When All Information Is Not Created Equal

by

Shashibhushan Prataprao Borade

S.M. Massachusetts Institute of Technology (2004)
B.Tech. Indian Institute of Technology, Bombay (2002)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author .....
Department of Electrical Engineering and Computer Science
June 24, 2008

Certified by..

Lizhong Zheng
Associate Professor
Thesis Supervisor

Accepted by.....

Terry P. Orlando
Chairman, Department Committee on Graduate Students

When All Information Is Not Created Equal

by

Shashibhushan Prataprao Borade

Submitted to the Department of Electrical Engineering and Computer Science
on June 24, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Following Shannon's landmark paper, the classical theoretical framework for communication is based on a simplifying assumption that all information is equally important, thus aiming to provide a uniform protection to all information. However, this homogeneous view of information is not suitable for a variety of modern-day communication scenarios such as wireless and sensor networks, video transmission, interactive systems, and control applications. For example, an emergency alarm from a sensor network needs more protection than other transmitted information. Similarly, the coarse resolution of an image needs better protection than its finer details. For such heterogeneous information, if providing a uniformly high protection level to all parts of the information is infeasible, it is desirable to provide different protection levels based on the importance of those parts.

The main objective of this thesis is to extend classical information theory to address this heterogeneous nature of information. Many theoretical tools needed for this are fundamentally different from the conventional homogeneous setting. One key issue is that *bits* are no more a sufficient measure of information. We develop a general framework for understanding the fundamental limits of transmitting such information, calculate such fundamental limits, and provide optimal architectures for achieving these limits. Our analysis shows that even without sacrificing the data-rate from channel capacity, some crucial parts of information can be protected with exponential reliability. This research would challenge the notion that a set of homogenous bits should necessarily be viewed as a universal interface to the physical layer; this potentially impacts the design of network architectures.

This thesis also develops two novel approaches for simplifying such difficult problems in information theory. Our formulations are based on ideas from graphical models and Euclidean geometry and provide canonical examples for network information theory. They provide fresh insights into previously intractable problems as well as generalize previous related results.

Thesis Supervisor: Lizhong Zheng
Title: Associate Professor

Acknowledgments

"Some people come into our lives and quickly go.
Some people move our souls to dance.
Some people make the sky more beautiful to gaze upon.
They stay in our lives for awhile,
leave footprints in our hearts,
and we are never, ever the same."

–Anon

I was privileged to meet so many such people during my six years at MIT. I am indebted forever to my advisor Prof. Lizhong Zheng for his continuous support, ideas, and confidence in me. He gave me complete freedom to pursue all my interests and also provided so many exciting directions to explore. Never there was any pressure to produce more results or meet any artificial deadlines. Like once I asked Lizhong in a February whether it was fast enough if I completed a journal version by June and he replied how about December! Behind the freedom he gave, there is his strong belief that the best research is done when you are having fun. In addition to his technical powers, what helped me a lot was his passionate approach to research, his intrepidity in attacking important hard problems, his enthusiasm for exploring new areas, and his emphasis on bold imagination and creativity. I will always remember my second-last semester when we were proving essentially a theorem a day.

He was a friend when I wanted to share something personal, a mentor when I needed help in career matters, and a math wizard when some proof was stuck for long. Lizhong's openness to my decisions and confidence in my abilities (even when I seemed to have lost it) made me reach much higher goals than I could have imagined. His infectious cheerfulness, zen-like attitude of dealing with challenges, and patience with random door-knocks would dissolve the worst of the stress. Proud to be his first student, I hope to keep in touch with this amazing mentor and friend.

I was also extremely fortunate throughout my graduate life to be under the caring wings of Prof. Bob Gallager. I am grateful for his genuine concern and help in my research and career. His crystal clear thinking, insightful research ideas, and very careful comments on my write-ups have been crucial for this thesis. In our weekly meetings, he was always encouraging and patient in understanding my often vague and random ramblings. An incredible thing about Bob is his ability to take a step back once a result has been proved and rethink what it means for the problem. For example, his thinking with such a feedback loop was the key behind one of my favorite results in this thesis (the sand-and-boulders theorem in Chapter 3).

He often reminded me that engineering insights should not be lost amidst the mathematical details. Also, his advice of simultaneously working on a variety of problems ensured that research never became boring. Bob's guidelines for good research have deeply influenced my choice of problems and the way I try thinking about them—strip away the unnecessary details. I will always remember "Any Tom Dick can answer a question, but the real genius lies in asking the right question". His philosophy has not only been tremendously helpful in my graduate research, but also during my summer stints in industry, and will keep helping me in the future. I will always cherish the kindness, ethics, and high standards Bob exemplifies.

I sincerely thank Prof. Dave Forney for his active interest in my progress and his encouraging advice on my career plans. His careful reading and suggestions on my thesis and papers have greatly improved their final versions. Many chapters of this thesis are influenced by his technical leadership, especially in simplifying error exponents and using them in wider engineering situations. One of the other committee members mirrors my view more poetically: "I used to think the best kind of research is fundamental research, but after reading Forney's papers, I realized the best research is romantic research." I will strive for a similar balance of engineering relevance and theoretical elegance of this personal hero. I also express my gratitude to Prof. Sanjoy Mitter for enriching us with his wide intellectual interests, deep insights, and passion for unifying various fields. His support and guidance in my research and career has been a great boost for both.

I thank Dr. Mitchell Trott his kind support, counsel, and a very enjoyable summer at Hewlett Packard Labs. He exposed me to many engineering issues which were a strong influence in defining the focus of this thesis. It is so much fun to experience Mitch's super-fast mind and razor-sharp wit. I am grateful to Prof. Emre Telatar for his never-ending help and hosting me for another wonderful summer in Switzerland. Emre's unlimited generosity, insightful thinking, and clarity of communication will always be an inspiration.

I am thankful to Prof. Greg Wornell for a his care, encouragement and stimulating advice—from literally my day one at MIT. I especially thank Greg (and Lizhong) for involving me in the launch of the Wireless Communication course. I also thank Prof. Devavrat Shah, who was always available for technical help and frank advice on many other things. The breadth and depth of Devavrat's knowledge continue to inspire me. Encouragement and counsel from Professors Muriel Medard, Rüdiger Urbanke, and Moe Win is also gladly acknowledged.

I owe a lot to Dr. Sreekar Bhaviripudi, Dr. Kranthi Gade, Swapnil Pawar, Dr. Vinod Prabhakaran, Dr. Vishwambhar Rathi and Sumit Singh for their friendship and support over so many years. I thank my longtime roommate and friend Dr. Amit Deshpande for his energy and interest in so many things. Special thanks to Dr. Sujay Sanghavi and Mrs. Laxmi for their friendship, contagious *joie de vivre*, wit, and infinite chat. At MIT, I was also lucky to be surrounded by Dr. Vijay Divi (and his crazy hairdos), Dr. Anand Srinivas (and his never-ending gossip), Dr. Jun Sun and Sujie Chang (and their exquisite tastes), Parikshit Shah, Dr. Pavithra and Ajay Deshpande, Dr. Sachin Katti, Prof. Rohit Karnik, Dr. Siddharth Ray, Urs Niesen and Dr. Ashish Khisti. Company and affection of these people ensured my sanity. I thank all members of LIDS, my academic home within MIT, for their help and camaraderie. I especially thank Barış Nakiboğlu for his tremendous mathematical powers, unshakable insistence on rigor, and many lively discussions. Many critical proofs in Chapter 3 and 4 were simply impossible without Barış.

Expressing my love to my parents, Prof. Pratap Borade and Mrs. Shashikala Borade, is impossible with words. My every success, big or small, is owed to their love,

support, and sacrifices. They provided me all the freedom to paint my own canvas and isolated me from any possible worries; their strength never stops to amaze me. I particularly thank Aai (and Pappa) for her Boston visits over two cold fall semesters, which provided me a lot of good food and company. I am truly blessed to have a loving sister Dr. Mrunmayee Vikhe. Her affection, trust, and courage have always been a source of power and sanity. The farther I go geographically, the closer we come. My super-cute (and equally dangerous) niece Tanishka keeps invigorating us with her surprising creativity and energy. Many thanks to my brother-in-law Dr. Vikram Vikhe, my mamalog, and my cousins (especially Chitrao and Mithil) for there unwavering support. Special thanks to my uncle-aunt, Mr. Kamalkishor and Mrs. Latika Kadam, for their very enjoyable visits.

Financial support from Hewlett Packard Graduate Fellowship (2005-2007) and MIT Presidential Fellowship (2002-2003) is thankfully acknowledged.

To the loving memory of

my grandparents,

for their love, support, and values.

Contents

List of Figures

Chapter 1

Introduction

"We hold these truths to be self-evident, that all men are created equal..."

— Declaration of American Independence

This thesis addresses a problem of interest to communication scenarios ranging from dynamic wireless networks to audio/video broadcasting to control systems: how to communicate efficiently when some pieces of information are more important than others and need better protection. We obtain the fundamental limits of protection for communicating such heterogeneous information. No matter how smart the transmitter and receiver are, these limits cannot be broken. We also provide the optimal architectures for achieving these limits.

Classical information theory on the other hand, following Shannon's seminal paper [1], has always assumed that all information is equally important. In the limit of infinitely long codes, this homogeneity assumption is extremely powerful and gives rise to the universal interface of *bits*, which is often viewed as Shannon's most significant contribution. This interface is optimal for sending any information source over any channel—provided the codelength is sufficiently large. This means that the source encoder converts the incoming information to a set of bits. For example, an MP3 encoder converts audio waveforms into bits and a JPEG encoder converts pictures to

bits. The channel encoder can simply treat these bits as independent coin-flips. It need not care what they mean or where these bits are coming from, they could be an MP3 file or an JPEG image or something else. Such separation of source coding and channel coding does not reduce efficiency in the infinite codelength limit.

However, when resources of delay and/or bandwidth are limited and codelength cannot approach infinity, we need to break away from this homogeneous view of information which oversimplifies the nature of information. With limited resources, protecting everything equally well is either inefficient or infeasible—one needs to prioritize. Now the heterogeneous nature of information should be leveraged for designing better communication systems. In other words, we should take advantage of the fact that not all information is created equal!

