ADVANCED CODING TECHNIQUES FOR MULTICASTING IN WIRELESS COMMUNICATIONS

Stefania Sesia

Thèse

Présentée pour Obtenir le Grade de

Docteur de l'École Nationale Supérieure des Télécommunications (ENST)

Spécialité: Communication et Electronique

Date de Soutenance: 29 Juin 2005
Composition du Jury:

Rapporteur	Prof. J. Hagenauer, (TUM, Munchen)
Rapporteur	Prof. R. Knopp, EURECOM (Sophia-Antipolis, France)
Examinateur	Prof. J.C Belfiore, ENST (Paris, France)
Examinateur	Prof. T. Ramstad, NTNU (Trondheim,Norway)
Examinateur	PhD. G. Vivier, MOTOROLA LABS (Gif-sur-Yvette, France)
Directeur de Thèse	Prof. Giuseppe Caire, EURECOM (Sophia-Antipolis, France)

Thèse réalisée au sein de l'Institut Eurecom et Motorola Labs

AUTHOR'S ADDRESS:
Stefania Sesia
Institut Eurecom
Departement de Communication Mobile
2229, route des Crêtes
B.P. 193
06904 Sophia-Antipolis – France
Email: Stefania.Sesia@eurecom.fr
URL: http://www.eurecom.fr/∼sesia

Acknowledgment

This thesis is the result of three year effort spent, partly at the Mobile Communications Department of institute Eurecom and partly at Motorola Labs in Paris, thus providing me with enriching both industrial and academic environments.

I would like to thank Motorola Labs for giving me the financial support and the technical environment to carry out this work. A special thank to my industrial supervisor, Guillaume Vivier, who created the proper conditions for me to develop my technical ideas within the Motorola group. The 15 months spent in Motorola Labs, in their Paris facilities, have been enjoyable both professionally and socially. I would like to thank the good friends I have met in Paris during that period for their help and for all the moments spent together. In particular Emilio, Pietro, Elisa, Jean-Christophe, among others, have made me really appreciate my stay in Paris.

Some people have given me invaluable help and scientific support during my thesis, among them I would like to express all my gratitude to my advisor Giuseppe Caire whose creativity has given me the opportunity to learn and explore different subjects related to source and channel coding.

The international and rich environment at Eurecom has surely played an important role in the realization of the thesis. The discussions with colleagues on both technical and more personal levels have contributed to a continued enjoyable and fruitful atmosphere. Among the persons I have met at Eurecom, I give Younes a special thanks. Many thanks also to David and all the other PhD students for their support.

Finally I would like to thank my family for their love and support.

Abstract

The thesis addresses some open problems in the area of efficient transmission of loss-sensitive and delay-sensitive data over wireless channels. The thesis mainly deals with coding techniques for multicast systems. Multicast differs from the information theoretic broadcast channel in that only common information is sent. In point-to-point transmission, reliability is achieved by means of Automatic Retransmission reQuest (ARQ). Forward Error Correcting (FEC) codes and ARQ are combined together in order to optimize the trade-off between reliability and efficiency. This approach is called Hybrid ARQ, (HARQ).

We consider HARQ schemes for point-to-point transmission with modern coding techniques (Low Density Parity Check codes, LDPC). The theoretical analysis shows that these codes ideally achieve optimal performance in terms of throughput. However, for practical finite-length codes, the scheme exhibits a loss in performance. Two different solutions are shown to recover most of this performance gap. Analysis of the complexity of HARQ schemes with LDPC decoding shows that a method based on asymptotic analysis (Density Evolution) yields considerable savings with respect to the other criteria that stop the iterations of the LDPC decoding.

In a multicast setting, however, HARQ protocols are inefficient. Strictly speaking, they are not fully scalable, i.e. the throughput goes to zero when the number of users increases. This motivates us to study the throughput per user of these protocols. In particular, HARQ based on Selective Repeat (SR) or Incremental Redundancy (IR) can be defined to be fully scalable if we allow for a fraction $x > 0$ of users that do not decode successfully. We show that under certain conditions the throughput achieved by the IR protocol equals the ergodic capacity at the expense of a delay that grows to infinity. Moreover, when the number of users increases, the performance of IR is identical to performance achieved by a FEC based system, in terms of delay, throughput and error probability. This makes questionable the interest of retransmission protocols in a multicast setting.

While in the first part of the thesis we have considered data communications, for which the relevant performance measure is error probability, in the second part we consider the trans-

mission of an analog source (for example an image). Existing practical solutions, mainly based on Shannon's separation theorem, are highly inefficient and in particular they are not robust to channel errors. Only few bits in error at the output of the channel decoder lead to catastrophic effects on the reconstruction quality. This requires very stringent conditions on the performance of the channel code and leads to suboptimal performance in terms of spectral efficiency.

In a multicast setting, moreover, it is important to design a scheme that guarantees good performance over a wide range of signal to noise ratio. Different users with different channel conditions can decode the source with acceptable reconstruction quality. Joint source-channel coding is a viable solution for robustness and efficiency in this context.

In this multicast environment we analyze and optimize three well-known strategies in a comprehensive manner: the first is based on an ideal successive refinement source code, coupled with a a progressive transmission scheme (time sharing); the second concatenates the same source encoder with a superposition broadcast coding scheme. These two fully digital schemes are compared with an optimized Hybrid Digital Analog (HDA) approach based on bandwidth splitting of the source and the combination of a digital and an "analog" encoder. By "analog" encoder we mean that the source signal is sent directly on the channel (with a suitable scaling, in order to meet the transmit power constraint) as in conventional analog amplitude modulation. These schemes are optimized such that the average overall distortion is minimized subject to both power and spectral efficiency constraints.

Finally, the problem of code construction for the HDA system is addressed in the last part of the thesis. Two schemes are proposed.

In the first case all the complexity relies on the quantizer scheme. The quantizer is defined in such a way that its performance is based on bit error rate (and not frame error rate) at the output of the channel decoder. We consider an embedded Multistage Trellis Quantizer (MTQ), based on standard binary convolutional codes. It is shown to achieve performance close to the theoretical limit and comparable to the best results found in literature. Moreover, thanks to the fact that convolutional codes have non-catastrophic encoder, it is very robust with respect to channel errors.

The second is consider a very simple quantizer scheme. Data compression and channel coding are combined and accomplished with a linear code. Here a multilevel compression scheme based on linear codes (Turbo Codes) is considered. Linear codes provide compression by exploiting the redundancy at the output of the quantizer. Traditionally the source/channel coding system is based on the concatenation of the quantizer, a lossless data compression stage and a standard channel code. Here, these last two stages are replaced by a single stage based on linear codes. This approach is shown to give remarkable improve-

ments compared to the traditional solution. This method can be used in the HDA scheme when concatenated to entropy constrained scalar or vector quantizers. Moreover, it can be easily adapted to work with more sophisticated and practical quantizers as Differential Pulse Code Modulation (DPCM) for transmission of images and it can be generalized to achieve progressive transmission through embedded quantization.

Publications and Patents

This work has led to the following publications and patents.

Publications

- S. Sesia, G. Caire, and G. Vivier, "Broadcasting of a common source: information theory results and system challenges," in *Proc. Winter School on Coding and Information Theory*, Monte Verita, Switzerland, February 2003.
- S. Sesia and G. Caire, "Incremental redundancy schemes based on LDPCs for transmission over Gaussian block fading channels," in *Proc. IEEE Information Theory Workshop (ITW 2002)*, Bangalore, India, October 2002.
- S. Sesia, G. Caire, and G. Vivier, "The throughput of LDPC-based incremental redundancy schemes with finite blocklength," in *Proc. IEEE International Symposium on Information Theory (ISIT 2003)*,Yokohama, Japan, June-July 2003.
- S. Sesia, G. Caire, and G. Vivier, "Reducing the average complexity of LDPC decoding," in *Proc. 3rd International Symposium on Turbo Codes and Related Topics*, Brest, France, September 2003.
- S. Sesia, G. Caire, and G. Vivier, "On the scalability of HARQ systems in wireless multicast," in *Proc. IEEE International Symposium on Information Theory (ISIT 2004)*,Chicago,Illinois, USA, June-July 2004
- S. Sesia, G. Caire, and G. Vivier, 'Incremental Redundancy Hybrid ARQ Schemes Based on Low-Density Parity-Check Codes," *IEEE Trans. on Communications*, Vol. 52, pp. 1311–1321, August 2004.
- S. Sesia and G. Caire, "Multistage trellis quantization and its applications," in *Proc. 42th Annual Allerton Conference on Communication, Control and Computing (ALLERTON 2004)*, September-October 2004.

- S. Sesia and G. Caire, “Optimized joint source-channel techniques for compound channel,” in *Proc. IEEE International Symposium on Information Theory (ISIT 2005)*, September 2005.
- A.N. Kim, S. Sesia, T. Ramstad, and G. Caire, “Combined unequal error protection and compression using Turbo codes for resilient image transmission, ” in *Proc. IEEE International Conference on Image Processing (ICIP 2005)*, Genova, Italy, September 2005.
- S. Sesia, G. Caire, and G. Vivier, “Optimized techniques for lossy broadcasting and code construction,” *to be submitted to IEEE Trans. on Wireless 2005*.
- S. Sesia, A. N. Kim, G. Caire, T. Ramstad, “Combined unequal error protection and compression using Turbo codes for resilient image transmission”, *to be submitted to IEEE Trans. On Communications 2005*.

Patents

- **13761 CML00490EP**, “Method for faster and lower power consumption of iterative decoding algorithms when using a type-II hybrid ARQ transmission scheme”, result: publishing defense.
- **16842 CML00891EP**, “Method to speed up the iterative decoding algorithm of compound codes”, result: filed
- **18143 CML01052EP**, “A method for efficient reliable point to multipoint wireless communications”, result: filed
- **22657 CML01608EP**, “Method for fast design of broadcast wireless systems”, result: publishing defense

CONTENTS

List of Tables

List of Figures

Acronyms

Here is a list of the main acronyms used throughout the manuscript.

AC: Arithmetic Code
ACK: ACKnowledgment
AEC: Adaptive Entropy Code
ARQ: Automatic Retransmission reQuest
AWGN: Additive White Gaussian Noise
BER: Bit Error Rate
BF-AWGN: Block Fading AWGN
BI-AWGN: Binary Input AWGN
BP: Belief Propagation
BWT-MDL: Burros Wheeler Transform-Minimum Description Length
cdf: cumulative density function
CE: Cross Entropy
CRC: Cyclic Redundancy Check
DE: Density evolution
DE-GA: DE-Gaussian Approximation
DPCM: Differential Pulse Code Modulation
ECSQ: Entropy Constrained Scalar Quantizer
ECTCQ: Entropy Constrained Trellis Coded Quantizer
FEC: Forward Error Correction
FER: Frame Error rate
HARQ: Hybrid ARQ
HDA: Hybrid Digital Analog
IR: Incremental Redundancy
IRA: Irregular Repeat and Accumulate
JSCC: Joint Source and Channel Code

LDPC: Low Density Parity Check Code
LLR: Log-Likelihood Ratio
ML: Maximum Likelihood
MMSE: Minimum Mean Squared Error
MTQ: Multistage Trellis Quantizer
NACK: Negative ACK
NN: Non recursive Non systematic
OSR: Outer Selective Repeat
pdf: probability density function
PSNR: Peak Signal to Noise Ratio
QoS: Quality of Service
RB: Random Binary
RCPC: Rate Compatible Punctured Code
RS: Recursive Systematic
RSNR: Reconstructed Signal to Noise Ratio
SLB: Shannon's Lower Bound
SNR: Signal to Noise Ratio
SPA: Sum Product Algorithm
SR: Selective Repeat
SSCC: Separated Source Channel Code
M-TCOM: Multilevel Turbo COMpression

CHAPTER 1

Introduction

1.1 CHALLENGES AHEAD FOR WIRELESS COMMUNICATIONS

The demand for new high-speed, reliable, wireless services is growing fast. Future wireless networks will provide added value by allowing a large variety of services. Real-life networks require the performance to be compliant with certain quality of service targets in forms of delay, error probability or fidelity in the reconstruction of the data. Depending on the applications, different measures of performance may be more critical than others. For example data-transmission (web browsing, data transfer, email) is not strictly delay sensitive but require a virtually error free link. Multimedia content (video streaming) can be more delay sensitive but tolerate some losses, or it can relax the conditions on the delay and accept some losses as in the case of image transmission. Advanced source and channel coding is the key technology that allows for the design of a bandwidth efficient transmission layer [1, 2, 3].

While in the single user case, families of codes exist that achieve capacity for increasing block-length [1, 2], the multiuser scenario, and in particular broadcast scenario, is still not so simple. Broadcast channels have been widely studied over several years, especially from the information theoretic stand-point, [4, 5]. Nevertheless, the capacity region for a general broadcast channel has not been fully characterized yet. Information-theoretic broadcast channels correspond to systems where one transmitter sends independent information to

different users and possibly some common information. Here however we limit ourselves to a system where the transmitter sends only the common information to *all* the users. To differentiate this setting from that of the conventional broadcast channel, we refer to this situation as a *multicast* setting.

As an example consider a multicast application where several users ask a server the same service. Practically speaking, the server will open a new connection for each user demanding the service. This is clearly inefficient when considering the bandwidth consumption. In fact, it can happen that when a new user asks the same service, the system refuses to provide the service because of lack of bandwidth. Schemes that exploits the multicast setting by sharing the bandwidth, are surely more bandwidth efficient but at the expense of a penalty in the throughput seen by each user.

Hence, one of the challenges of wireless communications is to design bandwidth efficient systems that satisfy quality of service requirements.

This thesis tackles some open problems in the area of efficient transmission of loss-sensitive and delay sensitive data over wireless channels in single user and multicast setting.

1.2 CODING FOR DATA TRANSMISSION

1.2.1 Hybrid Retransmissions Schemes in Single User Setting

Data transmission is very sensitive to noise/fading related errors. Therefore, it is essential that the MAC layer is able to correct deficiencies of the physical layer code. To do so, Forward Error Correction (FEC) is complemented with a retransmission protocol (Automatic Repeat reQuest (ARQ)) [6]. FEC consists in adding a fixed amount of redundancy to the data packet allowing the decoder at the receiver side to correct a certain number of transmission errors. ARQ consists in requesting a retransmission, when the receiver detects a corrupted packet.The gain of using ARQ, beyond the benefit of eventually obtaining error free packets is to decrease the required operating error rate for physical layer FEC algorithms, thus effectively lowering the Signal to Noise Ratio (SNR) requirement at the terminal and access points. However, this gain comes at the price of an added delay due to retransmissions [6, 7]. A recent trend is the joint optimization between the physical layer and the MAC layer. An example of such cross-layer optimization can be seen in the joint design of the FEC and the ARQ. This gives rise to hybrid ARQ (HARQ) schemes where the decoding function of the physical layer is handled jointly with the combining of retransmitted packets (see [6, 7] and references therein). In type I HARQ, the basic idea is that multiple disjoint coded versions of the same original packets are transmitted upon each

retransmission. The code design is such that the message is decodable upon reception of just a single coded packet. The different copies can be combined at the decoder in order to better exploit the diversity of the channel and to increase the probability of successful decoding. Packet combining can be based on hard decision [8, 9, 10] or on soft channel output [11, 12]. Throughout the thesis we refer to the type-I HARQ as the Selective Repeat (SR) protocol.
Type-II HARQ (also called Incremental Redundancy, (IR)), [13, 14] can be interpreted as a variable coding scheme where flexibility is realized by increasing or decreasing the coding rate depending on the channel conditions. The code rate of the first transmission is very high and whenever the system is asked for a retransmission it sends additional redundancy lowering the received rate. Furthermore the received packets are combined in order to exploit the time diversity of the channel since several coded blocks may experience independent fading coefficients. This principle can be implemented by using Rate Compatible Punctured Codes (RCPC) [13] where the transmitter progressively punctures the same "mother" low-rate code. Moreover, since the first transmission is always made using a very high rate code, in bad channel conditions, a retransmission always occurs and the delay is penalized [15, 16, 17].

An information-theoretic approach to study simple HARQ protocols over a slotted multiple-access Gaussian channel with fading is given in [18]. There, the authors analyze the throughput and average delay as well as the asymptotic behavior with respect to various system parameters. In [13, 14, 19, 20] rate compatible punctured convolutional codes are used in the IR scheme. The transmission starts with the highest code rate and additional redundancy is sent whenever requested. In [21] the authors suggested the use of Turbo codes [1] for type-II HARQ protocols where the incremental redundancy can be obtained by puncturing the parity bits.

1.2.2 Hybrid ARQ with LDPC

Recently, we have seen increasing interest in the class of Low Density Parity Check (LDPC) codes [2, 22, 23, 24, 25]. Together with the particular tradeoff offered by Hybrid ARQ techniques, this motivates us to analyze the throughput of incremental redundancy schemes with this class of codes.

LDPC codes were first studied by Gallager in his thesis [2], where he introduced an iterative message passing decoding technique that approximates Maximum Likelihood (ML) decoding. In fact, ML decoding becomes too complex for LDPC codes as the block length grows, while the message passing algorithm, also called *Sum-Product Algorithm* (SPA) [25, 26], has a complexity per iteration that is linear in the block length.

The term "*low density*" refers to the fact that the number of 1's in each row of the parity matrix is small, in particular linear in the block length[1]. These codes exhibit a threshold phenomenon: as the block length tends to infinity, an arbitrarily small Bit Error Rate (BER) can be achieved if the SNR is larger than a certain threshold [22]. Otherwise, the BER is bounded away from zero for any number of decoder iterations.

The idea in SPA is to calculate approximate marginal *a posteriori* probabilities by applying Bayes' rule locally and iteratively. In the case when the graph representing the code has no cycle, the SPA computes exact marginal a posteriori probabilities.
The asymptotic performance of the message-passing decoder and in particular of the SPA, are evaluated by using a method called *Density Evolution* (DE) [27, 25, 22, 24], that allows to compute the value of the threshold[2]. The concentration theorem [22] guarantees that the threshold computed via density evolution coincides with the exact asymptotic threshold in the limit of large block-length.

DE gives information about the performances in terms of BER. However, the computation of the throughput is based on the frame error rate (FER) more than the BER. It is also known that these codes exhibit very good waterfall performance in terms of BER, but bad results in terms of FER [28, 29] unless countermeasures are adopted such as concatenation or special graph constructions. Therefore the following issues arise regarding a practical design of LDPC-based HARQ protocols:

- Analysis of the behavior of ideal LDPC codes (infinite block length) and generalization of DE under HARQ protocols and block fading channels.
- Analysis of the behavior of practical (finite length) LDPC compared with infinite length counterparts.
- Definition of good countermeasures in order to have such codes nicely adapted to the HARQ framework, thus obtaining close to ideal performance.
- Finally, should more sophisticated constructions based on special graph design [30, 31, 32, 33] be considered for practical implementation ?

Some of these problems are addressed in chapter 2.

[1] For random linear codes the expected number of 1's is proportional to n^2, [22].

[2] The threshold is defined as the worst channel parameter such that the message distribution evolves in a way that the associated probability of error converges to zeros as the number of iterations goes to infinity.

1.2.3 Coding and Retransmission for the Multicast Channel

In the downlink of wireless networks, the transmitter mostly deals with multiuser situations, where multiple terminals need to be served at once in an efficient and reliable way. This situation is in general referred to as "broadcast" channel, [34, 35, 36, 37]. Here, however, we limit the analysis for the multicast scenario where only common information is sent to all the users. In this case, existing retransmission protocols are in general not efficient, [38, 39, 40, 41]. Because the HARQ has to adapt to the conditions of all the users in the cell including the worst one, the retransmission protocol can possibly be solicited many times and the effective coding rate of the system will be very low. This means that strictly speaking these kinds of protocols are not scalable with the number of users.

Very recently and independently from us, Gopala et al. [42] have studied retransmission protocols in the same multicast setting as here. The authors analyze the scaling low of the throughput and delay with respect to the number of users, when SR and IR HARQ schemes are considered. However, when they analyze the SR protocol, they assume perfect channel state information both at the transmitter and at the receiver, while the IR scheme does not require channel state information. They compare these two schemes with a method based on cooperation among users. They show that the three policies have progressively increasing complexity but also better throughput/delay scaling lows. They show that the SR scheme achieves optimal throughput scaling low when the transmitter targets the user with the "average" conditions in the cell.

In our case, for both IR and SR, the transmitter is not informed about the channel coefficient of each user, since the feedback is very easy and gives only information about the correctness of the received packets.

In this scenario, multiple open issues arise again, which we address in chapter 3:

- How can these protocols be made "fully scalable? By "fully scalable we means that the delay does not increase with the number of users.
- How do these protocols scale with the number of users? What is the limiting behavior of these protocols with respect to system or design parameters?
- An open issue is whether these retransmission protocols are viable solutions for the multicast setting.

These problems are tackled in chapter 3.

1.3 CODING FOR MULTIMEDIA SOURCES

Modern telecommunications very often involve the transmission of analog sources over digital channels. Paramount examples are digital TV, audio broadcasting (DTV, DAB) and transmission of still and moving pictures over wireless radio channels in 3G (and beyond) mobile devices.

In contrast to the error-sensitive data applications mentioned before, such applications can, by nature, be much more delay-sensitive (especially streaming applications), but at the same time more loss-tolerant (within the requirements in terms of quality of reconstruction). Other multimedia applications can have more relaxed constraints on delay, such as for instance the transmission of images over mobile devices.

In such setting, bit-error probability at the output of the channel decoder is no longer a good measure of performance. On the contrary, the end-to-end distortion is more representative of the quality of transmission.

In some lucky sporadic cases, such as a Gaussian source over a Gaussian channel, both with the same bandwidth, it is well-known that "analog transmission" is optimal [3, 45]. By analog transmission we mean that the source is scaled in order to meet the transmit power constraint and then it is sent over the channel. This can be seen as regular analog AM. Such conditions require mainly a particular match between channel and source. For example it requires equal source and channel bandwidth. The source bandwidth is in general given by nature while the channel bandwidth has to respect certain requirements. Therefore, it can be interesting to try to maximize the spectral efficiency, given as the ratio between source and channel bandwidth, or equivalently try to optimize to system to meet a particular spectral efficiency. In order to achieve high spectral efficiency (> 1), in general the source must be compressed, and at the same time protected against errors introduced by the channel [3, 46].

1.3.1 Lossy Transmission over Compound Channel

Consider a Block-Fading Additive White Gaussian Noise channel (BF-AWGN) where the channel gain is random but constant over the duration of a codeword. Under the assumption that the transmitter is not informed about the channel fading of the user but only of its statistic, the BF-AWGN channel can model a broadcast channel with infinite users each of one experiencing a different fading coefficient. Hence, given a certain statistic of the SNRs (or of the users), it is desirable to design a single transmitter that performs "well" for a wide range of SNR.

While analog schemes show a gradual change (graceful degradation) in the received signal

quality with changes in SNR, digital schemes suffer from the "threshold effect". The system can be designed to achieve asymptotically optimal performance at a given target SNR, but they perform poorly for SNR below this target and they do not take advantage of better channel conditions when the actual SNR is above the target SNR.

In general digital schemes are designed based on Shannon's separation principle that states that no loss in performance is incurred when designing source and error coding schemes separately [47, 46]. However, this does not take into consideration complexity and delay and it does not hold for non-ergodic scenario or multiuser setting as considered here. Consequently, much research has been done based on the *Joint Source and Channel Coding* (JSCC) principle that links and jointly optimizes the source and the channel strategy. In [48, 49] and references therein, the authors have shown that joint source-channel codes can solve the problem above for fixed complexity and delay and they are more robust to change in channel noise.

Consider the case when the transmitter sends the same analog source to all the users each of one having a different channel condition. Possibly, the user, by exploiting the "goodness" of its channel, can reconstruct the source at different quality levels. The key issue is to define one transmission scheme that works as close as possible to the theoretically achievable limit for a wide range of SNR, i.e for a large number of users.

In this setting, different well known transmission strategies can be analyzed. The easiest scheme is based on time sharing, also called "progressive transmission" of information. The source is splitted into independent layers of information each mapped onto a different channel codeword possibly with different channel coding rate. The codewords are sent trough the channel by using a time sharing strategy. The splitting of the source into independent layers can be implemented by using an *ideal successive refinement* [50, 51, 46, 52] source encoder. The concept of successive refinement consists of first approximating the data by using a few bits of information and then iteratively improving the approximation as more and more information is supplied. Under particular condition on the source [51], the successive refinement source code achieves optimal performance (the rate-distortion function) at each level. Here we consider such an ideal successive refinement source code.

In [34] it is shown that a superposition-based transmission strategy lies on an achievable rate region of the broadcast channel, greater than the region achieved by time-sharing. Such a scheme consists in superimposing a low-rate information for the 'bad' user into a higher-rate information that can be decoded only by the users with better channel conditions. The same ideal successive refinement source encoder can be coupled with the superposition strategy where each layer of information is mapped into a different channel codeword. The codewords are then superimposed and transmitter over the channel.

The third strategy considers an hybrid system that couples the benefits of a digital system, with graceful degradation in reconstruction quality offered by an "analog" (uncoded) scheme. These kind of schemes, called Hybrid Digital-Analog (HDA), have been analyzed from a theoretical point of view in [53, 54, 55, 56]. In [56], for example, the authors design HDA schemes that achieve for one target SNR asymptotically optimal performance. However in a multiuser environment, it seems more interesting to optimize the system in view of having a performance level (average distortion) as smooth as possible over the range of possible SNR values.

Given this background, some issues exist that should be addressed:

- The definition of an optimization problem for the theoretical analysis of the three transmission strategies described before (time sharing, superposition, HDA) is a key issue. The optimization problem is based on the minimization of the average distortion, where the average is done over the distribution of the SNRs. This becomes a power and rate allocation problem that allows for comparisons in terms of distortion versus instantaneous signal to noise ratio.

- The definition of algorithms that find the optimal allocation power and rate policies subject to total power and spectral efficiency constraints give guidelines for the practical construction of these systems.

- A big issue is, of course, the construction of codes (source code/channel code) that can approach the theoretical limit.

These problems are addressed in chapter 4.

1.3.2 Separated vs Joint Source-Channel Coding: Code Construction

Among the strategies briefly described in the previous section, the HDA scheme gives the best results, in the sense that it achieves smooth reconstruction quality for a wide range of SNRs. These schemes are made of a digital part (source-channel encoder) superimposed with the analog (uncoded) signal. We refer to the digital part as to "tandem encoder". The result of the optimization problem is the SNR threshold at which the digital code should be design to work. The analytical development considers an ideal source and channel code when the rate-distortion and the capacity-cost function are achieved. However, a key open problem is the construction of codes that work as close as possible to the theoretical limit. The tandem encoder can be designed separately or jointly.

Let us consider, first, the standard separated approach. In order to allow source compression, practical source encoders involve some linear transformation (e.g., Fourier or Wavelet subband decomposition), followed by some segmentation and decimation [3, 57]. Then the analog data may be quantized and the sequence of quantization indexes is lossless-compressed by an "entropy coding" stage, usually implemented by some form of adaptive arithmetic coding [58]. The best results known so far, in terms of quantization of memoryless sources are found in the family of Entropy Constrained Trellis Coded Quantizers (ECTCQ) [59]. These schemes uses the expanded signal set [60] and set partitioning ideas from coded modulation. The probability with which the points in the code alphabet are selected is not uniform. Hence, rate improvements can be achieved by concatenating this scheme with an arithmetic encoder and eventually a channel code to protect against the effects of the channel. Because of the variable-length coding stage, the source decoder is not robust to residual channel errors and a few wrong bits at its input may cause intolerable degradation of the reproduced source. This is the main weakness of Shannon's source-channel separation theorem. Moreover, the separation theorem does not generally hold in a non-ergodic environment such as the slowly fading AWGN channel. This leads to the analysis of the joint approach (JSCC), that links and jointly optimizes the source and the channel strategy. In [48, 49] and references therein, the authors have shown that JSCC can solve the problem above for fixed complexity and delay and they are more robust to change in channel noise. Different approaches for JSCC has been proposed so far, but we can summarize them mainly into two groups. The first tries to optimize the quantization step by deleting all the redundancy of the source. These schemes can be coupled with standard channel codes. The main issue here is to construct quantizer schemes such that the reconstruction quality relies on BER more than FER, i.e few bits in error at the output of the channel decoder do not have a catastrophic effect on the reconstruction quality. Examples of this group are Channel Optimized Scalar/Vector Quantizer [61, 62, 63].

The other group of JSCC schemes considers very easy quantizer scheme and the data compression and channel coding stage is performed jointly. This last stage achieves compression ratio by exploiting the residual redundancy at the output of the quantizer. However, the decoder needs the a priori information about the statistic of the indexes at the output of the quantizer. The joint data compression-channel coding stage can be efficiently implemented via linear codes, as shown in [64, 65].

The following open problems are addressed in chapter 5.

- Existing practical tandem encoder are not very close to the theoretical performance. However, source codes (ECTCQ) and channel codes (Turbo codes, LDPC) that perform very close to the ideal rate-distortion or capacity-cost functions exists. Their potential is still to be explored when jointly optimized.

- The design of non-catastrophic quantizer, i.e robust to channel errors is a key issue. A challenging features is the independence of performance from the source statistic. In practice, in fact the statistic of the source is not precisely known.
- Data compression/channel coding schemes based on linear codes have, as well, a big potential. An open problem is the design and optimization of codes for this compression/protection method.
- An open issue is whether soft reconstruction can be helpful when the compression method based on linear codes is considered.
- Parallel Concatenated Turbo codes are particularly suited for JSCC because by nature they are systematic and different coding rates can be achieved by puncturing the same mother code. However, analytical optimization is not straightforward. It could be interesting to optimize families of codes like LDPCs or Irregular Repeat and Accumulate (IRA) codes [66], where DE reveals all its advantages. In this case, modification of DE to take into account the compression scheme should be carried out.
- The analysis and study of the performance achieved by the best ECTCQ coupled with compression scheme based on optimized IRA codes is of great interest.
- An interesting point is to adapt this compression scheme in such a way that it can provide a graceful degradation of performance and thus can be adapted to the multiuser setting, also without the use of the analog signal.

1.4 CONTRIBUTIONS

We summarize our results around the two following axes of research: HARQ-based transmission over wireless channels and efficient transmission of multimedia content. The contributions in terms of publications are specified.

Throughout the chapters less intuitive acronyms are repeated, for the sake of clarity.

1.4.1 HARQ-based Transmission over Wireless Channels

(i) Chapter 2 .
The work is mainly inspired by [18]. By using *Renewal Reward* [67] theory, the throughput of LDPC codes ensembles with incremental redundancy protocol over slow fading channel

is studied, and the density evolution method is extended to the case of block fading channel and general retransmission protocols. The analysis shows that assuming infinite block length, LDPC codes yield almost the same performance as random binary codes.

As expected, the throughput performance of practical finite-length LDPC codes show a considerable loss with respect to the ideal behavior of the ensemble. Two original methods are presented and are shown to recover most of the gap between ideal performance and practical results. The first method is based on a simple special graph arrangement and the second is based on an outer selective-repeat protocol acting on smaller packets of information bits.

Finally we analyze the complexity of the IR protocol with LDPC codes. When a packet hits a deep fade, the iterative decoder may perform many iterations without converging to small error probability. This is a waste of computation time and battery energy. Ideally, it would be useful to trigger the iterative decoder only if the probability of successful decoding is high. A method based on the asymptotic analysis (density evolution) is shown to provide an asymptotic region of convergence of the IR scheme and provide some saving in complexity compared to standard stopping criteria.
This work led to the publications [68, 69, 70, 71] and two patents .

(ii) Chapter 3 .
In this chapter the scalability of the HARQ protocol is addressed. A multicast scenario is considered, where each user spans a fixed number of fading blocks. The results are given in terms of throughput per user vs number of users, when the IR and SR protocols are considered. These protocols, strictly speaking, are not scalable, i.e the average delay grows to infinity as the number of users augment. HARQ (SR or IR) can be made *fully* scalable if we allow for a fraction $x > 0$ of users that do not decode successfully. One of our results is that the IR scheme achieves asymptotically, for large number of users, performance greater or equal than the ergodic capacity of the system for fixed positive x. For the SR protocol, the optimal performance is always achieved with finite average delay, but at the expense of a penalty in throughput compared to the IR scheme. Moreover we show that the performance of IR and of FEC coding are identical in terms of delay, throughput and error probability, in the limit of a large number of users. Finally we target a streaming application and we give a simple practical example on the requirement in terms of buffer size at the receiver, based on the Birth-Death queue process. A part of this work gave rise to [72].

1.4.2 Efficient Transmission of Multimedia Content

(iii) Chapter 4 .
In this chapter, we define the optimization problem for three systems. The first is obtained by coupling an ideal successive refinement source coder with strategy based on time sharing and the second by coupling the same source encoder with a transmission scheme based on superposition. These two fully digital schemes are compared with the optimized HDA scheme based on superposition of the digital and the analog part. The function to be minimized is the average distortion under the transmit power constraint and total spectral efficiency. The algorithms gives the optimal power and rate allocation based on these constraints, as well as the optimal number of layers. We suppose, ideally, that the successive refinement scheme is able to provide independent levels of information and that it achieves the rate-distortion function at each level. For the sake of theoretical tractability ideal channel codes are considered. However, the algorithms can be generalized to practical schemes where the source and channel encoder are not ideal.
This work is presented in [73, 74].

(iv) Chapter 5 .
This chapter deals with the construction of practical coding schemes that approach the limits found in the previous chapter. We analyzed and compare two different schemes. The first belongs to the class of robust quantizer schemes that performs also data compression. Belonging to this class, we analyze a Multistage Trellis Quantizer (MTQ) based on a spherical dithering and on the scaled version of a "mother" convolutional (de)coder. The idea is to approximate the behavior of spherical codes with the convolutional code. In fact Lapidoth in [75] has shown that scaled spherical codes with minimum distance encoding are *robust* in the sense that they achieve the Gaussian rate distortion bound under very mild conditions on the source. This scheme, in the noiseless case, achieve performance close to the Gaussian rate distortion bound, independently from the statistic of the source. It is robust to channel errors because it inherits the property of the convolutional encoder to be non catastrophic. Moreover, its output is almost non redundant and thus, it is suited for concatenation with powerful standard Turbo codes or LDPC.

The second scheme that we analyze is based on the joint implementation of data compression and channel coding. In particular this can be achieved via linear codes as Turbo codes. In the following we refer to it as Multilevel Turbo COMpression (M-TCOM). The M-TCOM exploits the residual redundancy of the index at the output of the quantizer. Here, we consider a simple Entropy Constrained Scalar Quantizer (ECSQ) whose output indexes are redundant. The redundancy of these indexes is used as a-priori information to achieve

compression rate and drive the turbo decoder, this is called “source aided channel decoding” [64]. The indexes belonging to a $Q-$ary are mapped through a multilevel decomposition (bit planes) onto turbo codes that act on a per-level basis. The systematic bits of each turbo codes are punctured, together with a certain amount of parity bits in order to achieve the desired rate. A time-sharing approach is used for transmission over the channel.

This scheme outperforms standard concatenations that consider quantizer, entropy encoder and standard channel code. Results are given in terms of code optimization (polynomial generator of component convolutional codes and puncturing pattern). However, the analysis of ECTCQ concatenated with turbo compression and IRA codes, as well as the analytical optimization of such codes, is an on-going work.

This approach, by nature, can handle progressive transmission of information. Thus, by choosing the Q-ary to binary mapping such that it is embedded, the source can be reconstructed with different levels of distortion.

This scheme and the design of the coding rate are extended to the case of practical transmission of images over the wireless link. This is shown to give remarkable results when coupled with a modified Differential Pulse Code Modulation quantizer defined by Kim et al [57].
This work has partly led to [76, 77]. A comprehensive analysis and summary of the results as well as the on-going work contributes to [78]. The extensions to embedded quantization and the application to DPCM-based quantizer scheme contributes to [79].

CHAPTER 2

Throughput of Hybrid ARQ protocols with LDPC Codes

2.1 INTRODUCTION

This chapter is focused on the concept of reliability in the context of packet data transmission. Typically, data transmission is not strictly delay-sensitive but requires a virtually error-free link. In order to provide such level of reliability over wireless channels, affected by propagation impairments such as fading, ARQ schemes can be combined with channel coding (HARQ). In brief, when fading varies slowly over the duration of a codeword, coding takes care of the channel noise while retransmissions take care of bad channel conditions (deep fades).

This work is mainly inspired by [18] where the authors analyze, from an information theoretic point of view, throughput and average delay performance of some HARQ protocols over a slotted multiple access Gaussian channel with fading. The analysis is carried out by considering Gaussian codes for the shake of mathematical tractability.
Here we consider a point-to-point downlink scenario and we analyze the performance of HARQ schemes with the powerful class of LDPC codes, over block fading channel. Although very idealized, the simple block-fading model captures several aspects of wireless communications over fading channels (see the thorough discussion in [80], [81]). For example, this model applies to narrow-band transmission over a multipath fading channel with

slow frequency hopping (e.g., a GSM/GPRS system [82]). As illustrated in [80, 81], when fading is slowly-varying with respect to the duration of a codeword, each codeword experiences a fixed number of fading states (say M values). Under the realistic assumption of large number of dimension per block L, and small M,[1] the channel is not *information stable* and outage capacity[2], rather than standard ergodic capacity, describes the limits of reliable communications.

The analysis of the Incremental Rredundancy (IR) scheme with LDPC code ensembles with iterative belief-propagation decoding shows that, assuming infinite block length, LDPC codes yield almost optimal performance.

Unfortunately, practical finite-length LDPC codes incur a considerable performance loss with respect to their infinite-length counterpart. In order to reduce this performance loss, two effective methods are proposed : 1) using special LDPC ensembles designed to provide good *frame-error rate* (rather than just good *iterative decoding threshold*); 2) using an outer selective-repeat protocol acting on smaller packets of information bits. Surprisingly, these two apparently very different methods yield almost the same performance gain and recover a considerable fraction of the optimal throughput, thus making practical finite-length LDPC codes very attractive for data wireless communications based on incremental redundancy HARQ schemes.

In the last part of the chapter, the analysis of the complexity of HARQ scheme coupled with Belief Propagation decoding is carried out. When a packet hits a deep fade, the iterative decoder may perform many iterations without converging to small error probability. Eventually, a decoding failure is declared and a retransmission is requested. Therefore, when such event occurs, the iterations represent wasted computation time. Ideally, it would be useful to detect quickly whether the packet is likely to be correctly decoded or not, and trigger the iterative decoder only if the probability of successful decoding is high. While this is very easily obtained in simple ARQ protocols, where each packet is independently encoded and decoded and each retransmission is treated as a newly received packet, it is not so obvious in more sophisticated HARQ schemes that make use of packet combining [11] or incremental redundancy. In fact, when a data packet is LDPC-encoded and the resulting codeword is sent across the channel on a single fading block, it is sufficient to check if the instantaneous signal-to-noise ratio (SNR) at the receiver is larger than the LDPC iterative decoding threshold [22, 23], to know if the packet can be decoded successfully with high probability. On the contrary, if a codeword is transmitted over, say, m fading realizations,

[1] For example, in GSM $M = 8$ and $L \approx 100$, and in 64kbps downlink reference data channel for UMTS data-transmission modes, codewords are interleaved over $M = 2$ frames, and each frame may contains up to ≈ 1000 dimensions [83].

[2] Outage capacity is defined as the maximum information rate that can be achieved in any fading condition during non-outage

we would need an m-dimensional region of convergence such that if the SNR vector is in this region, then the iterative decoder is successful with high probability. Characterizing such multidimensional region of convergence for a given LDPC code and iterative decoder is not an easy task in general. A simple method to compute implicitly an *approximated* region of convergence of the belief-propagation decoder for a given LDPC code is presented. This method is based on the use of Density Evolution (DE). It is possible then, to check very efficiently if the vector of received SNRs is such that successful decoding is expected with high probability. Then, the iterative decoder is triggered only if the vector of received SNRs is in the region. This reduces the expected complexity of the decoder, with very important savings in both computation time and battery energy.

2.1.1 Summary of the Contributions

- Analysis of the throughput of IR schemes with infinite length LDPC codes and extension of density evolution method to the case of block fading channel and IR scheme.
- Countermeasures to recover the gap between the performance of infinite length LDPC ensemble and that achieved by finite length practical codes.
- Analysis of the complexity of IR scheme with Belief Propagation decoding.
- Definition of a approximate convergence region to lower the complexity of the decoder

2.1.2 Organization of the work

The rest of the chapter is organized as follows: in section 2.2 the system model and the retransmission protocol is introduced. Section 2.3 recalls the throughput analysis based on Renewal-Reward theory, while in section 2.4.2 and 2.5 the computation is particularized for random binary and LDPC codes. Section 5-22 give results in terms of throughput achievable by LDPC when ideal conditions are considered (infinite block length) and in section 2.7 two methods to fill in the gap between finite length and infinite length LDPC codes are given. The results are shown in section 2.8. Finally the complexity of LDPC decoding under IR retransmission protocol is analyzed in section 2.9 and section 2.10 concludes the chapter.

2.2 PROTOCOL AND MODEL USED

The system is composed by one transmitter and one receiver. Time is divided into *slots* each of duration T. In each slot the transmitter sends $L \simeq WT$ dimensions, where W is the two-sided signal bandwidth and we assume $WT \gg 1$. The fading is considered slowly time varying, in particular constant block fading on each slot. Moreover the channel gains over different slots are assumed statistically independent (figure 2-1). Denote $\mathbf{x}_s$ the transmitted

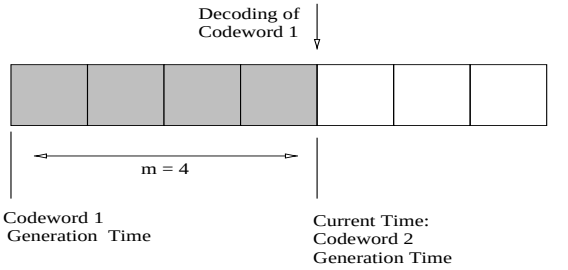


Fig. 2-1. Model representing the division of the codewords in bursts.

signal, $\mathbf{y}_s$ the received signal and $\boldsymbol{\nu}_s$ the background noise, during slot s,

$$\begin{aligned} \mathbf{x}_s &= (x_{s,1},\, x_{s,2},\, \ldots,\, x_{s,L}) \\ \mathbf{y}_s &= (y_{s,1},\, y_{s,2},\, \ldots,\, y_{s,L}) \\ \boldsymbol{\nu}_s &= (\nu_{s,1},\, \nu_{s,2},\, \ldots,\, \nu_{s,L}) \end{aligned} \tag{2-1}$$

The noise is assumed circularly symmetric Gaussian with i.i.d components with variance 1. The energy per symbol is constant and given by $\left[\|x_{s,l}\|^2\right] = 1$. The fading coefficient is normalized so that $\mathbb{E}[|c_s|^2] = 1$. The average received SNR is given by $\Gamma \stackrel{\Delta}{=} E_s/N_0$. For later use, we define also the fading power gain $\alpha_s \stackrel{\Delta}{=} |c_s|^2$ and the instantaneous received SNR over slot s, $\beta_s \stackrel{\Delta}{=} \alpha_s\Gamma$. The received signal over one slot is given by:

$$\mathbf{y}_s = \sqrt{\Gamma}\, c_s\, \mathbf{x}_s + \boldsymbol{\nu}_s \tag{2-2}$$

In the following we suppose that the decoder has a perfect knowledge of the channel gain c_s and of the SNR β_s.

The HARQ scheme under analysis is shown in figure 2-2. Roughly speaking, the transmitter keeps sending additional coded symbols (redundancy) until successful decoding is achieved.

For this reason, it is referred to as incremental redundancy protocol. The transmitter encodes information messages of b bits by using a channel code with codebook $\mathcal{C} \in \mathbb{C}^n$ of length $n = LM$ and coding rate $R = b/n$ bit/symbol. The codewords are divided in M blocks of length L symbols. Each block is sent over one slot. Let $\mathcal{C}_m$ denote the punctured code of length Lm obtained from $\mathcal{C}$ by "deleting" the last $M - m$ blocks. Without loss of generality, we enumerate the slots as $s = 1, 2, \ldots, M$. In order to transmit the current codeword, the transmitter sends the first block of L symbols on slot $s = 1$. The receiver decodes the code $\mathcal{C}_1$, by processing the corresponding received signal $\mathbf{y}_1$. If decoding is successful, a positive ACKnowledgment (ACK) is sent on a delay-free error-free feedback channel, the transmission of the current codeword is stopped and the transmission of the next codeword will start in the next slot (say, $s = 2$). On the contrary, if a decoding error is detected, a Negative ACK (NACK) is sent back and the next block of the current codeword is transmitted on slot $s = 2$. In this case, the receiver decodes $\mathcal{C}_2$ by processing the received signal $\{\mathbf{y}_1, \mathbf{y}_2\}$ and the same ACK/NACK procedure is repeated, until either successful decoding occurs, or all M blocks of the current codeword are transmitted without successful decoding (see figure 2-2).

If successful decoding occurs after $m \leq M$ blocks, the effective coding rate for the current codeword is $\frac{r}{m}$ bit/symbol, where we define the rate of the first block as $r \stackrel{\Delta}{=} b/L$. Therefore, the IR protocol implements easily an adaptive rate scheme that takes advantage of good instantaneous channel conditions. The throughput of the IR protocol is defined as the average number of bit/s/Hz successfully received. As far as the throughput is concerned, it is irrelevant whether codewords not successfully decoded after M blocks are retransmitted in some successive slots or if they are just discarded [18]. On the contrary, the packet loss rate and the average delay of the system are affected by the policy for handling decoding-failures. In general, for some information packet arrival model and some delay constraint we might seek a policy minimizing the delay subject to a packet loss probability constraint. This topic is out of the scope of this work and, for simplicity, we shall assume that the transmitter has an infinite number of information packets available (no packet arrival process) and applies the IR procedure to the current packet until decoding is successful or until a rate constraint violation happens.

2.3 THROUGHPUT ANALYSIS

In order to study the throughput the following optimistic assumptions are taken into account:

- The transmitter has an infinite number of messages to be sent.
- The ACK/NACK channel is delay and error free.

