobile communications systems have reshaped the way people interact and communicate with each other. From cellular and satellite phones enabling person to person calling from virtually anywhere in the world, to myriad portable devices allowing reliable Internet access while on the move, these mobile systems are unquestionably a technological, if not social, success.

This success comes despite a number of fundamental challenges associated with wireless mobility. One such challenge is the dynamic, time-varying nature of the signal propagation conditions through the air. Impairments caused by movement of the transmitter or receiver, reflections of the signal, and interference from other sources must all be handled in a reliable and efficient manner. For the past several decades research has been focused on developing solutions to the problems caused by these impairments and a large number of successful solutions have been implemented.

One of the fundamental tools used in this venue is channel coding; the use of structured redundancy in order to detect and correct errors induced by the channel. Coupled with modulation techniques, adaptive receiver designs and link-layer protocols, many wireless systems are able to achieve acceptable levels of reliability or efficiency. One need only to look at current cellular systems such as GSM/GPRS/EDGE [74] and UMTS [76], wireless LAN systems such as WiFi [77] and WiMAX [78], or short-range bluetooth systems [73] to witness the ubiquity and utility of coding.

Despite these advances, it remains a difficult challenge for current wireless systems to simultaneously maintain efficiency and reliability over dynamic channels. The instantaneous capacity of the channel, often a time-varying value, usually can not be tracked by the transmitter without sophisticated methods. As a result, measures of performance

such as throughput or probability of outage remain less than theoretically achievable. Current systems offer acceptable performance at the expense of signaling overhead and limited data rates.

A type of channel code, *fountain codes* have been developed for use over erasure channels. In this work we explore the possibility of using fountain coding principles for wireless channels. Specifically, we investigate the generalization of fountain codes which are referred to as *rateless codes*. We will show that the properties of rateless codes can help with some of the limitations of current wireless systems, resulting in improvements in throughput and outage.

1.1 Fountain Coding

A typical digital communication system may be broken down into the components shown in Figure 1.1. We wish to send a message source consisting of some binary sequence to the message sink at the other end of the system. We assume for the purposes of this work that the source sequence has been compressed and therefore we can assume that the source is an independent and identically distributed (i.i.d) random binary sequence with an equal probability of a one or a zero.

The source sequence is first mapped by the encoder to a codeword. This is the channel encoder which adds structured redundancy to the source sequence to provide resiliency to errors induced by the channel. The output of the encoder is fed into a modulator which maps the codeword to a channel-compatible form.

The channel introduces possible errors into the modulated message in a probabilistic manner. The role of the demodulator is to undo the mapping done by the modulator and to output estimates of the bits comprising the transmitted codeword. Finally, the decoder produces an estimate of the original source sequence given the input codeword estimate in an attempt to fix any errors that were induced by the channel.

We focus for a moment on the coding/decoding portions of Figure 1.1, whose combined purpose is to allow errors induced by the channel to be detected and/or corrected. Typically this operates in a block-based manner; some number k of information bits are mapped to some number $n \geq k$ of coded bits, where both k and n are fixed *a priori*. It is possible to create codes for which n is not fixed, but rather is an outcome based on the conditions of the channel and the code design. Codes of this type are often collectively referred to as rateless codes, and this nomenclature is adopted for this paper.

A specific class of rateless codes called fountain codes have received a great deal of

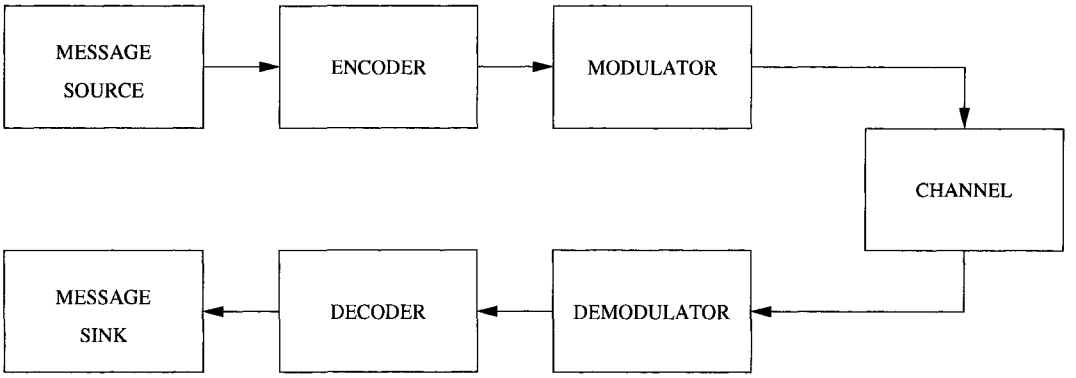

Figure 1.1: Digital communication system model

attention in the past few years. They not only have low computational complexity but also have capacity-achieving or capacity-approaching performance for certain classes of channel models. The term fountain is meant to invoke the sense of an infinite stream of information being communicated, much as a fountain continuously streams water.

The first fountain codes which are called *Luby Transform* (LT) codes were introduced by Luby in [48]. These codes were shown to achieve the capacity of the binary erasure channel (BEC) without any knowledge of the erasure probability at the encoder. However, LT codes are prone to significant performance degradation for small values of k. Raptor codes, the successor to LT codes, were introduced by Shokrollahi in [66] and use LT codes as an inner code with a high-rate low density parity check (LDPC) outer code. Raptor codes were shown also to be capacity achieving for the BEC, and with the beneficial properties of little or no error floor for smaller block lengths and a linear-time encoding and decoding computational complexity.

These codes have proven successful for Internet-based content distribution applications, and the company Digital Fountain [18], founded by Luby and Shokrollahi and holding many patents on these codes, is in the process of commercializing these applications. Other applications which may be served by fountain codes and are being commercialized include mobile communications and ad-hoc, mesh or sensor networks. However, these products are built on the assumption that the channel being used is an erasure channel of some type. That is, the channel is always modeled as an erasure channel, which is often either inefficient or not applicable for the application. We will show in this work that fountain codes can also be well suited to fading channels.

Unlike traditional fixed-rate codes which map k information bits to n coded bits, a fountain code operates on k information bits, generating coded bits individually in a

consecutive, sequential manner. The generation of a coded bit is a two-step process. The first step requires the encoder to draw a pseudo-random number d with some predetermined distribution. The encoder then pseudo-randomly chooses d distinct information bits with equal likelihood. These selected bits are combined together using the "exclusive OR" (XOR) operation, and the result is transmitted over the channel.

This process may be repeated an unlimited number of times, thereby generating codewords of essentially infinite length. In practice, the decoder, whose operation is described below, acknowledges the encoder when it has decoded and recovered the original k information bits through a feedback channel. In this way, the rate that is achieved by the system depends on both the channel and the code construction and implementation.

The decoder is capable of reproducing the identical pseudo-random realization that is used to by the encoder. It is then able to assign to each received symbol from the channel the list of information bits that were used to generate that bit. Essentially, the decoder reproduces the generator matrix that the encoder uses. The decoder also tracks the amount of information it has received from the channel. At some point it determines that it has sufficient information, and then decodes the codeword. If decoding is successful, it uses a feedback channel to notify the encoder. Otherwise, it simply collects more information from the channel and attempts decoding again at some future time.

A remarkable fact of these fountain codes is that, with proper construction, their performance is *universal* over the BEC. This means that, despite not knowing the erasure probability of the channel (in fact this probability may even vary during the course of the codeword transmission), these codes can achieve the capacity of the channel, regardless of what the erasure probability is. This powerful property is a key element to their success, and is a major motivation for their use in Internet-based applications. This and other useful properties of rateless codes provides further motivation to explore their potential over other channels and media, which is explored in the following chapters.

1.2 Overview and Thesis Contributions

This work focuses on the encoder/decoder and modulator/demodulator blocks of the communication system of Figure 1.1. Given a wireless channel modeled by an appropriate probabilistic transfer function, we seek a coding and modulation system that can overcome many of the limitations of current wireless systems as discussed above. Motivated by the properties of fountain codes, the ultimate goal of this work is to study and analyze the use of rateless concepts – both coding and modulation – in wireless

communication systems.

The nature of fading channels is investigated in Chapter 2 with a focus on the implications for rateless systems. Portions of this chapter were published originally in [10,9,14]. Over fading channels, achievable rates are random variables that are functions of the realizations of the random channel. As such, there will be a distribution of achievable rates induced by the channel. We provide an analytical perspective on the limits on the rate of reliable communication for fading channels, derived in the context of rateless systems. Traditional "fast" and "slow" fading models are considered, as well as fading rates that qualify as neither "fast" nor "slow". In the latter case, it is shown that the distribution of achievable rates takes a discontinuous form. This form further motivates the use of a rateless coding scheme (as opposed to a fixed-rate scheme) for such channels and all fading rates.

To motivate the work in Chapter 2, consider the following scenario. Suppose we wish to communicate k bits of information reliably over a fading channel without channel knowledge. When is the transmitted message decodable? This classical question has been dealt with for many types of fading channels. See, for example, [79] for a survey.

From a coding-theoretic perspective, this question is logical to ask only in the context of a variable-rate coding scheme. In typical fixed-rate coding schemes, the codeword length n is chosen *a priori*, hence the time of decoding is fixed. This restriction immediately implies that outage events are inevitable, and that the outage probability is bounded away from zero. With such a fixed-rate scheme, the only way to reduce outage probability is to reduce the communication rate, and the system designer is faced with a trade-off between efficiency (i.e., rate) and reliability (i.e, outage). So with fixed-rate coding, the question is not a matter of *when* a message is decodable, but rather *whether* a message is decodable at some rate.

From a rateless coding perspective, we characterize the distribution of the time of earliest successful decoding when only k is fixed *a priori*. We present a numerical approach to allow the computation of the distribution of decoding times for arbitrary block-independent fading channels. Using this form we show that the decoding time distributions are discontinuous. Decoding time distributions are computed for various underlying fading distributions, and consequences of these distributions are discussed.

Having investigated the nature of the channel, the next step is to consider whether rateless codes are in fact suited for wireless systems. This is also a subject of Chapter 2, where it is shown that these codes have properties that are well suited for wireless environments. In particular we define the types of wireless channels considered, provide some

measures used to compare the performance of different coding and modulation alternatives and then examine rateless codes. The properties of rateless codes are examined and we provide a proof of the existence of good rateless codes for the wireless channels considered. In this context we show that universally good rateless codes exist; i.e., rateless codes can achieve the (realized) capacity of a class of fading channels simultaneously for all possible channel realizations.

Based on this proof, we propose to use rateless codes for communication over fading channels when the channel state information is unknown at the transmitter. Our rationale for using rateless codes to achieve efficiency, reliability and robustness is in fact quite intuitive. First, unlike fixed-rate codes in which the rate k/n is a constant and independent of channel realizations, rateless codes realize a rate for each codeword depending on what channel realization is experienced by the codeword. Since the successful-decoding time n depends on the channel realization, a good rateless code may naturally adapt to the channel state thereby providing opportunities for the realized rate k/n to closely follow the realized capacity of the channel. Second, as suggested in [20], for a large class of discrete channels, choosing the codebook size 2^k to be sufficiently large may drive the outage probability to zero without sacrificing rate. Third, we note that knowledge of channel statistics at the transmitter plays no essential role with rateless codes. If a rateless code can realize a rate that is close to the realized capacity in every transmission, it can then closely follow the average realized capacity regardless of the channel statistics. Thus a rateless code can maintain its efficiency and reliability across all channel ensembles.

Given these promising results we then turn to the problem of modulation in Chapter 3. Here it is shown that when communicating over channels with unknown state at the transmitter, the choice of modulation rate plays an important role in the overall performance of the communication system. Focusing on M-ary modulation in one dimension, we introduce a new form of modulation called μ-PAM modulation and compare it with standard methods such as M-PAM. A portion of this work appeared originally in [11].

A novel solution to the problem of modulation for unknown channels is to use an adaptive demodulation scheme as described in [5], coupled with a Raptor code. Here, the modulation scheme is fixed, but the demodulation scheme is adaptive based on the channel state, thereby "compressing" the number of bits output from the demodulator. This approach automatically adapts its rate over a range of SNR such that the system targets a particular (high) SNR, and gracefully degrades when the SNR falls below this value.

We seek a modulation system that offers efficient and reliable communication over unknown channels with a somewhat different objective. Specifically, we would like a system that can perform well in very low SNR conditions, but adaptively increases its rate as the SNR increases.

The proposed strategy for this communications problem is to map some number of consecutive rateless-coded output bits to a point in a constellation of non-uniformly spaced, real-values. We generalize M-PAM modulation with a fractional-power modulation scheme called μ-PAM. This modulation scheme is similar to M-PAM, except that the symbol constellation no longer has an equal distance between adjacent symbols. Instead, the constellation points are generated in a recursive manner, dependent on μ, with $\mu \in (0, 1/2]$.

Since Raptor codes are able to produce a virtually infinite number of output symbols, we propose a method in which a large number of consecutive output symbols are mapped into a single real value, and this value is transmitted over the unknown channel. This process is repeated until the decoder is able to recover the information bits. The mapping from output symbols to transmitted symbol may be considered a form of non-uniform modulation, providing unequal error protection to the output symbols.

A simple demodulation technique is introduced which performs nearly as well as maximum likelihood detection. The algorithm is complimentary to the modulation structure in that it can be viewed as a recursive operation. Using this algorithm, performance of the modulation system is analyzed and studied in detail. In particular, we present bounds on the achievable rates at which the modulation system can communicate information, and present an asymptotic analysis of performance as M approaches infinity. We show that in many circumstances, this new modulation scheme provides benefits to the performance of the overall system, particularly for unknown channels. We compare this new modulation scheme with M-PAM and highlight the tradeoffs between the two in terms of achievable rates.

Having investigated theoretical limits of the rate of communication for rateless codes and modulation over fading channels, Chapter 4 presents an investigation of the application of rateless coding to fading channels when there is a delay constraint on the transmission of data. Part of this chapter was originally published in [14].

Delay constraints are a common and practical limit that is built into many communication systems. Their use may be a result of technical limitations, such as a minimum acceptable number or retries that a block-based coding system may impose, or may be due to quality-of-service guarantees. When a delay constraint is imposed over fading

channels with unknown state at the transmitter, it is almost certain that outage probability will be bounded away from zero. This is clearly true for fixed-rate communication schemes, but we will show that when rateless coding is applied, under certain circumstances it is possible that long-term outage probability can be driven to zero, even in the presence of delay constraints.

This will be seen in the context of a video streaming application where a delay constraint is imposed due to a quality-of-service requirement on the latency of transmission. For this application we compare fixed-rate coding schemes, some existing variable-rate coding systems, and rateless codes. We show that rateless codes offer the best overall performance when comparing the metrics of throughput, probability of outage and mean time to decode.

With such applications in mind, the channels we consider are fading channels with unknown states at the transmitter and the presence of a reliable feedback channel. The channel state may vary with time and the rate of this variation may result in different behaviors that require different communication strategies. For quasi-static fading channels, in which the channel conditions remain fixed for the duration of a codeword, one may use feedback in order to signal the channel state information to the transmitter. For fast-fading channels, in which channel conditions vary independently from symbol to symbol, one may rely on the law of large numbers and code for the expected case thereby achieving the ergodic capacity of the channel. Between these extremes, one may consider the block-fading channel with a delay constraint, where there are only a few fading blocks during the transmission of a codeword. In this case the above mentioned strategies fail to be efficient.

In this regime, the work in [56] examines the impact of causal feedback on the capacity of delay-constrained block-fading channels and presents an optimal power-control strategy under channel uncertainty. More generally, the work of [32] characterizes the capacity region of the delay-constrained multiple access channel, and presents an optimal resource allocation scheme. The works in [58, 69, 65] present performance analyses for variable-rate and rateless coding strategies using LDPC codes and fountain codes, demonstrating their applicability over a range of SNR.

It is worth noting that rateless coding may be considered as a type of Hybrid-ARQ scheme or variable-rate strategy, both of which have been used for communication under channel uncertainty. For example, using a block-incremental redundancy strategy, the throughput for the Gaussian collision channel is analyzed in [6]. Furthermore, the information theoretic analysis in [20] has demonstrated that rateless codes can achieve

the mutual information induced over the channel by the choice of input distribution for the class of all discrete memoryless channels with no channel state information.

The main contribution of this chapter is to demonstrate that rateless codes offer advantages in terms of throughput and outage probability for delay-constrained streaming applications over fading channels. Motivated by the video-streaming application, we compare rateless and fixed-rate transmission schemes and demonstrate the existence of a critical operating point for rateless coding strategies such that beyond this threshold, the outage (or frame-jitter) probability can be driven to zero in the long run. This makes rateless codes an attractive solution for such applications.

We provide a theoretical justification for the existence of the critical operating point and present a detailed characterization. The critical point depends on three system parameters, the SNR, the ratio of the delay constraint to coherence time, and the ratio of the information rate to coherence time. We present a number of plots demonstrating the impact of modifying these parameters and compare fixed-rate and rateless strategies.

Comparisons between rateless and fixed-rate systems are also made using mean throughput and mean time-to-decode. In both cases we show that rateless systems offer advantages. We also apply the μ-PAM modulation presented in Chapter 3, further highlighting some of the advantages that this modulation scheme provides when used in realistic applications.

The concepts developed in the early chapters of this work are expanded from a point-to-point setting to a multi-terminal setting in Chapter 5. Portions of this chapter appeared in [8]. As an extension of the results obtained so far, we present a novel framework for coding over relay channels using joint rateless codes and modulation. As far as we are aware, we are the first to propose the use of rateless coding for relay channels and relay networks. This framework is at the intersection of two active areas of research in communications, namely relay networks and rateless coding. We demonstrate that there is a very natural and useful fit between these two areas, and describe some design challenges and implementation considerations for this framework.

The use of relays in wireless communication networks provide a new dimension to the design space of wireless networks which promises enhancements to both the coverage and throughput of the network. In its simplest form, a relay network is a collection of terminals which are able to transmit, receive, and possibly assist the reliable delivery of information from source terminals to destination terminals. Thus, communication of data through a wireless relay network is not required to be direct; it may pass through a number of other terminals, though direct communication from source to destination is

not precluded. In fact, it is possible to simultaneously use single-hop, i.e. direct, and multi-hop communications paths.

Traditional wireless networks have predominantly used direct point-to-point or point-to-multipoint (e.g. cellular) topologies. The fundamentally different mode of transport, possible uncertainty in terminal geographical location, and difficulty in theoretical analysis of relay networks have kept them mainly to academic realms. However, there has been a number of recent theoretical results that may spur the use of relaying techniques in practical networks. The key behind these advances are mainly a result of research in multiple antenna systems.

Multiple antenna systems, or multiple-input, multiple-output (MIMO) systems have seen remarkable growth in recent years and can deliver significant throughput and coverage gains over wireless channels compared with single antenna systems. A fundamental observation to make with respect to wireless relay networks is that, with appropriate coordination or cooperation, communication between two terminals in the network can be viewed as a type of MIMO system. This may be achieved in a number of different ways, however all require some level of cooperation within the network. The relaying strategy used may impose limits on the similarity to a MIMO system.

In their two-part paper [63, 64], Sendonaris, Erkip and Aazhang introduce and examine the concept of user cooperation diversity. Here, the authors demonstrate that simple cooperation between transmitting users can increase throughput and coverage simultaneously. Other strategies are examined for wireless channels in [39]. Dohler *et al* introduce "virtual antenna arrays" in [19]. Here, groups of terminals cooperate to form a virtual MIMO system and exploit the spatial diversity that results. This is a similar concept to user cooperation, but focuses on different design aspects, such as link budget impact.

A natural extension of the basic implementation described in [63] is to use coded cooperation and this is described in [36] by Hunter and Nosratinia with further analysis and implementation details for wireless channels given in [37]. Further approaches using coded cooperation are given in [44, 43], where the authors propose a decode-and-forward (DF) scheme with many opportunistically cooperating terminals, and show that diversity gains scale in the number of potential relays rather than the actual number of participating relays.

In works that foreshadow the use of rateless codes for relay channels, Caire and Tuninetti in [6], and Zhao and Valenti in [89] propose and analyze the use of Hybrid-ARQ for relay channels. In [6], hybrid-ARQ protocols for the Gaussian collision channel

are studied. Notably, the results translate to relay channels. Also foreshadowing the application of rateless codes, Mitran, Ochiai and Tarokh present in [51] a two-phase communication scheme for wireless devices in a network with the half-duplex constraint along with an information-theoretic performance analysis.

Given these works, the use of rateless codes for wireless relay channels seems to be a synergistic match. We present a communication framework over relay channels using joint rateless codes and modulation. We first introduce the relay channel model that we study, then provide a justification for the framework and cooperation strategy proposed. An information-theoretic achievable rate region is presented and we show how the rateless framework can automatically attain the optimal achievable rate.

Building on this theory, we present a simple simulated implementation of the proposed framework. We simulate the basic relay channel in which a source wishes to communicate information to a destination, with possible assistance from an otherwise idle relay node. The Monte Carlo simulation is used to present performance curves of mean achieved rate, probability of outage and mean time to decode. It is shown that the diversity order of the system approaches that of a 2×1 MIMO system as expected.

The simulations also show that the rateless code is able to naturally and automatically adapt to the fluctuation, unknown conditions of the channel between each pair of nodes in the network. We use μ-PAM modulation presented in Chapter 3 in the simulation and explicitly present the advantages of using this modulation method compared to standard M-PAM in many conditions.

Chapter 6 concludes the thesis with a summary of the contributions made in this work. A number of extensions and suggestions for future research are also provided. Since the concept of rateless coding and modulation is relatively immature, there are a large number of open questions and directions for future work, covering information theory, coding theory, modulation design, relay channels and network design.

Chapter 2

Rateless Coding for Fading Channels

Channel coding is the primary means by which reliability is achieved in a communication system, and as such it has been the focus of intense research for over half of a century. Advances in coding theory have led to codes that are practical to implement and can communicate reliably within fractions of a decibel of the capacity of an ever-growing number of channels. Although the capacity of wireless channels are known and understood, their dynamic nature and wide range of models have made coding for these channel particularly difficult.

The goal of this chapter is to introduce and analyze rateless coding concepts, then apply them to wireless channels. Figure 2.1 shows the communication system with the encoding and decoding blocks highlighted, which are the focus of this chapter. The organization of this chapter is as follows. We introduce the class of channels considered in this work in Section 2.1. To achieve our goal, it is important to pose the problem we seek to solve clearly. One question we wish to answer as part of this work is "when can we decode a message"? This is studied in Section 2.2, where we show that the the time it takes to decode a message is a random variable induced by the channel. This result is the main motivation for the use of rateless codes.

Since we cannot know how long it will take to decode a message, we seek a coding scheme that can guarantee efficient performance *regardless* of the state of the channel. This is one of the powers of rateless codes, and in Section 2.3 we introduce rateless codes formally and provide some background and properties. In Section 2.4 we prove the existence of rateless codes which have the exact property that we are looking for; namely that there exist rateless codes which can achieve the capacity of a class of fading channels no matter what the realization of the channel is. To validate these results, Section 2.5

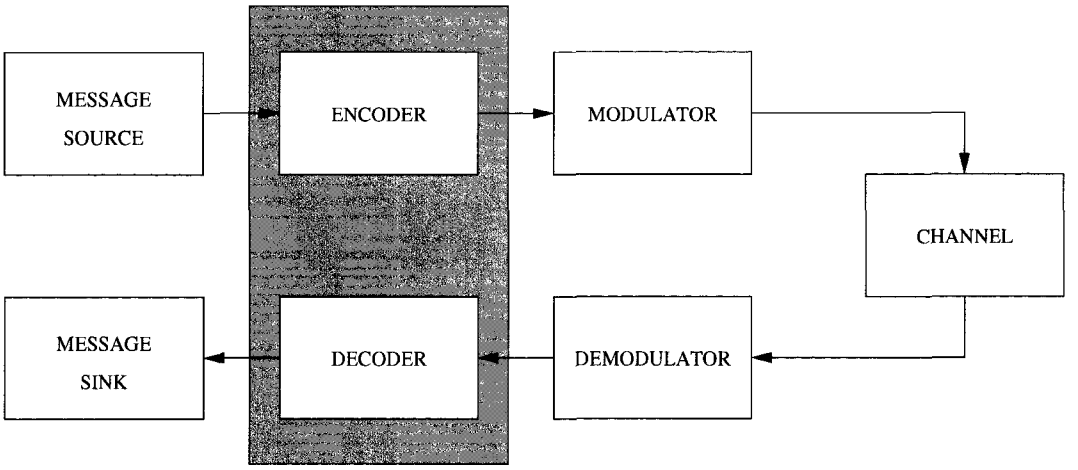

Figure 2.1: Digital communication system model highlighting coding functions

presents a simulated implementation of a rateless code over fading channels and we show that performance is very good across a wide range of channel conditions. We conclude the chapter with some discussion and summary in Section 2.6.

2.1 System Model

We consider point-to-point, single-antenna communication over fading channels where channel state information — the complex channel gain h_i — is available only at the receiver. The discrete channel model can be expressed as

$$Y_i = h_i X_i + N_i \tag{2.1}$$

where, at time i, Y_i and X_i are the received and transmitted complex symbols respectively, and N_i is complex additive white Gaussian noise. The time-varying characteristic of h_i is the block-fading model with coherence time T_c. Within each block of T_c channel uses h_i remains constant and between blocks h_i is drawn independently from some distribution over the set of complex numbers. For wireless fading channels, this distribution may be assumed to be a circularly-symmetric complex Gaussian distribution of unit variance. We impose no restriction on the form of this distribution beyond requiring that $0 < |h_i| < \infty$, which is needed for the lemma in section 2.2.1. The model in the limit of $T_c = \infty$ becomes the *quasi-static fading* model, which we may refer to subsequently as slow-fading. Similarly, at $T_c = 1$, the channel becomes the *fast-fading* model.

We will denote the sequence $(h_1, h_2, \ldots, h_n)$ from time instant 1 to time instant n by $\mathbf{h}^n$. Then $\mathbf{h}^\infty$, which we refer to as a *channel realization*, is characterized by a discrete-time random process $\{\mathbf{h}^\infty(\zeta) : \zeta \in \Omega\}$, where we treat $\mathbf{h}^\infty$ as a bijective function on some Ω and the probability measure on Ω induces the probability measure over the set of all channel realizations. The one-to-one correspondence between a channel realization $\mathbf{h}^\infty$ and ζ then allows us to identify a channel realization with a $\zeta \in \Omega$.

Given a channel realization ζ and any positive integers n_1 and n_2 ($n_1 < n_2$), we define the *realized capacity* (also called *instantaneous capacity* in some works, e.g. [1]) for channel realization ζ over the discrete-time interval $[n_1, n_2]$ as

$$C(n_1, n_2; \zeta) := \frac{1}{n_2 - n_1 + 1} \sum_{i=n_1}^{n_2} \log_2(1 + \gamma |h_i(\zeta)|^2) \quad \text{bits/channel use}, \tag{2.2}$$

where γ is the receive SNR. When n_1 and n_2 are interpreted respectively as the starting and ending time of transmitting a codeword, $C(n_1, n_2; \zeta)$ in (2.2) indicates the average "supported rate" over the duration of the codeword. Subsequently we may write $C(n; \zeta)$ in place of $C(1, n; \zeta)$ for simplicity. For sufficiently large n and when the number of fading blocks, n/T_c, approaches infinity, for any $\zeta \in \Omega$ $C(n; \zeta)$ approaches the deterministic ergodic capacity. That is, with probability approaching one,

$$C(n; \zeta) = E_h(\log(1 + |h|^2 \gamma)), \tag{2.3}$$

and that coding over a large number of fading blocks allows this capacity to be achieved. Note that this is the same capacity as that for the fast-fading channel. When $n/T_c \to 0$, realized capacity is a random variable given by

$$C(n; \zeta) = \log(1 + |h|^2 \gamma) \tag{2.4}$$

and is independent of n. This is the classical result for quasi-static fading.

The fact that both extremes of block-fading in terms of n/T_c result in capacities that are independent of n is interesting since between these two asymptotes realized capacities do indeed depend on n/T_c. This is explored in more detail in the following section.

In practice, where certain delay requirements dictate that a codeword cannot be transmitted over a large number of fading blocks, the notion of ergodic capacity is no longer useful. Such a scenario, under the condition that T_c and n are relatively large, is the primary interest of this chapter.

In this setting, when communicating a codeword of length n over the realized channel ζ, the *efficiency* of transmitting the codeword may be indicated by how closely the communication rate k/n approaches $C(n;\zeta)$, where k is the number of information bits carried in the codeword. In addition, when the codeword is decoded incorrectly, an outage occurs. Given that $C(n;\zeta)$ is the rate supported by the channel realization ζ, when using a codeword of length n to communicate k bits of information, the theoretical lower limit of the *outage probability* is given by [79]

$$P_{\text{out, lim}} = P[k/n > C(n;\zeta)]. \tag{2.5}$$

In practice, the outage probability for a system may be defined as the average probability of decoding error. The *reliability* of a communication scheme may then be indicated by how close the outage probability is to the limit in (2.5). The fact that when $k/n < C(n;\zeta)$, it is *in principle* [1] possible to have arbitrarily small probability of decoding error, and that when $k/n > C(n;\zeta)$, the error probability approaches one asymptotically.

Since the statistics of wireless channels typically vary with time and space, a communication scheme designed for a particular channel ensemble may become either less efficient or less reliable as these statistics vary. A system is said to be *robust* in this work if it is both reliable and efficient across channel statistics. As a limiting case, a system in which both efficiency and reliability are independent of the channel realization or statistics would be completely robust, and such a scheme would be considered *universal* in the usual sense over the given class of channels.

Note that in many wireless systems offering multimedia services, a number of possible fixed code rates are available for use, and the choice is dependent on the transmitter's estimation of the channel conditions. For example, UMTS systems use a turbo code with a number of different puncturing patterns, each of which results in a different actual code rate. Thus, depending on the service required and the channel conditions, a desired rate is chosen such that, along with power control and other control mechanisms, a certain reliability is maintained.

2.2 Decoding Times over Fading Channels

Imagine a communication system in which a transmitter desires to communicate k bits of information reliably over a fading channel without channel knowledge. We ask the

[1] This however needs the transmitter to select the rate based on channel realization h.

question: when is the transmitted message decodable? This classical question has been dealt with for many types of fading channels. See, for example, [79] for a survey.

From a coding-theoretic perspective, this question is logical to ask only in the context of a variable-rate coding scheme. In typical fixed-rate coding schemes, the codeword length n is chosen *a priori*, hence the time of decoding is fixed. This restriction immediately implies that outage events are inevitable, and that the outage probability is bounded away from zero. With such a scheme, the only way to reduce outage probability is to reduce the communication rate, and the system designer is faced with a trade-off between efficiency (i.e., rate) and reliability (i.e, outage). So with fixed-rate coding, the question is not a matter of *when* a message is decodable, but rather *whether* a message is decodable at some rate.

The reviving interest in feedback strategies [56] and incremental redundancy schemes [7], and the recent conceptual generalization of rateless codes from fountain codes [20] are now advocating a movement away from a fixed-rate methodology. That is, a message should only be decoded at a time when it *can* be decoded.

In this section we characterize the distribution of the time of earliest successful decoding when only k is fixed *a priori*. We are particularly interested in the scenario where k is of a similar order of magnitude as the channel coherence time, corresponding to slow-fading channels, delay-limited applications, or a combination of the two. We also demonstrate that these results degenerate to classical results in the limiting cases of fast-fading and quasi-static fading.

The scope of this work falls into the category of communication under channel uncertainty. Problems in this area are of practical significance, particularly for wireless channels. An information theoretic survey is given in [45]. Other researchers have made noteworthy contributions following this setting: The authors in [6] consider a non-ergodic block-fading model and determine that throughput is a discontinuous function of rate for multiple access channels under a hybrid-ARQ scheme. These results provide some additional validation of the discontinuous decoding times that will be presented in this section.

By examining the impact of the delay-constraints on block-fading channels, [1] presents a novel framework for determining and maximizing throughput for block coding methods. The work in [4] presents a graphical outage boundary-region concept which highlights the interaction between block coding and delay-constraints over block-fading channel models. It appears that this subject continues to attract research attention.

It is remarkable that even without any reference to feedback schemes or rateless

coding, the notion of earliest successful decoding time (referred to as decoding time hereafter) is a fundamental concept in its own right. Specifically one can view the minimal setup of a communication objective as the transfer of k bits reliably over a given channel. The decoding time is an intrinsic aspect of such a setup, indicating the intrinsic latency involved in communicating the k bits. This communication latency only depends on the channel characteristics and k, independent of whether there is any feedback or whether there is a delay, codeword length, or rate constraint. As such, one may view decoding time as a generic quantity underlying all communication problems with the same channel characteristics and information frame size k. Once decoding time is characterized, any performance metric can be derived. For example, one may immediately derive the outage probability by computing a (right) tail probability under the distribution curve of decoding times.

It is unfortunate that we see no closed-form expression obtainable for decoding time distributions in general. Nevertheless, we are able to characterize the decoding time distributions and present a numerical approach to allow the computation for arbitrary block-independent fading channels. Using this form we show that the decoding time distributions are generally discontinuous. This is the first time this observation is made, to the best of our knowledge. Decoding time distributions are computed for various underlying fading distributions, and are presented in the figures.

Examining the capacity equation in (2.2), we consider the impact of different fading models. We assume that the receiver can perfectly track the channel realizations. The transmitter is not assumed to have any knowledge of the channel realization. When the channel gain, h is random but remains constant for all time, then we have a slow fading situation where the delay requirement is short compared to the channel coherence time. Given some channel gain h, this is an AWGN channel with received SNR of $|h|^2\gamma$. The maximum rate of reliable communication supported by this channel is given by (2.4), a random variable. If the transmitter chooses some fixed-rate R to transmit at, regardless of how small this rate is, there will be a non-zero probability of outage. An alternative perspective is to say that the channel, conditioned on h, supports a rate equal to (2.4), and any rate of transmission smaller than this is achievable. Any rate of transmission greater than this will almost certainly result in an outage, i.e. it cannot be reliably supported.

Consider now extending this to time-varying fading models. Traditionally when dealing with time-varying fading, if we wish some target probability of outage (possibly zero), the length of the codeword must be large enough to average out both the Gaussian noise

and the channel gain h. Assuming this is possible, then the ergodic capacity (2.3) is achievable. To achieve this, the codeword must average out the fading, and thus must span many coherence time periods T_c. Depending on the nature of the fading, this can result in extremely long codewords. For applications which have a delay constraint that is small relative to the required codeword size, this ergodic capacity is not meaningful in practice. This practical issue is examined in greater detail in the following section, where the focus is on the number of channel uses – or codeword length – required to meet certain reliability criteria.

2.2.1 Distribution of Decoding Times

We are interested in characterizing the distribution of decoding times for the block-fading channel assuming that both k and T_c are very large and of similar magnitude. Since T_c is large then we can treat each block to behave as a quasi-static channel, for which there exists a code that can achieve the capacity given in (2.2). A proof for the existence of such a code is given in Section 2.4, but for now we will accept this assertion as true. Thus under an optimal coding scheme, in T_c channel uses we can transmit $C(T_c;\zeta)T_c$ bits of information. When dealing with the block-fading channel we will use the term $C_i := C((i-1)T_c+1, iT_c;\zeta)$ to simplify notation.

After N blocks, we are able to communicate up to

$$\sum_{i=1}^{N} T_c C_i \tag{2.6}$$

bits of information reliably. Alternatively, after $(N-1)T_c \leq n < NT_c$ channel uses, we are able to communicate up to

$$\sum_{i=1}^{N-1} T_c C_i + (n-(N-1)T_c)C((N-1)T_c+1, n;\zeta) \tag{2.7}$$

bits of information reliably, and it is simple to verify that the sum of (2.7) is less than or equal to (2.6). If k is less than or equal to the sum in (2.7) then it is possible to reliably decode after n channel uses.

Rewriting this condition, we obtain the inequality

$$\frac{k}{T_c} < \sum_{i=1}^{N-1} C_i + (\frac{n}{T_c} - (N-1))C((N-1)T_c+1, n;\zeta). \tag{2.8}$$

We are now in a position to tackle the original question: When can we reliably decode on this channel? Putting the question into the context of (2.8), suppose that we are given some large k and T_c. Then there exists a minimal n that satisfies (2.8), and since realized capacity is a random variable, so is n. Thus the complete answer to our question is the characterization of the distribution of n, or equivalently, of n/T_c. [2]

We now derive the form of the probability density function (pdf) $f_{n/T_c}(n/T_c)$, where we use the symbol f here to represent a distribution (probability measure) for continuous, discrete or mixed-typed random variables, and the random variable drawn from the distribution is labeled as the subscript.

It is useful at this point to provide some insight into the meaning of k/T_c. One may interpret this value in (2.8) as a sum-rate, which is partitioned into N sub-channels. From this perspective, it can then be seen as an achievable rate bound for independent users on a multiple access channel. It is the ratio k/T_c that is of importance and fundamental in this case, as opposed to either k or T_c considered individually.

Let a partial sum process S_x be defined as

$$S_x := \sum_{i=1}^{x} C_i, \tag{2.9}$$

with boundary condition $S_0 = 0$. We make use of the fact that n/T_c can be expressed in terms of an integer and fractional component. Specifically, let $n/T_c = L - 1 + r$, where L and r represent the integer and fractional components, respectively. Note that

$$L = \min\left\{x : S_x > \frac{k}{T_c}\right\} = \frac{n - (n \mod T_c)}{T_c} + 1,$$

$r := \frac{n \bmod T_c}{T_c}$, and n/T_c may be identified with the pair (L, r).

Rewriting (2.8) gives

$$\frac{k}{T_c} \leq \sum_{i=1}^{L-1} C_i + \frac{(n \mod T_c)}{T_c} C_L = S_{L-1} + rC_L.$$

and shows that the last block will (likely) be used for some fraction of T_c before equality

[2] We note that regarding n as the decoding time is valid if T_c is large. For relatively small T_c, n only approximates the decoding time.

is achieved. The expectation for this equation is

$$E\left[\frac{k}{T_c}\right] \leq E\left[\sum_{i=1}^{L-1} C_i + \frac{(n \bmod T_c)}{T_c} C_L\right] \tag{2.10}$$

$$= (L-1)E[C] + \frac{(n \bmod T_c)}{T_c} E[C] \tag{2.11}$$

$$= \frac{n}{T_c} E[C] \tag{2.12}$$

where we make use of the fact that $n = (L-1)T_c + (n \bmod T_c)$.

The problem now becomes one of finding the distribution for the number of T_c's required to transmit k information bits, assuming a reliable communication rate equal to the instantaneous capacity. We view the process S_x as a random walk and wish to find the time at which the walk exceeds k/T_c. Standard analysis of biased random walks with stopping times can be used to find $f_L(L)$ in a straightforward manner (see, for example [28]). Numerical methods are required, since the pdf of (2.9) can generally not be found analytically. In the case of computing $f_{L,r}(L,r)$, the process is a random walk, but one in which the walk is held at a fixed "speed" or "direction" for T_c time steps before a new value is chosen. This complicates the evaluation, but it is still possible and proceeds as follows.

The values S_L and S_{L-1} denote the partial sums as defined in (2.9) after L and $L-1$ blocks of channels uses respectively. Due to dependence between these random variables, we work with joint distributions. Given the joint distribution $f_{L,S_L,S_{L-1}}(L, S_L, S_{L-1})$ induced by the underlying fading process, we have the following lemma.