In the classical homogeneous view, any particular message being mistaken as any other is viewed to be equally costly. With such uniformity assumptions, the reliability of a communication scheme is measured by a single performance metric: the probability of error over all possible messages (either average or the worst case). In information theory literature, a communication scheme is said to be *reliable* if this error probability can be made vanishingly small. However, for communication scenarios ranging from wireless networks to video transmission to control applications, the performance metric is more appropriately a combination of different kinds of error probabilities. For example,

- Consider transmission of a multiple resolution source code like a JPEG image or a MPEG video. The coarse resolution needs a smaller error probability than the finer resolution, even though both resolutions are contained in the same file. This better protection ensures that at least the crude reconstruction is recovered after bad noise realizations.
- In a wireless network, protocol information like power control, channel state, and frequency allocation is often a precursor to delivering the payload data. Hence error probability for the protocol information should be smaller than the payload data. Thus even though the final objective is delivering the payload data, the physical layer should provide a better protection to the protocol information.

Similarly for the Internet, packet headers are more important for delivering the packet. Internet's TCP protocol falls apart without these headers[1]. Hence they should be protected better to ensure that the actual data gets through. In general, any layered architecture like the Internet, highlights the heterogeneous nature of information. A richer interface to the physical layer is needed which addresses this heterogeneous nature.

- Controlling unstable plants over noisy communication links [33] and compressing unstable sources [34] are further examples where different parts of information need different reliability.

For such situations of special bits, unequal error protection (UEP) is a natural generalization to the conventional content-blind information processing. However, when our goal is providing better protection to special parts of information, these parts of information need not be only *bits*. At finite codelengths, bits no longer suffice as a universal measure of information. Perhaps this was one of the major reasons why essentially no fundamental limits for UEP were previously known. A general formulation in this situation requires some additional notions for measuring information. Instead of some bits being special, some *messages* could be special.

To clarify this concept, consider a channel encoder which takes the input of k information bits, $\mathbf{b} = [b_1, b_2, \ldots b_k]$. This k-bit situation is equivalent to a random variable M taking values from the set $\{1, 2, 3, \ldots, 2^k\}$. Each element in this message set corresponds to a particular value of the bit-sequence $\mathbf{b}$. This set of possible values of M are referred to as *messages*. For transmission, a message is encoded into its corresponding codeword and sent over the channel. A decoding error is defined as the event that the receiver decodes to a message other than the transmitted message. In most information theory texts, when a decoding error occurs, the entire bit sequence $\mathbf{b}$ is rejected. That is, errors in decoding the message and in decoding the information bits are treated similarly.

Existing literature on unequal error protection only deals with special bits; infor-

[1]The packet header is analogous to the address label on a postcard. If the address label is wrong, the postcard is useless even if all other contents are correct.

mation bits are partitioned into subsets according to priority and decoding errors for different subsets of bits are viewed differently. For example, we may want to provide a better protection to the first bit b_1 by ensuring that errors in decoding b_1 are less probable than the other bits in $\mathbf{b}$. For one half of the 2^k messages, b_1 equals 0 and for the other half it equals 1. Better protection of b_1 demands that when a message from one half is sent, probability of decoding to a message in the other half should be minimized. This essentially means that codewords for the two halves should look like two distant clusters, which ensures that jumping to the wrong cluster is unlikely. We define such problems as "bit-wise UEP". Previous examples of packet headers, multiple resolution codes, etc. belong to this category of UEP.

However, in some situations, instead of *bits* one might want to provide a better protection to a subset of *messages*. For example, consider embedding a special message in the k-bit vector $\mathbf{b}$: out of the 2^k possible messages, say the first message $M = 1$ is special and requires smaller error probability. For concreteness, let $M = 1$ correspond to $\mathbf{b} = \mathbf{0}$ (the all zero bit-sequence). Note that the error event for this special message is not caused by error in any particular bit; instead it corresponds to the decoded bit-sequence being different from the all-zero sequence. To protect such special message, essentially all other codewords should be far away from the special codeword, which ensures that jumping towards other codewords is unlikely from the special codeword. This intuition already suggests that protecting special messages is quite different from protecting special bits (which required distant codeword clusters). Borrowing from hypothesis testing, we can further define two kinds of errors for special messages.

- We say that *missed-detection* of a message i occurs when that message is transmitted, but the receiver misses it by decoding to some other message $j \neq i$. Consider a special message indicating some system emergency which is too costly to be missed. Clearly, such special messages demand a small missed-detection probability. Note that the missed-detection probability of a message is the same as the conditional error probability after its transmission.
- We say that *false-alarm* of a message i occurs when some other message $j \neq i$

is transmitted, but the receiver decodes it to message i. Consider the reformat hard-disk command to a remote-controlled robot or satellite. False-alarm of such a message causes irreparable damage such as a formatted disk. Such irreversible instructions demand small false-alarm probability.

We denote such problems as "message-wise UEP". To illustrate the difference between bit-wise and message-wise notions, consider a 100-bit data packet for concreteness. In a bit-wise UEP example, the first bit (called b_1) of these 100 bits could be special, which corresponds to a short but important packet header. In a message-wise UEP example, the bit-sequence of 100 zeros could denote a special message, which corresponds to a system emergency. This bit-sequence is more important than all other $2^{100} - 1$ bit-sequences or messages.

In the conventional framework, every bit is as important as every other bit and every message is as important as every other message. In such a framework there is no reason to distinguish between bit-wise or message-wise error probabilities because message-wise error probability differs from bit-wise error probability by an insignificant factor. In the UEP setting however, it becomes necessary to differentiate between message-errors and bit-errors. We will see that in many situations, error probability of special bits and messages have very different behavior and it is usually much easier to protect special messages compared to special bits.

The most general formulation of UEP could be an arbitrary combination of protection demands from messages, where each message demands better protection against some specific kinds of errors. In this general definition of UEP, bit-wise UEP and message-wise UEP are simply two particular ways of specifying which kinds of errors are too costly compared to others. This thesis is restricted to these two notions of immediate practical interest, although these insights and framework will be also useful for addressing more general situations.

1.1 Previous work on UEP

There has been much work in UEP mechanisms from the coding theory perspective. Of course, the simplest approach to UEP is just allocating separate channels for different kinds of data. For example, many wireless networks allocate a separate control channel for transmitting protocol information and another data channel for sending payload data. However, a better performance is attainable by sending both on the same channel, where the protocol information is embedded within the payload data.

More systematic code designs on these lines can be found in the vast literature on UEP, which goes back to at least 1958 [15]. The first linear code was proposed in 1967 by Masnick and Wolf [16]. Rate compatible codes came up in 1988 by Hagenauer [19] and then multilevel codes came up from Calderbank and Seshadri [20]. Relatively recently, priority encoded transmission was proposed by Albanese et al [17], which is UEP for erasure channels like the internet. For high SNR wireless channels, diversity embedded codes were proposed by Diggavi and Tse [23] a few years ago. Besides the work mentioned here, numerous clever UEP code designs have been developed not only in communications literature but also video, computer systems and signal processing literature.

However, all these coding mechanisms were only focused on the notion of special bits, special messages remaining almost unaddressed. More importantly, past work was mostly about designing particular codes for specific channel models like Gaussian, erasure and so on. Optimality of these designs was essentially unknown from an information theoretic perspective. This seems to be partly due to the lack of a general framework for characterizing the optimal performance and partly due to the difficulty in proving converses, i.e., upper bounds on performance. A few exceptions addressed the issue of information theoretic limits of bit-wise UEP for the specific channel models they were considering, e.g., [21] for AWGN channels, [17] for erasure channels and [23] for high SNR wireless channels. For general channel models however, almost nothing was known.

Thus in short, for bit-wise UEP, essentially only coding theory results were known—no information theory results. For message-wise UEP, no results in coding theory were known and no results in information theory were known (with the sole exception of a result in [22]). This calls for a general understanding of the fundamental limits of unequal error protection, which hold true for a general channel model no matter how smart the coding mechanism is. This thesis develops such fundamental limits as well as optimal coding mechanisms for bit-wise UEP and message-wise UEP. These results provide practical guidelines and benchmarks for designing practical UEP codes for enhancing the overall bandwidth and/or energy efficiency.

1.2 Thesis Outline

Chapter 2 introduces classical error exponents (i.e., exponential error bounds) in various information theory problems using a geometric approach. Our discussion of these exponents [14] is based on a simple Pythagoras-like theorem in [35], which is intuitive yet rigorous. These exponents have been long known [2, 3, 4], so the value of our alternate derivations lies in their simplicity and the new geometric insights they bring. This background is useful in Chapters 3 to 4, where we use error exponents as a benchmark of protection. That is, the fundamental limits of UEP in those chapters are given in terms of the best error exponents achievable for various parts of information.

For conceptual clarity, Chapter 3 focuses on situations where the data-rate essentially equals the channel capacity. This analysis will address UEP issues for scenarios where data rate is a crucial system resource that cannot be compromised. In these situations, no positive error exponent in the conventional sense can be achieved. That is, if we aim to protect the entire information uniformly well, neither bit-wise nor message-wise error probabilities can decay exponentially fast with increasing code length. We ask the question then "can we make the error probability of a particular bit, or a particular message, decay exponentially fast with block length?"

For bit-wise UEP in this setting, we show even a single bit cannot achieve any

positive error exponent. Thus the data-rate must back off from capacity for achieving any error exponent even for a single bit. On the contrary, in message-wise UEP, positive error exponents can be achieved without giving up any data-rate. If only one message in a capacity achieving code is special and demands a (missed-detection) error exponent, its optimal value is equal to a new fundamental channel parameter called the *Red-Alert Exponent*. We then consider situations where an exponentially large subset of messages is special and each message in it demands a positive error exponent. Surprisingly, it turns out that these special messages can achieve the same exponent as if all the other (non-special) messages were absent. In other words, a capacity achieving code and an error exponent-optimal code below capacity can coexist without hurting each other. These results also shed a new light on the structure of capacity achieving codes.