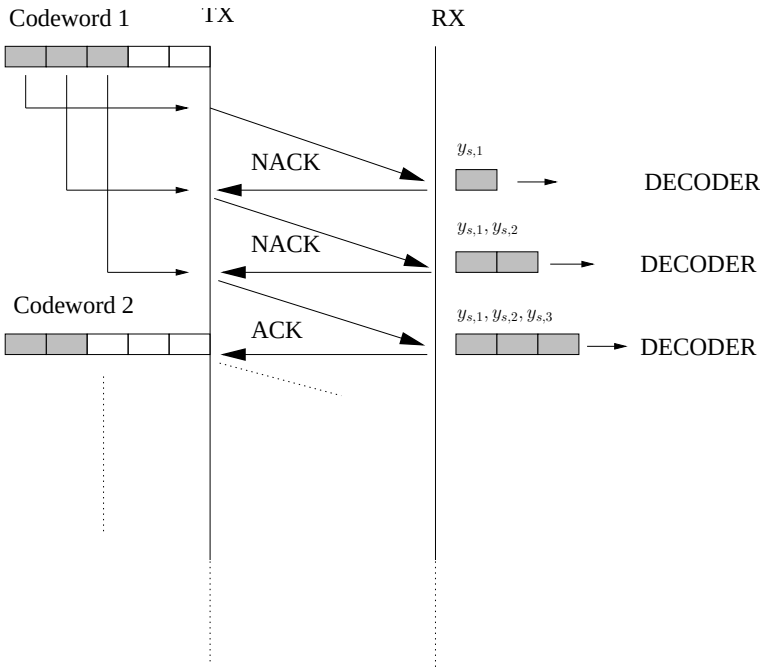


Fig. 2-2. Protocol HARQ

- The channel power gains α_s are i.i.d random variables for all the slots.

The throughput, expressed in bits per second per hertz is given by:

$$\eta = \lim_{t\to\infty} \frac{r(t)}{t} = \tag{2-3}$$

where $r(t) = \frac{b(t)}{L}$, t counts the number of slots and $b(t)$ the number of information bits successfully decoded up to slot t. As in [18] the throughput can be expressed using the *Renewal-Reward Theorem*, [67]. The event $\mathcal{E} = \{$The user stops transmitting the current codeword$\}$ is recognized to be a *recurrent event*.

A random *reward* $\mathcal{R}$ is associated to the occurrence of the recurrent event: $\mathcal{R} = r$ b/s/Hz if transmission stops because successful decoding and $\mathcal{R} = 0$ b/s/Hz if it stops because at step M it is not possible to successfully decode (violation of the rate constraint). Applying the Renewal Theorem we obtain $\eta = \frac{\mathrm{E}[\mathcal{R}]}{\mathrm{E}[\tau]}$ where τ is the random time, expressed in number of bursts, between two consecutive occurrences of the recurrent event. It is referred to as *inter-renewal time*. Define the event $\mathcal{A}_m = \{$successful decoding with m transmitted bursts$\}$, and $q(m)$ as the probability of having the first successful decoding at step m. The probability $q(m)$ can be expressed as:

$$\begin{aligned} q(m) &= \Pr\big(\overline{\mathcal{A}}_1, \overline{\mathcal{A}}_2, \ldots, \overline{\mathcal{A}}_{m-1}, \mathcal{A}_m\big) \\ &= \Pr\big(\overline{\mathcal{A}}_1, \overline{\mathcal{A}}_2, \ldots, \overline{\mathcal{A}}_{m-1}\big) - \Pr\big(\overline{\mathcal{A}}_1, \overline{\mathcal{A}}_2, \ldots, \overline{\mathcal{A}}_m\big) \\ &= p(m-1) - p(m) \end{aligned} \tag{2-4}$$

with $p(m) = \Pr\big(\overline{\mathcal{A}}_1, \overline{\mathcal{A}}_2, \ldots, \overline{\mathcal{A}}_m\big) = 1 - \sum_{i=1}^{m} q(m)$. A reward $\mathcal{R} = b$ is obtained when successful decoding occurs. This happens at step m-th with a probability $q(m)$. It follows that

$$\mathrm{E}[\mathcal{R}] = \sum_{m=1}^{M} r\, q(m) = r \sum_{m=1}^{M} q(m) = r\,[1 - p(M)] \tag{2-5}$$

The inter-renewal time is a random variable that takes the values m with probability $q(m)$ for $m = 0, \ldots, M$. In the case $m = M$ the transmission can stop not only because of successful decoding (it occurs with probability $q(M)$) but also because of the rate constraint (the decoding is not successful but the process stops because the complete codeword has already been sent). This occurs with probability $p(M)$. Finally the probability mass function is given by

$$f_\tau(m) = \Pr(\tau = m) = \begin{cases} q(m) & \text{if } m < M, \\ q(M) + p(M) & \text{if } m = M. \end{cases} \tag{2-6}$$

It follows that

$$\begin{aligned}\mathrm{E}[\tau] &= \sum_{m=1}^{M} m\,q(m) \;+\; M\,p(M) \\ &= \left(\sum_{m=1}^{M} m\,p(m-1) \;-\; \sum_{m=1}^{M} m\,p(m) \;+\; M\,p(M)\right) \\ &= \sum_{m=0}^{M-1} p(m) = 1 + \sum_{m=1}^{M-1} p(m) \end{aligned} \tag{2-7}$$

where the last step follows after some algebra and by setting $p(0) = 1$. Finally the throughput has the following expression

$$\eta \;=\; \frac{r\,(1 \;-\; p(M))}{\sum_{m=0}^{M-1} p(m)} \;=\; R\,M\,\frac{1 \;-\; p(M)}{1 + \sum_{m=1}^{M-1} p(m)} \tag{2-8}$$

The average delay (in slots) can be obtained either by simple direct calculation, or by noticing that the IR scheme where, in the presence of a decoding failure after M slot, the protocol is reset and the current codeword is transmitted again, corresponds to a newly defined renewal-reward process with deterministic reward RM. Therefore, from (2-8) it follows that the average inter-renewal time (i.e., the average delay) of this new process is clearly given by

$$\mu \;=\; \frac{1 + \sum_{m=1}^{M-1} p(m)}{1 \;-\; p(M)} \tag{2-9}$$

The variance of delay can be obtained by direct calculation and it yields:

$$\begin{aligned}\sigma_H^2 \;=\; & \frac{1 \;+\; \sum_{m=1}^{M-1} p(m) \;+\; 2\sum_{m=1}^{M-1} mp(m) \;-\; \left[1 \;+\; \sum_{m=1}^{M-1} p(m)\right]^2}{1 - p(M)} \;+ \\ & + \left(1 \;+\; \sum_{m=1}^{M-1} p(m)\right)^2 \frac{p(M)}{(1 - p(M))^2} \end{aligned} \tag{2-10}$$

2.4 THROUGHPUT BOUNDS: INFINITE LENGTH CODES

2.4.1 Conventional Coded ARQ

We take a short detour to compute the throughput of conventional ARQ schemes; this will be used in section 5-22 to motivate the effectiveness of IR with respect to these conventional

protocols. We shall consider two variants of conventional coded ARQ. In the first case, codewords of length L and rate $R = b/L$, spanning a single fading block, are used for transmission. In the presence of a decoding error (detected with arbitrarily large probability in the limit of large L, [18]), the codeword is retransmitted in some successive slot. Using the same arguments as before we can compute the throughput, the average delay (in slots) and the variance of the delay of this scheme; they are clearly given by

$$\begin{aligned} \eta_{\mathrm{SR}-1} &= R(1-p(1)) \\ \tau_{\mathrm{SR}-1} &= \frac{1}{1-p(1)} \\ \sigma^2_{\mathrm{SR}-1} &= \frac{p(1)}{(1-p(1))^2} \end{aligned} \tag{2-11}$$

where the subscript "SR-1" indicates *selective repeat* with coding over one block.

In the second case, codewords of length $n = LM$ and rate R are transmitted over M fading blocks and decoding is performed only after all M blocks are received. In the presence of a decoding error, the codeword is retransmitted in some successive group of M slots. The resulting throughput, average delay and variance of the delay are given by

$$\begin{aligned} \eta_{\mathrm{SR}-\mathrm{M}} &= R(1-p(M)) \\ \tau_{\mathrm{SR}-\mathrm{M}} &= \frac{M}{1-p(M)} \\ \sigma^2_{\mathrm{SR}-\mathrm{M}} &= M^2\frac{p(M)}{(1-p(M))^2} \end{aligned} \tag{2-12}$$

The subscript "SR-M" indicates *selective repeat* with coding over M blocks. It is immediate to see that $\eta_{\mathrm{SR}-\mathrm{M}} \leq \eta$ and $\mu_{\mathrm{SR}-\mathrm{M}} \geq \mu$ where η and μ are the throughput and average delay of the IR scheme given in (2-45) and (2-9).

In Section 5-22, we show by some examples that the IR scheme performs much better than the above SR schemes in terms of maximum throughput. The average delay and the variance of the delay are comparable for small values of throughput.

2.4.2 Random Binary (RB) Codes

Recall that we assume perfect channel knowledge at the receiver, i.e., the receiver knows perfectly the fading coefficients $\{c_s : s = 1, \ldots, M\}$. Let the *instantaneous* mutual information per input symbol on slot s be given by

$$J(\beta_s) \triangleq I(\mathbf{x}_s; \mathbf{y}_s|c_s) = \frac{1}{L}\mathbb{E}\left[\log_2 \frac{p(\mathbf{y}_s|\mathbf{x}_s, c_s)}{p(\mathbf{y}_s|c_s)}\right] \tag{2-13}$$

where $\mathbf{x}_s$ is distributed according to some input distribution $Q(\mathbf{x})$ and where

$$p(\mathbf{y}|\mathbf{x},c) = \frac{1}{(\pi N_0)^L} e^{-\frac{1}{N_0}|\mathbf{y}-c\mathbf{x}|^2}$$

is the channel transition pdf for given fading gain c. Given the sequence of fading gains $\mathcal{F}_m \stackrel{\Delta}{=} \{c_s : s = 1, \ldots, m\}$, we define the conditional probability of decoding error after m received slots $\Pr(\text{error}|\mathcal{F}_m, \mathcal{C}_m)$ given the code $\mathcal{C}$ and the fading sequence $\mathcal{F}_m$. In [18] it is shown that there exist families of codes $\mathcal{C}$ with increasing block length L such that

$$\lim_{L\to\infty} \Pr(\text{error}|\mathcal{F}_m, \mathcal{C}_m) = 0 \tag{2-14}$$

if $I_m \stackrel{\Delta}{=} \sum_{s=1}^{m} J(\beta_s) > r$. Moreover, for any L the error probability of any code is bounded away from zero if $I_m < r$. Finally, assuming typical-set decoding [46] the conditional probability of an undetected decoding error vanishes as $L \to \infty$ for any code $\mathcal{C}$ and any fading sequence $\mathcal{F}$.

Eventually, we can say that for large number of dimensions per slot L (i.e., large product WT) the error probability of the best possible code at each IR step m, for given fading sequence $\mathcal{F}_m$, is given by $\Pr(\text{error}|\mathcal{F}_m, \mathcal{C}_m) = 1\{I_m \le r\}$ where $1\{.\}$ is the indicator function. Hence, the average error probability (where average is with respect to the fading statistics), is given by

$$\Pr(\text{error}|\mathcal{C}_m) = \Pr(I_m \le r) \tag{2-15}$$

We define the probability $q(m)$ of successful decoding with m transmitted slots as

$$\begin{aligned} q(m) &\stackrel{\Delta}{=} \Pr(I_1 \le r,\, I_2 \le r,\, \ldots,\, I_{m-1} \le r,\, I_m > r) \\ &= p(m-1) - p(m) \end{aligned} \tag{2-16}$$

where $p(m)$ is defined as

$$p(m) \stackrel{\Delta}{=} \Pr(I_1 \le r,\, I_2 \le r,\, \ldots,\, I_m \le r) = 1 - \sum_{i=1}^{m} q(m) \tag{2-17}$$

In the random binary codes case the input distribution $Q(x)$ puts uniform probability on the binary antipodal alphabet $\{-\sqrt{E}, \sqrt{E}\}$. Because of non-negativity of mutual information, the sequence $(I_1, I_2, \ldots, I_m)$ is a non-decreasing sequence for all fading sequence realization. This yields

$$p(m) = \Pr(I_1 \le r, \ldots, I_m \le r) = \Pr(I_m \le r) = \Pr\left(\sum_{s=1}^{m} J(\beta_s) \le r\right) \tag{2-18}$$

For binary inputs the instantaneous mutual information $J(\beta_s)$ is given by

$$J(\beta_s) \;=\; 1 - \int_{-\infty}^{\infty} \log_2\Big(1 \;+\; e^{4\sqrt{\beta_s}\left(z-\sqrt{\beta_s}\right)}\Big)\, \frac{e^{-z^2}}{\sqrt{\pi}}\mathrm{dz} \tag{2-19}$$

Since the β_s's are i.i.d. random variables, the cumulative distribution function (cdf) (2-18) is obtained from the m-fold convolution of the probability density function (pdf) of $J(\beta_s)$, given by

$$f(x) = \frac{1}{\gamma} f_\alpha\left(J^{-1}(x)/\gamma\right)\left(\frac{\mathrm{d}J^{-1}(x)}{\mathrm{d}x}\right) \tag{2-20}$$

where $f_\alpha(x)$ is the pdf of the fading power gain α. The probability $p(m)$ is then:

$$p(m) = \int_0^r f_m(x)\mathrm{dx} \tag{2-21}$$

where $f_m(x) = \mathcal{F}^{-1}\{(\mathcal{F}\{f(x)\})^m\}$ and $\mathcal{F}\{.\}$ indicates the Fourier Transform.
In order to reduce the computation complexity of 2-21, for large m we can resort the Gaussian approximation or the Chernoff bound.

(v) Gaussian Approximation Using the Central Limit Theorem [84], for large m we have that

$$\frac{1}{\sqrt{m\sigma_G^2}} \sum_{i=1}^{m} I(\beta_i) \;-\; \sqrt{m}\frac{\mu_G}{\sigma_G} \underset{\text{distrib}}{\rightarrow} \mathcal{N}(0,1) \tag{2-22}$$

where

$$\mathrm{E}[I(\beta_i)] \;=\; \mu_G, \quad \mathrm{E}\Big[\big(I(\beta_i) \;-\; \mu_G\big)^2\Big] \;=\; \sigma_G^2$$

are the mean and the variance of the single letter mutual information $I(\beta_i)$. The value of $p(m)$ is then found using the integration of the cdf of the Gaussian RV with mean $m\mu_G$ and variance $m\sigma_G^2$:

$$p(m) \approx\; 1 \;-\; \mathrm{Q}\left(\frac{r - m\mu_G}{\sqrt{m}\sigma_G}\right) \tag{2-23}$$

for large m.

(vi) Chernoff Bound The Chernoff bound gives an upper bound on the probability $p(m)$:

$$\begin{aligned} p(m) &= \Pr\left(\sum_{i=1}^{m} I(\beta_i) \leq r\right) \\ &= \Pr\left(\sum_{i=1}^{m} I(\beta_i) - r \leq 0\right) \\ &= \mathrm{E}\left[\mathcal{I}\{\sum_{i=1}^{m} I(\beta_i) - r \leq 0\}\right] \end{aligned} \tag{2-24}$$

where $\mathcal{I}\{\mathcal{X}\}$ is the indicator function that is equal to 1 when the event $\mathcal{X}$ is verified and 0 otherwise.

This is upper-bounded by a negative exponential:

$$\begin{aligned} p(m) &\leq \mathrm{E}\left[e^{-\lambda\left(\sum_{i=1}^{m} I(\beta_i) - r\right)}\right] \\ &= e^{\lambda r}\, \mathrm{E}\left[e^{-\lambda \sum_{i=1}^{m} I(\beta_i)}\right] \\ &= e^{\lambda r} \prod_{i=1}^{m} \Phi_{I_i}(\lambda) \end{aligned} \tag{2-25}$$

where we have defined $\Phi_{I_i}(\lambda) = \mathrm{E}\left[e^{-\lambda I(\beta_i)}\right]$. The last step is due to the fact that the $I(\beta_i)$ are i.i.d RVs. Finally the upper bound is given by

$$p(m) \leq e^{\lambda r}[\Phi_{I_i}(\lambda)]^m \tag{2-26}$$

The minimization of λ gives tighter upper bound,

$$p(m) \leq \min_{\lambda} e^{\lambda r}[\Phi_{I_i}(\lambda)]^m \tag{2-27}$$

In this case the function $\Phi_{I_i}(\lambda)$ does not have a closed form solution, so we got the results through numerical simulations.

2.5 Low Density Parity Check Codes

In chapter 1 (LDPC) codes have been introduced together with the key properties and the concepts of iterative decoding algorithm based on message passing used to approximate maximum likelihood decoding, as well as the DE method used to evaluate bit error rate

performance in the limit of large blocklength. The message-passing decoder is called Sum Product Algorithm (SPA) or Belief Propagation algorithm (BP). In the following we refer to it as BP.

The parity-check matrix of a randomly selected instance $\mathcal{C}$ in a given LDPC ensemble is conveniently represented by a bipartite graph with the nodes on the left (bitnodes) corresponding to the coded symbols and the nodes on the right (checknodes) corresponding to parity-check equations. A bitnode v is connected to a checknode c if the corresponding v-th symbol participates in the c-th parity equation. The LDPC ensemble is defined by its left and right degree distributions $\lambda(x) \triangleq \sum_{i=2}^{d_v} \lambda_i x^{i-1}$ and $\rho(x) \triangleq \sum_{i=2}^{d_c} \rho_i x^{i-1}$, where λ_i (resp., ρ_i) is the fraction of edges in the graph connected to bitnodes (resp., checknodes) of degree i. The rate of the ensemble is given by $R = 1 - \frac{\int_0^1 \rho(x)dx}{\int_0^1 \lambda(x)dx}$.
Under the sum-product algorithm the variable and check nodes exchange messages iteratively. A check node gets messages from its d_c neighbors, processes the messages and sends it back to its neighbors. The same thing applies for the variable nodes. The essential constraint, necessary to have the correct marginal a posteriori probabilities, is that the output message of the variable (check nodes) is a function of all incoming messages to the node with the exception of the message coming from the node which the message will be sent to. After l iterations the variable node decodes the associated bit based on all the informations that it could get from the l-depth subgraph of its neighbors. In the limit of very long codes, it can be shown that the decoding neighborhood of a given variable node is *tree-like*, it does not contain any cycle; in this case all the random variable (the incoming messages to every node) are independent. The *Concentration Theorem* in [22], assures that, almost all randomly constructed codes behave alike. It follows that it is only necessary to determine the averages behavior of the ensemble. The average behavior is shown to be the cycle-free case. In the limit of infinite blocklength DE computes the correct marginal a posteriori probability.

The messages are indicated using a Log-Likelihood Ratio (LLR): $v = \log\frac{p(y|x=1)}{p(y|x=-1)}$ is the output message of the variable node and $u = \log\frac{p(y'|x'=1)}{p(y'|x'=-1)}$ is the output message of the check node, where x is the bit associated to the node, y is all the information available to the variable node, x' is the bit value of the node that gets the message and y' is all the information available to the check node. The idea, [22], is then to track the evolution of the messages distribution, (DE), instead than the messages itself.
Calculating thresholds using density evolution is computationally expensive because the iterative process involves a $n-$dimensional system. Some approximations methods are possible, for example the *Erasure-Channel* approximation [25] and the *Gaussian* approximation, [25], [24]. For the first case the threshold for the erasure channel is computed[3] and the

[3]The density evolution for the erasure channel becomes a one-dimensional evolution [85]

value is mapped into the threshold of the correct channel using the equal capacity curve. In the second case the threshold is estimated approximating message densities as Gaussian. In this case, by using the *symmetry condition* [25], it is shown that the mean of the Gaussian is the only information necessary to characterize the message density[4]. This allows to follow the evolution through the graph of one parameter instead than the complete characterization of the message density. Under the Gaussian approximations, others one-dimensional quantities, instead of the mean, have been considered to approximates the message density, as SNR [25] and *Mutual Information* [27]. In the following, because of numerical stability, the development is done in terms of mutual information.

In our analysis, we make the optimistic assumption that decoding is successful (the frame is error-free) with high probability if, after m received slots, the BER under BP decoding vanishes with the number of decoder iterations. Notice that vanishing BER does not necessarily imply vanishing FER in the limit of infinite block-length. However, arguments based on concatenation of LDPCs with outer *expander* codes [30] with very large rate show that, in principle, vanishing BER implies vanishing FER at least for such concatenated constructions. Furthermore, we assume that the convergence of the decoder to vanishing BER can be detected by the decoder, so that decoding failure is always revealed. Under these optimistic assumptions, we can use the same throughput formula (2-45) by redefining $p(m)$ as

$$p(m) = \Pr\left(\lim_{l\to\infty} \mathrm{BER}^{(l)}(1) > 0, \ldots, \lim_{l\to\infty} \mathrm{BER}^{(l)}(m) > 0\right) \tag{2-28}$$

where $\mathrm{BER}^{(l)}(m)$ is the BER at BP decoder iteration l with m received slots.

We assume that the coded symbols are randomly assigned to the M blocks so that the fraction of bitnodes of degree i on each m-th block is the same as for the total code. In other words, the fraction of edges connected to bitnodes of degree i on block m is equal to λ_i/M, for all $m = 1, \ldots, M$. Numerical examples supported our choice of distributing "uniformly" the left degrees on the blocks.

In order to compute $\lim_{l\to\infty} \mathrm{BER}^{(l)}(m)$ for given fading coefficients $(\alpha_1, \ldots, \alpha_m)$, we resort to a Gaussian Approximation (GA) of DE. Let $\mathcal{M}_k$ denote the channel observation message, in the form of the log-likelihood ratio for the symbol associated to the given bitnode, k, given the channel output. Assuming, without loss of generality, that the all-zero codeword is transmitted, if the symbol corresponding to the bitnode is transmitted on the s-th slot, it is easy to see that the initial message is equal to

$$\mathcal{M}_k = \log \frac{\Pr(y_k|x_k = 1, c_s)}{\Pr(y_k|x_k = -1, c_s)} = 4\sqrt{\Gamma}\mathrm{Re}\{yc_s^*\}$$

[4] Considering the symmetry condition, the variance of the Gaussian variable σ^2 is a function of to the mean.

where y_k is the corresponding channel output and c_s is the fading coefficient. It follows that the distribution of the initial message is $\mathcal{M}_k \sim \mathcal{N}(4\beta_s, 8\beta_s)$[5].

We define a random variable P that governs the distribution of the variable node belonging to the s-th block, so that P is uniformly distributed over $s = 1, \ldots, M$. Let X denote the bitnode variable and Y denote all the information available at the bitnode at a given iteration. Then, the mutual information between the output of the bitnode and the symbol X is given by

$$I(X, Y \mid P) \;=\; \underset{p(X,Y,P)}{\mathrm{E}}\left[\log \frac{p(X, Y \mid P)}{p(X \mid P)\, p(Y \mid P)}\right] \;=\; \sum_{s=1}^{M} \frac{1}{M}\, I(X, Y \mid P = s) \qquad \text{(2-29)}$$

From the Gaussian Approximation, it follows that

$$I(X; Y|P = s) = J((d-1)\theta + \beta_s)$$

for a bitnode of degree d transmitted on slot s, where $\theta = \frac{1}{4}\mathrm{E}[m_{c\to v}]$ is the the mean divided by 4 of the messages m that goes from the checknodes to the variable node. Call $I_{out,c}^{\ell-1}$ the mutual information of a message passed along a random edge from a check node to a variable node at iteration $\ell - 1$, than the average message $\mathrm{E}[m_{c\to v}]$ can be obtained as $\mathrm{E}[m_{c\to v}] = J^{-1}\Big(I_{out,c}^{\ell-1}\Big)$. Hence we can write

$$I_{out,v}^{l} = \frac{1}{M} \sum_{s=1}^{M} J\left((i-1)J^{-1}(I_{out,c}^{\ell-1}) + \beta_s\right) \qquad \text{(2-30)}$$

In order to find the mutual information transfer function for the checknodes, we use the so-called "approximate duality" (reciprocal channel mapping) relation [25]. With this approximation, a checknode can be replaced by a bitnode provided that its input mutual information I_{in} is transformed into $1 - I_{in}$ and its output mutual information I_{out} is transformed into $1 - I_{out}$ (see [66, 86] for a more rigorous motivation of this approximation). The function $\psi_x(.) : \mathbf{X} \to \mathbf{X}$ is defined as:

$$\psi_x(x) \;=\; C_x^{-1}(1 - C_x(x))$$

where $C_x(.)$ is the capacity function and $x \in \mathbf{X}$ is the channel parameter [25]. Hence, the mutual information transfer of a checknode of degree d is approximated by

$$I_{out,c}^{\ell-1} = 1 - J\Big((d-1)\; J^{-1}\Big(1 - I_{out,v}^{\ell-1}\Big)\Big) \qquad \text{(2-31)}$$

[5] Recall that $\beta_s \triangleq \alpha_s \Gamma$

By combining equations (2-30) and (2-31), we obtain the one-dimensional recursion

$$I_{out,v}^{l} = \frac{1}{M}\sum_{s=1}^{M} J\Big((d-1)J^{-1}\Big(1-J\Big((d-1)\ J^{-1}\Big(1-I_{out,v}^{\ell-1}\Big)\Big)\Big)+\beta_s\Big) \qquad (2\text{-}32)$$

with initial condition $I_{out,v}^{0} = 0$. Figure 2-3 and 2-4 visualize the message passing. The generalization to the irregular case is straightforward and follows by defining the fraction of edges emanating from a variable node with degree i that belongs to m-th slot as $\Pr(Z=i, P=m) = \frac{\lambda_i}{M}$, where Z is the random variable that governs the degree distribution of the variable nodes. The one-dimensional recursion that approximate the DE in the irregular case for IR scheme is given by

$$F_\lambda\Big(1-F_\rho\Big(1-I_{out,v}^{l-1},\,0\Big)\,,\,\beta_s\Big) \qquad (2\text{-}33)$$

where, for a general distribution $g(x)=\sum_{i\geq 2} g_i x^{i-1}$, $g(x)\in\{\lambda(x),\rho(x)\}$ and $b\geq 0$ we define the function

$$F_g(z,b) \triangleq \sum_{i\geq 2} g_i J\left((i-1)J^{-1}(z)+b\right) \qquad (2\text{-}34)$$

By defining the mapping function $\Psi(.)$ as

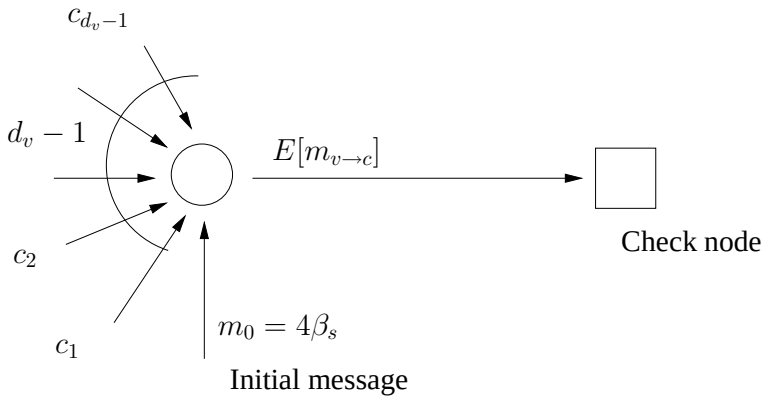


Fig. 2-3. Message flow through a variable node.

$$\Psi(z,\beta_1,\ldots,\beta_M) \triangleq \frac{1}{M}\sum_{s=1}^{M} F_\lambda(1-F_\rho(1-z,0)\,,\beta_s) \qquad (2\text{-}35)$$

we have that the condition of vanishing BER limit for given instantaneous SNRs $(\beta_1,\ldots,\beta_M)$ can be approximated by the condition that the one-dimensional dynamical system

$$z^{l} = \Psi\Big(z^{l-1},\beta_1,\ldots,\beta_M\Big), \quad l=1,2,\ldots \qquad (2\text{-}36)$$

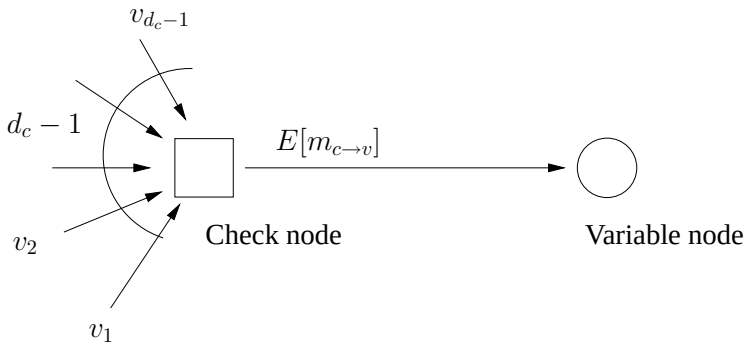


Fig. 2-4. Message flow through a check node.

with initial condition $z^0 = 0$ has a unique fixed-point $z^\infty = 1$.

The convergence behavior of the iterative decoding scheme can be seen using a mutual information transfer characteristic chart: the EXIT Chart introduced by S.ten Brink, [27] where we plot the curve $I_{out,v} = f(I_{in,v})$ and $I_{in,ch} = g(I_{out,ch})$: if the two curves have at least one intersection different from 1, the algorithm cannot converge.

At step m of the IR protocol, the decoder treats the not-yet received subblocks $s = m + 1, \ldots, M$ as erasures, i.e., as if the received signal was zero. In the DE-GA (Gaussian Approximation applied to Density Evolution) recursion for a given number of received blocks m with fading gains $\alpha_1, \ldots, \alpha_m$ is obtained by letting

$$\beta_s = \begin{cases} \Gamma\alpha_s & \text{for } s = 1, \ldots, m, \\ 0 & \text{for } s = m+1, \ldots, M. \end{cases} \tag{2-37}$$

in (2-35). It is possible to show that the function $\frac{1}{M}\sum_{s=1}^{M} F_\lambda(1 - F_\rho(1-z,\,0)\,,\,\beta_s)$ is non-decreasing with $z \in [0,1]$ and positive for $z = 0$. This implies that condition that (2-36) has unique fixed-point equal to 1 is equivalent to

$$\Psi(z, \beta_1, \ldots, \beta_M) > z, \quad \forall \;\; z \in [0,1) \tag{2-38}$$

2.6 ACHIEVABLE THROUGHPUT

In this section the results in terms of throughput average delay and variance of the delay of RB codes and infinite length LDPC are shown.

In all our numerical examples we assume Rayleigh fading, i.e., $f_\alpha(x) = e^{-x}$, and $M = 10$ fading blocks.

Figure 2-5 represents the behavior of the probability $p(m)$ for different values of m as a function of the coding rate R in the case when $\Gamma = 10$dB for RB codes. The $p(m)$ here has been computed by using equation (2-21), (by integrating of the pdf f_m obtained through the m-fold convolution of the pdf defined in equation (2-20)). Note that $p(m)$ is always equal to 1 for value of the rate R greater then m/M. In this case in fact the number of information bits is greater than the number of bits that has been sent. On the contrary, for $R \ll \frac{m}{M}$, ($b \ll Lm$) we have a very small probability of unsuccessful decoding. The outage probability shows a "step" behavior as long as the SNR increases.

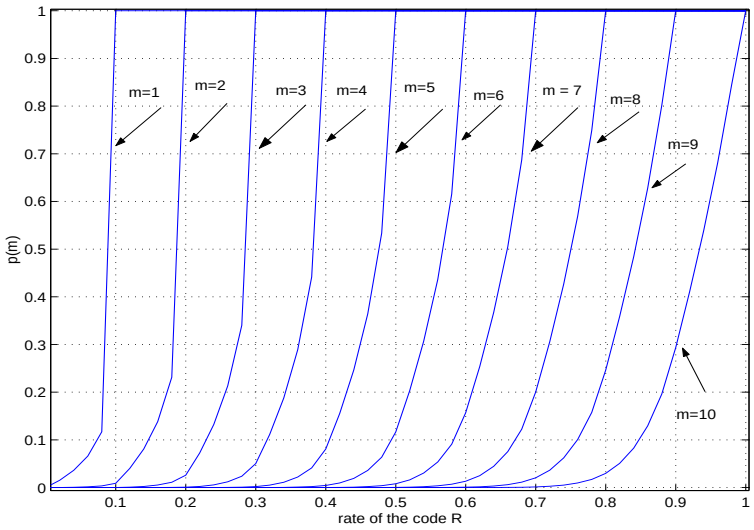


Fig. 2-5. $p(m)$ for $M = 10$, $\Gamma = 10dB$. The probabilities are generated using the convolution of the probability density function in (2-20)

Figures 2-6 and 2-7 show the throughput results in the case when the $p(m)$ is computed using equation (2-21) ("Convolutions"), using the Gaussian approximation defined in (2-22) ("Gaussian Approximation"), and finally when we use the Chernoff bound defined in (2-27) ("Chernoff Bound"), for $\Gamma = 3, 10$dB. The Chernoff bound gives a looser lower bound of the throughput while the Gaussian approximation becomes more precise in the region of high rate and for lower values of the SNR. Note that the throughput obtained using the "Convolutions" shows some picks: this is due to the particular "step" behavior of the probabilities $p(m)$. Consider for example a rate R between 0.1 and 0.2b/s/Hz. The throughput is given by $\eta = \frac{RM}{1+\sum_{m=1}^{M-1} p(m)} \backsimeq \frac{RM}{1+p(1)+p(2)}$, since from figure 2-5 $p(m)$ is negligible when $m > 2$. In particular at the extremes of this interval, since $p(1) = 1$ and $p(2) = 0$ for $R = 0.1$ and $p(1) = p(2) = 1$ for $R = 0.2$, the throughput is $1/2$ and $2/3$

respectively. In between we can see that $p(2)$ will not increase as much as the rate R, so that we can approximate the throughput as $\hat{\eta} \backsimeq RM/(2+\epsilon_R)$ with $\epsilon_R \backsimeq 0$ negligible. If higher values of rate are considered, for example $R = (0.7 - 0.8)$, this approximation does not hold anymore since the throughput can be approximated by $\hat{\eta} = RM/(7+\epsilon_R)$, with $\epsilon_R > 0$ (figure 2-5).

Figure 2-6 and 2-7 show also the throughput obtained when conventional coded ARQ systems are used. The comparison between IR and SR protocols is more evident by plotting the average delay vs. the throughput (see figure 2-9 in the case $\Gamma = 10$dB). From equations (2-8) and (2-9), we have expressions of η and or μ parameterized in the code rate $R \in [0, 1]$ (for given number of fading blocks M, fading gain statistics and SNR Γ). Hence, the curve $\mu = \mu(\eta)$ can be obtained in parametric form, by letting R varying in the interval $[0, 1]$. Since η is a non-monotone function of R, each value of η corresponds to possibly multiple values of μ. Clearly, in the presence of multiple values only the minimum is relevant. Figure 2-9 clearly shows that SR-M is not convenient. On the contrary, for a certain range of throughput SR-1 (which is also the simplest ARQ scheme) achieves almost the same average delay of IR. We have to notice that here SR-1 is coded over only one slot, meaning that the rate in that case $R_{SR-1} = \frac{b}{L}$. Consider figure 2-6 and 2-9: the highest throughput of SR-1 is $\eta_{SR-1} = 0.7$ is achieved when $R_{SR-1} \simeq 0.9$ with an average delay $\mu_{SR-1} \simeq 1$slots; the same throughput can be obtained with the IR scheme with the same delay and with mother code rate $R \simeq 0.09$. The IR in the average will send only the first coded burst with an effective coding rate $\frac{b}{Lm} \to \frac{b}{L}$. However, there is a range of high throughput that is not achievable by SR-1 while it can be achieved by the IR protocol at the cost of a very small average delay (from 2 to 6 or 7 slots). Figure 2-10 shows the average delay vs rate in the case of RB codes for $\Gamma = 3,\ 10$dB. Consider for example the curve for $\Gamma = 10$dB. The region of high throughput $R \simeq 0.7$ bit/symbol is achieved with an high average delay (8 or 9 slots). For some practical delay constraints we can be interested in considering regions where the average delay is small and the throughput is still good (region of $R \simeq 0.35$bit/symbol). As we expect, as the value of Γ augments the average delay is decreasing: in the limit of $\Gamma \to \infty$ the delay will be equal the minimum number of bursts necessary to contain all the information bits (RM). The variance of the delay, figure 2-11 increases when the value of Γ is decreasing. This is due to the fact that for high values of Γ the number of bursts between two transmissions of a codeword does not vary too much. In the limit of high signal to noise ratio, the number of iterations becomes a constant so that the variance goes toward zero. Figures 2-6, 2-7 show also the comparison between the throughput obtained using the Random Binary codes and the LDPC codes. In the case of LDPC, DE algorithm is used to obtain a evaluation of the performance. Each mark ($*$) in figures 2-6 and 2-7 is obtained by using an irregular LDPC ensemble with degree distributions λ, ρ optimized for the corresponding rate R and for the standard unfaded AWGN channel [87, 22]. No attempt was made to optimize the degree distributions to take into account the block-fading

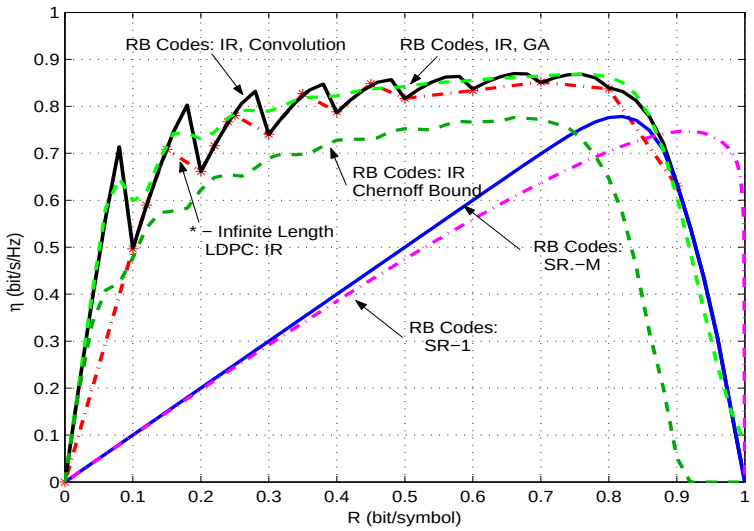


Fig. 2-6. Comparison between the throughput when we consider the convolution of the pdf, the Gaussian approximation and the Chernoff bound, when $\Gamma = 10dB$.

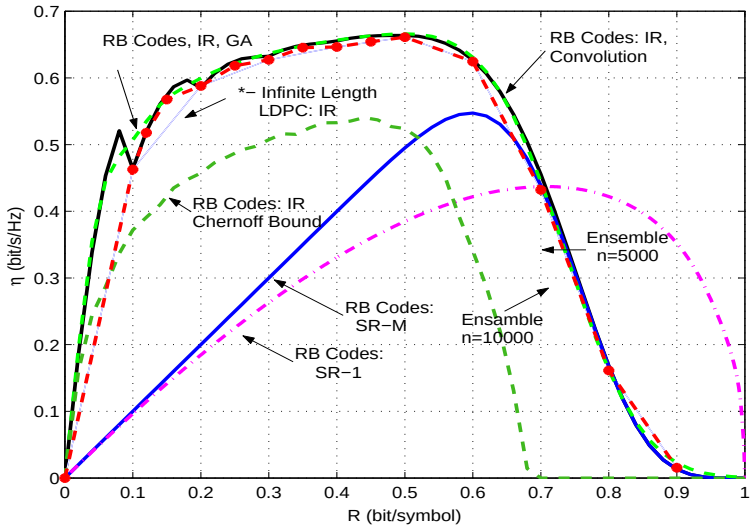


Fig. 2-7. Comparison between the throughput when we consider the convolution of the pdf, the Gaussian approximation and the Chernoff bound, when $\Gamma = 3dB$.

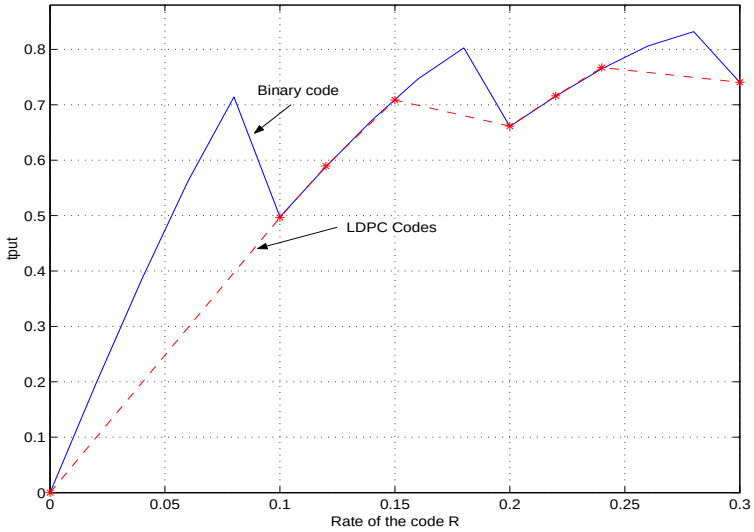


Fig. 2-8. Comparison between the throughput obtained using LDPC codes and RB codes, considering the convolution of the pdf, $\Gamma = 10dB$, zoom in the interval $R = (0.1, 0.3)$ b/s/Hz.

channel. Nevertheless, these results show that AWGN-optimized ensembles perform close to optimal and not much can be gained by further ensemble optimization.

2.7 Finite Length LDPC for HARQ

At this point, it is natural to ask how practical finite-length LDPC code perform on the block-fading channel under the IR protocol, by removing the optimistic assumptions (limit for large L, vanishing BER $\Rightarrow$ vanishing FER) that led to the outstanding results of the previous section. The first subsection shows the performance results of finite length LDPC without any countermeasure and finally two methods to overcome the performance loss due to finite length LDPC will be explained.

2.7.1 Finite Length LDPC from Complete Random Ensemble

Figures 2-12 and 2-13 show the throughput obtained by simulation of the IR protocol by using actual finite-length LDPC codes of length $n = 5000$ and $n = 10000$. For the sake of comparison we plot also the average throughput obtained using RB codes and infinite

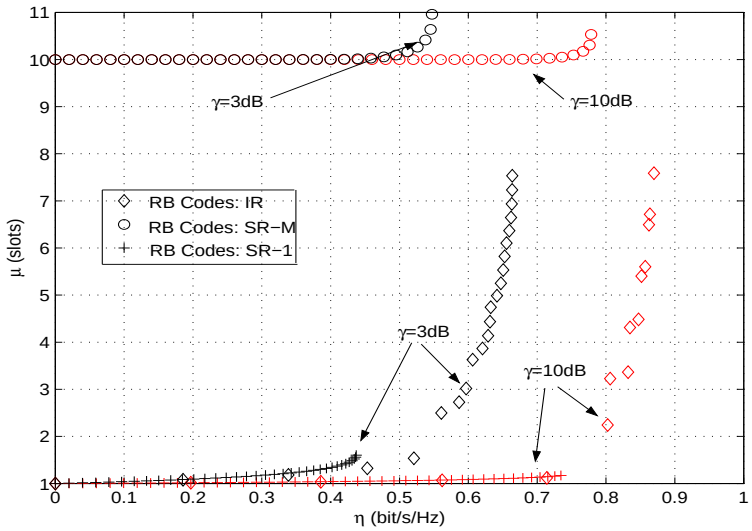


Fig. 2-9. Average delay μ vs. throughput η for IR, SR-1 and SR-M protocols with random binary codes for $\Gamma = 3, 10$dB.

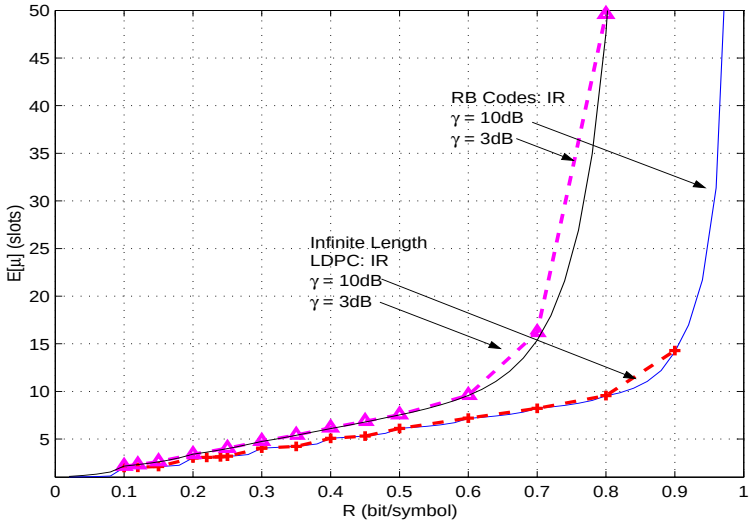


Fig. 2-10. Comparison between the average delay vs rate in the case of LDPC codes and RB codes for $\Gamma = 3,\ 10$dB.