Lemma 1 *For the block-fading channel given in (2.1) and large, fixed k and T_c, the probability density function $f_{L,r}(L,r)$ has the form*

$$f_{L,r}(L,r) = \int_{C_L} |C_L| \, f_{L,S_L,S_{L-1}}\left(L, (1-r)C_L + \frac{k}{T_c}, \frac{k}{T_c} - rC_L\right). \tag{2.13}$$

Proof:

Define $C_r := \frac{k}{T_c} - S_{L-1}$. This random variable represents the information per channel use that remains to be transmitted to the receiver after $(L-1)T_c$ channel uses and is used below to simplify the expression for r. Similarly, C_L represents the information per channel use which results in the total received information to be greater than or equal to k.

The joint probability density function $f_{L,C_L,C_r}(L, C_L, C_r)$ can then be computed as a transform of $f_{L,S_L,S_{L-1}}(L, S_L, S_{L-1})$ following standard pdf transform theorems (see for example, [28], Theorem 4.7.4). Thus,

$$f_{L,C_L,C_r}(L, C_L, C_r) = f_{L,S_L,S_{L-1}}(L, C_L - C_r + \frac{k}{T_c}, \frac{k}{T_c} - C_r).$$

The fractional component r can be rewritten as

$$r = \frac{C_r}{C_L}.$$

Now we can find the joint probability $f_{L,r}(L, r)$ by first transforming the joint pdf $f_{L,C_L,C_r}(L, C_L, C_r)$ into $f_{L,C_L,r}(L, C_L, r)$ given by

$$f_{L,C_L,r}(L, C_L, r) = |C_L| \, f_{L,C_L,C_r}(L, C_L, rC_L),$$

then by marginalizing over C_L, yielding

$$f_{L,r}(L, r) = \int_{C_L} f_{L,C_L,r}(L, C_L, r).$$

Finally, since $n = T_c(L-1+r)$ then $f_{n/T_c}(n/T_c)$ is found simply by a look-up of $f_{L,r}(L, r)$. This is a consequence of the fact that the domains of these two functions are bijective. Thus

$$f_{n/T_c}(n/T_c) = f_{L,r}\left(\frac{n - n \mod T_c}{T_c} + 1, \frac{n \mod T_c}{T_c}\right)$$

Simple arithmetic operations result in the form of the lemma.

□

Under the channel model given in (2.1), let f_h be the pdf of the fading coefficient. Let $\mathscr{F}_h$ be the space of all possible f_h. That is, $\mathscr{F}_h$ is the set of all functions f_h mapping the set of complex numbers to the set of non-negative real numbers with

$$\int_{h \in \mathbb{C}} f_h(h) dh = 1.$$

It is easy to see that $\mathscr{F}_h$ is an uncountable, convex set.

Let $\mathcal{P}$ be an arbitrary probability measure on $\mathscr{F}_h$ satisfying, that for any $f \in \mathscr{F}_h$, the event $\{f\}$ under $\mathcal{P}$ has probability 0. In other words, $\mathcal{P}$ is an arbitrary continuous distribution on $\mathscr{F}_h$.

Given this setup, a consequence of the above Lemma is the following theorem.

Theorem 1 *For any large, fixed k and T_c, draw f_h from $\mathcal{P}$ at random. Then $f_{n/T_c}(n/T_c)$ is discontinuous at every integer value with probability one.*

Proof: Let C be the random variable from which the i.i.d. sequence $C_1, C_2, \ldots$ is drawn. If C is discrete, then the theorem is clearly true. Thus assume C is a continuous random variable. Assume the support of $f_{L,r}(L,r)$ is not limited to $L = 1$. If this were true, then we have a static-fading channel and the theorem does not apply. Similarly, we assume that a majority of the probability density exists for finite L.

Let $\mathscr{F}$ be the space of all such $f_{L,r}$ and g be the mapping from f_h to $f_{L,r}$. It is clear that $\mathscr{F}$ is a continuous, uncountable space. Let Q be the continuous probability measure on $\mathscr{F}$ induced by $\mathcal{P}$ on $\mathscr{F}_h$.

The distribution $f_{L,r}$ is continuous if and only if

$$\lim_{r \to 1-} f_{L,r}(L,r) = \lim_{r \to 0+} f_{L,r}(L+1,r) \tag{2.14}$$

for every integer L in its support. The left-hand side and right-hand side of (2.14) have different parametric forms since the left-hand side involves one additional, independent random variable. As a result, the set of all $f_{L,r} \in \mathscr{F}$ satisfying (2.14) is a set with zero probability measure under Q. The left limit is a function of $L+1$ random variables, while the right limit is a function of the same $L+1$ variables and an additional independent random variable. As such, the parametric forms of the left and right limits differ, and by independence, there is no reason to expect that the two limits are equal. To prove this point, let Ω be the sample space of all possible ζ. Only a subspace of Ω can induce a distribution for C such that there is equality between left and right limits. Since Ω is a continuous space, any subspace must have zero "volume" and thus zero probability. Therefore $f_{L,r}(L,r)$ is almost surely discontinuous at these points.

□

Another consequence of Lemma 1 is simply that there exists a computable form for $f_{n/T_c}(n/T_c)$, i.e. we can readily determine the distribution of n. As such, we can make use of the distribution for the optimization of a communication strategy as will be mentioned in Section 2.2.2.

To give a complete recipe of the numerical procedure, we describe how $f_{L,S_L,S_{L-1}}$ may be computed. For each positive integer l, compute

$$f_{S_l}(s) := \overset{*}{\prod_{i=1,\ldots,l}} f_{C_i}(s),$$

where $\prod^*$ denotes the convolutional product [3] and f_{C_i} is the distribution f_C of the realized capacity over a fading block and independent of i. For each positive integer L, compute

$$f_L(L) := \int_{k/T_c}^{\infty} (f_{S_L}(s) - f_{S_L}(s))ds.$$

For each positive real number S_L and positive integer L, compute

$$f_{S_L|L}(S_L|L) := f_{S_L}(S_L) \cdot U(S_L - \frac{k}{T_c}) \int f_{S_L}(S_L) \cdot U(S_L - \frac{k}{T_c})ds$$

where $U(\cdot)$ is the unit-step function. Compute

$$f_{S_{L-1}|S_L,L}(S_{L-1}|S_L, L) := f_c(S_L - S_{L-1}) \cdot U(\frac{k}{T_c} - S_{L-1}) \int f_c(s)U(s - S_L + \frac{k}{T_c})ds.$$

Finally, compute

$$f_{L,S_L,S_{L-1}}(L, S_L, S_{L-1}) := f_L(L)f_{S_L|L}(S_L|L)f_{S_{L-1}|S_L,L}(S_{L-1}|S_L, L)$$

following the chain rule of probability. We note that the integration arising in the above equations may be computed numerically.

2.2.2 Discussion

The wireless applications for which the studied channel model might be applicable, e.g. high data-rate, frequency-hopped or OFDM systems, may have values of k/T_c that range from less than 0.1 to over 10. We present decoding time distributions in Figure 2.2 for underlying Rayleigh and low K-factor $K = 0.1$ Rician fading distributions at $k/T_c = 5$ and $\gamma = 5$dB[4]. Discontinuities are clearly seen at integer values of n/T_c as proved in Theorem 1. Also presented in the same figure is the decoding time distribution for a uniformly distributed realized capacity with mean (and variance) chosen to be equal to the mean capacity (and variance) induced by the Rayleigh fading distribution. This distribution was chosen to demonstrate that the discontinuities exist independent of the form of the underlying fading distribution. The similarity of this curve to the others reaffirms this fact. Figure 2.3 further demonstrates the nature of the decoding time distribution for $k/T_c = 2$ and $\gamma = 5$dB. Here we see the discontinuities more pronounced for the smaller value of k/T_c, particularly for the uniform capacity curve.

[3] The convolutional product is defined as the convolution of each of the terms in an analogous manner to the multiplicative product.

[4] High K-factor Rician distributions are of little interest here as they approach delta functions.

An interesting consequence of these discontinuities is that, for example, if a codeword has not yet been received after a number of channel uses greater than expected, it is more likely that the codeword will be received at the beginning of a subsequent coherence block than at the end of the current block.

The discontinuities are a consequence of the block-fading channel and coding over a small, finite number of blocks. One expects that these discontinuities would not appear in models with "smoother" fading characteristics, e.g., a Markov-modulated block-fading channel. Given this model, however, interesting decoding behavior may be observed. For decoding times less than the expected value, it is unlikely that the realized capacity over one block of T_c channel uses is sufficient to complete the reception of k-bits, but if this were to occur, it is more likely to require a majority of the T_c channel uses. For times greater than expected, it is likely that a small fraction of k bits remain to be received and so a small number of channel uses are likely required. If a poor fading state is realized for a block, it is more likely that reception will complete shortly after a new fading state is realized in the following block.

Asymptotically, as k/T_c goes to zero, the channel behaves as though it were quasi-static fading. Then n/T_c will necessarily remain less than unity and the decoding time distribution will be precisely the expected result induced by the underlying fading distribution. A similar observation is made with respect to the limit as k/T_c grows large, where one may see the distribution becoming smooth and approaching a delta function at the decoding time (rate) corresponding to the ergodic capacity of the channel. This then becomes equivalent to a fast-fading model, and demonstrates the validity of Lemma 1 over a broad range of channel conditions. An example of this behavior is given in Figure 2.4, where the distribution is given as a function of n/k (or inverse rate) for an underlying Rayleigh fading distribution and varying values of k/T_c. The trend toward a smooth, delta function is seen in the figure as k/T_c increases from 1 to 20.

The use of Lemma 1 allows us to determine when to decode from an information theoretic perspective, i.e. assuming ideal coding, and it provides some insight into more practical issues such as coding strategies, throughput and outage optimization.

The support of these decoding time distributions may span many blocks and this suggests that fixed-rate coding schemes will suffer when the channel realization is not known at the transmitter. In fact, transmitting a codeword when n/T_c is chosen *a priori* will almost surely result in either an outage or an inefficient use of the realized capacity. Standard approaches to combat this problem, for example ARQ or block-incremental redundancy schemes, while making up for much of the deficiency of fixed-rate schemes,

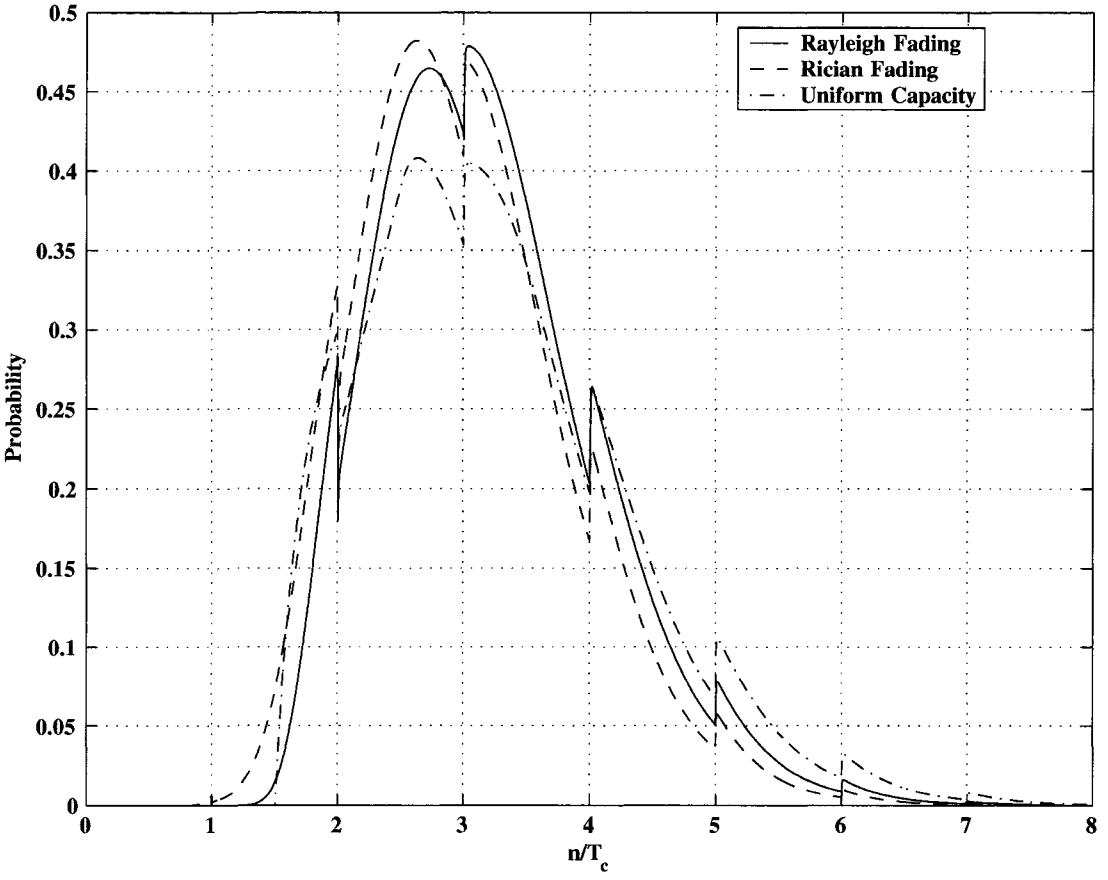

Figure 2.2: Distribution of decoding times, $k/T_c = 5$, $\gamma = 5$dB

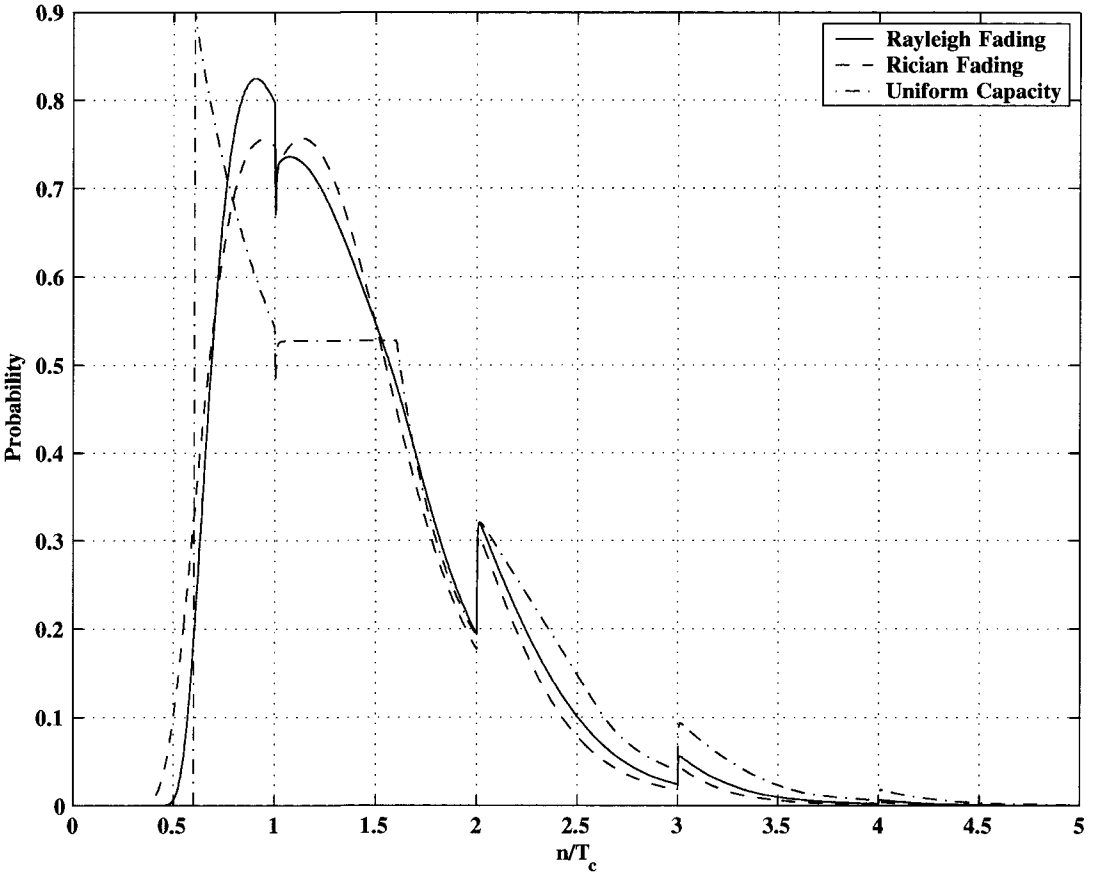

Figure 2.3: Distribution of decoding times, $k/T_c = 2$, $\gamma = 5$dB

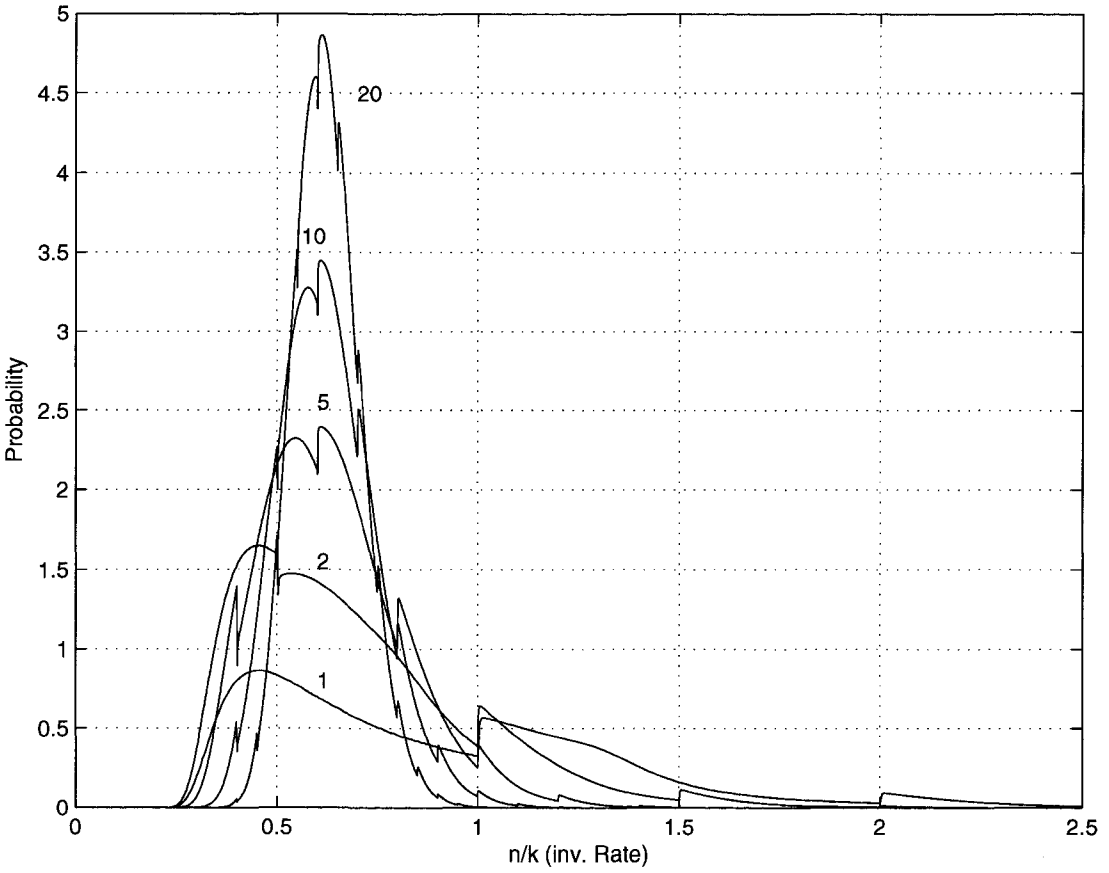

Figure 2.4: Distribution of decoding times for varying k/T_c, Rayleigh fading, $\gamma = 5$dB

fail to fully exploit the decoding opportunities in these distributions. This is particularly severe for smaller k/T_c as can be seen for example in Figure 2.3.

Coding strategies which can exploit all decoding opportunities are rateless, e.g., Raptor codes [66], which have been shown to perform well over wide ranges of SNR for fading and AWGN channels [59, 9], though they are generally not capacity-achieving. We address this issue in detail in Section 2.4.

Since the decoding time distribution is readily computable, it is a simple matter to determine the theoretical outage probability that some delay-constraint imposes. When the delay-constraint is expressed in terms of channel uses, it is a simple matter to compute the outage probability by determining the amount of probability mass beyond the constraint. In a similar manner, these distributions can easily be applied to determine throughput of various communications protocols over block-fading channels [14]. This is discussed in detail in Chapter 4.

Reliable communication over unknown channels is a problem of practical concern for wireless communications. The notion of transmitting codewords without fixed n *a priori* is a fundamental concept that is gaining attention and we have presented a characterization of the distribution of decoding times for block-fading channels. Additionally, we have demonstrated that these distributions are generally discontinuous and have shown that they are computable numerically.

These characteristics of these distributions suggest that it is advantageous to use rateless codes for reliable communications, and computability of the distributions makes them a useful tool for the analysis of rateless code performance. Metrics including outage probability and throughput can be determined for any channel configuration, from slow to fast fading.

2.3 Rateless Coding

The previous section provided us with a solid motivation for the use of rateless codes for wireless channels. Given the uncertainty in the time or rate at which a codeword may be reliably decoded, we seek a coding framework that allows us to adapt to the realizations of the channel. Rateless codes provide just such a means, which is now formally introduced and studied in detail.

Rateless codes are codes that encode a finite number of messages but have an infinitely long block length and are thus parametrized by a single number k, the length in bits of the information block. Comparatively, fixed-rate block codes are parametrized by the pair

(k, n), where n defines the codeword length. The transmission of a rateless codeword is terminated when the receiver decodes the message and uses a feedback channel to communicate an `ACK` to the transmitter. As indicated by its name, a rateless code does not have a fixed rate, but rather the rate is determined "on the fly" by the time at which the receiver decodes the message.

As the first efficient class of rateless codes, fountain codes not only have low complexity but also have capacity-achieving or capacity-approaching performance for several classes of channel models. LT Codes, introduced by Luby in [48] were shown to achieve capacity for any BEC. In practice, LT codes are prone to have a noticeable error floor for small k. Raptor codes were introduced by Shokrollahi in [66] and use LT codes as an inner code with a high-rate LDPC outer code. Raptor codes were also shown to be capacity achieving for the BEC, but also to have the beneficial properties of little or no error floor for small block lengths, and a linear-time encoding and decoding computational complexity.

Given their many useful properties, Raptor codes and their performance over other channels have been investigated, see e.g., [59, 23, 9]. Very good performance has been found for many other channels including the AWGN channel and various types of fading channels.

As described in [66], a Raptor code is defined by parameters $(k, \mathcal{C}, \Omega)$, where k is an information block length, $\mathcal{C}$ is a linear code of block-length n' and dimension k acting as the outer code, and Ω is a degree distribution controlling the LT inner code. The code $\mathcal{C}$ is called the pre-code of the Raptor Code. The input symbols of a Raptor Code are the k symbols used to construct the codeword in $\mathcal{C}$ consisting of n' intermediate symbols. The output symbols are the symbols generated by the LT-Code from the n' intermediate symbols.

An example of a fountain (LT) code over GF(2) operating over the binary erasure channel is described here, with these restrictions chosen for clarity of exposition. The transmitter and receiver are assumed to be synchronized in some manner so that the receiver is able to correctly reproduce the generator matrix used by the transmitter and identify received symbols. This may be as simple as sharing a common clock source. Both transmitter and receiver are initialized at time zero. Given some block of k-bits of information, the transmitter consecutively generates codeword symbols, and each symbol transmission corresponds to a single time-step.

The generation of a codeword symbol is a two-step process. The first step requires the transmitter to draw a pseudo-random number from some *a priori* known distribution Ω

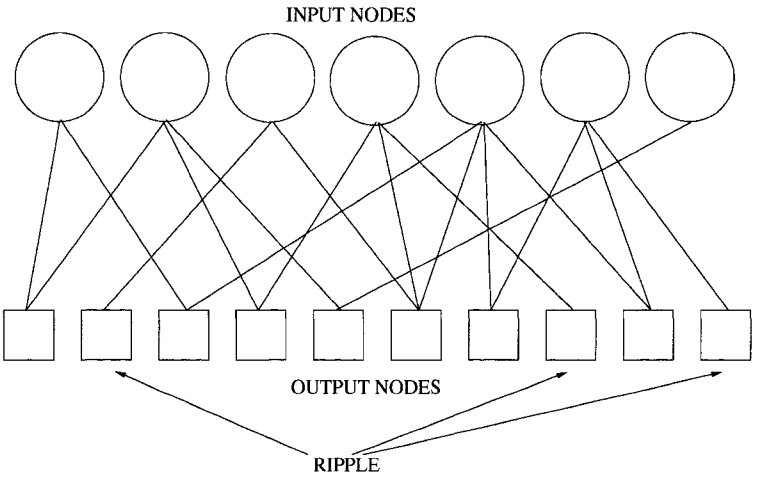

Figure 2.5: Factor graph decoding of a fountain code

over the positive integers. The receiver also knows this distribution and is assumed to be able to independently draw the identical value as the transmitter. Given this value $d \leq k$, the transmitter pseudo-randomly selects d distinct information bits uniformly. These bits are XOR'ed together, and the result is transmitted over the channel. This operation is equivalent to pseudo-random, "on-the-fly" generation of the generator matrix $\mathbf{G}$.

The receiver is capable of reproducing the identical pseudo-random realization that is used to generate codeword symbols at the transmitter and in this way is able to maintain track of $\mathbf{G}$. The receiver also tracks, in some manner, the amount of information it has received. At some point it determines that it may be possible to successfully decode, and then attempts to do so. The algorithm used to decode may depend on the implementation, though the iterative belief propagation algorithm is often acceptable [40]. In the case of the binary erasure channel, belief propagation amounts to a graph pruning procedure, which is described in detail in [48].

Following the terminology of [48, 66], define the factor graph [40] associated with a rateless code at time t to consist of k variable nodes called the *input nodes*, and x check nodes called the *output nodes*. The number of check nodes x is equal to the number of non-erased symbols received at the decoder at time t. Each of the x output nodes has edges connecting it to every input node that was used to generate it, based on the distribution Ω.

An example factor graph for a rateless code is presented in Figure 2.5 for $x = 10$. For the binary erasure channel, decoding attempts to prune the graph to "cover" all of the input nodes. An input node that is covered is a node for which there is no uncertainty

about the value it must take, and occurs whenever it is connected to a single output node. The decoding process operates as follows: Define the set of all degree-one output nodes to be the *ripple*. If this set is empty, terminate decoding. Choose a node from the ripple and cover the unique, connected input node with the value of this node. Remove both the output node and input node from the graph. If no input nodes remain in the graph, decoding has completed successfully and the received message is given by the value of each of the covered input nodes.

If decoding is successful, the decoder sends an `ACK` to the transmitter using a feedback channel to terminate transmission. Otherwise, the decoder will simply collect some further number of codeword symbols and attempt decoding again. Using this approach, it is almost always possible for the receiver to successfully decode the transmitted codeword.

The choice of distribution Ω is critical to decoding performance. It was shown by Luby in [48] that for the binary erasure channel there exists an Ω which results in a rateless code that achieves the capacity of *any* erasure rate on the channel. This remarkable distribution is termed the soliton distribution for its resemblance to soliton pulses. Unfortunately this distribution is only optimal for $k \rightarrow \infty$. However, Luby also defined the robust-soliton distribution, a slight modification of the soliton distribution that works very well for finite k, and can achieve rates arbitrarily close to capacity on the binary erasure channel.

Over other channels, there is no known distribution that is "universal" in the sense that the soliton distribution is for the binary erasure channel. In fact, it was shown in [23] that there does *not* exist such a distribution for the AWGN channel. Despite this fact, excellent results over large ranges of channel parameters have been shown for rateless codes over the AWGN and binary symmetric channel [59, 23]. Detailed performance results over fading channels are provided in Section 2.5.

Given these promising results, we wish to investigate the theoretical properties of rateless codes. In particular, we explore their properties over fading channels and show in the following section that "good" rateless codes exist over fading channels.

2.4 Rateless Codes for Fading Channels

The goal of this section is to prove the existence of "good" rateless codes over fading channels. By good, we refer to the ability of the rateless code to approach the capacity of the unknown channel for *all* possible channel realizations, simultaneously. This is similar in concept to the notion of universality that is used to describe source coding [17], and we

use the term universal here to have a loose, but similar meaning; that the performance of the code is in essence independent of the channel realization.

Generalizing the definition of a Raptor code, we say that a rateless code $\mathcal{C}(k, f, \{\phi_n : n = 1, 2, \ldots\})$ is parametrized by an integer number of bits k, an encoding function f and a sequence of decoding functions $\{\phi_n : n = 1, 2, \ldots\}$. The encoder maps each message $w \in \{1, 2, \ldots, 2^k\}$ to a codeword $(x_1, x_2, \ldots)$ — a semi-infinite sequence of complex numbers — using encoding function f and sequentially transmits the codeword symbols across the channel. In this way the transmitter acts as a continuous source of codeword symbols, invoking the "fountain" terminology commonly used to describe these codes.

There is usually a prescribed set $\mathcal{N}$ of time instants at which the decoder may attempt to decode. At some time instant $n \in \mathcal{N}$, the decoder may attempt decoding using the received complex numbers $\{y_1, y_2, \ldots, y_n\}$ and decoding function ϕ_n, computing the message $\hat{w} := \phi_n(y_1, y_2, \ldots, y_n)$. If the decoder is sufficiently confident that $\hat{w}$ is the transmitted message according to a criterion to be specified, it declares $\hat{w}$ as the transmitted message and sends an `ACK` to the transmitter to terminate the transmission of this codeword. The feedback is assumed reliable and instantaneous. If the decoder is not confident, it waits until the next decoding opportunity in $\mathcal{N}$ and may attempt to decode again. This process repeats until the decoder can confidently decode. The decision whether to attempt decoding or not is entirely up to the decoder. See [21] for a treatment of this area. For simplicity we drop f and $\{\phi_n\}$ when referring to a rateless code, and simply denote it by $\mathcal{C}(k)$.

Assuming the transmission of a codeword begins at time 1, we use $P_e(\zeta, n; \mathcal{C}(k))$ to denote the maximal probability of error for rateless code $\mathcal{C}(k)$ over channel ζ at decoding time n, where an error event at time n is the event that the decoder chooses to decode and declared message $\hat{w}$ at time n is not equal to the transmitted message w. The maximization is over all $w \in \{1, 2 \ldots, 2^k\}$. Note that $P_e(\zeta, n; \mathcal{C}(k))$ is well defined even for $n \notin \mathcal{N}$, and generalizes to arbitrary transmission times.

For any given $\epsilon > 0$, we define the *ϵ-error realized rate* (or simply *realized rate*) of code $\mathcal{C}(k)$ over channel ζ as

$$R(\epsilon, \zeta; \mathcal{C}(k)) := \frac{k}{N(k, \epsilon; \mathcal{C}(k))},$$

where $N(k, \epsilon; \mathcal{C}(k))$ is defined to be the smallest $n \in \mathcal{N}$ such that $P_e(\zeta, n; \mathcal{C}(k)) < \epsilon$.

Theorem 2 *Under the quasi-static fading model, for any ϵ and $\delta > 0$, there exists a sufficiently large k and rateless code $\mathcal{C}(k)$ whose realized rates satisfy $R(\epsilon, \zeta; \mathcal{C}(k)) > C(N(k, \epsilon; \mathcal{C}(k)); \zeta) - \delta$ for all channel realizations $\zeta \in \Omega$ simultaneously.*

In this theorem, ϵ characterizes a notion of reliability, δ characterizes a notion of efficiency, and the results that reliability and efficiency can be achieved for all $\zeta \in \Omega$ simultaneously characterizes a notion of "universality". The proof of this theorem is given in Appendix 2.A.

We note that for the quasi-static fading channel, $C(n;\zeta)$ depends only on ζ. In this setting, a natural strategy is to use the feedback channel to signal the channel information to the transmitter and then use a fixed-rate block code, achieving $C(n;\zeta)$.

The following theorem is a generalization of Theorem 2 to block-fading channels.

Theorem 3 *Under the block-fading model, there exists a $\tilde{T}$ and $\tilde{k}$ such that for all $T_c > \tilde{T}, k > \tilde{k}$, and any ϵ and $\delta > 0$, there exists a rateless code $\mathcal{C}(k)$ whose realized rates satisfy $R(\epsilon,\zeta;\mathcal{C}(k)) > C(N(k,\epsilon;\mathcal{C}(k));\zeta) - \delta$ for all channel realizations $\zeta \in \Omega$ simultaneously.*

An overview of the proof is provided here. Details of the proof are given in Appendix 2.A. The receiver in the proof decodes at the time corresponding to the smallest $n \in \mathcal{N}$ such that the realized rate $R = \frac{k}{n} < C(n;\zeta)$. Since the decoder has the channel state information, it knows the realized capacity up to the end of the current coherence block, and whether the codeword rate will be less than this value by the end of the block. If the codeword rate is not sufficiently low by the end of the block, it waits for reception of the next block. Once it determines that decoding is possible, the decoder treats the entire received sequence as a single punctured codeword, where punctured symbols correspond to those not yet received. It is shown that successful decoding is always possible under standard constraints for all possible channel realizations.

Although sufficient for proving the existence of a rateless code that is efficient and reliable for all ζ, a limitation of the decoding scheme used in the above proof is that T_c must be large for a main ingredient of the proof — the Asymptotic equipartition property (AEP) — to hold, particularly in dealing with the last, possibly partially received, block. Maximum likelihood decoding can be used to circumvent this limitation [21].

Another consequence of Theorem 2 is the existence of good rateless codes for the AWGN channel. Fixing the channel gain h to a value of 1 reduces this channel to the AWGN channel, and thus Theorem 2 applies directly.

Having shown the existence of good rateless codes, we turn to the application and implementation of these codes over fading channels.

2.5 Rateless Code Implementation

The previous section dealt with the existence of good rateless codes, but the proofs are not constructive. Fortunately, there exist implementations of rateless codes such as Raptor codes which have been shown to be efficient for the binary erasure channel. We are interested in simultaneous efficiency and reliability in this work, and explore in this section these aspects for rateless code implementations.

Although there has been research literature addressing some of the above aspects (e.g., see [72, 90, 49, 6]), we are not aware of any theoretical study that considers the aspects of efficiency, reliability, and robustness jointly at the physical layer. Commonly, these issues have been examined at higher layers. We argue that when channel state information is not available at the transmitter, conventional fixed-rate communication schemes suffer in all three aspects.

In typical fixed-rate schemes, the transmitter chooses a fixed code rate k/n and code length n based on given channel statistics and delay constraints. First, since the realized channel capacity $C(n;\zeta)$ is a random variable from the transmitter's perspective, there is no hope for the chosen rate k/n to closely follow the realized capacity. Second, for any given positive rate k/n, the outage probability is strictly bounded above zero. This simply follows from the non-zero probability that $k/n > C(n;\zeta)$. This phenomenon is particularly severe when the delay constraints are such that the number of fading blocks n/T_c is relatively small and the codewords experience a limited subset of the possible channel behavior. Third, with a fixed-rate scheme, there is an unavoidable tradeoff between efficiency and reliability. More specifically, we must pay a penalty in rate in order to ensure some desired level of reliability, and conversely, if we wish to transmit at a higher rate, we are forced to accept larger outage probability. When channel statistics are known at the transmitter, a fixed-rate system may in principle operate at an optimal tradeoff point, i.e., at the highest rate subject to a certain outage requirement, but when the channel statistics are unknown, inaccurately estimated or varying, this operating point is no longer optimal and the system becomes either less efficient or less reliable.

Inspired by results in Section 2.4, we propose to use rateless codes for communication over fading channels when the channel state information is unknown at the transmitter. Here we focus on single-antenna systems and the fading channel model expressed by (2.1), though generalization of rateless codes to vector-valued codeword symbols (i.e., space-time codes) for multiple antenna systems is straightforward.

Our rationale of using rateless codes to achieve efficiency, reliability and robust-

ness is in fact quite intuitive. First, unlike fixed-rate codes in which the rate k/n is a constant and independent of channel realizations, rateless codes realize a rate for each codeword depending on what channel realization is experienced by the codeword. Since the successful-decoding time n depends on the channel realization, a good rateless code may naturally adapt to the channel state thereby providing opportunities for the realized rate k/n to closely follow the realized capacity $C(n;\zeta)$. Second, as suggested in [20], for a large class of discrete channels, choosing the codebook size 2^k to be sufficiently large may drive the outage probability to zero without sacrificing rate. Third, we note that knowledge of channel statistics at the transmitter plays no essential role with rateless codes. If a rateless code can realize a rate that is close to the realized capacity in every transmission, it can then closely follow the average realized capacity regardless of the channel statistics. Thus a rateless code can maintain its efficiency and reliability across all channel ensembles, i.e. it can be robust.

2.5.1 Fountain Code Implementation

We simulate a rateless code for a single antenna system and over the channel described earlier. For a proof-of-concept to validate the results of Section 2.4, we choose a Raptor code as presented in [66] and follow the construction given in [59] with $k = 9,500$ and QPSK modulation. The fading variable h_i is real-valued with variance following a Rayleigh distribution. We choose the coherence time T_c to be $1000, 5000$, and ∞ (quasi-static fading), and simulate various transmit SNR settings. For each transmitted codeword, belief propagation decoding is first attempted after $k/2$ channel uses, and then periodically at intervals of 100 channel uses. Given the value of k, this represents an approximate resolution of 0.1% of k which is sufficiently accurate to generate reasonable performance metrics. The initial messages for belief propagation are calculated based on the received signal and the receiver's (perfect) knowledge of channel gains and noise variance. In every decoding attempt, as the passing of messages iterates, the decoder keeps examining whether the hard decisions on the messages form the transmitted codeword; when this is the case, a successful decoding is declared and the transmission of this codeword is terminated; when the hard decisions on the messages do not form the transmitted codeword within 100 iterations, the current decoding attempt is stopped, and the decoder waits until the next decoding attempt to decode again. We note that in this rule for stopping a decoding attempt and declaring a decoding success, we essentially assume that the receiver knows whether it decodes correctly. In practice, this may

correspond to the case where CRC bits are embedded within the k bits, and we note that the rate loss in such an implementation is negligible. Other rules for stopping a decoding attempt and for declaring a decoding success, such as those based on the reliability of the decoded word, are also possible. A study of this and other practical issues related to the decoding of rateless codes are presented in [35, 34, 15]

Figure 2.6 presents the curves of averaged realized rate and averaged realized capacity (under QPSK input constraint) versus average receive SNR. Here the averaged realized rate is defined as the total number of transmitted information bits divided by the total number channel uses for transmitting these information bits. Notice that for all three choices of coherence time T_c, the averaged realized rate continuously scales with the averaged realized capacity over the broad range of SNR, with some loss (about $10-15\%$) due to code suboptimality. This plot demonstrates the efficiency of this scheme and the advantage of rateless codes over fixed-rate codes. We note that this advantage is not only in terms of rate, but also in terms of outage — as discussed earlier, there is simply no outage event. As the SNR increases, one may observe that the capacity curve reaches the asymptote of 2 bits/channel use, which is governed by the limit enforced by the QPSK modulation. It is interesting to note that the best performance is observed at $T_c = 5000$. Deserving further investigation, this non-monotonic performance trend with respect to T_c may possibly suggest that the design of rateless codes need to take channel coherence time into consideration, and that a given rateless code may perform the best over a certain range of T_c.

Although Raptor codes are not universal for fading channels in the strict sense, these results suggest a behavior that tracks capacity across the studied range of channel statistics, subject to the constraints of QPSK modulation. Similar results are also observed in Figure 2.7, where along with the realized capacity upper limit (dashed lines), the realized rate versus realized capacity for each codeword transmission is plotted for an average receive SNR equal to 0 dB and $T_c = 1000$ and ∞ respectively. One should observe a much tighter variation in the scattering of rate-capacity pairs for the smaller value of T_c. This is because for the same average receive SNR, when T_c is smaller, codewords span a larger number of fading blocks, and the realized capacities are more concentrated around the ergodic capacity.