These insights for the case without feedback become useful in Chapter 4, where we investigate similar problems assuming perfect causal feedback from the receiver to the transmitter. Such feedback creates some fundamental connections between bit-wise UEP and message-wise UEP. Now even for bit-wise UEP, positive error exponent can be achieved at capacity. Now a single special bit can achieve the same exponent as a single special message—the Red-Alert Exponent. When the number of special bits increases, their error exponent decays linearly from Red-Alert Exponent to 0 as their rate increases from 0 to channel capacity. A successively refinable version of this linear tradeoff is also achievable when there are multiple levels of specialty — most-special bits, second-most-special bits and so on.

Although the error exponent for a single special bit increases from 0 to the Red-Alert Exponent due to feedback, the error exponent for a single special message exponent cannot increase beyond the Red-Alert Exponent in spite of feedback. Then we address the case of exponentially many messages. Many special messages obviously cannot achieve a better exponent compared to a single special message. However, we show that at rates beyond certain threshold, the special messages can achieve the same error exponent with feedback as if all other messages were absent. We also address the best false-alarm exponent for a special message (like a disk-format command).

In Chapter 5, we treat data rates strictly below capacity. Sacrificing the data-rate from capacity should provide additional reliability. Here the question of interest is the optimal tradeoffs between the reliability of the crucial parts of information vs. that of the ordinary parts. We formulate this question in two ways. This chapter addresses the first formulation, which is based on error exponents as in the Chapters 3 and 4—what are the optimal error exponents achievable simultaneously for different parts of information? Many of the same questions in bit-wise and message-wise UEP for rates approaching capacity are revisited here. For example, we generalize the classical sphere-packing exponent for bit-wise UEP situations. We also extend the notion of Red-Alert Exponent for rates below capacity.

In Chapter 6, the case of rates below capacity is addressed from a different viewpoint—network information theory. In earlier chapters, a point-to-point channel is considered and UEP was used for extra protection for the crucial parts against large deviations of the channel noise. In other words, UEP was aimed at providing better error exponents for those crucial parts. Alternatively, UEP can be used in a broadcast network where the crucial parts should be protected against channel fading or movements of users. That is, the crucial parts should be received by all users, including those far from the base station or those with bad fading. However, the better off users (which are near the base station or experiencing little fading) should be able to decode the ordinary parts as well. Chapter 6 mainly focuses on the bit-wise notion of UEP for networks. This problem is equivalent to the network information theory problem of broadcast with degraded message sets [42].

Although this problem has been long open in general, we show how ideas from graphical models and Euclidean geometry can provide fresh insights by simplifying the problem. In particular, we first show that the simplification with Euclidean geometry becomes useful for "very noisy" situations discussed in [48],[44]. This demonstrates how the Euclidean approach can provide canonical problems for information theory which are easy to solve but still shed light on some general issues. For situations that are not very noisy, the simplification via graphical models becomes useful. Graphical models provide a framework to systematically think about broadcast situations. They

enable us to solve some new classes of broadcast networks, which generalize many previously solved networks. At the end of Chapter 6, we briefly discuss message-wise UEP over networks. This problem is shown to be connected to the problem of compound channel capacity. A few simple analogs of earlier results in error-exponent formulation are also discussed briefly.

We conclude in Chapter 7 by discussing some future directions and implications of our results for network architectures.

Chapter 2

Error Exponents in Information Theory: A Geometric Introduction

"Those unversed in geometry shall not enter."
— engraving on the entrance of Plato's Academy

This chapter introduces a geometric approach for analyzing error exponents in various information theory problems. Our approach [14] is based on a simple Pythagoras-like theorem called I-Projection and hence more intuitive than classical algebraic derivations [2]. By illuminating on the hidden geometrical structure, it also clarifies the distribution of the log likelihood for correct and incorrect codewords.

A large number of information theoretic problems can be written as optimization of Kullback-Liebler (KL) divergence. This includes most calculations of channel capacities, rate-distortion functions, and error exponents. With a handful of famous exceptions, most of these calculations can only be carried out numerically. Especially, the results of many multi-user information theory problems are given in the form of multi-dimensional optimizations, with little effort spent in finding the structure of their solutions.

As an example of such divergence minimization problems, the calculation of the error exponent is known to have two different (but equivalent) forms of solutions.

The original solutions by Gallager *et al* [2, 3] were derived using techniques similar to the Chernoff bound. Those solutions take the form of optimization over the input distribution and a scalar parameter ρ. While these results are concise and relatively easy to compute, it is hard to capture the intuition behind their derivations.

In comparison, Csiszár and Korner took a conceptually more tractable approach. They used large deviations to study decoding errors in a discrete memoryless channel (DMC) and provided error exponents in the form of KL divergence minimizations. The solution to these minimization problems has an important operational meaning—they characterize the typical error event. This approach is much more intuitive and hence widely used in a variety of information theory problems. However, instead of a scalar parameter, such solutions require a high dimensional optimization over the space of channel realizations, which is often harder to compute.

Forney analyzed this problem [5] for symmetric channels like the Binary Symmetric Channel (BSC). The optimizations over the space of channel realizations are very clean and intuitive there, because one only needs to focus on noisier versions of the original BSC. Non-symmetric realizations of the channel do not matter. Thus we only need to focus on a single dimensional family of noisier channel realizations—BSCs with increasing crossover probability.

For a general DMC however, it is not clear which single-dimensional family of 'noisier' channels should be considered for calculating error exponents. Our analysis answers this in terms of a single dimensional *exponential* family defined later. For a BSC, this exponential family is equivalent to the family of noisier BSCs. Thus an exponential family generalizes the notion of 'noisier BSCs for general DMCs.

It is worth pointing out that most of the results we derive can also be obtained from direct algebraic calculations. Thus much of the value of our approach lies in the simplicity of the solution and the new geometric insights it brings. Now we will discuss the Pythagoras-like theorem for KL divergences—the *I-Projection* theorem. It will be our main tool for deriving error exponents in this chapter. As a warmup before the channel coding exponents, we will analyze error exponents in Chernoff bound and binary hypothesis testing.

2.1 Preliminaries

We first introduce the notions of a linear family and an exponential family of probability distributions. These two families happen to be orthogonal to each other in a certain sense. Later, this orthogonality property called I-projection becomes very useful. We will use $h(n) \doteq g(n)$ to denote exponential approximation at large n.

$$h(n) \doteq g(n) \quad \Leftrightarrow \quad \lim_{n\to\infty} \frac{\log h(n)}{n} = \lim_{n\to\infty} \frac{\log g(n)}{n}$$

This means that $h(n)$ and $g(n)$ are equal up to a sub-exponential factor in n. To illustrate the $\doteq$ approximation, we now describe Stein's lemma for the binary case. Consider a random binary sequence $Z^n = (Z_1, \ldots, Z_n)$ be drawn i.i.d. with distribution $(p, 1-p)$. The probability of the empirical output distribution, i.e., output type being $(q, 1-q)$ equals $\binom{n}{nq} p^{nq}(1-p)^{n(1-q)}$. Here $\binom{n}{nq}$ is the number of sequences having nq ones and $p^{nq}(1-p)^{n(1-q)}$ is the probability of each such sequence. This probability of output type $(q, 1-q)$ can be upper and lower bounded as follows using bounds on $\binom{n}{nq}$ in [4].

$$(n+1)^{-2} e^{nH(q)} e^{-n(D(q\|p)+H(q))} \leq \binom{n}{nq} p^{nq}(1-p)^{n(1-q)} \leq e^{nH(q)} e^{-n(D(q\|p)+H(q))}$$

Here $H(q)$ denotes the entropy of distribution q and $D(q\|p)$ denotes the KL divergence between q and p. Hence the probability of observing output type q is $\doteq e^{-nD(q\|p)}$. Similar result holds for non-binary alphabets as well.

2.1.1 Orthogonality of linear and exponential families

Consider a discrete random variable[1] Z taking values from the finite set $\mathcal{Z}$. Given a function $f : \mathcal{Z} \to \mathcal{R}^K$ and a constant $\alpha \in \mathcal{R}^K$, a family $\mathcal{L}_{f,\alpha}$ of probability

[1]Unless mentioned otherwise, random variables are denoted by capital letters and their values by small letters.

distributions of Z is defined as:

$$\mathcal{L}_{f,\alpha} \equiv \left\{ q \mid \sum_{z \in \mathcal{Z}} q(z) f(z) \equiv \mathrm{E}_q [f(Z)] = \alpha \right\} \tag{2.1}$$

where $\mathrm{E}_q [f(Z)]$ denotes the expectation of $f(Z)$ under distribution q. In other words, $\mathcal{L}_{f,\alpha}$ is the set of q satisfying K linear (i.e., expectation) constraints. This family is called the *linear family*[2] of distributions in which the expected value of $f(Z)$ equals α. Since $\mathcal{L}_{f,\alpha}$ satisfies K expectation constraints as well as the probability simplex constraint, it is a $|\mathcal{Z}| - K - 1$ dimensional hyperplane contained in the probability simplex in $\mathcal{R}^{|\mathcal{Z}|}$. For $\alpha \neq \beta$, the $\mathcal{L}_{f,\alpha}$ and $\mathcal{L}_{f,\beta}$ hyperplanes are related by a parallel shift.

For the same function $f(\cdot)$ and a given probability distribution p, the corresponding *exponential family* $\mathcal{E}_{f,p}$ is defined as:

$$\mathcal{E}_{f,p} \equiv \left\{ q \mid q(z) = \frac{p(z) \cdot \exp\left(\sum_{i=1}^{K} \theta_i f_i(z)\right)}{k(\theta)}, \theta \in \mathcal{R}^K \right\} \tag{2.2}$$

where $f_i(\cdot)$ denotes the i'th component of $f(\cdot)$ and $k(\theta) = \sum_z p(z) \exp\left(\sum_{i=1}^{K} \theta_i f_i(z)\right)$ is the normalization factor to ensure $\sum_z q(z) = 1$. The parameter θ corresponding to a point (i.e., distribution) in $\mathcal{E}_{f,p}$ is called the exponential parameter of that point. Since $\theta \in \mathcal{R}^K$, note that $\mathcal{E}_{f,p}$ is a K-dimensional surface contained in the probability simplex in $\mathcal{R}^{|\mathcal{Z}|}$.