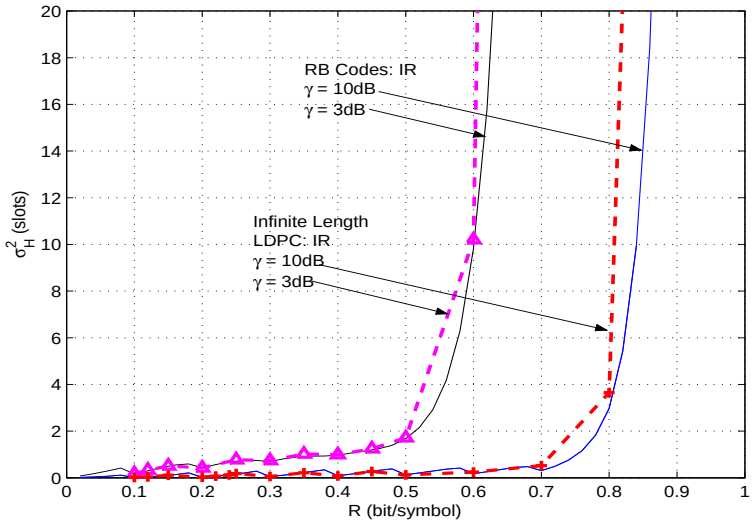


Fig. 2-11. Comparison between the variance of the delay vs rate in the case of LDPC codes and RB codes for $\Gamma = 3,\ 10$dB.

length LDPC discussed before. The finite-length results are obtained by averaging over the channel fading, the noise and the ensemble of codes, i.e., a new parity-check matrix is randomly generated according to the given left and right degree distributions λ, ρ defining the ensemble for each transmitted information packet. The throughput formula for finite-length codes is still given by (2-45) where $p(m)$, for a given LDPC ensemble with degree distributions λ, ρ, is expressed by

$$p(m) = \underset{\mathcal{C}(n,\lambda,\rho),\boldsymbol{\alpha}}{E}[Pr(\overline{\mathcal{A}}_1, \overline{\mathcal{A}}_2, \ldots, \overline{\mathcal{A}}_m \,|\boldsymbol{\alpha}, (\lambda, \rho))] \tag{2-39}$$

where, $\boldsymbol{\alpha}$ is the sequence of fading gains, $\mathcal{A}_s$ is the event of successful decoding at step s and where the code parity-check matrix is randomly generated with uniform probability over all bipartite graphs with degree distributions λ, ρ (a method for generating such random graphs is given in [22]). Successful decoding is defined by the event that, after a given maximum number of BP decoder iterations, all information bits are correct.

The throughput performance loss of finite-length ensembles with respect to their infinite-length counterpart can be explained by observing that, typically, irregular finite-length LDPC codes with bitnodes of degree 2 have very poor FER performance, despite the fact that they perform well in terms of BER. This is because typical decoding errors involve a very small number of bit-errors per frame error [29].

In principles, throughput higher than the ensemble average performance can be achieved by careful selection of a particularly good realization of the code parity-check matrix. However, selecting such matrix is not a simple task in general. A standard technique is to generate the matrix into some restricted (or "expurgated") ensemble where codes with good FER performance can be found with high probability. An example of this approach will be detailed later, in Section 2.7.2.

Another remarkable fact evidenced by figures 2-12 and 2-13 is that codes with block length $n = 5000$ slightly outperform codes with $n = 10000$. This is surprising since in standard AWGN settings (without ARQ) BER is known to improve with the code block length [22]. Indeed, irregular LDPC codes are commonly believed to provide good performance only for extremely large block length. The above results show that in the presence of time-varying channels and retransmission schemes this is not the case, as FER and not BER determines the throughput performance. Next, we propose two approaches to improve the performance of IR with finite-length practical LDPC codes. The first approach acts directly on the code design and leaves the IR protocol unchanged. As anticipated above, it consists of selecting the code parity-check matrix in some appropriate ensemble with good FER properties. The second approach acts on the IR protocol and leaves the code design unchanged. It consists of dividing the information packet into subpackets, performing error detection on each of the subpackets and using an outer selective-repeat protocol only for the subpackets in error. Interestingly, although these approaches are quite different, they yield almost the same performance improvement and recover a considerable fraction (up to 80% at SNR $= 10$ dB) of the loss due to finite with respect to infinite length.

2.7.2 Special graph construction

Solutions to improve the FER performance of LDPCs consist of finding special constructions based on expander graphs [88, 30, 31, 89, 32, 33, 90], or a deterministic arrangement of the edges adjacent to degree-2 bitnodes [87].

Due to its simplicity, we follow this second method. Good FER codes can be obtained constructing the graph such that the edges emanating from a bitnode of degree 2 are placed semi-deterministically. Let R denote the rate of the code and $\tilde{\lambda}_2 = \frac{\lambda_i/i}{\sum_j \lambda_j/j}$ be the fraction of bitnodes of degree 2. For $\tilde{\lambda}_2 < \frac{(1-R)}{2}$ we connect each edge emanating from one of these bitnodes to a different checknode so that there are no checknodes adjacent to more than one bitnode of degree 2. For $\frac{(1-R)}{2} < \tilde{\lambda}_2 < 1 - R$ we arrange the $\tilde{\lambda}_2 n$ deg-2 bitnodes and $\tilde{\lambda}_2 n$ checknodes into a cycle of girth $2\tilde{\lambda}_2 n$, as shown in the example of figure 2-14.

As an example of this construction, consider a standard unfaded AWGN channel and the en-

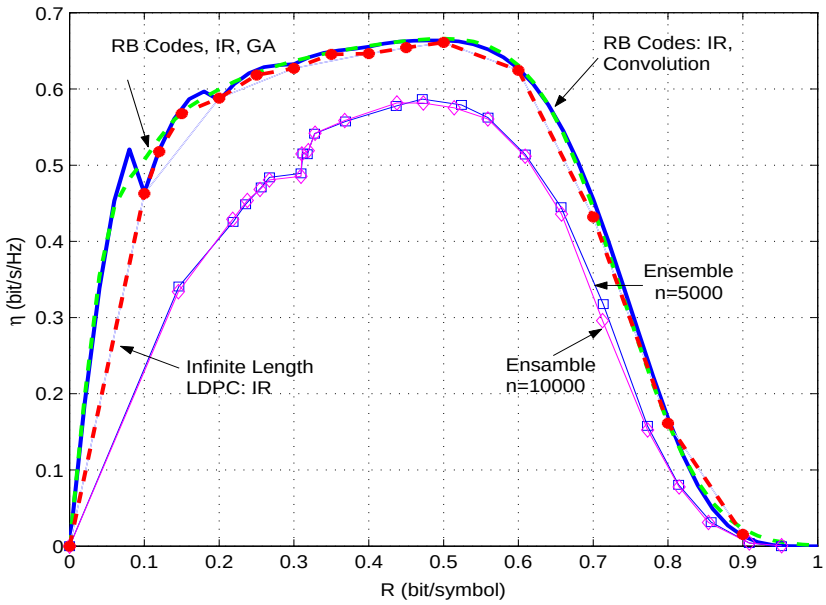


Fig. 2-12. Throughput vs. code rate R for $\Gamma = 3$dB. IR protocol with RB codes, infinite length LDPC codes with degree distributions taken from [87], finite length LDPC with $n = 5000,\ 10000$.

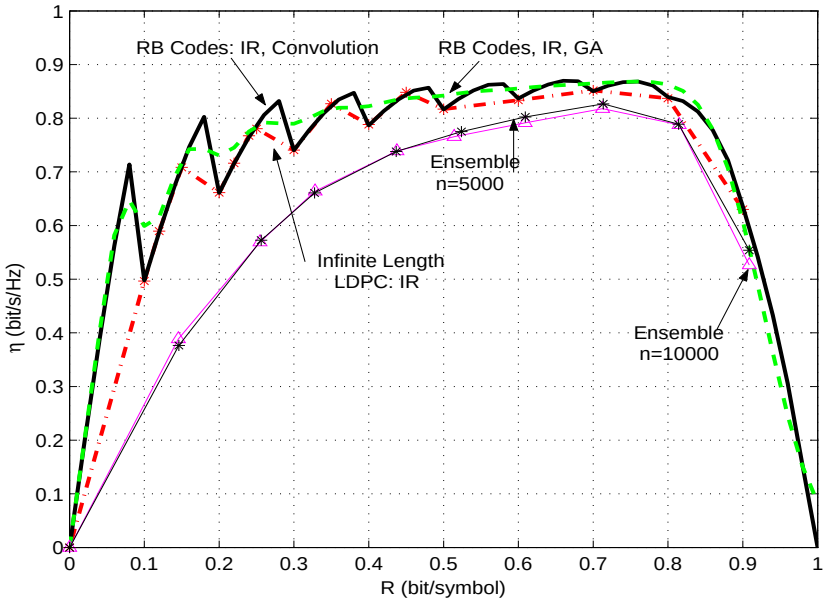


Fig. 2-13. Throughput vs. code rate R for $\Gamma = 10$dB. IR protocol with RB codes, infinite length LDPC codes with degree distributions taken from [87], finite length LDPC with $n = 5000,\ 10000$.

semble of codes with variable and check node degree sequences defined defined in [87], for a rate $R = 0.3$ bit/symbol, maximum left degree $d_v = 100$, average right degree $a_r = 6.9$ and block length $n = 10000$. Figure 2-15 shows the BER and the FER obtained by averaging over all graphs with given degree distributions (Total ensemble) and by averaging over all graphs with special cyclic arrangement of the edges connected with degree-2 bitnodes (Modified ensemble). It is clear that the modified ensemble yield much better FER and BER performance.

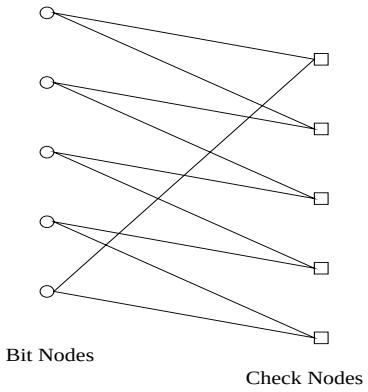


Fig. 2-14. Cyclic arrangement of the edges adjacent to bitnodes of degree 2.

2.7.3 Outer Selective Repeat System (OSR)

Our second approach to close the gap between infinite and finite length LDPCs stems from the following observation: for standard irregular LDPC codes, most frame errors involve a very small number of bit errors. Therefore, by dividing the information packet into smaller subpackets, only a few of them will contain errors after decoding. Hence, an Outer Selective-Repeat (OSR) protocol acting on these smaller subpacket units can recover subpacket errors without having to retransmit the whole codeword, only the erroneous packets are retransmitted together with new information packets. The optimistic assumption underlying this approach is that subpackets in error can be perfectly detected. The concept of the concatenated selective-repeat scheme is represented in figure 2-16. Let us focus on step m of the IR protocol. If the iterative BP decoder, processing the received signal $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ with instantaneous SNRs $\Gamma\alpha_1, \ldots, \Gamma\alpha_m$, works below the iterative decoding threshold, the decoded codeword after a given (large) number of iterations might be either error-free or contain a small number of residual errors. These few residual errors are the main cause of performance loss of finite-length LDPC codes, since even a single bit-error would generate a NACK and the IR protocol would proceed to block $m+1$ of the current codeword instead

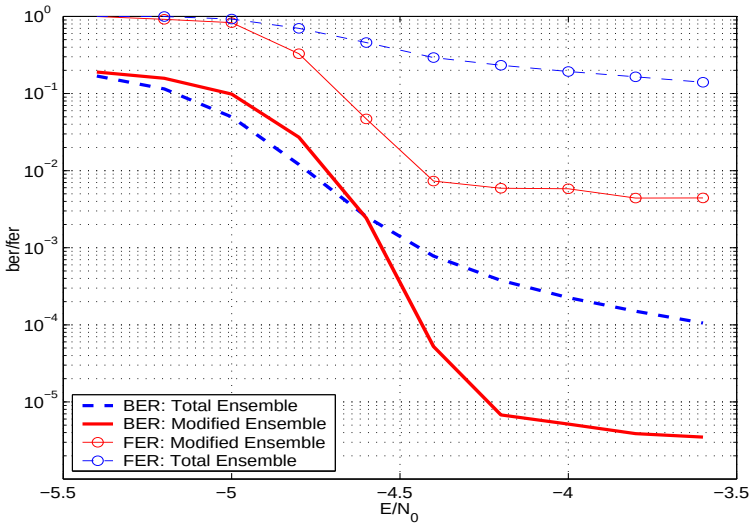


Fig. 2-15. BER and FER of the LDPC ensemble with degree distributions given in [87] for a rate $R = 0.3$ bit/symbol, maximum left degree $d_v = 100$, average right degree $a_r = 6.9$ and length $n = 10000$, over the AWGN channel. The curves labeled as "total ensemble" are obtained by averaging over all code graphs with the given degree distributions. The curves labeled by "modified ensemble" are obtained by averaging over the graphs with degree-2 edges arranged in a cycle, as shown in figure 2-14.

of starting with block 1 of the next codeword. However, the typical case of a few bit-errors implies that only a small number of data subpackets are in error, which can be handled by the OSR.

Let P denote the subpacket length in bits, and $n_p = b/P$ be the number of subpackets per LDPC codeword. At step m of the IR protocol, after a given number of decoder iterations, let e_m denote the number of subpackets in error. We shall consider "successful" decoding (i.e., the IR protocol stops the transmission of the current codeword at step m) if $e_m \leq \delta$. Otherwise, if $e_m > \delta$, a NACK is sent and the block $m + 1$ of the current codeword is sent on the next slot. The system throughput can be optimized with respect to the threshold $\delta \in [0, n_p]$. Notice that setting $\delta = 0$ is equivalent to the IR alone, without the OSR. Therefore, this system is expected to provide a throughput gain with respect to the basic IR protocol with finite-length LDPCs.

We shall compute the throughput of the concatenated OSR-IR protocol by using again the Renewal-Reward theorem, by appropriately defining the random reward $\mathcal{R}$ and the inter-renewal time τ. Let $\mathcal{E} = \{$The user stops transmitting the current codeword $\}$ be again the

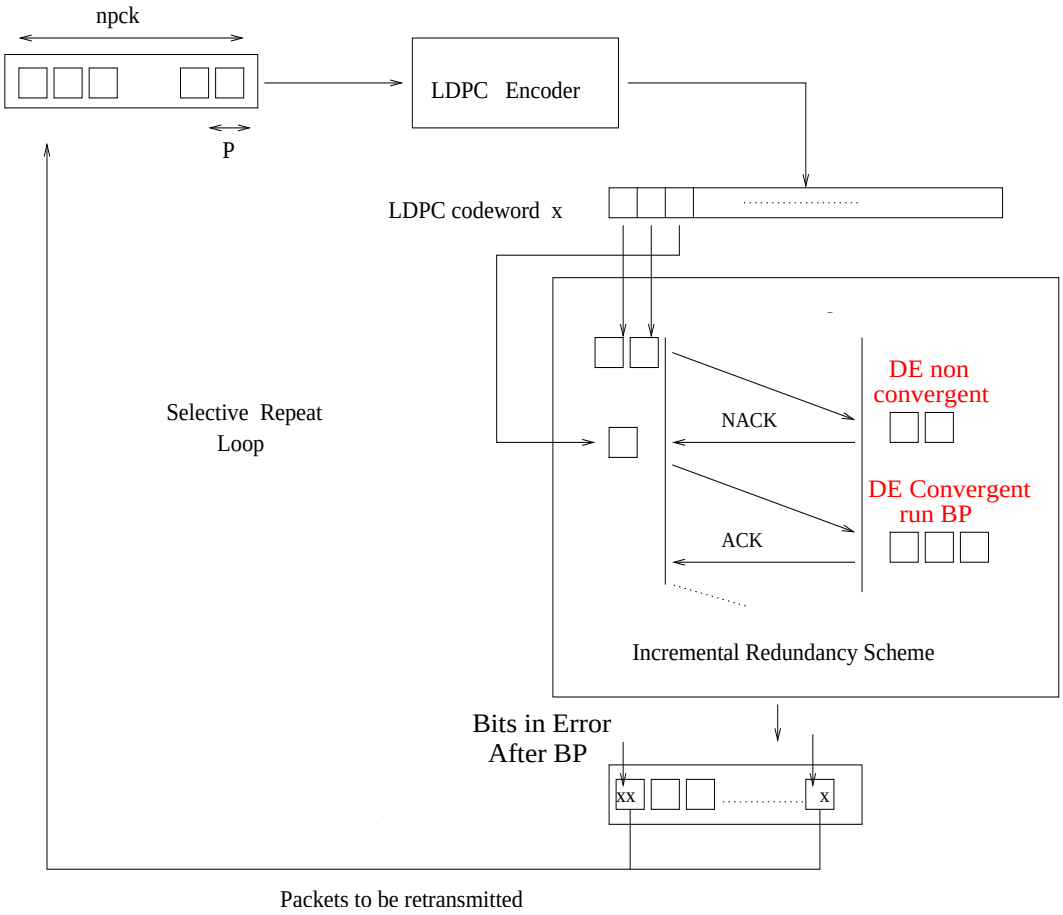


Fig. 2-16. Outer Selective Repeat scheme.

recurrent event, and $\hat{q}(m)$ be the probability that the BP algorithm ends with a number of erroneous subpackets $e_m \leq \delta$. Defining $\mathcal{B}_s = \{e_s < \delta\}$ for $s = 1, \ldots, M$, we have

$$\hat{q}(m) = \Pr\big(\overline{\mathcal{B}}_1, \ldots, \overline{\mathcal{B}}_{m-1}, \mathcal{B}_m\big) \tag{2-40}$$

The recurrent event probability is given by

$$\begin{cases} \Pr(\mathcal{E}_m) = \hat{q}(m) & \text{if } m \leq M-1, \\ \Pr(\mathcal{E}_M) = 1 - \sum_{m=1}^{M-1} \hat{q}(m) & \text{if } m = M. \end{cases} \tag{2-41}$$

Defining $\hat{p}(m) = \Pr\big(\overline{\mathcal{B}}_1, \ldots, \overline{\mathcal{B}}_{m-1}, \overline{\mathcal{B}}_m\big)$, we have $\hat{q}(m) = \hat{p}(m-1) - \hat{p}(m)$ and substituting this in (3-7) we get $\Pr(\mathcal{E}_M) = \hat{p}(M-1)$.

The average inter-renewal time (in slots) is given by:

$$\mathrm{E}[\tau] = \sum_{m=1}^{M} m \cdot \Pr(\mathcal{E}_m) = \sum_{m=1}^{M-1} m\hat{q}(m) + M\hat{p}(M-1) = 1 + \sum_{m=1}^{M-1} \hat{p}(m) \tag{2-42}$$

The reward $\mathcal{R}$ is a random variable that takes values in the range $\{0, S/L, \ldots, n_p S/L\}$. Recalling the definition of e_m as the number of erroneous packets after decoding at IR step m, we can write

$$\begin{aligned} \mathrm{E}[\mathcal{R}] &= \frac{P}{L} \sum_{m=1}^{M} \sum_{e=0}^{n_p} (n_p - e) \Pr(e_m = e \,|\, \mathcal{E}_m) \Pr(\mathcal{E}_m) \\ &= \frac{P n_p}{L} \left(1 - \sum_{m=1}^{M-1} r_m \hat{q}(m) - r_M \hat{p}(M-1) \right) \end{aligned} \tag{2-43}$$

where we define

$$r_m = \frac{1}{n_p} \sum_{e=0}^{n_p} e \Pr(e_m = e | \mathcal{E}_m)$$

to be the average fraction of subpackets in error after decoding at step m, given the recurrent event. Recalling that $P n_p / L = RM$, we obtain the desired throughput expression as

$$\eta = RM \frac{1 - \sum_{m=1}^{M-1} r_m \hat{q}(m) - r_M \hat{p}(M-1)}{1 + \sum_{m=1}^{M-1} \hat{p}(m)} \tag{2-44}$$

The above formula can be evaluated after computing by Monte Carlo simulation the probabilities $\hat{p}(m)$ and the fractions r_m.

2.8 FINITE LENGTH LDPC: ACHIEVED PERFORMANCE AND COUNTERMEASURES

In this section we show the throughput resulting from the modified LDPC ensemble, from the use of an OSR protocol, or from a combination of both techniques. In all the following examples, we fixed the subpacket length of the OSR protocol equal to $P = 48$ bits (6 bytes).

Clearly, the throughput achieved by OSR depends on the threshold δ. Analytical optimization of δ is difficult if not impossible. Hence, we exhaustively searched for the best threshold value. Figure 2-18 shows the throughput as a function of $\delta \in [0, 1]$ for the same setting as in figure 2-15 and $\Gamma = 10$dB. We notice that the performance of the OSR is quite insensitive to the value of δ (unless δ is either very close to 0 or very close to 1). We plotted also the throughput achieved by the same ensemble with infinite length, with finite length without any countermeasure and with finite length by averaging over the modified ensemble. These results are shown as horizontal lines as they do not depend on δ.

Both the OSR and the modified ensemble are able to recover a large fraction of the loss incurred by finite length LDPCs (until 80%). It is natural to wonder about the benefit of using jointly the OSR protocol and a modified LDPC ensemble. Unfortunately, the answer to this question is negative. In figure 2-18, the curve labeled by "OSR-Modified Ensemble" refers to this case and we notice that the obtained throughput is slightly inferior to that obtained by using OSR with the total ensemble. This fact can be explained by noticing that for a typical code in the modified ensemble a frame-error corresponds to a large number of bit errors (i.e., a large number of subpackets to retransmit). Hence, using an outer SR protocol does not improve the throughput.

The almost constant behavior of throughput of OSR over a wide range of values of the threshold δ is explained by observing the statistics of the number of subpackets in error e_m after decoding. For example, figure 2-19 shows the probability mass function of e_m conditioned on the event that the decoder works above its iterative threshold decoding (i.e., subject to the event that DE with m received blocks converges to vanishing BER), with $m = 4$ received blocks. We notice that the number of packets in error is mostly concentrated below 10% and above 90%. This behavior can be observed for all m. Therefore, the throughput is almost constant for $\delta \in (0.1, 0.9)$.

2.9 REDUCING THE AVERAGE COMPLEXITY OF LDPC DECODING

This section is focused on the analysis of the complexity of the decoder when IR scheme is coupled with LDPC codes. We show that a method based on DE lowers the complexity of the decoder and can ideally achieve all the trade-offs between complexity and performance.

For standard time-invariant channels (e.g., the binary-input AWGN channel) and for a given ensemble (λ, ρ) it can be shown that there exists a value SNR*, the *iterative decoding threshold*, such that if the signal to noise ratio is below SNR* then the DE error probability limit is bounded away from zero (even after an infinite number of iterations) while if it is above SNR* the DE error probability limit is zero. For LDPC codes of practical length (say, between $n = 5000$ to $n = 10000$), in order to predict if iterative decoding is successful with high probability, we can just compare the instantaneous channel SNR with the iterative decoding threshold SNR*.

This idea does not carry over straightforwardly for a time-varying channel such as in the case of the IR-HARQ protocol considered here. The BP decoder at slot m "sees" a time-varying channel defined by the instantaneous SNRs $\{\beta_1, \ldots, \beta_m\}$ and by the fact that the symbols in slots $m+1, \ldots, M$ are erased (i.e., the corresponding channel outputs are zero). Clearly, no simple threshold criterion for convergence of BP can be applied here. Indeed, we might define a region of convergence for the decoder in slot m as an m-dimensional region $\mathcal{R}_m \subset \mathbb{R}_+^m$, such that if $(\beta_1, \ldots, \beta_m) \in \mathcal{R}_m$, then DE converges to vanishing BER and the BP decoder applied to the actual finite-length code with channel observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ yields successful decoding with high probability. For a given ensemble (λ, ρ) and a given average SNR Γ, in principles one could determine the region of convergence $\mathcal{R}_m$ by running the DE algorithm for all values of $(\beta_1, \ldots, \beta_m) \in \mathbb{R}_+^m$. This is clearly not an easy task, since the SNR vector takes on values in a continuous and unbounded m dimensional real set. In order to overcome this problem, it is possible to use a on-line low-complexity approximation of DE and run it in real-time at each newly received slot before activating the BP decoder. Hopefully, the approximate DE is able to approximate accurately the convergence region $\mathcal{R}_m$ for all $m = 1, \ldots, M$. Therefore, if the approximate DE converges to zero error probability, the BP decoder is triggered and actual decoding is performed, otherwise a NACK is sent without actually performing decoding. This results into a tremendous saving in decoder average complexity without affecting the average throughput.

2.9.1 Average Throughput and Complexity

It is important to note that this method may, in general, decrease the average throughput since it declares a decoding failure whenever DE does not converge, while there is a chance

that the actual BP decoder is successful even if DE does not converge. Recall that the throughput ca be written as

$$\eta = R\,M\,\frac{1-p(M)}{1+\sum_{m=1}^{M-1}p(m)} \tag{2-45}$$

We define $p^{BP}(m)$ and $p^{T}(m)$ as the outage probability obtained when running always the BP algorithm and when using the DE-test, and $q^{BP}(m)$, $q^{T}(m)$ as the probability of successful decoding at step m for the two methods; thus, redefining $\mathcal{A}_m$ as the event $\{$The BP algorithm converges at step $m\}$, and $\mathcal{B}_m$ the event $\{$The DE test converges at step $m\}$ it follows that $q^{BP}(m) = \Pr\big(\overline{\mathcal{A}}_1, .., \overline{\mathcal{A}}_{m-1}, \mathcal{A}_m\big) = p^{BP}(m-1) - p^{BP}(m)$ and

$$\begin{aligned} q^{T}(m) &= \sum_{i=1}^{m} \Pr\big(\overline{\mathcal{B}}_1, \ldots \overline{\mathcal{B}}_{i-1}, \mathcal{B}_i, \overline{\mathcal{A}}_i, \ldots \overline{\mathcal{A}}_{m-1}, \mathcal{A}_m\big) \\ &= \sum_{i=1}^{m} q_c^{T}(m|i)\cdot q^{\infty}(i) = p^{T}(m-1) - p^{T}(m) \end{aligned} \tag{2-46}$$

where we have defined

$$q^{\infty}(i) \stackrel{\Delta}{=} \Pr\big(\overline{\mathcal{B}}_1, \ldots \overline{\mathcal{B}}_{i-1}, \mathcal{B}_i,\big)$$

and

$$q_c^{T}(m|i) \stackrel{\Delta}{=} \Pr\big(\overline{\mathcal{A}}_i, .., \overline{\mathcal{A}}_{m-1}, \mathcal{A}_m \,|\, \overline{\mathcal{B}}_1, .., \overline{\mathcal{B}}_{i-1}, \mathcal{B}_i,\big)$$

We call η^{BP} the average throughput obtained when using always BP algorithm and η^{T} when using the DE-test based decoder; they can be obtained substituting $p^{BP}(m)$ and $p^{T}(m)$ in (2-45) respectively. Let us consider the simple case when we let the BP decoder run for a maximum number of iterations without stopping criterion, and for DE, we evaluate the mapping curve Ψ over a fine grid of points over the interval $[0, 1]$ and we detect if these points are all above the diagonal or if there is intersection [27]. Under these simplifying hypothesis, figure 2.10 shows the comparison between $q^{BP}(m)$ and $q^{T}(m)$ vs rate as a function of the number of transmitted slots m for $\Gamma = 3dB$. As we expect, the DE-test method does not decrease the performance of the iterative decoder since the probability of successful decoding for the two method are very close (see figure 2-22).

2.9.2 Average Complexity: Independent Case

In the following we demonstrate that using this method the average complexity is greatly reduced with respect to the case when we perform always the BP decoder. First we consider the simple case described before when the complexity of both BP and DE are independent on the fading realization and on the index m. We compute the average complexity under

the following hypothesis: $1)$ Look-up table cost zero (i.e., evaluating $J(x)$ costs zero), $2)$ Additions, multiplications and comparisons cost the same (one binary operation), $3)$ We consider I_d points equally spaced on $[0,1]$ to evaluate the mapping function equals, $4)$ We call I_b the maximum number of iterations of the BP algorithm.

In this simple case, let C_{DE} and C_{BP} the complexity of the DE test and the complexity of BP algorithm. Calling c and d the different degrees in the check node and variable node degree distribution, $\mathcal{L}$ the number of binary operations to compute a logarithm and exponential, n_e the number of edges in the graph, and n the length of the code, it follows that

$$C_{DE} = C_{DEi} I_d = [4 \cdot c + d \cdot (2 + 3M) + 3\,]\, I_d \tag{2-47}$$

$$C_{BP} = C_{BPi} I_b = [(2\mathcal{L} + 6)\, 2n_e + 2n\,]\, I_b \tag{2-48}$$

given in number of binary operations. Let τ be the RV denoting the number of slots needed for stopping the transmission of the current codeword, $h = 1, \ldots M$ the number of slots needed to have DE convergent, $\phi = 0, \ldots M - h + 1$ the number of slots after DE have converged (including the slot for which DE converges) until BP converges, thus $\tau = h + \phi - 1$. We call $\bar{C}_{std}$ and $\bar{C}_{test}$ the average complexity when using BP always and when using the proposed method. We can always write the expected complexity conditioning on the value of τ, thus $\mathbb{E}[C] = \mathbb{E}[\,\mathbb{E}[C|\tau]\,]$, yielding[6]

$$\begin{aligned} \bar{C}_{std} &= \mathbb{E}[C_{std}] = \mathbb{E}[\,\mathbb{E}[C_{std}|\tau]\,] = \mathbb{E}[C_{BP}\,\tau] \\ &= C_{BP}\mathbb{E}[\tau] \overset{(a)}{=} C_{BP}\left[1 + \sum_{m=1}^{M-1} p^{BP}(m)\right] \end{aligned} \tag{2-49}$$

where (a) follows from equation (2-45) noticing that $\eta \propto 1/\mathbb{E}[\tau]$, with $p^{BP}(m)$ defined above.

The average complexity of the DE-test method can be obtained averaging over the value of h and ϕ, yielding

$$\begin{aligned} \bar{C}_{test} &= \mathbb{E}[C_{test}] = \mathbb{E}[\,\mathbb{E}[\,\mathbb{E}[\,C_{test}|h,\phi\,]\,]\,] \\ &= C_{DE}\mathbb{E}[h] + C_{BP}\mathbb{E}[\phi] \end{aligned} \tag{2-50}$$

where $\mathbb{E}[h]$ is the average number of slots after which DE converges $E[h] = 1 + \sum_{m=1}^{M-1} p^{\infty}(m)$

[6] In the following whenever there is no confusion, we skip the index that denotes the RV with respect to whom we are averaging.

By definition $E[\phi] = \sum_{i=0}^{i=M} \phi \Pr(\phi = i)$ and

$$\begin{aligned}\Pr(\phi = i) &= \sum_{k=1}^{M} \Pr(\phi = i | h = k) \Pr(h = k) \\ &= \sum_{k=1}^{M} q_c^T(k + i - 1|k) \cdot q^{\infty}(k) \end{aligned} \tag{2-51}$$

All this quantity $q^{BP}(m)$, $q_c^T(m|i)$ and $q^{\infty}(m)$ are computed by Monte Carlo simulations.

Figure 2-22 gives the average complexity $\bar{C}_{test}$ and $\bar{C}_{std}$ as a function of the rate when we consider $I_d = 100$ and $I_b = 200$. The curve referred to as $\bar{C}_{mod}$ represents an amelioration of the proposed method. The rationale is the following: we can precompute a threshold for β_1 by running DE offline, so that for the step $m = 1$ there is no need for performing the DE test. In this case, the complexity $\bar{C}_{test}$ can be further reduced to

$$\begin{aligned}\bar{C}_{mod} &= \mathbb{E}[C_{test}] = \mathbb{E}[\,\mathbb{E}[\,\mathbb{E}[\,C_{test}|h, \phi]]\,] \\ &= C_{DE} \sum_{i=2}^{M} h \Pr(h) + C_{BP} \mathbb{E}[\phi] \\ &= C_{DE} \mathbb{E}[h] - C_{DE} q^{\infty}(1) + C_{BP} \mathbb{E}[\phi] \end{aligned} \tag{2-52}$$

As we can see the average complexity is drastically decreased when using DE-test based method while this further amelioration $\bar{C}_{mod}$ does not introduce important gain with respect to $\bar{C}_{test}$.

2.9.3 Average Complexity: BP Stopping Criteria

Consider now the case when C_{BP} and C_{DE} are not constant but depend on the fading realization and index m: this includes cases when we consider stopping criteria for BP algorithm and when we run DE as a dynamical system. Consider at first BP algorithm: in [91, 92, 93, 94] the authors analyzed criteria to stop the iteration process in turbo decoding, in particular in [91] this is studied in the context of type I HARQ; these criteria are based essentially on Cross Entropy (CE) computation of the estimates at the outputs of the decoder after each iteration. The decoder declares a packet acceptable if the cross entropy at the ith iteration falls below a particular threshold. It is important to notice that in [91, 92, 93, 94] the authors find a criterion to stop the iterations of BP algorithm while here we propose a method to trigger the iterative decoding only if the instantaneous SNR is inside the convergence region; obviously the stopping criteria studied in [91, 92, 93, 94] can be jointly applied with the DE-test method, yielding further savings. They have also

introduced a method based on the Cyclic Redundancy Check, (CRC). After each iteration of the turbo decoder the CRC bits are used to detect any error that remains. The packet is accepted when the CRC decoder declares the packet to be error free. These criteria reduce the number of iterations of the iterative decoding technique but they accept performance degradations that can be important in the case of CRC based method. We have also to notice that the CE based stopping criteria requires evaluating average Kullbach-Leibler distance over the whole block: this costs roughly as an additional iteration, per iteration (additional cost per iteration proportional to n). Here we do consider an other stopping criterion based on syndrome computation at each iteration: it costs very little since it is an operation in the binary field. When we consider the DE-test based method, we can run DE as a dynamical system stopping the iterations when a fixed point is reached. This includes also the possibility to modify the initialization at each new iteration for the DE test: in the standard algorithm the initialization is $I^0_{out,v} = 0$; let us consider the case when at step m there is a fixed point different from 1, $I^{I_d}_{out,v}(m)$, when I_d is as before the maximum number of iterations. The decoder needs more redundancy to be able to decode. At step $m+1$ we change the initialization setting $I^0_{out,v}(m+1) = I^{I_d}_{out,v}(m)$. This speed up the algorithm without loosing in performance since it can be shown that the mutual information sequence is monotonically increasing in $\Gamma\alpha_m$. We call $\bar{C}_{stdm}$ and $\bar{C}_{testm}$ the average complexity using the classical method and the DE-test method with CE or syndrome computation as stopping criteria (in the figure 'm' will be substituted with CE or SY respectively). In the case when we consider the BP algorithm, call I^{BP}_τ the sum of all the iterations done after receiving $1, ..\tau$ slots, given the fact that BP converges at step τ,

$$\begin{aligned}\bar{C}_{stdm} &= \mathbb{E}[C_{stdm}] = \mathbb{E}[\,\mathbb{E}[C_{stdm}|\tau]\,] \overset{(a)}{=} C_{BPi}\mathbb{E}\big[\mathbb{E}\big[I^{BP}_\tau\big]\big] \\ &\overset{(b)}{=} C_{BPi}\mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{\tau} i^{BP}_k\right]\right] = C_{BPi}\mathbb{E}\left[\sum_{k=1}^{\tau} \bar{i}^{BP}_k\right] \end{aligned} \tag{2-53}$$

where (a) is due to the fact that the average complexity given τ is equal to the complexity of one iteration times the total number of iterations done over all the τ slots I^{BP}_τ; (b) is due to the fact that $\mathbb{E}\big[I^{BP}_\tau\big] = \mathbb{E}\big[\sum_{k=1}^{\tau} i^{BP}_k\big] = \sum_{k=1}^{\tau}\mathbb{E}\big[i^{BP}_k\big] = \sum_{k=1}^{\tau}\bar{i}^{BP}_k$. With analogy we say that the average complexity when using the DE test is given by:

$$\begin{aligned}\bar{C}_{testm} &= \mathbb{E}[C_{testm}] = \mathbb{E}\Big[\mathbb{E}\Big[\,\mathbb{E}\Big[C'_{test}|h,\phi\Big]\Big]\Big] \\ &= C_{DEi}\,\mathbb{E}\left[\sum_{k=1}^{h}\bar{i}^{DE}_k\right] + C_{BPi}\,\mathbb{E}_{h,\phi}\left[\sum_{l=1}^{\phi}\bar{i}^{BP}_{h+l-1}\right] \end{aligned} \tag{2-54}$$

where $\bar{i}^{DE}_k$ ($\bar{i}^{BP}_{h+l-1}$) is the average number of iteration necessary to know if DE (BP) converges when transmitting k ($h+l-1$) slots. Figure 2.10 shows results in terms of throughput

for all the methods considered here, BP without any countermeasure η_{std}, BP with stopping criteria based on Cross Entropy and Syndrome computation (η_{stdCE}, η_{stdSY}), the DE-test based method for the three cases above (η_{test}, η_{testCE}, η_{testSY}); for the sake of comparison we plot also the throughput obtained when using DE, η_{DE}. As we can see all this methods does not significantly modify the throughput, as shown also in figure 2.10 that shows the probability of successful decoding for the cases above mentioned. However the average complexity shown in figure 2-22 for all the cases considered shows that DE-test based system decreases considerably the average complexity. It is interesting to notice that even if CE based stopping criterion introduces more complexity per iteration, on the other hand it reduces a lot the average number of iterations when choosing the correct threshold value. We can see that DE-test method with stopping criterion based on syndrome computation achieves the best average complexity, thus it is a good candidate for practical implementation.

2.9.4 Modified DE-Test

As we have seen above the DE-test method proposed does not penalize the throughput while reducing a lot the complexity.

In this section we modify the DE-test in order to find a trade off between throughput and complexity. Clearly if we use a test with a very high rejection rate the average complexity can be made as small as we want but the throughput will also go to zero. Therefore by penalizing the DE-test we can reduce the complexity at the price of accepting some throughput degradation. In order to control the trade off η vs C we introduce the penalized SNR $\Gamma_\Delta|_{dB} = \Gamma|_{dB} - \Delta|_{dB}$ where the parameter Δ can be interpreted as an SNR margin of the actual BP decoder over the ideal BP decoder applied to an infinite length LDPC code. In order to find the fixed points of DE, we now iterate the recursion (2-33) substituting Γ_Δ to Γ, thus for each received slot we run the following modified one-dimensional recursion

$$I^l_{out,v} = \frac{1}{M}\sum_{m=1}^{M} F_\lambda\Big(1 - F_\rho\Big(1 - I^{l-1}_{out,v},\, 0\Big)\,,\, \alpha_s\Gamma_\Delta\Big) \tag{2-55}$$

The introduction of the parameter $\Delta \geq 0$, reduces the probability of having DE convergent for a certain vector $(\alpha_1, \ldots, \alpha_m)$ at step m, thus reducing the average number of bursts processed with BP algorithm. Figure 2-23 and 2.10 show the average throughput and complexity as a function of Δ when $R = 0.3$bit/symbol and $\Gamma = 3$dB. It is interesting to notice that using this simple method there are values of $\Delta \simeq 2$dB for which the throughput loss is negligible (~ 0.55bit/sec/Hz vs ~ 0.53bit/sec/Hz) while reducing the complexity of a factor of about 50%.

2.10 CONCLUSIONS

This work extends the previous analysis in [18], where the authors study the performance of HARQ-IR protocol based on infinite length Gaussian code. Here the analysis is extended to ideal infinite length binary codes (random binary and LDPC) and to practical finite block length LDPC codes. We have shown that irregular LDPC ensembles with degree distribution optimized for the standard AWGN channel [22] provide performance very close to the information-theoretic limit given by random binary codes.

Although infinite-length LDPC codes provide near-optimal throughput, practical finite-length LDPC codes incur a considerable performance loss if used in the IR scheme without any countermeasure. We proposed two methods to overcome this problem and to make practical LDPC codes effective for the IR protocol: the first method consists of constructing the LDPC code with a special arrangement of the edges of left-degree 2, in order to improve the FER performance. The second method is based on the concatenation of an outer selective-repeat loop acting on smaller information packet units. We have shown that both methods are able to recover a significant fraction of the loss and provide approximately equivalent performance. Hence, they can be regarded as two valuable alternatives for the system designer.

Finally we have shown an easy to implement method that lower the complexity of the decoder in the context of HARQ protocols. The method consists on introducing a test based on DE prior to decode that prevent using the iterative decoder if it is likely to be non convergent. The proposed algorithm reduces considerably the average complexity without degrading the performance, and modification of the same algorithm allows to achieve all range of trade-offs between throughput and complexity.

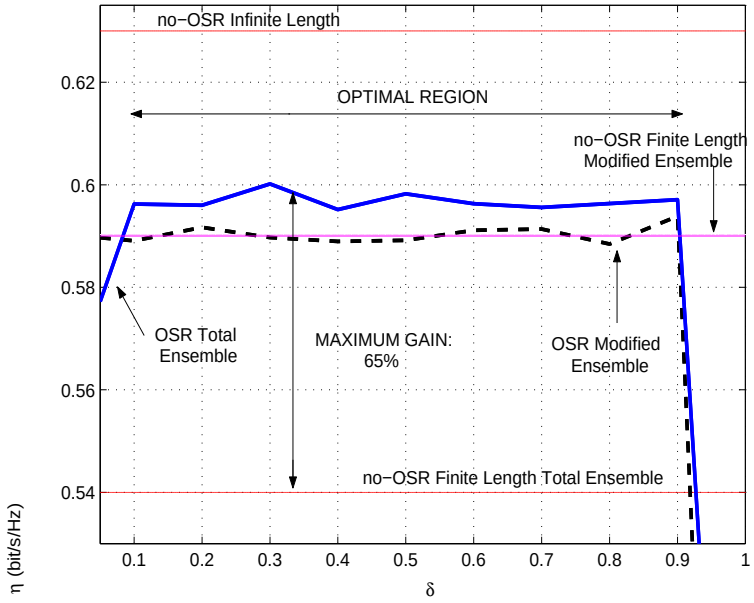


Fig. 2-17. Throughput vs. δ of OSR for $\Gamma = 3$dB and $R = 0.3$bit/symbol for the LDPC codes with $n = 10000$. The throughput without OSR (labeled "no-OSR") for finite and infinite length are shown for comparison as horizontal lines.

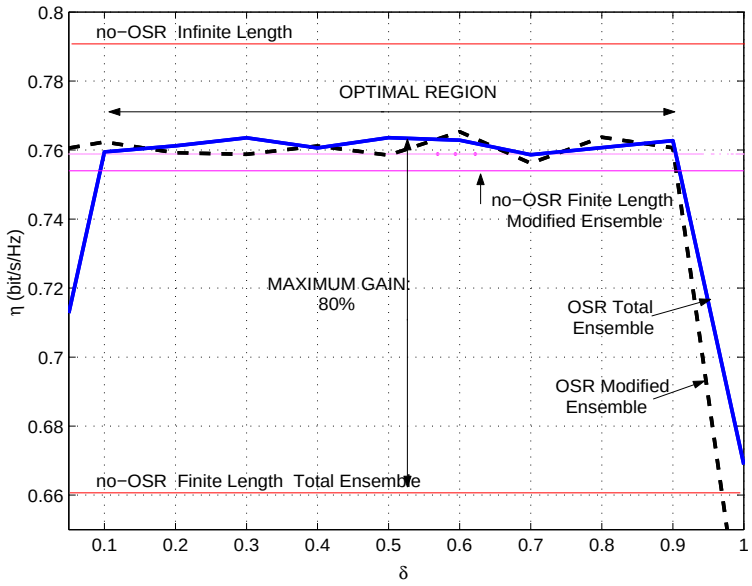


Fig. 2-18. Throughput vs. δ of OSR for $\Gamma = 3$dB and $R = 0.3$bit/symbol for the LDPC codes with $n = 10000$. The throughput without OSR (labeled "no-OSR") for finite and infinite length are shown for comparison as horizontal lines.

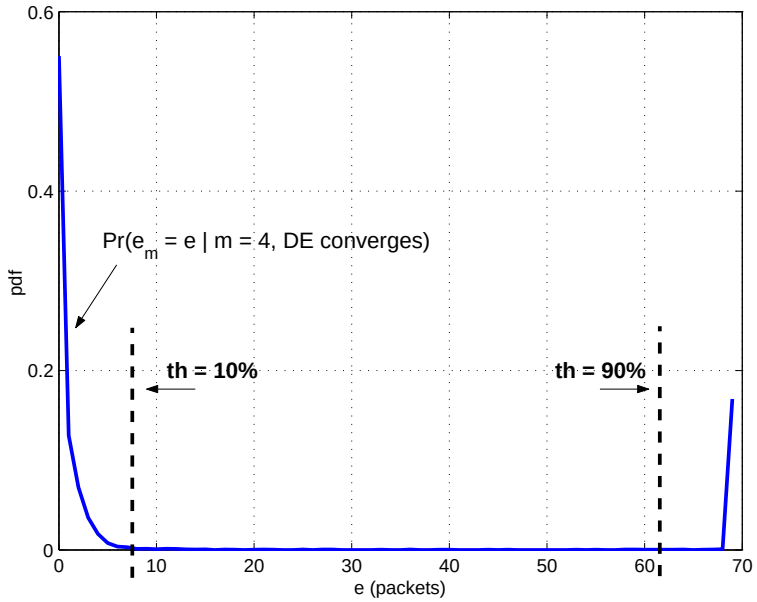


Fig. 2-19. Probability mass function $\Pr(e_m = e \,|\, DE_m \text{ converges})$ for $m = 4$, $R = 0.3$bit/symbol, $\Gamma = 10$dB and $n = 10000$.

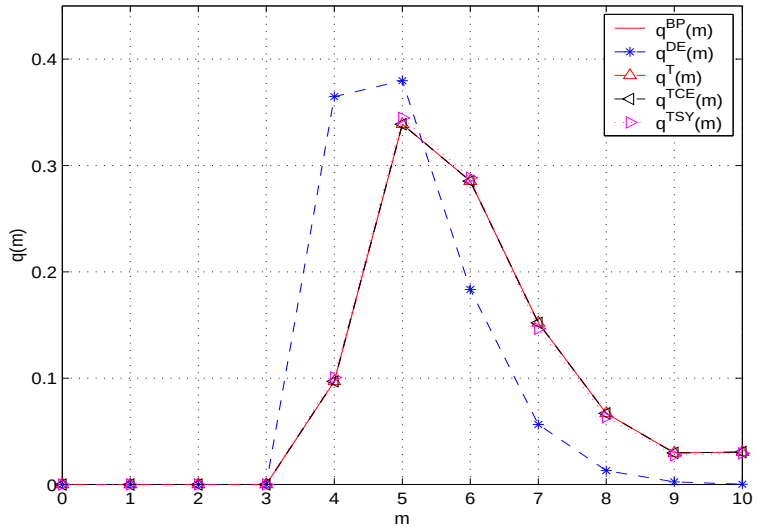


Fig. 2-20. Comparison between $q^{BP}(m)$, $q^{\infty}(m)$, $q^{T}(m)$; $q^{TCE}(m)$ and $q^{TSY}(m)$ represent the DE-test based method when using stopping criteria based on CE and syndrome computation respectively.