It is instructive to plot a curve of outage probability to make comparison with the theoretical limits (2.5). Here for a given communication rate target R_T, the natural definition of realized outage probability for a rateless code is

$$P_{\text{out}} = P[n > k/R_T], \tag{2.15}$$

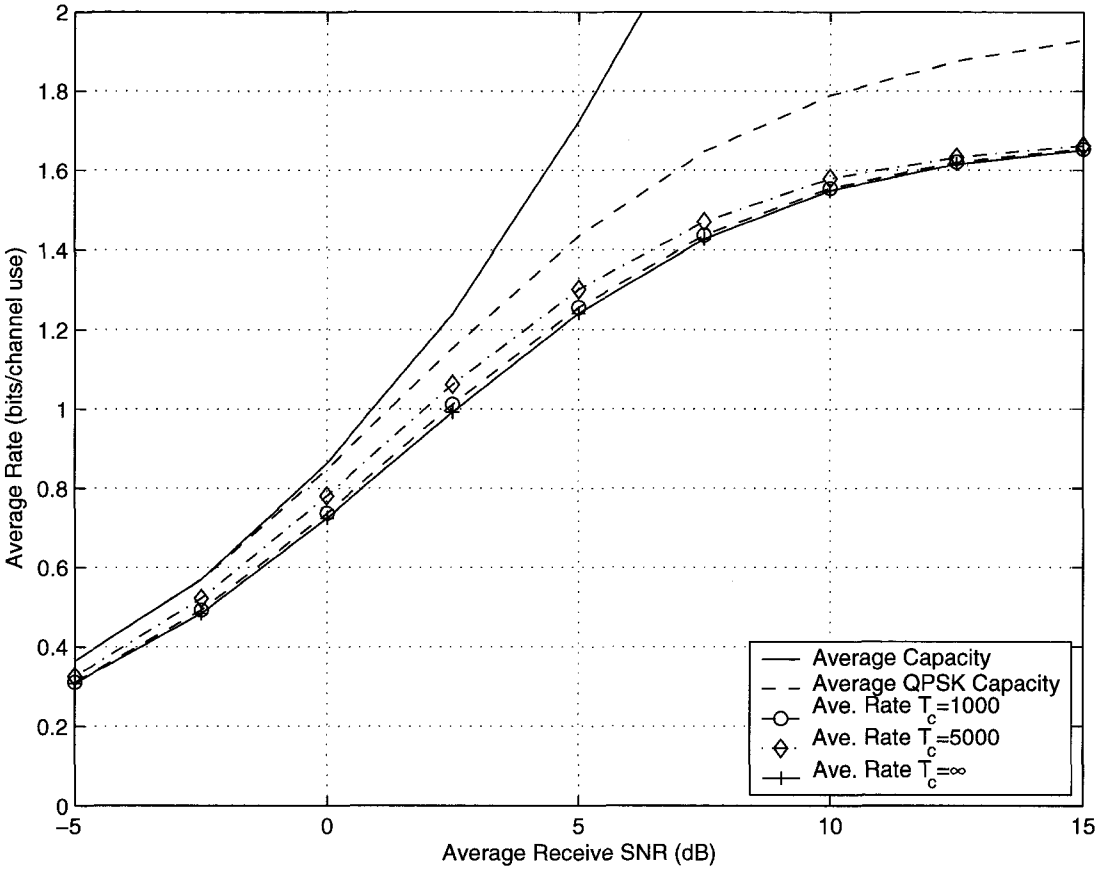

Figure 2.6: Average realized rate and average realized capacity vs. average receive SNR

i.e., the probability that a codeword can not be decoded before time k/R_T (or equivalently the realized rate k/n is lower than the target rate R_T). In this context, the theoretical limit of outage probability in (2.5) becomes

$$P_{\text{out}} = P[R_T > C(n;\zeta)], \tag{2.16}$$

where for a fair comparison with the realized outage probability in a rateless coded scheme, ζ is interpreted as the realized channel sequence $(h_1, h_2, \ldots, h_n)$ experienced by a rateless codeword decoded at time n. Figure 2.8 contains the realized outage probability curves for the rateless code simulated and the corresponding theoretical limit with $R_T = 0.5$. The actual outage probability follows the theoretical curve quite closely over the entire SNR range. We note however that this plot does not truly reflect the outage behaviors of rateless codes, since after all there are no outage events in rateless coding scheme, given a sufficient number of channel uses. A proper interpretation of this figure should be from the "user-satisfaction" perspective. That is, suppose the user desires to communicate at rate R_T subject to certain delay requirement, then these curves indicate how often the user is dissatisfied under the optimal fixed-rate coding strategy and under the simulated rateless coding strategy. — In terms of the true sense of outage, rateless codes always beat the outage limits for fixed-rate codes. This will be explored further in Chapter 4.

2.6 Discussion and Conclusions

The proposed framework is motivated by practical communication over fading channels. Our approach demonstrates advantages in efficiency, reliability and robustness for slow fading channels where certain delay constraints disallow coding over a large number of blocks. On fast fading channels, we remark that fixed-rate schemes and rateless schemes may in principle achieve the same level of efficiency and reliability. However, rateless coding still has an advantage in robustness, since rate is not a design parameter and need not to be set *a priori* according to the channel statistics.

From a complexity perspective, rateless codes are of the same decoding complexity as most LDPC codes. However, given that decoding may be attempted many times before successful decoding, the total number of computations, assuming independence between decoding attempts, can be orders of magnitude greater than a fixed-rate code. Work has been done to address this issues, notably [35].

The framework presented above may be generalized to multiple-antenna settings. It is well known that under the fast fading model, a fixed-rate system can achieve the ergodic capacity of the channel with zero probability of outage. However, fixed-rate codes are not robust in this setting since either efficiency or reliability must suffer when the transmitter does not know the channel statistics. An important distinction in that case is that communication efficiency and reliability should be defined in terms of mutual information rather than capacity. In the multiple-antenna setting, rateless codes are then expected to have the same advantages over fixed-rate codes as they do in the single-antenna case.

It is worth noting that the notion of rateless coding is conceptually similar to Hybrid ARQ (see [46] and the references therein). The difference is mainly in code construction. In a sense, one may view rateless codes as a "continuous" version of Hybrid ARQ schemes. It appears that the reviving interest in Hybrid ARQ and the extension to rateless coding are leading to new techniques in various areas of wireless communications (see, for example, [6, 89, 8]). We explore this connection more closely in Chapter 4.

2.A Proof of Theorems 2 and 3

Proof: We follow a very similar development as the proof in [6, Lemma 1]. As explicitly noted by the authors of [6], their setup is directly applicable to a block fading channel with Gaussian input distribution. As such, we start with our system model given by (2.1), with the following generalization: We allow the block length to vary randomly from block-to-block. This is done as a mathematical convenience to combine the proofs of Theorem 2 and Theorem 3. Let s denote the block index, $s \in \{1, 2, \ldots\}$, and let γ_s be a positive value on $(0, 1]$ representing a fraction of the total block size T_c such that for this "block" fading setup, the channel remains constant for $\gamma_s T_c$ channel uses. In this context, one may think of a channel realization as the pair (H_s, γ_s), where the elements of this pair are drawn independently from different distributions. The value γ_s is drawn randomly i.i.d. for every s and known at the receiver only. The channel model can then be represented by

$$\mathbf{Y}_s = H_s \mathbf{X}_s + \mathbf{N}_s,$$

where s denotes the block number index. To recover the block fading model in (2.1), we set the distribution for γ_s to be a delta function at 1. To recover quasi-static fading, draw H at time zero and hold it constant for the duration of the codeword transmission. The

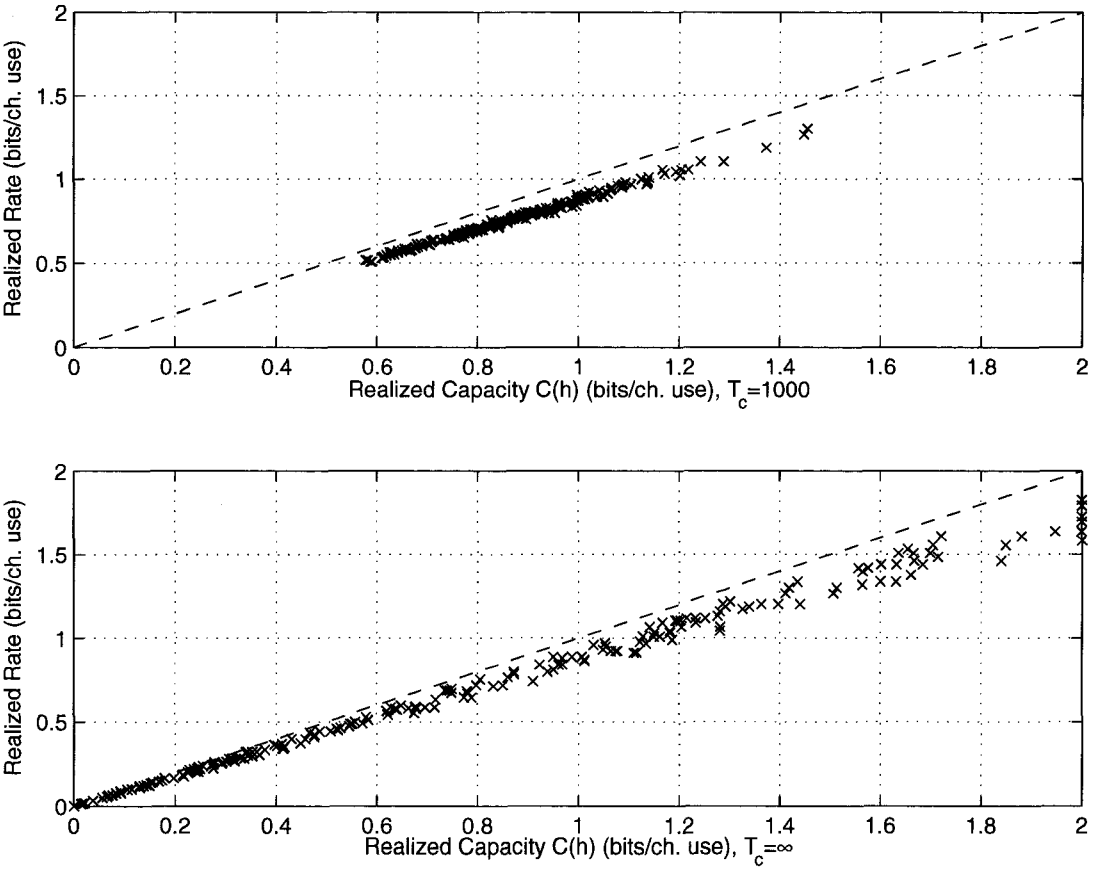

Figure 2.7: Realized rate vs. realized capacity, for average receive SNR=0dB: Top, T_c=1000, Bottom, quasi-static

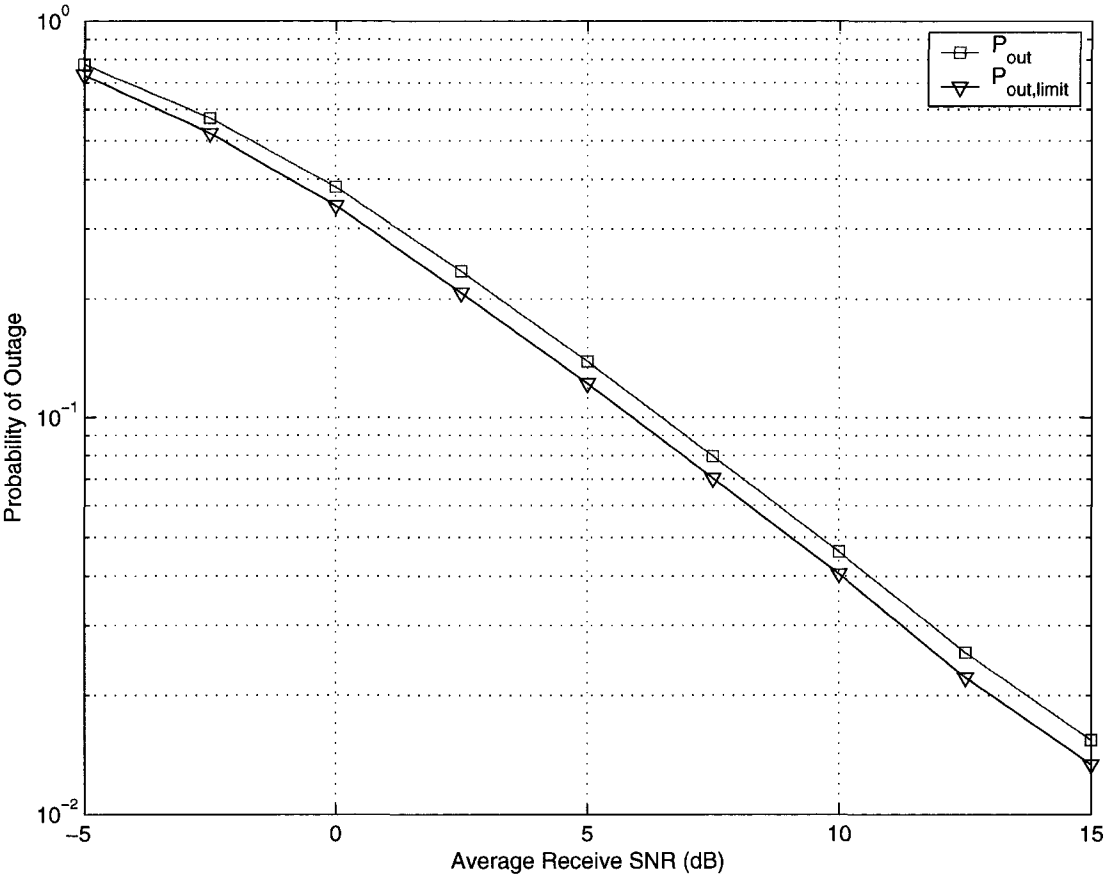

Figure 2.8: Probability of outage vs. average receive SNR, for $R_T = 0.5$

purpose of γ_s in quasi-static fading context is to prove that rateless codes are good at any rate, not just those corresponding to block boundaries, subject to standard constraints.

Using continuity arguments [26], we deal with the discrete version of this channel, with single-letter transition pmf of the channel represented by $p_s(Y|X)$. Fix a sequence of block length fractions to be

$$\mathcal{G} := \{\gamma_s : \gamma_s > 0\}.$$

In the same manner, fix a sequence of channel realizations to be

$$\mathcal{P} := \{p_s(Y|X)\}.$$

Let

$$P(\mathbf{X}_s, \mathbf{Y}_s), P(\mathbf{X}_s) \text{and} P(\mathbf{Y}_s)$$

be the joint and marginal probability distributions induced by the input distribution $q(\mathbf{X})$. We then have the following modified expressions for the weak law of large numbers:

$$\lim_{T_c\to\infty} \frac{1}{T_c} \log_2 P(\{\mathbf{X}_s, \mathbf{Y}_s\}) \rightarrow -\sum_s \gamma_s H_s(X,Y) \tag{2.17}$$

$$\lim_{T_c\to\infty} \frac{1}{T_c} \log_2 P(\{\mathbf{X}_s\}) \rightarrow -\sum_s \gamma_s H(X) \tag{2.18}$$

$$\lim_{T_c\to\infty} \frac{1}{T_c} \log_2 P(\{\mathbf{Y}_s\}) \rightarrow -\sum_s \gamma_s H_s(Y), \tag{2.19}$$

where convergence is in probability, and where

$$H_s(X,Y) := -\sum_{x,y} q(x)p_s(y|x) \log_2 q(x)p_s(y|x) \tag{2.20}$$

$$H(X) := -\sum_x q(x) \log_2 q(x) \tag{2.21}$$

$$H_s(Y) := -\sum_{x,y} q(x)p_s(y|x) \log_2 \sum_{x'} q(x')p_s(y|x') \tag{2.22}$$

represent the joint, input and output entropies in channel block s.

Using the standard definition of typicality – as in [6] – we follow the usual steps to prove the achievability portion of the channel coding theorem (see Cover and Thomas [17, Theorem 8.7.1]). Let

$$I(q(x), p_s(y|x)) := H(X) + H_s(Y) - H_s(X,Y).$$

It then follows that any rate less than

$$\sum_{s=1}^{m} \frac{\gamma_s}{m} I(q(x), p_s(y|x))$$

is ϵ-achievable. In particular, letting $m = M$ and $\hat{M} := \sum_s \gamma_s$, then for a given sequence of channels $\mathcal{P}$ and sequence $\mathcal{G}$, and for sufficiently large T_c there exists codes $\mathcal{C}_k$ of length $T_c\hat{M}$ and rate $k/T_c\hat{M}$ with error probability less than ϵ if

$$R := k/T_c < \sum_{s=1}^{M} \gamma_s I(q(x), p_s(y|x)). \tag{2.23}$$

It remains to be shown that a single code exists satisfying (2.23) which is simultaneously good for all $\mathcal{P}$ and $\mathcal{G}$, where "good" denotes the property of having a maximum probability of error less than ϵ. Furthermore, it must be shown that such for a code, for all $1 \leq m \leq M$, if

$$R < \sum_{s=1}^{m} \gamma_s I(q(x), p_s(y|x)),$$

then a punctured version of the code using only the first m blocks must also be good.

Let a codebook be denoted by $\mathcal{C}$, and a decoding error for $\mathcal{C}$ be given by $\mathscr{E}$. The strong converse given in [6, Lemma 2] shows that the following limit holds in probability:

$$E_{\mathcal{C}}[\Pr(\mathscr{E}|\mathcal{P}, \mathcal{G}, \mathcal{C})] \to I[R \geq \sum_{s}^{m} \gamma_s I(q(x), p_s(y|x))] \tag{2.24}$$

as $T_c \to \infty$, where $I[\cdot]$ is the indicator function and $E_{\mathcal{C}}$ denotes expectations taken over all codebooks of size k and length $T_c\hat{M}$. Although $\hat{M}$ is a random variable not known by the transmitter, this has no impact on the existence of a good codebook. The receiver, knowing $\mathcal{G}$, decodes at block boundaries, regardless of where they occur.

We take expectations of (2.24) over all possible channel sequences $\mathcal{P}$ and $\mathcal{G}$ simultaneously. For the same reasons stated in [6] – that the integrand is non-negative and bounded by one – we swap order of expectations, giving

$$E_{\mathcal{C}}[E_{\mathcal{P}}[E_{\mathcal{G}}[\Pr(\mathscr{E}|\mathcal{P}, \mathcal{G}, \mathcal{C})]]] \to \Pr\left(\sum_{s \in \mathcal{S}_{k,M}} \gamma_s I(q(x), p_{k,s}(y|x)) \leq R \right). \tag{2.25}$$

Then there exists a family of codes $\mathcal{C}^*$ for increasing T_c such that

$$E_{\mathcal{P}}[E_{\mathcal{G}}[\Pr(\mathscr{E}|\mathcal{P}, \mathcal{G}, \mathcal{C}^*)]] \leq \Pr\left(\sum_{s \in \mathcal{S}_{k,M}} \gamma_s I(q(x), p_{k,s}(y|x)) \leq R \right). \tag{2.26}$$

It follows, again from [6], that there must exist $\Pr(\mathscr{E}|\mathcal{P},\mathcal{G},\mathcal{C}^*) \to 0$ for all $\mathcal{P}$ and $\mathcal{G}$ such that

$$R < \sum_{s\in\mathcal{S}_{k,m}} \gamma_s I(q(x), p_{k,s}(y|x)). \tag{2.27}$$

It also follows that any sub-sequence with $m < M$ satisfying this equation must also be decodable with error approaching zero. This is true even as $M \to \infty$. This can be seen by considering the fact that any sub-sequence can be decoded by decoding the mother code of block-size M, treating the $M - m$ unused blocks as erasures. Since (2.27) is satisfied, such a procedure results in the desired probability of error.

Let $M \to \infty$ so the rate of the code approaches 0. However, assuming that $I(q(x), p_s(y|x)) > 0$, then certainly there exists some $\tilde{M} < M$ such that

$$R/\tilde{M} < \frac{1}{\tilde{M}} \sum_{s\in\mathcal{S}_{k,m}} \gamma_s I(q(x), p_{k,s}(y|x)). \tag{2.28}$$

Virtually any rate – subject to the above constraints – is then achievable as $M \to \infty$. Therefore, subject to (2.27), it is always possible to decode at a rate arbitrarily close to the induced mutual information of the channel, $\frac{1}{\tilde{M}} \sum_s \gamma_s I(q(x), p_s(y|x))$.

To finish the proofs for both Theorem 2 and Theorem 3, we apply the above result to the channels considered. For Quasi-static fading, choose

$$\hat{\mathcal{P}} \subset \mathcal{P}, \hat{\mathcal{P}} = \{p_{k,s}(y|x) : s \in \mathcal{S}_{k,m}, p_{k,s}(y|x) = p_{k,t}(y|x), \forall s,t \in \mathcal{S}_{s,m}\}.$$

That is, the channel sequence is fixed for the duration of the codeword transmission. Since we have shown a code exists which is good for all channel sequences, and over all sequences of γ_s, then this code is also good for the quasi-static fading channel. Similarly, block fading is a special case by setting $\gamma_s = 1, s = \{1,2,\ldots\}$. Again, since a code exists which is good over all sequences of γ_s, it must be good for this channel as well.

□

Chapter 3

Modulation for Rateless Systems

The preceding chapter focused on the problem of good rateless codes for digital communication over fading channels. This chapter focuses on modulation, the other major component in communication systems. Figure 3.1 highlights the relationship of modulation and demodulation in the digital communication system model.

This chapter explores both modulation and demodulation when viewed in the rateless communication setting. It will be shown that traditional "fixed-rate" modulation methods limit the efficiency of the system when used over unknown channels such as those considered here. Adaptive modulation techniques have been the focus of a great deal of research and may be considered as a means of mitigating this limitation. However, the goals and results of such works are not easily adapted to the rateless system that is the focus of this work. Thus, to address such limitations, we introduce a new baseband modulation technique that provides some benefits to efficiency. This new approach, called μ-PAM modulation, is analyzed and we demonstrate some of the advantages it provides relative to pulse amplitude modulation (PAM).

The remainder of this chapter is organized as follows: Section 3.1 presents the motivation for this chapter, followed by Section 3.2 which outlines the system we propose and describes the proposed modulation scheme. An analysis of the proposed scheme is given in Section 3.3. A Monte-Carlo simulation is presented in Section 3.4. We conclude the chapter in Section 3.5.

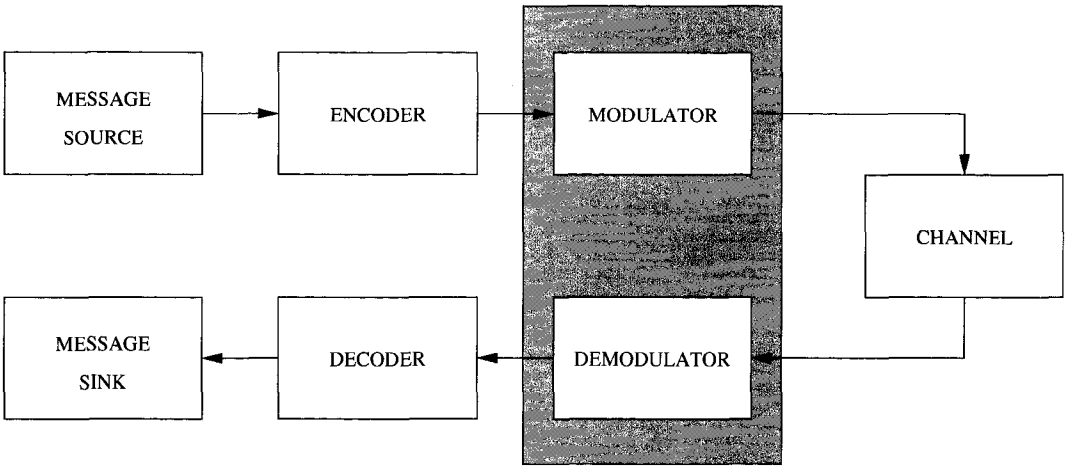

Figure 3.1: Digital communication system model highlighting modulation functions

3.1 Motivation

In many practical communications systems, it is difficult for the transmitter to know *a priori* the conditions, or state, of the channel it is communicating over. For instance, wireless channels often have a fading loss that is random, as well as a propagation loss that is not known by the transmitter without feedback from the receiver. As another example, to model Internet packet communication, a packet erasure channel with time-varying and unknown erasure probabilities are commonly used. In cases such as these, it is difficult, if not impossible for the transmitter to tune its transmission rate to suit the channel. Robust coding systems are required to operate without this knowledge.

When the unknown channel supports a possibly large range of rates, we show that coding and modulation must be considered simultaneously. Coding schemes which are efficient for low-rate, low-SNR operation are likely inefficient for high-rate, high-SNR operation. As a simple example, consider a channel that may be in one of two states, unknown to the transmitter. The channel may be in a "good" state, and support very high communication rates, or it may be in a "bad" state and support only a fraction of the rate supported by the good state. In such a scenario, modulation may have a large impact on the efficiency of the system.

Rateless coding schemes, which include ARQ-based and incremental redundancy methods, are a common means for providing reliability for such unknown channels. As demonstrated in Chapter 2, these codes offer robustness over wireless channels when only the receiver has channel state information, with the code rate naturally adapting to the

conditions of the channel.

Studies of the performance of Raptor codes for different channels [22, 59, 9] have mainly been limited to binary-input symmetric channels – whose capacities are limited to 1-bit per channel use – where the use of BPSK modulation is not detrimental to performance. For channels such as the AWGN channel at SNRs supporting higher rates of communication, binary coding and modulation may severely limit efficiency.

A novel solution to this problem is to use an adaptive demodulation scheme as described in [5], coupled with a Raptor code. Here, the modulation scheme is fixed, but the demodulation scheme is adaptive based on the channel state, thereby "compressing" the number of bits output from the demodulator. This approach automatically adapts its rate over a range of SNR such that the system targets a particular (high) SNR, and gracefully degrades when the SNR falls below this value.

We seek a coding and modulation system that offers efficient and reliable communication over unknown Gaussian channels with a somewhat "reverse" objective. Specifically, we would like a system that can perform well in very low SNR conditions, but adaptively increases its rate as the SNR increases. To that end, we consider a method of mapping k information bits to a constellation point in a signal space with dimension that increases with the number of channel uses.

Since Raptor codes are able to produce a virtually infinite number of output symbols, we propose a method in which a large number of consecutive output symbols are mapped into a single real value, and this value is transmitted over the unknown channel. This process is repeated until the decoder is able to recover the information bits. The mapping from output symbols to transmitted symbol may be considered a form of non-uniform modulation, providing unequal error protection to the output symbols.

Other approaches one may envision include using a schedule of decreasing modulation orders. For example, the transmitter and receiver agree upon the use of a high modulation order initially, which decreases as the number of channel uses increases to match an expected channel state.

We consider the former option, making use of a non-uniform modulation scheme along with rateless codes. We show that the performance of such a system is good over a very broad range of SNRs, suggesting their practical use in such environments.

3.2 System Model

We consider the standard discrete-time, real-valued AWGN channel, expressed as

$$y_i = x_i + n_i,$$

where x_i and y_i are the input and output symbols of the channel at time i respectively, and n_i is additive zero-mean, white Gaussian noise of variance σ^2. Only the receiver knows the value of σ^2.

The reason to consider the AWGN channel rather than the fading channels used in the rest of this work is to simplify the problem for the purpose of analyzing modulation performance. The fact that the AWGN channel is unknown to the transmitter makes this channel essentially equivalent to the quasi-static fading channel for our purposes.

The capacity of this channel C, in bits per channel use, is well known:

$$C = \frac{1}{2}\log_2(1 + \text{SNR}),$$

where SNR is $\frac{P}{\sigma^2}$, and P is the average power of the transmitted signal. In order to achieve the capacity of the channel, the input distribution must be Gaussian. Any other input distribution will incur a loss in capacity. For example, the use of a uniform input distribution incurs an asymptotic loss (as SNR increases) of $\pi e/6$ (1.53dB) relative to C. We neglect this aspect of the design problem in this chapter, though interested readers are directed to [24] for a treatment on the subject.

We are concerned with *unknown* AWGN channels in this chapter; i.e., we assume that the transmitter has no knowledge of the variance of the noise at the receiver. Consequently, there is a large range of possible operating SNRs and supported rates. A survey of the information theoretic aspects for communicating reliably over unknown channels is given in [45]. We assume in this work that a reliable, rate-limited feedback channel exists[1]. Our goal is a system for reliable communication that is efficient at low SNR and is able to automatically adapt to exploit higher rates efficiently.

3.2.1 Modulation Description

Consider the problem of modulating a set of M possible messages, where M is a power of two. Associate with the integer value $0 < i \leq M$ a bit-vector representation of

[1]We note that this is not required for the modulation method presented here. It is only required for the rateless code component.

the value $\mathbf{b}$. Canonical M-ary PAM modulation, or M-PAM for simplicity, maps a vector of bits $\mathbf{b}$ of length $L := \log_2 M$ to a modulation symbol from the finite set $A := (d_0/2)\{-M+1, -M+3, \ldots, M-1\}$, where d_0 represents the distance between adjacent symbols, and will dictate the average transmit power P of the system. Gray mapping from bits to symbols is typically used, such that the Hamming distance between the bit-vectors associated with adjacent symbols is equal to one.

The signal energy required per dimension is denoted by E_n. For some N-dimensional signal set constructed from points within hypercubes, the rate of communication possible in terms of bits per channel use per dimension is bounded by [82]

$$R_0^* := \frac{\log_2 e}{2}\left[1 + \frac{E_n}{N_0} - \sqrt{1 + \left(\frac{E_n}{N_0}\right)^2}\right] + \frac{1}{2}\log_2\left[\frac{1}{2}\left(1 + \sqrt{1 + \left(\frac{E_n}{N_0}\right)^2}\right)\right]. \quad (3.1)$$

As M-PAM modulation uses a 1-dimensional energy constrained signal set, (3.1) bounds the rate at which it may reliably communicate information. It is also notable that $\frac{1}{2}C \leq R_0^* < C$ [82]. The discrepancy between R_0^* and C is due to use of the probability of union bound in the computation of R_0^*. Nonetheless it serves as a meaningful limit on performance for our purposes.

Therefore we know that M-ary PAM, or any PAM-like modulation system is bounded by (3.1). Figure 3.2 presents the performance of M-PAM for different values of L for a uniform probability of selecting each constellation point.

We will assume a uniform probability of selecting constellation points, neglecting opportunities for shaping gain. This simplifies the analysis, though we are sacrificing potential improvements in achievable rates by neglecting shaping. The use of non-uniform probabilities have been investigated for many different applications. For coded applications, non-uniform QAM constellations were used in [25, 57] to provide shaping gain. For multicast problems, where different quality-of-service levels are to be provided, non-uniform modulation techniques have been explored and found to perform well, see e.g. [60, 54]. However these approaches are not obviously applicable to the problem considered in this chapter.

The proposed strategy for this communications problem is to map some number of consecutive rateless-coded output bits to a point in a constellation of non-uniformly spaced, real-values. The number of output bits used to map to a constellation point may, in general, depend on a number of factors including the number of previous channel uses, but for the purposes of this work we choose a constant number L. That is, every L consecutive bits output from a rateless code encoding k information bits is mapped onto

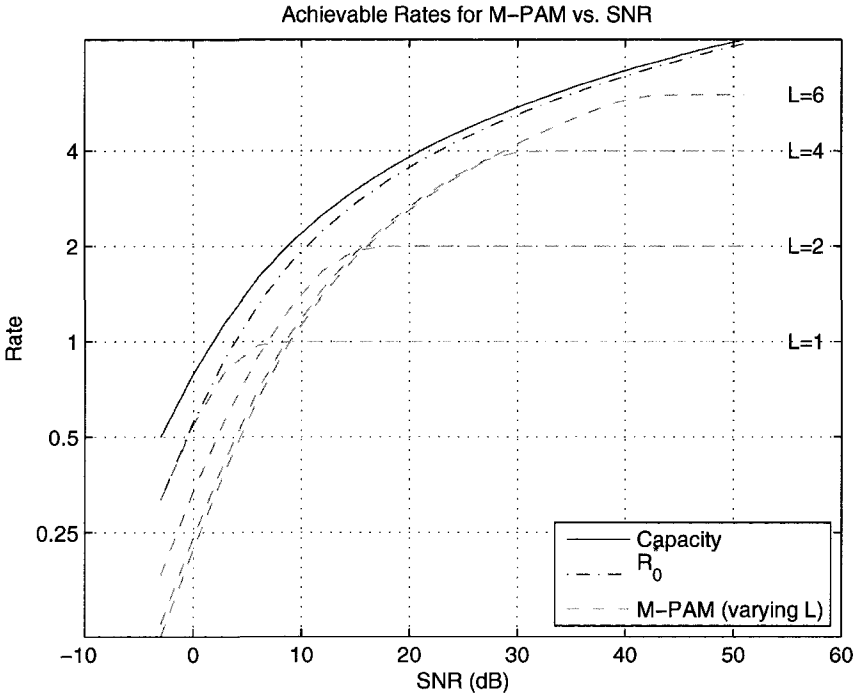

Figure 3.2: Achievable rates of M-PAM vs. SNR

a single constellation point and transmitted over the channel.

We generalize M-PAM modulation with a fractional-power modulation scheme called μ-PAM. This modulation scheme is similar to M-PAM, except that the symbol constellation no longer has an equal distance d_0 between adjacent symbols. Instead, the constellation points are generated in a recursive manner, dependent on μ, with $\mu \in (0, 1/2]$.

Figure 3.3 gives an example of the construction of a μ-PAM signal constellation for $\mu = 1/3$. The constellation begins as a standard BSPK signal set as seen in the top panel with $L = 1(M = 2)$. As with BPSK, one bit of information is used to discriminate between the two symbols in the constellation. In a recursive manner, additional bits are embedded into the current constellation set. Each operation of embedding one bit splits the constellation set by splitting each symbol and moving them by $\pm\mu^{L-1}$.

In the figure, the two BSPK symbols in the upper panel are split and moved by $\pm\mu^1$ to generate the middle panel ($M = 4$) which we say has an *embedding depth* of 2. Similarly, the four symbols in the middle panel ($M = 4$) are split and moved by $\pm\mu^2$ to generate the lower panel ($M = 8$) which has an embedding depth of 3. The embedding

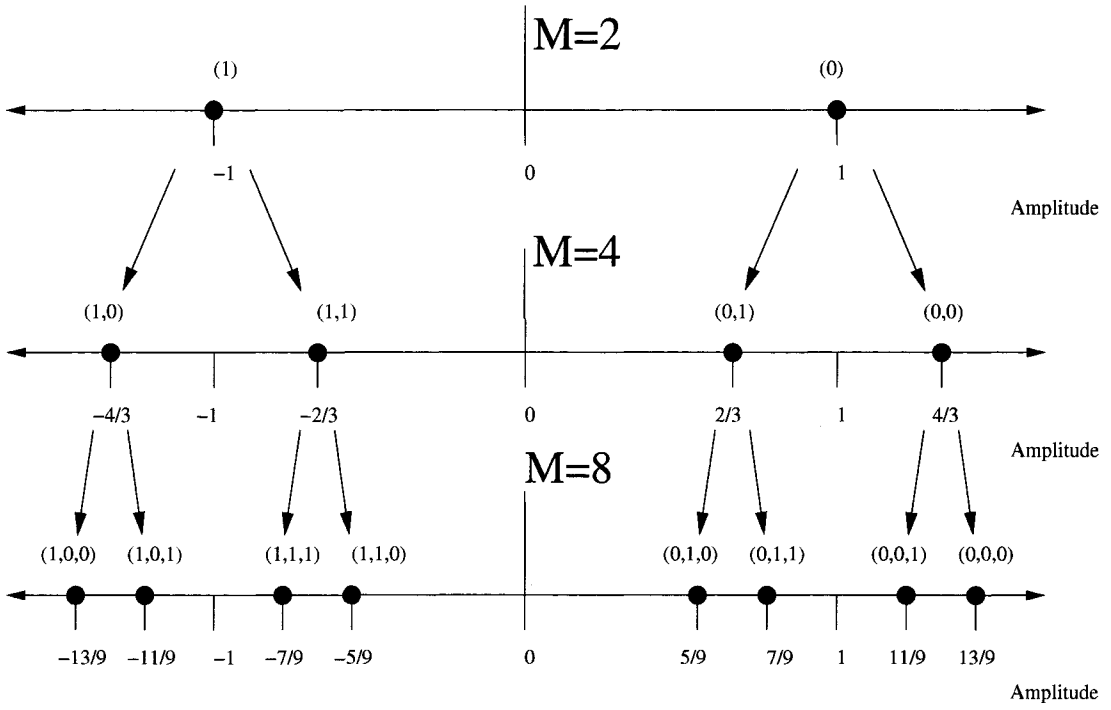

Figure 3.3: Example signal constellation construction of μ-PAM, $\mu = 1/3$. The top panel corresponds to $L = 1(M = 2)$. The panels below show the recursive construction of the constellation for each doubling of M.

operation is continued until some desired number of bits are embedded. The figure shows embedding only to $M = 8$, but there is no limit to the depth that may be used.

The resulting constellation structure has a self-similar nature that assists in the analysis of the modulation scheme. Comparing the positive (or negative) half-constellation for $M = 8$ in Figure 3.3 with the entire constellation for $M = 4$, we see that they are equivalent except for a scaling factor of μ and an offset. This is true for all values of M and, due to the recursive nature in which the constellation is constructed, it is straightforward to find similar relationships between any two embedding depths.

To gain some further insight into this modulation structure, Figure 3.4 shows two examples of μ-PAM modulation for $M = 8$, and $\mu = 1/2$ (upper) and $\mu = 1/3$ (lower). A few remarks are due based on these examples. First, choosing $\mu = 1/2$ recovers M-PAM up to scale. Second, the smaller the value of μ, the greater the non-uniformity, and the non-uniformity provides a form of unequal error protection. Note however that this form of non-uniformity is not intended to provide shaping gain, as is the case in, e.g., [25].

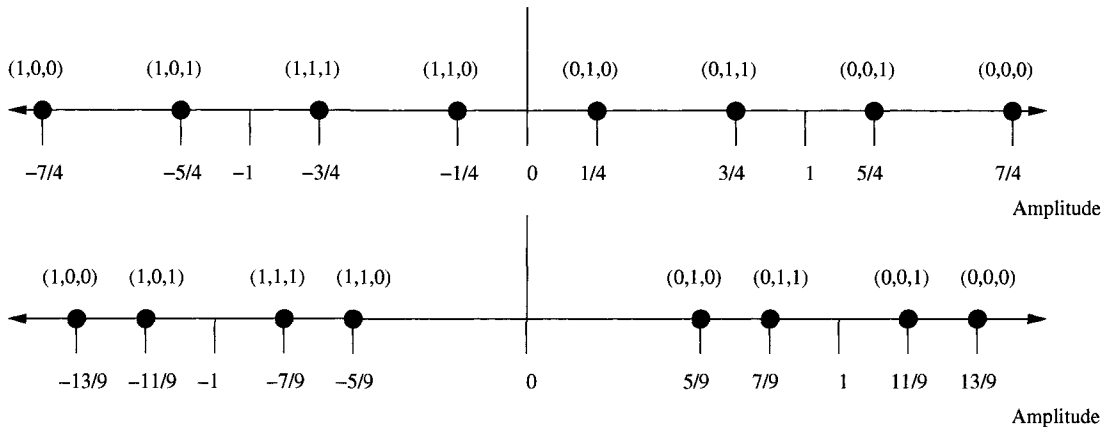

Figure 3.4: Signal constellations for μ-PAM, $M = 8$ with $\mu = 1/2$(upper) and $\mu = 1/3$(lower).