For a given p and a linear family $\mathcal{L}_{f,\alpha}$, define q^* as the projection of p on $\mathcal{L}_{f,\alpha}$.

$$q^* \equiv \arg \min_{q \in \mathcal{L}_{f,\alpha}} D(q \| p)$$

We can now state the I-projection theorem (see [35] for a proof) which is illustrated in Fig. 2-1.

[2]The term *linear* should not be interpreted in the algebraic sense of linearity. It only indicates that $\mathcal{L}_{f,\alpha}$ satisfies the linear constraints in (2.1) because $\mathrm{E}_q [f(Z)]$ is linear in q.

Theorem 1 *For any* $q \in \mathcal{L}_{f,\alpha}$,

$$D(q\|p) = D(q\|q^*) + D(q^*\|p)$$

Moreover, the projection q^* *lies on the exponential family:* $q^* \in \mathcal{E}_{f,p}$.

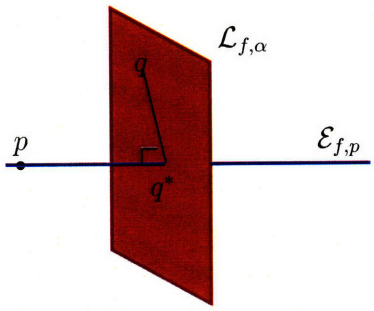


Figure 2-1: The exponential family and the corresponding linear family are 'perpendicular' to each other in terms of minimizing KL divergence. Note that although $\mathcal{E}_{f,p}$ is illustrated as a straight line, it is not straight (or flat) in the Euclidean sense. Recall that $\mathcal{L}_{f,\alpha}$ is a $|\mathcal{Z}| - K - 1$ dimensional hyperplane.

Since $\mathcal{E}_{f,p}$ is a K dimensional surface, Fig. 2-1 depicts the case of $K = 1$. This case of $K = 1$, i.e., a scalar f, is particularly useful for analyzing error exponents. In this case, I-projection reduces the divergence minimization problem into a search over a scalar parameter θ. The following example shows that the Chernoff bound, which often shows up in Gallager's derivations of the error exponents, is directly related to this geometric picture.

Example: Chernoff Bound

Let $Z^n = (Z_1, \ldots, Z_n)$ be drawn i.i.d. from distribution p. Consider the event $\{\frac{1}{n}\sum_{i=1}^n f(Z_i) \geq \alpha\}$, which can be rewritten as Z^n taking an empirical distribution q such that $E_q[f] \geq \alpha$. By Sanov's theorem [35], this probability decays exponentially in n as follows:

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) \geq \alpha\right) \doteq \exp\left(-n \min_{q:\ E_q[f]\geq\alpha} D(q\|p)\right)$$

The above optimization over q differs slightly from the I-projection theorem in Fig. 2-1. The inequality constraint means we are projecting q on the half-space to the

right of $\mathcal{L}_{f,\alpha}$ in Fig. 2-1 instead of just projecting it on the $\mathcal{L}_{f,\alpha}$. For the moment, we assume that the two optimizations are equivalent and the optimum q^* lies on $\mathcal{L}_{f,\alpha}$ in both of them. For this to happen, the average $E_q[f]$ should be monotonic along $\mathcal{E}_{f,p}$ and moreover, the divergence $D(q\|p)$ should be monotonically increasing as move away from p along $\mathcal{E}_{f,p}$. Both of these facts will soon be clear in the next sub-section. Now by the I-projection theorem, $q^* \in \mathcal{E}_{f,p}$.

$$q^*(z) = \frac{p(z) \cdot e^{\theta f(z)}}{k(\theta)}$$

where θ is chosen to satisfy $E_{q^*}[f] = \alpha$, i.e., $q^* \in \mathcal{L}_{f,\alpha}$.

$$\begin{aligned} \text{Now} \quad D(q^*\|p) &= E_{q^*}\left[\log \frac{q^*(Z)}{p(Z)}\right] \\ &= E_{q^*}\left[\theta f(Z)\right] - \log k(\theta) \\ &= \theta\alpha - \log\left(\sum_z p(z) \cdot e^{\theta f(z)}\right) \end{aligned}$$

which gives the same exponent as the familiar Chernoff bound. This shows that, although an upper bound in general, Chernoff's bound is exponentially tight.

2.1.2 Binary hypothesis testing

In the rest of this section, we will focus on a particular kind of exponential family, one connecting two given distributions. This exponential family can be thought of as a "line" connecting the two distributions. We then apply this concept to understand error exponents for binary hypothesis testing.

Consider a sequence Z^n of n discrete random variables. Under hypothesis H_0, Z^n is drawn i.i.d from distribution p_0 and under hypothesis H_1, it is drawn i.i.d. from p_1. The MAP test makes the decision by comparing the average log-likelihood ratio (LLR) $\frac{1}{n}\sum_{i=1}^{n} L(Z_i)$ (where $L(z) = \log \frac{p_1(z)}{p_0(z)}$) to a threshold α. For LLR less than α, it chooses H_0 and vice versa. If H_0 is the true hypothesis, this threshold test makes

an error when

$$\frac{1}{n}\sum_{i=1}^{n} L(Z_i) \equiv E_q[L] \geq \alpha$$

where q denotes the output type observed. Similarly, if H_1 is the true hypothesis, an error happens if $E_q[L]$ is less than α. For applying the I-projection theorem, we now define $f(z)$ to be the LLR $L(z)$. The two types of error events have probability

$$\Pr(H_0 \to H_1) \doteq \exp\left(-n \min_{q:\ E_q[L]\geq\alpha} D(q\|p_0)\right) \tag{2.3}$$

$$\Pr(H_1 \to H_0) \doteq \exp\left(-n \min_{q:\ E_q[L]\leq\alpha} D(q\|p_1)\right) \tag{2.4}$$

The I-projection theorem implies that the optimum q for each of the two optimizations above lies on $\mathcal{E}_{L,p_0}$ and $\mathcal{E}_{L,p_1}$ respectively. Since $L(\cdot)$ is the LLR function between p_1 and p_0, these two exponential families are in fact the same. With slight abuse of notation, let $\mathcal{E}_{p_0,p_1}$ denote this exponential family connecting p_0 and p_1.

$$\mathcal{E}_{p_0,p_1} = \left\{ p_t \mid p_t(z) = \frac{p_0(z)\exp[tL(z)]}{k(t)} = \frac{p_1^t(z)p_0^{1-t}(z)}{k(t)} \right\}$$

where t is the scalar exponential parameter $\theta \in \mathcal{R}$. Thus the solutions of (2.3) and (2.4) are indeed the same distribution p_{t^*}, where t^* is chosen so that[3] $E_{p_{t^*}}[L] = \alpha$. The two exponents are then given by $D(p_{t^*}\|p_0)$ and $D(p_{t^*}\|p_1)$, respectively. For convenience, we usually limit the range of t to be within $[0, 1]$. Clearly, $t = 0$ corresponds to p_0 and $t = 1$ to p_1. We can thus visualize the exponential family as a straight line segment connecting p_0 and p_1.

There are four quantities that are particularly important for this family:

- t, the exponential parameter (the same as θ before) preveious
- $\psi \equiv \log k(t) = \log \sum_z p_1^t(z)p_0^{1-t}(z)$, the log normalization factor[4]
- $\eta \equiv E_{p_t}[L]$, the average log-likelihood ratio (between p_1 and p_0) at p_t
- $D(p_t\|p_0)$, the K-L divergence corresponding to p_t.

[3]Note that we have again used the I-projection theorem for the inequality constraint.
[4]This parameter is directly related to the free energy in statistical physics [40].

The following relations between these quantities are easy to check. They are depicted in Figure 2.1.2, where ψ and η are plotted as functions of $t \in [0, 1]$.

$$\frac{\partial \psi}{\partial t} = \eta \tag{2.5}$$

$$\frac{\partial D(p_t \| p_0)}{\partial \eta} = t \tag{2.6}$$

$$t \cdot \eta = D(p_t \| p_0) + \psi \;\Rightarrow\; D(p_t \| p_0) = t\eta - \psi \tag{2.7}$$

$$\text{Similarly, we have } D(p_t \| p_1) = (t-1)\eta - \psi \tag{2.8}$$

Note from (2.6) that the exponential parameter t signifies the sensitivity of divergence w.r.t. average log-likelihood ratio η.

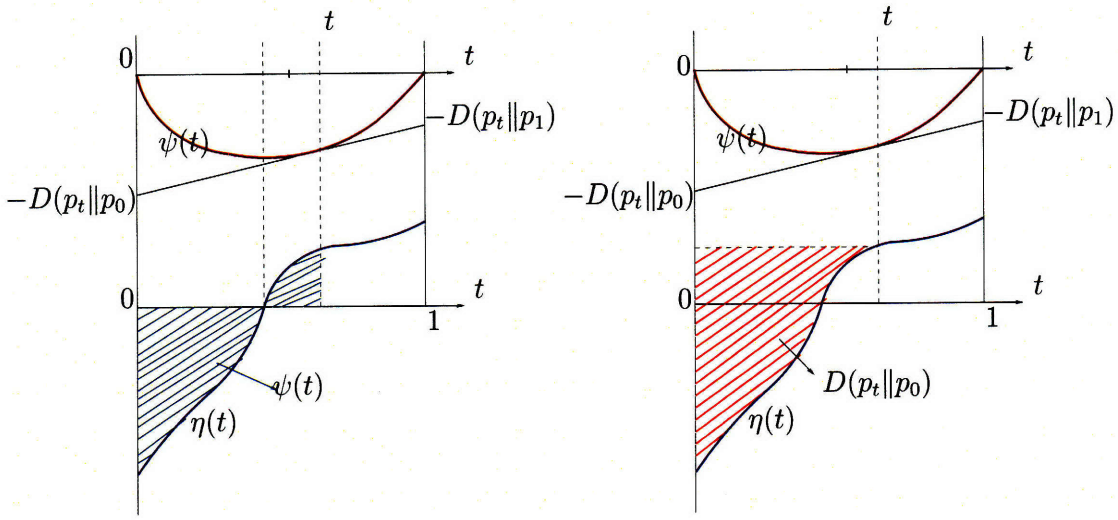


Figure 2-2: The top curves denote ψ and the bottom curves denote their derivative η as a function of t. The intersects of the tangent to $\psi(t)$, at $t = 0$ and $t = 1$, give exponents of the two errors in hypothesis testing.