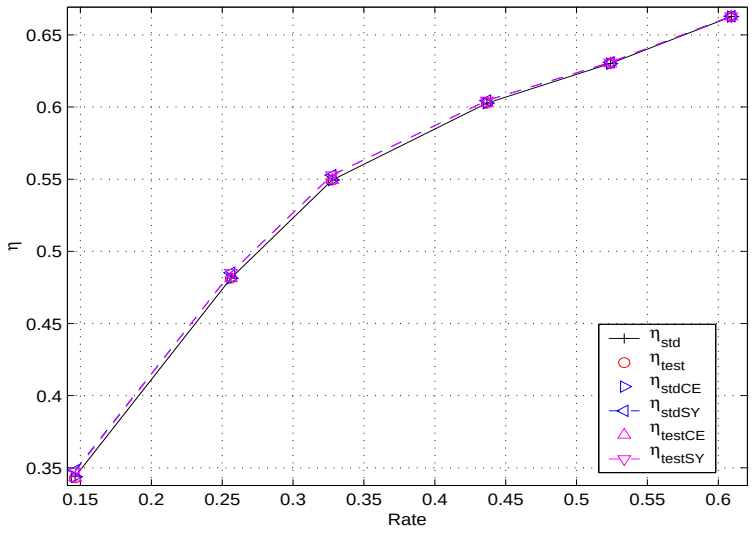


Fig. 2-21. Comparison of average throughput between η_{std}, η_{DE}, η_{stdCE}, η_{stdSY}, η_{test}, η_{testCE} and η_{testSY}.

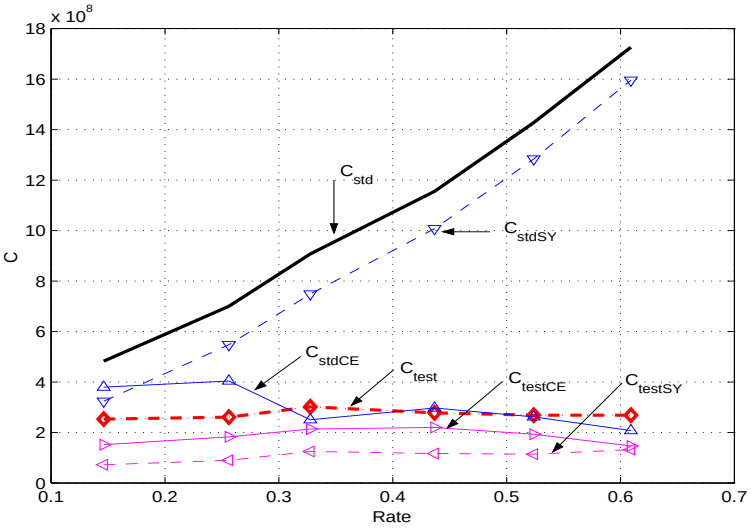


Fig. 2-22. Comparison of average complexity between C_{std},C_{test}, C_{mod}, C_{stdCE}, C_{stdSY}, C_{testCE} and C_{testSY}.

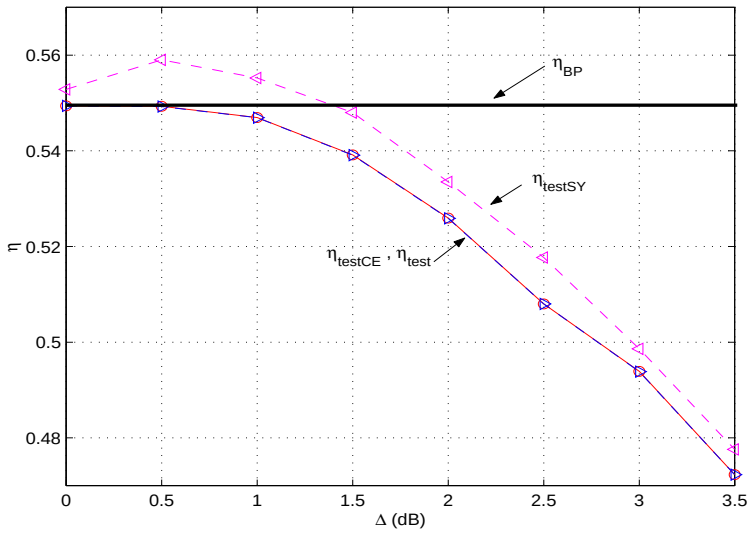


Fig. 2-23. Throughput of the modified DE-test method vs Δ (dB) for $R = 0.3$bit/symbol and $\Gamma = 3$dB.

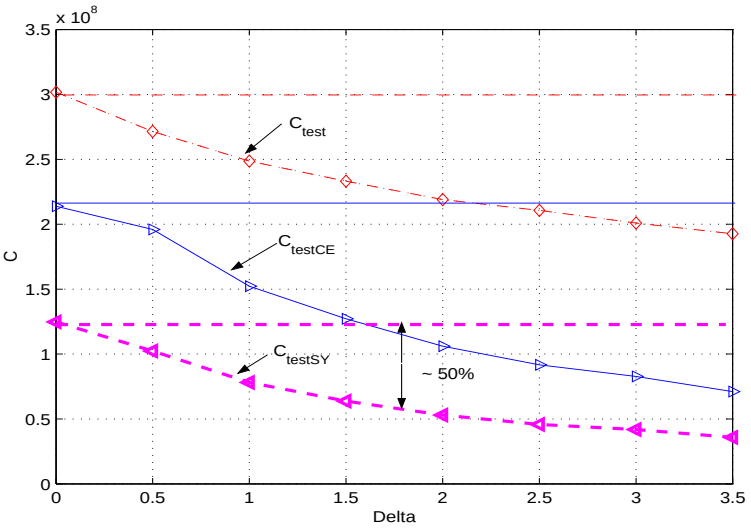


Fig. 2-24. Complexity of the modified DE-test method vs Δ (dB) for $R = 0.3$bit/symbol and $\Gamma = 3$dB.

CHAPTER 3

Feedback Systems for Multicasting Common Information

3.1 INTRODUCTION

Consider a multicast wireless downlink scenario with N users, where the transmission is slotted and the channel is slowly block-fading. Recall that "multicast"means that the base station sends the same information to all the users.

Traditionally the transmitter opens a new connection for each users with an extremely high waste in terms of bandwidth. In this case the transmitter optimizes the transmission parameters for the particular users but it does not exploit the multicast setting.

This chapter deals with the computation of throughput, delay and limiting behavior for large N, when simple HARQ (IR and SR) protocols are considered in a multicast environment.

Several HARQ schemes have been proposed for a point-to-point environment [6]. Recently these point-to-point HARQ techniques have been extended to the case of multicast links, [38, 40, 41, 39, 95], where the authors find ameliorations of standard HARQ schemes to achieves good performance for a particular point to multi-point link. But the study of the achievable performance in terms of throughput for simple HARQ protocols has not been carried out. HARQ protocols better exploit the characteristics of the wireless link such as

independence of the channel seen by different users and the peculiarity of the multicast setting that is to send only common information. In chapter 2, IR and SR protocols have been introduced. For the sake of completeness we recall here that under SR the transmitter sends disjoint copies of the same packets possibly combined at the receiver , while under IR the transmitter sends additional redundancy at each new retransmission.

Since the users are completely symmetric and information is the same for all, the optimal delay-unconstrained coding transmission strategy is trivially given by coding at rate as close as desired to the ergodic capacity of the channel (the same for all users). However if we assume that, because of delay constraints, codewords span a finite number of fading blocks, than reliable communication based on pure FEC coding is impossible. With a FEC based system, in order to achieve vanishing error probability, every codeword must span an arbitrarily large number of fading blocks. This is true even if we back-off in rate and we accept a non vanishing gap from the ergodic capacity. Hybrid ARQ schemes, on the contrary, ensure reliable communications but require explicit ACK/NACK feedback. In the single user case [18] shows that zero error probability can be achieved for finite average delay[1] for any spectral efficiency with fixed gap to capacity by using a HARQ protocol based on IR [18].

However, in the multicast setting, the delay of an HARQ scheme that keeps on sending the same information message until all users have successfully decoded, goes to infinity as the number of users increases. Gopala et al. [42] very recently have analyzed the scaling low, with respect to the number of users, of throughput and delay for three protocols. The first is a "static" SR where they assume that both the transmitter and the receiver have perfect channel state information. At each retransmission, the coding rate is designed in order to target a fixed fraction of users. The protocol reset when all the users are satisfied. They compare this scheme with the IR protocol, assuming, as here, that the transmitter is not aware of the fading coefficients of the users. They show that the average delay of the IR scheme grows to infinity slower than the average delay of the SR protocol. Here, for both SR and IR protocol, we consider that the transmitter has no channel state information. Strictly speaking, these HARQ protocols are not scalable with the number of users, in the sense that the average delay grows to infinity as long as the number of users increases. We show that if we optimize the system in order to achieve a target throughput equal to a given fraction of the ergodic capacity, the delay increases very slowly with the number of users if the gap from capacity is not too small. Hence, if we are not too ambitious in spectral efficiency, the system becomes *practically* scalable up to typical values of the number of users in a cell of a wireless cellular system. In order to make the IR and SR scheme scalable in a strict sense (meaning that the delay tends to a finite limit as $N \to \infty$, for target throughput $\eta < C(\Gamma)$), we have to accept that a fixed fraction of users $x \in (0, 1)$ will not be able to correctly decode the information. We refer to these users as *unfulfilled*. A receiver moves from the

[1] Delay is measured in slots, i.e., in multiples of T.

unfulfilled to the fulfilled state when its channel coefficients are such that it can decode the information, (see Figure 3.3). The transmitter stops sending the current codeword and move to the next codeword if the number of unfulfilled users is not larger than xN. Note that $x = 0$ means to wait until all the users are satisfied.

The main difference between the SR protocol analyzed here and the one in [42] relies on the fact that they consider perfect channel state information while here we do not. This means that the coding rate is adjusted at each retransmission depending on the scheduling algorithm, while here the coding rate is a fixed parameter that does not change from one retransmission to the other.

The IR scheme they consider is the same as here, but they focus on the analysis of the scaling low for $x = 0$.

In this chapter the general expression of the throughput as a function of x, the number of users and the parameter $R = b/L$ is given. Recall that b is the number of information bits per codeword and L is the number of dimensions per slot. We study the behavior of the average delay versus N for given target throughput, when we let R be a design parameter to be optimized. Then, we study the limit $\lim_{N\to\infty} \eta$ for given $x > 0$ and we show that this limit coincides with the spectral efficiency of FEC coding over a number of slots equal to the delay of the IR system (that becomes a deterministic quantity for large number of users) and with error probability precisely equal to x. Hence, for $x > 0$ and $N \to \infty$ the IR scheme has the same performance (in terms of throughput, delay and error probability) of a FEC coding system. We notice also that in this limit, due to the *large-system hardening*, no explicit feedback channel is needed unless the transmitter needs to know, for some reason such as billing, the identity of the unfulfilled users. Hence, under IR protocol considered here, the optimal policy is to accept a fraction of unfulfilled users x that equals the outage probability that minimize the average throughput when the FEC scheme is used.

3.1.1 Summary of the Contributions

- General expression of the throughput as a function of the fraction of unfulfilled users, N and R.
- Analysis of the limiting behavior of the throughput of IR and SR under various system parameters, x, N and R.
- It is shown that for certain values of x when $N \to \infty$, the achievable throughput of IR equals the ergodic capacity at the expense of an average delay that grows to infinity. However if we accept a gap from the ergodic capacity the average delay becomes a

constant. For the other values of x the maximum throughput is always achieved for finite average delay.

- The SR protocol is shown to achieve the maximum throughput always for finite average delay. This maximum is obviously less than the maximum achieved by IR.
- The comparison of IR protocol with FEC based scheme is carried out in the limit of large number of users. It is found that, in this limit, the two schemes are identical, for equal error probability.
- A simple example, based on Birth-Death process [96], gives an idea of buffer size requirement at the receiver side for a streaming application.

3.1.2 Organization of the Work

This chapter is organized as follows: section 3.2 describes the multicast model, section 3.3 compute the throughput of SR scheme with a Markovian model. Then the general expression of the throughput for SR and IR is given by using the Renewal-Reward theory and the limiting behavior for large number of users is analyzed. In section 3.6.1 the comparison between IR and FEC is carried out, and finally we conclude the chapter with an example of buffer requirements calculation at the receiver, when a streaming application is considered.

3.2 SYSTEM MODEL

We consider a wireless multicast system where a sender (base station) wishes to transmit reliably *the same* information to N users. The channel is block-fading Gaussian. Transmission is slotted, every slot spans $L \approx WT \gg 1$ complex dimensions (where W is the two-sided bandwidth and T is the duration of a slot) and the channel fading coefficients for all users are i.i.d., constant on each slot. The signal received by user u on slot s is given by

$$\mathbf{y}_{s,u} = c_{s,u}\sqrt{\Gamma}\mathbf{x}_s + \boldsymbol{\nu}_{s,u} \tag{3-1}$$

where $c_{s,u}$ is the fading coefficient, $\mathbf{x}_s \in \mathbb{C}^L$ is the transmitted signal belonging to Gaussian codebook, and $\boldsymbol{\nu}_{s,k} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I})$ is a complex circularly-symmetric white Gaussian noise. With the normalization $\mathbb{E}[|c_{s,k}|^2] = \frac{1}{L}\mathbb{E}[|\mathbf{x}_s|^2] = 1$, Γ takes on the meaning of average received SNR. The transmitter is not aware of the fading channel coefficients, while the receivers have perfect channel state information.

For later use we define here the ergodic capacity,

$$C(\Gamma) = \mathbb{E}\left[\log(1 + \Gamma\alpha)\right] \tag{3-2}$$

where $\alpha \sim f_\alpha(z)$ is a random variable distributed as the channel power gain $|c_{s,u}|^2$. To simplify the presentation we assume that $f_\alpha(z) = e^{-z}1\{z \geq 0\}$, i.e., Rayleigh fading, where $1\{\cdot\}$ is the indicator function. The results of this work hold under mild conditions on the fading distribution and apply to other fading distributions.

In the N users case, due to the fact that information is the same for all users and that the users are completely symmetric, we will measure spectral efficiency (or "throughput") from the base station viewpoint. Letting $b(s)$ be the number of transmitted information bits up to slot s, the throughput is given by

$$\eta = \lim_{s\to\infty} \frac{b(s)}{sL} \quad \text{bit/dim.} \tag{3-3}$$

We expect that the average delay necessary to achieve any desired throughput η tends to infinity as N increases. Intuitively, the probability that at least one user out of N is not able to decode successfully after m slots tends to 1 for any finite m and $N \to \infty$. Hence, the transmitter will eventually send "for ever" additional redundancy of the same codeword. In the next section we provide an exact throughput analysis of the SR and IR scheme with N users and show that indeed this intuition is correct. However, it is interesting to notice that the delay necessary to achieve throughput $\eta = (1-\delta)C(\Gamma)$, increases quite slowly if δ is not too small. Hence, we argue that for typical values of N in a cellular system, and for target throughputs not too close to the ergodic capacity, the IR scheme is a viable solution for reliable multicast.

3.3 MARKOV MODEL

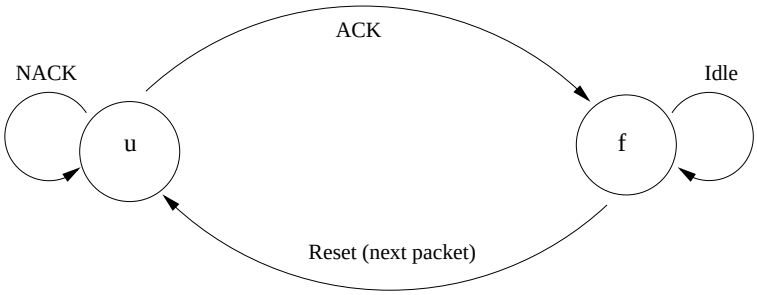


Fig. 3-1. Receiver model.

In the following we consider the SR scheme and we compute the throughput at the base station using a Markovian model. Note that under the SR scheme the transmitter encodes

b information bits by using a channel code with codebook $\mathcal{C} \in \mathbb{C}^L$, length L and rate $R = b/L$ bit/symbol. The codeword is sent over one slot. We associate to each user an error probability, a given by the probability that the mutual information per input symbol on slot s ΔI_s is less or equal than the coding rate [18], i.e.

$$a = \Pr(\Delta I_s < R) = \Pr(\log_2(1+\gamma\alpha_s) < R) = F_\alpha\left(\frac{2^R-1}{\gamma}\right) \tag{3-4}$$

where $\alpha_s = |c_s|^2$ is the fading power gain and $F_\alpha(.)$ is the cdf of α; equation (3-4) is justified by assuming $\mathcal{C}$ a Gaussian random code and L sufficiently high.

Define U a random variable that counts the number of unfulfilled users. The system keeps sending packets until the number of unfulfilled users is less or equal to a fraction of the total number of users, i.e. it is less or equal than n. The system can be modeled as a Markov chain with $N-n+1$ states where each state v_i for $i = 0, \dots N-n-1$ represents U, i.e $s_i = N-i$; the last state s_{N-n+1} represents the successful event, the number of unfulfilled users is $U \leq n$. We call $s_{N-n+1} = S$. Whenever the "successful" state is reached the system is reset and the transmitter begins sending a new codeword. Figure 3.3 shows the Markov chain for $N = 2$ and $n = 0$ users.

$$\mathbf{\Phi} = \begin{bmatrix} a^2 & 2a(1-a) & (1-a)^2 \\ 0 & a & 1-a \\ 1 & 0 & 0 \end{bmatrix} \tag{3-5}$$

and solving $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{\Phi}$ using the normalization property of the stationary probability vector, we obtain $\pi_S = \frac{1-a^2}{(2+2a-a^2)}$ and consequently the throughput is given by $\eta_{2,0} = R\frac{1+2a}{1-a^2}$.

The generalization for an arbitrary value of N and n is cumbersome. Moreover for the IR the Markovian description is much more complicated because the state is a N-dimensional vector $\in \mathbb{R}_+^N$ where each element is the mutual information accumulated up to a certain step. This becomes a N-dimensional discrete time Markov process. Fortunately the throughput of SR and IR can be computed by resorting the Renewal theory [97, 18].

3.4 THROUGHPUT ANALYSIS BASED ON RENEWAL THEORY

The SR scheme can be seen as a particular case of IR scheme, so in the following we apply the Renewal theory to the IR scheme and we particularize it for the SR case.

We define the event $\mathcal{E}_m = \{$The transmitter stops transmitting the current codeword after m slots$\}$. This is seen to be a recurrent event, since the system resets. Let U_m denote the

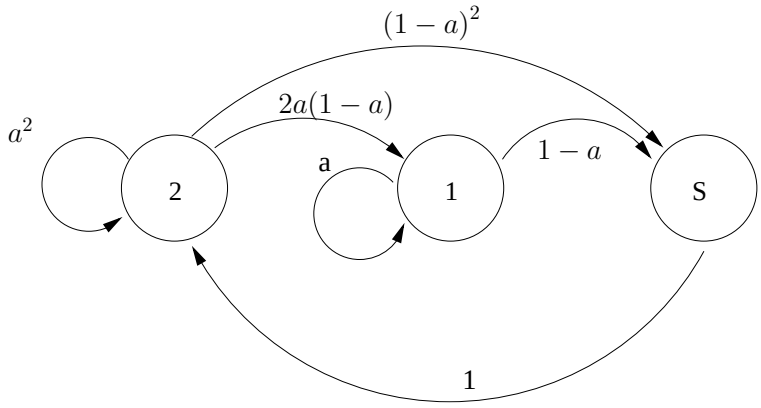


Fig. 3-2. Markov Chain $N = 2$, $n = 0$.

number of unfulfilled users after m transmitted blocks of the current codeword, and define the event $\mathcal{D}_m = \{U_m \leq xN\}$. Then,

$$\mathcal{E}_m = \bar{\mathcal{D}}_1 \cap \cdots, \cap \bar{\mathcal{D}}_{m-1} \cap \mathcal{D}_m \tag{3-6}$$

(the bar indicates complement event). The probability of the recurrent event is given by

$$\Pr(\mathcal{E}_m) \;=\; \tilde{p}(m-1) - \tilde{p}(m) \tag{3-7}$$

where we define $\tilde{p}(m) \;\stackrel{\Delta}{=}\; \Pr\big(\overline{\mathcal{D}}_1,\, \ldots,\, \overline{\mathcal{D}}_{m-1},\, \overline{\mathcal{D}}_m\big)$.

From the renewal theorem [97], the throughput seen at the transmitter is given by the ratio between the number of information bits/dimension R (it is the reward associated to the occurrence of $\mathcal{E}_m$) and the average inter-renewal time $\bar{\tau} = \mathbb{E}[\tau]$, defined as the number of slots between two occurrence of the recurrent event. We have

$$\begin{aligned} \eta(N, x, R, \Gamma) = \frac{R}{\bar{\tau}} &= \frac{R}{\sum_{m=1}^{\infty} m \Pr(\mathcal{E}_m)} \\ &= \frac{R}{1 + \sum_{m=1}^{\infty} \tilde{p}(m)}. \end{aligned} \tag{3-8}$$

Notice that $\bar{\tau}$ is also the average delay measured in slots.

By observing that $\overline{\mathcal{D}}_1 \;\supseteq\; \ldots \;\supseteq\; \overline{\mathcal{D}}_{m-1} \;\supseteq\; \overline{\mathcal{D}}_m$, we obtain

$$\tilde{p}(m) = \Pr\big(\overline{\mathcal{D}}_m\big) = \Pr(U_m > xN)\,. \tag{3-9}$$

Let $\Delta I_{s,u}$ denote the average mutual information (in bits per dimension) in slot s for user u. Assuming a Gaussian codebook, we have

$$\Delta I_{s,u} = \frac{1}{L} I(\mathbf{x}_{s,u}; \mathbf{y}_{s,u} | c_{s,u}) = \log_2(1 + \Gamma \alpha_{s,u})$$

The mutual information accumulated up to slot s by user u is given by

$$\frac{1}{m} \sum_{s=1}^{m} \Delta I_{s,u}. \tag{3-10}$$

Following [18], for sufficiently large L, user u decodes successfully at step m if the mutual information (3-10) is larger than the effective coding rate, given by R/m, while it cannot decode successfully if the mutual information is below R/m. Define the event $\mathcal{A}_{u,m} = \{\text{The user } u \text{ decodes at step } m\}$, it follows that

$$\begin{aligned} \Pr(\mathcal{A}_{u,m}) &= \Pr\left(\frac{1}{m} \sum_{i=1}^{m} \Delta I_{i,u} \geq \frac{b}{Lm} \right) = \Pr\left(\sum_{i=1}^{m} \Delta I_{i,u} \geq \frac{b}{L} \right) \\ &= 1 - p(m) \end{aligned} \tag{3-11}$$

where $p(m)$ is defined, as in chapter 2, as

$$p(m) \stackrel{\Delta}{=} \Pr\big(\overline{\mathcal{A}}_{u,1}, \overline{\mathcal{A}}_{u,2}, \ldots \overline{\mathcal{A}}_{u,m}\big) \stackrel{(a.)}{=} \Pr\big(\overline{\mathcal{A}}_{u,m}\big) = \Pr\left(\sum_{i=1}^{m} \Delta I_{u,i} \leq R \right) \tag{3-12}$$

and where $(a.)$ is because $\overline{\mathcal{A}}_{u,1} \supseteq \overline{\mathcal{A}}_{u,2} \supseteq \cdots \supseteq \overline{\mathcal{A}}_{u,m}$.

The probability $\tilde{p}(m)$ has the following expression

$$\begin{aligned} \tilde{p}(m) &= \Pr(U_m > xN) \\ &= \Pr\left(\sum_{u=1}^{N} 1\left\{ \sum_{s=1}^{m} \Delta I_{s,u} \leq R \right\} > xN \right) \\ &\stackrel{(a.)}{=} \sum_{k=0}^{N - \lceil xN \rceil} \binom{N}{k} (1 - p(m))^k \, p(m)^{N-k} \end{aligned} \tag{3-13}$$

where $(a.)$ follows by noticing that the random variables $1\{\sum_{s=1}^{m} \Delta I_{s,u} \leq R\}$ for $u = 1, \ldots, N$ are independent Bernoulli random variables with parameter $p(m)$ defined in (3-12). By using (3-7) and (3-13) in (3-8), the throughput is given by

$$\eta(N, x, R, \Gamma) = \frac{R}{1 + \sum_{m=1}^{\infty} \sum_{k=0}^{N - \lceil xN \rceil} \binom{N}{k} (1 - p(m))^k \, p(m)^{N-k}}. \tag{3-14}$$

Notice that the probabilities $p(m)$ depends on both R and Γ.

When the total codeword of length L is retransmitted instead of additional redundancy, we obtain the SR scheme. The SR scheme can be seen as a special case of the general IR scheme where the codewords are obtained by the concatenation of a code of length L with an arbitrarily long repetition code. Hence, the throughput of SR lower-bounds the throughput of IR. For SR, user u is unfulfilled after m slots if the event $\cap_{s=1}^{m}\{\Delta I_{s,u} \leq R\}$ occurs. Since the $\Delta I_{s,u}$ are i.i.d. random variables, we obtain an explicit expression for the probability $p(m)$ as

$$p(m) = \left(1 - e^{-\frac{2^R-1}{\gamma}}\right)^m \tag{3-15}$$

where we have used the fact that $\alpha_{s,u}$ is exponentially distributed.

It is easy to verify that for $N = 2$, $n = 0$ we obtain again $\eta(2, 0, R, \gamma) = R\frac{1+2a}{1-a^2}$ in accordance with the previous Markov chain analysis.

3.5 THROUGHPUT FOR FINITE NUMBER OF USERS N

When $x = 0$, the throughput in (3-14) is given by

$$\eta(N, 0, R, \Gamma) = \frac{R}{\sum_{m=0}^{\infty}\left[1 - (1 - p(m))^N\right]}. \tag{3-16}$$

It is easy to see that, for any $R < \infty$,

$$\lim_{N\to\infty} \eta(N, 0, R, \Gamma) = 0. \tag{3-17}$$

In [42] the authors show that for IR scheme $\eta = \Theta\left(\frac{\log\log N}{\log N}\right)$.

The limit (3-17) is valid for all $R < \infty$. Hence, by letting first $N \to \infty$ and then $R \to \infty$ it still holds. On the contrary, by following in the footsteps of the analysis in [18], it is not difficult to see that for any fixed $N < \infty$ we obtain

$$\lim_{R\to\infty} \eta(N, 0, R, \Gamma) = C(\Gamma). \tag{3-18}$$

where also the average delay tends to infinity $\bar{\tau} = \Theta(R)$. Hence, by reversing the order of the limits and letting first $R \to \infty$ and then $N \to \infty$ we find that the throughput is not vanishing, although the average delay still tends to infinity.

At this point it is natural to ask about the behavior of average delay with respect to N when we let the throughput equal to a given fraction of the ergodic capacity, i.e., when we set R as the solution of the equation

$$\eta(N, 0, R, \Gamma) = (1 - \delta)C(\Gamma)$$

Figure 3-3 shows R and the resulting $\bar{\tau}$ needed to achieve the above equality for $\delta \sim 3\%, 7\%, 15\%$. As expected, the average delay grows very fast if the target throughput is close to the ergodic capacity. On the contrary, it increases very slowly (except for an initial transient where it increases roughly linearly) when we allow for a certain non-negligible gap from capacity. This gap from capacity is the price to pay to achieve reliable communications (vanishing error probability) in the block-fading channel under a delay constraint, with this simple protocol.

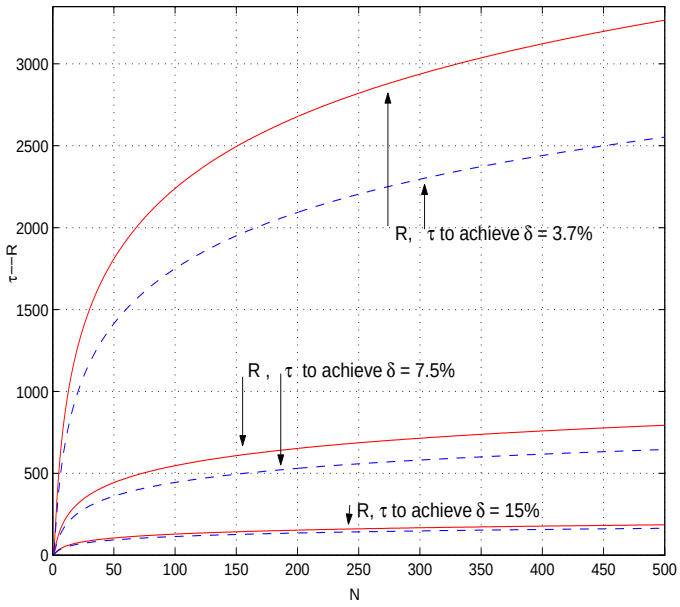


Fig. 3-3. R and τ s.t $\eta(N, 0, R, 3dB) = (1 - \delta)C(\Gamma)$ vs N for $\Gamma = 3$dB for different value of δ.

3.6 Limiting Throughput for Large Number of Users

This section analyzes the behavior of (3-14) in the limit of large number of users ($N \to \infty$) with $n/N = x$ where $x \geq 0$ when we schedule the transmission to a fraction of users

that experiences favorable channel conditions. The reset of the renewal reword process is adjusted such that each transmission by the base station can be decoded only by a fraction x of users.

The results of this section are manifold. First the best achievable throughput is analyzed and it is shown that for particular values of x the ergodic capacity is achievable but at the expense of large delays. On the other end if the delay is fixed then all possible throughput $\eta \in (0, \infty)$ can be achieved depending on x. However, we can define the average throughput, computed from the user point of view, as $\bar{\eta} = (1-x)\eta$. It is easy to show that there is a particular x that yields the best $\bar{\eta}$.

Define

$$\mathcal{V}(p, N, x) = \sum_{k=0}^{N-\lceil Nx \rceil} \left(\begin{array}{c} N \\ k \end{array} \right) (1-p)^k (p)^{N-k}$$

We have

$$\begin{aligned} \eta_\infty(x, R, \gamma) &= \lim_{N\to\infty} \eta(N, x, R, \gamma) \\ &= \lim_{N\to\infty} \frac{R}{1 + \sum_{m=1}^{\infty} \mathcal{V}(p(m), N, x)} \\ &= \frac{R}{1 + \lim\limits_{N\to\infty} \sum_{m=1}^{\infty} \mathcal{V}(p(m), N, x)}. \end{aligned} \tag{3-19}$$

It is straightforward to see that it is possible to exchange the limit in (3-19) with the infinite summation w.r.t m. Note that $\mathcal{V}(p(m), N, x)$ can be seen as the cdf of a Binomial random variable X_m computed in $N - \lceil xN \rceil$, i.e $X_m \sim \mathrm{Bin}(N, 1-p(m))$. Hence

$$\lim_{N\to\infty} \mathcal{V}(p(m), N, x) = \lim_{N\to\infty} \Pr(X_m \leq N - \lceil xN \rceil)$$

The mean and the variance of X_m are given by $\mu_X = \mathbb{E}[X] = N(1-p(m))$ and $\sigma_X^2 = N(1-p(m))\,p(m)$. Appendix 7.1 shows the following Lemma

Lemma 1: : The limit for large number of users of the function $\mathcal{V}(p(m), N, x)$ is given by

$$\lim_{N\to\infty} \mathcal{V}(p(m), N, x) = 1\{x \leq p(m)\} - \frac{1}{2}\delta(x - p(m)) \tag{3-20}$$

□

Eventually, the limiting throughput is given by

$$\eta_\infty(x, R, \Gamma) = \frac{R}{1 + \sum_{m=1}^{\infty} 1\{x \leq p(m)\} - \frac{1}{2}\delta(x - p(m))} \tag{3-21}$$

For the SR scheme, using (3-15), we obtain the explicit formula

$$\eta_\infty(x,R,\Gamma)=\frac{R}{1+\lfloor\frac{\log x}{\log a}\rfloor-\frac{1}{2}\delta\Big(\lfloor\frac{\log x}{\log a}\rfloor-\frac{\log x}{\log a}\Big)} \tag{3-22}$$

where $\lfloor . \rfloor$ means the integer part and $a \stackrel{\Delta}{=} \left(1-e^{-\frac{2^R-1}{\Gamma}}\right)$.

For each m, $p(m)$ is an increasing function of R. For $k=0,1,2,\ldots$, we define

$$R(k) \stackrel{\Delta}{=} \sup\left\{R \;:\; \sum_{m=1}^{\infty} 1\{x \le p(m)\} - \frac{1}{2}\delta(x-p(m)) \le k\right\} \tag{3-23}$$

The following result holds:

Theorem 2: The supremum over $R \ge 0$ of $\eta_\infty(x,R,\Gamma)$ is given by $R(k)/(1+k)$ for some $k=0,1,\ldots$, that in general depends on x and Γ. □

The proof is given in Appendix 7.2.

For $N=1$, [18] shows that the IR throughput is an increasing function of R and that

$$\sup_{R\ge 0}\eta(1,0,R,\Gamma) = \lim_{R\to\infty}\eta(1,0,R,\Gamma) = C(\Gamma) \tag{3-24}$$

Appendix 7.3 shows that the limiting throughput for large number of users of IR protocol, for a positive fraction of unfulfilled users is a constant that for same particular values of x equals the ergodic capacity.

Theorem 3: For independent Rayleigh fading SNR Γ and IR protocol, define $G_m(z)$ the cdf of the random variable $\frac{1}{m}\sum_{i=1}^{m}\Delta I_{i,u}$. Define $x_d \stackrel{\Delta}{=} \min(G_m(C(\Gamma)))$. Then for all $x \in (0,x_d)$, $\eta_\infty(x,R,\Gamma)$ is increasing with R. Therefore, $\sup_{R\ge 0}\eta_\infty(x,R,\Gamma)$ is achieved for $R\to\infty$ and $\bar{\tau}\to\infty$, and it is equal to $C(\Gamma)$. Also, for all $x \in (x_d,1)$ $\sup_{R\ge 0}\eta_\infty(x,R,\Gamma)$ is achieved for finite delay. □

The derivation of a closed form expression for x_d as a function of Γ is not straightforward. Nevertheless, it can be observed that for a wide range of SNR Γ x_d is very close to 0.5.
Note that Theorem 2 guarantees the existence of $\eta_{\sup,x}$ while Theorem 3 gives the value of $\eta_{\sup,x}$ and the rate and average delay necessary to achieve it.

For the SR protocol, Appendix 7.4 shows the following theorem.

Theorem 4: For the SR protocol, $\sup_{R\geq 0} \eta_\infty(x, R, \Gamma)$ is always achieved for finite R and delay. In particular the optimal R is given by $R(k) = F(k+1) - \epsilon$ where k is found as the index that maximizes the following sequence

$$b[i] = \frac{1}{i+1}[F(i+1) - \epsilon] \tag{3-25}$$

and

$$F(i) = \left[\log_2\left(1 - \Gamma \log\left(1 - x^{\frac{1}{i}}\right)\right)\right]$$

for arbitrarily small ϵ. □

Figure 3-4 shows the behavior of the sequence $b[i]$ vs i for $\Gamma = 0$dB parametrized in x, when $\epsilon = 10^{-3}$. We can clearly see that for all x the maximum is a small value of i. It becomes 0 when $x \to 1$.

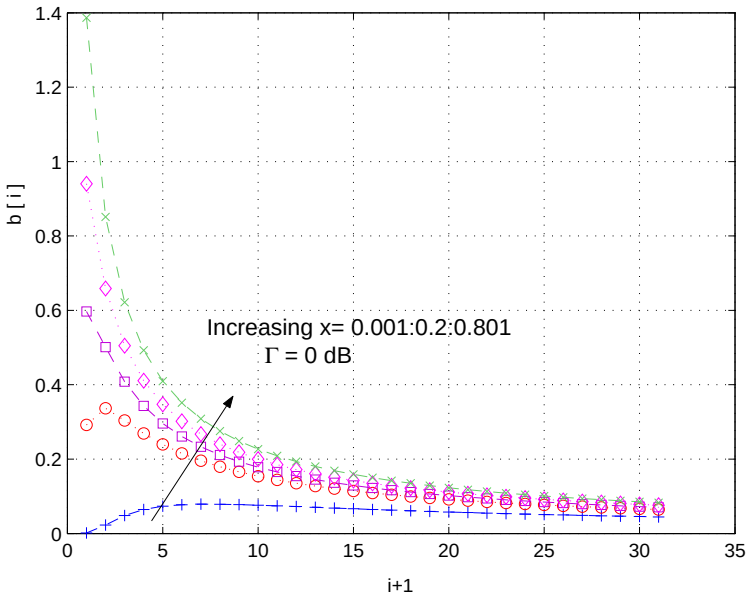


Fig. 3-4. Sequence $b[i]$ vs $i+1$ for $\Gamma = 0$dB parametrized in x, $\epsilon = 10^{-3}$.

3.6.1 Comparison with the "FEC only" System

In this section the comparison between IR scheme and a system that broadcasts the same information to all the users using only channel code without ARQ is carried out. This scheme

is referred to as "FEC only" system. The information bits are coded using a code of length BL symbols transmitted over B different slots. The throughput seen by the transmitter is simply given by the code rate $\eta = R_{FEC,x}$ for which the average error probability is x. Note that the "FEC only" system has a fixed delay equal to the number of slots used to send the codeword, B, while SR and IR have a variable delay τ.

In order to make a fair comparison we consider the average delay of the IR (or SR) equal to the fixed delay of the "FEC only" scheme, we set $B = \lfloor \bar{\tau} \rfloor$ where $\bar{\tau}$ is the average delay (in slots) of the IR system, and such that users have error probability x. This is a fair comparisons for large N, since both the delay and the fraction of unfulfilled users become deterministic and equal for the two systems (if $\bar{\tau}$ is an integer). The following theorem holds.

Theorem 5: Let $\bar{\tau}$ be the integer delay of IR protocol such that the error probability is x. Consider a "FEC only" system with $B = k + 1$ and let the error probability be equal to x. Then the spectral efficiency of "FEC only" system equals the throughput of the IR scheme in the limit for large number of users, i.e.

$$\eta^{FEC} = \eta_\infty(x, R(k), \Gamma) \tag{3-26}$$

□

The proof follows from the fact that the spectral efficiency of FEC coding satisfies the equation

$$\Pr\left(\sum_{s=1}^{B} \Delta I_{s,u} \leq B\eta^{\mathrm{FEC}}\right) = x \tag{3-27}$$

We notice that $R(k)$ defined in (3-23) must satisfy the equation

$$\Pr\left(\sum_{s=1}^{k+1} \Delta I_{s,u} \leq R(k)\right) = x \tag{3-28}$$

By comparing (3-27) and (3-28) we conclude that for all integer delays $\tau = 1 + k$, $R(k) = (1 + k)\eta^{\mathrm{FEC}}$ and, from Theorem 1, $\eta_\infty(x, R(k), \Gamma) = \eta^{\mathrm{FEC}}$. In particular, since the throughput is maximized for some k, we find that the maximum throughput (with respect to delay, for x and Γ given) of the IR and of the FEC coding systems is identical in the limit of a large number of users.

3.7 RESULTS

It is interesting to note that, with a very simple binary feedback scheme from each user, whenever a positive fraction x of users to be unfulfilled is accepted, the throughput seen by

the transmitter is positive even for an infinite number of users.

Consider, without loss of generality, the SR case: the throughput approaches R (the coding rate) whenever the parameter a goes to zero (lossless channel), or x goes to 1 (we accept 100% of unfulfilled users), and it goes to 0 whenever we want all the users to be fulfilled ($x \to 0$). Note also that $\eta_{\infty,x} = R$ when $\frac{\log(x)}{\log(a)} < 1$, that means when $a < x$.

Figure 3-5 and 3-6 show the comparison of the throughput maximized over R versus x, in the limit for $N \to \infty$, for $\Gamma = 3, 10$dB for the two protocols SR and IR.

In the case of $\Gamma = 3$dB $x_d = 0.5$. Moreover for $x \in (0.552, 1)$ $\sup_R \eta_\infty(x, R, \Gamma) = \log_2(1 - \Gamma \log(1 - x))$ achieved with average delay $\bar{\tau} = 1$. For $x \in (x_d, x_u)$ it is only possible to conclude that $\bar{\tau}$ is a decreasing function of x.

When $\Gamma = 10$dB, instead $x_d = G_1(C(\Gamma)) = 0.48$, and moreover $G_1^{-1}(x) > G_m^{-1}(x)$ for all $x > x_d$. This tells that for $x > x_d$ $\sup_R \eta_\infty(x, R, \Gamma) = \log_2(1 - \Gamma \log(1 - x))$ achieved with average delay $\bar{\tau} = 1$.

Since for high x the optimal throughput is obtained for $\bar{\tau} = 1$, and the throughput of the SR protocol lower-bounds the performance of IR, then SR and IR achieve the same results.

Figure 3-7 shows $\eta_\infty(x, R, \Gamma)$ as a function of R for fixed x and $\Gamma = 3$dB. The throughput is a non decreasing function of the rate for small x. On the contrary, for large x there exist a finite value of R which maximizes the throughput, and this maximum is larger than the ergodic capacity, as stated in theorem 4.

Figure 3-8 represents the optimal throughput $\sup_k \eta_\infty(x, R(k), \Gamma)$ for SR and the optimal constrained throughput for the IR, when we set $\bar{\tau}_{IR} = \bar{\tau}_{SR}$. Also in that suboptimal case the IR still gains with respect to SR in region $0 < x < 0.24$. The figure also shows the rate $R(k)$ for IR and SR and the average delay for SR $\bar{\tau}_{SR}$.

Figure 3-9 shows the convergence of $\sup_k \eta(N, x, R(k), \Gamma)$ vs N for SR when $x = 0.2$. The rate of convergence slows down as long as we select R equal to the optimal value, $R(k)$. Further analysis on the number of users necessary to achieve a certain error between $\sup_k \eta_\infty(x, R(k), \Gamma)$ and $\sup_k \eta(N, x, R(k), \Gamma)$ shows that it depends on $\frac{1}{(p(k+1)-x)^2}$ that is in general very small ($p(k+1)$ also depends on $R(k)$). Therefore, as long as we accept a small rate loss $R^* = R(k) - \epsilon$ the rate of convergence increases.

Figures 3-10 and 3-11 show the throughput $\eta(N, x, R, \Gamma)$ for $x = 10^{-2}$, $\Gamma = 3$dB and $\bar{\tau} = 10, 50$ respectively. R is set equal to $R(\bar{\tau} - 1)$ defined in (3-23). We can notice that FEC and IR coincides only in the limit of large number of users even though they are always quite close for small N. Surprisingly the throughput of IR is not a decreasing function of N. This is due to the fact that we allow $x > 0$ and that the delay becomes a deterministic

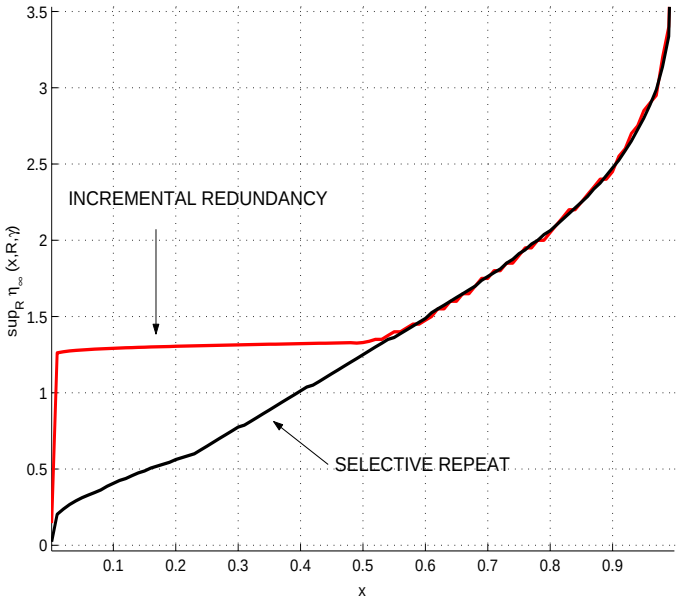


Fig. 3-5. $\sup_k \ \eta_\infty(x, R, \Gamma)$ vs x, for $\Gamma = 3$dB.

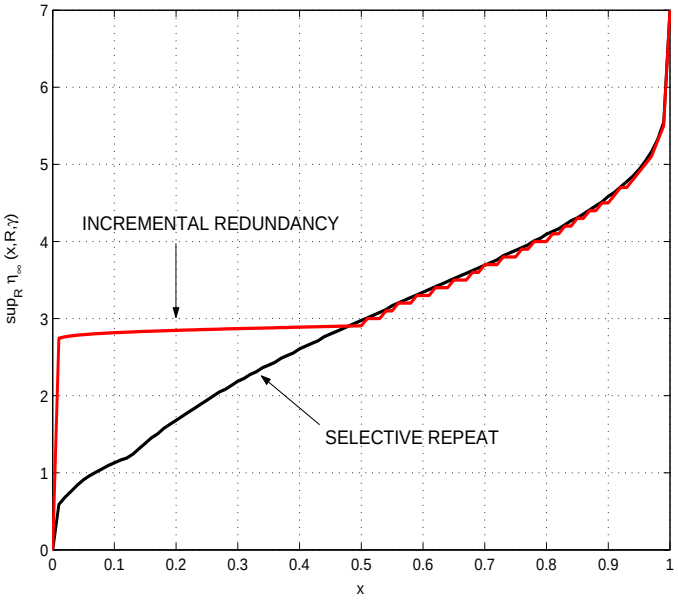


Fig. 3-6. $\sup_k \ \eta_\infty(x, R, \Gamma)$ vs x for $\Gamma = 10$dB.