The mapping from a vector of bits $\mathbf{b}$ of length $L = \log_2 M$ to a μ-PAM symbol s can equivalently be defined as follows. Let $\hat{\mathbf{b}}$ be the vector of bits mapped from $\{0, 1\} \rightarrow \{+1, -1\}$. Then

$$s(\mathbf{b}) := \sum_{i=1}^{L} (\mu)^{i-1} \prod_{j=1}^{i} \hat{b}_j, \tag{3.2}$$

where $\hat{b}_j$ denotes the j-th element of the vector $\hat{\mathbf{b}}$. The above equation results in a Gray mapping between bit vectors and constellation symbols. As such, at high SNR the average (uncoded) bit error probability will approximately be equal to the average symbol error probability divided by L.

Gray mapping is not optimal for non-uniform constellations such as μ-PAM, and therefore we will expect some small performance degradation relative to optimal bounds. However, Gray mapping allows us to use a relatively simple demodulation technique which is described in Section 3.2.2.

We define the μ-PAM constellation – the set of all possible points given values μ and L – as

$$\mathbf{s}(\mu, L) := \{s(\mathbf{b}_1), \ldots, s(\mathbf{b}_M)\}$$

which will be used in the following sections.

3.2.2 Demodulation Description

For demodulation, it is straightforward to derive the *log-likelihood ratio* (LLR) expressions for each bit after reception of a modulation symbol. Many iterative decoders, such as those used by Raptor decoders, operate efficiently with LLRs. The demodulation of μ-PAM is identical to that of M-PAM. The only difference is the constellation signal set used. As a result, the complexity of LLR computation for M-PAM and μ-PAM are identical for identical values of M.

To compute the log-likelihood ratio, the Euclidean distance from the received symbol $\hat{s}$ to each of the constellation symbols in $\mathbf{s}(\mu, L)$ is first computed. Then the likelihood of each constellation point, given the received symbol and assuming a Gaussian noise model of known variance is determined. Finally, for each of the L bits comprising the received symbol, the likelihood ratio of the bit is a straightforward computation. Let $\lambda(i)$ be the LLR value for the i-th bit. Then the overall computation can be expressed as

$$\lambda(i) = \log\left(\frac{P\left[i=0|\hat{s}\right]}{P\left[i=1|\hat{s}\right]}\right).$$

Further details of this procedure may be found in, e.g., [5].

3.3 Analysis of μ-PAM

This modulation scheme essentially allocates power to bits unevenly in the mapping from bits to symbols. The first bit is given the largest portion of the power, subsequent bits are allocated fractionally smaller amounts, while the last bit is given the least. It is possible to allow the modulation order in μ-PAM to grow to infinity[2]. Thus it is possible to pack an essentially infinite number of bits onto a single symbol and, in the absence of noise, this is recoverable in an error-free manner. We provide the following theorem showing that the transmit power of such a constellation remains bounded.

Lemma 2 *The maximum instantaneous transmit power P^* of μ-PAM, as $M \to \infty$ is given by*

$$P^* = \frac{1}{1-\mu},$$

which is bounded for all $0 \leq \mu < 1$.

[2]We note the passing similarity here with Arithmetic source coding.

Proof: The proof of this theorem follows simply from convergence properties of series. The maximum transmit power of the constellation corresponds to the constellation point(s) with maximum amplitude. It is easy to see the all 0's bit-vector is such a point. From (3.2), the transmit power of this point is

$$P^* = \left(\sum_{i=1}^{L} (\mu)^{i-1} \right)^2 .$$

The expression within the brackets is a geometric series with a limiting value as $M \to \infty$ of $\frac{1}{1-\mu}$.

Since the maximum transmit power of μ-PAM is finite and power is a non-negative value, it is clear that the average transmit power is also finite and bounded. □

The performance of a modulation scheme is characterized by its distance properties, particularly the *minimum distance* of the scheme. The minimum distance of a modulation scheme is defined as the smallest Euclidean distance between any pair of points in the signal constellation set. Given the unequal protection that μ-PAM provides to the input bits, it is useful to determine an expression of the minimum distance as a function of the bit-position in the input vector $\mathbf{b}$.

Lemma 3 *The minimum distance d^1_{min} between points in the μ-PAM constellation corresponding to the first bit in $\mathbf{b}$ is given by*

$$d^1_{min}(L, \mu) := 2 \left(1 - \sum_{i=1}^{L-1} \mu^i \right) .$$

The proof of this theorem follows directly from the properties of geometric progression as used in the proof of Lemma 2.

We are then able to use μ-PAM with some assurance of the reliability of the system at low SNR relative to a BPSK system. Clearly, as M or μ increase, the worst-case degradation increases. However, the amount of information that may be communicated may in fact increase. This tradeoff is interesting to characterize, and is explored further in the next section which presents details of a test of the proposed system for finite M.

It is possible to generalize the minimum distance bound for Lemma 3 for any bit in the vector. Let $d^i_{\min}$ be the minimum distance between all adjacent symbols for which the i-th bit in $\mathbf{b}$ differs.

Theorem 4 *The minimum distance d^i_{min} between points in the μ-PAM constellation corresponding to the i-th bit in* $\mathbf{b}$ *is given by*

$$d^i_{min}(L,\mu) := 2\left(\mu^{i-1} - \sum_{j=i}^{L-1}\mu^j\right). \tag{3.3}$$

Proof: We prove by induction over i. The theorem is true for $i = 1$ due to Lemma 3. Suppose that it is true for $i = n$. By the recursive generation of the constellation set, either half – the positive or negative symbols – of the constellation can be viewed as an entire constellation set with an appropriate modification of parameters. In particular, the subset is shifted by some amount ± 1, scaled in amplitude by amount μ and has depth $L-1$. We can then express the minimum distance for $i = n+1$ to be

$$\begin{aligned}
d^{n+1}_{\min}(L,\mu) &= 2\mu^L + \mu d^n_{\min}(L,\mu) \\
&= 2\mu^L + 2\mu\left(\mu^{n-1} - \sum_{j=n}^{L-1}\mu^j\right) \\
&= 2\mu^L + 2\mu\left(\mu^{n-1} - \left[\mu^n + \ldots + \mu^{L-1}\right]\right) \\
&= 2\mu^L + 2\left(\mu^n - \left[\mu^{n+1} + \ldots + \mu^L\right]\right) \\
&= 2\left(\mu^n - \left[\mu^{n+1} + \ldots + \mu^{L-1}\right]\right) \\
&= 2\left(\mu^n - \sum_{j=n+1}^{L-1}\mu^j\right) \\
&:= d^{n+1}_{\min}(L,\mu).
\end{aligned}$$

□

It is straightforward to express these distances in the limit as $L \to \infty$. This can be used to bound the performance of the modulation system asymptotically.

Corollary 1 *The minimum distance d^i_{min} between points in the μ-PAM constellation corresponding to the i-th bit in* $\mathbf{b}$ *as $L \to \infty$ is given by*

$$d^i_{min}(\infty,\mu) := \frac{2\mu^{i-1}(1-2\mu)}{1-\mu}.$$

Proof: The proof follows by simply by allowing $L \to \infty$ in (3.3) and using the geometric sum,

$$\begin{aligned} d^i_{\min}(\infty, \mu) &= \lim_{L\to\infty} 2\left(\mu^{i-1} - \sum_{j=i}^{L-1} \mu^j\right) \\ &= 2\left(\mu^{i-1} - \frac{\mu^i}{1-\mu}\right) \\ &= \frac{2\mu^{i-1}(1-2\mu)}{1-\mu}. \end{aligned}$$

□

If we let $\mu = 1/2$, then in the limit, the minimum distance approaches zero for all values of i. For any smaller value of μ, the minimum distance is bounded above zero, and the fractal nature of the modulation scheme is seen again; all distances at depth i are a factor of μ smaller than at the previous depth.

3.3.1 Simplified Demodulation Method

Based on the above results, it is possible to find bounds on the probability of error and achievable rates of communication that are possible with this modulation technique. Error events for this modulation scheme will be dominated by errors between the closest pairs of symbols. For values of $\mu < 1/2$, this will be for the pairs of symbols that differ only in the last bit. An upper bound expression for this probability of error is

$$P_s(\mathscr{E}) \leq (M-1)Q\left(\sqrt{\frac{\left(d^L_{\min}\right)^2}{2N_0}}\right), \tag{3.4}$$

which is simply the worst-case probability of error for an M-PAM modulation scheme. Since for $\mu < 1/2$ every symbol has exactly one neighbor of minimum distance, this is a loose bound. An exact analysis of the probability of error is beyond the scope of this work, but interested readers are directed to [24] for a general treatment of the subject. However, it is possible to derive some bounds on the error probabilities that are tighter than what is given in (3.4).

To do so, we model the communication channel as a system of parallel, though dependent Gaussian channels. That is, given a μ-PAM modulation scheme operating with some chosen L, decompose the transmission of a single symbol as transmission of individual bits over L parallel virtual AWGN channels, depicted in Figure 3.5. We view the

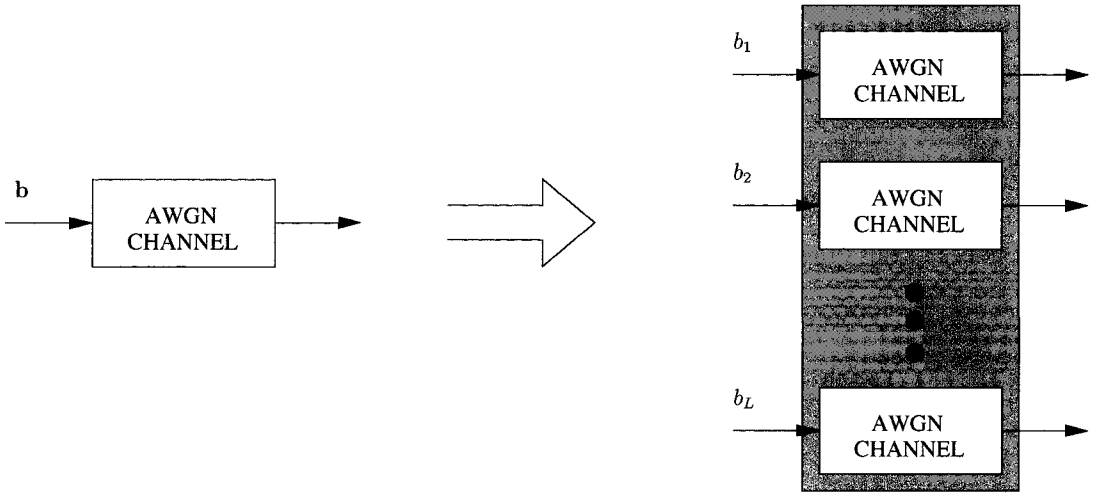

Figure 3.5: Decomposition of AWGN channel into L virtual channels

i-th bit as traveling over the i-th virtual channel. This is an *approximate* decomposition, since the parallel channels are most certainly not independent. Additionally, it is likely that each virtual channel is not truly Gaussian. However, Monte-Carlo tests were run showing empirical distributions for the virtual channels which are extremely close to Gaussian. In fact, no visible difference was noticeable beyond what would be expected due to the random nature of the Monte-Carlo test.

A consequence of Corollary 1 is that, for sufficiently large L, the minimum distance of the i-th bit is a factor of μ larger than the minimum distance of the $(i+1)$-th bit. From the perspective of the virtual channel, it appears that each subsequent bit passes through a channel of SNR which is a factor of μ smaller than the previous bit. There is of course dependence between these channels, which makes it difficult to analyze using LLR decoding described in Section 3.2.2, particularly for large values of L.

Instead, we use the following simplified decoding algorithm, in which estimation is done on a per-bit basis, in a recursive manner which mirrors the construction of the μ-PAM constellation set. Define a decision boundary D, and initialize it to 0. Let the received value from the channel be y. Let $\tilde{b}_i$ represent the LLR estimate of the i-th bit, conditioned on receiving y. Then we can determine the values of $\tilde{b}_i$ given y using the following steps: For each bit starting with the 1-st bit, estimate the transmitted value assuming that the transmitted modulator was BPSK with an offset of D and an amplitude (relative to D) of μ^{i-1}. Then hard decisions or LLR computations for the i-th bit are quite simple and involve only two possible transmitted symbols. After each

Input:		
y	:=	output of channel
Initialization:		
D	=	0
i	=	1
For i from 1 to L:		
$\{s_1, s_2\}$	=	$\{D - \mu^{i-1}, D + \mu^{i-1}\}$
$\hat{b}_i$	=	$\log\left(\frac{\lvert s_1 - y\rvert}{\lvert s_2 - y\rvert}\right)$
D	=	$D + \mu^{i-1}\text{sign}(\hat{b}_i)$
End		

Table 3.1: Simplified decoding procedure; Pseudo-code

estimation is made, the value of D is incremented by μ^{i-1} in the direction of the estimated bit. Pseudo-code for the decoding procedure is given in Table 3.1.

It is notable at this time to make a few remarks about the computational complexity of this decoding method. Standard Maximum Likelihood (ML) LLR computations as those described in Section 3.2.2 have a computational complexity that grows in proportion to 2^L. Every additional bit doubles the number of symbols for which distance computations are required. In contrast, this simplified method has a computational complexity that is linear with L. Each additional bit requires only one additional LLR computation involving a pair of symbols, as well as a few addition and multiplication computations. For very large L, the difference in overall complexity of the decoding schemes can be quite dramatic.

As it is to be expected, this simplified decoding method is suboptimal relative to ML LLR computations. In addition, this simplified method, though easy to implement, is nevertheless difficult to characterize analytically. The dependencies between the virtual channels used to model this scheme makes an accurate bounding of probability of error non-trivial. Bounds which we found based on minimum distance properties are in fact quite loose and therefore not informative in any way.

As an alternative, a numerical approach is used to determine performance of μ-PAM with the simplified demodulation method. The general approach is to make use of the virtual channel decomposition model to treat the transmission of each bit in the μ-PAM modulated symbol as passing over an AWGN channel of known SNR. Taking into account

the dependencies between the bits, we can treat the estimation of successive bits to have a Markov dependency structure.

Our goal will be to determine the mutual information between the input symbol $\mathbf{b}$ and the output symbol $\tilde{\mathbf{b}}_\mathbf{i}$. This mutual information represents the upper bound on the rate at which we can communicate information reliably with this modulation scheme and simplified decoding method.

Formally, denote by S^i the state of the demodulator before demodulation of the i-th bit of the received symbol y using the simplified demodulation method. Before the 1-st bit is demodulated, there is only one possible state, $S^1 = 0$. Before the 2-nd bit is demodulated, $S^2 \in \{0, 1\}$, denoting whether the previous bit was estimated to be a 0 or a 1 respectively. In general, the cardinality of the state S^i is 2^{i-1}, representing all possible bit-sequences of length $i - 1$. The demodulation of each successive bit induces a state transition from $S^i \rightarrow S^{i+1}$. The transition is probabilistic based on the noise realization on the AWGN channel, but the state transition probability matrix $T^{i \rightarrow i+1}$ is relatively easy to compute numerically, conditioned on an input symbol x.

Let x be a fixed modulated symbol which is the μ-PAM mapping from the bit vector $\mathbf{b}$ of length L. After passing over the AWGN channel, the symbol y is received by the simplified demodulator. Let the sequence of states in the Markov chain modeling the demodulator be $\{S^1, \ldots, S^L\}$. We model the 1-st bit to pass over the virtual channel 1 of some SNR^1, a function of the (approximately) AWGN noise and $d^1_{\min}$. Based on standard analysis of BPSK channels, namely that the probability of error is

$$P_{\mathrm{BPSK}}(\mathscr{E}) = Q\left(\sqrt{2\mathrm{SNR}^1}\right),$$

we can determine $T^{1 \rightarrow 2}$. In this case, assuming that that the first bit of x is $b_1 = 0$, then

$$\begin{aligned} T^{1\rightarrow 2}(S^1 = 0, S^2 = 1 | b_1 = 0) &= Q\left(\sqrt{2\mathrm{SNR}^1}\right) \\ T^{1\rightarrow 2}(S^1 = 0, S^2 = 0 | b_1 = 0) &= 1 - Q\left(\sqrt{2\mathrm{SNR}^1}\right). \end{aligned}$$

This fully characterizes the probability of error for the 1-st bit in the received symbol given some x. We note again that these equations are approximations due to the assumptions of independence between the virtual channels.

We then proceed, following the same procedure at each step for $i = 1, \ldots, L$. At each step, the state-space increases by a factor of two and the number of non-zero elements of the state transition probability matrix also increases by a factor of two. Upon determining

all of the state transition probability matrices from 1 to L, they can be combined to form a conditional channel probability matrix $p(\tilde{\mathbf{b}}|\mathbf{b})$. This is done by determining, for each $i \in \{1, \ldots, L\}$, the overall probability of each of the two possible values (0,1), that $\tilde{b}_i$ may take.

To do this, we marginalize over all state sequences which contain the desired value of $\tilde{b}_i$. First, define a *path* as a particular sequence of states, $\{S^1, \ldots, S^L\}$. For example, for $L = 3$, one path may be $\{0, 0, 2\}$. Let $\mathscr{P}_i$ be the set of all paths which at the i-th element have a state which is consistent with $\tilde{b}_i$. Let some path in $\mathscr{P}_i$ be denoted by $\mathcal{P}$. Identify with a particular element of $q \in \mathcal{P}$ the following values: Let $l(q)$ be the initial depth, $l(q) \in \{1, \ldots, L\}$. Let $S^I(q), S^F(q)$ be the initial and final states associated with element q in path $\mathcal{P}$, respectively.

With these definitions, the conditional probability matrix of the channel is given by

$$p(\tilde{\mathbf{b}}_\mathbf{i}|\mathbf{b}) = \sum_{\mathcal{P} \in \mathscr{P}} \prod_{q \in \mathcal{P}} T^{l(q) \to l(q)+1}(S^I(q), S^F(q)|b_{l(q)}).$$

We note that $p(\tilde{\mathbf{b}}|\mathbf{b})$ is an L-dimensional matrix of dimension $2 \times 2 \times \ldots \times 2$.

Computation of the conditional probability matrix of the channel for all possible values of $\mathbf{b}$, along with a distribution $p(\mathbf{b})$ over $\mathbf{b}$ is sufficient to define the joint probability matrix of the channel given by

$$p(\tilde{\mathbf{b}}, \mathbf{b}) = p(\tilde{\mathbf{b}}|\mathbf{b})p(\mathbf{b}). \tag{3.5}$$

This joint probability of the channel allows us to compute the mutual information of the simplified modulation technique.

Recall the definition of mutual information [17]

$$I(X;Y) = \sum_{x,y} \log p(x,y) \frac{p(x,y)}{p(x)p(y)} \tag{3.6}$$

As is typical for input distributions for binary input modulation schemes, we assume a uniform distribution for $p(\mathbf{b})$. It is a simple matter to marginalize the joint probability matrix of the channel, and so we have everything required to determine the mutual information $I(\tilde{\mathbf{b}}; \mathbf{b})$. The following section presents the results of the numerical computations.

3.3.2 Numerical Results

Equation (3.6) can be used with (3.5) to numerically determine the rates that are achieved by μ-PAM with the simplified demodulation method described in Section 3.3.1. For a

number of different configurations, this section presents some results and compares with theoretical limits for PAM modulation systems.

Figure 3.6 presents the achievable rates of μ-PAM with the simplified demodulation scheme as a function of SNR for $L = 2, 4, 6$, and $\mu = 1/3$. (A detail of the low SNR region is given in Figure 3.7). Also presented are some M-PAM achievable rates based on Figure 3.2. The uppermost curve is R_0^* from (3.1). Comparing M-PAM and μ-PAM for the same value of L, we can see common features. First, performance of M-PAM is superior to μ-PAM at high SNR, but is inferior at low SNR. This observation appeals to intuition in the following way; we designed μ-PAM to be robust over AWGN channels with unknown SNR, and as such expect to see a certain robustness at low SNR due to the relatively large $d^1_{\min}$ compared to M-PAM. The price we pay is performance at higher SNR where the relatively small minimum distances of the encoded bits close to the value of L. At high SNR then, μ-PAM cannot support as high a rate as M-PAM.

For each value of L we can identify a crossover SNR above which M-PAM provides superior results. The following table lists some of these crossover values between μ-PAM and M-PAM for $\mu = 1/3$.

L	Crossover SNR
2	6.4 dB
4	15.7 dB
6	17.9 dB

Note that the crossover SNR increases with increasing SNR. This is consistent across all tested values of μ and L.

Comparing μ-PAM for different values of L, we see that there is a tradeoff in performance for different values of L. This is demonstrated in Figure 3.7 where the rates achieved for μ-PAM are presented for $\mu = 1/3$ and $L = 2, 4, 6$. At low SNR below 1.1dB, the best choice is $L = 2$. Above that $L = 4$ is provides the best performance until an SNR of 6.3dB above which $L = 6$ gives the best performance. In all cases, however, it should be noted that the difference in performance is relatively small. This highlights the fact that there is a tradeoff to be made in choice of L in addition to choice of μ.

A common design problem is choosing the best modulation scheme for a given operating point or SNR. For an AWGN channel with known SNR, it is simply a matter of choosing the modulation scheme that achieves the highest rate at that particular SNR. An equivalent criterion is choosing the scheme which achieves the highest *fraction* of channel capacity. This form of comparison is often easier to visualize. The fraction of

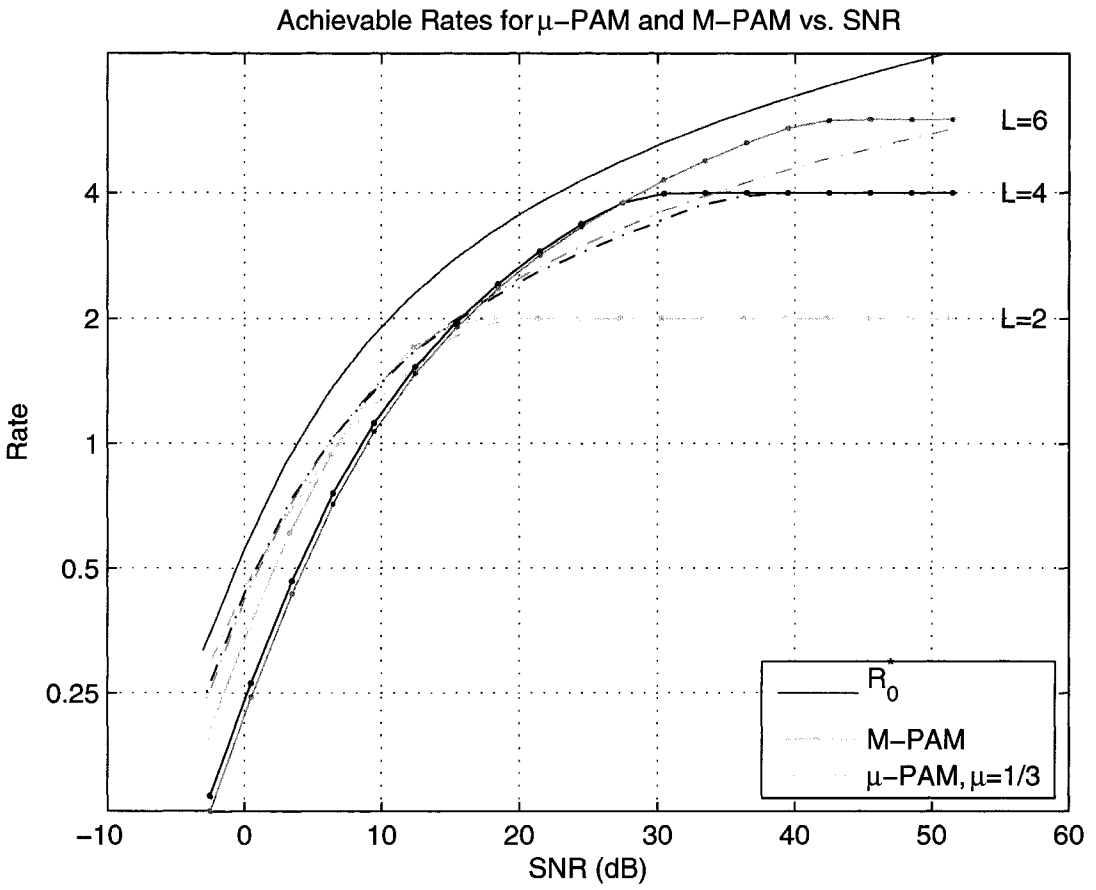

Figure 3.6: Achievable rates for μ-PAM vs. SNR

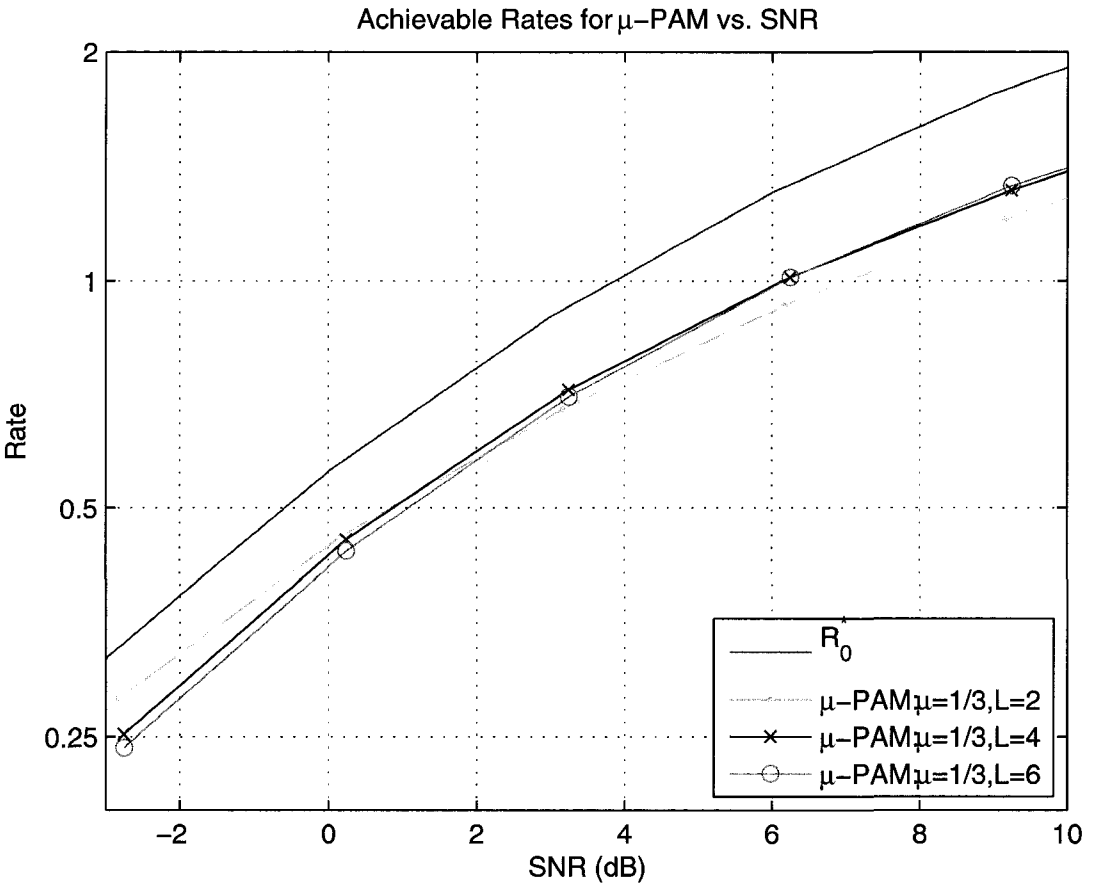

Figure 3.7: Comparison of μ-PAM rates vs. SNR

capacity is computed as $\frac{R}{C}$, the ratio of the achieved rate to the capacity of the AWGN channel at a particular SNR.

Figures 3.8 and 3.9 present the fraction of capacity achieved as a function of SNR for some values of L and μ. Figure 3.8 compares M-PAM and μ-PAM for $\mu = 1/3$ and a number of values of L. It can be seen that for most SNRs a suitable choice of M-PAM performs better than smaller values of μ. However of the range of SNR from about 5 to 15 dB, a μ-PAM scheme with $\mu = 1/3$ outperforms any M-PAM scheme.

It is interesting to note the peaks when using M-PAM. For a given choice of M, there is clearly a small range of SNR for which performance can be said to be the best in terms of fraction of capacity achieved. This is not the case with $\mu = 1/3$. Although performance is not as good as M-PAM in these 'peak' SNR ranges, there is a much broader range of SNRs over which a reasonably large fraction of capacity is achieved. This supports the idea that μ-PAM may be better suited for channels with unknown SNR.

A similar observation may be made from the results in Figure 3.9 which compares M-PAM and μ-PAM for $\mu = 1/3$ and $\mu = 2/5$. Here we see that the inclusion of $\mu = 2/5$ as a modulation option greatly increases the range over which μ-PAM is the best choice from the design problem perspective, though the value of μ is a key design parameter.

The above design problem assumes a *known* SNR over the channel. However as we are ultimately interested in unknown channels such as those that may be encountered in wireless environments we consider the case in which the SNR of the AWGN channel is unknown. In this case, the design criterion must be modified to account for the uncertainty of the channel SNR.

We model the channel SNR as a random variable taken from some distribution. In this case, a natural metric which can be used to compare performance of different schemes is the mean achieved rate. The mean achieved rate, $\bar{R}$, is merely the expectation of the achieved rate, R, taken over the channel SNR distribution,

$$\bar{R} := E[R]. \tag{3.7}$$

Maximizing this metric is equivalent to optimizing the overall throughput of the system from the perspective of the transmitter.

Figure 3.10 presents a comparison of the mean achieved rate as a function of L for M-PAM and μ-PAM assuming that the channel SNR is drawn from an exponential distribution with parameter β. In the figure, three values of β are shown; $\beta = 35$dB, $\beta = 20$dB and $\beta = 6$dB, representing high, medium and low average SNR conditions

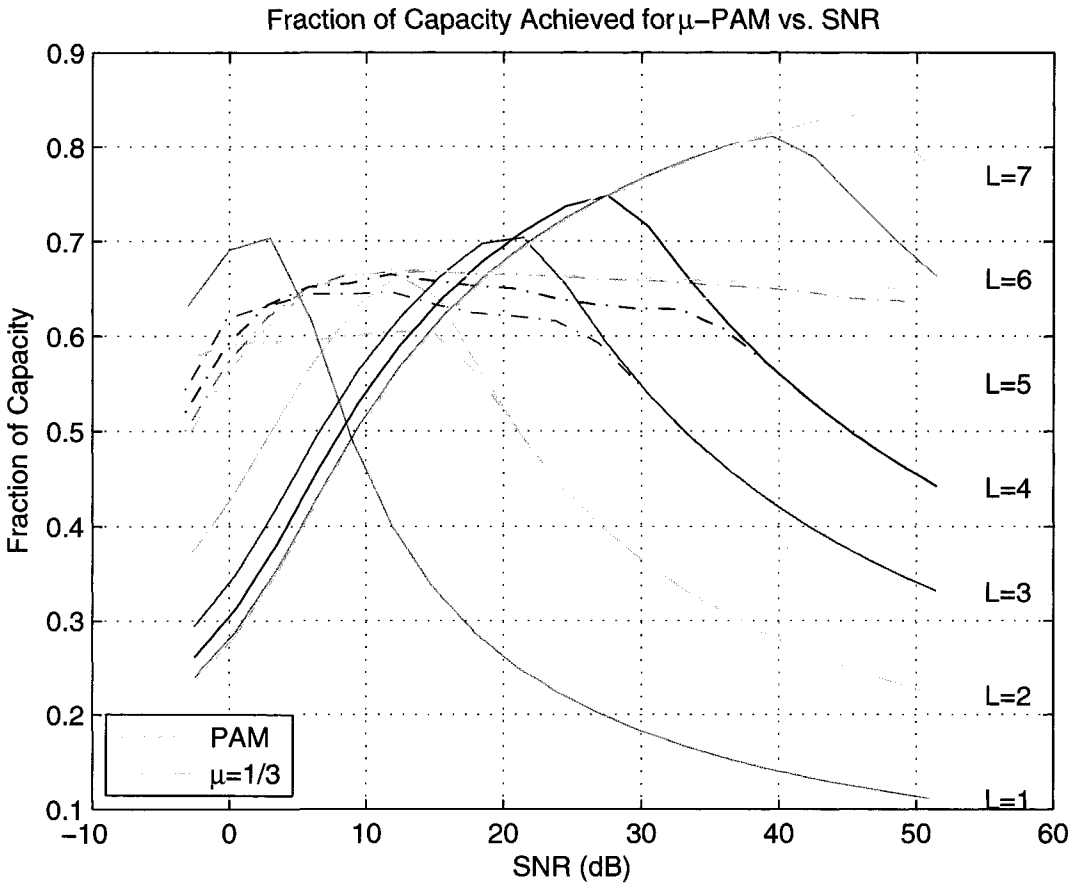

Figure 3.8: Fraction of capacity achieved for M-PAM and μ-PAM vs. SNR

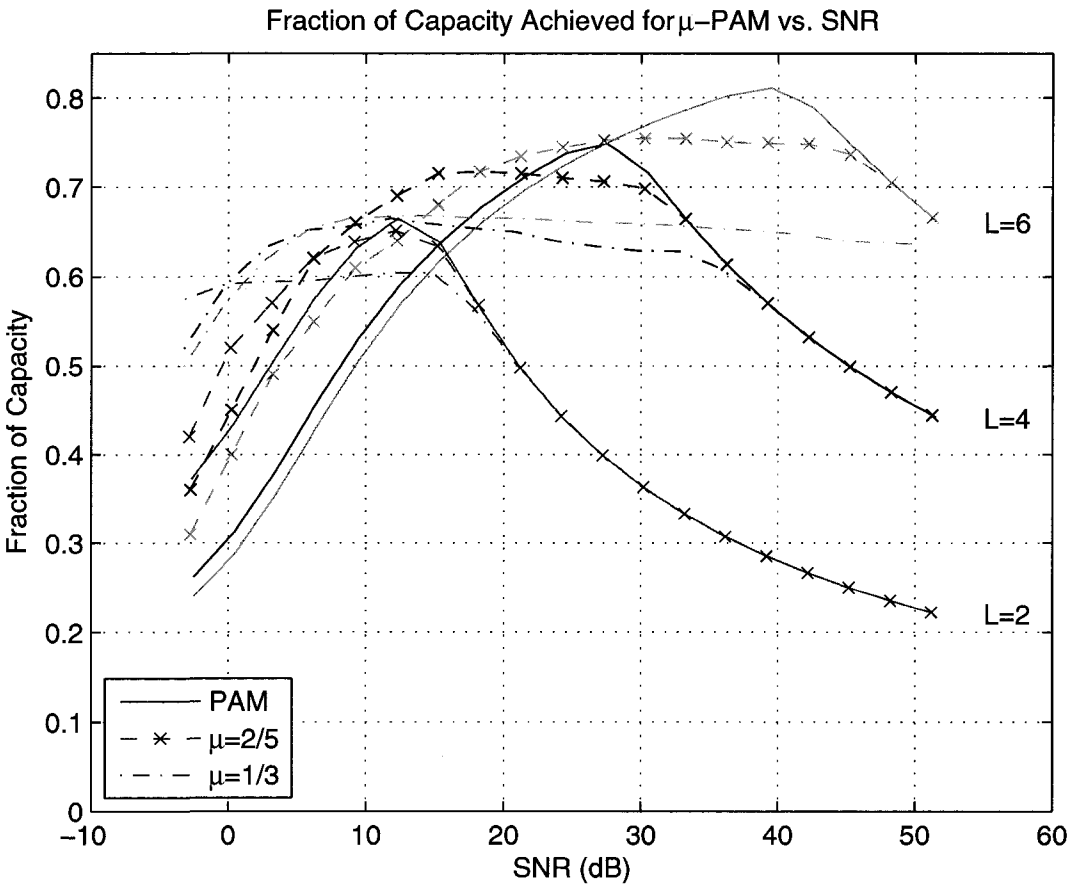

Figure 3.9: Comparison of the fraction of capacity achieved for μ-PAM vs. SNR

respectively.

For $\beta = 35$dB, M-PAM is the best choice for all values of L. This is due to the performance penalty paid by μ-PAM for robustness at low SNR. In this SNR scenario, the performance of μ-PAM suffers increasingly as μ decreases from 1/2. For $\beta = 20$dB, we begin to see $\mu = 2/5$ take over as the best choice for most values of L. At the low SNR condition of $\beta = 6dB$, the best choice depends on L and is either $\mu = 1/3$ for large L or $\mu = 2/5$ for smaller values of L.

These results are consistent with the previous results; when the SNR is relatively small, μ-PAM offers performance advantages compared to M-PAM. However, relatively high SNRs are best suited for use with M-PAM.

This relationship is further investigated in Figure 3.11, where the mean achieved rate is presented as a function of μ for different values of L and β. Here, the fading scenarios presented are for $\beta = 20$dB and $\beta = 6$dB representing a medium and low SNR scenario, respectively. In all four curves presented, an optimal value of μ can be found. For $\beta = 20$dB, the optimal values of μ are 0.5 (M-PAM) for $L = 3$ and $\mu \approx 0.42$ for $L = 7$. For $\beta = 6$dB, the optimal values of μ are ≈ 0.45 for $L = 3$ and $\mu \approx 0.42$ for $L = 7$.

In the case of unknown SNR, it can be seen that, depending on the distribution from which the SNR is taken, μ-PAM can offer performance benefits over that of M-PAM. Subsequently we shall explore how μ-PAM can be integrated with Rateless Coding presented in Chapter 2.

3.4 μ-PAM Modulation and Rateless Codes

We have explored and analyzed the μ-PAM modulation scheme for AWGN channels in which the SNR is not known at the transmitter. The purpose for doing this is to find a modulation method that is suited for communication over unknown wireless channels, and for use with the rateless codes presented in Chapter 2. We describe how μ-PAM may be used jointly with rateless codes to provide a complete communication system.

On the transmitter side, the combination of encoding and modulation is quite natural. Recall that a rateless code produces a potentially infinite number of coded bits. Sequential bits in multiples of L output by the encoder are mapped into a single modulated symbol and is transmitted on the channel.