In the left figure, shaded area denotes $\psi(t)$, which is the integral of η w.r.t. t. It is the net sum of the negative area below the t-axis and the positive area above the t-axis. In the right figure, shaded area (gross) denotes $D(p_t \| p_0)$, which is the integral of t w.r.t η. It is the gross sum of shaded area below the t-axis and the shaded area above the t-axis. Net addition of the shaded regions in these two figuress gives a rectangle of area $t \cdot \eta = \psi + D(p_t \| p_0)$ because the shaded area below t-axis cancels out.

The Fisher information for this one-dimensional exponential family $\mathcal{E}_{p_0,p_1}$ param-

eterized by t is:

$$g_t = E_{p_t}\left[\left(\frac{\partial}{\partial t}\log p_t(Z)\right)^2\right] = \text{variance of } L(Z)$$

One can also check that the derivative of $\eta(t)$ is equal to g_t. This shows that $\eta(t)$ is increases monotonically along the exponential family—a fact we assumed previously in discussing the Chernoff bound.

$$\frac{\partial^2\psi}{\partial t^2} = \frac{\partial\eta}{\partial t} = \frac{\partial}{\partial t}\left[\sum_z p_t(z)L(z)\right] = g_t$$

and similarly $\frac{\partial^2 D(p_t\|p_0)}{\partial\eta^2} = \frac{\partial}{\partial\eta}t = 1/g_t$.

This gives a simple relation between $D(p_t\|p_0)$ and the Fisher information as

$$D(p_t\|p_0) = \int\int \frac{1}{g_{\tilde{t}}} d\tilde{\eta}d\hat{\eta} = \int_0^t s g_s ds \tag{2.9}$$

$$\text{Similarly, we get } D(p_t\|p_1) = \int_t^1 (1-s)g_s ds \tag{2.10}$$

The double integral also demonstrates that $D(p_t\|p_0)$ increases with t, i.e., the divergence increases as we move farther on the exponential family. This also proves the fact used earlier: Replacing the inequality constraint in the I-projection theorem by an equality constraint does not change the optimum solution.

Remarks: Similar to the results in [39], (2.9) gives a relation between an information theoretic quantity and an estimation theoretic quantity. However, this result involves a double integral. We do not expect any close connection between the two results.

The simplest case of (2.9) is when g_t remains constant (say at g) along $\mathcal{E}_{p_0,p_1}$. Such constant is approximation of g_t is good when p_0 is very close to p_1, which corresponds to *very noisy* hypothesis testing problems. The double integral yields a simple quadratic relation now.

$$D(p_t\|p_0) = \frac{1}{2}gt^2, \qquad D(p_t\|p_1) = \frac{1}{2}g(1-t)^2 \tag{2.11}$$

We will revisit this relation when deriving the error exponents for very noisy channels.

2.2 Random Coding exponent for DMC

Now as a more advanced application of I-projection, let us consider the random coding exponent for a DMC. Error exponents for discrete memoryless channels with random coding were analyzed in the seminal work of Gallager [38] and later in [4], [6] and others. More recently, [7] derived these results using large deviation theory and Lagrange multipliers. We use the same random i.i.d. coding formulation[5] as in [7].

A random i.i.d. code of length n and rate R (nats/symbol) consists of $\lceil e^{nR} \rceil$ codewords of length n. For transmitting message $k \in \{1, 2 \cdots, \lceil e^{nR} \rceil\}$, the corresponding codeword denoted by $\bar{x}^n(k) \equiv (\bar{x}_1(k), \bar{x}_2(k) \cdots, \bar{x}_n(k))$ is transmitted. The symbols of every codeword $\bar{x}^n(k)$ are chosen i.i.d. with distribution P_X. The output of the channel takes values from the finite set $\mathcal{Y}$. The channel transition probability is denoted by $W_{Y|X}$, i.e., $W_{Y|X}(y|x)$ gives the probability of observing output $y \in \mathcal{Y}$ for input $x \in \mathcal{X}$.

Without loss of generality, we assume that message 1 was transmitted. Channel memorylessness implies that the probability of output sequence y^n conditioned on $\bar{x}^n(1)$ is

$$\Pr(y^n | \, \bar{x}^n(1)) = \prod_{i=1}^{n} W_{Y|X}(y_i|\bar{x}_i(1)) \tag{2.12}$$

$$\Rightarrow \Pr(y^n, \bar{x}^n(1)) = \prod_{i=1}^{n} \left(P_X(\bar{x}_i(1)) \cdot W_{Y|X}(y_i|\bar{x}_i(1))\right) = \prod_{i=1}^{n} P_{XY}(\bar{x}_i(1), y_i) \tag{2.13}$$

where P_{XY} denotes the joint XY distribution $P_X W_{Y|X}$ of this channel. The last step followed because symbols in $\bar{x}^n(1)$ are generated i.i.d. with joint distribution P_X. Hence the pair $(\bar{x}^n(1), y^n)$ of the correct codeword and the output sequence is an i.i.d. sequence generated with distribution P_{XY}. Let the corresponding marginal

[5] Although we only consider random i.i.d. codes here, error exponents for randomly chosen fixed composition codes can be also obtained on similar lines using I-projection. Similarly, error exponents for List-of-L decoding can also be obtained .

distribution of Y be denoted by P_Y

$$P_Y(y) = \sum_{x \in \mathcal{X}} P_{XY}(x, y)$$

We can also write P_{XY} as $P_Y P_{X|Y}$, where $P_{X|Y} = \frac{P_X W_{Y|X}}{P_Y}$ denotes the reverse channel from Y to X.

Since the codewords are generated independently, the output sequence is independent of any incorrect codeword $\bar{x}^n(j)$, where $j \neq 1$. Hence the pair $(\bar{x}^n(j), y^n)$ of the incorrect codeword and the output sequence is an i.i.d. sequence generated by the independent distribution $P_X P_Y$.

$$\Pr(\bar{x}^n(j), y^n) = \prod_{i=1}^{n} P_X(\bar{x}_i(j))\ P_Y(y_i)$$

2.2.1 Error exponent conditioned on the output type

We now analyze the error probability when the output sequence y^n has a given type Q_Y. This analysis will give us the error exponent $E_r(R, Q_Y)$ at rate R conditioned on the output type Q_Y. Since the number of output types is polynomial in n, the overall error exponent can be obtained later by a minimization over Q_Y.

When the received output type is Q_Y, our space of possible joint (X, Y)-types is all the distributions which ensure that the marginal distribution of Y is Q_Y. It is easy to check that this space of distributions is a linear family. With little abuse of notation, we denote this family of joint distributions by $\mathcal{L}_{Q_Y}$. Any point in this family has the form $Q_Y Q_{X|Y}$ for some reverse channel type $Q_{X|Y}$. The divergence between a point in this family and the distribution P_{XY} (related to the correct input-output pair) is equal to

$$\begin{aligned} D(Q_Y Q_{X|Y} \| P_{XY}) &= \sum_{x,y} Q_Y(y) Q_{X|Y}(x|y) \log \frac{Q_Y(y) Q_{X|Y}(x|y)}{P_{XY}(x,y)} \\ &= \sum_{x,y} Q_Y(y) Q_{X|Y}(x|y) \log \frac{Q_Y(y)}{P_Y(y)} \frac{Q_{X|Y}(x|y)}{P_{X|Y}(x|y)} \\ &= D(Q_Y \| P_Y) + D(Q_Y Q_{X|Y} \| Q_Y P_{X|Y}) \\ &\geq D(Q_Y \| P_Y) \end{aligned}$$

The last step is met with equality when $Q_{X|Y} = P_{X|Y}$. Hence, the projection of P_{XY} on this linear family is given by $Q_Y P_{X|Y}$. Thus only the marginal distribution is changed from P_Y to Q_Y but the reverse channel type is the same as $P_{X|Y}$.

On similar lines, the divergence between a point in this family and the distribution $P_X P_Y$ (corresponding to an incorrect input and output pair) equals

$$D(Q_Y Q_{X|Y} \| P_Y P_X) = D(Q_Y \| P_Y) + D(Q_Y Q_{X|Y} \| Q_Y P_X) \;\geq\; D(Q_Y \| P_Y) \quad (2.14)$$

Thus the projection of $P_Y P_X$ on $\mathcal{L}_{Q_Y}$ is given by changing the Y-marginal to Q_Y and keeping the reverse channel type the same as the (trivial reverse channel) P_X.