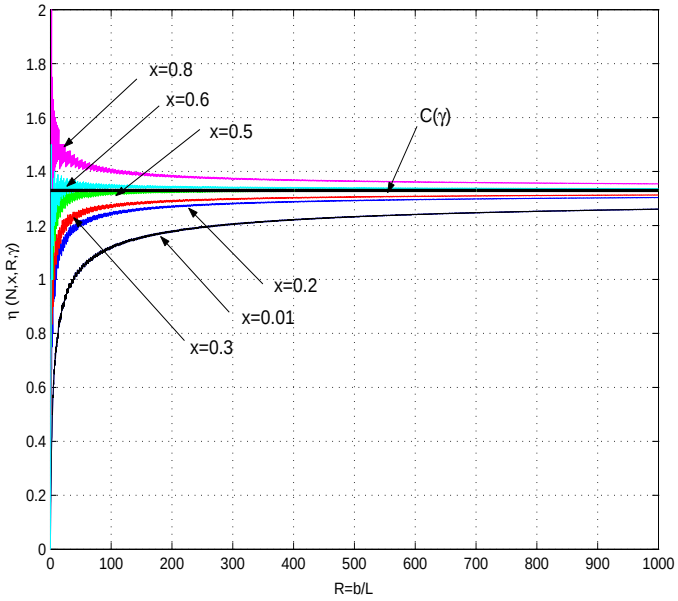


Fig. 3-7. $\eta_\infty(x, R, \Gamma)$ vs $R = b/L$ for different value of x for $\Gamma = 3$dB.

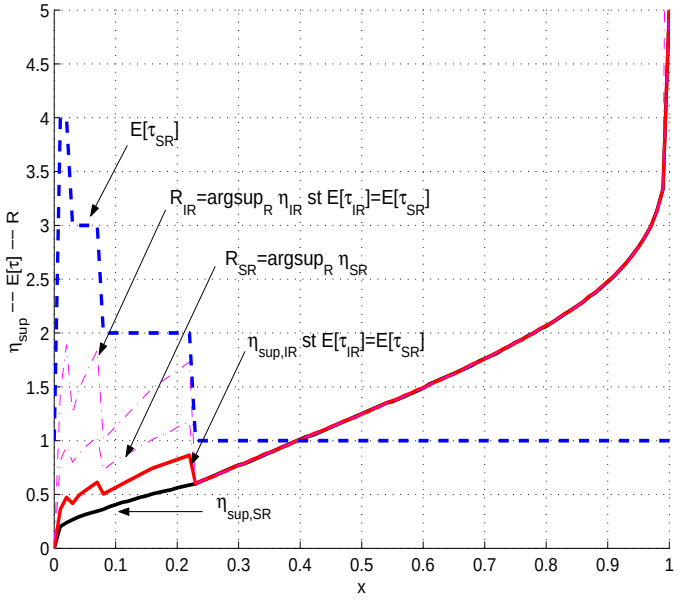


Fig. 3-8. $\sup_k \eta_\infty(x, R(k), \Gamma)$ for SR and IR vs x when we fix $\bar{\tau}_{IR} = \bar{\tau}_{SR}$. We plot also the rate $R(k)$ and $\bar{\tau}_{SR}$.

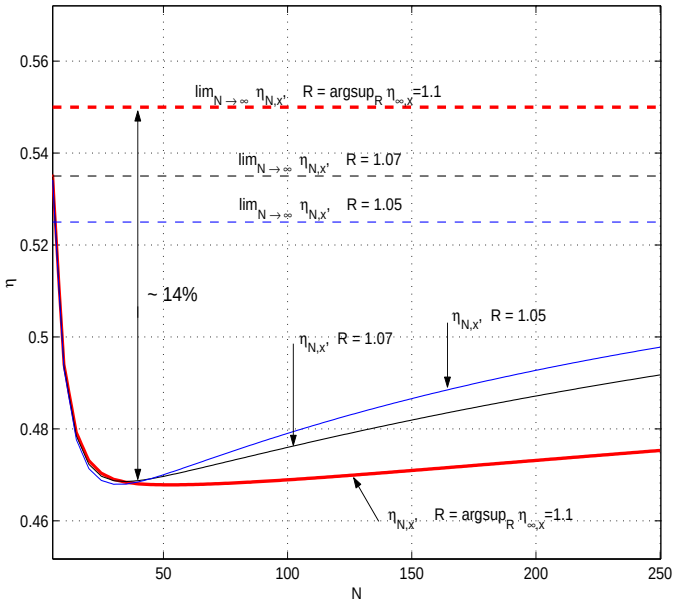


Fig. 3-9. $\sup_k \eta(N, x, R(k), \Gamma)$ for SR and $\sup_k \eta_\infty(x, R(k), \Gamma)$ vs N. Convergence vs limit for SR when $x = 0.2$, $R(k) = 1.06$bit/symbols and $\Gamma = 3$dB.

variable only for large number of users while for small N this is a random variable that depends on N.

Finally consider a fixed delay B and the IR protocol. $\eta(\infty, x, R(B), \Gamma)$ is an increasing function of x. However we can compute the total average throughput as

$$\eta_{\text{tot}} = (1 - x)\eta(\infty, x, R(B), \Gamma)$$

It is possible to show that the $\lim_{x\to 1} \eta_{\text{tot}} = 0$ implying the existence of x^* that yields optimal total throughput. This turns out to be a known problem, i.e find the optimal outage probability that minimize the total throughput when a FEC system is considered.

3.8 Dimensioning the Pre-fetching buffer for streaming application

Consider an application like video streaming where different users request the same information. Traditionally the system opens a connection for each user and it adapts the transmission parameters to the channel condition of that user. This is very expensive in

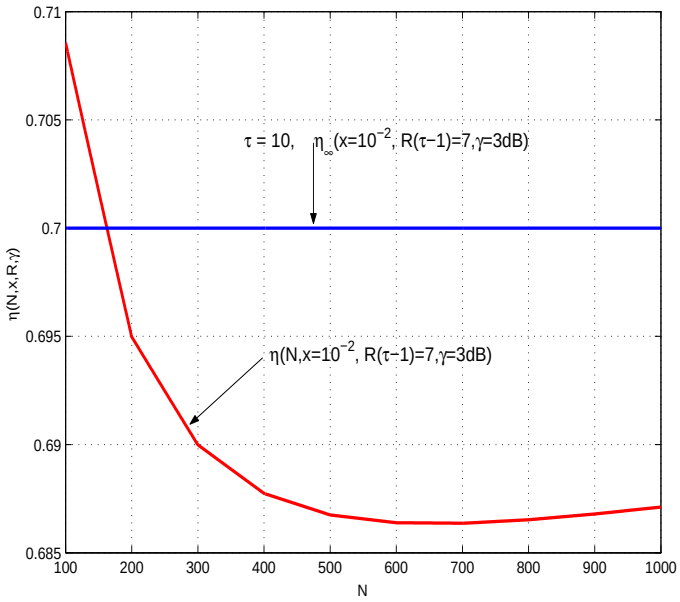


Fig. 3-10. $\eta_\infty(x, R, \Gamma)$ and $\eta(N, x, R, \Gamma)$ vs N for $x = 10^{-2}$, $\Gamma = 3$dB $\bar{\tau} = 10$ and $R = R(\bar{\tau} - 1)$.

terms of bandwidth and can cause problems when a new user tries to access to the same service. Eventually the system refuses to open the connection for the new user. In this case the system is not exploiting the multicast setting and in particular the fact that all the users needs the same information. A better solution, in terms of bandwidth efficiency would be to share the bandwidth among the users at the expense of a penalty in throughput of each user. Hence, we can consider the use of IR or "FEC only" protocol with a fixed fraction of users targeted at each transmission or equivalently fixed error probability x.

Suppose that each user is equipped with a buffer with E elements where E packets can be stored. Moreover, suppose that the application use one packet per time instant. The queue is filled in with probability $(1 - x)$, i.e the probability that the user can decode the information or equivalently that the user is the "fulfilled" set. Suppose also that at time instant 0 the queue is completely filled up. We want to find the probability that the buffer is empty before a certain time instant θ. This gives rise to a Birth-Death process [96] shown in figure 3.8. The last state E is an absorbing state. When this state is reached an outage occurs and a resynchronization is necessary. Define T the time necessary to reach the state 'E', and consider that at time instant 0 the system is in the state 0, i.e the buffer is full. It is

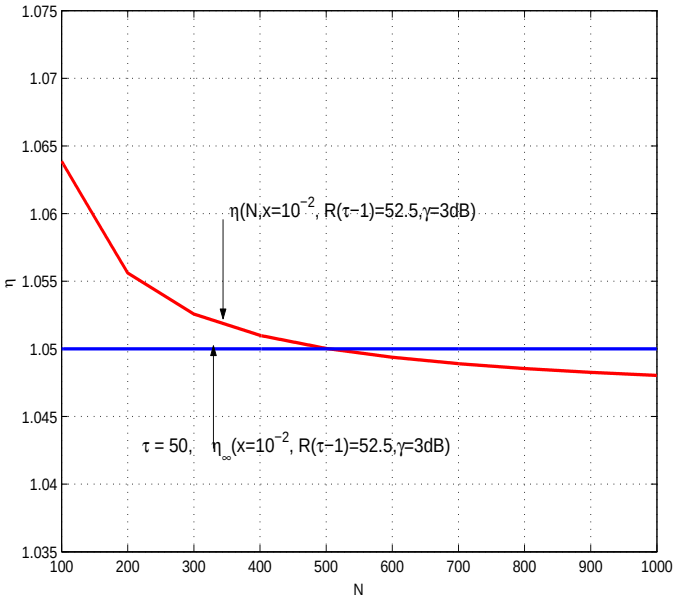


Fig. 3-11. $\eta_\infty(x, R, \Gamma)$ and $\eta(N, x, R, \Gamma)$ vs N for $x = 10^{-2}$, $\Gamma = 3$dB $\tau = 50$ and $R = R(\bar{\tau} - 1)$.

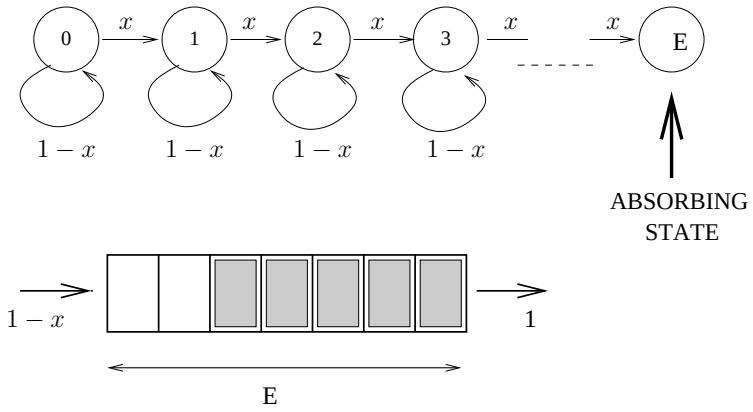


Fig. 3-12. Birth Death process.

easy to see that the probability mass function of T is given by

$$\Pr(T=t) = \begin{cases} 0 & \text{if } t < E \\ \left(\begin{array}{c} t-1 \\ E-1 \end{array} \right) x^E (1-x)^{t-E} & \text{else} \end{cases} \tag{3-29}$$

and finally the cumulative density function $\Pr(T < \theta)$ is

$$\Pr(T \leq \theta) = \begin{cases} 0 & \text{if } \theta < E \\ \sum_{t=E}^{\theta} \left(\begin{array}{c} t-1 \\ E-1 \end{array} \right) x^E (1-x)^{t-E} & \text{else} \end{cases} \tag{3-30}$$

Moreover $\Pr(T \leq \theta) = 1$ for $E = 0$ and $\forall\ \theta > 0$.

Define $S_{k,p}$ a binomial random variable with parameter p, i.e $S_{k,p} \sim \text{Bin}(k,p)$. It is straightforward to see that (3-30) can be written as

$$\Pr(T \leq \theta) = \begin{cases} 0 & \text{if } \theta < E \\ x \sum_{t=E}^{\theta} \Pr(S_{t-1,x} = E-1) & \text{else} \end{cases} \tag{3-31}$$

Suppose to fix the time threshold θ and define the outage probability as the probability of an empty buffer before time instant θ, i.e $\Pr(T \leq \theta)$. It is possible to find the the minimum buffer size, as a function of x, such that the outage probability is less that a certain threshold value p_0, $\Pr(T \leq \theta) \leq p_0$. The optimal buffer size is then found as

$$E_{\text{opt}} = \inf_{E} \left\{ E :\ x \sum_{t=E}^{\theta} \Pr(S_{t-1,x} = E-1)\, 1\{E \leq \theta\} \leq p_0 \right\}$$

Figure 3-13 shows an example of buffer size requirement as a function of the constraint time θ when $p_0 = 10^{-8}$. The time here is measured in packets. It is interesting to notice that the buffer size increases slowly with the constraint θ if the value of x is sufficiently small. Note that when $\theta = E$ than $\Pr(T \leq \theta) = \Pr(T = \theta) = x$. If $x < p_0$ than the optimal buffer size is $E_{\text{opt}} = theta + 1$.

3.9 CONCLUSIONS

The results of this analysis are mixed. On one hand, since reliable packet transmission in delay-limited wireless communications is currently obtained by using HARQ protocols, one would like to keep the same protocol for reliable multicast. On the other hand, it is clear from this analysis and from [42], that the scalability of such protocol in a multicast

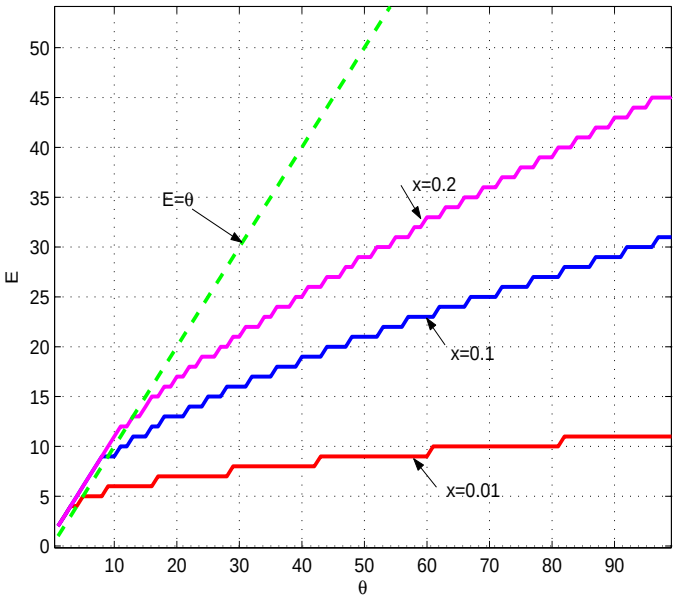


Fig. 3-13. Buffer requirement versus θ for $p_0 = 10^{-8}$ parametrized in x.

environment is questionable. If the underlying application can afford a non-vanishing probability of error (expressed by the fraction x of unfulfilled users), then IR and SR schemes are fully scalable and guidelines for the choice of optimal parameters are given. However, the conclusion is that FEC coding targeted to achieve error probability x, without any explicit ACK/NACK feedback channel, is to be preferred since it achieves the same performance when the number of users gets large, with less complexity.

If full reliable transmission is required, the HARQ scheme is *practically* scalable for typical number of users per cell in a cellular environment, if one accepts a certain non-vanishing gap from ergodic capacity.

These conclusions might be radically different in a rate-distortion setting, where the same source can be multicasted to several users at different distortions.

CHAPTER 4

Lossy Broadcasting Common Information: Optimization of Some Transmission Strategies

4.1 INTRODUCTION

This chapter is focused on the analysis and the optimization of some strategies for the transmission of an analog source over the Gaussian multicast channel. In this case, bit-error probability at the output of the channel decoder is no longer a good measure of performance. On the contrary, the end-to-end distortion is more representative of the quality of transmission.

If we restrict to the case of band-limited Gaussian sources to be transmitted on an additive band-limited Gaussian channel, it is well known that when the source and the channel are particularly matched to each other (the channel bandwidth W_c equals the source bandwidth W_s and when a Gaussian source has to be sent over a Gaussian channel), the uncoded transmission achieves optimal performances. In [45] it is shown that, allowing for a single letter mapping, sufficient and necessary condition for the optimality of uncoded transmission can be found. The main result is a criterion to check whether the single letter code performs optimally for a given source-channel pair. It is also shown that when a single Gaussian source is sent to two different users through Gaussian channels, uncoded transmission gives

a distortion that is Pareto optimal[1] and that lies strictly outside the distortion region for the separation based approach when superposition coding is assumed. Superposition coding consists on embedding high-rate information on low-rate information [46, 36]. However, it is easy to find some practical and not negligible examples when Gastpar's conditions do not hold, for example when the source bandwidth W_s is different from channel bandwidth W_c. As an example, compare the bandwidth of an analog TV signal with the bandwidth of the FM TV signal, clearly $W_c > W_s$ (where we have called W_c the band of the channel and W_s that of the source). When $W_c > W_s$ coding of the source becomes necessary to exploit the additional channel bandwidth.

One of the advantages of using analog schemes is that they achieve gradual changes in the reconstruction quality when changing the SNR while digital schemes show the "threshold effect" described in section 1.3.1.

Shannon's separation theorem states that separating the coding into two steps, source coding and channel coding is optimal. This does not take into account delay and complexity issues and in general it does not hold for multiuser communication or non-ergodic scenario [98]. In [48, 49] and reference therein, the authors have shown that Joint Source-Channel Codes (JSCC) outperform codes designed based on the separation theorem, for fixed complexity and delay, and they are more robust to change in channel noise.

In this chapter we consider the simplest possible scenario of this kind, which is, nevertheless, not yet fully solved. We consider a Gaussian i.i.d. source with bandwidth W_s that has to be transmitted over a band-limited channel with bandwidth W_c under the end-to-end quadratic distortion criterion. As motivated before, we assume spectral efficiency $\eta \stackrel{\Delta}{=} W_s/W_c > 1$.

Again, the BF-AWGN channel is considered, for which the channel gain is random but constant over the duration of a codeword. The coding block length is assumed large enough such that any rate below the instantaneous channel capacity for the given fading realization can be decoded with negligible probability of error, while any rate above the instantaneous channel capacity yields probability of error close to 1. The BF-AWGN channel is a useful mathematical abstraction that models very slowly-varying fading channels, as for example, stationary terminals such as TV receivers, or the path loss determined by the distance to the base station in a mobile cellular communications. In these cases, the fading changes much more slowly than the coding delay and the channel behaves non-ergodically (see [80] for a thorough discussion). The BF-AWGN channel, under the assumption, made here,

[1] Optimality criterion for optimization problem with multi-criteria objectives. A state A is said to be Pareto optimal if there is no other state B dominating the state A w.r.t. the state of objectives functions. A state A dominates the state B if A is better than B in at least one objective function and not worse w.r.t all other objective functions.

that the transmitter is not informed of the channel fading (but it knows its statistics), may also model a Gaussian broadcast/multicast channel [46] with $K \to \infty$ users, such that the empirical distribution of the users SNRs converges almost everywhere to the fading cumulative density function. For this scenario, we optimize and compare three strategies: time-sharing-based transmission, superposition coding and a Hybrid Digital/Analog scheme (*HDA*). We optimize these schemes by minimizing the average end-to-end distortion for given transmit power Γ and fading statistics (assuming a *continuous* pdf $f_A(z)$ and cdf $F_A(z)$).

A transmission strategy widely used in the broadcast setting consists of the so-called "progressive transmission". The source is splitted into L parallel streams mapped onto channel codewords with different coding rate and possibly different energy per channel symbol. This achieves unequal error protection for each level of information. The codewords are sent through the channel by a time-sharing strategy. In [99] the optimization analysis is carried out for Binary Symmetric Channel (BSC) and Binary Erasure Channel (BEC). Using this principle, [100] characterizes an achievable average distortion region for the broadcasting of a common source. The splitting of the source is represented by an ideal successive refinement source encoder that provides independent levels of information each of one conveying the same amount of information bits per source symbols.

A *broadcast approach* to the BF-AWGN channel was proposed and analyzed in [101] (and references therein) in order to maximize the average transmission rate. It consists again of splitting the information message into $L \to \infty$ parallel streams and mapping each stream onto a layer of a superposition coding scheme. Each layer is modulated with a power level $\gamma(a)$, and optimized under the overall power constraint $\mathbb{E}[\gamma(A)] \leq \Gamma$ such that the average successfully received rate is maximized. Following the approach of [101] yields unmanageable expressions due to the fact that the distortion is a non-linear function of the rate and the elegant solution of [101] based on Euler integral does not apply.

The last strategy that we analyze is a Hybrid Digital-Analog (HDA) JSCC. These hybrid systems have been proposed in [53, 54, 55, 56]. These schemes couple the graceful degradation in reconstruction quality with changes in SNR offered by the analog part with the error correcting capability of the digital part.

Shamai et al., in [53], show that *systematic joint source channel coding* (a type of bandwidth splitting HDA) is optimal for a wide class of source and channels. They analyze the capacity of the channel consisting of the digital channel in parallel with the analog channel. They call systematic those source/channel codes which transmit the raw uncoded source in addition to the encoded version (see figure 4.1). When a Gaussian source and channel are considered and when $W_c > W_s$, they show that the conditions for optimality of systematic coding techniques are respected. In [55] the authors show that when Gaussian mixture source are

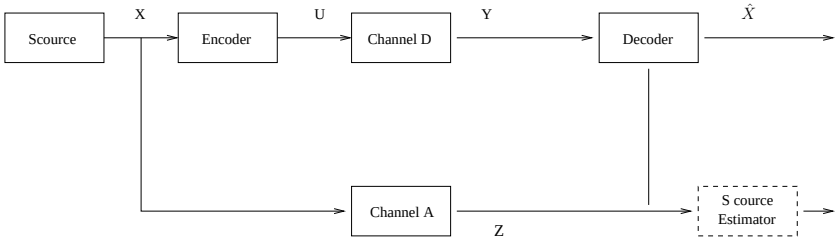


Fig. 4-1. Systematic Source-Channel coding.

transmitted over a Gaussian broadcast channel, the HDA scheme is asymptotically optimal.

Other types of HDA systems has been studied by Mittal et al. in [56], where the authors analyze in terms of average distortion, several coding system based on dimension (bandwidth) splitting and/or power splitting of the source. Also in this case Gaussian source and AWGN channel is considered. They provide examples of "nearly" robust systems[2]. They show that the HDA scheme proposed outperforms in terms of distortion region a time-sharing system and the purely digital system and in certain cases they outperform the systematic code introduced above.

In the following we compute analytically the average distortion of the systems in the ideal case when the source code achieves rate-distortion and the channel code achieve the capacity-cost functions[3]. The system is optimized for block fading AWGN channel. We give an algorithm that can be generalized to take into account more practical scenario where the source and the channel codes are not ideal. Moreover we compare the progressive and superposition scheme with an distortion-based optimized version of the nearly robust HDA scheme proposed in [56].

4.1.1 Summary of the Contribution

- Definition of the optimization problem based on the minimization of the average distortion for progressive, superposition and HDA based scheme, for BF-AWGN channel.

[2] "Nearly robust" means that the system asymptotically operates at the rate-distortion limit for a particular SNR value

[3] This notion is meaningless since a single code cannot *achieve capacity*. However, what we mean here is that $\mathcal{C}$ is a member of a sequence of codes that work arbitrarily close to the capacity limit for increasing block length.

- Algorithms for superposition and progressive approach that give the optimal transmission power and/or coding rate.

4.1.2 Organization of the work

The chapter is organized as follows: In section 4.2 the general definition of the optimization problem is given and the computation of the optimal power allocation and coding rate policy that achieves minimum average distortion is carried out for time-sharing based scheme. Section 4.3 finds the optimal power allocation policy in the case of the superposition approach. In section 4.3.1 the optimization of the HDA scheme is carried out. Section 4.5 gives the results and compare the different strategies in terms of average distortion versus average and instantaneous SNR when Rayleigh fading is considered.

4.2 PROGRESSIVE-BASED TRANSMISSION STRATEGY

We consider a discretized system with L layers, where the number of source code layers coincides with the number of channel codewords. Each level has a source coding rate r_s bits/source sample and it is mapped onto a codeword belonging to a channel code $\mathcal{C}_i'$, modulated at different power levels. In general $\mathcal{C}_i'$ is identified by the rate SNR-threshold pair (r_i, τ_i) such that for SNR larger than τ_i the code yields *acceptable* performance (roughly speaking, low-enough bit-error rate).

The successively refinability property of the source allows to achieve the distortion-rate function at each level $D_\ell = 2^{2r_s\ell}$ and $D_0 = 1$, where ℓ is the number of layer successfully decoded .

Codeword i has coding rate $r_i = \frac{k}{n_i}$, where k is the number of bits output by the i-th layer of the multiresolution source coder and n_i is the blocklength. Figure 4-2 illustrates this scheme. The spectral efficiency is given by

$$\eta = \frac{1}{r_s \sum_{i=1}^{L} \frac{1}{r_i}} \quad \Rightarrow \quad \sum_{i=1}^{L} \frac{1}{r_i} = \frac{1}{\eta r_s}$$

Define γ_ℓ as the energy per channel symbol used to transmit the $\ell-$th codeword. The total power Γ can be computed as

$$\Gamma = \sum_{\ell=1}^{L} \frac{n_\ell}{\sum_{i=1}^{L} n_i} \gamma_\ell = \eta r_s \sum_{\ell=1}^{L} \frac{\gamma_\ell}{r_\ell}$$

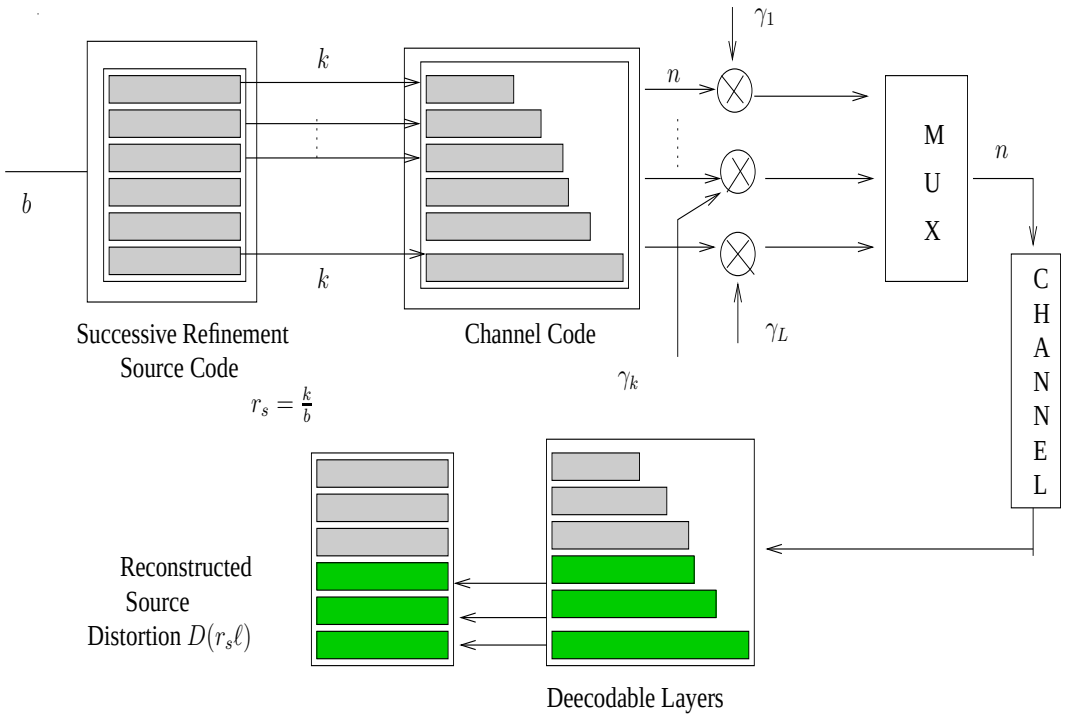


Fig. 4-2. Progressive transmission scheme.

Define a set of fading thresholds $0 < a_1 < \cdots < a_L$ (where $a_{L+1} = \infty$) such that layers up to ℓ can be decoded if $A \in [a_\ell, a_{\ell+1}]$. Assuming ideal Gaussian channel codes, the ℓ-th code rate is given by $r_\ell = \log_2(1 + a_\ell \gamma_\ell)$. Call now $z_\ell = \frac{1}{r_\ell}$ and $y_\ell = \frac{\gamma_\ell}{r_\ell}$. It follows that

$$a_\ell = \frac{\left(2^{1/z_\ell} - 1\right) z_\ell}{y_\ell} \tag{4-1}$$

We wish to minimize the resulting average distortion $D_{av}(r_s, \mathbf{z}, \mathbf{y})$ with respect to $\mathbf{z}$ and $\mathbf{y}$ subject to the constraints

$$\sum_{i=1}^{L} z_\ell = \frac{1}{\eta r_s}; \quad \sum_{i=1}^{L} y_\ell = \frac{\Gamma}{\eta r_s} \tag{4-2}$$

where $D_{av}(r_s, \mathbf{z}, \mathbf{y})$ has the following expression

$$D_{\text{av}}(r_s, \mathbf{z}, \mathbf{y}) = F_A(a_1) + \sum_{\ell=1}^{L} D_\ell \left(F_A(a_{\ell+1}) - F_A(a_\ell)\right) \tag{4-3}$$

with a_ℓ defined in (4-1).

The associated Lagrangian functional Φ is given by

$$\Phi = D_{av}(r_s, \mathbf{z}, \mathbf{y}) + \lambda \sum_{i=1}^{L} z_i + \rho \sum_{i=1}^{L} y_i$$

For the Kuhn Tucker's conditions it follows that the partial derivative with respect to z_ℓ and y_ℓ has to be greater or equal to 0,

$$\frac{\partial \Phi}{\partial z_\ell} = \Delta D_\ell f_A(a_\ell) \left(\frac{\left(2^{1/z_\ell} - 1\right) z_\ell - 2^{1/z_\ell} \ln 2}{z_\ell y_\ell} \right) + \lambda \geq 0$$

$$\frac{\partial \Phi}{\partial y_\ell} = \Delta D_\ell f_A(a_\ell) \left(-\frac{z_\ell \left(2^{1/z_\ell} - 1\right)}{y_\ell^2} \right) + \rho \geq 0 \tag{4-4}$$

(4-5)

y_ℓ can be found as a function of ρ, λ and z_ℓ

$$y_\ell = \frac{1}{\mu} \, \frac{\left(2^{1/z_\ell} - 1\right) z_\ell^2}{2^{1/z_\ell} \ln 2 - \left(2^{1/z_\ell} - 1\right) z_\ell} \tag{4-6}$$

where μ is defined as $\mu \triangleq \frac{\rho}{\lambda}$. z_ℓ is then obtained as

$$-\Delta D_\ell g(z_\ell, \mu) + \lambda = 0 \tag{4-7}$$

where $g(z_\ell, \mu)$ is defined as

$$g(z_\ell, \mu) \stackrel{\Delta}{=} f_A\left(\mu \frac{2^{1/z_\ell}\ln 2 - \left(2^{1/z_\ell}-1\right) z_\ell}{z_\ell}\right) \cdot \left(\mu \frac{\left(2^{1/z_\ell}\ln 2 - \left(2^{1/z_\ell}-1\right) z_\ell\right)^2}{z_\ell^3\left(2^{1/z_\ell}-1\right)}\right) \quad (4\text{-}8)$$

4.3 SUPERPOSITION-BASED TRANSMISSION STRATEGIES

Consider now a superposition-based approach where each level is mapped onto an independently selected codeword of "a basic channel code" $\mathcal{C}'$ modulated at different power levels. Each layer has source coding rate r_s bits/source sample and channel coding rate r_c bit/channel uses, so that $\eta = r_c/r_s$. Figure 4-3 shows the block diagram of the superposition scheme. The mother code $\mathcal{C}'$ is identified by the rate SNR-threshold pair (r_c, τ).

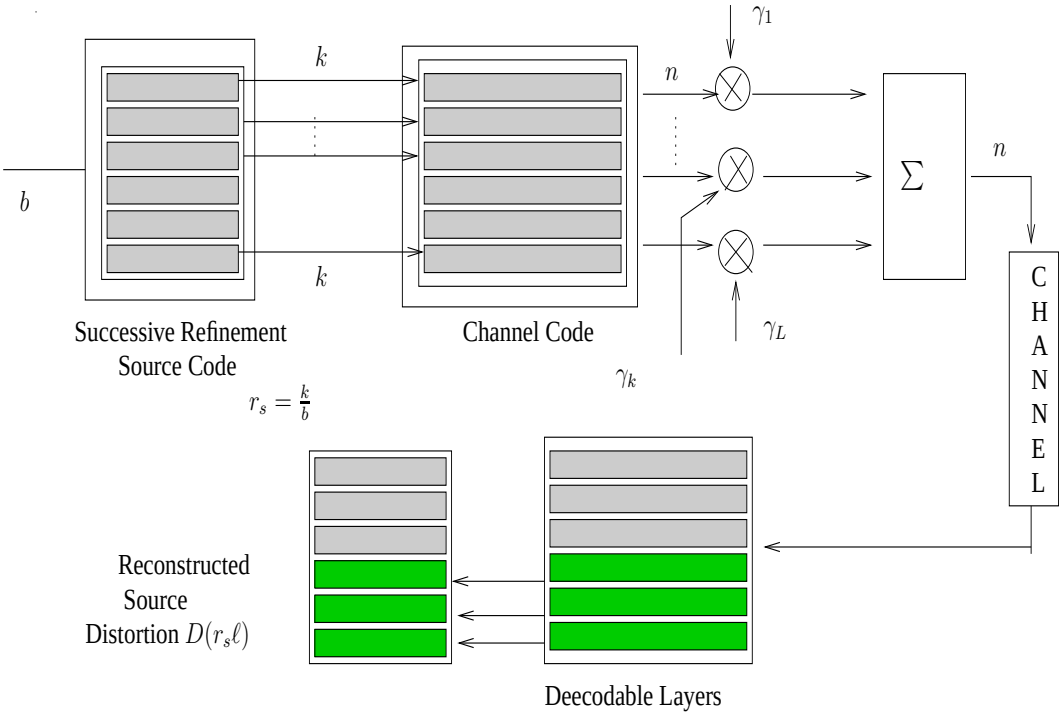


Fig. 4-3. Superposition scheme.

The transmitted superposition codeword is given by $\mathbf{x} = \sum_{\ell=1}^{L} \sqrt{\gamma_\ell}\mathbf{c}'_\ell$ where γ_ℓ and $\mathbf{c}'_\ell$ are the power level and the codeword of $\mathcal{C}'$ associated to level ℓ, respectively.

Define again the set of fading thresholds $0 < a_1 < \cdots < a_L$ (where $a_{L+1} = \infty$) such that layers up to ℓ can be decoded if $A \in [a_\ell, a_{\ell+1}]$. The resulting average distortion is given by

$$D_{\text{av}}(r_s, \boldsymbol{\gamma}) = F_A(a_1) + \sum_{\ell=1}^{L} D_\ell \left(F_A(a_{\ell+1}) - F_A(a_\ell)\right) \tag{4-9}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$. We want now to solve the following minimization problem

$$\min_{r_s} \ \min_{\boldsymbol{\gamma}} \ D_{\text{av}}(r_s, \boldsymbol{\gamma}) \quad \text{s.t} \quad \sum_{i=1}^{L} \gamma_i = \Gamma \tag{4-10}$$

The condition for successive decodability of the superposition code up to layer ℓ is given by

$$\frac{a_\ell \gamma_\ell}{1 + a_\ell \sum_{j=\ell+1}^{L} \gamma_j} \geq \tau \tag{4-11}$$

The levels a_ℓ are uniquely defined by the power levels γ_ℓ by imposing the constraint (4-11) with equality,

$$a_\ell = \frac{\tau}{\gamma_\ell - \tau \sum_{j=\ell+1}^{L} \gamma_j} \tag{4-12}$$

Conversely, the γ_ℓ's can be expressed in terms of the a_ℓ's by solving the triangular linear system $a_\ell \gamma_\ell - \tau a_\ell \sum_{j=\ell+1}^{L} \gamma_j = \tau$ for all $\ell = 1, \ldots, L$ which yields

$$\gamma_\ell = \tau x_\ell + \tau^2 x_{\ell+1} + \sum_{j=\ell+2}^{L} \tau^2 x_j \left(1 + \tau\right)^{j-\ell-1} \tag{4-13}$$

where $x_\ell \stackrel{\Delta}{=} \frac{1}{a_\ell}$. We wish to minimize the average distortion (4-9) with respect to $\{\gamma_1, \ldots, \gamma_L\}$ subject to the constraint $\sum_\ell \gamma_\ell = \Gamma$.

The associated Lagrangian functional is

$$\Phi = D_{\text{av}}(r_s, \gamma_1, \ldots, \gamma_L) + \lambda \sum_{\ell=1}^{L} \gamma_\ell \tag{4-14}$$

The ℓ-th partial derivative is given by

$$\frac{\partial \Phi}{\partial \gamma_\ell} = \sum_{j=1}^{\ell-1} \Delta D_j \frac{1}{x_j^2} f_A\left(\frac{1}{x_j}\right) - \Delta D_\ell \frac{1}{\tau} \frac{1}{x_\ell^2} f_A\left(\frac{1}{x_\ell}\right) + \lambda \tag{4-15}$$

where $\Delta D_\ell = D_{\ell-1} - D_\ell$. From the Kuhn-Tucker conditions, we look for the values of x_ℓ such that the derivative is non-negative.

With the substitution $w_\ell = \frac{1}{x_\ell^2} f_A\left(\frac{1}{x_\ell}\right)$, the system given by $\frac{\partial \Phi}{\partial \gamma_\ell} = 0$ is linear and lower triangular, and the solution is given by

$$w_\ell = \frac{\lambda \tau \left(1+\tau\right)^{\ell-1}}{D_{\ell-1} - D_\ell} = \lambda \mathcal{G}_\ell \tag{4-16}$$

where we have defined $\mathcal{G}_\ell \stackrel{\Delta}{=} \frac{\tau(1+\tau)^{\ell-1}}{D_{\ell-1}-D_\ell}$.

Finally, the derivative is greater or equal to zero if

$$-\frac{\frac{1}{x_\ell^2} f_A\left(\frac{1}{x_\ell}\right)}{\mathcal{G}_\ell} + \lambda \geq 0 \tag{4-17}$$

4.3.1 Hybrid Analog/Digital Scheme

The scheme of the encoder is shown in figure 4-4 while the decoder is shown in figure 4-5. The source bandwidth is splitted so that $W_{s,A}$ dimensions are sent through an analog (uncoded) branch while $W_{s,D} = W_s - W_{s,A}$ is sent through the digital encoder, called tandem encoder. The two outputs are modulated by power levels γ_A and γ_1 respectively, superimposed and sent to the channel. γ_A, γ_1 are such that $\gamma_A + \gamma_1 = \Gamma$. The total transmitted signal is given by

$$\mathbf{y} = \sqrt{\gamma_1}\mathbf{s}_1 + \sqrt{\gamma_A}\mathbf{s}_A + \boldsymbol{\nu}$$

where $\mathbf{s} = [\mathbf{s}_1, \mathbf{s}_A]$ is the total analog source with bandwidth W_s, $\boldsymbol{\nu}$ is the complex circularly symmetric AWGN with components $\boldsymbol{\nu}_i \sim \mathcal{N}(0,1)$. At the received side the digital decoder decodes the information by considering the analog signal as noise. The estimated information is re-encoded and removed from the received signal. This constitutes the signal from which the analog decoder estimates the analog information.

$$\mathbf{y}_A = \sqrt{\gamma_A}\mathbf{s}_A + \sqrt{\gamma_1}(\mathbf{s}_1 - \hat{\mathbf{s}}_1) + \boldsymbol{\nu} \tag{4-18}$$

When the digital decoder can not decode the information with vanishing error probability, than the analog decoder will estimate $\mathbf{s}_A$ by considering the second term in 4-18 as additive noise.

In [56] the system is designed such that to satisfy the "nearly robust" constraint, i.e to give asymptotically optimal performances for a particular value of $\mathrm{SNR} = \mathrm{SNR}^\star$. This yields a particular power splitting between the digital and the analog part. In this work, instead of constraining the system to be nearly robust, we find the power allocation policy that minimize the average distortion subject to a total power constraint. The "analog code" is

a linear coder/decoder with coefficients that minimize the mean squared error. In [56] the authors consider a matched tandem encoder as digital system. The encoder is said to be matched if the channel input is a scaled version of the first n components of the quantizer output, where n is the blocklength. When the SNR is such that the tandem decoder can not decode the information (the SNR is lower than the threshold of the code), then the decoder becomes a linear decoder that estimates the first n symbols.

Here we limit the analysis and the results for the non matched case. However note that the construction of such codes is not straightforward and that they lead to small improvement in the performances only for SNR $<$ SNR*, i.e very small values of fading coefficients.

This scheme has only one digital layer so we will define only one fading threshold a_1 s.t. $0 < a_1 < +\infty$. The threshold a_1 is such that if the actual fading value is above that threshold, then the digital decoder can recover the information with vanishing error probability, while on the contrary the digital decoder acts as noise for the analog decoder. The channel code is defined by the pair (r_c, τ), where in the ideal case $\tau = 2^{r_c} - 1 = 2^{(\eta-1)r_s} - 1$ and the condition for successful decodability is given by

$$\frac{a_1\gamma_1}{1 + a_1\gamma_A} \geq \tau = \text{SNR}^* \tag{4-19}$$

where the analog layer is treated as additional noise by the digital decoder. By imposing the equality in (4-19) and substituting the total power constraint, we obtain $\gamma_A = \frac{\Gamma}{(1+\tau)} - \frac{\tau}{a_1(1+\tau)}$. Note that for a given fading value a, if $a < a_1$ the digital signal cannot be decoded correctly, i.e the output of the channel encoder acts as noise for the linear estimator of the analog layer. The distortion due to the analog layer is given by

$$\begin{aligned} D_A(r_s, a_1) =& \frac{1}{1 + a\gamma_A} \text{ if } a \geq a_1 \\ =& 1 - \frac{a\gamma_A}{1 + a\Gamma} \text{ else} \end{aligned} \tag{4-20}$$

The average distortion can be written as

$$\begin{aligned} D_{ave}(r_s, a_1) =& \frac{\eta - 1}{\eta} \left(F_A(a_1) + D_1(1 - F_A(a_1))\right) + \frac{1}{\eta} \\ & \left[F_A(a_1) - \int_0^{a_1} \frac{a\gamma_A f_A(a)da}{1 + a\Gamma} + \int_{a_1}^{\infty} \frac{f_A(a)da}{1 + a\gamma_A}\right] \end{aligned} \tag{4-21}$$

The result of the unconstrained minimization of (4-21) is obtained by finding a_1 solution of

the following equation

$$f_A(a_1) - D_1 f_A(a_1)\frac{\eta-1}{\eta} - \frac{\gamma_A a_1 f_A(a_1)}{\eta(1+a_1\Gamma)} - \frac{f_A(a_1)}{\eta(1+a_1\gamma_A)} - \frac{F_A(a_1)\tau}{\eta(1+\tau)a_1^2\Gamma} + \\ + \frac{\tau}{\eta(1+\tau)a_1^2}\left[\int_0^{a_1}\frac{f_A(a)da}{1+a\Gamma} - \int_{a_1}^{\infty}\frac{f_A(a)da}{(1+a\gamma_A)^2}\right] = 0 \qquad (4\text{-}22)$$

4.4 SHANNON'S SEPARATION THEOREM

For comparison, a system based on the separation theorem that transmits a single layer is considered. This can be regarded as the baseline system representative of today's technology, such as terrestrial and satellite DTV, or DAB. We consider the minimization of average distortion with respect to the source coding rate r_s. The average distortion of the one-layer digital system is given by

$$D_{\text{sep}} = F_A\left(\frac{2^{\eta r_s}-1}{\Gamma}\right) + 2^{-2r_s}\left[1 - F_A\left(\frac{2^{\eta r_s}-1}{\Gamma}\right)\right] \qquad (4\text{-}23)$$

In the general case, the optimal value $r_s^\star$ is given by the solution of

$$f_A\left(\frac{x^\eta-1}{\Gamma}\right)\frac{x^\eta\eta}{\Gamma}\log 2 - 2\log(2)\frac{1}{x^2}\left(1 - F_A\left(\frac{x^\eta-1}{\Gamma}\right)\right) - \frac{1}{x^2}f_A\left(\frac{x^\eta-1}{\Gamma}\right)\frac{x^\eta\eta}{\Gamma}\log 2 = 0$$

with $x = 2^{r_s^\star}$.

4.5 ON ACHIEVABLE RSNR: RAYLEIGH FADING

In this section we show the results of the optimization problems defined above. We consider Rayleigh fading so that the pdf of the fading power gain in given by $f_A(a) = e^{-a}$. The spectral efficiency η is fixed to 3 complex source symbol per channel use. Note that in all the above systems r_s is left as a design parameter and numerical optimization w.r.t r_s is further carried out. For the superposition strategy by letting L arbitrarily large with r_s arbitrarily small our numerical computable solution will approach arbitrarily closely the optimal solution of [101] when the average distortion is minimized instead of maximizing the average rate. For the progressive transmission approach, the optimal performance is, as well, obtained for $r_s \to 0$. For the HDA scheme however the optimal r_s is a fixed non vanishing value $r_s^\star$.

Figure 4-6 shows a graphical representation of (4-17) for the superposition scheme, for Rayleigh fading. The set of solutions x_ℓ is to be found in the region defined as 'valid

solutions' in the figure 4-6, i.e the region where the property $x_1 \geq x_2 \ldots \geq x_L$ holds. We would like to stress the fact that $\lambda\mathcal{G}_\ell$ is an increasing sequence with respect to ℓ, for fixed λ. Hence, the number of levels that yield optimal performance, is given by $L = \sup\{\ell : w_\ell \leq \lambda\mathcal{G}_\ell\}$, for the choice of λ that satisfy the power constraint Γ.