At the decoder, the integration of the demodulator and decoder is only slightly more complex. Recall that rateless codes are well suited to handling arbitrary erasures on the channel. Since the receiver knows the channel state (SNR in the case of AWGN channels),

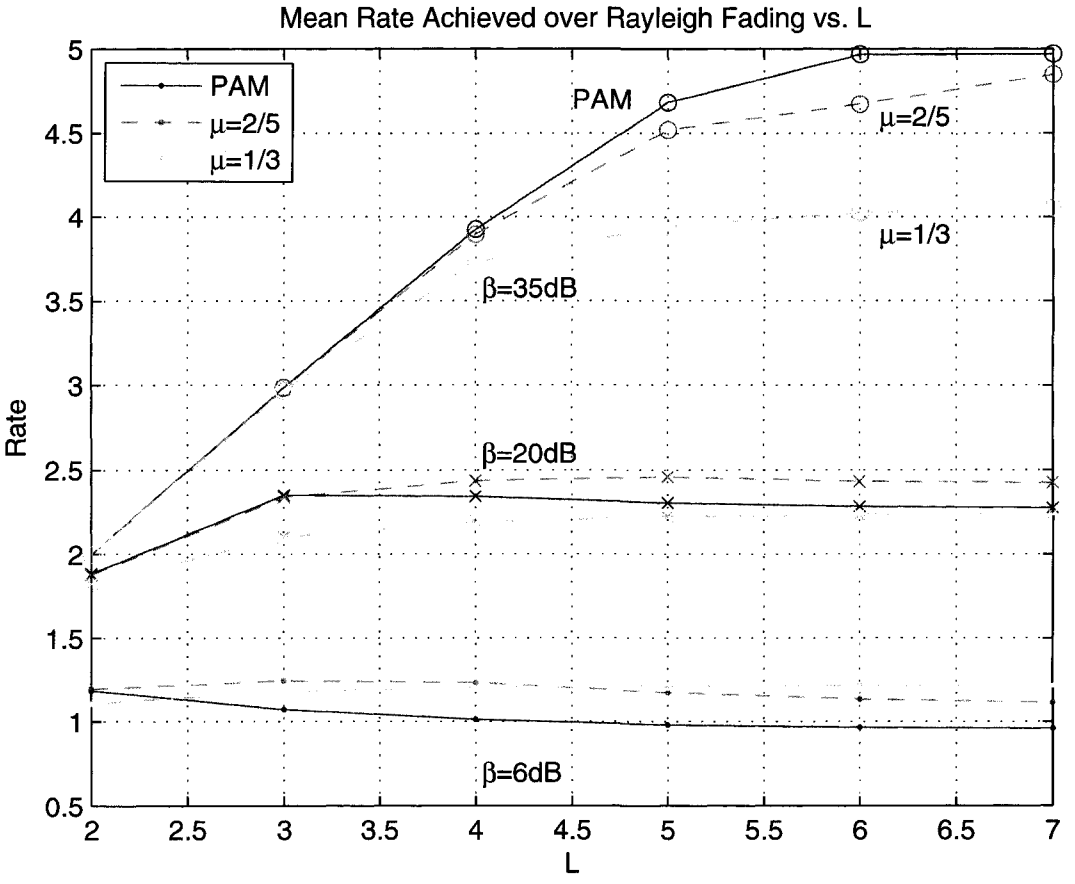

Figure 3.10: Mean achieved rates vs. L for the Rayleigh fading model

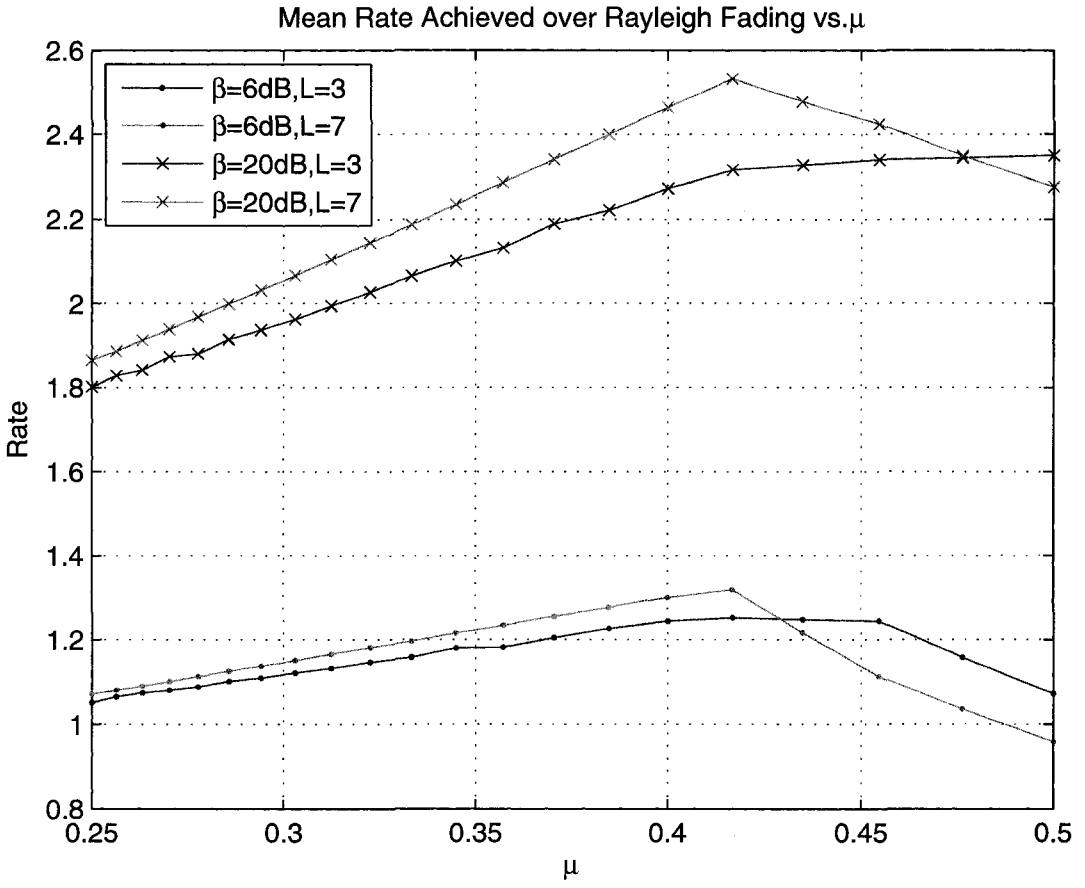

Figure 3.11: Mean achieved rates vs. μ for the Rayleigh fading model

M	μ	$d^1_{\min}$	Name
2	1/2	2	BPSK
16	1/2	1/8	16-PAM
16	1/3	$82/81 \approx 1$	16-1/3-PAM
16	1/4	$342/256 \approx 4/3$	16-1/4-PAM

Table 3.2: Configurations for μ-PAM simulation

it can choose an appropriate number of bits to decode. For example, if the channel supports only 1.3 bits per channel use, there is on average no benefit to demodulating more than the first two bits of the received symbol. In this case, the demodulator may choose to only decode 2 of L bits and set the remainder of the bits to erasures. This has a two-fold benefit. First, demodulated bits which have virtually no information content are prevented from being included at the decoding stage, which helps performance. Second, the smaller number of non-erased bits at the decoder results in a smaller computational requirement to decode.

This approach differs from previous works in the following ways. First, our focus is on the reliable transfer of information over unknown channels. Other work with this focus, in particular [5], uses hard-decisions at the demodulator. The μ-PAM modulation scheme is well suited to providing soft-valued outputs. Other works that use hierarchical modulation methods such as [60, 54] make use of non-uniformity to induce non-equal error protection for different data streams. Here we note that μ-PAM allows a virtually infinite amount of bits to be embedded into a symbol, which makes it well suited for rateless communication systems.

In this section we simulate a joint rateless coding and modulation system to demonstrate the performance of such a communication scheme and to verify the results presented in the previous sections. In order to test the proposed system, we perform a Monte-Carlo simulation using the Raptor code presented in [66] and following the construction given in [59], with $k = 10,000$. The LDPC component of the Raptor code has rate 0.95, and we use the same degree distributions as those in [66] for the outer (LDPC) and inner (LT) components of the Raptor code. The performance of this code has been reported in [59] for several lossy channel models including binary-input AWGN channels.

Belief propagation is used for decoding attempted after a specified set of channel uses. This set of channel uses corresponds to rates equal to capacity and all 1% incremental reductions thereof. Decoding is successful if the decoder converges to the correct trans-

mitted codeword within 100 iterations of belief propagation. The highest rate at which decoding is successful is the *achieved* rate for that trial. There is no upper limit on the number of decoding attempts that may be made by the decoder.

Four different modulation types were tested which are given in Table 3.2. Note that the values of $d^1_{\min}$ values are not directly comparable, as each configuration may have a different average transmit power. Nevertheless, it serves as a useful estimate of worst-case degradation at low SNR.

Figure 3.12 presents the curves of mean achieved rate $\bar{R}$ versus SNR for different modulation orders and values of $\mu = \{1/3, 1/4\}$. Mean achieved rate for T codeword transmissions is computed as

$$\bar{R} := \frac{Tk}{\sum_{t=1}^{T} n_t},$$

where n_t is the required number of channel uses by the t-th codeword to successfully decode. Note that this is simply the empirical form of (3.7). In all cases the mean rate follows the capacity of the channel, with some loss, subject to the constraint of the modulation order. For $L = 4$ ($M = 16$), we see as a function of μ that there is a tradeoff in performance at low and high SNR. The M-PAM results, equivalent to $\mu = 1/2$, performs best at high SNR, but worst at low SNR. Conversely, the $\mu = 1/4$ performs best of the $L = 4$ modulation types at low SNR, but worst at high SNR. As suggested by Theorem 3, the $L = 4$ modulation schemes are inferior to BPSK modulation at low SNR. The results for $\mu = 1/3$ offers a tradeoff, providing the best performance at moderate SNR values, without much loss (relatively) at either high or low SNR.

This comparison is more clearly seen in Figure 3.13, where the gap to capacity (in dB) is plotted as a function of channel SNR. Gap to capacity is used here to be the difference (in dB) between the SNR at which C and the implemented mean rate are equal in value. The μ-PAM modulation scheme with $\mu = 1/3$ shows very good performance at low SNR, while having a relatively graceful degradation in performance relative to capacity. The $M = 16$ M-PAM results are relatively poor at low SNR, but at relatively high SNR, have a fairly constant gap to capacity. This constant gap corresponds to the range of SNR at which M-PAM is commonly used for fixed-rate communications schemes. The fact that the gap to capacity actually shrinks as the SNR increases from 20dB to 23dB is simply reflective of the fact that this particular choice of M is very good in this SNR range. Balance between performance at a particular SNR regime and graceful degradation can be achieved by judicious choice of M and μ.

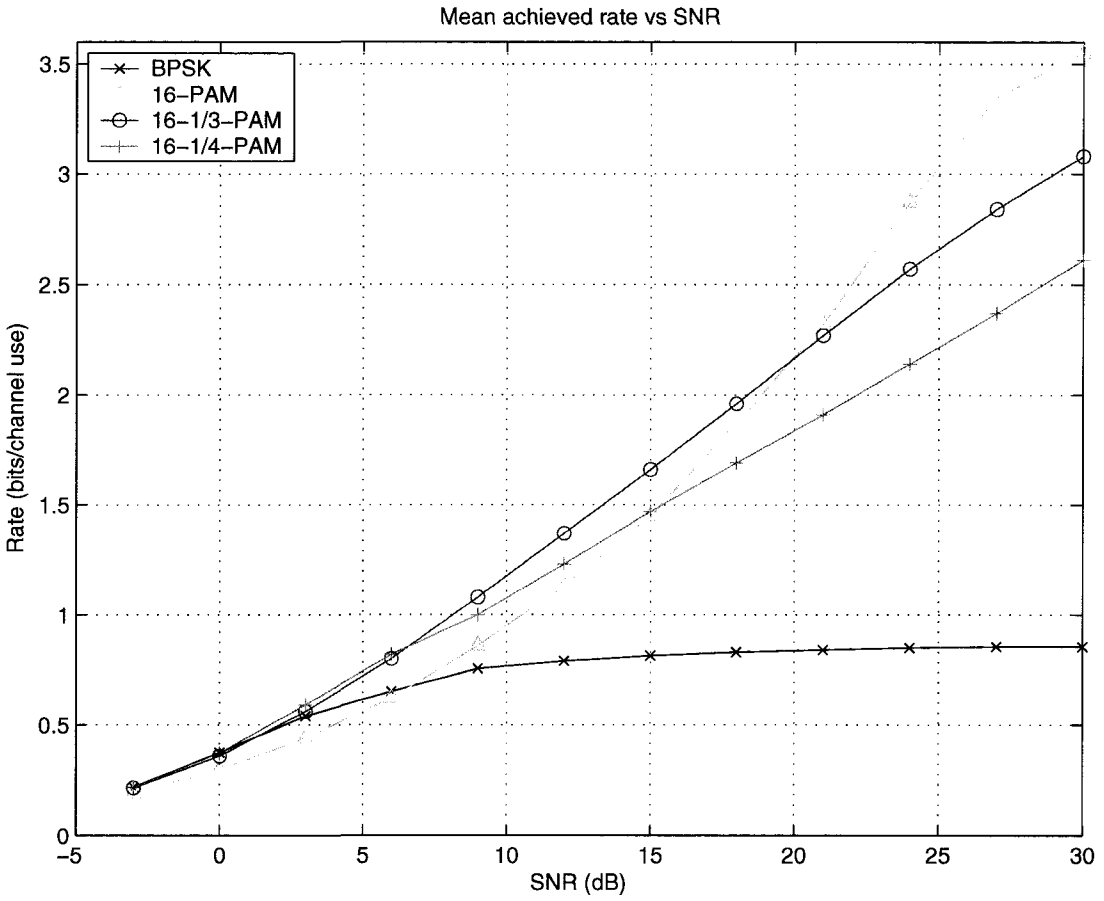

Figure 3.12: Mean achieved rate vs. SNR

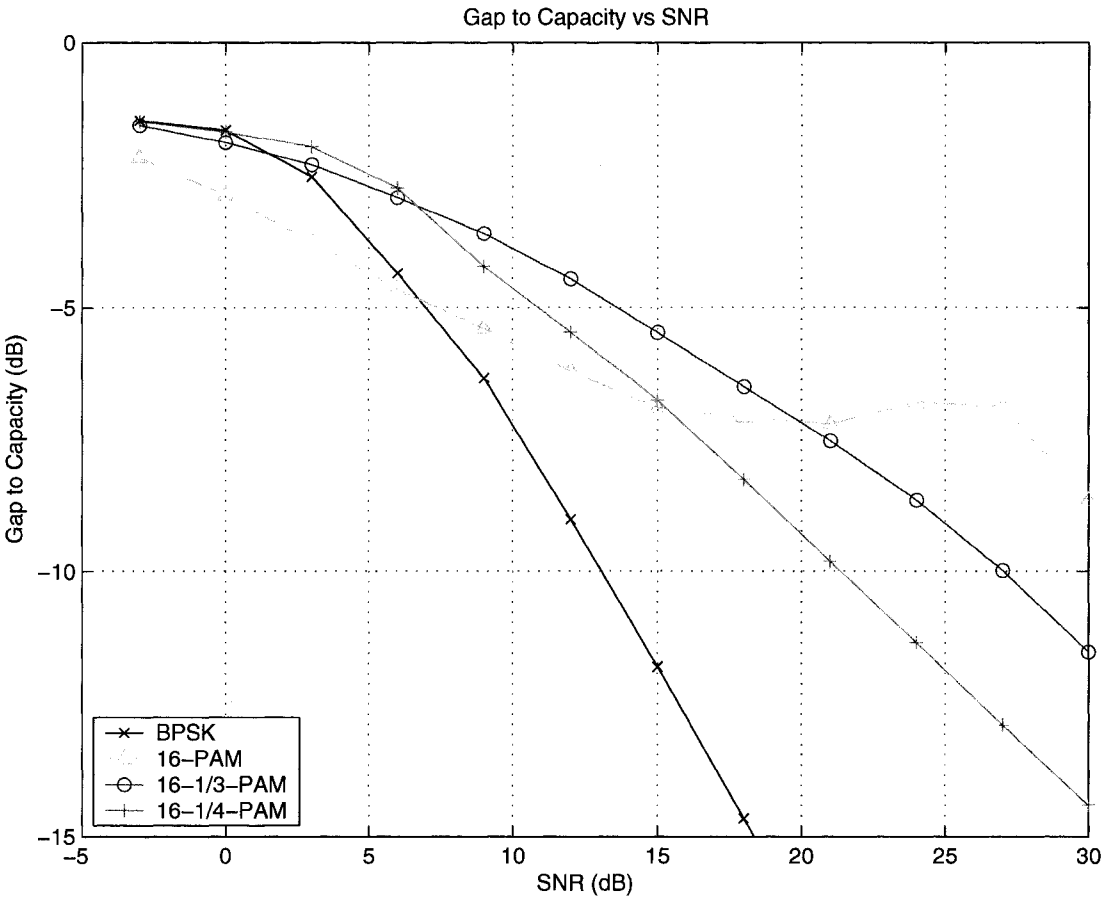

Figure 3.13: Gap to capacity for different modulation types vs. SNR

3.5 Conclusions

We have presented a modulation system which can be used for reliable communication over unknown AWGN channels that performs well over a broad range of channel conditions. Using the μ-PAM modulation scheme, a non-uniform reliability across the bits comprising a constellation symbol is induced. Coupling this modulation scheme with rateless codes allows for good performance in terms of mean achieved rate. Simulations were performed showing that the use of this modulation scheme can be configured to provide good performance at low SNR, while adapting naturally to allow higher rate communication as the SNR increases. Degradation in performance relative to the capacity of the channel was shown to be graceful.

An extension of this work would be to adapt M automatically as a function of the number of channel uses. The use of higher-dimensional modulation, thereby recovering some shaping gain may be interesting to investigate. The extension to two dimensions is identical to the extension of PAM to QAM. Finally, one may consider that certain bits of the modulation symbol may be too unreliable to be used in decoding, and so some erasure techniques, possibly similar to that used in [5] may be beneficial.

As seen in this chapter, joint rateless coding and modulation methods can be used which may result in more robust system performance over a broad range of operating points. Traditional modulation approaches, which map a fixed number of coded bits to modulated symbols, typically have only a relatively small range of SNR over which they can closely follow the capacity of the channel. At too high of an SNR, the system is limited by the rate of the modulation type. At too low of an SNR, the system performance suffers due to the low reliability of the demodulated symbols.

Indeed, even the results presented above can be said to suffer from some of the same limitations. Practically, one would typically put an upper limit on the number of modulated symbols. As such, the operating range of the proposed scheme, though greater than fixed-modulation techniques, is still limited. Other design options such as the extensions discussed above will likely result in larger operational ranges exhibiting acceptable performance.

The joint use of rateless coding and modulation holds promising potential for reliable communication over unknown channels. We have shown that the rateless codes presented in Chapter 2 and the modulation schemes presented in this chapter are well suited for use together. In subsequent chapters we will explore more properties of this joint communication system in different environments and applications.

Chapter 4

Rateless Coding with Delay Constraints

We have seen in Chapter 2 that rateless codes have properties that make them well suited for use over fading channels such as those in wireless systems. In particular when the transmitter has no knowledge of the channel we saw that rateless codes are able to achieve the realized capacity of the channel.

We also explored the problem of when a message is decodable over a fading channel and saw that the answer is random. We determined the distribution for the decoding time, and presented some results of a simulation of a Raptor code over a slow fading channel.

This chapter focuses on the situation in which there is a delay constraint in the system. Delay constraints are a common and practical limit that is built into many communication systems. Their use may be a result of technical limitations, such as a maximum acceptable number of retransmissions that a block-based coding system may impose, or may be due to quality-of-service guarantees. For example, a communication system may impose a limit on the latency of communication from source to destination.

When a delay constraint is imposed over fading channels with unknown state at the transmitter, it is almost certain that outage probability will be bounded away from zero. This is clearly true for fixed-rate communication schemes, but we will show that when rateless coding is applied, under certain circumstances it is possible that long-term outage probability can be driven to zero, even in the presence of delay constraints.

This will be seen in the context of a video streaming application where a delay constraint is imposed due to a quality-of-service requirement on the latency of transmission.

For this application we compare fixed-rate coding schemes, some existing variable-rate coding systems, and rateless codes. We show that rateless codes offer the best overall performance when comparing the metrics of throughput, probability of outage and mean time to decode.

The remainder of this chapter is organized as follows. Section 4.1 presents some background of the problem of reliable communication with delay constraints. The framework using rateless codes for delay constrained communication is presented in Section 4.2, where performance metrics are introduced and some key results are developed. We also present some simulation results of the video-streaming application. Section 4.3 expands on the results of the investigation using the modulation method presented in Chapter 3. We conclude the chapter in Section 4.4.

4.1 Background

The development of fountain codes [48, 66] inspire us to consider using rateless coding ideas for reliable communication over channels with receiver-only state information and delay constraints. An overview of communication under channel uncertainty is presented in [45]. Some applications in this area include, e.g., the works in [20,71,58]. Here we take channel uncertainty to mean that only the receiver is aware of the state of the channel (or channel law) and therefore the transmitter is unable to choose an optimal codebook *a priori*. We also assume the existence of a limited-capacity feedback channel. By limiting the capacity of the feedback channel, we limit the effectiveness of the strategy in which the receiver directly signals an optimal codebook to the transmitter.

This problem is of practical concern since many wireless communication systems utilize asymmetric duplexing in which the feedback channel is allocated fewer resources: power, bandwidth, etc. Additionally, channel uncertainty is a common characteristic in wireless settings due to the time-varying nature of the channel. Finally, many applications that operate over wireless channels seek to deliver latency-sensitive information, so delay constraints naturally exist.

With such applications in mind, the channels we consider are fading channels with unknown states at the transmitter. The channel state may vary with time and the rate of this variation may result in different behaviors that require different communication strategies. For quasi-static fading channels, in which the channel conditions remain fixed for the duration of a codeword, one may use feedback in order to signal the channel state information to the transmitter. For fast-fading channels, in which channel conditions vary

independently from symbol to symbol, one may rely on the law of large numbers and code for the expected case thereby achieving the ergodic capacity of the channel. Between these extremes, one may consider the block-fading channel with a delay constraint, where there are only a few fading blocks during the transmission of a codeword. In this case the above mentioned strategies fail to be efficient.

In this regime, the work in [56] examines the impact of causal feedback on the capacity of delay-constrained block-fading channels and presents an optimal power-control strategy under channel uncertainty. These results are extended in [7, 80] for multiple access channels with variable-rate coding strategies. More generally, the work of [32] characterizes the capacity region of the delay-constrained multiple access channel, and presents an optimal resource allocation scheme.

The works in [58, 69, 65] present performance analyses for variable-rate and rateless coding strategies using LDPC codes and fountain codes, demonstrating their applicability over a range of SNR. It is worth noting that rateless coding may be considered as a type of Hybrid-ARQ scheme or variable-rate strategy, both of which have been used for communication under channel uncertainty. For example, using a block-incremental redundancy strategy, the throughput for the Gaussian collision channel is analyzed in [6]. Furthermore, the information theoretic analysis in [20] has demonstrated that rateless codes can achieve the mutual information induced over the channel by the choice of input distribution for the class of all discrete memoryless channels with no channel state information. Similar results are found in [71, 67].

The main contribution of this work is to demonstrate that rateless codes offer advantages in terms of throughput and outage probability for delay-constrained streaming applications over fading channels. In particular, we show the existence of a phase-transition effect for rateless coding strategies and that beyond the phase-transition threshold, the outage (or frame-jitter) probability can be driven to zero. This makes rateless codes an attractive solution for such applications.

4.2 Rateless Codes for Delay-Constrained Communication

The problem of reliable streaming of data over the block-fading channel is useful to study as it is particularly well matched to streaming data and other delay-sensitive applications over wireless channels. We structure the problem in a similar manner to the settings

in [80] or [6], where the channel gains follow a Rayleigh distribution.

Our setting models, e.g., a video-streaming application where the transmitter has an infinite buffer of video data to send and the receiver has an infinitely large buffer to store the received data. Each video frame consists of k bits and has a *playback deadline*, which requires that at time lD, $l \in \{1, 2, \ldots\}$, the l^{th} frame must be displayed, or otherwise used by the application. An *outage* event $\mathscr{E}_l$ is said to occur if the l^{th} frame is not decoded correctly by its playback deadline. This notion of outage corresponds to a 'frame jitter' in the video-streaming context, and reduces to the standard notion of outage for fixed-rate codes with block length D.

Since we consider a fixed coherence time T_c, we often express the pair (D, k) as $(D/T_c, k/T_c)$ to remove the dependency of coherence time in our analysis. It is useful to examine typical values of k/T_c and D/T_c one might find in current wireless systems such as UMTS and WiMAX [76, 78]. The information data rate for UMTS is typically about 1Mb/s. Using the channel models proposed by UMTS, coherence times for mobility devices in the 1.9 to 2.4 GHz range can typically fall in the range of 10ms to 100ms. Using these numbers, a value for k/T_c may lie between 0.1 and 10. Similar values can be found for WiMAX and other wireless standards based on their suggested channel models, despite different channel structures. Typical values for D/T_c are commonly between 0.1 and 100, depending on the standard. OFDMA based systems such as WiMAX would typically have smaller values of D/T_c due to the relatively slow over-the-air OFDM symbol rate, whereas direct sequence CDMA based systems such as UMTS would have larger values.

For the streaming application, we study three strategies: a fixed-rate (FR) coding strategy and two rateless coding strategies. The rateless strategies are a continuous-incremental-redundancy (CIR) scheme and a block-incremental-redundancy (BIR) scheme.

In the fixed-rate coding strategy, each frame of k-bits is encoded into a codeword of length D and frames are transmitted consecutively. In the CIR and BIR schemes, a frame is encoded and transmitted in a rateless manner until the receiver decodes or an outage occurs. The difference between CIR and BIR is that in CIR, decoding is attempted at every channel use, whereas in BIR, decoding is attempted at time instants that are integer multiples of T_c. In all three strategies, the l^{th} frame is transmitted immediately after the $(l-1)^{\text{th}}$ frame is decoded or after the playback deadline for the $(l-1)^{\text{th}}$ frame has passed, whichever occurs earlier.

Given any of the coding schemes above, let $A_l(\zeta)$ be the time at which the last (codeword) symbol of the l^{th} frame is transmitted over channel realization ζ, and $A_0(\zeta) :=$

0. Let $n_l(\zeta)$ be the number of channel uses required to transmit the l^{th} frame over channel realization ζ, assuming no play-back deadline. Then $A_l(\zeta) := \min(A_{l-1}(\zeta) + n_l(\zeta), lD)$. For the fixed-rate scheme, $n_l(\zeta) = D$ for all l due to the deterministic codeword length. For the CIR scheme, we take $n_l(\zeta)$ as the smallest n such that the capacity $C(A_{l-1}(\zeta) + 1, A_{l-1}(\zeta) + n; \zeta) \geq k/n$; and for the BIR scheme, we take $n_l(\zeta)$ as the smallest integer multiple n of T_c such that $C(A_{l-1}(\zeta) + 1, A_{l-1}(\zeta) + n; \zeta) \geq k/n$.

Thus, for the purposes of comparing the three schemes, we assume they are ideal. The two rateless schemes are assumed ideal in the sense that the correct decoding occurs only when the realized capacity supports the realized rate. The fixed-rate scheme is assumed ideal in the same sense, i.e., that an outage event $\mathscr{O}_l$ occurs if and only if $C((l-1)D + 1, lD; \zeta) < k/D$.

The two rateless schemes allow for the utilization of any left-over channel uses made available by early decoding of prior codewords. This can have a dramatic effect on the long-term behavior of decoding statistics. For example, if a rateless codeword is successfully decoded before its playback deadline, then there exist some extra channel uses that the transmitter may opportunistically exploit to begin the transmission of the next codeword. This in turn increases the probability that the next codeword will be decoded early. Thus there is a compounding effect. As the frame number increases, an essentially infinite number of extra channel uses may become available for transmission, driving the outage probability to zero.

For a given rateless scheme as described above, we use $S_l(\zeta) := lD - A_l(\zeta)$ to denote the time "saved" after the transmission of the l^{th} frame, compared to the l^{th} playback deadline. We treat $S_l(\zeta)$ as a random walk with increments given by $\Delta S_l(\zeta) := S_l(\zeta) - S_{l-1}(\zeta) = D - \Delta A_l(\zeta)$, where $\Delta A_l(\zeta) := A_l(\zeta) - A_{l-1}(\zeta)$. It is straightforward to show that the following property of $n_l(\zeta)$ holds: [68].

Lemma 4 *When treated as a random process, if the support set of $n_l(\zeta)$ is continuous, then $n_l(\zeta)$ is* asymptotically *stationary and ergodic in l.*

The proof is given in Appendix 4.A.

As a consequence, $\Delta A_l(\zeta)$ and $\Delta S_l(\zeta)$ are both asymptotically stationary and ergodic. Then as l approaches infinity, the sequence $\{n_l(\zeta)\}$ of random variables converges to some random variable $n^*(\zeta)$ in distribution.

Theorem 5 *For the rateless schemes, if $E[n^*(\zeta)] < D$, where expectations are taken over channel and noise realizations, then for any $\delta > 0$ the probability that the number*

of saved channel uses exceeds any finite $M > 0$ can be lower bounded as

$$\Pr[S_l > M] > 1 - \delta, \tag{4.1}$$

for sufficiently large l. Conversely, if $E[n^*(\zeta)] > D$, then for any $\delta > 0$, there exists some finite $M > 0$ such that for all l,

$$\Pr[S_l < M] > 1 - \delta. \tag{4.2}$$

The proof of this theorem is provided in Appendix 4.B.

This theorem demonstrates that the outage probability is characterized by a phase transition. We define the average outage rate for transmitting L frames as

$$P_{\text{out}}^L := E\left[\frac{\sum_{l=1}^{L} I[\mathscr{E}_l]}{L}\right],$$

which approaches the outage probability as $L \to \infty$, and where $I[\cdot]$ denotes the indicator function.

Corollary 2 *For the rateless schemes, if $E[n^*(\zeta)] > D$, then for all L sufficiently large, $P_{out}^L > \alpha$ for some $\alpha > 0$. If $E[n^*(\zeta)] < D$, then for any $\epsilon > 0$, $P_{out}^L < \epsilon$ for sufficiently large L.*

The proof of this corollary is given in Appendix 4.C. From this corollary we see the effect of the phase-transition about the critical point $E[n^*(\zeta)] = D$. As long as $E[n^*(\zeta)] < D$, the long-term behavior of rateless codes exhibits virtually zero outage probability, as if no delay constraint exists. This forms a significant contrast with fixed-rate codes, for which the outage probability is always bounded away from zero. We note that given a rateless coding strategy, $E[n^*(\zeta)] = f(\gamma, k)$ for some function f of SNR γ and frame size k, where $f(\gamma, k)$ always increases with k and decreases with γ. It follows that any two values of the triple $(\gamma, k/T_c, D/T_c)$ define the critical value for the third parameter via the equality $f(\gamma, k) = D$ for a given rateless scheme.

In order to compare the performance of the three coding strategies – fixed-rate, block-incremental-redundancy and continuous-incremental-redundancy – we numerically analyze each strategy for the data streaming application assuming $L = 100,000$ codewords. For each strategy and channel configuration, the probability mass function of decoding times is readily computable, and this is used as the basis to determine the outage probability and throughput presented subsequently. We note that in the following figures, the

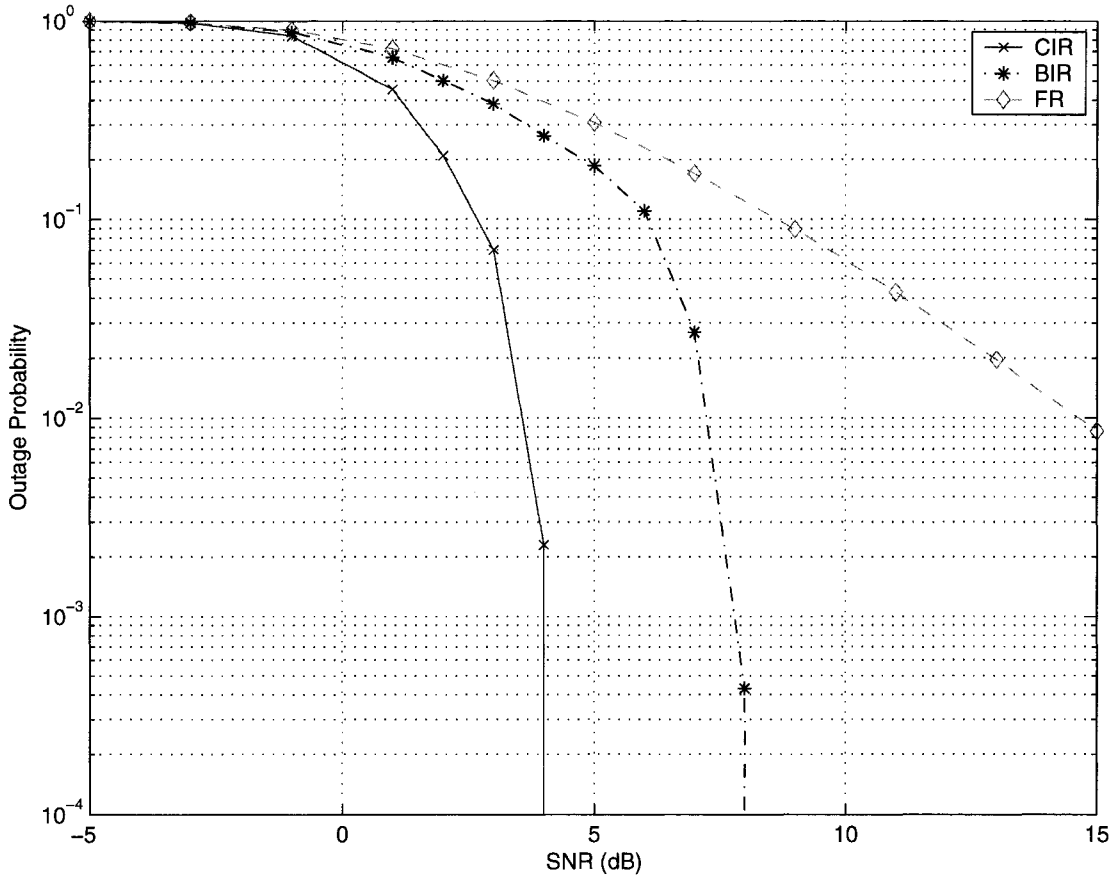

Figure 4.1: Average probability of outage vs. SNR, $k/T_c = 2$, $D/T_c = 1.5$.

parameters were chosen to contrast the different coding strategies. In practice, a code designer may be able to select a set of parameters that are better suited to a particular coding strategy.

Numerical computations are made to derive all of the figures in this section. The channel model used is given by (2.1), with block-fading length of T_c. Random values are drawn to derive channel realizations, ζ, and these are used to determine the performance metrics such as probability of outage, etc., assuming that the codes are capacity achieving. For FR-based coding, each $h \in \zeta$ maps to a realized capacity, each of which are used to attempt to transfer k-bits of information. Assuming an ideal code, the k-bits are received correct if and only if the realized capacity exceeds the code rate, k/D.

Similar approaches are used for the BIR and CIR strategies. However in this case, we keep track of how many channel uses are required before either the transmission is completed or the end of a coherence block is reached. This way we can account for every channel use. Despite the fact that we use the unrealistic assumption of ideal codes, these results serve to provide useful insight into the relative performance of the different strategies and their properties.

Using the described numerical computation, the outage probability as a function of SNR (calculated for $L = 100,000$) is given in Figure 4.1. We see that as SNR increases to pass their respective critical points, the outage probability curves for the two rateless schemes fall abruptly. For the CIR strategy, this critical point SNR* occurs at about 4dB and is about 8dB for the BIR strategy. For all SNRs the outage probabilities for the rateless strategies are less than that of the fixed-rate strategy. The CIR scheme also performs better in comparison with the BIR scheme, both in terms of the magnitude of outage probability and in terms of lower SNR threshold for the phase-transition. Intuitively, this follows due to the greater number of opportunities that the CIR has for decoding before the playback deadline. Of course, this is achieved at the expense of greater complexity. One may verify that the fixed-rate outage probability curve matches the theoretical curve of error probability for a block-fading channel with the given parameters.

The outage probability of the system is presented as a function of k/T_c in Figure 4.2. The phase-transition effect is clearly seen for both incremental redundancy strategies, with the critical value of k/T_c of about 4.2 for CIR and 2.7 for BIR. The greater value allows a greater resolution (or higher data rate) for the same delay. No such behavior is seen with the FR strategy, and the outage probability scales as expected as the rate of the code decreases with decreasing k/T_c.

As a function of D/T_c, outage probability is presented in Figure 4.3. For the given

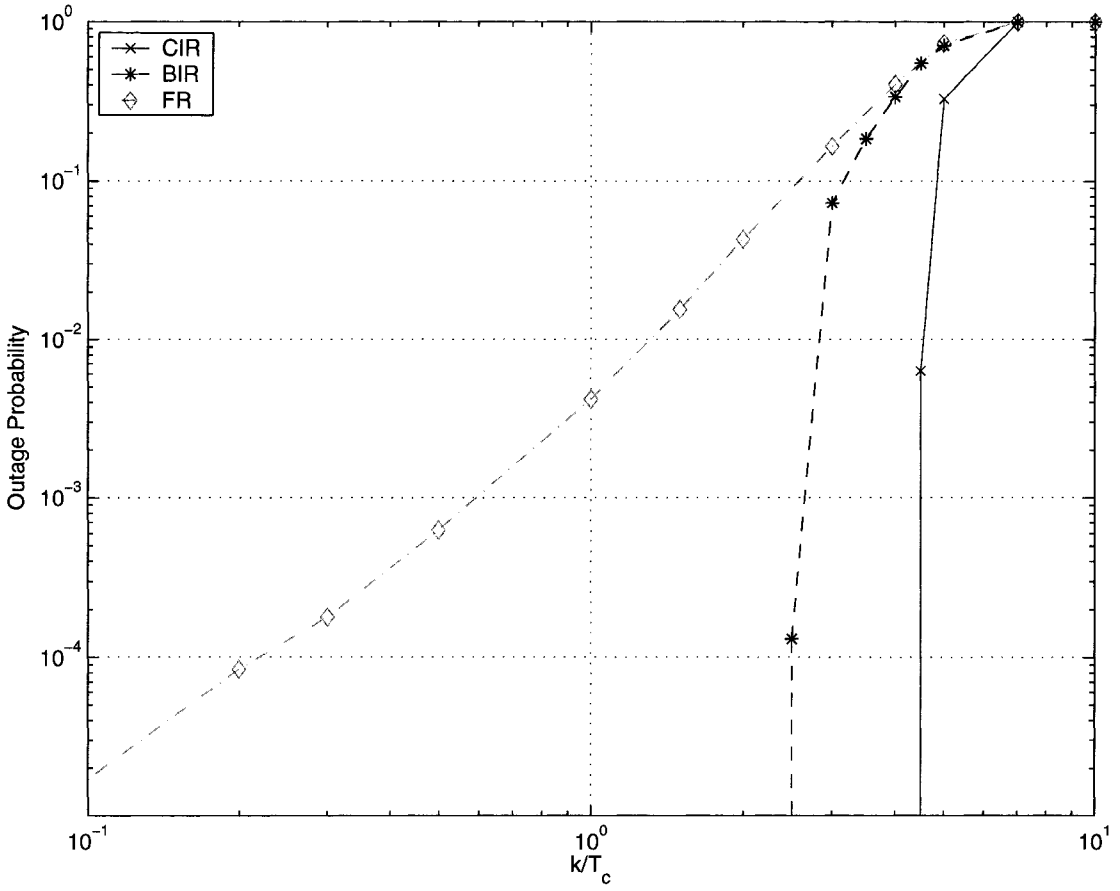

Figure 4.2: Average probability of outage vs. k/T_c, SNR=11dB, $D/T_c = 1.5$.