Now we show that Maximum-Likelihood decoder can also be thought as a Maximum-LLR decoder. This is because for given output sequence Y^n of type Q_Y, the decoded message $\hat{m}$ by the ML decoder is

$$\begin{aligned} \hat{m} &= \arg \max_{1 \leq k \leq \lceil e^{nR} \rceil} P(y^n | \bar{x}^n(k)) \\ &= \arg \max_{1 \leq k \leq \lceil e^{nR} \rceil} \sum_{i=1}^{n} \log P_{Y|X}(y_i | \bar{x}_i(k)) \quad \text{(memorylessness)} \\ &= \arg \max_{1 \leq k \leq \lceil e^{nR} \rceil} \sum_{i=1}^{n} \log \frac{P_{X|Y}(\bar{x}_i(k)|y_i)}{P_X(\bar{x}_i(k))} \qquad \text{(Baye's rule)} \end{aligned}$$

$$= \arg \max_{1 \leq k \leq \lceil e^{nR} \rceil} \sum_{i=1}^{n} \log \frac{P_{X|Y}(\bar{x}_i(k)|y_i)}{P_X(\bar{x}_i(k))} \frac{Q_Y(y_i)}{Q_Y(y_i)}$$

Dividing this by n gives

$$\hat{m}=\arg \max_{1\le k\le \lceil e^{nR}\rceil} \sum_{x,y} Q_Y(y)Q^k_{X|Y}(x|y)\log\frac{Q_Y(y)P_{X|Y}(x|y)}{Q_Y(y)P_X(x)}$$

where $Q_YQ^k_{X|Y}$ denotes the joint type of the k'th codeword and the output sequence $(\bar{x}^n(k), y^n)$. Thus ML decoding is equivalent to decoding the codeword with the largest normalized log-likelihood-ratio between the two distributions $Q_YP_{X|Y}$ and Q_YP_X. Recalling the notation in the previous section, let the joint distribution $Q_YP_{X|Y}$ be denoted by p_1 and let Q_YP_X be denoted by p_0. Their log-likelihood ratio can be denoted by $L(x,y)$.

$$L(x,y)=\log\frac{p_1(x,y)}{p_0(x,y)}=\log\frac{P_{X|Y}(x|y)}{P_X(x)}=\log\frac{W_{Y|X}(y|x)}{P_Y(y)}$$

Notice that $L(x,y)$ only depends on the channel $W_{Y|X}$ and the input distribution P_X, so it is independent of the observed channel type. The LLR decoder chooses the codeword with the largest LLR score.

$$\hat{m}=\arg\max_{1\le k\le\lceil e^{nR}\rceil} \mathrm{E}_{Q_YQ^k_{X|Y}}[L(X,Y)] \equiv \arg\max_{1\le k\le\lceil e^{nR}\rceil} S_k$$

where $Q^k_{X|Y}$ is the reverse channel type for pair $(\bar{x}^n(k), y^n)$ and $S_k \equiv \mathrm{E}_{Q_YQ^k_{X|Y}}[L(X,Y)]$ is the LLR score of the k'th codeword. Note that each score S_k is a random variable depending on the channel noise and codeword $\bar{x}^n(k)$. An error happens if and only if the score S_1 of the correct codeword is less than the score S_j of any incorrect codeword. To analyze the error exponent, we should know the distribution of these score variables. We need to find the reverse channel type $Q^k_{X|Y}$ which is the dominant cause of errors.

Consider a sub-family of $\mathcal{L}_{Q_Y}$ where the expectation of $L(X,Y)$ equals η. This is a linear family within $\mathcal{L}_{Q_Y}$ and we denote it by $\mathcal{L}_{Q_Y,\eta}$. It corresponds to the dashed line in Fig. 2-3. Applying the I-projection theorem, we see that the projection of p_0 (or p_1) on $\mathcal{L}_{Q_Y,\eta}$ is given by $Q_YP^{(t)}_{X|Y}$ for some $t\in[0,1]$, where the reverse channel

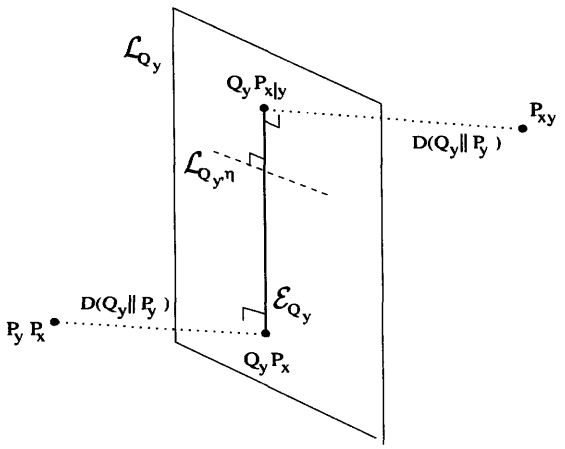


Figure 2-3: The rectangular frame represents the linear family $\mathcal{L}_{Q_Y}$, which is the set of all joint XY distributions for which the Y marginal distribution equals Q_Y. The dashed line represents $\mathcal{L}_{Q_Y,\eta}$, although note that $\mathcal{L}_{Q_Y,\eta}$ need not be single dimensional.

$P^{(t)}_{X|Y}$ for each $y \in \mathcal{Y}$ is

$$P^{(t)}_{X|Y}(x|y) = \frac{P^t_{X|Y}(x|y)P^{1-t}_X(x)}{k_y(t)} \tag{2.15}$$

$$\text{where} \quad k_y(t) = \sum_{x\in\mathcal{X}} P^t_{X|Y}(x|y)P^{1-t}_X(x) \tag{2.16}$$

The superscript t of $P^{(t)}_{X|Y}(x|y)$ is parenthesized to distinguish it from $P^t_{X|Y}(x|y)$, the t^{th} power of $P_{X|Y}(x|y)$. Conditioned on output type Q_Y, an incorrect codeword is generated i.i.d. with distribution P_X. Hence applying Sanov's's theorem and I-projection gives the following exponent for the score S_j of a wrong codeword exceeding η. It is obtained by optimizing the reverse channel type $Q^j_{X|Y}$ to the j'th codeword.

$$\lim_{n\to\infty} -\frac{\log \Pr(S_j \geq \eta |\ Q_Y)}{n} = \min_{Q^j_{X|Y}: S_j \geq \eta} D(Q_Y Q^j_{X|Y} \| p_0) \tag{2.17}$$

$$= D(Q_Y P^{(t)}_{X|Y} \| p_0) \quad \text{(by I-projection)} \tag{2.18}$$

$$= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) \tag{2.19}$$

$$\text{where } t \text{ satisfies } \eta = \mathrm{E}_{Q_Y P^{(t)}_{X|Y}}[L(X,Y)] \equiv \eta_{Q_Y}(t) \tag{2.20}$$

Thus, given output type Q_Y, the dominating manner in which a wrong codeword's

score S_j crosses η is when the joint type of $(\bar{x}^n(j), y^n)$, given by $Q_Y Q^j_{X|Y}$, lies on the exponential family $\mathcal{E}_{Q_Y}$ defined as

$$\mathcal{E}_{Q_Y} = \{Q_Y P^{(t)}_{X|Y} \text{ for } t \in [0,1]\}$$

This family connects $Q_Y P_{X|Y}$ and $Q_Y P_X$ within $\mathcal{L}_{Q_Y}$. It is illustrated in Fig. 2-3 as a straight line connecting $Q_Y P_{X|Y}$ and $Q_Y P_X$ (although note that it is not straight in the Euclidean sense).

Similar steps can be repeated for the score S_1 of the correct codeword. Conditioned on the output type Q_Y, symbols of the correct codeword are chosen i.i.d. with distribution $P_{X|Y}$. Again apply Sanov's theorem and I-projection to optimize the reverse channel type $Q^1_{X|Y}$ to the correct codeword $\bar{x}^n(1)$.

$$\lim_{n\to\infty} -\frac{\log P(S_1 \leq \eta_{Q_Y}(t)|\ Q_Y)}{n} = \min_{Q^1_{X|Y}:\ S_j \leq \eta_{Q_Y}(t)} D(Q_Y Q^1_{X|Y} \| p_1) \quad (2.21)$$

$$= D(Q_Y P^{(t)}_{X|Y} \| p_1) \quad (2.22)$$

$$= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_{X|Y}) \quad (2.23)$$

Thus given output type Q_Y, the dominating manner in which the correct codeword's score S_1 is smaller than η is when the joint type of $(\bar{x}^n(1), y^n)$, given by $Q_Y Q^1_{X|Y}$, lies on the same exponential family $\mathcal{E}_{Q_Y}$.

By the union bound, the exponent of the probability that one or more wrong codewords have a score crossing the threshold $\eta_{Q_Y}(t)$ is given by[6] $[D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) - R]^+$, where $[\mathrm{x}]^+ = \max\{0, x\}$. Since all codewords, are drawn independently of each other, the exponent $E(t, R, Q_Y)$ for the joint probability of $S_1 < \eta_{Q_Y}(t)$ and $S_j \geq \eta_{Q_Y}(t)$ (for some $j \neq 1$) is the sum of the exponents for these independent events.

[6]Since all codewords are drawn independently of each other, this exponent of union the bound is precise. This follows by standard arguments like Chebyshev's inequality as in [41].

$$E(t, R, Q_Y) \tag{2.24}$$

$$\equiv -\lim_{n\to\infty} \frac{\log P(S_1 \leq \eta_{Q_Y}(t),\ \exists j \neq 1 \text{ s.t. } S_j \geq \eta_{Q_Y}(t)|\ Q_Y)}{n} \tag{2.25}$$

$$= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_{X|Y}) + [D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) - R]^+ \tag{2.26}$$

The error exponent $E_r(R, Q_Y)$ conditioned on Q_Y is obtained by minimizing the above expression over the LLR $\eta_{Q_Y}(t)$ or equivalently minimizing it over t. This minimization corresponds to finding the LLR which dominates the error event for this Q_Y.

Using (2.7) and (2.8), we can prove the following:

$$D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) = D(Q_Y P^{(t)}_{X|Y} \| p_0) = \ t\eta_{Q_Y}(t) - \psi_{Q_Y}(t) \tag{2.27}$$

$$\text{where } \ \psi_{Q_Y}(t) \equiv \ \sum_{y\in\mathcal{Y}} Q_Y(y) \log k_y(t) \tag{2.28}$$

$$\text{Similarly, } \ D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_{X|Y}) = \ D(Q_Y P^{(t)}_{X|Y} \| p_1) \tag{2.29}$$

$$= (t-1)\eta_{Q_Y}(t) - \psi_{Q_Y}(t) \tag{2.30}$$

Recall that for each y, $k_y(t)$ is the normalization constant for reverse channel $P^{(t)}_{X|Y}$.