For the progressive transmission approach $g(z_\ell, \mu)$ in (4-8) has the same behavior as (4-17). Moreover, analogous analysis on set of solutions and the optimal number of levels, fixed μ, holds. Figure 4-7 shows the results in terms of RSNR defined as $10\log_{10}\frac{1}{D}$ where D is the average distortion vs the average channel signal to noise ratio Γ. In the figure the average performances of superposition, progressive and the HDA ('HDA_D' in the figure) scheme are compared. Also plotted are the scheme based on the separation theorem and the nearly robust HDA ('HDA_{NR}' in the figure). For the superposition and progressive schemes, when vanishing r_s is optimal we plot the average performances for a r_s "small enough", $r_s = 1/20$. For practically small rate the gap from the optimal performances becomes negligible.

Reducing r_s increases the optimal number layer and since a small enough value of r_s allows to achieve optimal performance, practically there is no need to make too many layers. For the separated approach and for the HDA the value of r_s is fixed to $r_s^\star$. Finally, notice that the HDA_D outperforms all the other schemes and that there is practically no difference between the separation theorem based approach and the progressive transmission in terms of average distortion.

Figure 4-8 shows the performances in terms of RSNR vs instantaneous SNR, for average SNR $\Gamma = 20$dB. For comparison the performance of the nearly robust non matched HDA scheme is plotted. The theoretical limit, in terms of distortion, is given by Shannon as $D_{Sh} = \frac{1}{(1+a\Gamma)^{2/\eta}}$. As announced before, the enhancement of the performances due to the matched encoder is small and concentrated in a range SNR $<$ SNR* because in that range the tandem decoder becomes a linear decoder that estimates the symbols. The HDA_D schemes and the superposition scheme gives smooth performances for a wide range of SNR providing more graceful degradation under mismatched channel condition compared to the separation theorem based scheme and the progressive approach . Finally most of the gain in RSNR of the HDA_D is due to the presence of the linear encoder.
Note that even if the separated approach yields the same performance as the progressive scheme in terms of average distortion vs average SNR, the situation is different when the average distortion is plotted as a function of the instantaneous SNR. The progressive based scheme shows advantages in terms of graceful degradation of performances over a wide range of SNRs and therefore it is more suited to multiuser applications.

Further computation with different fading statistic have shown that, when considering a uniform distribution of users over a circular cell and attenuation due to path loss, the super-

position scheme yields closed form solution for the optimal power allocation policy.

4.6 CONCLUSIONS

In this chapter a general optimization method for three transmission schemes for compound channel is given: the first scheme is based on ideal multilevel quantizer and time-sharing transmission scheme, the second couples the multilevel quantizer with a superposition scheme. Finally, the third is an hybrid digital-analog scheme. These schemes are optimized in order to minimize the average overall distortion under total transmitted power constraint and spectral efficiency. The algorithms give the optimal power/rate allocation policy that minimize the average distortion, as well as the optimal number of layers. The algorithms are derived for ideal source/channel code behavior but can be generalized to take into account more practical setting where the source and the channel codes are not ideal. Under the assumption, considered here, that the transmitter is not informed about the instantaneous fading condition, the compound channel can model a broadcast (multicast) scenario where each users has a particular fading coefficient. The algorithms found in this chapter can be used for the design of good joint source channel codes that approach theoretical limits in a multicast setting.

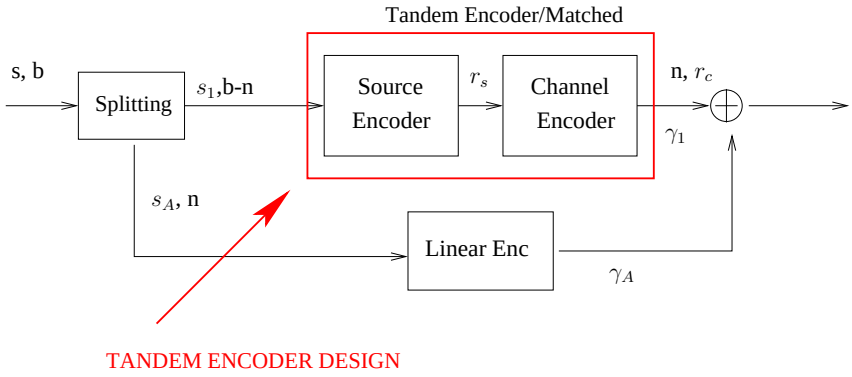


Fig. 4-4. Hybrid digital-analog scheme.

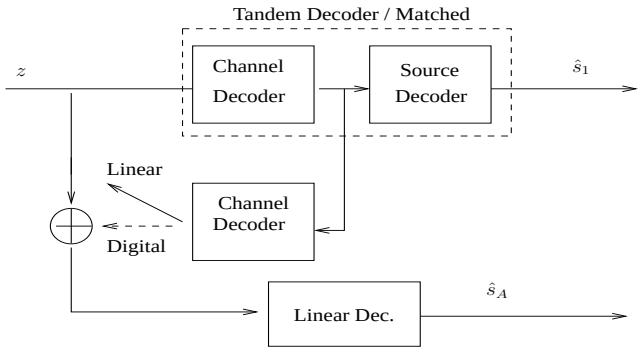


Fig. 4-5. Hybrid digital-analog scheme.

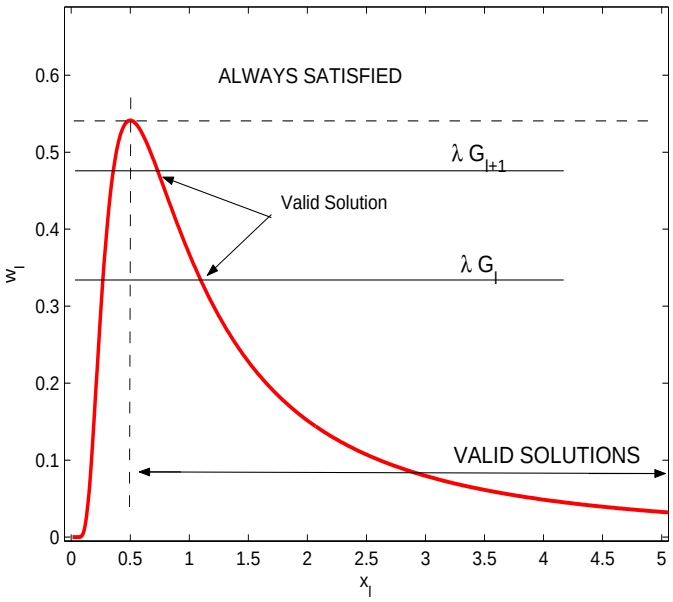


Fig. 4-6. Function to find the minima for superposition scheme.

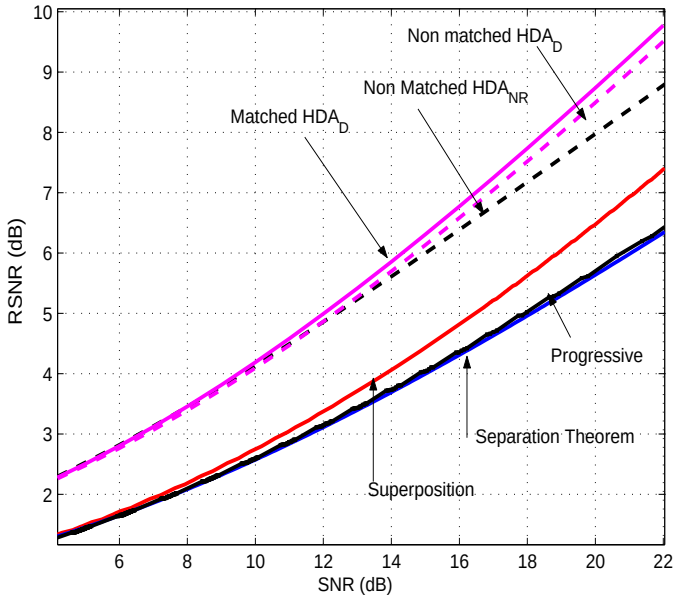


Fig. 4-7. RSNR vs channel average SNR (Γ) for the superposition, progressive and HDA approach. Also the scheme based on the separation theorem and the optimized nearly robust HDA is plotted for comparison.

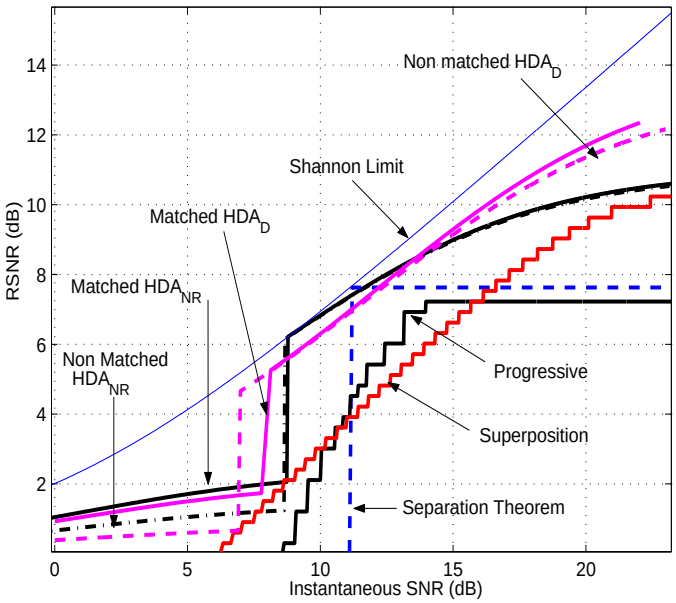


Fig. 4-8. RSNR vs channel instantaneous SNR for $\Gamma = 20$ dB for the superposition, progressive and HDA approach. Also the scheme based on the separation theorem and the optimized nearly robust (matched and unmatched) HDA is plotted for comparison.

CHAPTER 5

Practical Code Constructions

5.1 INTRODUCTION

Chapter 4 shows that HDA schemes outperform fully digital strategies based on time-sharing and superposition. Hence, here, we focus on the practical construction of HDA schemes. Figure 5-1 recall the encoding structure of the HDA scheme. Recall that the HDA is based on the splitting of the source bandwidth such that one part is uncoded and the second part is encoded by a tandem encoder. The two signals are modulated by different power levels, superimposed and passed through the channel. A tandem encoder is a general term to indicate both the source and the channel encoder. The optimization of the HDA scheme in chapter 4 yields the value of a threshold SNR^* for which the tandem encoder should be designed. Therefore, a key issue is the design of a tandem encoder robust to channel errors, such that it works as close as possible to the theoretical limit.

We consider here two different strategies. The first is based on sophisticated quantizer schemes as vector quantization-based constructions, that output almost non-redundant bits. These can be protected against channel errors by using standard channel codes, like Turbo codes or LDPC. The issue, here, is to design quantizer schemes that are non-catastrophic, in the sense that few bits in error at the output of the channel decoder should not lead to catastrophic effects on the reconstruction quality. Examples of such schemes can be found in [102, 103, 104]. Note that best quantizer schemes known in literature, are found among

the family of Entropy Constrained Trellis Coded Quantizer (ECTCQ) [113]. However, since the indexes at the output of the quantizer have residual redundancy, traditionally they are coupled with a variable rate source code. This makes the scheme very sensible to channel errors, since it is well known that arithmetic encoders have catastrophic inverse.

Here we propose a scheme, based on convolutional codes and unitary transformations, efficiently implemented by FFT/IFFT and interleaving. This scheme offers performance comparable to the best known Trellis Coded Quantizer (TCQ) [59] and very fine granularity of rates. This scheme inherits from convolutional codes the property of being non-catastrophic, thus it is robust to residual errors of the channel decoder. It can be coupled with standard Turbo codes or LDPC. Note also that this quantizer is embedded by construction, i.e it implements a successive refinable source code. Consequently it can work also with time-sharing/superposition transmission strategy.

The second class of tandem encoder that we consider exploits the residual redundancy of the indexes at the output of the quantizer. For this, scalar or vector quantizers whose output is redundant, can be considered. Data compression and channel protection are jointly performed. A key issue is the design and the optimization of such codes so that they are robust to channel errors. The idea of exploiting the redundancy of the source coder output to increase performance is well known. Sayood et al. [105] suggested to use this redundancy for error protection. Hagenauer et al. in [106] proposed to use it to modify the soft information processed by the decoder. Another increasingly popular scheme involves dual-functional channel codes. It was shown in [107] that fixed-to-fixed length data compression of a discrete source using linear codes is strongly related to transmission via linear codes on a discrete additive noise channel where the noise has the same statistic of the source. This analogy can be exploited by using linear error correction channel codes such as LDPC codes [107] or Turbo codes, for data compression. In [108, 65] Garcia-Frias et al. proposed a scheme for data compression for both single memoryless source and correlated sources where the desired compression ratio can be achieved by properly puncturing turbo codes, in particular by puncturing the information bits and the parity bits. A priori probability of the source is used to modify the extrinsic information in the iterative decoding process.

Here we consider a data-compression/ channel protection scheme based on Turbo codes, and we refer to it as Multilevel Turbo COMpression (M-TCOM). This is realized by coupling a scalar/vector quantizer, whose indexes still contain redundancy, with a compression/protection multilevel scheme based on Turbo Codes. For simplicity we consider here Entropy Constrained Scalar Quantizer (ECSQ), but the same scheme can be generalized to work with Entropy Constrained Vector Quantizer (ECVQ) like ECTCQ.

A $Q-$ary to binary mapping transforms the output of the ECSQ into L bit-streams. Each bit level is, then, mapped on different Turbo codes where the systematic bits are punctured,

together with a certain amount of parity bits in order to achieve the desired rate. The codewords are multiplexed and transmitted over the channel. We assume that the decoder is aware of the "conditional statistic" of each bit-plane. By "conditional statistic" we mean the conditional probability of the bits at level ℓ given the previous levels.

Optimization of turbo codes is necessary in terms of polynomial generator of the component convolutional code and puncturing pattern. This can be carried out by using EXIT Chart-based method [27].

The M-TCOM approach, by nature, can be generalized to yield progressive transmission of information. By choosing the Q-ary to binary mapping and the quantizer scheme such that it is embedded, the source can be reconstructed with different levels of distortion. The scheme and the design of the code are extended to the case of practical transmission of images over the wireless link. This is shown to give remarkable results when coupled with a modified Differential Pulse Code Modulation-based (DPCM) quantizer defined by Kim et al [57].

Finally, MTQ concatenated with Turbo codes and M-TCOM scheme are compared. MTQ based scheme yields performance closer to the theoretical limit (Shannon limit). The poorer performance of M-TCOM compared to MTQ are mainly due to two factors. First, in the noiseless case the M-TCOM scheme can achieve only the performance of ECSQ (poor performance in low rate region). Second, the optimization of Turbo codes is not straightforward. An open issue is the analytical optimization of codes belonging to the Irregular Repeat and Accumulate family, through Density Evolution, and the use of ECTCQ, instead of ECSQ. This, potentially, will approach, in a better way, the Shannon's limit.

5.1.1 Summary of the Contribution

- Definition of the Multistage Trellis Quantizer (MTQ) based on unitary transformation and convolutional code.
- Analysis of the behavior of the MTQ in the noiseless and noisy case.
- Construction and analysis of a compression scheme based on Turbo codes.
- Guidelines for optimization of Turbo codes.

5.1.2 Organization of the Work

This chapter is organized as follows Section 5.2 describes the construction of the MTQ. In section 5.2.1 the necessary background is given while section 5.2.2 explains the code

design. Section 5.3 deals with lossy transmission over noisy channel when considering the MTQ scheme.

Section 5.4 introduces the compression scheme based on linear codes and give the results of the optimization of the component codes. An example is given in section 5.5 when this scheme is coupled with more practical quantizer (Differential Pulse Coded Modulation, DPCM) for lossy transmission of image over noisy channel. Finally 5.6 concludes the chapter and discuss some open points and extensions.

5.2 MULTISTAGE TRELLIS QUANTIZER

5.2.1 Background

Consider a source $S \in \mathbb{R}$ with rate-distortion function $R(D)$ with respect to a certain distortion measure $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, inducing the distortion measure on $\mathbb{R}^k \times \mathbb{R}^k$ according to

$$d(\mathbf{s}, \widehat{\mathbf{s}}) = \frac{1}{k} \sum_{i=1}^{k} d(s_i, \widehat{s}_i) \tag{5-1}$$

An L-level successive refinement source code of block length k is defined by the encoding functions $g_\ell : \mathbb{R}^k \to \{1, \ldots, M_\ell\}$ and by the reconstruction functions

$$\phi_\ell : \{1, \ldots, M_1\} \times \cdots \times \{1, \ldots, M_\ell\} \to \mathbb{R}^k$$

The rate L-tuple of the successive refinement code is given by

$$R_\ell = \sum_{j=1}^{\ell} \log_2 M_j : \ell = 1, \ldots, L$$

and the achieved distortion L-tuple is given by

$$\{D_\ell = \mathbb{E}\left[d(\mathbf{s}, \phi_\ell(g_1(\mathbf{s}), \ldots, g_\ell(\mathbf{s})))\right] \ : \ \ell = 1, \ldots, L\}$$

The successive refinement structure of the code manifests itself in the fact that distortion level D_ℓ is obtained by *refining* the coarser description at level $\ell - 1$ by incorporating additional information at rate increment $R_\ell - R_{\ell-1}$ bits/source symbol.

The source S is said *successively refinable* [51, 50] if, for any desired integer L, distortion L-tuple $D_1 < D_2 < \cdots < D_L$, $\epsilon > 0$ and sufficiently large k, there exists an L-level

successive refinement source code of block length k with rate L-tuple $(R_1, \ldots, R_L)$ such that

$$R_\ell \leq R(D_\ell) + \epsilon \quad \forall \ \ell = 1, \ldots, L \tag{5-2}$$

and

$$\mathbb{E}[d(\mathbf{s}, \phi_\ell(g_1(\mathbf{s}), \ldots, g_\ell(\mathbf{s})))] \leq D_\ell + \epsilon \quad \forall \ \ell = 1, \ldots, L \tag{5-3}$$

In other words, the L-tuple of *optimal* rate-distortion pairs $\{(R(D_\ell), D_\ell) : \ell = 1, \ldots, L\}$ is achievable by successive refinement.

In the rest of this work we restrict to the quadratic distortion measure $d(s, \widehat{s}) = |s - \widehat{s}|^2$ and to sources with mean zero and variance 1 (different variances can be handled by normalization).

It is well-known that a Gaussian i.i.d. source $S \sim \mathcal{N}(0, 1)$ is successively refinable [51, 50]. It is also well-known that, in the Gaussian case, optimal successive refinement codes have an additive structure [109], i.e., the ℓ-th level representation vector $\widehat{\mathbf{s}}_\ell$ for the source vector $\mathbf{s}$ is given by

$$\widehat{\mathbf{s}}_\ell = \sum_{j=1}^{\ell} \psi_j(g_j(\mathbf{s})) \tag{5-4}$$

where $\psi_\ell : \{1, \ldots, 2^{k(R_\ell - R_{\ell-1})}\} \to \mathbb{R}^k$ denotes the reconstruction *increment* function at level ℓ.[1]

Now, consider a spherical codebook

$$\mathcal{C} = \left\{ \mathbf{c}_q \in \mathbb{R}^k \ : \ q = 1, \ldots, 2^{kr_s} \right\} \tag{5-5}$$

where r_s is a design parameter. The codewords of $\mathcal{C}$ lie on a k dimensional sphere of squared radius k. Consider $\Delta \in (0, 1]$ and let $Q_\alpha : \mathbb{R}^k \to \{1, \ldots, 2^{kr_s}\}$ denote the minimum Euclidean distance decoder for the scaled code $\alpha\mathcal{C}$, i.e.,

$$Q_\alpha(\mathbf{s}) = \arg \min_q d(\mathbf{s}, \alpha \mathbf{c}_q) \tag{5-6}$$

Lapidoth [75] showed that for $r_s > \frac{1}{2}\log_2 \frac{1}{\Delta}$, $\alpha = \sqrt{1 - \Delta}$, $\epsilon > 0$ and for sufficiently large k there exist spherical codes $\mathcal{C}$ such that

$$\mathbb{E}[d(\mathbf{s}, \alpha \mathbf{c}_{Q_\alpha(\mathbf{s})})] \leq \Delta + \epsilon \tag{5-7}$$

This result holds for any source S, not necessarily Gaussian, i.i.d., or even ergodic, under the condition that $\frac{1}{k}|\mathbf{s}|^2 \to 1$ in probability [75]. In some sense, scaled spherical codes

[1] We define $R_0 = 0$ and $D_0 = 1$.

with minimum distance encoding are *robust* in the sense that they achieve the Gaussian rate distortion bound under very mild conditions on the source. On the other hand, for these codes all sources appear as hard to compress as the Gaussian i.i.d. source.

We shall construct L-levels successive refinement codes from a single spherical code $\mathcal{C}$, denoted as the "basic code". Fig. 5-2 provides a pictorial representation of the geometry of the proposed construction. The encoding function at level ℓ is based on the minimum distance decoder of the basic code. It computes the ℓ-level index

$$g_\ell(\mathbf{s}) = Q_{\alpha\sqrt{\Delta^{\ell-1}}}(\mathbf{s} - \widehat{\mathbf{s}}_{\ell-1}) \tag{5-8}$$

where

$$\widehat{\mathbf{s}}_\ell = \sum_{j=1}^{\ell} \alpha\sqrt{\Delta^{j-1}}\mathbf{c}_{g_j(\mathbf{s})} \tag{5-9}$$

is the representation vector at level ℓ. Such multistage structure can achieve the rate L-tuple $\{R_\ell = \ell r_s : \ell = 1, \ldots, L\}$ with distortion L-tuple $\{D_\ell = \Delta^\ell : \ell = 1, \ldots, L\}$.

Lastras and Berger [52] showed that any well-behaved source can be encoded by successive refinement incurring a bounded rate penalty at each level. In particular, let S be an arbitrary i.i.d. source with mean zero, variance 1, finite differential entropy $h(S)$ and rate-distortion function $R(D)$. The distortion L-tuple $(D_1, \ldots, D_L)$ can be achieved by successive refinement at rates $(R_1, \ldots, R_L)$ such that

$$R_\ell \leq R(D_\ell) + \frac{1}{2}\log_2 \frac{1}{\mathcal{P}_S} \tag{5-10}$$

where $\mathcal{P}_S = \frac{2^{2h(S)}}{2\pi e}$ is the *entropy power* of S, i.e., it is the variance of a Gaussian source with the same differential entropy of S.

The multistage spherical code can achieve $D_\ell = \Delta^\ell$ at rate $R_\ell = \frac{\ell}{2}\log_2\frac{1}{\Delta}$. By using the Shannon lower bound on the rate distortion function [46], we find that the rate penalty is bounded by

$$R_\ell - R(D_\ell) = \frac{\ell}{2}\log_2\frac{1}{\Delta} - R(\Delta^\ell) \leq \frac{\ell}{2}\log_2\frac{1}{\Delta} - \frac{1}{2}\log\frac{\mathcal{P}_S}{\Delta^\ell} = \frac{1}{2}\log_2\frac{1}{\mathcal{P}_S} \tag{5-11}$$

which coincides with the bound in (5-10). In other words, the behavior of the proposed scheme is good in the sense that it meets Lastras and Berger bound for any source for which Lapidoth result [75] holds. In practice, a successive refinement code that *approaches* the Gaussian rate-distortion bound for any target distortion L-tuple and any well-behaved source is highly desirable. This is pretty much all what we can hope for in practical applications, when the statistics of the source is not known a priori and might not be ergodic.

A typical example is provided by image coding, where the statistics of the output of the "analog" part of the encoder, essentially given by a linear transformation followed by segmentation and decimation, gives origin to blocks of signal $\mathbf{s}$ to be quantized, that are nearly uncorrelated and whose statistics may change from image to image and it is usually estimated adaptively [110].

5.2.2 Code Design

Suppose that we are given a "capacity achieving" spherical code $\mathcal{C}$, for the real AWGN channel with SNR τ. Then, we choose $r_s = \frac{1}{2}\log_2(1+\tau)$, $\Delta = 1/(1+\tau)$ and $\alpha = \sqrt{\tau/(1+\tau)}$. We can write the source vector as

$$\mathbf{s} = \sum_{\ell=1}^{L} \alpha\sqrt{\Delta^{\ell-1}}\mathbf{c}_{g_\ell(\mathbf{s})} + \mathbf{e}_L \tag{5-12}$$

where $\mathbf{e}_L$ is the representation error vector at level L. By interpreting (5-12) as the output of a multiple-access channel with background noise $\mathbf{e}_L$, we notice that the levels are successively decodable by stripping in the order $1,\ldots,L$. In fact, the interference plus noise ratio (SINR) seen by stage ℓ of the stripping decoder is given by

$$\frac{\alpha^2\Delta^{\ell-1}}{\Delta^L + \alpha^2\sum_{j=\ell+1}^{L}\Delta^{j-1}} = \tau \tag{5-13}$$

Unfortunately, spherical codes that work very close to the AWGN capacity and admit minimum distance decoders with practical complexity (say, polynomial in the block length) have not been found so far. If they were available, both the problems of channel coding and of source coding would have been already solved. Hence, driven by complexity considerations, we propose to use as basic code a trellis-terminated binary convolutional code with binary antipodal modulation (i.e., mapping the alphabet $\{0,1\}$ onto $\{+1,-1\}$). In this case, the minimum distance decoder $Q_\alpha(\cdot)$ is efficiently implemented by the Viterbi algorithm.

Since trellis-terminated convolutional codes with fixed (not increasing with the block length) trellis complexity do not approach the AWGN capacity, the choice of α and Δ according to a threshold SNR τ outlined at the beginning of this section is not optimal any longer. On the contrary, for a given basic code we find the optimal scaling factor α and the resulting optimal distortion Δ numerically. Let $\mathbf{s}$ be Gaussian i.i.d. $\sim \mathcal{N}(0,1)$. By Monte Carlo simulation, we find

$$\alpha = \arg\ \min_{\beta\geq 0}\ \mathbb{E}\left[d(\mathbf{s},\beta\mathbf{c}_{Q_\beta(\mathbf{s})})\right] \tag{5-14}$$

and the resulting distortion is given by $\Delta = \mathbb{E}\left[d(\mathbf{s}, \alpha \mathbf{c}_{Q_\alpha(\mathbf{s})})\right]$. Fig.5-3 shows $\mathbb{E}\left[d(\mathbf{s}, \beta \mathbf{c}_{Q_\beta(\mathbf{s})})\right]$ versus β, where the optimal pair (α, Δ) is clearly evidenced.

It is interesting to observe that for the optimal value of α and Δ, the convolutional code works "above capacity". More precisely, suppose that we can write the source $\mathbf{s}$ as

$$\mathbf{s} = \alpha \mathbf{c}_{Q_\alpha(\mathbf{s})} + \mathbf{e}_1 \tag{5-15}$$

where $\mathbf{e}_1$ is the representation error signal, such that $\mathbb{E}[\frac{1}{k}|\mathbf{e}_1|^2] = \Delta$ (by definition). If we interpret (5-15) as a binary-input AWGN channel, its SNR is given by α^2/Δ. The corresponding capacity, $C_{\text{biawgn}}(\alpha^2/\Delta)$, is found to be less than the rate r_s of the basic code $\mathcal{C}$. For example, for the code of rate $r_s = 1/4$ and 128 states of Fig.5-3 we find $\alpha = 0.52$ and $\Delta = 0.729$, yielding capacity $C_{\text{biawgn}}(\alpha^2/\Delta) = 0.2268$, which is less than $1/4$.

In some sense, this explains why using LDPC [26] or Turbo Codes [1] as quantizers is hopeless. These codes have a very sharp behavior around their SNR threshold. For SNRs larger than the threshold they achieve very small bit-error probability, while for SNRs smaller than the threshold their error probability is very large. The iterative Belief-Propagation decoder is clearly unable to find a codeword if the channel SNR is below the code threshold. Since the code threshold is strictly larger than the capacity SNR threshold, and since for quantization we have to work with a "test channel" whose SNR is *less* than the SNR capacity threshold, it is clear that codes under Belief-Propagation iterative decoding cannot work as quantizers.

Another countermeasure we take to partially compensate for the gap of binary convolutional codes from the AWGN capacity consists of introducing unitary transformations at each level such that the signals input to the Viterbi decoders look like Gaussian. In particular, let $\mathbf{U}_\ell$ denote a unitary transformation of $\mathbb{R}^k$. Each Viterbi decoder at level ℓ computes

$$g_\ell(\mathbf{s}) = Q_{\alpha\sqrt{\Delta^{\ell-1}}}(\mathbf{U}_\ell(\mathbf{s} - \widehat{\mathbf{s}}_{\ell-1})) \tag{5-16}$$

Then, the representation vector at level ℓ is given by

$$\widehat{\mathbf{s}}_\ell = \widehat{\mathbf{s}}_{\ell-1} + \alpha\sqrt{\Delta^{\ell-1}}\mathbf{U}_\ell^{-1}\mathbf{c}_{g_\ell(\mathbf{s})} \tag{5-17}$$

Ideally, we should select the unitary transformations independently at each level, according to the Haar measure, i.e., uniformly distributed on the manifold of unitary $k \times k$ matrices. This approach requires common randomness between encoder and decoder, and might be seen as a spherical version of the *dithering* approach commonly used in lattice quantizers [111]. In fact, since lattices are additive groups, randomization with lattice quantizers is obtained by *translating* the source vector by a dither vector $\mathbf{u}$ uniformly distributed over the

lattice Voronoi cell. In our case, since spherical codes obtained from binary convolutional codes are *multiplicative* groups, we obtain randomization by *rotating* the source vector by a unitary matrix $\mathbf{U}$ uniformly distributed over the unit sphere (Haar measure), and hence also over the code Voronoi cell because of the geometric uniformity property.

Notice that both translations and rotations are *isometries* of $\mathbb{R}^k$, therefore, they preserve Euclidean distance (distortion). This means that the only effect of randomization via the unitary transformation is to present to each level Viterbi decoder a signal whose statistics is more *adapted* to the basic code.

We notice also that this approach might be extended to other families of spherical geometrically uniform codes, such as linear trellis codes over $\mathbb{Z}_M$ mapped to the M-PSK constellation [112].

In practice, sampling elements from the Haar measure is quite computationally intensive for large dimension k. Moreover, matrix-vector multiplications have complexity $O(k^2)$ and matrix inverse $O(k^3)$. Also, precomputing and storing $k \times k$ real matrices with no special structure is highly impractical for large k. Hence, for the sake of complexity and practical implementation, we propose the use of structured unitary transformations given by

$$\mathbf{U}_\ell = \mathbf{\Pi}_\ell \begin{bmatrix} \mathbf{C} & -\mathbf{S} \\ \mathbf{S} & \mathbf{C} \end{bmatrix} \tag{5-18}$$

where $\mathbf{\Pi}_\ell$ is a random permutation of size k (interleaving), $\mathbf{C} + j\mathbf{S} = \sqrt{\frac{2}{k}}\mathbf{F}$ and where $\mathbf{F}$ is the Fourier matrix of dimension $k/2$, with (n, m) elements $e^{-j\frac{4\pi}{k}mn}$, for $m, n \in \{0, \ldots, k/2 - 1\}$. In this way, the product $\mathbf{U}_\ell\mathbf{x}$ can be efficiently computed by FFT and interleaving. Fig. 5-4 shows the block diagram of the proposed Multistage Trellis Quantizer (MTQ). In standard TCQ [59], a trellis code defined over a multilevel alphabet is used. The resulting code is similar to Ungerboeck TCM [60]. It turns out that the probability with which the points in the code alphabet are selected is not uniform. Hence, rate improvement can be obtained by binary labeling the points with variable-length labels. A modified Max-Lloyd algorithm that exploits Viterbi decoding and training vectors is used in order to optimize the code alphabet and the binary representation of the points. This approach is generally known as ECTCQ, entropy-constrained TCQ. The best known trellis quantizers for standard i.i.d. sources such as Gaussian, uniform and Laplacian, are found in the family of ECTCQ [113]. It is natural to ask if some rate improvement can be achieved in our scheme by applying entropy coding on the quantization indexes $g_\ell(\mathbf{s})$. Notice that $g_\ell(\mathbf{s})$ is the sequence of information bits (input to the convolutional encoder) that corresponds to the codeword found by the Viterbi algorithm in (5-16). We run some experiments by applying the Burros Wheeler Transform-Minimum Description Length (BWT-MDL) source modeler of [114]. This modeler identifies the tree source model that best explains the bi-

nary sequence $g_\ell(\mathbf{s})$ by using the Burrows-Wheeler transform and the Minimum Description Length principle, i.e., the tree source model for which the overall description length (including coding and model redundancy) of $g_\ell(\mathbf{s})$ is minimized. We simulated 2000 independent source sequence of length $k = 1000$ and we computed the empirical entropy of $g_\ell(\mathbf{s})$ according to the BWT-MDL model. For all simulated frames this was always equal to 1 bit per symbol. This shows that the output of our multistage quantizer is close to an i.i.d. sequence of fair bits and that, in practice, post-processing entropy coding cannot improve performance.

Figures 5-5, 5-6 and 5-7 show the performance of the multistage trellis quantizer for Gaussian, Laplacian and uniform sources, in terms of RSNR defined as $-10\log_{10}\Delta^\ell$ vs $R_s = \ell r_s$, with $r_s = 1/4$ and 128 states. The performance is compared with the optimal RSNR achieved by the distortion rate function. In the case of Laplacian and uniform sources we plot also the Shannon's lower bound (SLB) [46] and the RSNR obtained with the Gaussian distortion rate function. We can see that uniform and Laplacian sources achieve exactly the Gaussian performance, respecting Lapidoth's result [75]. Note that the deviation of the MTQ with respect to the limit in high rate region is not due to simulations but to the method itself. In fact, recall that the MTQ can theoretically achieve distortion $D_\ell = \Delta^\ell$ at rate $R_\ell = \ell r_s$, while the distortion-rate function is given by $D_G = 2^{-2\ell r_s}$. If we consider the RSNR it follows that $-\log D_\ell = \ell \log 1/\Delta$ and $-\log(D_G) = 2\ell r_s \log(2)$, where $\Delta > 2^{-2r_s}$. The different slope of the two curves is due to the fact the scaled convolutional code is only an appoximation of the ideal sperical code.

5.2.3 Soft Reconstruction? Systematic Recursive Convolutional Codes or not?

As far as the reconstruction is concerned, several recent works focused on soft reconstruction, where the channel decoder provides soft-output symbol-by-symbol information and this is used by the source decoder to mitigate the effect of residual post-decoding channel errors. In the same spirit of ubiquitous "EXIT charts" [27], we may model the channel decoder soft output as provided by a BI-AWGN *extrinsic channel*. Let $q_{j,\ell} \in \{0, 1\}$ denote the j-th binary symbol of the source encoder index $g_\ell(\mathbf{s})$, for $j = 1, \ldots, r_s k$. We model the posterior log-likelihood ratio provided by the channel decoder for the (ℓ, j)-th symbol as

$$\mathcal{L}_{j,\ell} = \mu(1 - 2q_{j,\ell}) + \sqrt{2\mu}\mathcal{N}(0, 1) \tag{5-19}$$

where $\mu > 0$ is a parameter. Let $J(\mu) = I(q_{j,\ell}; \mathcal{L}_{j,\ell})$ denote the mutual information (as a function of μ) of $q_{j,\ell}$ and $\mathcal{L}_{j,\ell}$. Optimal soft reconstruction of the ℓ-th level codeword in the Minimum Means Squared Error (MMSE) sense, given the sequence of (independent) LLRs defined above, is obtained by computing the non-linear MMSE estimator of each i-th

codeword symbol as

$$\widehat{c}_i = \sum_{\mathbf{c}\in\mathcal{C}} c_i P(c_i|\{\mathcal{L}_{j,\ell} : j = 1,\ldots,r_s k\}) \tag{5-20}$$

We notice that (5-20) involves only the symbol-by-symbol posterior probabilities $P(c_i|\{\mathcal{L}_{j,\ell} : j = 1,\ldots,r_s k\})$, that can be readily and efficiently computed by the BCJR algorithm [115] acting on the trellis of the basic code $\mathcal{C}$ with input $\{\mathcal{L}_{j,\ell} : j = 1,\ldots,r_s k\}$ for the information bits (convolutional encoder input) and input zero for the coded bits (convolutional encoder output).

In Fig.5-8 we show the reconstruction SNR versus the extrinsic channel mutual information $J(\mu)$, for hard reconstruction (corresponding to making hard decisions $\widehat{q}_{j,\ell} = 1\{\mathcal{L}_{l,\ell} < 0\}$ and feeding these into the convolutional encoder) and for soft reconstruction based on the BCJR algorithm, when the basic code of the multistage scheme has non-recursive non-systematic (NN) and recursive systematic (RS) encoders. We notice that there is almost no difference between hard and soft reconstruction in both cases. Hence, the more complex BCJR reconstruction is not needed. However, there is a noticeable difference between NN and RS realizations of the encoders (notice that the code is the same for both realizations, so its distortion in the absence of the noisy channel is identical in both cases). Not surprisingly, Fig. 5-8 shows that the NN encoder has better conditioned inverse than the RS encoder.

5.3 LOSSY ADAPTIVE TRANSMISSION OVER NOISY CHANNELS

We consider the transmission of a source S over a channel $P_{Y|X}$. The decoder must provide a reproduction of the source such that end-to-end distortion is minimized.

Practical source encoder and decoder are too sensitive to channel errors, this implies very strong requirements in terms of residual BER at the output of the channel decoder. This is mainly due to the catastrophic behavior of the source encoding inverse function. The non-catastrophic behavior of convolutional *encoders* has been widely studied. We know that convolutional codes admits non-catastrophic encoders such that small Hamming distance between encoder input sequences cause small distance in encoded sequences, and vice versa. In particular, this is the case of feedback-free non-catastrophic convolutional encoders [116]. Our multistage source encoder inherits the property of having well-conditioned inverse function from its basic code component.

Driven by this consideration, we shall consider the concatenation of the multistage source code with a channel code.

In general, the best possible performance is achieved by separation. Namely, let η denote

spectral efficiency measured by the number of source symbols per channel use (equivalently, by the ratio of the (discrete-time) source bandwidth over the (discrete-time) channel bandwidth). Let $R(D)$ denote the source rate-distortion function and $C(\Gamma)$ denote the channel capacity-cost function. Hence, spectral efficiency η can be achieved with distortion D and input cost Γ if and only if

$$\eta \leq \frac{C(\Gamma)}{R(D)} \tag{5-21}$$

For fixed spectral efficiency, the best achievable distortion as a function of the channel input cost is given by $D_{\rm opt} = R^{-1}(C(\Gamma)/\eta)$. In our example, for simplicity, we fix the channel to be a binary-input AWGN (BI-AWGN) channel, defined by

$$y = \sqrt{\Gamma}x + \nu \tag{5-22}$$

where $x \in \{-1, +1\}$ with energy per symbol E_s, $\nu \sim \mathcal{N}(0, 1/2)$, Γ is the signal to noise ratio $\Gamma = E_s/N_0$. and the source to be Gaussian i.i.d. with quadratic distortion. Notice that in this case the conditions of [45] do not hold, hence we *have* to code the source and the channel in smart ways. The multistage source encoder produces the indexes $(g_1(\mathbf{s}), \ldots, g_L(\mathbf{s}))$ in the form of binary sequences. Namely, $g_\ell(\mathbf{s})$ is the sequence of information bits corresponding to the codeword $\mathbf{c}_{g_\ell(\mathbf{s})}$ selected by the Viterbi decoder at level ℓ. As channel codes we may consider any family of good binary codes for the BI-AWGN channel. In particular, in our example we considered convolutional codes with 64 states and rates $1/4$, $1/3$, $1/2$, $2/3$, $3/4$, $5/6$, and the turbo code with component generators (37,21) (octal notation) taken from [1] with interleaving size 65536 and puncturing in order to have rates $1/3$, $1/2$, $2/3$, $3/4$, $5/6$, $11/12$. We run experiments by using LDPC codes with optimized right and left degree distributions [87]. In this case we can scan the rates with higher granularity and we consider all the possible channel rates such that $\eta = \frac{R_c}{Lr_s}$, in particular $R_c = L/12$ and $L = 1, \ldots, 12$.

The source code is based on the convolutional code of rate 1/4 and 128 states already used in Fig. 5-3.

Figures 5-9 and 5-10 show the resulting distortion for $\eta = 1/3$ versus the channel SNR, defined as $\Gamma = E_s/N_0$. The separation limit is shown for comparison. Remarkably, the performance of the turbo-coded and LDPC system is quite close to the theoretical optimum. Note also the LDPC codes achieve slightly better performances than Turbo codes. This difference is due to the fact that LDPC codes do not need puncturing to obtain different rate but they are optimized for the given coding rate. Degradation comes from two effects: a horizontal displacement due to the SNR gap of the punctured turbo codes with respect to their capacity limit, and a vertical displacement due to the gap of the multilevel source code with respect to its rate-distortion limit.

In practice, coupling our multilevel source code with channel coding of different rates can easily implement a variable-quality scheme that operates at fixed target spectral efficiency and adapts itself to the user SNR condition.

5.4 JOINT SOURCE CHANNEL CODE BASED ON TURBO CODES: M-TCOM

This section deals with the practical construction of JSCC based on Turbo codes. The scheme is realized by coupling the best scalar quantizer known so far, the Entropy Constrained Scalar Quantizer and a compression/protection multilevel scheme based on Turbo Codes. The performance is given for AWGN channel.

With properly chosen rate and puncturing, the system is able to outperform the conventional Separated Source Channel Coding (SSCC) setup that consists on the concatenation of the same source encoder, an arithmetic encoder and the best Turbo code. EXIT Chart [27] analysis gives us insight on the choices of particular puncturing pattern and component codes for Turbo codes. The concatenation of this scheme with practical source encoder, like modified DPCM, validates its advantages over conventional schemes.

Shannon's source coding theorem states that a binary memoryless source $\mathrm{U} = u_1, u_2, ...u_n$ can be lossy compressed up to its entropy $H(\mathrm{U})$ [46]. When the compression rate is lower than the entropy of the source, then the compression introduces a distortion. Obviously when the source is not discrete the quantization introduces always a distortion. It was shown in [107] that there is strong correlation between almost noiseless fixed-length data compression and almost noiseless coding of a discrete signal-additive noise whose noise has the same statistics as the source. This analogy can be exploited by using linear error correction channel codes such as LDPC codes for data compression [107]. Turbo codes which is able to give near Shannon limit performance with the low complexity iterative decoding is another well suited candidate. Garcia-Frias showed that desired compression ratio can be achieved by properly punctured Turbo codes [108], in particular by puncturing the information bits and the parity bits. A priori probability of the source is used to modified the extrinsic information in the iterative decoding process. When channel error is present, the rate of the Turbo codes has to be selected such that it complies with Shannon's channel coding theorem, in which case less puncturing is needed for better error protection.

5.4.1 M-TCOM System Structure

The proposed system is illustrated in Figure 5-11. In the following the development and the rationale is carried out for ECSQ, but it can be easily extended to ECTCQ. Conventional separated approaches (figure 5-12) achieve rate reduction by using arithmetic coding at the output of the quantizer. When the block length is sufficiently high, the rate of the arithmetic codes approaches the entropy of the sources [117]. However, they are very sensible to channel error. A single bit in error at the output of channel decoder can eventually propagate forever. That is the reason why these schemes need very strong BER conditions at the

output of channel decoder. Instead of arithmetically encoding the quantization indexes, the JSCC scheme provides compression and error protection via turbo codes as explained in the following.

Call $\mathbf{q} \in \mathbb{N}^{N'}$ the quantization indexes vector, $q \in (0, \dots Q-1)$. Suppose further that $Q = 2^L$. The sequence $\mathbf{q}$ is an independently identically distributed sequence with a probability mass function (pmf) $P_Q(q)$. The indexes are mapped into binary bitstream by using a one-to-one mapping $\mu : \mathbb{Z}_Q \rightarrow \mathbb{F}_2^L$ such that $\mu(q) = (\mu_1(q), \dots, \mu_L(q))$. Call $\mathbf{B}_\ell$ the ℓ-th bitplane obtained by applying the mapping μ_ℓ to the sequence $\mathbf{q}$ componentwise, i.e $\mathbf{B}_\ell = \mu_\ell(\mathbf{q})$, for $\ell = 1, \dots, L$. After interleaving, each bitplane is mapped onto a distinct channel codeword $\mathbf{x}_\ell$, such that the composition of all the codewords $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_L)$ represents the transmitted codeword. The total codeword is then passed through the AWGN channel and the observation is given by $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_L)$,

$$\mathbf{y} = \sqrt{\Gamma}\mathbf{x} + \boldsymbol{\nu}$$

where $\boldsymbol{\nu}$ is the circularly symmetric AWGN with per component variance $1/2$, Γ is the signal to noise ratio.