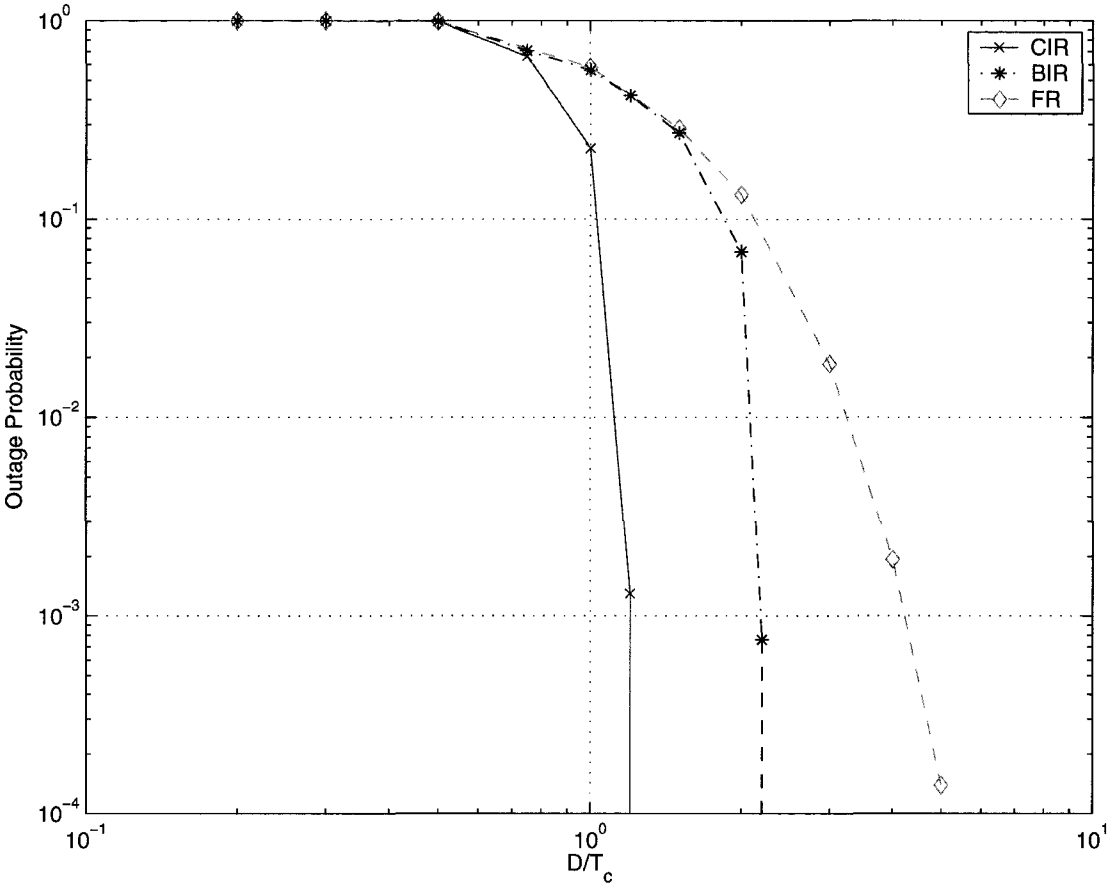

Figure 4.3: Average probability of outage vs. D/T_c, SNR=3dB, $k/T_c = 1.5$.

configuration, a complementary behavior can be seen as compared with the results in Figure 4.2. The critical value of D/T_c is about 1.4 for CIR and 2.5 for BIR. Here, for fixed k/T_c, a smaller delay is beneficial, though more difficult to achieve. In all of the figures presenting outage probability, the benefits of the incremental redundancy strategies, particularly CIR, can be seen. These benefits come as a result of the codes' ability to naturally adapt rate to the channel realization and the presented figures verify the phase-transition effect predicted by Corollary 2.

We also compare the performance of the three schemes in terms of throughput, where the throughput of a given coding strategy refers to the expected number of information bits that are *correctly* received per channel use. More precisely, the throughput ν_L in transmitting L frames is defined as

$$\nu_L := E_\zeta \left[\frac{\sum_{l=1}^{L} k \cdot I[\mathscr{E}_l]}{A_L(\zeta)} \right]. \tag{4.3}$$

The throughput for the three schemes are computed numerically for $L = 100,000$. Throughput curves as functions of SNR, k/T_c and D/T_c are plotted in Figures 4.4, 4.5 and 4.6, respectively.

In Figure 4.4, the advantage of rateless schemes over the fixed-rate scheme can be clearly seen. For comparison, we also plot the throughput curve of CIR under the $D = \infty$ limit. The effect of delay constraints diminish as the SNR increases to pass a threshold, which in fact corresponds to the phase-transition point. The BIR curve is bounded away from the CIR curve with $D = \infty$ since the maximum throughput of the BIR scheme is limited by the first decoding attempt at time T_c. For the given parameters, this causes the protocol to saturate at a throughput of 2 bits/channel use.

Figure 4.5 plots the dependence of throughput on k/T_c. Different values of k/T_c may correspond to different qualities of video frames, whereas the value of throughput indicates a compound effect of user experience in terms of frame-jitter rate (or outage events) and video quality. For large values of k/T_c, we see in the figure that the given SNR and delay constraint preclude the possibility of achieving high throughput — the channel simply cannot support it.

At low values of k/T_c, corresponding to a low quality video stream, the throughput of both the BIR and fixed-rate strategies are limited by the minimum length of their codewords. The channel supports much higher rates, so these low values of k/T_c represent a poor matching of the coding scheme to channel conditions. However the CIR scheme is able to maintain a relatively large throughput, mainly as a consequence of the ability

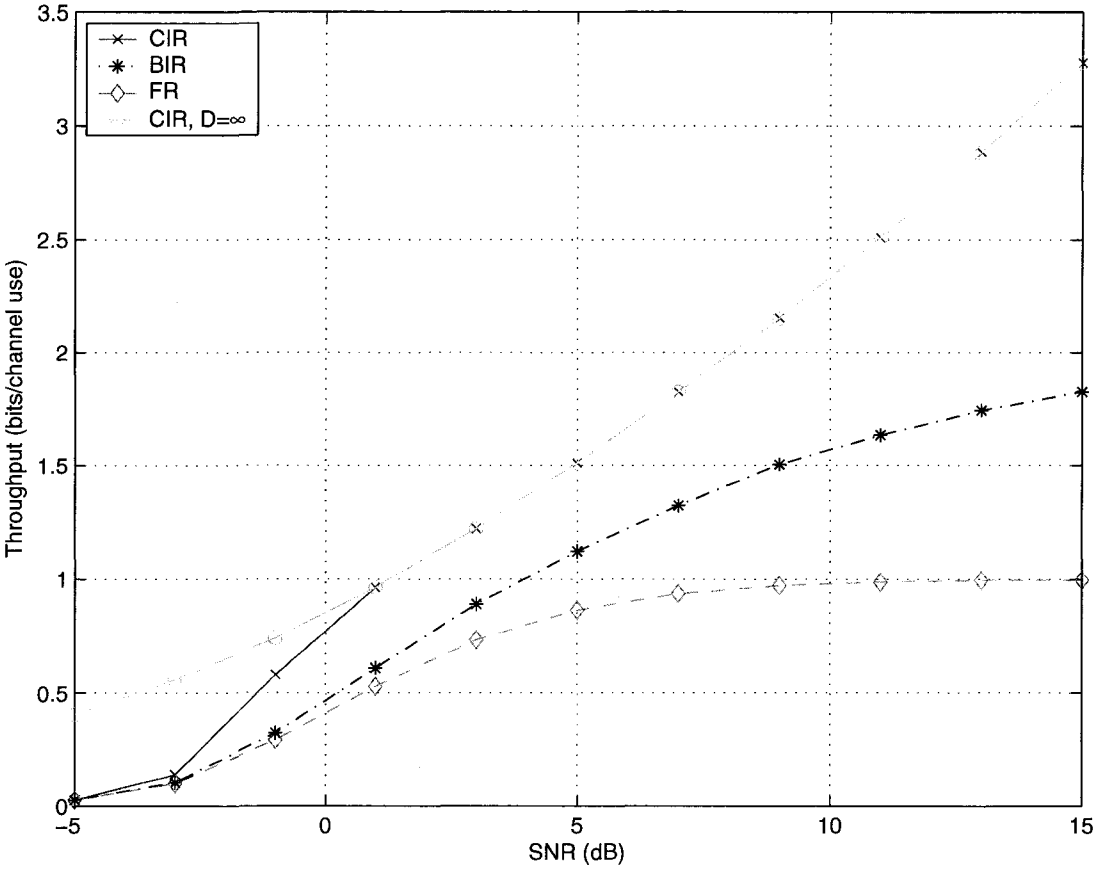

Figure 4.4: Throughput vs. SNR, $k/T_c = 2$, $D/T_c = 2$.

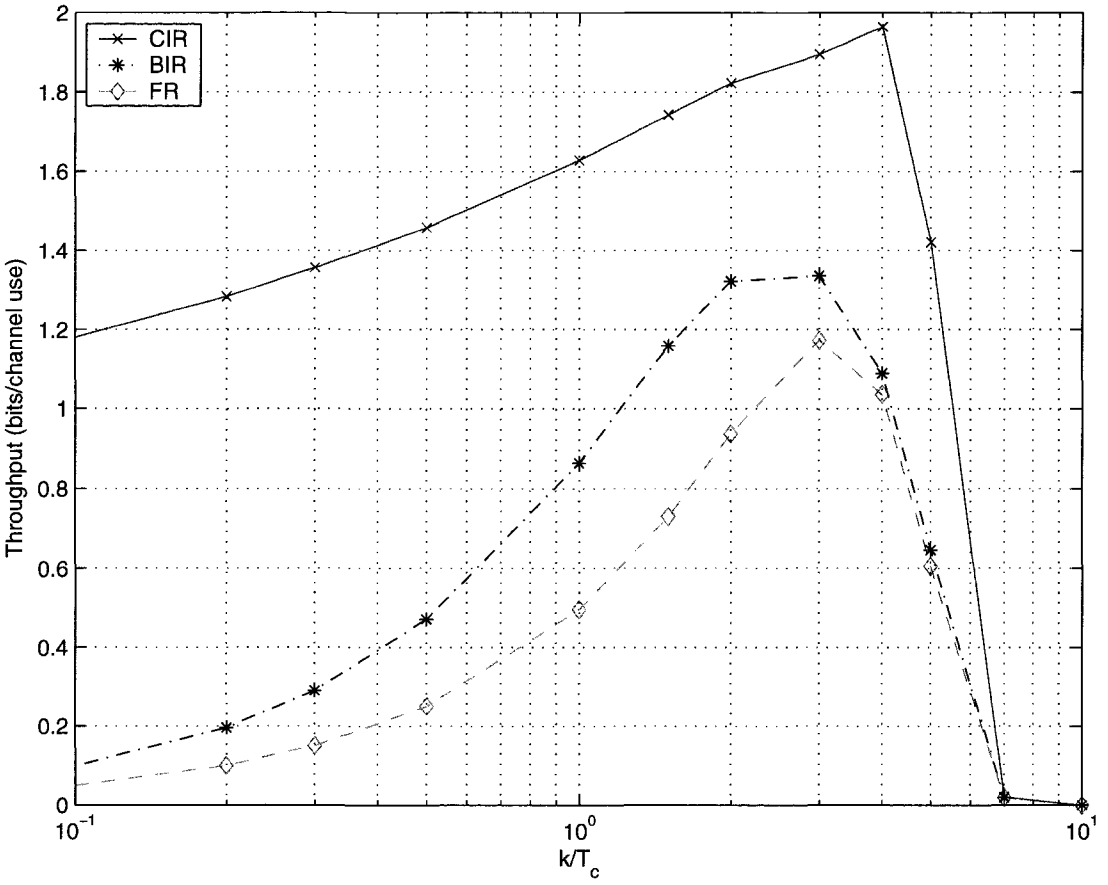

Figure 4.5: Throughput vs. k/T_c, SNR=7dB, $D/T_c = 2$.

to decode before T_c channel uses, but also due to lower outage probability.

Figure 4.6 presents the throughput of the three coding strategies as a function of D/T_c at a fixed SNR and k/T_c. At low values of D/T_c throughput is limited by the large outage probability, as the minimum allowed rate is not supported by the channel. At high values of D/T_c, the incremental redundancy strategies reach the maximum throughput at the rate supported by the channel SNR and k/T_c. The point at which this value is reached is determined by the critical value for D/T_c which is about 1.3 for CIR and 5 for BIR. The FR strategy suffers at high values of D/T_c due to the inefficient coding rate dictated by D/T_c.

In each of the presented figures the advantages of rateless coding strategies are apparent for the wireless streaming application studied. In terms of both efficiency and reliability, the CIR strategy is clearly the best of the three studied strategies, although both the CIR and BIR provide benefits compared to the FR strategy.

Despite their relative benefits, the results of the two rateless coding schemes are limited by the way in which feedback is used. Here, feedback is used only to indicate the successful decoding of a codeword. However, it is useful to consider the option of feeding back channel states to implement power-allocation or rate-allocation policies, and this may open up another dimension of the solution space providing additional benefits. This dimension is explored in, e.g., [7, 56, 69, 65].

4.3 Joint Coding and Modulation

We now expand upon some of the previous results to include μ-PAM modulation discussed in Chapter 3. In that chapter we presented results demonstrating that μ-PAM offers advantages in terms of mean achieved rate when communicating over unknown channels. In the context of communication with delay constraints, mean achieved rate is no longer the most meaningful metric. As used in the previous sections, throughput and mean decoding times are more applicable.

Appealing to the results of Theorem 5, recall that the relationship of mean decoding time $E[n^*(\zeta)]$ with the delay constraint D determines in the long run we if expect the streaming application considered in this chapter to have a positive outage probability. We see here that the mean decoding time is critical to the overall user experience.

To that end, we compare results of the mean decoding time achievable for M-PAM and μ-PAM for different values of L and μ. To be consistent with the previous results, we consider a block-fading channel with channel gain that is drawn from a Rayleigh

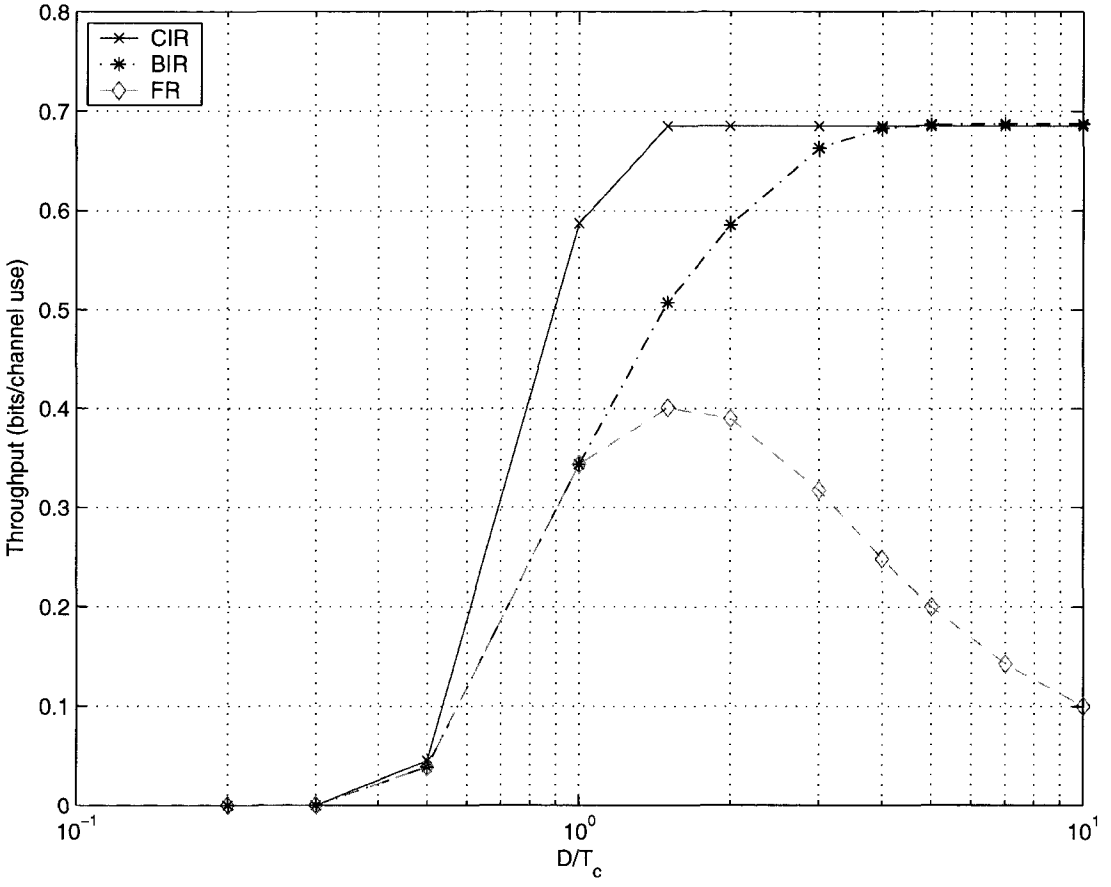

Figure 4.6: Throughput vs. D/T_c, SNR=-1dB, $k/T_c = 1$.

distribution with parameter β, known to the receiver only. The mean decoding time is determined using the mutual information bounds given by R from (3.6) and is defined as

$$E[n^*] = E\left[\frac{1}{R}\right].$$

Figure 4.7 presents the normalized mean decoding times in units of channel uses per information bit as a function of L for different SNR settings (β) and different values of μ. In the figure, three values of β are shown; $\beta = 35$dB, $\beta = 20$dB and $\beta = 6$dB, representing high, medium and low average SNR conditions respectively. It can be seen that for the high SNR condition, M-PAM is the best choice, while at low SNR, M-PAM provides the largest (worst) mean decoding times. At the medium mean SNR condition, $\beta = 20$dB, the best choice of modulation depends on L, with $\mu = 1/3$ being the best for large values of L.

This relationship is further investigated in Figure 4.8, where the mean decoding time is presented as a function of μ for different values of L and β. Here, the fading scenarios presented are for $\beta = 20$dB and $\beta = 6$dB representing a medium and low SNR scenario, respectively. In all four curves presented, an optimal value of μ can be found. For $\beta = 20$dB, the optimal values of μ are 0.47 for $L = 3$ and $\mu \approx 0.42$ for $L = 7$. For $\beta = 6$dB, the optimal values of μ are ≈ 0.42 for $L = 3$ and $\mu \approx 0.40$ for $L = 7$.

These figures demonstrate that integration of μ-PAM and the rateless codes described in this chapter can yield performance benefits from the perspective of improved mean decoding times. A consequence of this is that critical points representing $E[n^*(\zeta)] = D$ can be found at smaller SNRs than M-PAM for many situations. This is of course beneficial to the overall performance of a delay constrained application such as the video streaming application considered in this chapter.

4.4 Conclusions

We have examined the use of rateless coding concepts for delay-constrained communication applications over unknown fading channels. Ideal rateless coding schemes have been applied to a delay-constrained wireless streaming application, and under the transmission protocol studied, we demonstrate the existence of a phase-transition in the outage probability for rateless codes. Beyond this threshold the protocol essentially guarantees zero outage probability. A comparison of rateless codes and fixed-rate codes has demonstrated the improvements possible in throughput and outage probability when rateless

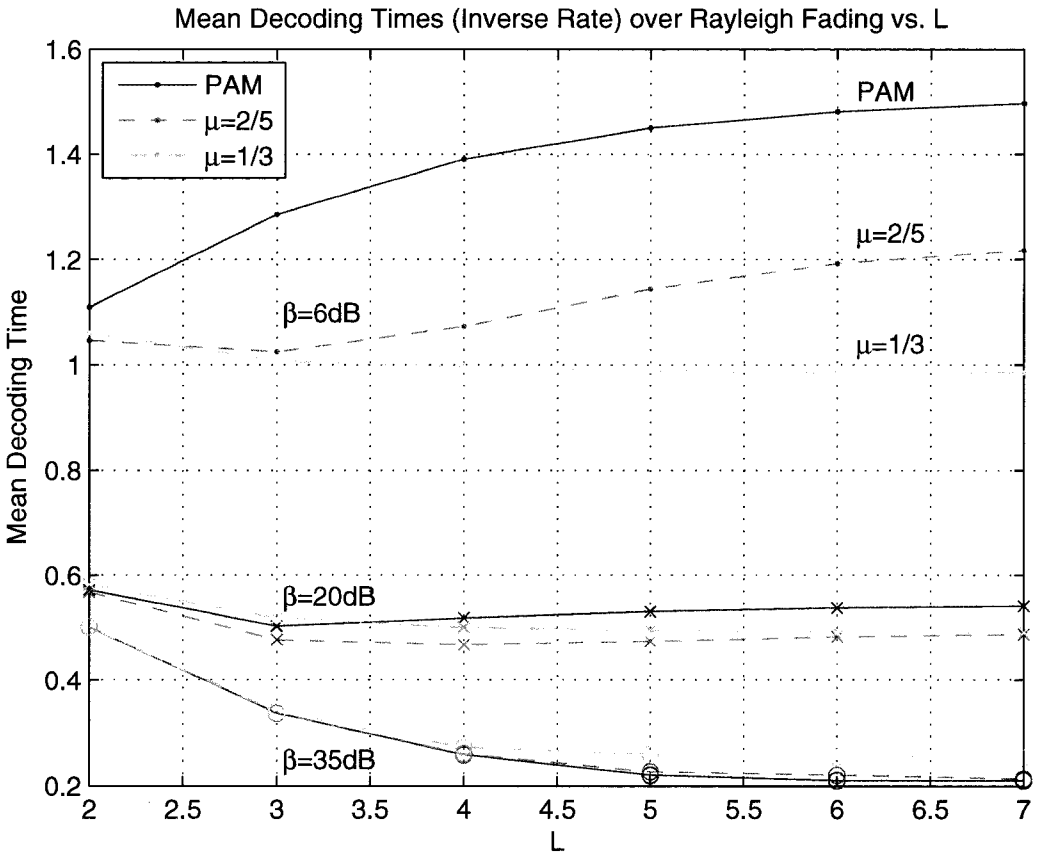

Figure 4.7: Mean decoding times vs. L

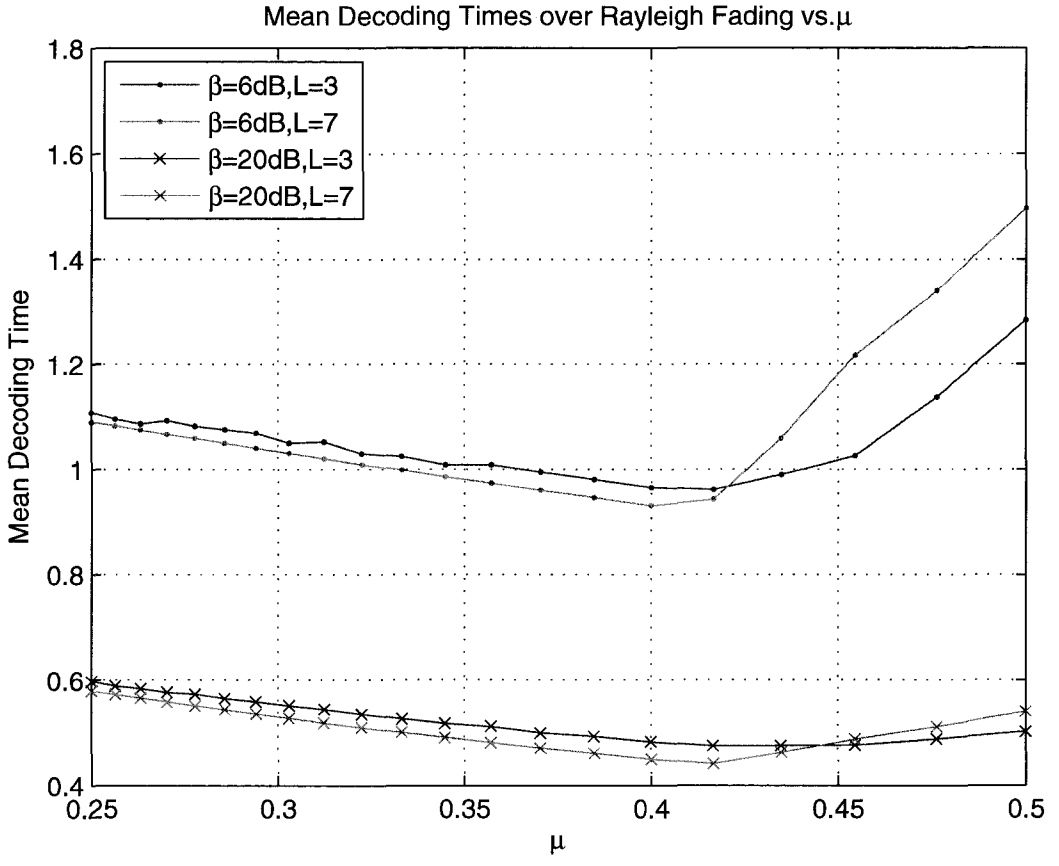

Figure 4.8: Mean decoding times vs. μ

coding is applied.

We also applied μ-PAM modulation along with rateless codes to the delay-constrained video streaming application. Using mean decoding time as the metric of comparison, we showed that μ-PAM modulation, when used with rateless coding, offers further improvements in the critical values needed for outage-free communication.

The recent revival of research activities on hybrid-ARQ schemes and the *new* ideas of rateless coding, including the work of this chapter, seem to suggest that rateless codes can be exploited as a viable solution for wireless communications. In particular, the results and analysis presented in this chapter may be potentially incorporated into link-layer considerations, as in [83,38], to allow rateless codes to be more readily integrated into practical communication systems.

4.A Proof of Lemma 4

Proof: Consider the vector-valued process containing $n_l(\zeta)$:

$$\left(\begin{array}{c} n_1(\zeta) \\ H_{A_1(\zeta)} \end{array}\right), \left(\begin{array}{c} n_2(\zeta) \\ H_{A_2(\zeta)} \end{array}\right), \ldots, \left(\begin{array}{c} n_l(\zeta) \\ H_{A_l(\zeta)} \end{array}\right), \ldots.$$

This is a Markov process since, given the value of the process at time l, the process at time $l+1$ is independent of that at time $l-1$. Furthermore, since h is drawn i.i.d., the transition matrix for this Markov process is time-invariant and thus the process is homogeneous. Also, since we assume a continuous support set of $n_l(\zeta)$, the process is irreducible since all states intercommunicate. Finally, all states are non-null aperiodic since all states have a finite mean recurrence time. As a result, by standard theorems of Markov chains, this process is ergodic and has an asymptotically stationary distribution.

Now marginalizing the states of the Markov process over $H_{A_l(\zeta)}$ leaves only the $n_l(\zeta)$ variable. This process too has an asymptotically stationary distribution. Therefore $n_l(\zeta)$ is asymptotically stationary and ergodic. □

4.B Proof of Theorem 5

Proof: Define the unbounded version of the random walk S_l to be $\mathscr{S}_l(\zeta)$ as follows: Let the unbounded version of A_l to be $\mathscr{A}_l(\zeta) := \mathscr{A}_{l-1}(\zeta) + n_l(\zeta)$, and $\Delta\mathscr{A}_l(\zeta) = n_l(\zeta)$.

Then simply,

$$\mathscr{S}_l(\zeta) = \sum_{i=1}^{l}(D - \Delta\mathscr{A}_i(\zeta)) = \sum_{i=1}^{l}(D - n_l(\zeta)).$$

It is straightforward to see that $S_l(\zeta) \geq \mathscr{S}_l(\zeta)$ for all l. Define

$$\mathscr{D}_l := S_l(\zeta) - \mathscr{S}_l(\zeta).$$

This difference is strictly non-negative and monotonic non-decreasing. Furthermore, $\mathscr{D}_l$ remains constant as a function of l except those values of l for which $S_l(\zeta) = 0$.

For the case of $E[n^*(\zeta)] < D$, $E[D - \Delta\mathscr{A}_i(\zeta)] > 0$ and so by properties of random walks, (see, e.g., [31] Theorem II.8.2), , $\mathscr{S}_l(\zeta) \to \infty$ almost surely as $l \to \infty$. Since $S_l(\zeta) \geq \mathscr{S}_l(\zeta)$ then $S_l(\zeta) \to \infty$, and clearly $Pr[S_l(\zeta) > M] > 1 - \delta$.

Turning to case of $E[n^*(\zeta)] > D$, $E[D - \Delta\mathscr{A}_i(\zeta)] < 0$ and so by the same properties, , $\mathscr{S}_l(\zeta) \to -\infty$ almost surely as $l \to \infty$. Furthermore, by other properties of random walks (see, e.g., [31] Theorem II.10.1) ,

$$\max_l \mathscr{S}_l(\zeta) < M$$

for some finite M, i.e., $\mathscr{S}_l(\zeta)$ is upper bounded by some finite value.

Since these random walks are memoryless, it is possible to 'reset' the walk at any step by setting its value to the initial condition (in this case 0) without affecting the long-term behavior of the walk. For every step $L = \{1, 2, \ldots\}$, generate a 'reset' random walk $\mathscr{S}_l^L(\zeta)$ defined for $l \geq L$ and initial condition $\mathscr{S}_L^L(\zeta) = 0$. Now it can be seen that for each l,

$$\max_{i=\{1,2,\ldots,l\}} \mathscr{S}_l^i(\zeta) = S_l(\zeta).$$

However every walk $\mathscr{S}_l^L(\zeta)$ has a finite upper bound since each has the same long-term behavior as $\mathscr{S}_l(\zeta)$. Therefore

$$\max_l S_l(\zeta) < M$$

almost surely for some finite M, and thus $Pr[S_l(\zeta) < M] > 1 - \delta$.

□

4.C Proof of Corollary 2

Proof: For the case of $E[n^*(\zeta)] < D$, from Theorem 5 it is clear that for sufficiently large l there exists an infinite amount of extra time, and therefore probability of error for

any codeword is zero. Averaging over a sufficiently large number of codewords ensures the results stated in the corollary.

For the case of $E[n^*(\zeta)] > D$, the extra decoding time remains finite, and no asymptotic limits can ensure an arbitrarily small probability of error. Therefore there exists some small $\alpha > 0$ such that $P_{\text{out}}^L > \alpha$, proving the corollary.

□

Chapter 5

Rateless Coding for Relay Networks

We have seen that rateless codes and μ-PAM modulation offer some benefits when communicating over wireless channels with unknown state at the transmitter. In terms of time to decode, throughput and probability of outage, we have shown in Chapter 4 that rateless systems offer a greater degree of robustness. In wireless systems, robustness is often obtained through diversity methods and the application of multiple antenna principles, and this is indeed the natural progression for rateless systems for wireless communications. Among the myriad diversity strategies that exist for wireless systems, relaying remains one which is actively researched and contains many practical advantages and so is the focus of this chapter.

As an extension of the results obtained so far, we present a framework for coding over relay channels using joint rateless codes and modulation. This framework is the intersection of two active areas of research in communications, namely relay networks and rateless coding. We demonstrate that there is a very natural and useful fit between these two areas, and describe some design challenges and implementation considerations for this framework.

The use of relays in wireless communication networks provide a new dimension to the design space of wireless networks which promises enhancements to both the coverage and throughput of the network. In its simplest form, a relay network is a collection of terminals which are able to transmit, receive, and possibly assist the reliable delivery of information from source terminals to destination terminals. Thus, communication of data through a wireless relay network is not required to be direct; it may pass through a number of other terminals, though direct communication from source to destination is not precluded. In fact, it is possible to simultaneously use single-hop, i.e. direct, and

multi-hop communications paths.

A question then arises: How does one code for and coordinate the various transmissions that various relays may make? This becomes a particularly difficult challenge when channel information is unavailable at the transmitting terminal, as is typically the case with time-varying, wireless channels.

As will be explained in detail in the following sections, the use of rateless codes in this setting is a promising strategy; it provides some answers to the question above and overcomes many of the problems that typically arise in relay networks. In general, a rateless code is a code that has a rate determined by the number of transmitted symbols required before the decoder is able to decode. The rate then is not known *a priori* as it is in typical fixed-rate block codes. Existing rateless codes, namely classes of fountain codes, exhibit the property of naturally adapting to channel conditions without requiring channel knowledge at the transmitter. This alone suggests their usefulness in a relay network, and is explored in detail in this chapter.

The remainder of the chapter is organized as follows: Section 5.1 provides a background and survey of relay channels and relay networks. Section 5.2 presents the system model utilizing rateless codes for relay channels. An achievable rate region for the system we consider is provided and analyzed. Section 5.3 provides results of a Monte Carlo simulation of a simple relay system, demonstrating that gains in throughput and outage probability can be achieved simultaneously. Extensions and practical implementation considerations are also discussed. We conclude the chapter in Section 5.4 with some final remarks.

5.1 Background

5.1.1 Relay Channels and Relay Networks

Traditional wireless networks have predominantly used direct point-to-point or point-to-multipoint (e.g. cellular) topologies. The fundamentally different mode of transport, possible uncertainty in terminal geographical location, and difficulty in theoretical analysis of relay networks have kept them mainly to academic realms. Occasional attempts by industry to reap the benefits of relaying have been made, such as those by Ricochet Networks in the late 1990's, but have met with limited success[1]. However, there has

[1]Riccochet Networks is still providing service based on a relay network topology. Other vendors, such as Bel-Air Networks, are also introducing similar relay-based networks.

been a number of recent theoretical results that may spur the use of relaying techniques in practical networks. The key behind these advances are mainly a result of research in multiple antenna systems.

Multiple antenna systems, or multiple-input, multiple-output (MIMO) systems have seen remarkable growth in recent years and can deliver significant throughput and coverage gains over wireless channels compared with single antenna systems. Wireless standards such as the IEEE 802.16e (WiMAX) [78] and IEEE 802.11n [75] depend heavily on MIMO principles to achieve promised throughput and reliability targets that are being demanded in the marketplace.

These systems, by using multiple antennas at the source and destination nodes, are able to achieve spatial diversity in addition to temporal diversity that traditional coding provides, as introduced in [72]. It is also possible to exploit the multiplicity of spatial channels in MIMO systems to increase the throughput beyond what would be possible in single antenna systems.

A fundamental observation to make with respect to wireless relay networks is that, with appropriate coordination or cooperation, communication between two terminals in the network can be viewed as a type of MIMO system. This may be achieved in a number of different ways, however all require some level of cooperation within the network. The relaying strategy used may impose limits on the similarity to a MIMO system. Also, constraints on the system – such as the half-duplex constraint, i.e., no terminal may receive and transmit simultaneously – may limit the degree to which communication through a relay network behaves like a MIMO system.

There are three "classic" relaying strategies that are commonly considered: *amplify-and-forward* (AF), *decode-and-forward* (DF) and *compress-and-forward* (CF). In AF, a relay is a repeater, amplifying and transmitting the received signal. In DF, the relay attempts to decode received signals. If successful, it re-encodes the information and transmits it. Finally, CF attempts to generate an estimate of the received signal. This is then compressed, encoded and transmitted in the hope that the estimate may assist in decoding the original codeword at the destination.

In their two-part paper [63,64], Sendonaris, Erkip and Aazhang introduce and examine the concept of user cooperation diversity. Here, the authors demonstrate that simple cooperation between transmitting users can increase throughput and coverage simultaneously. The implemented strategy uses a pair of transmitting, full-duplex users who cooperate in sending independent data from both users to a common destination. This is accomplished by a multi-period transmission process. Initially, each user sends its own

data, while listening to the other users' transmission. After some time, each user will allocate some amount of power to send an estimate of what it received from the other. In essence, each user is acting as a relay for the other and using the AF relaying strategy. This approach demonstrates some of the gains that may be had through cooperation, though it relies on some channel information at the transmitter.

The DF and CF strategies are thoroughly examined for wireless channels in [39]. In addition to providing a thorough survey of relay networks, they show, under certain conditions, that the DF strategy is capable of achieving rates up to the ergodic capacity of the channel. Furthering these results, Lai, Liu and Gamal in [42] combine the use of DF and CF to achieve the same bounds, but with fewer restrictions on the system.

Dohler *et al* introduce "virtual antenna arrays" in [19]. Here, groups of terminals cooperate to form a virtual MIMO system and exploit the spatial diversity that results. This is a similar concept to user cooperation, but focuses on different design aspects, such as link budget impact. Further work done in the area of signal design for relay networks is by Nabar, Bolcskei and Kneubuhler in [55], where the authors consider code design aspects of DF and AF strategies.

A natural extension of the basic implementation described in [63] is to use coded cooperation and this is described in [36] by Hunter and Nosratinia with further analysis and implementation details for wireless channels given in [37]. Further approaches using coded cooperation are given in [44, 43], where the authors propose a DF-based scheme with many opportunistically cooperating terminals, and show that diversity gains scale in the number of potential relays rather than the actual number of participating relays.

As the concept of coded cooperation has grown, implementations of the concept have begun to appear. In [88], Zhang, Bahceci and Duman present a strategy based on turbo-codes for communication on relay channels. There, they design the code and iterative decoder, and show that the performance can be close to an achievable rate bound. Hu and Duman apply a similar concept in [33] using LDPC codes, and these also show very efficient performance.

In works that foreshadow the use of rateless codes for relay channels, Caire and Tuninetti in [6], and Zhao and Valenti in [89] propose and analyze the use of Hybrid-ARQ for relay channels. In [6], hybrid-ARQ protocols for the Gaussian collision channel is studied. Notably, the results translate to relay channels. This is noted in [89], where the authors propose the use of a hybrid-ARQ-type protocol for relay networks using orthogonal signaling slots with the half-duplex constraint. They demonstrate that their protocol provides significant improvements in throughput and average transmission delay.

Also foreshadowing the application of rateless codes, Mitran, Ochiai and Tarokh present in [51] a two-phase communication scheme for wireless devices in a network with the half-duplex constraint along with an information-theoretic performance analysis. The two phases are the *listening* phase, in which the source node broadcasts and other nodes listen, and the *collaboration* phase, in which multiple nodes cooperate to transmit to the destination. It is assumed that the channel state information is not available at the transmitters but is available at the receivers. The results suggest that such a collaborative communication scheme can lead to significant diversity enhancement compared to direct communication between source and destination.

From an information theoretic perspective, the work by Cover and El Gamal in [16] remains the foundational treatment on the relay channel. There, the authors present a number of theorems for this channel under different conditions. Unfortunately, the capacity is not solved for the general case; only an achievable rate is provided. Capacity results are provided by the authors only for certain degraded channels. A partial converse to the general case was communicated in [87], though the full solution to this problem remains open.

As the idea of relay networks attracted attention, researchers began to investigate their information theoretic aspects. In particular, determining achievable rate regions was investigated under a number of different assumptions. As a fairly general case, Gupta and Kumar in [30, 29] present an achievable rate region for arbitrary relay networks. Further results are available in [42, 84, 41, 27] under different network assumptions and topologies.

More recently, attention has turned to generalizations of relay networks, particularly the MIMO relay network. In this area, Wang *et al* in [81] present bounds on the capacity of MIMO relay channels. This is furthered by Tang *et al* in [70].

As the connections between relay channels and MIMO systems matured, it became clear that the fundamental properties of MIMO channels are also applicable to the relay case. The well-known diversity-multiplexing trade-off (DMT) presented by Zheng and Tse in [90] was extended to the relay case by Yuksel and Erkip in [85]. When constrained to half-duplex channels, Azarian, El Gamal and Schniter demonstrated in [3] a novel AF scheme that achieves a DMT bound for AF strategies, and provides insight into some of the obstacles facing practical relay network implementations. Finally, [86] demonstrates that he CF strategy is the only known strategy capable of achieving the full duplex relay DMT.

Given these properties, the use of rateless codes for wireless relay channels seems

to be a synergistic match. Building on the ideas developed here which are published in [8, 13, 12], Liu in [47] presents a rateless code based protocol that can achieve even better efficiencies combining DF and CF concepts. Another approach, building on the work presented in this chapter and published in [8] is described by Molisch, Mehta, Yedidia and Zhang in [52, 53]. Focusing on multi-relay networks, the authors present two protocols utilizing rateless coding for a large numbers of potential relays. The first protocol requires some fixed number of relays to receive the source message in the listening phase before the collaboration begins. Practical implementation issues relating to the coordination of relays are addressed by the authors, and simulations of the proposed protocols are presented demonstrating the benefits of the approach, including an analysis of the expected transmission energy.