Now we will minimize $E(t, R, Q_Y)$ in (2.26) over t to obtain $E_r(R, Q_Y)$. Let $\hat{t}$ be the solution to the following equation

$$D(Q_Y P^{(\hat{t})}_{X|Y} \| Q_Y P_X) = D(Q_Y P^{(\hat{t})}_{X|Y} \| p_0) = R \tag{2.31}$$

For any $t < \hat{t}$, the exponent $E(t, R, Q_Y) > E(\hat{t}, R, Q_Y)$. This is because the first term in (2.26) increases with decreasing t and the second term will remain at 0 for $t < \hat{t}$. Hence the optimum solution lies in $[\hat{t}, 1]$. Since $D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) \geq R$ for $t \in [\hat{t}, 1]$,

$$\begin{aligned} [D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) - R]^+ &= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) - R \\ \Rightarrow \quad E(t, R, Q_Y) &= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_{X|Y}) + D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) - R \\ &= (2t-1)\eta_{Q_Y}(t) - 2\psi_{Q_Y}(t) - R \quad \text{(from (2.27) and (2.30))} \end{aligned}$$

Differentiating this w.r.t. t and equating it to 0 gives,

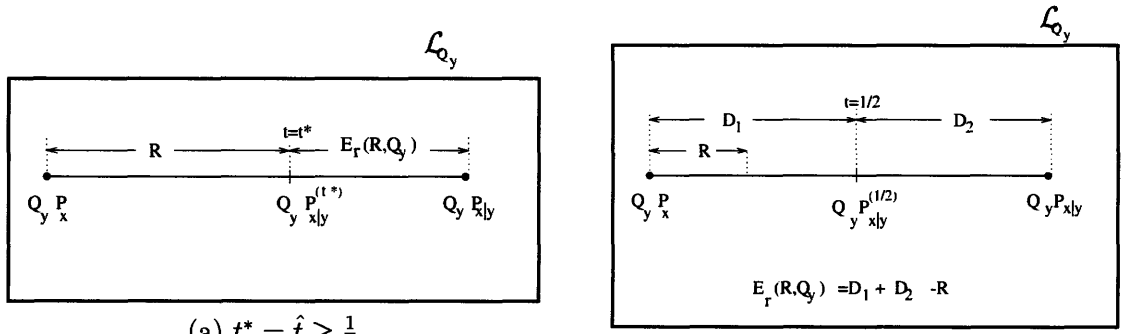
$$\begin{aligned}(2t-1)g_{Q_Y}(t) + 2\eta_{Q_Y}(t) - 2\frac{\partial\psi_{Q_Y}(t)}{\partial t} &= 0 \\ \Rightarrow \qquad (2t-1)g_{Q_Y}(t) &= 0 \qquad \left(\text{because } \frac{\partial\psi_{Q_Y}(t)}{\partial t} = \eta_{Q_Y}(t)\right)\end{aligned}$$

Since the Fisher information $g_{Q_Y}(t)$ is strictly positive, the optimum $t^* = \frac{1}{2}$, provided $\frac{1}{2} \in [\hat{t}, 1]$. Otherwise, if $\frac{1}{2} < \hat{t}$, then $E(t, R, Q_Y)$ is strictly increasing in $[\hat{t}, 1]$ because its derivative $(2t-1)g_{Q_Y}(t)$ is always positive. Then the optimum t^* equals $\hat{t}$. The optimum $t^* \geq \frac{1}{2}$ in either case. This phenomenon reflects the union bound constraint of $\rho \leq 1$ in Gallager's analysis. In fact the ρ in that analysis and t in this analysis are related as $t = 1/(1+\rho)$.

Thus we get the error exponent conditioned on output type Q_Y as follows:

$$\begin{aligned}E_r(R, Q_Y) &= D(Q_Y P_{X|Y}^{(\hat{t})} \| Q_Y P_{X|Y}) \quad \text{if } \hat{t} \geq \frac{1}{2} && (2.32)\\ &= D(Q_Y P_{X|Y}^{(1/2)} \| Q_Y P_{X|Y}) && (2.33)\\ &\quad + D(Q_Y P_{X|Y}^{(1/2)} \| Q_Y P_X) - R \quad \text{if } \hat{t} \leq \frac{1}{2} && (2.34)\end{aligned}$$

where $\hat{t}$ is the solution to $D(Q_Y P_{X|Y}^{(\hat{t})} \| Q_Y P_X) = R$. This solution is depicted in the figure below.



(a) $t^* = \hat{t} \geq \frac{1}{2}$ (b) $t^* = \frac{1}{2} > \hat{t}$

Figure 2-4: Geometric interpretation of error exponent conditioned on output type Q_Y. The rectangular frame in each figure denotes the linear family $\mathcal{L}_{Q_Y}$. The solid line in it represents the exponential family $\mathcal{E}_{Q_Y}$. The distance between two points corresponds to their KL divergence (in appropriate order).

Note from (2.31) that $\hat{t}$ increases with R. Hence the case of $\hat{t} \geq \frac{1}{2}$ corresponds to high (enough) rates R and vice versa. Thus the equations above (re)derive the following phenomenon in [5]:

1. The dominant cause of error for high (enough) rates (i.e. $\hat{t} \geq \frac{1}{2}$) is when a large number of incorrect codewords can be confused with the correct one.

2. The dominant cause of error for lower rates ($\hat{t} < \frac{1}{2}$) is when a single incorrect codeword is confused with the correct one.

The expression in (2.32) also provides an upper bound to the actual random coding exponent $E_r(R, Q_Y)$. This bound is related to the sphere-packing exponent. This expression is an upper bound because it is equivalent to relaxing the minimization constraint $t^* \geq 1/2$ (or $\rho < 1$) and assuming t^* always equals $\hat{t}$.

Remark 1: Note that for each output letter y, the dominant reverse channel type $Q_{X|Y}(\cdot|y)$ lies on the exponential family (in the space of distributions on $\mathcal{X}$) connecting $P_X(\cdot)$ and $P_{X|Y}(\cdot|y)$. Analysis in this section, based on the spaces of joint XY distributions, trivially shows the following coupling phenomenon between the dominant reverse channel type for all output letters y. The exponential parameter t is the same for the reverse channel type from each letter y, which creates a coupling between these reverse channel-types. Thus the dominant reverse channels for all output letters are equally tilted, where t corresponds to the common tilt parameter.

2.2.2 Optimizing over output types

Previously we found the error exponent $E_r(R, Q_Y)$ for a given output type Q_Y. Since the output sequence is generated i.i.d. according to P_Y, Stein's lemma implies that exponent of observing type Q_Y equals $D(Q_Y \| P_Y)$. Hence the overall exponent of error corresponding to output type Q_Y is given by

$$E_r(R, Q_Y) + D(Q_Y \| P_Y)$$

The effective error exponent is given by minimizing the above expression over all Q_Y.

$$E_r(R) = \min_{Q_Y} D(Q_Y \| P_Y) + E_r(R, Q_Y) \tag{2.35}$$

Let the optimum (or dominating) Q_Y be denoted by Q_Y^*. For a symmetric channel (like the BSC) with uniform input distribution P_X, conditional error exponent $E_r(R, Q_Y)$ is independent of Q_Y. Hence the Q_Y^* equals P_Y. Thus previous subsection is enough to understand the symmetric channel case.

However, for non-symmetric channels, the optimum Q_Y^* need not be simply P_Y. Let the joint type which dominates the error event be given by $Q_Y^* P_{X|Y}^{(t^*)}$, where Q_Y^* optimizes (2.35) and conditioned on Q_Y^*, the reverse channel type $P_{X|Y}^{(t^*)}$ achieves $E(R, Q_Y^*)$ in (2.32,2.33). We saw previously that dominating error event conditioned on the output type happens when the reverse channel type lies on the exponential family $\{P_{X|Y}^{(t)}$ for $t \in [0,1]\}$ of reverse channels. This family connects the trivial reverse channel P_X and actual reverse channel $P_{X|Y}$. It turns out that the dominating output type Q_Y^* also has such interpretation in terms of a certain exponential family. Refer to Appendix A for a simple proof based on I-projection.

Theorem 2 *Consider the exponential family connecting $p_1 = P_{XY}$ to $p_0 = Q_Y^* P_X$. The joint type $Q_Y^* P_{X|Y}^{(t^*)}$ dominating the error event lies on this exponential family (see Fig. 2.2.2).*

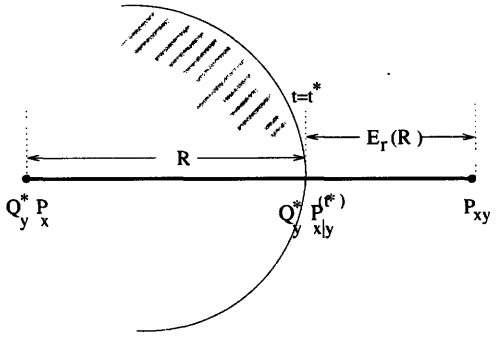
$$\begin{aligned} Q_Y^*(y) P_{X|Y}^{(t^*)}(x|y) &= \frac{p_1^{t^*}(x,y) p_0^{1-t^*}(x,y)}{k(t^*)} \\ \text{where } k(t^*) &= \sum_{x,y} p_1^{t^*}(x,y) p_0^{1-t^*}(x,y) \end{aligned}$$

The interesting part of this theorem is although Q_Y^* need not be the Y-marginal throughout this exponential family in space of joint XY distributions. Nonetheless, it is indeed the Y-marginal at the optimum t^* on this exponential family: optimality guarantees consistency. Also note that reverse channel at any t on this exponential family is the same as $P_{X|Y}^{(t)}$ as before. Recall that $P_{X|Y}^{(t)}$ was the reverse channel seen in

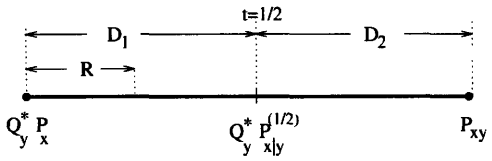
previous subsection (where Y-marginal was fixed to Q_Y) for exponential family $\mathcal{E}_{Q_Y}$ connecting $Q_Y P_X$ and $Q_Y P_{X|Y}$.

Figure 2-5: Geometric interpretation of the dominant Q_Y and the random coding exponent $E_r(R)$. Here plane of the paper represents the space of all joint distributions. The solid line in each figure represents the exponential family connecting its two ends (although it is not straight in the Euclidean sense).

The distance between two points corresponds to their KL divergence and the shaded ball in Figure (a) denotes all joint distributions within KL divergence R from $Q_Y^* P_X$. As an aside, Figure 2.2.2(a) is very similar to the figure in [14] for the source-coding error exponent. That source-coding exponent is obtained similarly by I-projection on a divergence ball.



(a) $R \geq R_c$



$E_r(R) = D_1 + D_2 - R$

(b) $R \leq R_c$

The distance between two points corresponds to their KL divergence and the shaded ball in Figure (a) denotes all joint distributions within KL divergence R from $Q_Y^* P_X$. As an aside, Figure 2.2.2(a) is very similar to the figure in [14] for the source-coding error exponent. That source-coding exponent is obtained similarly by I-projection on a divergence ball.