Define $b_{1:\ell} \stackrel{\Delta}{=} (\mu_1(q), \dots, \mu_\ell(q)) = \mu_1^\ell(q)$. For each bit plane we define the conditional marginal probability at level $\ell = 1, \dots, L$ as

$$\begin{aligned} P_\ell(0|b_{1:\ell-1}) \stackrel{\Delta}{=} & P(\mu_\ell(q) = 0 | \mu_{\ell-1}(q) = b_{\ell-1}, \dots, \mu_1(q) = b_1) \\ = & \frac{\sum_{q \in \mathbb{Z}_Q : \mu_\ell(q)=0, \mu_1^{\ell-1}(q) = b_{1:\ell-1}} P_Q(q)}{\sum_{q \in \mathbb{Z}_Q : \mu_1^{\ell-1}(q) = b_{1:\ell-1}} P_Q(q)} \end{aligned} \tag{5-23}$$

By applying the chain rule, we can express the entropy of the samples as:

$$\begin{aligned} H(q) = & \sum_{\ell=1}^{L} H(\mu_\ell(q) | \mu_{\ell-1}(q), \dots, \mu_1(q)) \\ = & \sum_{\ell=1}^{L} \sum_{\mu_1^{\ell-1}} P(b_{1:\ell-1}) H(\mu_\ell(q) | \mu_1^{\ell-1}(q) = b_{1:\ell-1}) \\ = & \sum_{\ell=1}^{L} \sum_{b_{1:\ell-1}} \sum_{q \in \mathbb{Z}_Q : \mu_1^{\ell-1}(q) = b_{1:\ell-1}} P_Q(q) h(P_\ell(1|b_{1:\ell-1})) \\ = & \sum_{\ell=1}^{L} H_\ell \end{aligned} \tag{5-24}$$

where the second summation is done over all the possible vector realizations of the vector $b_{1:\ell-1}$, $P(b_{1:\ell-1})$ is the probability that $\mu_1^{\ell-1} = b_{1:\ell-1}$ and $h(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function and where H_ℓ is defined as

$$H_\ell \stackrel{\Delta}{=} \sum_{b_{1:\ell-1}} \sum_{q\in\mathbb{Z}_Q:\mu_1^{\ell-1}(q)=b_{1:\ell-1}} P_Q(q)h(P_\ell(0|b_{1:\ell-1}))$$

Consider a binary linear systematic code of rate $R = k/n$ defined by a generator matrix $\mathbf{G}$, and let $\mathbf{c} \in \mathbb{F}_2^k$ denote the information vector. The corresponding codeword is obtained as $\mathbf{x} = \mathbf{cG}$ that splits into the *systematic part* and the *parity vector*, called here $\mathbf{u}$. Examples of powerful systematic binary linear codes are turbo parallel concatenated codes [1], and irregular repeat-accumulate codes [66, 118]. In [65, 107, 64], linear codes and iterative *Belief Propagation* decoding are shown to be able to provide data compression. The idea is as follows. Consider for a moment a BSC channel. With slight modification, the rationale can be extended to AWGN channel. Let $\mathbf{c}$ denote a sequence of i.i.d. symbols such that $\Pr(c_i = 1) = p$. Then, we can produce the codeword $\mathbf{x}$ and retain only the parity part $\mathbf{u}$. This is our compressed sequence, of length $n - k$. The compression rate is given by $R_c = 1 - k/n = 1 - R$. Now, if the code is very powerful and is able to approach the capacity $C = 1 - h(p)$ of a BSC with parameter p, then the compression rate is as close as desired to $1 - R = 1 - 1 + h(p) = h(p)$, i.e., to the source entropy. Of course, we have to ensure that the source sequence $\mathbf{c}$ can be reconstructed from the parity sequence $\mathbf{u}$. Let us suppose that the parity part of the code is transmitted via another BSC with crossover probability ρ, then it can be shown (see [64]) that the decoder is fully equivalent to a channel decoder that observes $\mathbf{u}$ via the BSC channel with parameter ρ and $\mathbf{c}$ via a BSC with crossover probability p. In other words, the statistics of the source yields an "equivalent" noise statistics. The achievable transmission rate, in terms of source symbols per channel use, is given by $(1 - h(\rho))/h(p)$.

We can get back now to our case. Suppose that we have a collection of linear binary channel codes with systematic encoders, with information length N' and rate $R_1, \ldots, R_L$. Let $\mathbf{u}_\ell$ denote the parity part of code at level ℓ of length $m_\ell = N'(\frac{1}{R_\ell} - 1)$. Suppose that level-$\ell$ code can recover with high probability the bitplane $\mathbf{B}_\ell$ from the output $\mathbf{y}_\ell$ and using the a priori probability $P_{\ell,m}(0|b_{1:\ell-1})$ defined in (5-23) and the knowledge of the previous bitplanes $\mathbf{b}_1, \ldots, \mathbf{b}_{\ell-1}$, that have already been recovered at the previous decoding stages. This yields a *necessary* condition on the coding rate R_ℓ for successful decoding with high probability, i.e

$$N'/m_\ell \leq (C(\Gamma))/H_\ell \quad \Rightarrow \quad R_\ell \leq \frac{C(\Gamma)}{H_\ell + 1 - h(\rho)} \tag{5-25}$$

where $C(\Gamma)$ is the capacity of the BI-AWGN channel. The overall coding rate of the scheme

is upper-bounded by

$$\begin{aligned} \eta &= \frac{N'}{\sum_{\ell=1}^{\mathcal{L}} m_\ell} \\ &\leq \frac{N'(C(\Gamma))}{N' \sum_{\ell=1}^{\mathcal{L}} \bar{H}_\ell} = \frac{C(\Gamma)}{H(q)} \end{aligned} \tag{5-26}$$

Note that for a perfect channel and an ideal entropy compressor applied to the sequence of quantization indexes $\mathbf{q}$, we obtain a rate of $1/(H(q))$ source symbols per coded bit, which is exactly the best achievable by the lossy source encoder (ECSQ) alone, without the channel.

Each turbo decoder will output the soft information, a posteriori probability, on the information bits for a particular bit-plane. Decoding the L-th bitplanes will provide the a posteriori probabilities of the quantization indexes, $\mathrm{APP}_i(q)$. These soft values are used for the MMSE estimate of the quantizer reconstruction values. The 'soft' reconstruction sequence is given by

$$\hat{s}_i = \sum_{q \in \mathbb{Z}_Q} \phi(q) \mathrm{APP}_i(q) \tag{5-27}$$

is determined for $i = 1, \ldots, N'$, where $\phi(q)$ is the dequantizer operation.

5.4.2 Simulation's Results

In this section we discuss some results in terms of optimization of the Turbo codes and Reconstructed SNR (RSNR) of the M-TCOM approach. We compare the results with the separate scheme SSCC and the MTQ coupled with turbo codes. In the following, the source is Gaussian memoryless. As stated in the previous section in the noiseless case the M-TCOM scheme can achieve the same performance of ECSQ. Call $R(D)$ the Shannon rate-distortion function and $\tilde{H}(D)$ the rate-distortion curve achieved by ECSQ, then, in the limit of large rate, it is possible to show that $\tilde{H}(D) - R(D) \sim 0.25$ bits [117]. Figure 5-13 shows the comparison between the reconstructed signal to noise ratio achieved by ECSQ and MTQ vs rate, Shannon rate-distortion function and the TCQ scheme with 128 state and $R + 1$ Lloyd-Max output points [59]. Note that MTQ outperforms ECSQ for small rate while as long as the rate increases the MTQ diverges from the rate distortion function while the ECSQ has a fixed asymptotic gap from $R(D)$. This shows that in low rate region the M-TCOM scheme coupled with ECSQ will be penalized with respect to MTQ based system. Moreover, note that the performance of MTQ and TCQ is comparable.

We compare, now the M-TCOM scheme and the MTQ, coupled with Turbo codes, when the channel is AWGN. For this matter, we fix a target spectral efficiency $\eta = 1/3$ and a target

Thresholds	-2.66	-1.61	-0.66	0.28	1.22	2.21	3.26	
Probability	$3.9\,10^{-3}$	$4.9\,10^{-2}$	$2\,10^{-1}$	$3.57\,10^{-1}$	$2.78\,10^{-1}$	$9.7\,10^{-2}$	$1.28\,10^{-2}$	$5.5\,10^{-4}$

TABLE 5-1. THRESHOLD VALUES AND PROBABILITY MASS FUNCTION OF THE INDEXES AT THE OUTPUT OF THE ECSQ.

$P_0(0) : 0.611$			
$P_1(0\|0) : 0.087$	$P_1(0\|1) : 0.965$		
$P_2(0\|00) : 0.074$	$P_2(0\|01) : 0.74$	$P_2(0\|10) : 0.36$	$P_2(0\|11) : 0.96$

TABLE 5-2. CONDITIONAL PROBABILITY PER BIT-LEVEL.

RSNR, for example $\mathrm{RSNR}^* = 11.5$dB. ECSQ achieves this RSNR for $H(Q) \sim 2.16$bits. However, since from (5-26) $\eta \leq \frac{C(\Gamma)}{H(Q)}$, the target signal to noise ratio is given by $\Gamma^* \sim 0$dB.

The ECSQ algorithm optimizes the codebook and the value of the quantization thresholds with the constraint that the rate equals the desired entropy [117, 3]. Table 5-1 shows the thresholds and the probability mass function of the indexes at the output of the ECSQ with target entropy $H(Q) = 2.16$bits.

The $Q-$ary to binary mapping is defined simply as the natural binary mapping that transform an index into a binary stream. The definition of the mapping allows to compute the a-priori conditional probability per bit-plane. These are shown in table 5-2. Finally table 5-3 gives the value of the average entropy per bit-plane (first column) and the bound on the rate of the Turbo code per level given by (5-25) (second column).

Note that the matrix of a-priori probability is a side information that needs to be sent error-free to the decoder. However, a slight modification of the blind decoder defined in [65] can estimate, at each iteration, the value of the a-priori probabilities.

However it is not easy to find the exact rate, shown in the second column of table 5-3, by puncturing rate $1/3$ Turbo codes. The results are given by conservatively choosing the rate of the codes. These are shown in the third column of table 5-3. Finally the last column show the polynomial generator used for the simulations.

In [119, 120], the authors show that a priori probability-based polynomial generator selection gives better thresholds behavior. The optimization of Turbo codes is not straightforward and, due to the lack of analytical tools, exhaustive search is needed in order to find good polynomial generator and/or puncturing pattern. The EXIT Chart [27] provides with an approximate and asymptotic threshold and reduces the complexity of this exhaustive search. Figures 5-14 and 5-15 show results in terms of threshold obtained through EXIT Chart when different polynomial generator are considered for level 0 and level 1. It is clear that for the

ℓ	H_ℓ	R_ℓ	$\bar{R}_\ell$	Polynomial
1	0.964	0.428	0.42	$(5,7)$
2	0.345	0.67	0.66	$(31,23)$
3	0.859	0.456	0.44	$(5,7)$

TABLE 5-3. AVERAGE ENTROPY PER BIT-LEVEL, NOMINAL CHANNEL CODE RATE ($\frac{N'}{N'+m_\ell}$), QUANTIZED VALUES AND POLYNOMIAL GENERATOR FOR $H(Q) = 2.16$BITS.

first level polynomial generator $(5,7)$ yields better performance than the others while for the second level, polynomial generator $(31,23)$ outperforms all the others. The threshold found by EXIT Chart is $\Gamma_{th,s} \sim 1.2$dB for the three layers. Further ameliorations of performance can be obtained by asymmetrically puncturing the turbo codes. The number of parity bits punctured at the output of the first convolutional encoder is different with respect to the number of parity bits punctured at the output of the second encoder. The puncturing pattern are randomly generated at each simulated frame. Hence, the results are obtained by averaging over different puncturing pattern. However, it can happen that a particularly "bad" realization of the puncturing pattern occurs, which dominates the performance. Figures 5-16 and 5-17 show the performance in terms of threshold of asymmetric punctured Turbo codes for layer 0 and layer 1 when the best polynomial generator is considered. The performance achieved by symmetric puncturing are also shown for comparison. The EXIT Chart are computed for $\Gamma = \Gamma_{th,s}$. From figure 5-16 we can see that for low rate and polynomial generator $(5,7)$ asymmetric puncturing allows for a better threshold, $\Gamma_{th,a} \sim 0.9$ dB (first level), while figure 5-17 shows clearly that no improvement in the performance is achieved by asymmetrically puncturing the component codes in the second level. The system, however, is limited by the threshold of the worst layer. If the Turbo code at layer ℓ works above its threshold, the propagation of errors decreases the probability of correct decoding at levels $\ell+1,\ldots,L$. However the use of asymmetric puncturing on layer 0 prevents from propagation of errors from layer 0 to 1.

Figure 5-18 show the comparison between MTQ coupled with Turbo codes and M-TCOM. Also shown are the Shannon's bound, the performance achieved by ECSQ in the noiseless case and the threshold obtained via EXIT Chart. The performance suffers from an horizontal and vertical loss mainly due to two factors. First because of the sub-optimality of ECSQ and second because analytical optimization of Turbo codes is not possible and too many parameters influence the behavior of these codes. An open issue is the analytical optimization of different families of codes, as IRA or LDPC, where DE can be written in closed form.

Finally, 5-19 shows the comparison between M-TCOM and SSCC scheme in terms of RSNR versus channel SNR. The Turbo code, in the SSCC, is designed such that the two schemes achieve the same spectral efficiency. In particular the rate of the Turbo code is 0.7

information bits per channel use, and the polynomial generator is $(37, 21)$. We can clearly see the difference between the separated scheme and the joint approach. The separated scheme needs a much higher input mutual information in order to achieve high RSNR performances. This is due to the fact that the performances of the arithmetic code are strongly dependent on the FER performances of Turbo codes and not on BER. The entropy decoder can recover the indexes when the output of the turbo codes is error free, so it needs very low FER performances in order to achieve close to optimal RSNR.

5.5 DPCM AND TURBO COMPRESSION

In this section we briefly generalize the analysis before in order to take into account the particularities of a more practical Differential Pulse Code Modulation scheme for robust transmission of images over the wireless link.

In [57] Kim et al. introduced a low bit-rate predictive image coder, which consists of a modified DPCM coder using multi-rate processing and the Wiener filter. Further rate reduction is achieved through allocating different entropy coders to different areas of the image. The joint decoder exploits the residual redundancy of the channel encoder input bits and produces soft-bits for the reconstruction to be used in the predictive source decoder. This scheme is again to be compared with SSCC setup where the source coder output (the output of the entropy coder) is protected from channel noise by conventional turbo codes. Figure 5.6 illustrates the proposed system. In the predictive encoder, the source image $\mathbf{s} \in \mathbb{R}^N$ is first low-pass filtered and down-sampled two dimensionally. It is then fed through the DPCM encoder, which outputs the difference between the down-sampled image and its prediction. The prediction error image U is quantized using a uniform quantizer and fed back through the prediction loop. The prediction error image still contains residual redundancy and this can be exploited through a classification process based on its varying local statistics. The prediction error image is first divided into blocks of fixed size, typically 4x4 pixels. The block variance σ^2 is used as classification criterion. The image can then be classified into M sub-sources based on the probability distribution of the block variances. A Lloyd-Max like algorithm is used to find the optimal variance representation values of each class such that the over-all average description length is minimized at a given distortion. More details can be found in [58].

For the SSCC approach it is possible to exploit the non-stationarity of the source and to achieve rate reduction by using the Adaptive Entropy Coding (AEC) method [58], i.e by designing M different entropy codes, one for each class. The average code word length

using M sub-sources, $\mathcal{C}_M$, can be expressed as:

$$\mathcal{C}_M = \sum_{m=1}^{M} \mathcal{C}_{\mathrm{EC}_m} = \sum_{m=1}^{M} \int_{\sigma^2_{d,m-1}}^{\sigma^2_{d,m}} \mathcal{C}_m(\sigma^2) p_{\Sigma^2}(\sigma^2) d\sigma^2 \tag{5-28}$$

where $\mathcal{C}_{\mathrm{EC}_m}$ is the average code word length from entropy coder m, $\mathcal{C}_m(\sigma^2)$ is the code word length function for coding class m and $p_{\Sigma^2}(\sigma^2)$ is the probability density function (pdf) of the block variances. The output of the M entropy encoders are concatenated and fed into punctured Turbo codes to obtain rate R_T such that the spectral efficiency is η. Instead of applying arithmetic code as entropy code to the quantization indexes of each class followed by channel coding, the JSCC scheme described above is concatenated to obtain joint compression and protection against channel errors. The analytical formulation defined in section 5.4.1 still holds with slight modification that take into account that now the sequence of indexes is piecewise independently identically distributed where each segment has a probability mass function (pmf) $P_{Q_m}(q)$. Figure 5-21 show an example when $M = 3$.

Without loss of generality, the indexes that belongs to the same class are grouped together, i.e. $\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_M)$ and $q \in \{0, ...Q_m - 1\}$ for $m = 1, \ldots, M$, where $Q_m = 2^{L_m}$. The total number of levels is $\mathcal{L} = \max_m(L_m)$. The total entropy of the source is computed averaging over all the classes

$$\begin{aligned} H(q) &= \sum_{\ell=1}^{\mathcal{L}} \sum_{m=1}^{M} H_m(\mu_\ell(q)|\mu_{\ell-1}(q), \ldots, \mu_1(q))\Pi_m \\ &= \sum_{\ell=1}^{\mathcal{L}} \bar{H}_\ell \end{aligned} \tag{5-29}$$

where we have defined the average entropy per class as

$$\begin{aligned} &H_m(\mu_\ell(q)|\mu_{\ell-1}(q), \ldots, \mu_1(q)) \\ &= \sum_{\ell=1}^{L_m} \sum_{b_{1:\ell-1}} \sum_{q \in \mathbb{Z}_{Q_m} : \mu_1^{\ell-1}(q) = b_{1:\ell-1}} P_{Q_m}(q) h(P_{\ell,m}(0|b_{1:\ell-1})) \end{aligned} \tag{5-30}$$

and the average entropy per level as

$$\bar{H}_\ell \triangleq \sum_{m=1}^{M} H_m(\mu_\ell(q)|\mu_{\ell-1}(q), \ldots, \mu_1(q))\Pi_m \tag{5-31}$$

and where Π_m is the probability of the m-th class and $P_{\ell,m}(0|b_{1:\ell-1})$ is the conditional probability at level ℓ and for the class m, defined in 5-23.

5.5.1 DPCM's Results

In this section we discuss some results in terms of entropy and bound on the rate and in terms of the Peak Signal to Noise Ratio (PSNR) for 8 bit gray scale source image, defined as $10\ \log_{10}\frac{255^2}{D}$ where D is the mean squared distortion per pixel, vs channel capacity. The simulations are run for BSC channel.

The test image is the monochrome 512×512, "Lena" image and the JSCC system has been design to work at a nominal channel parameter $\rho = 0.05$. For the SSCC setup we use arithmetic codes as entropy coders and the coder outputs are coded by same Turbo codes. Figure 5-22 shows the comparison between the JSCC with the rates shown in table 5-4 and the SSCC scheme with rate $R_T = 0.72$ and generator polynomial $(37, 21)$ and $(5, 7)$. The true spectral efficiency (by considering the second column of table 5-4) is $\eta = 1.25$ source sample per channel use, while the actual spectral efficiency (due to quantization of the rate) is equal to $\eta \sim 1.1$ source sample per channel use.

The predictive coding scheme generally has the problem of error propagation. Our bitplane setup ease the problem by relying on low BER performance of turbo codes at the same time avoid spending much synchronization bits as arithmetic codes need. Hence, we can achieve a PSNR equal to the noiseless case for lower input mutual information. Here the results are given for hard reconstruction of the source, further improvements can be obtained through using so-called 'soft-bits' for reconstruction. The investigation of "soft-reconstruction" is an interesting open issue.

5.6 CONCLUSIONS AND FUTURE WORK

This chapter has dealt with the construction of joint source and channel code that achieve close to optimal performance when AWGN channel is considered. These schemes are practical implementations of the tandem encoder used in the HDA scheme introduced in the previous chapter.

The first part has been focused on the construction of multistage source code that guarantee successive refinement of information and almost non-redundant layers. This scheme is shown to be very resistant to channel errors, due to the fact that convolutional codes have non-catastrophic encoders. This makes the scheme suited for concatenation with good channel code as Turbo codes or LDPC to implement adaptive transmission over noisy channel.

The second part has dealt with the construction of joint source and channel codes where a simple quantizer is concatenated with a data compression/channel protection scheme. It is based on Turbo codes. This scheme exploits the redundancy at the source encoder output

and it is less sensitive to channel errors than variable length-based source encoder schemes. Hence, it yields much better results compared to SSCC.

However, the comparison between the multistage trellis quantizer and the Turbo compression based scheme, when the channel is AWGN, reveals that MTQ achieves better performance. We think that analytical optimization of other families of codes different from Turbo codes and the use of ECTCQ could achieve better performance. Hence, the potential of this scheme is not fully explored. Further results on the application of these schemes in the HDA system are working progress.

An interesting generalization is to extend this scheme to achieve progressive image transmission through embedded quantization. The proposed scheme encodes bit-plane by bit-plane and decodes them in sequence, such that bit-plane ℓ can be recovered after having received the corresponding channel output $\mathbf{y}_\ell$ and after having decoded the previous bit-planes at levels $1, \ldots, \ell - 1$. Hence, the scheme is suitable for *progressive transmission*. By choosing the Q-ary to binary mapping μ such that it is embedded, the source can be reconstructed at different levels of distortion from $1, 2, \ldots, \ell$ bit-planes. If the reconstruction operation (from $\hat{\mathbf{q}}$ to $\widehat{\mathbf{s}}$) is linear, then the reconstruction of the bitplanes can be simply added after interpolation for finer resolution. Several open issues are still to be explored in this area, of which some are subject to the author's on-going research, and are discussed in the next chapter.

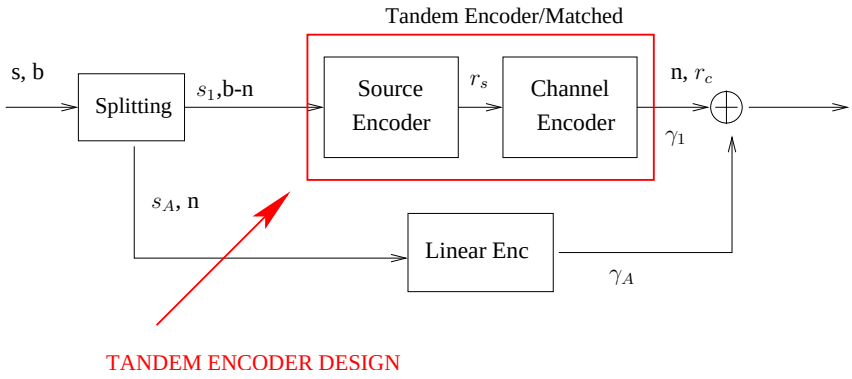


Fig. 5-1. Hybrid digital-analog scheme.

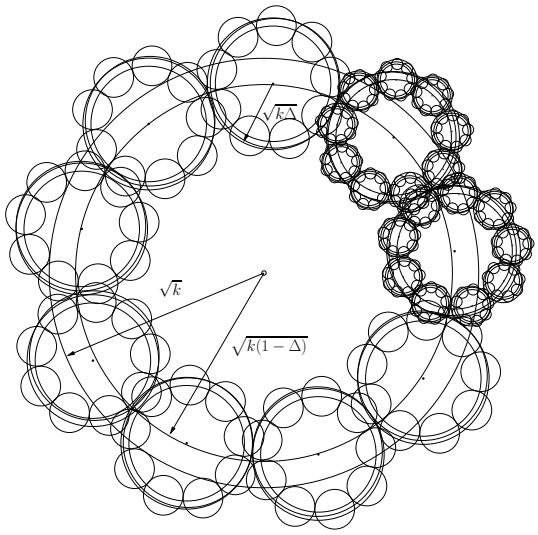


Fig. 5-2. Geometry of a successive refinement source code based on spherical code.

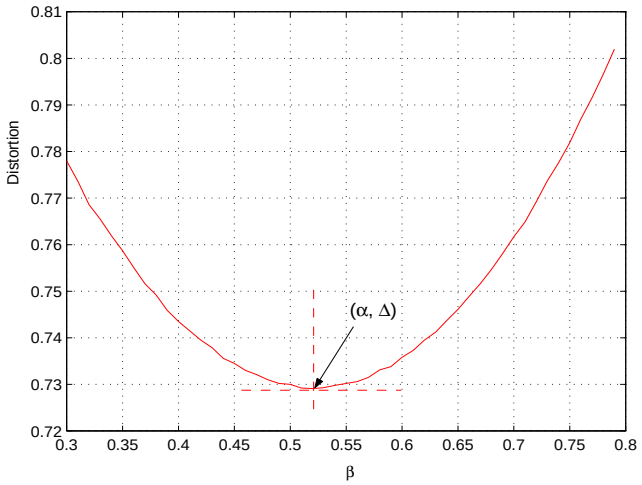
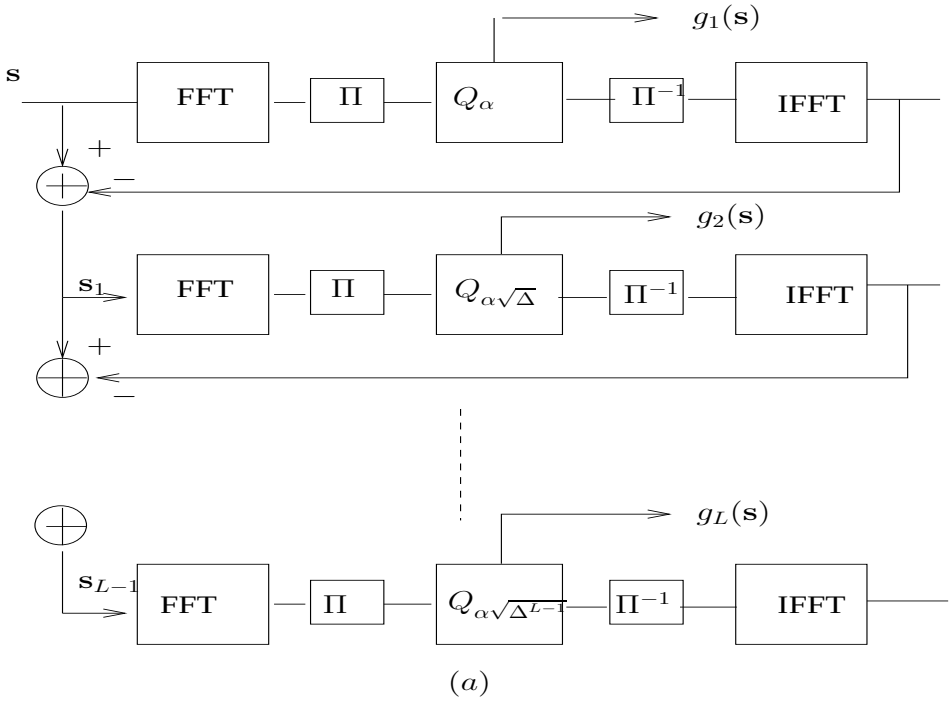
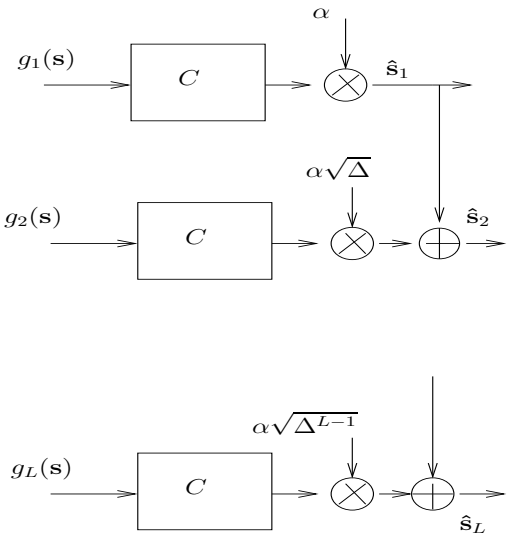


Fig. 5-3. Distortion vs β for $r_s = 1/4$ and 128 states for Gaussian sources.



(a)



(b)

Fig. 5-4. Block diagram of MTQ with FFT/IFFT and interleaving (a) and reconstruction (b).

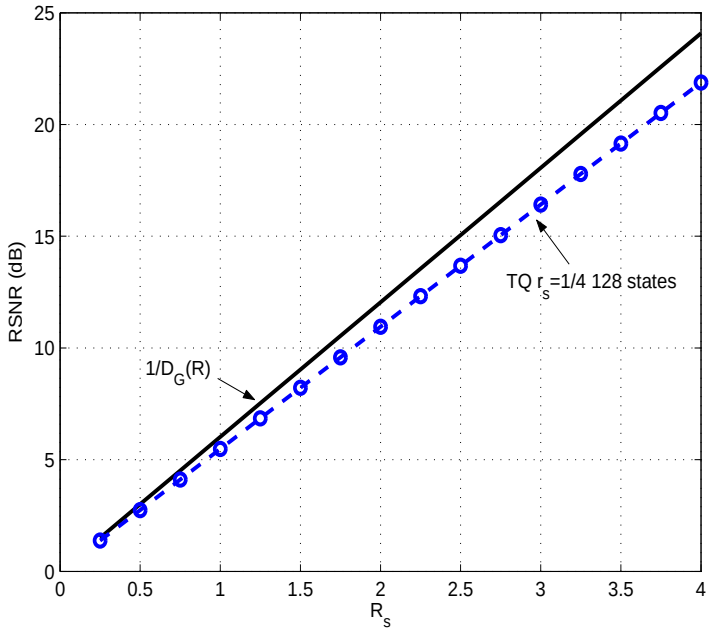


Fig. 5-5. RSNR vs R of MTQ scheme for rate $1/4$ and 128 states for Gaussian sources.

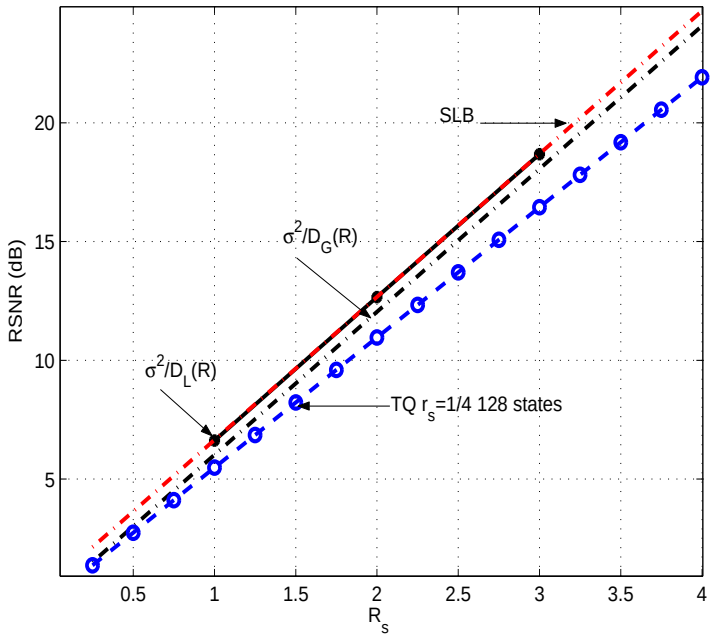


Fig. 5-6. RSNR vs R of MTQ scheme for rate $1/4$ and 128 states for Laplacian sources.

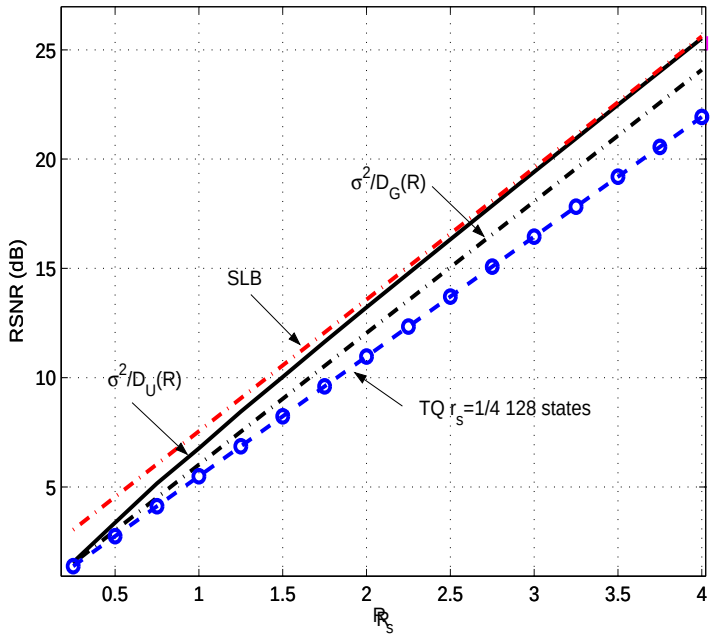


Fig. 5-7. RSNR vs R of MTQ scheme for rate $1/4$ and 128 states for uniform sources.

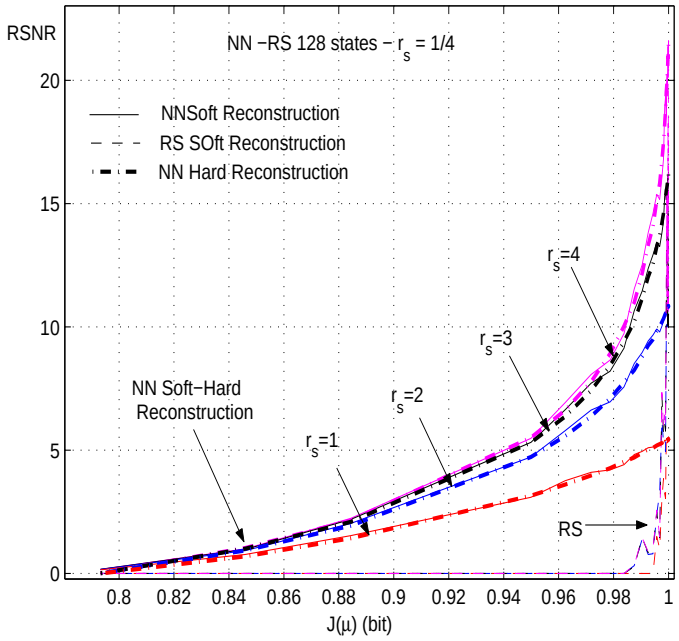


Fig. 5-8. RSNR vs mutual information for hard and soft reconstruction and NN and RS encoders, $r_s = 1/4$ and 128 states.

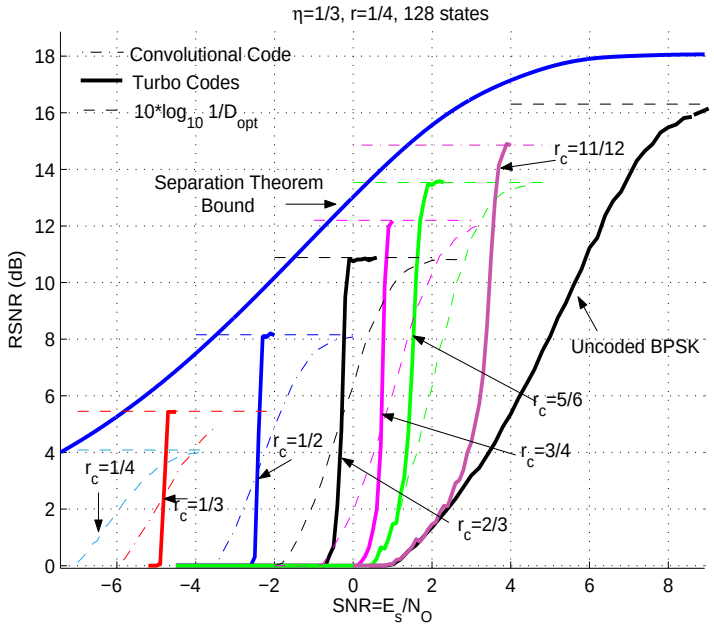


Fig. 5-9. RSNR vs SNR for $r_s = 1/4$ with 128 states. The channel codes are 64 states convolutional codes and $(37, 21)$ turbo codes [1], punctured to obtain different rates. The bound based on separation theorem and the performances of uncoded BPSK transmission are also plotted.

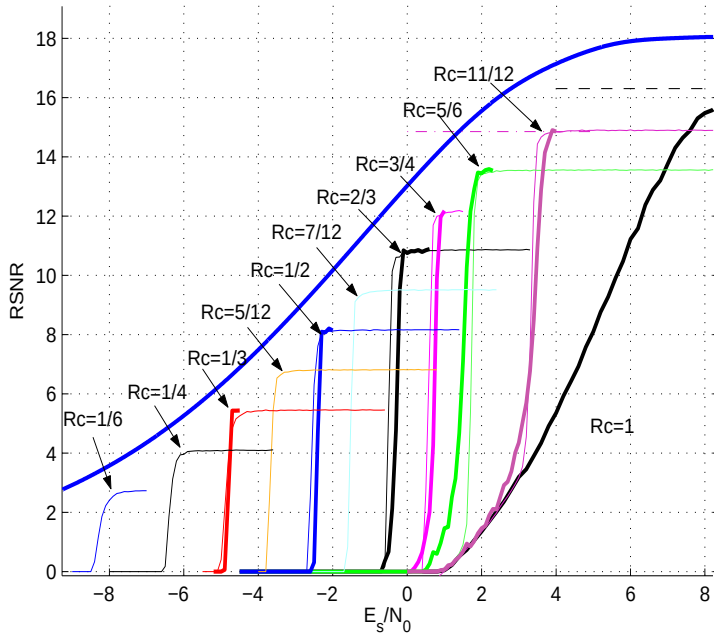


Fig. 5-10. RSNR vs SNR for $r_s = 1/4$ with 128 states. The channel codes are LDPC codes with rate $\eta r_s L$ and $L = 1, \ldots 12$. The results obtained with turbo codes are also plotted for comparison. The bound based on separation theorem and the performances of uncoded BPSK transmission are also plotted.

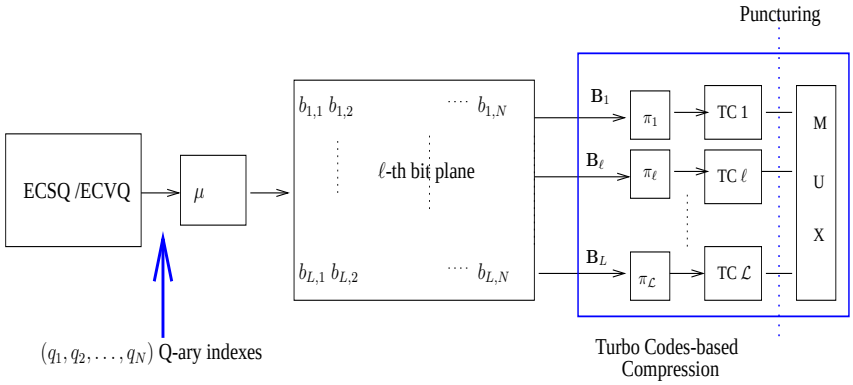


Fig. 5-11. JSCC Using Turbo Compression and Error Protection.

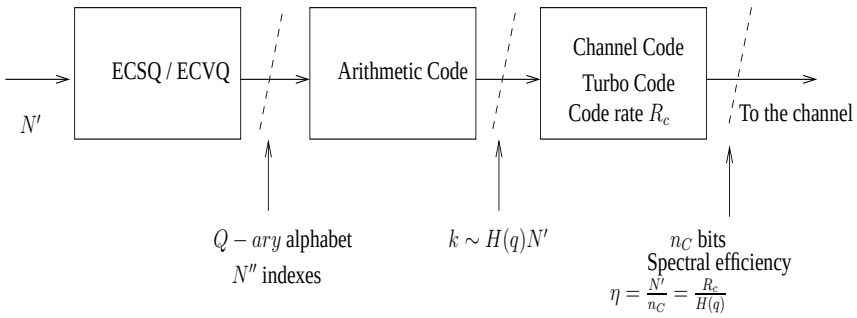


Fig. 5-12. Conventional SSCC scheme.

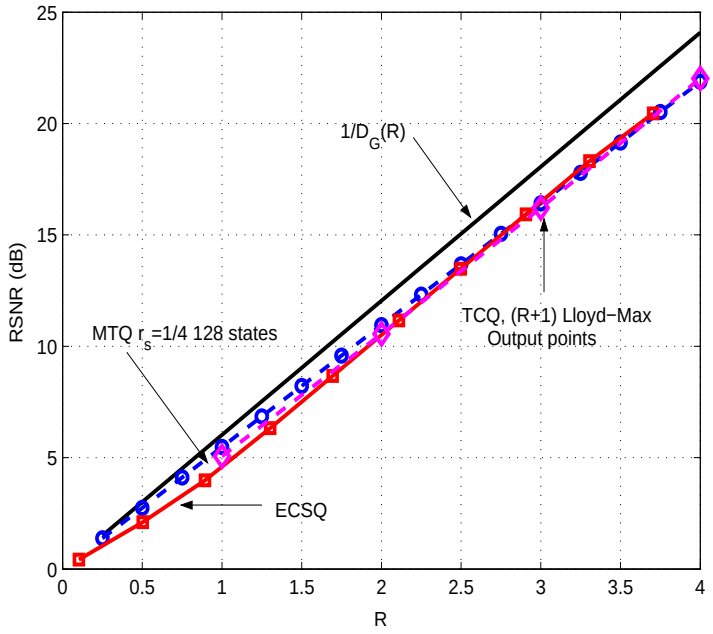


Fig. 5-13. Comparison between $R(D)$, $\tilde{H}(D)$ and the reconstructed signal to noise ratio achieved by MTQ vs rate.

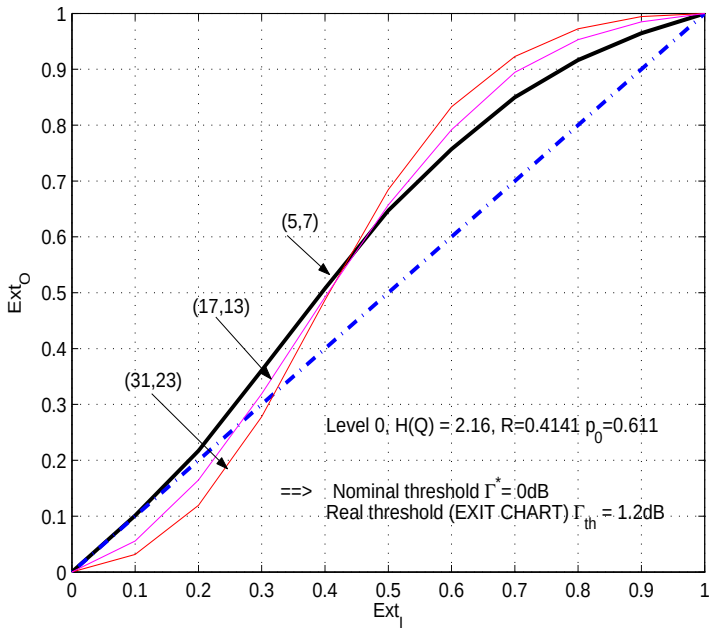


Fig. 5-14. EXIT Chart optimization of polynomial generator for level 0 (see table 5-2).

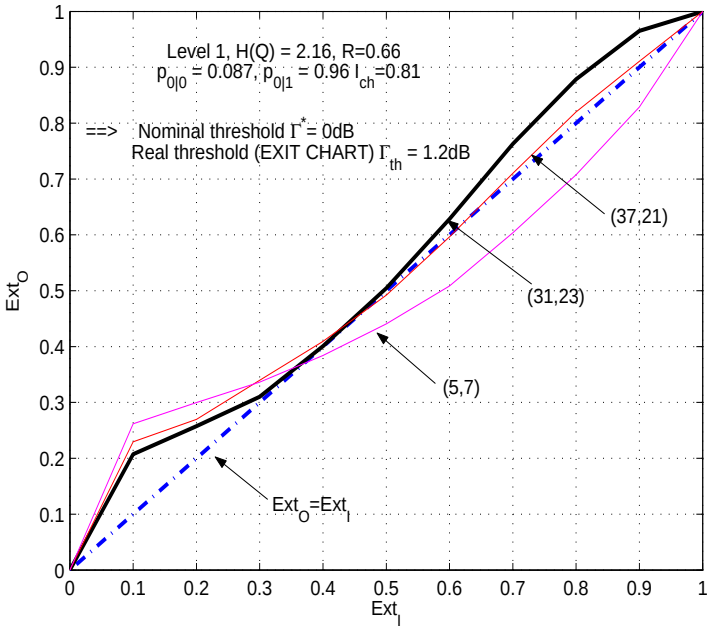


Fig. 5-15. EXIT Chart optimization of polynomial generator for level 1 (see table 5-2).

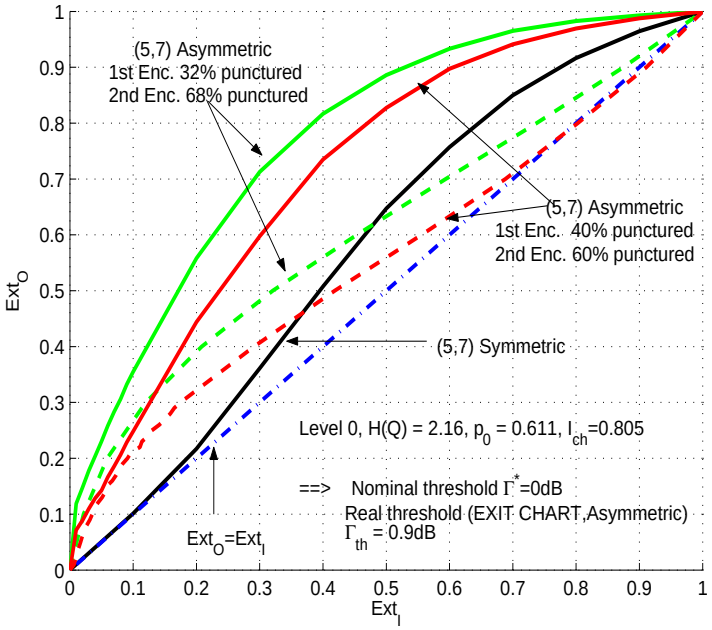


Fig. 5-16. EXIT Chart optimization of polynomial generator for level 0 (see table 5-2).

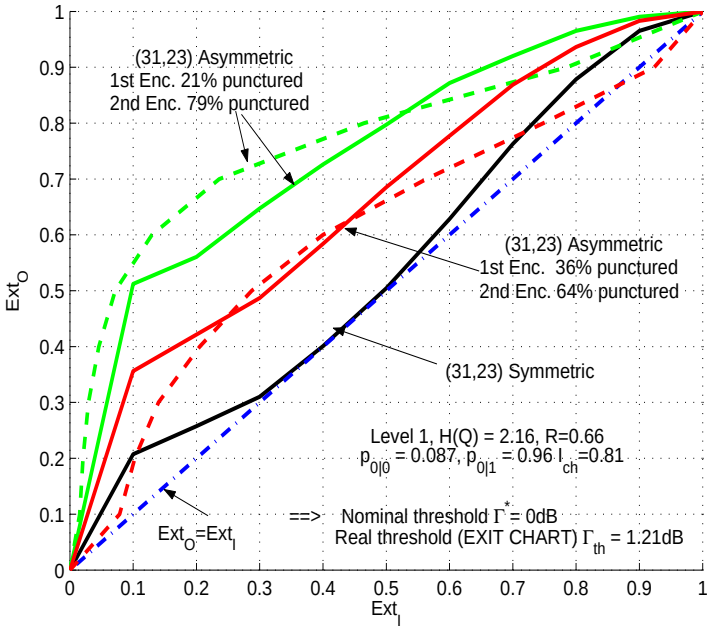


Fig. 5-17. EXIT Chart optimization of polynomial generator for level 1 (see table 5-2).