5.1.2 Motivation for Rateless-coded Relay Networks

Approaching the communication problem from a coding-theoretic perspective, one seeks to design a practical coding framework that effectively implements a collaboration strategy to maximize achievable rate and minimize outage probability. [2]

It appears that no *fixed-rate* coding system is capable of driving the outage probability to zero without channel state information at the source. Also, unless operating at a low efficiency (i.e. low rate), no fixed-rate coding system is robust to the variation of channel statistics typical in a wireless setting. With the use of a feedback channel it is possible to provide channel information to the transmitter and thereby overcome many of these limitations. However, for large relay networks this overhead becomes significant compared to the small number of bits needed for acknowledgments in a rateless system, particularly for time-varying channels or moving terminals. Further analysis comparing fixed-rate and rateless approaches in the relay network setting are given in [52]. In particular, the benefits of a rateless system for large relay networks are highlighted.

Usual solutions to this problem have been to utilize ARQ or hybrid-ARQ methods such as described in [89]. Like rateless coding, feedback channels are required to signal successful reception of codewords at the destination. In fact, rateless coding may be considered a form of "continuous incremental redundancy", as described in Chapter 4, compared to the block-based incremental redundancy provided by hybrid-ARQ.

The use of rateless codes for relay networks naturally accounts for communication

[2] One may also, depending on the ultimate application, wish to minimize transmission latency, transmit power, etc.

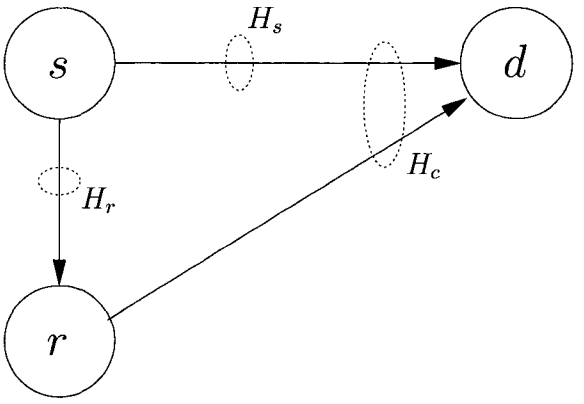

Figure 5.1: Three node wireless relay network

efficiency and system robustness simultaneously. We review some theoretical bounds resulting from this framework and show the diversity and throughput gains that may be had. We present a simulated implementation of this framework based on fountain codes as a tutorial to demonstrate these advantages. Finally, we discuss a number of practical issues and limitations of this system, suggesting alternatives, such as those provided in [52].

5.2 Rateless Relay System Model

To demonstrate the usefulness of rateless codes in relay networks, we will focus on the base system model presented in [51] and extended for rateless codes in [8].

Consider the system shown in Figure 5.1 with three wireless devices. This configuration forms the building block for all relay networks and so is useful to first understand before considering general networks. The source s wishes to communicate with the destination d, possibly with the help of the third device, relay r. We consider the half-duplex scenario, where the relay does not receive from the source and transmit to the destination at the same time. We assume that each transmitted symbol from a source antenna or from a relay antenna has the same average energy E_s, though in general this need not be the case. Let $X[i]$ and $U[i]$ be respectively the symbol vectors at time i transmitted from the source antennas and the relay antennas. Consider the quasi-static Rayleigh fading model and use H_r, H_s and H_c to denote the channel gain matrices for the source-to-relay channel, source-to-destination channel, and the compound channel from the combined antennas of source and relay to the antennas of the destination. Let $Y[i]$ and $Z[i]$ be the

received signals at time instant i at the relay and destination respectively. During each codeword transmission,

$$\begin{aligned} Y[i] &= H_r X[i] + N_Y[i], \\ Z[i] &= H_c[X[i]^T \; U[i]^T]^T + N_Z[i], \end{aligned}$$

where N_Y and N_Z are zero-mean (vector-valued) white complex Gaussian noise processes received at the relay and the destination respectively. We note that H_s is a sub-matrix of H_c and if the relay does not assist the source at time instant i, the received signal $Z[i]$ reduces to

$$Z[i] = H_s X[i] + N_Z[i].$$

We assume that the entries of H_r and H_c are drawn i.i.d. respectively from complex Gaussian distributions with variances G and 1. This restricts the source-to-destination channel and relay-to-destination channel to have the same fading statistics. Such a restriction plays no essential role in the applicability of this work, and merely serves as a simplification assumption, also making these results comparable with those of [51]. The values H_r and H_c are assumed to be *known* at the respective receivers but *unknown* at each transmitter. The variance of N_Y and N_Z are both $N_0/2$, also known at the receivers. It is also assumed that the relay is capable of synchronizing with the source at the symbol level. Synchronization issues will be discussed in Section 5.3.4.

For such a setup, an information-theoretic analysis of block-coded schemes is presented in [51], which may be summarized as follows. The source selects a rate R to transmit a block of k-bits of information using a block code of length $n := k/R$. The relay, aware of the channel state H_r, decides a number $f \in (0, 1]$ such that after listening for $n_1 := fk/R$ symbols, the relay is able to decode the information block and collaboratively transmits to the destination.

The case that $f = 0$ is operationally equivalent to the situation where the relay has the knowledge of source codeword *non-causally*, and therefore equivalent to a MIMO system where the antennas of the source and those of the relay jointly form the set of transmit antennas. On the other hand, the case that $f = 1$, is equivalent to the situation where the relay is not able to decode before the destination and thus does not collaborate.

Under this scheme, Theorem 1 of [51] shows that given a channel realization (H_r, H_c), and given a choice of f, any rate R satisfying

$$R < fC(H_s, \gamma) + (1 - f)C(H_c, \gamma) \tag{5.1}$$

and

$$R < fC(H_r, \gamma) \tag{5.2}$$

simultaneously, or satisfying

$$R < C(H_s, \gamma) \tag{5.3}$$

is achievable, where $C(H, \gamma) := \log_2 \det(I + \gamma H H^T)$ is the MIMO capacity formula under equal power allocation across antennas, and γ in our setup equals E_s/N_0. Since H_r and H_c are both random variables, the authors of [51] argue that for any fixed R and any f, the outage probability may be defined as the probability that R is outside the interval prescribed by (5.1), (5.2) and (5.3), under the channel law for H_r and H_c. Furthermore, they show that there exists an optimal choice $\hat{f}$ of f — in the sense of *minimizing the outage probability for a fixed rate* R. Such a choice can also be seen as improving the diversity order, as shown in [3].

This fixed-rate strategy is challenged by practical issues. Without channel knowledge, as long as the source selects a transmission rate R, there is no opportunity for the relay to improve the rate, even when the channel supports much higher rates. The only benefit that the relay offers is a decreased outage probability. It is also notable that no matter what rate is chosen, the outage probability is bounded above zero for quasi-static or slow fading channels. With the use of feedback in the system, it is possible to signal a suitable R to the source, though knowledge and coordination of relays remains a difficult obstacle. In fact, the optimal choice of rate requires that the source be aware of the existence and participation of the relay as well as channel knowledge. Furthermore, such a choice is fragile to the variation of channel conditions and relay availability.

To handle efficiency and robustness simultaneously, rateless codes may be used in place of the fixed-rate codes. The source encodes a block of k information bits using a rateless code with parameter k and sequentially broadcasts the codeword symbols to both the destination and the relay. The relay attempts to decode the information until it succeeds. At this time, if the destination has not decoded the information, the relay then collaborates with the source by transmitting to the destination using another rateless code. Starting from time instant 1, the destination also attempts to decode the source information, and whenever it can decode, it sends an `ACK` back to the source and the relay in order to terminate the current transmission. Here we note that the one-bit feedback for acknowledgment entails little implementation difficulty as long as the feedback channel exists. Note that it is straightforward to generalize this rateless coding concept to multiple relay, multiple antenna systems and to other channel models.

In this strategy, the source does not need to be aware that a relay exists. The rateless nature of the coding scheme allows the source to communicate with the destination at a rate adapted to the channel conditions, to the availability of the relay, and to the collaborating strategy of the relay. Furthermore, although we are dealing with fading channels, the outage probability can be made arbitrarily small. This is because the successful decoding almost surely occurs at a time corresponding to a rate supported by the channel.

A fundamental question then arises: what rates are achievable with the presented scheme? Here, we present an achievable rate result built on the work of [51].

Let n be the time needed for the destination to decode a message. Similarly, we denote by n_1 the time needed for the relay to decode a message. We define the *realized rate R of a transmission* by $R := k/n$ bits/channel use. Notice that both n_1 and n are random variables depending on channel realization (H_r, H_c) and noise realizations. Therefore the realized rate R is also a random variable. We say that a rate R is *achievable* for a given channel realization (H_r, H_c) if there exists a family of rateless codes (each parametrized by a different k) such that after k/R channel uses, the decoding error at the destination can be made arbitrarily small for sufficiently large k.

We now present a theorem giving the optimal value of f in the sense of maximizing the achievable rate of the relay channel.

Theorem 6 *Let*

$$\tilde{f} := \begin{cases} \frac{C(H_c,\gamma)}{C(H_c,\gamma)+C(H_r,\gamma)-C(H_s,\gamma)} & \text{if } C(H_r,\gamma)>C(H_s,\gamma) \\ 1 & \text{otherwise} \end{cases} \tag{5.4}$$

and let

$$\tilde{R} := \tilde{f}C(H_s,\gamma) + (1-\tilde{f})C(H_c,\gamma). \tag{5.5}$$

Then for any $\delta > 0$, there exists a block coding scheme at rate $\tilde{R} - \delta$ such that with increasing block length, the decoding error probability is driven arbitrarily close to 0.

Proof: First consider the case when $C(H_r,\gamma) \leq C(H_s,\gamma)$. Then $\tilde{f} = 1$ and $\tilde{R} = C(H_s)$. By (5.3) or a standard MIMO capacity argument, rate $\tilde{R} - \delta$ is achievable with arbitrarily small decoding error for sufficiently large block length. Now consider the case when $C(H_r,\gamma) > C(H_s,\gamma)$. Referring to Figure 5.2, by noting that $C(H_c,\gamma) > C(H_s,\gamma)$, it can be verified that the supremum $\tilde{R}$ of R is defined by the intersection of boundary of constraint (5.1) and that of constraint (5.2). Therefore, we can solve for the intersecting point by solving the two linear equations. This gives $\tilde{f}$ and $\tilde{R}$ in the theorem. With

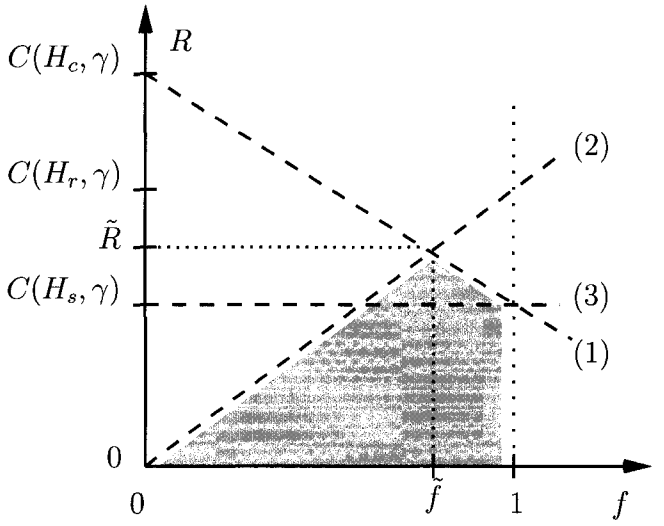

Figure 5.2: Sketch of the relationship between (5.1), (5.2) and (5.3) for $C(H_r,\gamma) > C(H_s,\gamma)$.

the standard information theoretic argument, for any rate $\tilde{R}-\delta$ and the corresponding $\tilde{f}$, the decoding error probability at the destination can be made arbitrarily small for sufficiently large block length. □

Some insights may be obtained by distinguishing $\tilde{f}$ with $\hat{f}$. Given a rate R, $\hat{f}$ minimizes outage probability resulting from channel uncertainty, and such minima are strictly bounded above zero. However, knowing channel realizations at the source, $\tilde{f}$ maximizes the achievable rate dictated by (5.1), (5.2) and (5.3) and allows virtually zero outage probability.

Although this achievable rate is only shown to be valid for block coding schemes and for the case in which the channel is *known* at the transmitters, we demonstrate in the following section that rateless codes can also achieve this rate *without* channel knowledge at the transmitter.

5.3 Rateless Code Implementation

The presented rateless coding scheme may be implemented with fountain codes. We implement such a scheme given in Figure 5.3 for illustrative purposes using a Raptor code as an outer code concatenated with a space-time inner code to form a rateless code. The presented system makes simplifying assumptions that real systems need to address.

These are discussed below in Section 5.3.4.

This architecture is tested over a simplified channel model, where the source, relay and destination each have only one antenna, and are time-synchronized at the symbol level. The channel is defined in Section 5.2, which is a quasi-static fading model such that, for the duration of the transmission from the source, all channel realizations are constant. Each new codeword transmission by the source uses independently drawn realizations for each of the three channels. The relative gain between the source to relay channel and the source to destination channel is denoted by G, i.e., $G = \frac{H_r}{H_s}$.

A Raptor code with $k = 9,500$ is used as the inner code both at the source and the relay. The LDPC component code of the Raptor code has rate 0.95. We use the degree distributions as in [66] for the inner (LDPC) and outer (LT) components of the Raptor code. The performance of Raptor codes with these parameters has been reported in [59] for several lossy channel models including the AWGN channel. Consecutive output symbols from the Raptor code are then QPSK modulated, chosen for simplicity.

The rateless code used by the relay is the same Raptor code also with QPSK modulation. When in the collaboration phase, the relay aligns its output symbols in time to correspond to the source's output symbols.

The output symbols of the source and relay are input to the distributed space-time inner code. The space-time inner code uses the Alamouti scheme [2] and works as follows. During both the listening and collaboration phases, the source acts as the first antenna in an Alamouti system. The relay, once in the collaboration phase, acts as the secondary antenna in which consecutive pairs of input symbols are transformed according to the Alamouti scheme.

This distributed space-time code is received and decoded at the destination. During the listening phase the destination receives symbols only from the source, and during the collaboration phase the destination performs standard Alamouti decoding using the received symbols from both source and destination.

For each transmitted codeword, belief propagation decoding is attempted periodically at intervals of 100 channel uses. For each attempt, the initial messages for belief propagation are calculated based on the received signal and the receiver's (perfect) knowledge of channel gains and noise variance. During each iteration of decoding, the decoder examines whether hard decisions on the messages form a codeword. When this occurs, the transmission of the codeword is terminated. If the hard decisions do not form a codeword within 100 iterations, the current decoding attempt is stopped and the decoder waits for the next decoding attempt. For the purposes of the simulation we assume that

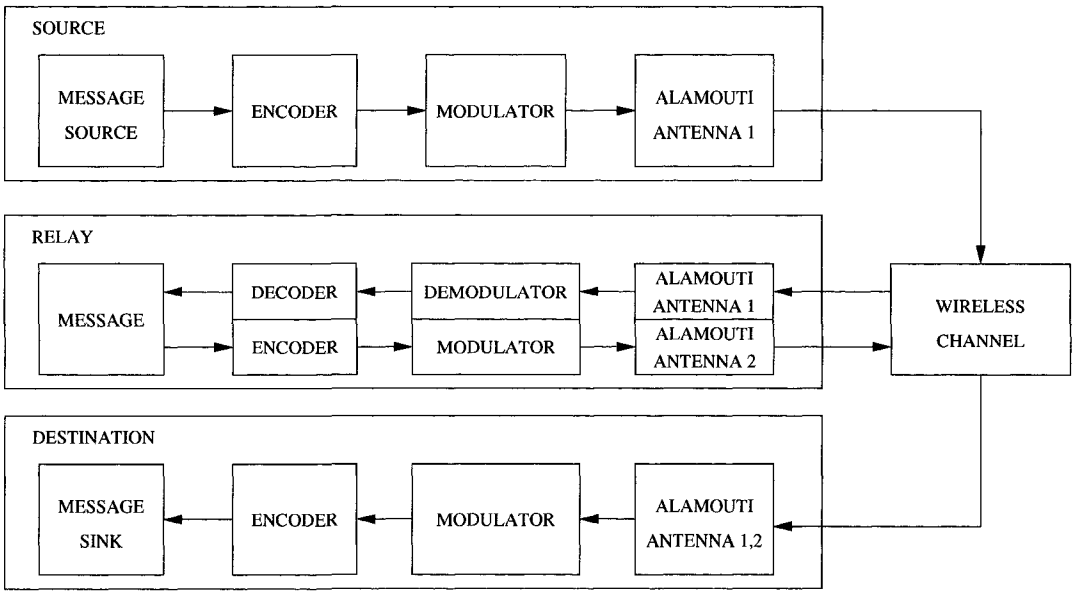

Figure 5.3: System diagram for simulated relay framework

the receiver knows whether it decodes correctly. In practice, this may correspond to the case where CRC bits are used in the message. [3] Other rules for stopping a decoding attempt and for declaring a decoding success (such as those based on the reliability of the decoded word) are also possible.

5.3.1 Simulation Results

We perform a Monte Carlo simulation of the presented rateless code implementation over a range of SNR from -5dB to 15dB. For the purpose of analyzing our results, we denote for each codeword transmission the decoding times at the relay and destination by n_1 and n respectively. The realized fraction of time that the relay was listening is defined as $f := \min\{1, n_1/n\}$. The maximum achievable realized rate for a given realized f is defined as

$$R^* = fC(H_s, \gamma) + (1-f)C(H_c, \gamma). \tag{5.6}$$

For any given f, R^* indicates the maximal achievable rate prescribed by (5.1), (5.2) and (5.3). Then the gap between $\tilde{R}$ and R^* indicates the suboptimality of *realized* f in comparison to the optimal choice $\tilde{f}$, for a given channel realization. To an extent, this gap also reflects the spectral efficiency limitation associated with the modulation

[3] We note that the entailed rate loss is negligible for large k.

scheme. The gap between R^* and R, the realized rate, indicates the coding loss due to the suboptimality of Fountain code.

Following [51] and since the entries of H_c have unit variance, the average receive SNR at the destination can be defined by

$$\text{SNR} := (2 - E[f])\gamma, \tag{5.7}$$

and similarly, the average optimal SNR at the destination associated with the optimal choice $\tilde{f}$ can be defined as

$$\widetilde{\text{SNR}} := (2 - E[\tilde{f}])\gamma. \tag{5.8}$$

Figure 5.4 presents the curves of R versus SNR, R^* versus SNR, and $\tilde{R}$ versus $\widetilde{\text{SNR}}$ averaged over the ensemble of channel realizations, as well as the standard 2×1 MIMO capacity bound as a function of receive SNR, for $G = -5$dB. First notice from the figure that R^* converges with $\tilde{R}$ at low SNR regime, which indicates that the realized f is nearly optimal. As SNR increases, the R^* curve reaches the asymptote of 2 bits/channel use, which is governed by the QPSK modulation scheme. There is an approximately constant fraction of rate loss across all SNR when comparing R with R^*. This is better visualized in Figure 5.5, where the same performance metrics unfold along the G dimension at $\gamma = -5$dB. Combining these two figures one can conclude that there is only about 10%-15% rate loss due to code suboptimality. This is in fact consistent with the results of [59] for Gaussian channels. Clearly, at high SNR, both R^* and R can be increased by using higher-order modulations. The 2×1 and 1×1 MIMO capacity values are also given in Figure 5.5, and bound $\tilde{R}$.

Although outage is not a concern with rateless coding schemes, we still plot a curve of outage probability to make comparison with the theoretical limits presented in [51] for fixed-rate codes. Here the natural definition of outage probability for the rateless code at a given rate α is the relative frequency at which the decoder can *not* decode the message before time k/α. Figure 5.6 contains outage probability curves for our rateless code, the theoretical limits of [51], and the standard theoretical limits for 1×2 and 1×1 MIMO system for $R = 0.5$ and some values of G. From these plots, it can be seen that our results are only about 1 dB away from the fixed-rate code outage limits of [51], achieve the same order of diversity, and significantly improve upon a 1×1 MIMO system, particularly at high SNR.

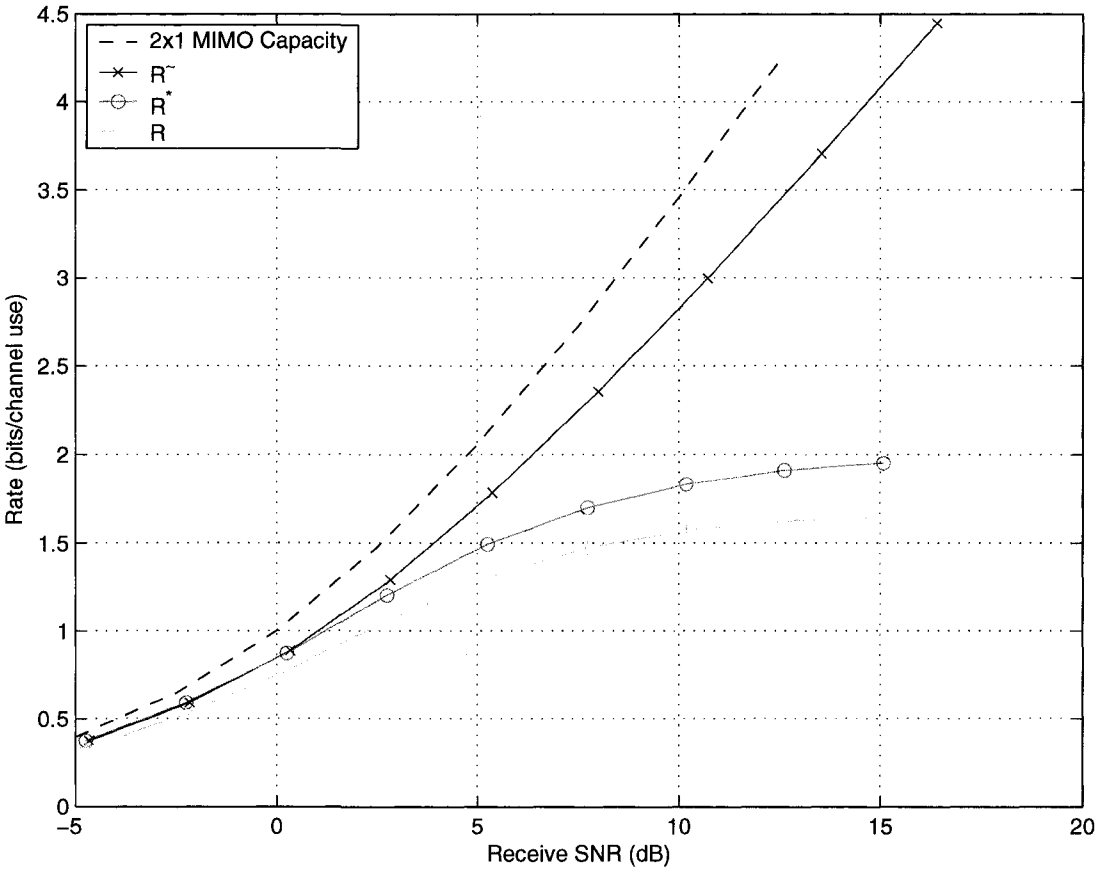

Figure 5.4: Mean system rates vs. average receive SNR, $G = -5$dB

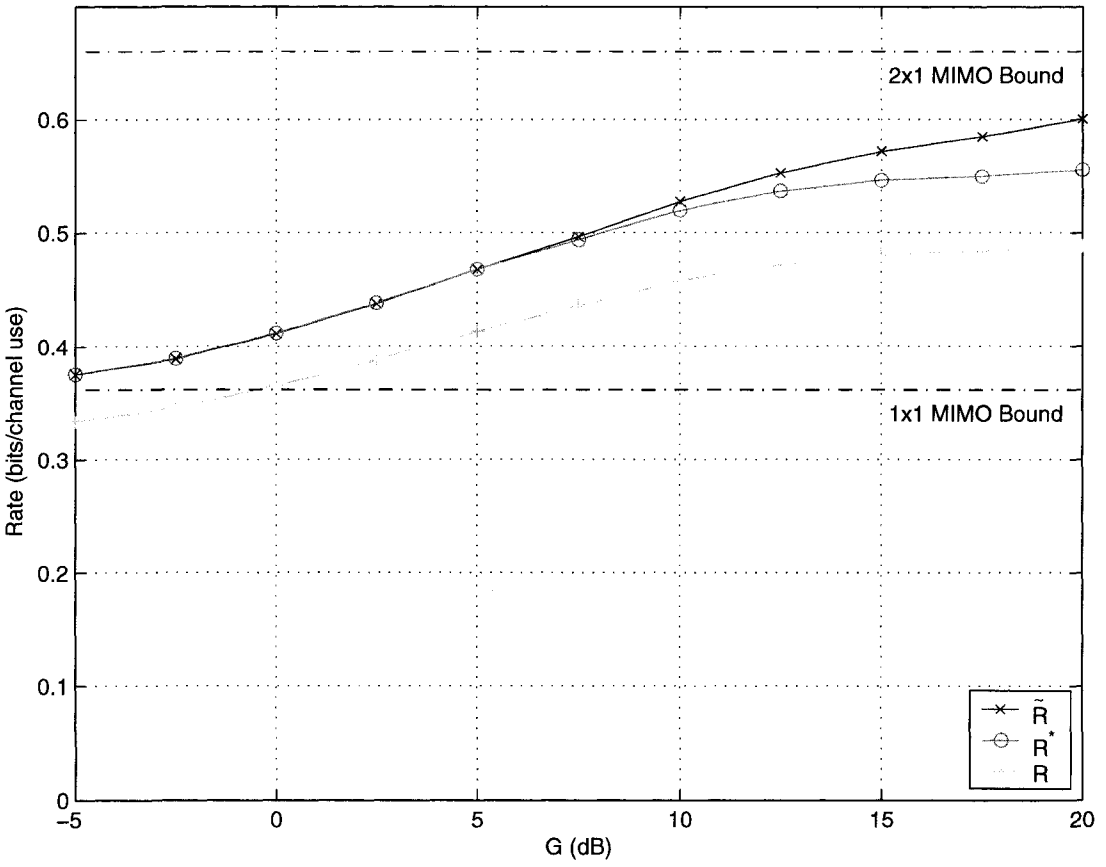

Figure 5.5: Mean system rates vs. G, $\gamma = -5$dB

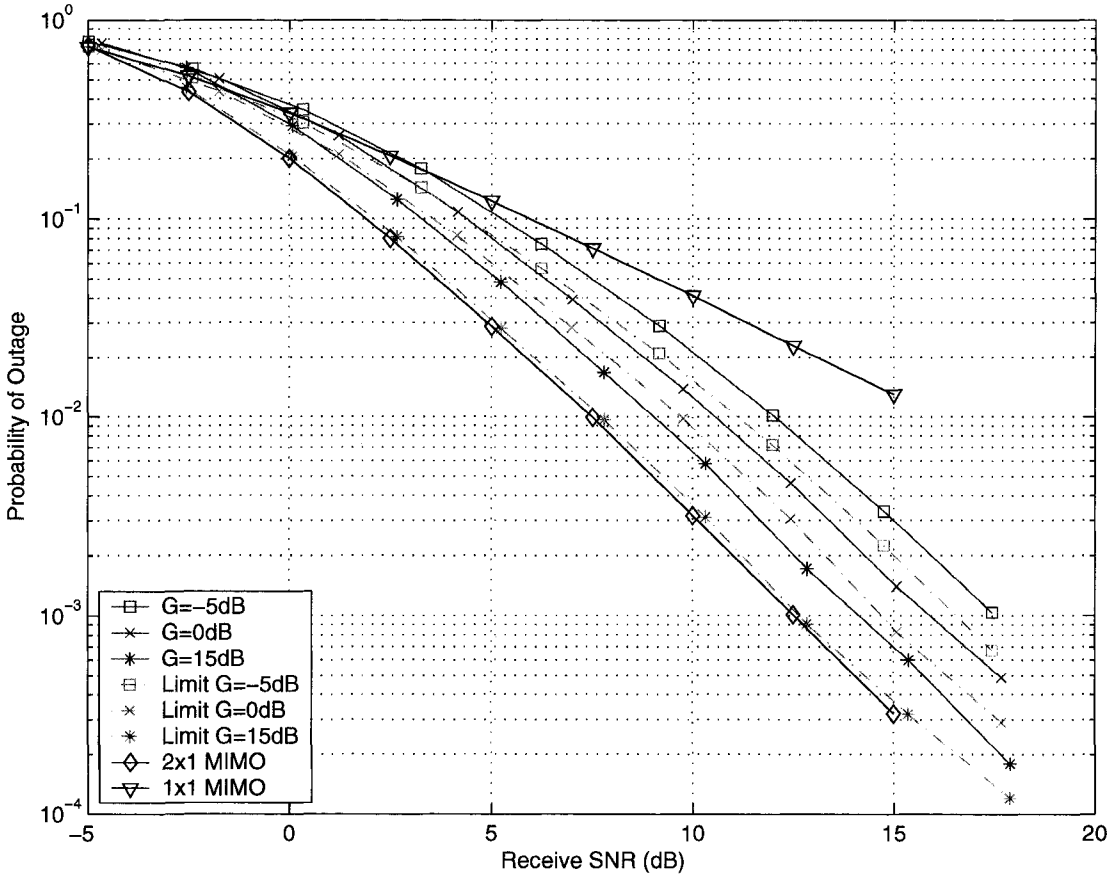

Figure 5.6: Probability of outage vs. average receive SNR, $R = 0.5$

5.3.2 Modulation Considerations

The simulation results presented in the previous section used QPSK as the modulation scheme and was thus limited to a maximum communication rate of 2 bits per channel use. QPSK modulation was chosen for simplicity of simulation. Given the relatively low complexity modulation scheme of μ-PAM introduced in Chapter 3, we investigate its use in the relay setting considered in this chapter.

To compare the performance of M-PAM and μ-PAM we take two approaches. First, we build on the results of Chapter 3 and the numerical analysis used there to find numerical estimates for the achievable rates of the 3-node relay network. Second, we extend the simulation results of the previous section to include μ-PAM.

To investigate the achievable rates for the relay channel considered in this chapter and compare different modulation methods, we assume that there is a capacity achieving rateless code in the system. We assume quasi-static fading with coefficients for the three channels in the system to be drawn independently from a Rayleigh distribution. The mean SNR of the relay-destination and source-destination are set to be equal, and the relative gain of the mean SNR of the source-relay channel is denoted by G. As in Chapter 3 we determine the mean rate by taking the expectation over Rayleigh distribution, except here we use $\tilde{R}$ as the function we evaluate.

Figure 5.7 presents and compares the mean achievable rates as a function of average receive SNR for $L = 2$ (recall $L = \log_2 M$), M-PAM and μ-PAM with $\mu = 2/5, 1/3$. Here $G = 15$dB, a situation in which it is expected that the relay will be of significant benefit on average. Also presented is the no-relay case assuming M-PAM. Clearly the presence of the relay in this configuration makes a large difference in performance. With the presence of the relay, the three modulation settings achieve similar rates, with the best setting depending on the average SNR. Similar to the results of Chapter 3, low SNR scenarios are best suited to smaller μ, and here $\mu = 1/3$ achieves the best performance in this regime. At higher SNR either $\mu = 2/5$ or M-PAM yield the best mean rates.

The same data is presented in Figure 5.8, the only difference being $L = 7$ instead of $L = 2$. Here we see the same relationships hold as in Figure 5.7, and in fact these observations are valid for any valid of L tested. It is notable that at high SNR the no relay case actually achieves higher mean rates than the $\mu = 1/3$ μ-PAM configuration. This is due to the fact that for the large value of L and small value of μ, very little energy is being dedicated to the last of the embedded bits so that, even in a high SNR condition, it is difficult to reliably determine their value.

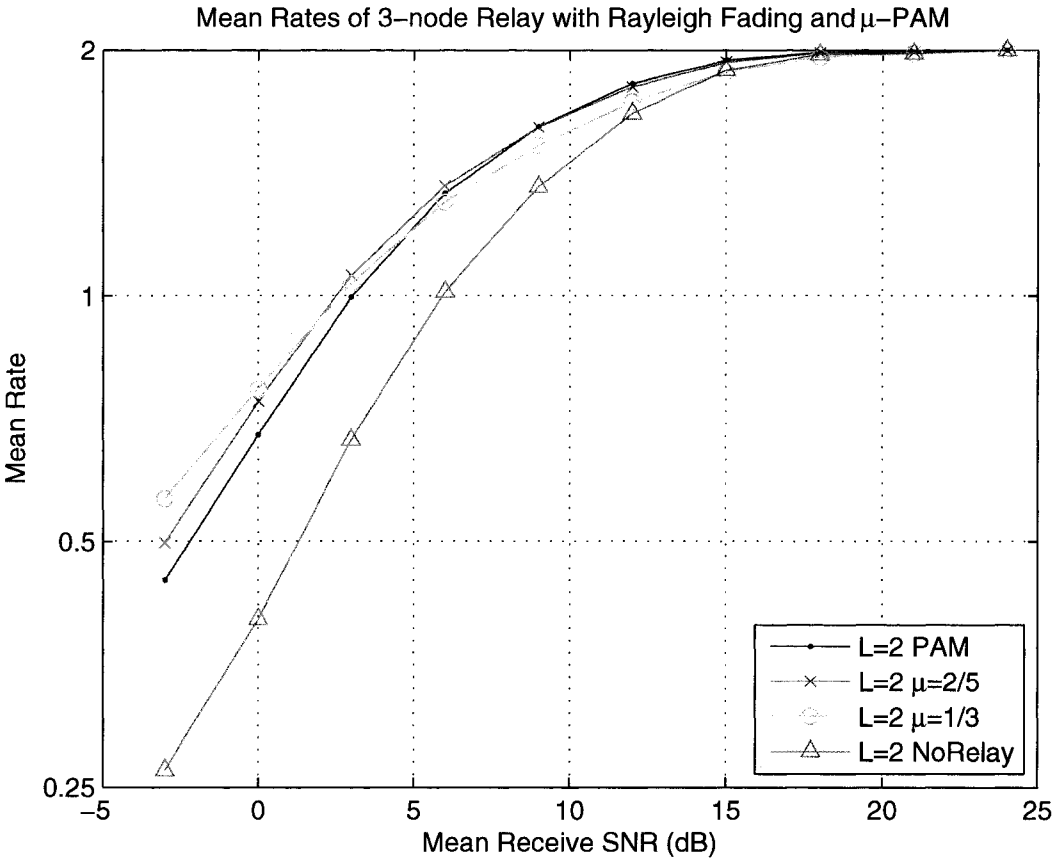

Figure 5.7: Mean system rates vs. average receive SNR, $L = 2, G = 15\text{dB}$

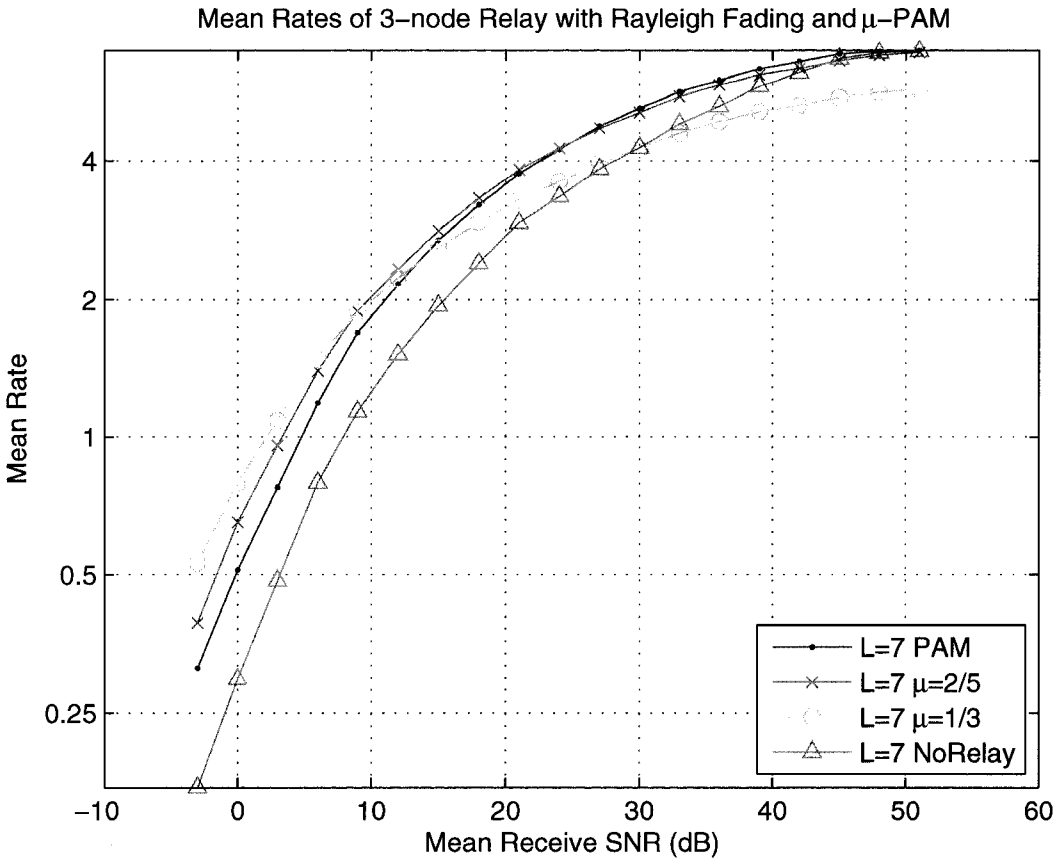

Figure 5.8: Mean system rates vs. average receive SNR, $L = 7, G = 15\text{dB}$

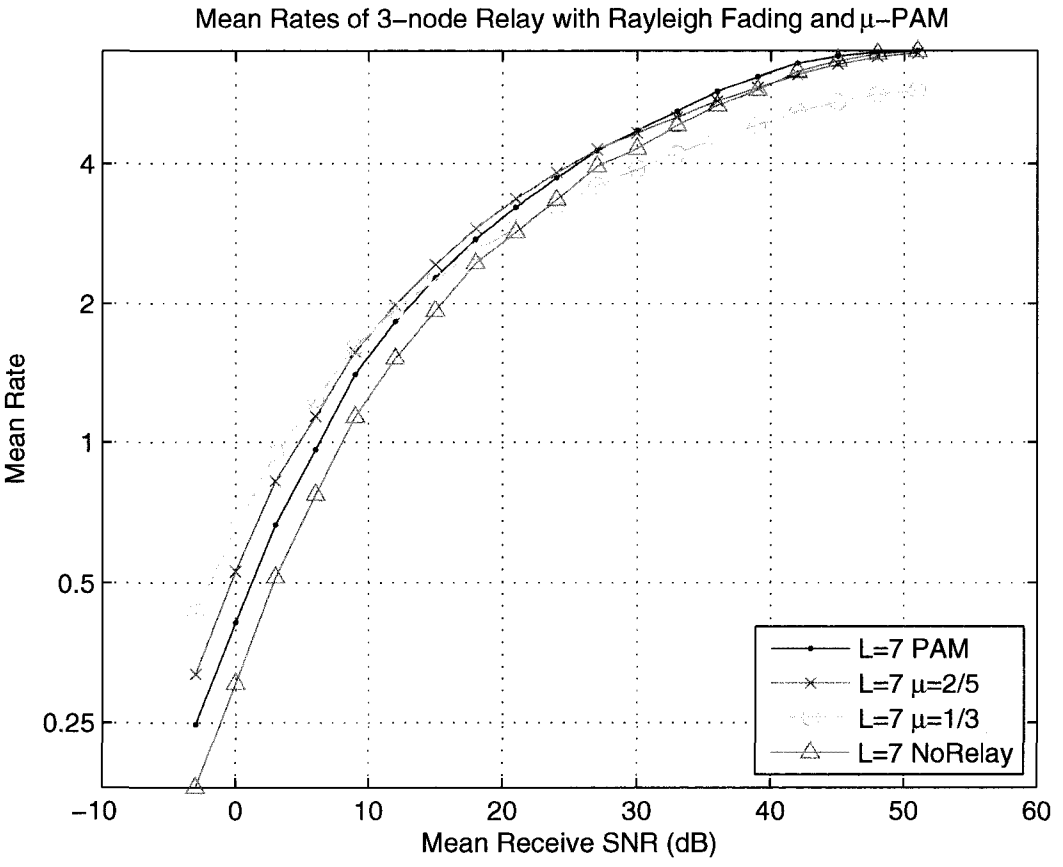

Figure 5.9: Mean system rates vs. average receive SNR, $L = 7, G = 5\text{dB}$

To understand the impact of the parameter G, the same results for $L = 7$ are presented in Figure 5.9 with $G = 5$dB. This corresponds to a situation in which, on average, the relay is beneficial, but is often unable to cooperate in the message transmission from source to destination. In general, compared to Figure 5.8, the relay-enabled rates are reduced such that they are not very much larger than the no-relay configuration.

As argued in Chapter 4, mean decoding times are a often a more meaningful measure of performance than mean rates, particularly from the perspective of the destination terminal. In the case of the relay channel, the normalized mean decoding time $\bar{n}$ is defined as

$$\bar{n} = E\left[\frac{1}{\tilde{R}}\right],$$

and is expressed in channel uses per bit.

Figure 5.10 presents the mean decoding times of the system for the three modulation configurations and $L = 2$. In a reciprocal manner to Figure 5.7, the results are intuitive. Unlike the situation when looking at mean achieved rates, the $\mu = 1/3$ μ-PAM modulation scheme provides the best performance or near-to-best performance over the entire SNR range studied. This is due to the fact that mean decoding times are typically dominated by low SNR realizations, resulting in a long decoding time, in which $\mu = 1/3$ is often the best choice. At low SNR it is notable that there is a dramatic difference in mean decoding times between the with-relay and no-relay configurations, with a maximum separation of over a factor of two.

As a comparison to the results of Figure 5.8, Figure 5.11 presents the mean decoding times of the system for the three modulation configurations with $L = 7$. Very similar results are seen here as with the $L = 2$ scenario. Here we see that the maximal difference in performance at low SNR is greater than a factor of three, suggesting that μ-PAM coupled with the rateless coding framework presented here can offer significant improvements.

With the addition of a Raptor code, we re-perform the simulations of Section 5.3.1 with μ-PAM modulation. This is done to verify overall system performance for the relay channel with both coding and modulation taken into consideration. Since the previous simulations used QPSK modulation, taking advantage of both I and Q components of a real channel, we apply M-PAM or μ-PAM independently on both I and Q channel for these simulations. The mean SNRs have been normalized to account for the extra energy in the transmitted, complex, constellations. It is assumed that no distortion takes place over the channel, and so the two components can be modulated and demodulated independently.

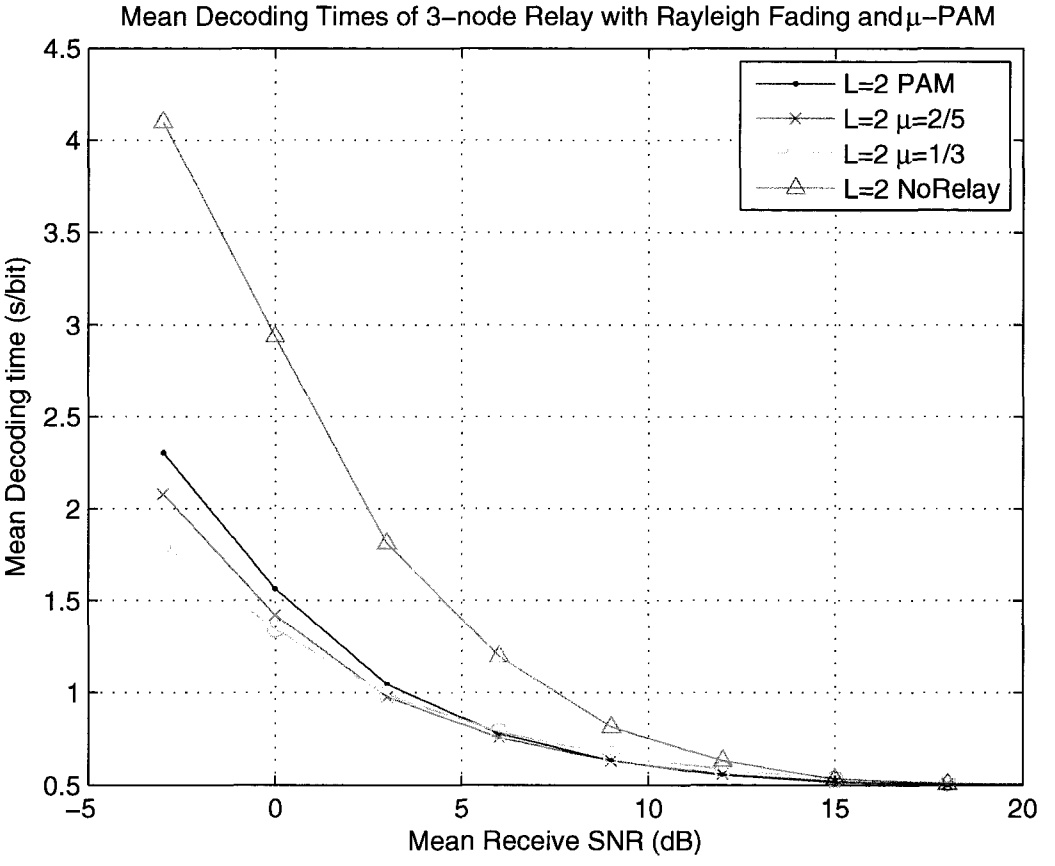

Figure 5.10: Mean decoding times vs. average receive SNR, $L = 2, G = 15$dB

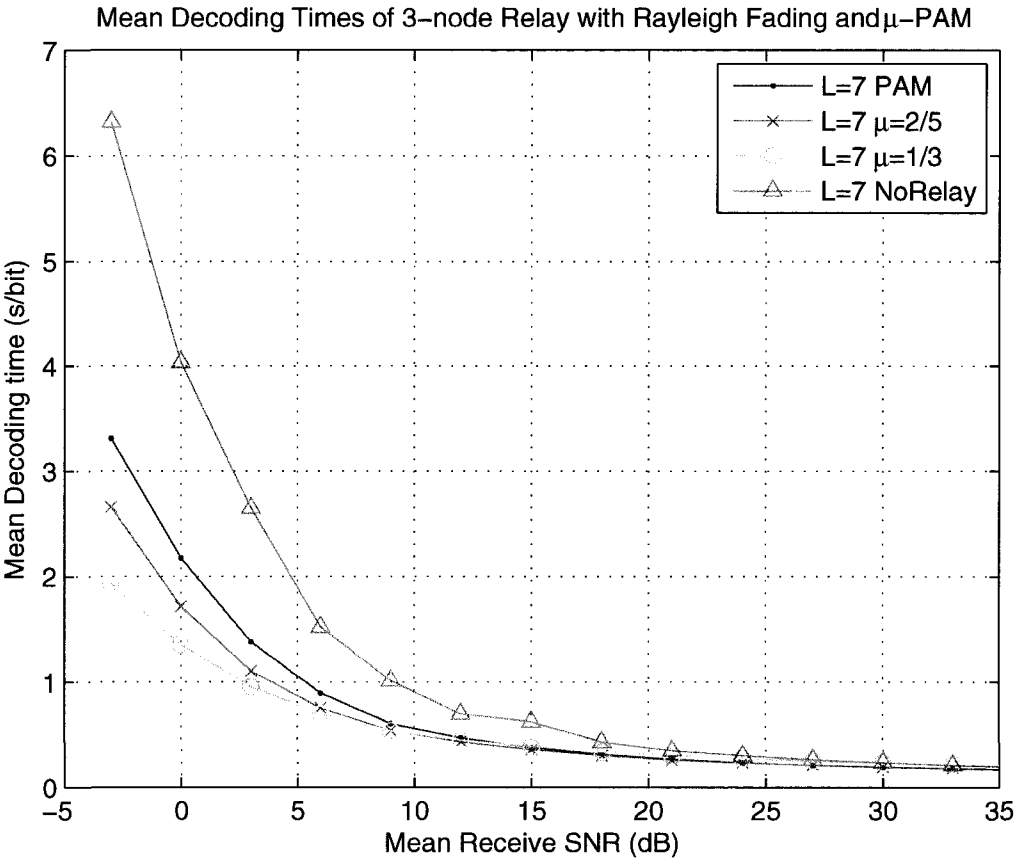

Figure 5.11: Mean decoding times vs. average receive SNR, $L = 7, G = 15$dB

Figure 5.12 extends the results presented in Figure 5.4 using $M = 7$ and μ-PAM values of 1/3 and 2/5. In addition, M-PAM results for $L = 2$ and $L = 7$ are shown for comparison purposes. The results are consistent with the previous figures comparing μ-PAM in this section. Given that the simulation was run at a relatively low SNR (-5dB to 15dB), the μ-PAM performance numbers are superior over the entire SNR range. In fact, $\mu = 1/3$ is the best performing setting for all SNRs.

In general, the simulation results reinforce the fact that Raptor codes and μ-PAM modulation can work well together over the relay channel, and can achieve a large fraction of the capacity of the channel, even when the realizations of the channel are not known to any of the transmitters.

5.3.3 Extension to Multiple Relays

The presented framework may naturally be extended to include an arbitrary number of relays, corresponding to the "asynchronous protocol" in [52]. All terminals in the network which are neither the source or destination terminal may act as a receiver for the codeword transmission from the source and other relays.

Each relay accumulates information from all other current collaborators. As soon as it is able to decode, it begins collaboration with the source and other relays already collaborating. This approach can match the achievable rate region for arbitrary relay networks found in [29].

Intuitively one can think about the flow of information through the network beginning at the source and ending at the destination node. As wireless is a broadcast medium, all transmissions may be received by any other node in the network subject to availability and SNR conditions. As relay nodes decode incoming messages from the source and other relays, it "joins the flow" in some coordinated manner, thereby increasing the overall rate at which information is flowing. This continues in a cascade manner, with more and more relays adding to the information flow rate, until the destination node receives the message sent originally from the source node.

In addition to this approach, other options are detailed in [52]. Specifically, one may coordinate relays in the network to not collaborate until a sufficient number have successfully decoded the source message. This offers some potential benefits in ensuring some minimum level of quality or redundancy in the network. This in turn allows for reliability against routing errors or relays dropping out of the network, making it more suited to ad-hoc systems. Tradeoffs between these two extensions of the basic framework

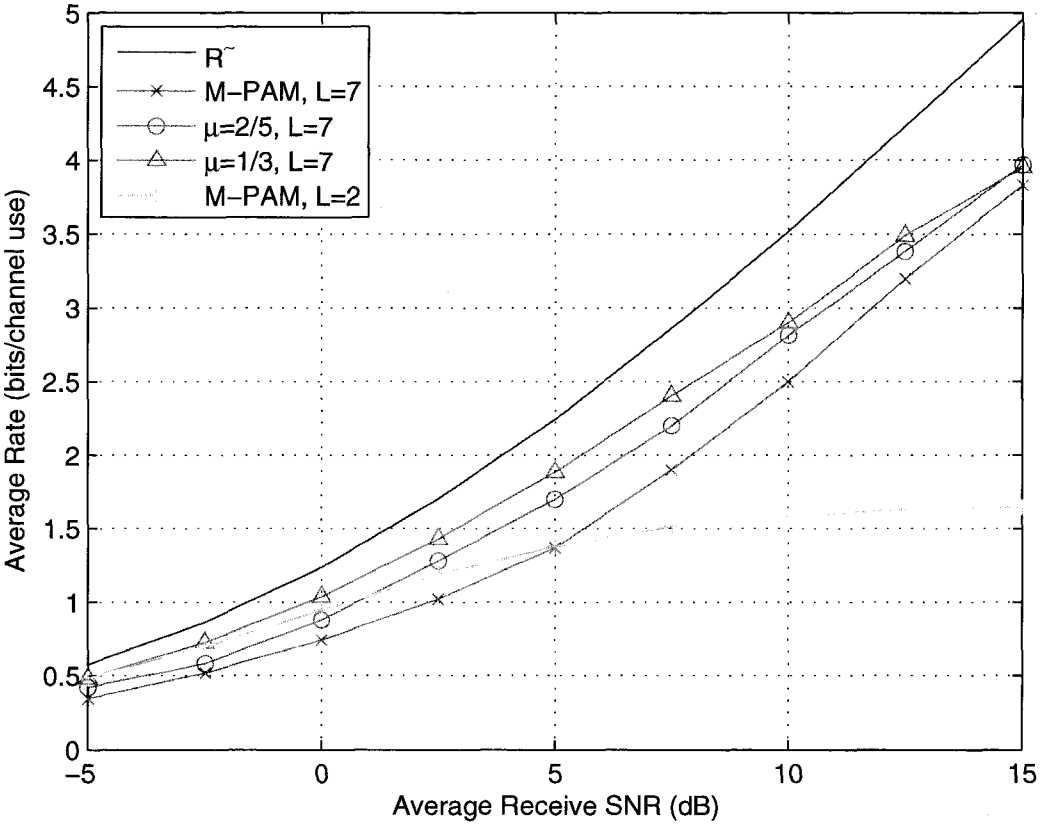

Figure 5.12: Mean system rates vs. average receive SNR with μ-PAM, $G = -5$dB

are discussed in detail in [52].

5.3.4 Practical Considerations

Some aspects of this framework have been idealized or otherwise simplified for the sake of convenience. In practical settings, any such collaboration scheme must consider at least the coordination and synchronization of relays, and provide an operational feedback channel.

Coordination of relays, and of the system as a whole, is of fundamental importance. By coordination, we mean the allocation of unique resources to collaborating relays. For example, if using space-time codes for collaboration, how should a relay choose which column (in the case of space-time block codes) or component of the code to use? Do the source or relays need to be notified of the participation of other relays? It may be possible to uniquely identify each relay in a small network, and use brute-force detection methods at the receiver, but this in not scalable to larger networks.

The work in [61] presents one possible option which the authors term as "opportunistic large arrays", in which collaborating relays flood the network. The random distribution of relays and channel realizations result in both space and time diversity gains. Other alternatives utilizing space-time codes may be found in [55, 50].

A different approach found in [52] uses CDMA spreading codes allocated uniquely per relay which, at the expense of bandwidth expansion, provides orthogonality at the destination. Of course the scalability of such a scheme is limited by the number of orthogonal spreading codes that are available, which is proportional to the degree of bandwidth expansion. A Rake receiver is used for recovery of the desired symbols from the collaborators.

The work in [3] suggests the use of an *artificial Inter-Symbol Interference (ISI)* channel instead of an orthogonalization method, which they show can achieve the bounds of the DMT. Here, they propose the use of self interference in the form of ISI as a means of providing diversity. Despite this interference, the DMT may still be achieved. This is due in great part to the fact that the ISI is self induced in a controlled manner and not something that is randomly added by the channel.

In the presented system, it is assumed that symbol-level synchronization is maintained between received signals. In relay networks with unknown or variable distances between terminals, this may not be simple to achieve, although a number of potential solutions may be used. For example, at the expense of bandwidth expansion, OFDM may be

applied, effectively reducing the symbol rate (for the same data rate) to a suitable degree.

Synchronization is not vital for operation of the system, though lack of it may limit achievable rates. Many wireless systems operate in an asynchronous manner; the suitability of such a design decision depends on the various tradeoffs that must be made between performance and complexity of the implementation. This degree to which synchronization impacts performance is developed in [52] and the some achievable rates are presented there.

Finally, a feedback channel is an essential feature of any rateless system. Practically, the feedback channel must exist orthogonally to the forward channel. This requirement may have cost or complexity implications for the terminals in the network. Natural options are to use either TDD of FDD between forward and feedback channels.

The amount of feedback required is relatively minimal; a single bit must be fed back (broadcast) to the source and any collaborating relays. However, this feedback must be reliable. If the feedback channel is noisy, then coding is likely required. This will induce some latency into the feedback channel. The effect of this latency is to directly reduce the throughput of the system by the amount of the latency. Of course, more sophisticated feedback schemes may mitigate this impact.

5.4 Conclusions

We have presented a framework for collaboration over wireless relay channels based on rateless codes that is simultaneously robust against outage and efficient in rate. In particular, we have shown that this system can be implemented using rateless codes and is capable of performing at rates approaching theoretical limits across a wide variety of channel configurations.

The proposed framework consists of two phases, the listening and the collaboration phase. Assuming that rateless codes are used, we determined the achievable rate region for this framework. We showed that there is a maximal achievable rate and that the relay naturally begins to collaborate such that this maximal rate can be achieved.

We demonstrated through simulations that this framework does indeed provide gains in terms of throughput and probability of outage. This was shown to hold across a large range of SNR, though neither the source or the relay knew the channel they were transmitting over. We then made use of μ-PAM modulation detailed in Chapter 3 and showed that their can provide an overall improvement in performance for many SNR ranges. Performance in terms of throughput and mean time to decode was presented.

The combination of the ideas of rateless codes and wireless relay networks provides for a method of increasing the throughput and reliability of wireless networks. Rateless codes, with their ability to adapt to the conditions of the channel–without requiring channel knowledge at the transmitter–are well suited for the channel conditions that wireless relay networks typically operate under.

A number of extensions were discussed and we provided some ideas on how the three-node relay network studied here can be used to build large, practical wireless relay network systems based on this framework. Limitations due to the need to coordinate the space-time codes of the relays and the need for synchronization were also discussed, though each of these issues present no fundamental limitation.

Relay networks, with their ability to exploit the spatial diversity that physically separated terminals provide, are capable of providing increased coverage and throughput in the system. Together, these two technologies complement each other, and promise to help multi-terminal networks to become economically viable means for wireless communications.

Chapter 6

Conclusions

In this work we have investigated rateless coding and modulation for wireless communication systems. The communication problem we explored used a rather general block-fading model in which we assumed that only the receiver had information about the channel. The coherence time (or coherence block size in a digital communication system) of the channel was assumed to take values that were of a similar magnitude as the duration (or size in discrete time) of the transmitted codeword. It was in this regime, where neither of the well known results of quasi-static fading and fast fading were applicable, that we focused our attention.

We argued in Chapter 2 that the time to decode a message of length k was of fundamental importance, an we examined the form that this random value over the block-fading channel. The discontinuous distributions that resulted were shown to be useful for measuring other performance metrics of interest such as the probability of outage. It was clear that fixed-rate coding schemes could not be robust in such a system; either efficiency or reliability must be sacrificed. These observations served as a motivation to investigate the use of rateless codes in place of fixed-rate codes for this class of channels.

We then introduced rateless coding as a viable method for reliable communication over fading channels when only the receiver has information about the channel. Rateless codes were shown to have some beneficial properties that make them well suited to the problem. In particular we proved the existence of good rateless codes, where the term good was used to imply simultaneous robustness to all possible channel realizations.

An implementation of rateless codes in the form of a Raptor code over fading channels was presented using basic PAM modulation. Simulation of the Raptor code showed that it was able to automatically adapt its rate to the realized capacity of the channel. Overall

performance across a broad range of SNRs was simulated and achieved rates across these SNRs tracked capacity well, albeit with some implementation loss.

Reliable digital communication systems require both coding and modulation, and we turned our attention to the problem of modulation for rateless systems in Chapter 3. Here we argued that typical modulation methods such as M-ary PAM are not well suited to communication over unknown channels. The fact that one must choose M *a priori* and the uniform distances between adjacent symbols results in such systems having poor performance if the SNR of the channel is not well matched to the choice of M. This characteristic makes M-PAM a possibly poor choice for unknown channels.

To alleviate this problem we introduced μ-PAM modulation, a PAM-based modulation system which allocated bits to symbols in a non-uniform, recursive manner. The resulting non-uniform constellation was shown to have scalable minimum distance properties as a function of the input bits that constitute a symbol. In particular, we showed that, due to the recursive nature in which μ-PAM is defined, the minimum distance between symbols conditioned on a particular bit was equal to the minimum distance of the previous bit times a factor of μ. We demonstrated how μ controlled the degree of non-uniformity, and in effect, the amount of protection allocated to different bits in a symbol.

A simplified demodulation method was introduced and analyzed for use with μ-PAM constellations. Using a numerical method to estimate the mutual information between input bits and output estimated bits over an AWGN channel, we presented and compared μ-PAM and M-PAM performance metrics. When the transmitter does not have knowledge of the channel, it was shown that μ-PAM is often a better choice compared to M-PAM due to better performance over a larger range of SNR. This conclusion was also borne out by simulations made comparing μ-PAM and M-PAM modulation methods across a broad range of SNRs. This was exactly the result we were aiming to achieve for integration with rateless codes.

We then turned our attention in Chapter 4 to the practical problem of communication over unknown fading channels with delay constraints. When constrained by a delay limit probability of outage can no longer be driven to zero as was shown to be otherwise possible with rateless codes in Chapter 2. However, we investigated a streaming communication problem with delay constraints, which was practically represented by a video streaming application. Given the sequential transmission of codewords, we showed that rateless codes offer significant benefits when compared to fixed-rate systems. We also compared with block-based incremental redundancy schemes based on Hybrid-ARQ.

The probability of outage characteristics were explored for the streaming application, and it was found that for rateless systems it is possible to define a critical SNR. At SNRs greater than this critical value it was demonstrated that, over sufficiently long periods, probability of outage could indeed be driven to zero if a rateless or rate-adaptive system was used. Otherwise, we showed that probability of outage was bounded above zero.

We then presented an analysis and comparison of fixed-rate and rateless coding for the video streaming application. Comparisons were made using performance metrics of throughput and probability of outage. We identified the critical SNR values and other proxies for this critical value in terms of normalized information rate k/T_c and normalized delay constraint D/T_c. The results showed that rateless codes were the best choice in all scenarios.

In Chapter 5 we explored the use of a joint rateless coding and modulation system for relay networks. We provided a background on the area of relay channels and relay networks. We then presented a framework for reliable communication over unknown, half-duplex relay channels. This was built upon previous information theoretic results using a two-phase communication scheme that did not consider the possibility of rateless codes. We extended this work to incorporate the use of rateless codes and presented an optimal information theoretic achievable rate region for the relay channel.

Based on this framework, we simulated an implementation using rateless codes and μ-PAM modulation with a single relay. The results showed that performance tracked the theoretical achievability limits reasonably well, and the implementation loss remained reasonably constant relative to these limits. In addition to achievable rates, we also argued that mean time to decode is an important metric in this setting, and we showed that μ-PAM offered improvements compared to M-PAM.

We outlined how this framework can be extended to multiple relays in a very natural way; each additional relay operates independently and the improvement in diversity is proportional to the number of relays. A number of practical issues were also discussed. The need for synchronization and coordination between nodes in the network was discussed and citations to works in this area offering potential solutions were given.

This work has presented a framework for reliable communication over a class of fading channels when the channel is known only at the receiver. The use of rateless codes, with their ability to adapt automatically to the dynamic, fluctuating channel state was shown offer numerous benefits in terms of throughput, time to decode, and probability of outage. Coupled with μ-PAM modulation, the joint communication system was shown to be effective to both delay constrained communication systems and relay networks.

6.1 Extensions

The work presented here can be extended in a number of ways, and there are a number of challenges that must be solved before the framework proposed in this work can be put into practice. This section briefly touches on a few of these.

The inclusion of Raptor codes or other rateless codes into communications standards are potentially hampered by three factors. The first is that practical decoding costs are actually quite large. Despite Raptor codes having linear decoding complexity as a function of codeword length, the fact that many unsuccessful decoding attempts are often made before the realized rate is sufficiently low results in a high overall decoding complexity. There are many options to solve this problem, and this is an area of active research. Different decoding schedules on the factor graph representation, and the reuse of previous decoding estimates such as proposed in [35,34] offer directions toward solving this problem.

Second, the fact that rateless codes may be decoded after any number of channel uses, and thus have no *a priori* rate, is a major change in typical thinking about coding and one which does not easily fit within the framework of many communication standards. For example, many wireless standards segment communication into deterministic, fixed-time periods (equivalently fixed symbol lengths). Unfortunately this is not well suited to the nature of rateless codes. One may envisage a radio system which behaves somewhat more like the rateless codes themselves, transmitting coded symbols until receiving an acknowledgment from a receiver. This is similar to the approach used in Ethernet, and it seems that the use of rateless codes for fading channels will necessitate similar types of physical layer and radio architectures.

Finally, we have discussed the need for a feedback channel to enable robust rateless codes. At this time we have not explored how this feedback channel can be implemented, or what the effect of imperfections of this channel will have on the system. It is required that it be reliable, so presumably some coding will be needed over this channel in addition to the forward channel itself. Foundational work in this area can be found in [62]. Practical aspects of the feedback channel; errors, latency, and supported rates will all impact the overall performance of the forward channel, and protocols must be developed at the link layer or medium-access layer to support the overall robustness of the rateless system.

We introduced μ-PAM modulation as a means to better support many of the characteristics that rateless codes offer. In particular μ-PAM offers reasonable performance

over a broader range of SNRs than M-ary PAM. However to this point we have only explored single dimensional modulation, and have neglected gains that may be had by shaping. Thus an area of future research is to explore higher-dimensional modulation techniques. Mapping some number of rateless coded bits to a point on a hypersphere has the potential to offer many of the same benefits of μ-PAM with improved performance. Of course higher dimensional mapping will entail a higher complexity for both modulation and demodulation.

The idea of "twisted modulation" as described in [82] may be particularly well suited for rateless coding as a very large number of coded bits may be encoded into a single symbol, and adaptive demodulation may extract only as many as the channel will support. Unfortunately the non-linear nature of this modulation scheme makes it potentially difficult for practical implementation.

In the space of relay networks, there are a number of areas which may be investigated with respect to rateless coding. Relay networks are really distributed multi-antenna systems, so the application of rateless codes to other MIMO and cooperative networks can be explored with reasonable hope of success. However, the problem of the coordination of relays remains a practical obstacle that must be overcome. Some ideas and current research results in this area were discussed in Section 5.3.4, such as the allowance for bandwidth expansion to enable unique spreading codes to be used for the relays. Relay networks and multi-terminal systems is a very active research area, and new results are continuing to reduce the gap between theory and application.

The use of joint rateless coding and modulation for wireless communications offers potentially significant benefits in application. Although obstacles remain before such systems are realizable in practice, it appears that there are no fundamental issues which will prevent these systems from becoming a reality. The work we have presented we hope will help serve as evidence and motivation to continue to explore the use of rateless concepts for practical communication schemes.

Bibliography

[1] N. Ahmed and R. G. Baraniuk. Throughput measures for delay-constrained communications systems in fading channels. In *Proc. 41st Annual Allerton Conference on Communication, Control and Computing*, Oct. 2003.

[2] S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE J. Select. Areas Commun.*, 16:1451–1458, Oct. 1998.

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

[4] J. Boutros, A. Guillién i Fàbregas, and E. Strinati. Analysis of coding on non-ergodic channels. In *Proc. 43rd Annual Allerton Conference on Communication, Control and Computing*, Sept. 2005.

[5] J.D. Brown, S. Pasupathy, and K.N. Plataniotis. Adaptive demodulation using rateless erasure codes. *IEEE Trans. Commun.*, 54(9):1574–1585, Sept. 2006.

[6] G. Caire and D. Tuninetti. The throughput of hybrid-ARQ protocols for the Gaussian collision channel. *IEEE Trans. Inform. Theory*, 47:1971–1988, Jul. 2001.

[7] G. Caire, D. Tuninetti, and S. Verdu. Variable-rate coding for slowly fading Gaussian multiple-access channels. *IEEE Trans. Inform. Theory*, 50(10):2271–2292, Oct. 2004.

[8] J. Castura and Y. Mao. Rateless coding for wireless relay channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 810–814, 2005.

[9] J. Castura and Y. Mao. Rateless coding over fading channels. *IEEE Commun. Lett.*, 10(1):46–48, Jan. 2006.

[10] J. Castura and Y. Mao. When is a message decodable over fading channels? In *Communications, 23rd Biennial Symposium on*, pages 59 – 62, May 2006.

[11] J. Castura and Y. Mao. Non-binary rateless codes for unknown gaussian channels. In *Proc. IEEE Canadian Workshop on Inform. Theory*, Jun. 2007.

[12] J. Castura and Y. Mao. Rateless coding and relay networks. *IEEE Signal Processing Mag.*, 24(5):27–35, 2007.

[13] J. Castura and Y. Mao. Rateless coding for wireless relay channels. *IEEE Trans. Wireless Commun.*, 6(5):1638–1642, 2007.

[14] J. Castura, Y. Mao, and S. C. Draper. On rateless coding over fading channels with delay constraints. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 1124–1128, 2006.

[15] Z. Cheng, J. Castura, and Y. Mao. On the design of raptor codes for binary-input Gaussian channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, June 2007.

[16] T. Cover and A. El Gamal. Capacity theorems for the relay channel. *IEEE Trans. Inform. Theory*, 25(5):572–584, Sept. 1979.

[17] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[18] Digital Fountain Incorporated. `http://www.digitalfountain.com`.

[19] M. Dohler, E. Lefranc, and H. Aghvami. Virtual antenna arrays for future mobile communication systems. In *IEEE ICT 2002, Beijing, China*, Jun. 2002.

[20] S. C. Draper, B. J. Frey, and F. R. Kschischang. Efficient variable length channel coding for unknown DMCs. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 379, 2004.

[21] S. C. Draper, B. J. Frey, and F. R. Kschischang. Rateless coding for non-ergodic channels with decoder channel state information. Submitted to *IEEE Trans. Infor. Theory*, 2006.

[22] O. Etesami, M. Molkaraie, and A. Shokrollahi. Raptor codes on symmetric channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 38, 2004.

[23] O. Etesami and A. Shokrollahi. Raptor codes on binary memoryless symmetric channels. *IEEE Trans. Inform. Theory*, 52(5):2033–2051, May 2006.

[24] G.D. Forney Jr. and G. Ungerboeck. Modulation and coding for linear Gaussian channels. *IEEE Trans. Inform. Theory*, 44(6):2384–2415, Oct. 1998.

[25] C. Fragouli, R.D. Wesel, D. Sommer, and G.P. Fettweis. Turbo codes with non-uniform constellations. In *IEEE International Conference on Communications 2001*, volume 1, pages 70–73, Jun. 2001.

[26] R. Gallager. *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[27] M. Gastpar and M. Valenti. On the capacity of large Gaussian relay networks. *IEEE Trans. Inform. Theory*, 51(3):765–779, Mar. 2005.

[28] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, 2001.

[29] P. Gupta and P. R. Kumar. Towards an information theory of large networks: An achievable rate region. *IEEE Trans. Inform. Theory*, 49(8):1877–1894, Aug. 2003.

[30] P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Trans. Inform. Theory*, 46(2):388–404, Mar. 2000.

[31] A. Gut. *Stopped Random Walks: Limit Theorems and Applications.* Springer-Verlag, 1988.

[32] S. Hanly and D. Tse. Multiaccess fading channels-part II: Delay limited capacities. *IEEE Trans. Inform. Theory*, 44(7):2816–2831, Nov. 1998.

[33] J. Hu and T. Duman. Low density parity check codes over half-duplex relay channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 972–976, 2006.

[34] K. Hu, J. Castura, and Y. Mao. Reduced-complexity decoding of Raptor codes over fading channels. In *IEEE GLOBECOM 2006*, pages 1–5, Nov. 2006.

[35] K. Hu, J. Castura, and Y. Mao. Performance-complexity tradeoffs of Raptor codes over Gaussian channels. *IEEE Commun. Lett.*, 11(4):343–345, Apr. 2007.

[36] T. E. Hunter and A. Nosratinia. Cooperation diversity through coding. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 220, 2002.

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

[38] H. Jenkac, T. Stockhammer, and W. Xu. Asynchronous and reliable on-demand media broadcast. *IEEE Network*, 20(2):14–20, Mar. 2006.

[39] G. Kramer, M. Gastpar, and P. Gupta. Cooperative strategies and capacity theorems for relay networks. *IEEE Trans. Inform. Theory*, 51(9):3037–3063, Sept. 2005.

[40] K.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47(2):498–519, Feb. 2001.

[41] S.R. Kulkarni and P. Viswanith. A deterministic approach to throughput scaling in wireless networks. *IEEE Trans. Inform. Theory*, 50(6):1041–1049, Jun. 2004.

[42] L. Lai, K. Liu, and H. El-Gamal. The three-node wireless network: achievable rates and cooperation strategies. *IEEE Trans. Inform. Theory*, 52(3):805–828, Mar. 2006.

[43] J. N. Laneman, D. N. C. Tse, and G. W. Wornell. Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *IEEE Trans. Inform. Theory*, 50:3062–3080, Dec. 2004.

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

[45] A. Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, 44(6):2148–2175, Oct. 1998.

[46] S. Lin and D. Costello. *Error control coding.* Prentice Hall, Apr. 2004.

[47] Y. Liu. A low complexity protocol for relay channels employing rateless codes and acknowledgement. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 1244–1248, 2006.

[48] M. Luby. LT codes. In *43rd Annual IEEE Symp. on the Found. of Comp. Sci.*, pages 271–280, 2002.

[49] M. Medard. The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel. *IEEE Trans. Inform. Theory*, 46(3):933–946, May 2000.

[50] B.S. Mergen and A. Scaglione. Randomized space-time coding for distributed cooperative communication. In *IEEE International Conference on Communications*, volume 10, pages 4501–4506, Jun. 2006.

[51] P. Mitran, H. Ochiai, and V. Tarokh. Space-time diversity enhancements using collaborative communications. *IEEE Trans. Inform. Theory*, 51(6):2041–2057, Jun. 2005.

[52] A. F. Molisch, N.B. Mehta, J. S. Yedidia, and J. Zhang. Cooperative relay networks using fountain codes. In *Global Telecommunications Conference*, pages 1 – 6, Nov. 2006.

[53] A. F. Molisch, N.B. Mehta, J. S. Yedidia, and J. Zhang. Cooperative relay networks with mutual-information accumulation. *to appear, IEEE Trans. Wireless Comm.*, 2007.

[54] R.H. Morelos-Zaragoza, M.P.C. Fossorier, S. Lin, and H. Imai. Multilevel coded modulation for unequal error protection and multistage decoding. II. asymmetric constellations. *IEEE Trans. Commun.*, 48(5):774 – 786, May 2000.

[55] R.U. Nabar, H Bolcskei, and F.W. Kneubuhler. Fading relay channels: performance limits and space-time signal design. *IEEE J. Select. Areas Commun.*, 22(6):1099–1109, Aug. 2004.

[56] R. Negi and J.M. Cioffi. Delay-constrained capacity with causal feedback. *IEEE Trans. Inform. Theory*, 48(9):2478–2494, Sept. 2002.

[57] N.H. Ngo, S.S. Pietrobon, and S.A. Barbulescu. Performance of non-uniform 16QAM modulation over linear and nonlinear channels. *IEEE Electronics Letters*, 42(9):544 – 546, Apr. 2006.

[58] J.M. Ooi and G.W. Wornell. Fast iterative coding techniques for feedback channels. *IEEE Trans. Inform. Theory*, 44(7):2960–2976, Nov. 1998.

[59] R. Palanki and J. S. Yedidia. Rateless codes on noisy channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 37, 2004.

[60] M.B. Pursley and J.M. Shea. Nonuniform phase-shift-key modulation for multimedia multicast transmission in mobile wireless networks. *IEEE J. Select. Areas Commun.*, 17(5):774–779, May 1999.

[61] A. Scaglione and Y.W. Hong. Opportunistic large arrays: cooperative transmission in wireless multihop ad hoc networks to reach far distances. *IEEE Trans. Signal Processing*, 51(8):2082 – 2092, Aug. 2003.

[62] J.P. Schalkwijk and T. Kailath. A coding scheme for additive noise channels with feedback, Pt. I: No bandwidth constraint. *IEEE Trans. Inform. Theory*, 12(2):72–182, Apr. 1966.

[63] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity. part I. system description. *IEEE Trans. Commun.*, 51(11):1927–1938, Nov. 2003.

[64] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity. part II. implementation aspects and performance analysis. *IEEE Trans. Commun.*, 51(11):1939–1948, Nov. 2003.

[65] S. Sesia, G. Caire, and G. Vivier. Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes. *IEEE Trans. Commun.*, 52(8):1311–1321, Aug. 2004.

[66] A. Shokrollahi. Raptor codes. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 36, 2004.

[67] N. Shulman. Communication over an unknown channel via common broadcasting, 2003. Ph.D. Thesis.

[68] M. J. Sobel and D. P. Heyman. *Stochastic Models in Operations Research, Vol. I.* Dover Publications, 2003.

[69] E. Soljanin, N. Varnica, and P. Whiting. LDPC code ensembles for incremental redundancy hybrid ARQ. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 995–999, Sept. 2005.

[70] T. Tang, C-B Chae, R. Heath Jr., and S Cho. On achievable sum rates of a multiuser MIMO relay channel. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 1026–1030, 2006.

[71] A. Tchamkerten and E. I. Telatar. Optimal feedback schemes over unknown channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 378, 2004.

[72] I.E. Telatar. Capacity of multi-antenna Gaussian channels. *European Trans. on Telecomm.*, 10(6):585–595, 1999.

[73] The Bluetooth SIG. `http://www.bluetooth.org`.

[74] The GSM Association. `http://www.gsmworld.com`.

[75] The IEEE 802.11 Working Group. `http://www.ieee802.org/11/`.

[76] The UMTS Forum. `http://www.umts-forum.org`.

[77] The Wi-Fi Alliance. `http://www.wi-fi.org`.

[78] The WiMAX Forum. `http://www.wimaxforum.org`.

[79] D. Tse and P. Viswanath. *Fundamentals of wireless communication.* Cambridge University Press, Jun. 2005.

[80] D. Tuninetti and G. Caire. The effect of delay constraint and causal feedback on the wideband performance of multiaccess block-fading channels. In *36th Asilomar Conf. on Signals, Systems and Computers*, Nov. 2001.

[81] B. Wang, J. Zhang, and A. Host-Madsen. On the capacity of MIMO relay channels. *IEEE Trans. Inform. Theory*, 51(1):29–43, Jan. 2005.

[82] J.M. Wozencraft and I.M. Jacobs. *Principles of Communication Engineering.* John Wiley and Sons, 1965.

[83] D. Wu and R. Negi. Effective capacity: a wireless link model for support of quality of service. *IEEE Trans. Wireless Commun.*, 2(4):630–643, Jun. 2003.

[84] L.L. Xie and P.R. Kumar. A network information theory for wireless communication: scaling laws and optimal operation. *IEEE Trans. Inform. Theory*, 50(5):748–767, May 2004.

[85] M Yuksel and E. Erkip. Diversity-multiplexing tradeoff in multiple-antenna relay systems. In *Proc. IEEE Int. Symp. on Inform. Theory*, pages 1154–1158, 2006.

[86] M Yuksel and E. Erkip. Multi-antenna cooperative wireless systems: A diversity-multiplexing tradeoff perspective. *IEEE Trans. Inform. Theory*, pages 3371–3393, Oct. 2007.

[87] Z. Zhang. Partial converse for a relay channel (corresp.). *IEEE Trans. Inform. Theory*, 34:1106–1110, Sept. 1988.

[88] Z. Zhang, I. Bahceci, and T. M. Duman. Capacity approaching codes for relay channels. In *Proc. IEEE Int. Symp. on Inform. Theory*, page 2, 2004.

[89] B. Zhao and M. C. Valenti. Practical relay networks: A generalization of hybrid-ARQ. *IEEE J. Select. Areas Commun.*, 23:7–18, Jan. 2005.

[90] L. Zheng and D. N. C. Tse. Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *IEEE Trans. Inform. Theory*, 49(5):1073– 1096, May 2003.