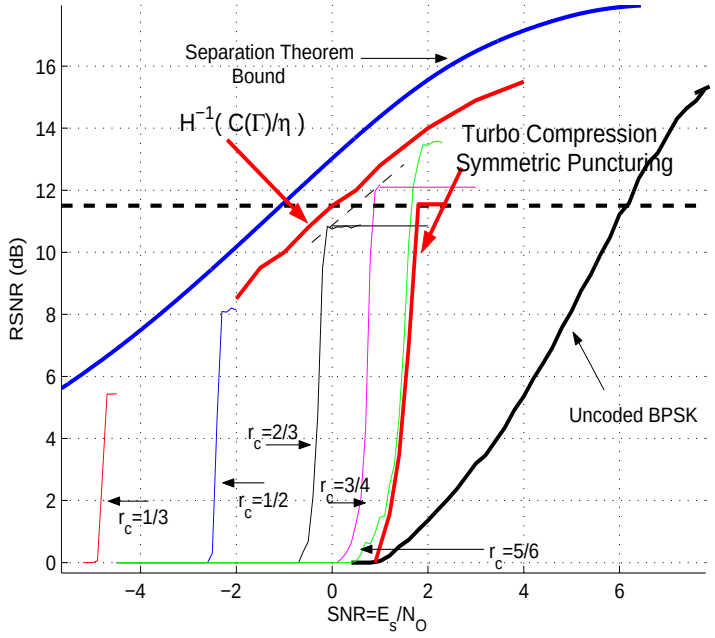


Fig. 5-18. Comparison between MTQ ans M-TCOM.

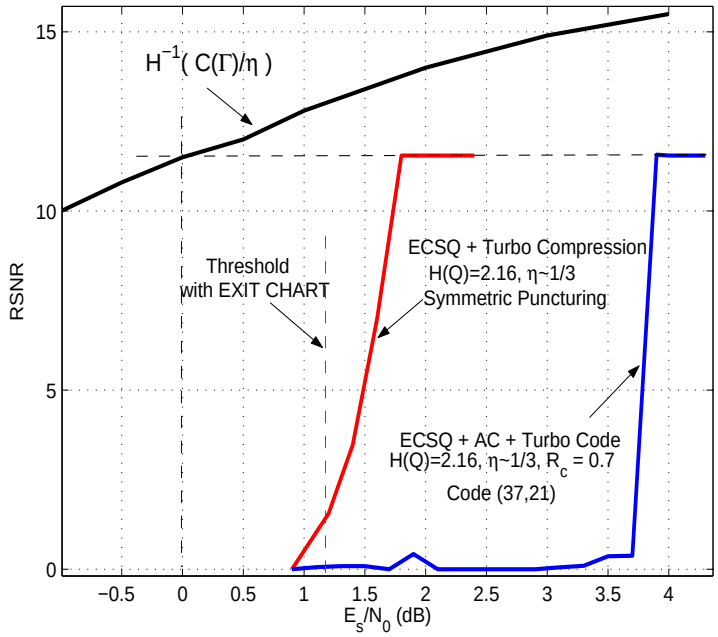


Fig. 5-19. Comparison of the M-TCOM and SSCC based on arithmetic source code.

ℓ	$\bar{H}_\ell$	R_ℓ	$\bar{R}_\ell$	Polynomial
1	0.071	0.9	0.88	$(37, 21)$
2	0.374	0.65	0.64	$(37, 21)$
3	0.734	0.49	0.46	$(5, 7)$
4	0.805	0.46	0.44	$(5, 7)$
5	0.174	0.8	0.78	$(37, 21)$
6	0.079	0.9	0.88	$(37, 21)$

TABLE 5-4. AVERAGE ENTROPY PER BIT-LEVEL, NOMINAL CHANNEL CODING RATE ($\frac{N'}{N'+m_\ell}$) AND QUANTIZED VALUES FOR DPCM AND POLYNOMIAL GENERATOR.

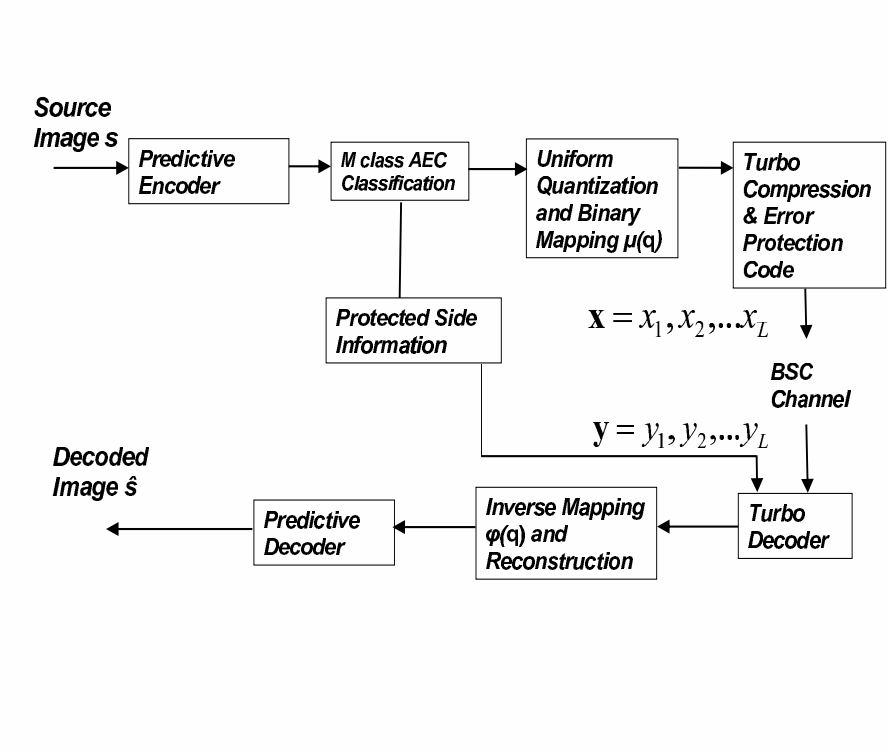


Fig. 5-20. JSCC Using Turbo Compression and Error Protection

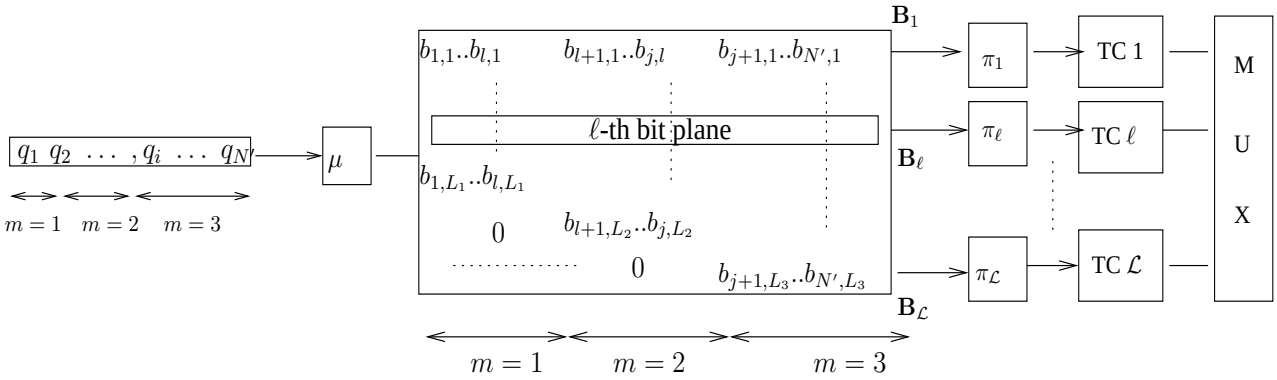


Fig. 5-21. JSCC Using Turbo Compression and Error Protection applied to DPCM output.

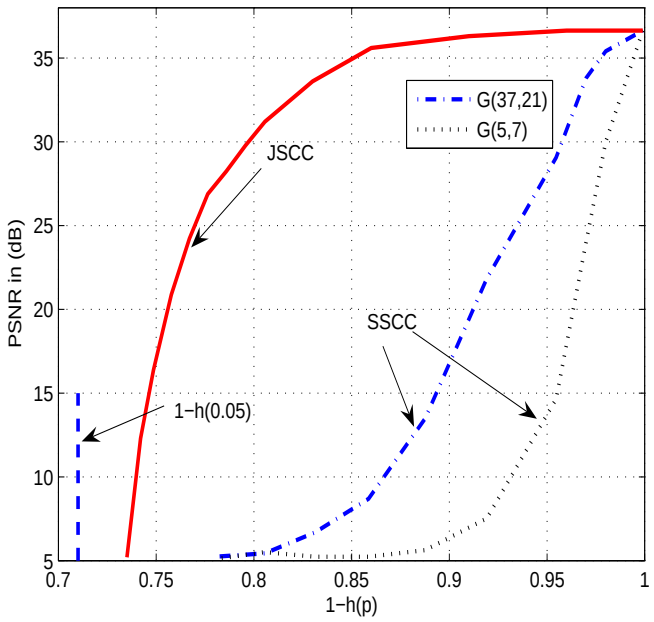


Fig. 5-22. Comparison between JSCC and SSCC. Here the JSCC is with natural binary code as binary mapping. The SSCC are arithmetic codes with turbo codes using the corresponding generator polynomials

CHAPTER 6

Conclusions

In this thesis we have tackled some of the open problems first discussed in chapter 1, related to the concept of efficient transmission of loss- and delay-sensitive data over wireless channels. We address in particular a multicast setting, where the transmitter sends the same common information to all the users in the cell. Depending on the application, certain measures of signal delivery performance (distortion, BER, delay, ..) will be more critical than others. For example in packet oriented transmission the data are not always delay-sensitive but they typically require a quasi error-free link, while analog sources can be delay sensitive but error-tolerant or they can have more relaxed constraints on the delay.

In point to point scenarios, a good trade-off between reliability and efficiency is obtained by coupling ARQ protocols and FEC. This gives rise to hybrid schemes (HARQ) than can easily adapt to channel conditions. FEC handles most frequently occurring errors while ARQ solves remaining FEC decoder failures with a retransmission request. Motivated by the increasing interest in iterative decodable codes we have analyzed the performance of HARQ schemes coupled with LDPC in a single user setting and slowly variant fading. We have shown that ideally these codes approach optimal performances. In this case the analysis is done by means of powerful tools like DE, generalized to take into account block fading conditions and HARQ protocols. However we have shown that practical finite length codes exhibits a considerable loss in performance due to the bad FER behavior. Two effective methods to recover this gap are given and interestingly they achieve almost equal performance, making LDPC codes attractive for implementation with HARQ schemes. The

analysis of the complexity shows however that some saving can be obtained by triggering the decoder only when the probability to decode is high. A method based on asymptotic analysis is considered prior to decode. It prevents using the iterative decoder if it is likely to be non convergent, and it can be coupled with standard methods to stop the iterations to further reduce the complexity. Additionally, some simple modification of the same algorithm allows to achieve all range of trade offs between throughput and complexity. Open issues and extensions could be to explore more sophisticated graph construction in order to improve the FER performance.

However a critical issue is to achieve good trade-offs between reliability, delay, performance in a multicast setting. Hence we have analyzed the throughput performance of HARQ protocols in such a framework. Strictly speaking these protocols are not scalable with the number of users. However, if we are not too ambitious and put a reasonable limit on the performance requirements, these protocols can be made practically scalable. In order to make HARQ protocols fully scalable an expurgated ensemble of users needs to be considered, by selecting only a fraction of users to which the transmission is intended. We show that under particular conditions the throughput of incremental redundancy schemes equals the ergodic capacity of the system but with delay that grows to infinity. For selective repeat based protocols, one achieves optimal performance with finite average delay but at the expense of a penalty in throughput compared to incremental redundancy based scheme. We show that the performance of IR and of FEC coding are identical in terms of delay, throughput and error probability, in the limit of a large number of users.

In many cases, the sources that are transmitted over the network are analog, for example transmission of images, video, voice over the wireless link. The schemes that are practically implemented nowadays are based on the separation principle that states that no loss in performances is incurred by separating source and channel code design. However it does not take into consideration complexity and delay and it does not hold in a non ergodic scenario or in a multiuser scenario. Consequently joint source channel coding technique are attracting a lot of interest in our field of research. These schemes achieve better performance by linking together the source and channel code design. Here we model the multicast scenario with a compound channel. In fact, the compound channel, under the assumption that the encoder is aware of the channel coefficient of the user but it knows the statistic of the fading, can model a Gaussian broadcast/multicast channel with an infinity of users each of one experiencing a different channel coefficient. Further we assumed that the decoder has perfect channel state information.

In this setting we study three different strategies; the first is based on a successive refinement source encoder coupled with a time-sharing transmission scheme, the second couples the same source encoder with a superposition transmission technique. Finally, the third is an Hybrid Digital Analog (HDA) scheme based on bandwidth and power splitting that super-

imposes the output of a digital tandem encoder with an analog encoder. These three schemes are optimized in order to yield minimal end-to-end average distortion and are compared in terms of average distortion vs. average and instantaneous signal to noise ratio. The key conclusions are that superposition schemes and progressive schemes give graceful degradation of performance. Furthermore, superposition schemes achieve better average distortion results. However the hybrid scheme is very close to the OPTA (optimal performance theoretical attainable) curve for a wide range of instantaneous signal to noise ratios, showing very clearly that most of the gain is due to the analog branch. The algorithms that give the optimal transmission parameters in the three cases are also provided and analyzed.

However, for the analytical analysis ideal source/channel codes are considered. Hence, a big issue is the construction of practical codes that can achieve performance close to the limits mentioned above. In particular, we have analyzed the construction of tandem encoder that can be used in the HDA scheme.

A Multistage Trellis Quantizer (MTQ) based on the scaled version of a unique convolutional encoder is shown to give results very close to the distortion rate function. Moreover the results are independent from the statistic of the source, yielding always the same performance as in the Gaussian distributed source case, which is a very useful robustness property in practice. Notably this can be interesting in implementations where the probability density function (pdf) of the source is a mixture of different pdf modeled, in general, as Gaussian. The results obtained with this scheme in an ideal noiseless channel are comparable to the best results found in literature, that have to be found in the family of Trellis Coded Quantizer (TCQ). The latter schemes are known to be very sensitive to channel-related errors, while on the contrary, the multistage scheme proposed here is very robust to errors. This advantage is due to the fact that the convolutional encoder can be non-catastrophic.

Another joint source channel coding scheme based on Multilevel Turbo COMpression (M-TCOM) is analyzed. In this scheme a linear code (Turbo code) is used to compress a redundant digital source. This scheme exploits the fact that the input bits are not fair coins. This a-priori probability is considered to be known at the decoder. This technique can be coupled with entropy constrained scalar/vector quantizers, or the best Entropy Constrained TCQ by mapping the output into binary streams. Practical results on the transmission of images over a BSC channel are also given by coupling our algorithm with a Differential Pulse Code Modulation based quantizer. It then shows remarkable results, especially when compared to the standard approach that consists on concatenating the quantizer with an arithmetic code and a powerful channel code. This scheme is also well suited to progressive transmission of information when we consider an embedded quantizer instead than ECSQ.

However, several issues are the subject of on-going work. First of all more extensive results on the use of the MTQ and M-TCOM scheme in a HDA system, will give more insights

into the behavior of such methods, and it will lead to more accurate conclusions on the construction of these codes.

So far we have shown that the MTQ and the M-TCOM have big potential, but the comparison between the two has been carried out in a special case, i.e. when the output of the ECSQ is compressed with a scheme based on Turbo codes. The results are in favor of the MTQ scheme. However, we think that better results, that could lead to opposite conclusions, can be achieved by considering other families of codes as LDPC or Irregular Repeat and Accumulate (IRA). The degree distribution of these codes, in fact, can be analytically optimized via DE, taking into account the particular structure of the compression method. This is shown to provide advantages compared to the use of Turbo codes.

Moreover, the outstanding results of ECTCQ motivate us to analyze the concatenation of the compression scheme based on optimized IRA with these kind of quantizers, for the transmission of images over the wireless link. This concatenation will surely give remarkable gains, compared to the results given here, and have the potential to approach closely the Shannon's bound.

Overall, our conclusions tend to indicate that Multilevel Turbo/IRA compression and Multistage Trellis Quantizer can be considered as a viable solution for the problem of transmission of images over a wireless link in a multicast setting, where the property of graceful degradation of performance with respect to different signal to noise ratios, is fundamental.

CHAPTER 7

Feedback Systems for Multicasting Common Information

7.1 COMPUTATION OF THE LIMIT FOR $N \to \infty$ OF $\mathcal{V}(p(m), N, x)$

In this section we want to show that

$$\lim_{N\to\infty} \mathcal{V}(p(m), N, x) = \lim_{N\to\infty} \Pr(X_m \leq N - \lceil Nx \rceil) = \ell_m \tag{7-1}$$

where $\ell_m = 1$ if $x < p(m)$, $\ell_m = 1/2$ if $x = p(m)$ and 0 otherwise. The case when $x = p(m)$ is straightforward since we are computing the probability that a Binomial random variable is less than its mean. Let us restrict to the case when $x < p(m)$ meaning that $N - \lceil Nx \rceil > \mathbb{E}[X_m]$. In order to compute the limit in (7-1) standard bounds on the tails of binomial distribution can be used: here we applied the exponential Hoeffding's bound [121] to the Binomial RV $X_m \sim \mathrm{Bin}(N, 1 - p(m))$

$$\Pr(X_m > \mathbb{E}[X_m] + \rho) \leq e^{-\frac{\rho^2}{4\mathrm{Var}(X_m)}} \tag{7-2}$$

In order to show the limits we need to show that $\forall \epsilon > 0$ arbitrarily small $\exists\ N_0$ such that if $N > N_0$ than $|\ell_m - \Pr(X_m \leq N - \lceil Nx \rceil)\,| < \epsilon$. The limit holds by using (7-2) and

setting $\rho = N(p(m) - x) - 1$. In fact,

$$|\ell_m - \Pr(X_m \leq N - \lceil Nx \rceil)\,| = \Pr(X_m \geq N - \lceil Nx \rceil)) \\ \leq \Pr(X_m \geq N - Nx - 1)) \leq e^{-\frac{(N(p(m)-x)-1)^2}{4N(1-p(m))p(m)}} = \epsilon \tag{7-3}$$

The result is shown by setting N_0 solution of $e^{-\frac{(N_0(p(m)-x)-1)^2}{4N_0(1-p(m))p(m)}} = \epsilon$. Analogously for the condition $x > p(m)$ by noticing that $\Pr(X_m \leq N - \lceil Nx \rceil) < \Pr(X_m \leq N - Nx)$.

7.2 PROOF OF THEOREM 2

Theorem
The supremum over $R \geq 0$ of $\eta_\infty(x, R, \Gamma)$ is given by $R(k)/(1+k)$ for some $k = 0, 1, \ldots$, that in general depends on x and Γ. □

Proof: Suppose that the maximum throughput is achieved by selecting $R = R(k) - \delta$ where δ is such that the average delay is still given by $\tau = k + 1$, then

$$\frac{R(k) - \delta}{k+1} < \frac{R(k) - \delta/2}{k+1}$$

contradicts the fact that $R = R(k) - \delta$ is the rate for which the throughput is maximum. Suppose now that the maximum throughput is achieved with $R = R(k) + \delta$. By definition $\tau = k + 2$ and then

$$\frac{R(k) + \delta}{k+2} < \frac{R(k)}{k+1}$$

for δ sufficiently small. This shows that $R(k)$ is a stationary point of $\eta_{\infty,x}$ and the rate that maximize the throughput is one of the $R(k)$.

7.3 PROOF OF THEOREM 3

Theorem
For independent Rayleigh fading SNR Γ and IR protocol, define $G_m(z)$ the cdf of the random variable $\frac{1}{m}\sum_{i=1}^{m} \Delta I_{i,u}$. Define $x_d \stackrel{\Delta}{=} \min(G_m(C(\Gamma)))$. Then for all $x \in (0, x_d)$, $\eta_\infty(x, R, \Gamma)$ is increasing with R. Therefore, $\sup_{R \geq 0} \eta_\infty(x, R, \Gamma)$ is achieved for $R \to \infty$ and $\bar{\tau} \to \infty$, and it is equal to $C(\Gamma)$. Also, for all $x \in (x_d, 1)$ $\sup_{R \geq 0} \eta_\infty(x, R, \Gamma)$ is achieved

for finite delay. □

Proof: Recall that $\eta_{\infty(x,R,\Gamma)}$ is given by

$$\eta_\infty(x,R,\Gamma) = \frac{R}{1+\sum_{m=1}^{\infty} 1\{x \le p(m)\} - \frac{1}{2}\delta(x-p(m))}$$

In the following we skip the dependency of the parameters and we call it simply η_∞ . By definition $p(m) = G_m(\frac{R}{m})$, it follows that

$$\begin{aligned}\eta_\infty &= \frac{R}{1+\lfloor \frac{R}{G_m^{-1}(x)} \rfloor - \frac{1}{2}\delta\Big(m - \frac{R}{G_m^{-1}(x)}\Big)} \\ &\ge \frac{R}{\lfloor \frac{R}{G_m^{-1}(x)} \rfloor} \ge G_m^{-1}(x) \end{aligned} \tag{7-4}$$

We need to show that $\forall\, x \in (0,1)$

$$\lim_{m\to\infty} G_m^{-1}(x) = C(\Gamma) \tag{7-5}$$

Call Z_m the RV defined as $Z_m \stackrel{\Delta}{=} \frac{1}{m}\sum_{i=1}^{m} \Delta I_{i,u}$, with mean $\mathbb{E}[Z_m] = C(\Gamma) = \mu$. For the definition of limit we need to show that $\forall \epsilon\ \exists\, m_0$, such that if $m > m_0$ than $|G_m^{-1}(x) - C(\Gamma)| < \epsilon$. It follows that

$$|G_m^{-1}(x) - C(\Gamma)| < \epsilon \;\;\Rightarrow\;\; \mu - \epsilon < G_m^{-1}(x) < \mu + \epsilon \tag{7-6}$$

The function $G_m(.)$ is continuous and monotonically increasing and thus it follows that $G_m(\mu-\epsilon) < x < G_m(\mu+\epsilon)$. For the central limit theorem Z_m converge in probability to a Gaussian random variable with the same mean and variance. This means that $\forall\ \epsilon_1 > 0$ it is possible to find $m > m_1$ such that $G_m(\mu+\epsilon) \ge 1-\epsilon_1$; $\forall\ \epsilon_2 > 0$ it is possible to find $m > m_2$ such that $G_m(\mu-\epsilon) \le \epsilon_2$. This implies $\epsilon_2 \le x \le 1-\epsilon_1$. Thus it is sufficient to take $m > \max(m_1, m_2)$ s.t $x \in (\epsilon_2, \le 1-\epsilon_1)$ for the limit to hold.

For definition of x_d, it follows that $\forall\, x \in (0, x_d)\ G_m^{-1}(x) \le C(\Gamma)$. In particular $\sup_m G_m^{-1}(x) = C(\Gamma)$ and $\mathrm{argsup}_m\, G_m^{-1}(x) = \infty$ for (7-5).

Also note that $G_m^{-1}(x) = \frac{R(m)}{m}$. If the limit for $m \to \infty$ is a constant different from zero and infinity, than also $R = \Theta(m)$. This means that

$$\sup_R\ \eta_\infty = \sup_m\ \eta_\infty = C(\Gamma)\ \forall\ \ x \in (0, x_d)$$

For $x \in (x_d, 1)$, for definition it exists a value $m < \infty$ such that $G_m^{-1}(x) > C(\Gamma)$, this already show that $\sup_R \eta_\infty > C(\Gamma)$ and $\mathrm{argsup}_R\, \eta_\infty < m$.

7.4 PROOF OF THEOREM 4

Theorem:
For the SR protocol, $\sup_{R\geq 0}\eta_\infty(x,R,\Gamma)$ **is always achieved for finite** R **and delay. In particular the optimal** R **can be found as** $R(k)=F(k+1)-\epsilon$ **where** k **is the index that maximize the following sequence**

$$b[i]=\frac{1}{i+1}[F(i+1)-\epsilon] \tag{7-7}$$

and

$$F(i)=\left[\log_2\left(1-\Gamma\log\left(1-x^{\frac{1}{i}}\right)\right)\right]$$

for arbitrarily small epsilon. □

The throughput is defined as

$$\eta_\infty=\eta_\infty(x,R,\Gamma)=\frac{R}{1+\lfloor\frac{\log x}{\log a}\rfloor-\frac{1}{2}\delta\left(\lfloor\frac{\log x}{\log a}\rfloor-\frac{\log x}{\log a}\right)}$$

and recall that $a=\left(1-e^{-\frac{2^R-1}{\Gamma}}\right)$. Consider the values of R such that

$$k<\frac{\log x}{\log a}<k+1$$

After some algebra we find that R must satisfy the following inequality

$$F(k)<R<F(k+1)$$

where

$$F(k)=\left[\log_2\left(1-\Gamma\log\left(1-x^{\frac{1}{k}}\right)\right)\right]$$

When $F(k)<R<F(k+1)$ the average delay is a constant equal to k and $\eta_\infty=\frac{R}{1+k}$. Moreover,

$$R(k)\stackrel{\Delta}{=}\sup_{R\in(F(k),F(k+1))}\eta_\infty=F(k+1)-\epsilon$$

with ϵ arbitrarily small. Define now $b[i]\stackrel{\Delta}{=}\frac{R(i)}{1+i}$, it follows directly form computation that

$$\lim_{k\to\infty}b[k]=0\quad\text{and}\quad\lim_{k\to 0}b[k]=\log_2[1-\Gamma\log(1-x)]-\epsilon$$

This implies that the $\mathrm{argsup}_k b[k]<\infty$.

Bibliography

[1] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: turbo-codes," *IEEE Trans. on Communications*, Vol. 44, pp. 1261–1271, October 1996.

[2] R. G. Gallager, *Low-Density Parity-Check Codes*, PhD thesis, Cambridge, MA: MIT Press, 1963.

[3] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, New Jersey: Prentice Hall, 1971.

[4] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. on Information Theory*, Vol. 25, pp. 306–311, May 1979.

[5] H. Sato, "An outer bound to the capacity region of broadcast channels (Corresp.)," *IEEE Trans. on Information Theory*, Vol. 24, pp. 374–377, May 1978.

[6] D. J. Costello and S. Lin, *Error and Control Coding: Fundamentals and Applications*, Prentice Hall, 1983.

[7] D. Costello, J. Hagenauer, and H. Imai, "Applications of error-control coding," *IEEE Trans. on Information Theory*, Vol. 44, pp. 2531–2560, October 1998.

[8] P.S. Sindhu, "Retransmission error control with memory," *IEEE Trans. on Communications*, Vol. 25, pp. 473–479, May 1977.

[9] G. Benelli, " An ARQ scheme with memory and soft detection," *IEEE Trans. on Communications*, Vol. 33, pp. 285–288, March 1985.

[10] E. Yli-Juuti, S.S. Chakraborty, and M. Liinaharja, "An adaptive ARQ scheme with packet combining," *IEEE Communication Letters*, Vol. 2, pp. 200–202, July 1998.

[11] D. Chase, "Code combining, a maximum likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. on Communications*, pp. 385–393, May 1985.

[12] B.A. Harvey and S.B. Wicker, " Packet combining system based on the Viterbi decoder," *IEEE Trans. on Communications*, Vol. 42, pp. 1544–1557, Feb./Mar./Apr. 1994.

[13] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. on Communications*, Vol. 36, pp. 389–400, April 1988.

[14] S. Kallel, "Analysis of a type-II hybrid ARQ scheme with code combining," *IEEE Trans. on Communications*, Vol. 38, pp. 1133–1137, August 1990.

[15] M. Zorzi and R.R. Rao, "Throughput performance of ARQ selective-repeat with time diversity in Markov channel with unreliable feedback," *Wireless Network*, Vol. 2, pp. 63–75, 1996 March.

[16] M. Zorzi and R.R. Rao, "Performance of ARQ go-back-N protocol in Markov Channels with Unreliable Feedback," *Mobile Networks and Applic.*, Vol. 2, pp. 183–193, 1997.

[17] T.F. Wong, Q. Zhang, and J.S. Lehnert, "Performance of type-II hybrid ARQ protocol in slotted DS-SSMA packet radio systems," *IEEE Trans. on Communications*, Vol. 47, pp. 281–290, February 1999.

[18] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. on Information Theory*, Vol. 47, pp. 1971–1988, July 2001.

[19] S. Kallel, "Complementary punctured convolutional (CPC) codes and their cpplications," *IEEE Trans. on Communications*, Vol. 43, pp. 2005–2009, June 1995.

[20] C.-F. Leanderson and G. Caire, "The performance of incremental redundancy schemes based on convolutional codes in the block-fading Gaussian collision channel," *IEEE Trans. on Wireless Communication*, Vol. 3, pp. 843–854, May 2004.

[21] K.R. Narayanan and G.L. Stuber, "A novel ARQ technique using the Turbo coding principle," *IEEE Communication Letters*, Vol. 1, pp. 49–51, 1997 March.

[22] T. J. Richardson and R. L Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. on Information Theory*, Vol. 47, pp. 599–618, February 2001.

[23] T. J. Richardson, M. A Shokrollahi, and R. L Urbanke, "Design of capacity-approaching irregular low-density parity-check Codes," *IEEE Trans. on Information Theory*, Vol. 47, pp. 619–637, February 2001.

[24] S. Y. Chung, T. J. Richardson, and R. L Urbanke, "Analysis of sum-product decoding of low-density parity-check codes using a Gaussian approximation," *IEEE Trans. on Information Theory*, Vol. 47, pp. 657–670, February 2001.

[25] S. Y. Chung, *On the Construction of Some Capacity-Approaching Coding Scheme*, PhD thesis, Massachussetts Institute of Technology, September 2000.

[26] T. J. Richardson and R. L Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. on Information Theory*, Vol. 47, pp. 599–618, February 2001.

[27] S. T. Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. on Communications*, Vol. 49, pp. 1727–1737, October 2001.

[28] S. ten Brink, G. Kramer, and A. Ashikhmin, "Design of low-density parity-check codes for modulation and detection," *IEEE Trans. on Communications*, Vol. 52, pp. 670–678, April 2004.

[29] C. Di, R. Urbanke, and T. Richardson, "Weight distribution: how deviant can you be?," in *International Symposium on Information Theory*, Washington, DC,June 24-29, 2001.

[30] D. Burshtein and G. Miller, "Expander graph arguments for message-passing algorithms," *IEEE Trans. on Information Theory*, Vol. 47, pp. 782–790, February 2001.

[31] G. A. Margulis, "Explicit constructions of graphs without short cycles and low density codes," *Combinatorica*, Vol. 2, No. 1, pp. 71–78, 1982.

[32] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Trans. on Information Theory*, Vol. 42, pp. 1710–1722, November 1996.

[33] G. Zemor, "On expander codes," *IEEE Trans. on Information Theory*, Vol. 47, pp. 835–837, February 2001.

[34] T.M. Cover, "Comments on broadcast channel," *IEEE Trans. on Information Theory*, Vol. 44, pp. 2524–2530, October 1998.

[35] P.P. Bergmans and T.M. Cover, "Cooperative broadcasting," *IEEE Trans. on Information Theory*, Vol. 20, pp. 317–324, May 1974.

[36] P.P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. on Information Theory*, Vol. 19, pp. 197–207, March 1973.

[37] R. G. Gallager, "Capacity and coding for degraded broadcast channel," *Probl. Pered. Inform. Transm*, Vol. 10, pp. 3–14, July-September 1974.

[38] S. R. Chandran and S. Lin, "Selective-repeat-ARQ schemes for broadcast links," *IEEE Trans. on Communications*, Vol. 40, pp. 12–19, January 1992.

[39] J. He, K.R. Subramanian, L. Zhang, and K.K. Ma, "Analysis of a full-memory multidestination ARQ protocol over broadcast links," *IEEE Trans. on Communications*, Vol. 49, pp. 1889–1894, November 2001.

[40] H. Djandji, "An efficient hybrid ARQ protocol for point-to-multipoint communication and its throughput performance ," *IEEE Trans. on Vehicular Technology*, Vol. 48, pp. 1688–1698, September 1999.

[41] R.H. Deng, "Hybrid ARQ schemes for point-to-multipoint communication over nonstationary broadcast channels ," *IEEE Trans. on Communications*, Vol. 41, pp. 1379–1387, September 1993.

[42] P.K. Gopala and H. El Gamal, "On the throughput-delay tradeoff in cellular multicast," submitted to *Trans. on Communications*.

[43] D.N.C. Tse, "Optimal power allocation over parallel Gaussian channels," in *Proc. International Symposium on Information Theory*, June 1997.

[44] R. Knopp and P. A. Humblet, "Maximizing diversity on block-fading channels, " in *Proc. IEEE International Conference on Communications*, 1997.

[45] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. on Information Theory*, Vol. 49, pp. 1147–1158, May 2003.

[46] T. Cover and J. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

[47] C. E. Shannon, "A mathematical theory of communication," *Bell. System Tech. J.*, vol 27, pp. 379-423 and 623-656, 1948.

[48] N. Favardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. on Information Theory*, Vol. 35, pp. 827–838, November 1987.

[49] N. Phamdo, N. Favardin, and T Moriya, "A unified approach to tree-structured and multistage vector quantisation for noisy channels," *IEEE Trans. on Information Theory*, Vol. 39, pp. 835–850, May 1993.

[50] B. Rimoldi, "Successive refinement of information: characterization of the achievable rates," *IEEE Trans. on Information Theory*, Vol. 40, pp. 253–259, January 1994.

[51] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. on Information Theory*, Vol. 37, pp. 269–275, March 1991.

[52] L. Lastras and T. Berger, "All sources are nearly successively refinable," *IEEE Trans. on Information Theory*, Vol. 47, pp. 918–926, March 2001.

[53] S. Shamai, S. Verdu, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. on Information Theory*, Vol. 44, pp. 564–579, March 1998.

[54] Z. Reznic, M. Feder, and R. Zamir, "Distortion bounds for broadcasting with bandwidth expansion," *40th Annual Allerton Conference on Communication Control and Computing*, Monticello, Illinois, October 2002.

[55] Z. Reznic, M. Feder, and R. Zamir, "Joint source-channel coding of a Gaussian mixture source over Gaussian broadcast channel," *submitted to IEEE Trans. on Information Theory*, 2002.

[56] U. Mittal and N. Phamdo, "Hybrid digital-analog (HDA) joint source-channel codes for broadcasting and robust communications," *IEEE Trans. on Information Theory*, Vol. 48, pp. 1082–1102, May 2002.

[57] A.N. Kim and T.A. Ramstad, "Practical low bit rate predictive image coder using multi-rate processing and adaptive entropy coding," in *Proc. of International Conference on Image Processing*, Singapore, October 2004.

[58] J.M. Lervik and T.A. Ramstad, "Optimality of multiple entropy coder systems for nonstationary sources modelled by a mixture distribution," in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc.*, 1996.

[59] M.W. Marcellin, *Trellis Coded Quantization: An Efficient Technique for Data Compression*," PhD thesis, Texas A&M Univ, December 1987.

[60] G. Ungerboeck, Channel coding with multilevel/phase signal," *IEEE Trans. on Information Theory*, Vol. 28, pp. 56–67, January 1982.

[61] N. Favardin, "A study of vector quantization for noisy channel," *IEEE Trans. on Information Theory*, Vol. 36, pp. 779–809, July 1990.

[62] K. Zeger and A. Gersho, "Pseudo-gray coding," *IEEE Trans. on Communications*, Vol. 38, pp. 2174–2158, December 1990.

[63] J.W. Modestino and D. G. Daut, "Combined source-channel coding of images," *IEEE Trans. on Communications*, Vol. 27, pp. 779–809, November 1979.

[64] G. Caire, S. Shamai and S. Verdu, "Noiseless data compression with low-density parity check codes," in *Proc. DIMACS workshop*, July 2004.

[65] J. Garcia-Frias and Y. Zhao, "Compression of binary memoryless sources using punctured turbo codes," *IEEE Communication Letters*, Vol. 6, pp. 394–396, September 2002.

[66] A. Roumy, S. Guemghar, G. Caire, and S. Verdu, " Design methods for irregular repeat-accumulate codes," *submitted to IEEE Trans. on Information Theory*, October 2002.

[67] M. Zorzi and R. R. Rao, "On the use of renewal theory in the analysis of ARQ protocols," *IEEE Trans. on Communications*, Vol. 44, pp. 1077–1081, September 1996.

[68] S. Sesia and G. Caire, "Incremental redundancy schemes based on LDPCs for transmission over Gaussian block fading channels," in *IEEE Information Theory Workshop*, October 2002.

[69] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Trans. on Communications*, Vol. 52, pp. 1311–1321, August 2004.

[70] S. Sesia, G. Caire, and G. Vivier, "The throughput of LDPC-based incremental redundancy schemes with finite blocklength," in *IEEE International Symposium on Information Theory*, June-July 2003.

[71] S. Sesia, G. Caire, and G. Vivier, "Reducing the average complexity of LDPC decoding," in *3rd International Symposium on Turbo Codes and Related Topics*, September 2003.

[72] S. Sesia, G. Caire, and G. Vivier, "On the scalability of HARQ systems in wireless multicast," in *IEEE International Symposium on Information Theory*, June-July 2004.

[73] S. Sesia, G. Caire, and G. Vivier, "Broadcasting of a common source: information theory results and system challenges," in *Winter School on Coding and Information Theory*, February 2003.

[74] S. Sesia and G. Caire, "Optimized joint source-channel techniques for compound channel," in *IEEE International Symposium on Information Theory*, Adelaide, Australia, September 2005.

[75] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. on Information Theory*, Vol. 43, pp. 38–47, January 1997.

[76] S. Sesia and G. Caire, "Multistage trellis quantization and its applications," in *42th Annual Allerton Conference on Communication, Control and Computing*, September-October 2004.

[77] A.N. Kim, S. Sesia, T. Ramstad, and G. Caire, "Combined unequal error protection and compression using Turbo codes for resilient image transmission," in *Proc. of International Conference on Image Processing*,Genova, Italy, September, 2005.

[78] S. Sesia, G. Caire, and G. Vivier, "Superposition vs progressive vs hybrid approaches for lossy transmission over BF-channels and Code Construction," *To be submitted to IEEE Trans. on Wireless*, 2005.

[79] S. Sesia, A. N. Kim, G. Caire, T. Ramstad, "Combined unequal error protection and compression using Turbo codes for resilient image transmission", *to be submitted to IEEE Trans. On Communications 2005*.

[80] E. Biglieri, J. Proakis, and S. Shamai, " Fading channels: information-theoretic and communications aspects," *IEEE Trans. on Information Theory*, Vol. 44, pp. 2619–2692, October 2001.

[81] L. H. Ozarow, S. Shamai, and A.D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. on Vehicular Technology*, Vol. 43, pp. 359–378, May 1994.

[82] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*, Cell&Sys, 1992.

[83] H. Holma and A. Toskala, *WCDMA for UMTS, 2nd Edition*, John Wiley and Sons, 2002.

[84] P. Billingsley, *Probability and Measure*, Wiley Interscience, 1995.

[85] M. G. Luby and M. Mitzenmacher, "Efficient erasure correcting codes," *IEEE Trans. on Information Theory*, Vol. 47, pp. 569–584, February 2001.

[86] A. Ashikhmin, G. Kramer, and S. T. Brink, "Extrinsic information transfer functions: A model and two properties," in *Proc. IEEE Int. Symposium Information Theory*, Lausanne, Switzerland, p. 115, July 2002.

[87] R.L. Urbanke et al., "http://lthcwww.epfl.ch/research/ldpcopt/"

[88] M. G Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Improved low density codes and improved designs using irregular graphs," *IEEE Trans. on Information Theory*, Vol. 47, No. 2, pp. 585–598, 2001.

[89] J. Rosenthal and P. O. Vontobel, "Costructions of regular and irregular LDPC codes using Ramanujan graphs and ideas from Margulis," *in Proc. of the 38th Allerton Conference on Communication, Control and Computing*, pp. 248–257, 2000.

[90] J. Lafferty and D. Rockmore, "Codes and iterative decoding on algebraic expander graphs," *International Symposium on Information Theory and Its Application*, Honolulu, Hawaii, U.S.A, November 2000.

[91] A. Banerjee, D. J. Costello, and T. E. Fuja, "Comparison of different retransmission strategies for bandwidth efficient hybrid ARQ scheme using Turbo codes," *IEEE Trans. on Information Theory*, pp. 1–10, July 2001.

[92] R. Y. Shao, S. Lin, and M. P. C. Fossorier, "Two simple stopping criteria for Turbo decoding," *IEEE Trans. on Communications*, Vol. 47, pp. 1117–1120, August 1999.

[93] M. Hagenauer, "Iterative decoding of binary block length and convolutional codes," *IEEE Trans. on Information Theory*, pp. 1–10, July 2001.

[94] M. Moher, "Decoding via cross-entropy minimization," *IEEE Trans. on Information Theory*, pp. 1–10, July 2001.

[95] A. Shiozaki, "Adaptive type-ii hybrid broadcast ARQ system ," *IEEE Trans. on Communications*, Vol. 44, pp. 420–422, April 1996.

[96] J.J. Nelson, *Probability, Stochastic Processes and Queueing Theory*, Springer-Verlag, 1998.

[97] W. Feller, *An Introduction of Probability Theory and Its Applications*, New York: Wiley, 1968.

[98] C. E. Shannon, "Communication in the Presence of Noise," *in Proc. IRE.*, Vol. 37, pp. 10–21, January 1949.

[99] V. Chande and N. Favardin, "Progressive transmission of images over memoryless noisy channels," *IEEE Journal on Selected Areas in Communications*, Vol. 18, pp. 850–860, June 2000.

[100] N. Sarshar and X. Wu, "Broadcasting with fidelity criteria," in *Information Theory Workshop*, October 24-29 2004.

[101] S. Shamai, "A broadcast strategy for the Gaussian slowly fading channel," in *Proc. IEEE Int. Symposium Information Theory*, Ulm, Germany, June 1997.

[102] M. Skoglund, N. Phamdo, and F. Alajaji, "VQ-based hybrid digital-analog coding joint source-channel coding" in *Proc. IEEE International Symposium on Information Theory*, Sorrento, Italy, June 2000.

[103] M. Skoglund, N. Phamdo, and F. Alajaji, "Hybrid digital-analog coding for bandwidth compression/expansion using VQ and Turbo codes," in *Proc. IEEE International Symposium on Information Theory*, Washington, D.C., USA, June 2001.

[104] M. Skoglund, N. Phamdo, and F. Alajaji, "Design and performance of VQ-based hybrid digital-analog joint source-channel codes," *IEEE Transaction on Information Theory*, vol. 48, no.3, pp. 708-720, March 2002.

[105] K. Sayood and J.C. Borkenhagen, "Use of residual redundancy in the design of joint source/channel coders," *IEEE Transactions on Communications*, Vol. 39, pp. 394–396, June 1991.

[106] W. Xu.; J. Hagenauer and J. Hollmann, "Joint source-channel decoding using the residual redundancy in compressed images," in Proceedings of the *IEEE International Conference on Communications*, 1996.

[107] G. Caire, S. Shamai and S. Verdu, "Universal data compression with LDPC codes," in *Proceedings of 3rd International Symposium on Turbo-Codes*, 2003.

[108] J.Garcia-Frias and Y. Zhao, "Data compression of unknown single and correlated binary sources using punctured Turbo codes," in Proceedings of *the 39th Annual Allerton Conference on Communication, Control, and Computing*, 2001.

[109] E. Tuncel and K. Rose, Additive successive refinement," *IEEE Trans. on Information Theory*, Vol. 49, pp. 1983–1991, August 2003.

[110] D.S. Taubman and M.W. Marcellin, "JPEG2000: standard for interactive imaging," in *Proceedings of the IEEE*, pp. 1336–1357, August 2002.

[111] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizer," *IEEE Trans. on Information Theory*, Vol. 38, pp. 428–436, March 1992.

[112] G. D. Jr. Forney, "Geometrically uniform codes," *IEEE Trans. on Information Theory*, Vol. 37, pp. 1241–1260, September 1991.

[113] T.R. Fischer and M. Wang, "Entropy-constrained trellis-coded quantization," *IEEE Trans. on Information Theory*, Vol. 38, pp. 415–426, March 1992.

[114] D. Baron and Y. Bresler, "An $O(N)$ semipredictive universal encoder via the BWT," *IEEE Trans. on Information Theory*, Vol. 50, pp. 928–937, May 2004.

[115] L.R. Bahl, J. Cocke, F. Jelinek, and J Raviv, " Optimal decoding of linear codes for minimizing symbol error rate," in *IEEE Trans. on Information Theory*, Vol. 20, pp. 284–287, March 1974.

[116] G. D. Jr. Forney, "Convolutional codes I: algebraic structure," *IEEE Trans. on Information Theory*, Vol. 16, pp. 720–738, November 1970.

[117] A. Gershpo and R.M. Gray, *Vector Quantization and Signal Compression*, Boston/Dordrecht/London: Kluwer Academic Publishers, 1991.

[118] S. ten Brink and G. Kramer, "Design of repeat-accumulate codes for iterative detection and decoding," *IEEE Transactions on Signal Processing*, Vol. 51, pp. 2762–2772, November 2003.

[119] Zhu, G.-C. and Alajaji, F. "Design of turbo codes for nonequiprobable memoryless sources," *39th Annual Allerton Conference on Communication, Control, and Computing*, 2001.

[120] G.-C. Zhu and F. Alajaji, "Turbo codes for non uniform memoryless sources over noisy channel," *IEEE Communication Letters*, Vol. 6, pp. 64–66, February 2002.

[121] W. Hoeffding, *Probability Inequalities for Sums of Boundend Random Variables*, American Statistical Association Journal, Vol. 58, pp. 13–30, 1963.