Recalling $P_{X|Y}^{(t^*)}(x|y) = \frac{P_{X|Y}^{t^*}(x|y) P_X^{1-t^*}(x)}{k_y(t^*)}$ from (2.15) and plugging this in the above

theorem gives

$$
\begin{aligned}
(Q_Y^*(y))^{t^*} &\propto (P_Y(y))^{t^*} k_y(t^*) && (2.36)\\
\Rightarrow \quad Q_Y^*(y) &\propto P_Y(y) \cdot k_y^{1/t^*}(t^*) && (2.37)\\
&= P_Y(y) \cdot \left(\sum_x P_{X|Y}^{t^*}(x|y) P_X^{1-t^*}(x) \right)^{1/t^*} && (2.38)\\
\text{(by Baye's rule)} \quad &= \left(\sum_x P_X(x) P_{Y|X}^{t^*}(y|x) \right)^{1/t^*} && (2.39)
\end{aligned}
$$

This is the same solution as [7] for the dominant output type Q_Y^* for error events. To emphasize the dependence of Q_Y^* on t^*, let it be denoted by $Q_Y^{(t^*)}$. The optimum t^* equals $\frac{1}{2}$ for rates below the critical rate given by

$$R_c = D(Q_Y^{(1/2)} P_{X|Y}^{(1/2)} \| Q_Y^{(1/2)} P_X) \tag{2.40}$$

For higher rates, t^* is the solution to

$$D(Q_Y^{(t^*)} P_{X|Y}^{(t^*)} \| Q_Y^{(t^*)} P_X) = R \tag{2.41}$$

Since t^* and $Q_Y^{(t^*)}$ are dependent on each other through (2.39) and (2.41), a closed for expression cannot be given for either of them. However, iterating between (2.39) and (2.41) converges to the optimum t^* and $Q_Y^{(t^*)}$. This gives an algorithm similar to Blahut-Arimoto for calculating error exponents at $R > R_c$.

To summarize, we derived the random coding error exponent:

$$
\begin{aligned}
E_r(R) &= D(Q_Y^{(1/2)} P_{X|Y}^{(1/2)} \| P_{XY}) + && (2.42)\\
&\quad D(Q_Y^{(1/2)} P_{X|Y}^{(1/2)} \| Q_Y^{(1/2)} P_X) - R \quad \text{for } R < R_c && (2.43)\\
&= (Q_Y^{(t^*)} P_{X|Y}^{(t^*)} \| P_{XY}) \qquad\qquad \text{for } R \geq R_c && (2.44)
\end{aligned}
$$

2.2.3 Very noisy channels

We define a very noisy channel as a channel for which the distribution P_X is very close to $P_{X|Y}(\cdot|y)$ for each given $y \in \mathcal{Y}$. Thus the conditional distributions $P_{X|Y}(\cdot|y)$, denoted in short as $P_{X|Y=y}$, are very close to the unconditional distribution P_X.

Equivalently for each $y \in \mathcal{Y}$, the Fisher information $g_y(t)$ is constant[7] for the exponential family of $\mathcal{X}$-distributions joining $P_{X|Y=y}$ and P_X. Let this constant Fisher information for output y be denoted by g_y. For a given output type Q_Y this implies

$$\begin{aligned} D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) &= \sum_{y\in\mathcal{Y}} Q_Y(y) D(P^{(t)}_{X|Y=y} \| P_X) \\ &= \sum_{y\in\mathcal{Y}} Q_Y(y) \frac{g_y t^2}{2} \quad \text{(using (2.11))} \\ &\equiv t^2 C_{Q_Y} \end{aligned}$$

where C_{Q_Y} is a shorthand for $\left(\sum_{y\in\mathcal{Y}} \frac{Q_Y(y) g_y}{2}\right)$. Substituting $t = 1$ shows that C_{Q_Y} equals $D(Q_Y P_{X|Y} \| Q_Y P_X)$. Similarly using (2.11), we can show that

$$\begin{aligned} D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_{X|Y}) &= \sum_{y\in\mathcal{Y}} Q_Y(y) D(P^{(t)}_{X|Y=y} \| P_{X|Y=y}) \\ &= (1-t)^2 C_{Q_Y} \end{aligned}$$

Recalling that the error exponent conditioned on output type Q_Y for $R \geq R_c$ equals

$$\begin{aligned} E_r(R, Q_Y) &= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_{X|Y}) = (1-t)^2 C_{Q_Y} \\ \text{where } t \text{ satisfies, } R &= D(Q_Y P^{(t)}_{X|Y} \| Q_Y P_X) = t^2 C_{Q_Y} \end{aligned}$$

[7]This happens when the channel's conditional output distributions $W_{Y|X}(\cdot|x)$ for different $x \in \mathcal{X}$ are very close. That is, for some probability distribution q_Y, Gallager defines [2]

$$W_{Y|X}(y|x) = q_Y(y) + \epsilon T(x, y) \tag{2.45}$$

where ϵ tends to zero in the very noisy limit and $T(\cdot,\cdot)$ is a fixed matrix. Every row of T sums to zero to ensure that $W_{Y|X}(\cdot|x)$ remains a valid distribution. The channel capacity and Fisher Informations $g_y(t)$ are of the order $O(\epsilon^2)$ in this setup. Up to this order, the Fisher information $g_y(t)$ remains constant in the range of interest for t.

Eliminating t from the two equations gives,

$$E_r(R, Q_Y) = C_{Q_Y}\left(1 - \sqrt{R/C_{Q_Y}}\right)^2 = (\sqrt{C_{Q_Y}} - \sqrt{R})^2$$

The overall error exponent is obtained by optimizing over the output type

$$E_r(R) = \min_{Q_Y} D(Q_Y \| P_Y) + E_r(R, Q_Y) = \min_{Q_Y} D(Q_Y \| P_Y) + (\sqrt{C_{Q_Y}} - \sqrt{R})^2$$

Choosing $Q_Y = P_Y$ for the capacity achieving P_Y gives the approximate solution in[8] [2].

For the above minimization over Q_Y, since $E_r(R, Q_Y)$ depends on Q_Y only through its effect on $C_{Q_Y} = \mathrm{E}_{Q_Y}[g_y/2]$, the optimum Q_Y^* will be on the exponential family $\mathcal{E}_{f,P_Y}$ of $\mathcal{Y}$-distributions going through P_Y for $f(y) = g_y = 2D(P_{X|Y=y} \| P_X)$ (due to the very noisy assumption). This family is denoted by $\mathcal{E}_{g_y,P_Y}$ (see Fig. 2-6). I-projection theorem implies the optimum Q_Y^* is of the form

$$Q_Y^*(y) = \frac{1}{k(\theta)} P_Y(y) \exp(\theta g_y)$$

for some $\theta \in \mathcal{R}$ where $k(\theta)$ is the normalization constant. A general (not very noisy) channel need not have Q_Y^* of this form.

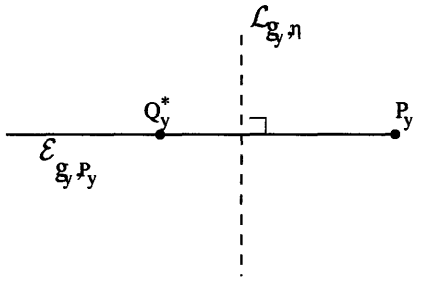


Figure 2-6: The space of Y-distributions: solid line shows the exponential family $\mathcal{E}_{g_y,P_Y}$ and dashed line shows an orthogonal linear family $\mathcal{L}_{g_y,\eta} = \{Q_Y | \mathrm{E}_{Q_Y}[g_Y] = \eta\}$.

[8] In fact, for the very noisy channel in [2], this substitution gives the right answer in the ϵ^2 scaling of interest as ϵ in (2.45) tends to zero.

2.3 Error exponent of the expurgated ensemble

Snce I-projection gives us a tool to address high-dimensional optimizations, let us analyze the error exponent of Gallager's expurgated ensemble. We first create an i.i.d. random code of rate R, generated using input distribution P. With some abuse of notation, let x^n denote the correct codeword $\bar{x}^n(1)$ and let z^n denotes an incorrect codeword $\bar{x}^n(i)$. Let Q_{XZ} denote the joint type of (x^n, z^n) and let $Q_{Y|XZ}(\cdot|xz)$ denote the conditional output type of y^n where the correct input is x and the incorrect input is z.

From this random code, let us throw away the "*bad*" codeword pairs which are "*too close*". More specifically, we expurgate codeword pairs such that $D(Q_{XZ}\|P\otimes P) > R$, where $P \otimes P$ denotes independent X and Z, with marginal distribution P for both. Since the fraction of such codewords will be exponentially small, this rule guarantees that the rate of the expurgated code is not smaller than R. After expurgation all codeword pairs will satisfy $D(Q_{XZ}\|P \otimes P) \leq R$. The error exponent $E_{ex}(R)$ of this expurgated code equals

$$E_{ex}(R) = \min_{Q_{XZ}:\ D(Q_{XZ}\|P\otimes P)\leq R} E_x(Q_{XZ}) + (D(Q_{XZ}\|P \otimes P) - R) \tag{2.46}$$

where $E_x(Q_{XZ})$ denotes the error exponent for a codeword pair with type Q_{XZ}. The second term is due to the union bound[9] and the fact that the exponent of observing the joint type Q_{XZ} in an i.i.d. random code equals $D(Q_{XZ}\|P \otimes P)$. Now let us analyze $E_x(Q_{XZ})$, which is given by this minimization over conditional types $Q_{Y|XZ}$.

$$E_x(Q_{XZ}) = \min_{Q_{Y|XZ}:\ \text{error}} \sum_{x,z} Q_{XZ}(x,z) D\left(Q_{Y|XZ}(\cdot|xz)\|W_{Y|X}(\cdot|x)\right) \tag{2.47}$$

where $W_{Y|X}(\cdot|x)$ denotes the actual channel distribution from the correct input and error happens when log-likelihood of the correct codeword is smaller than that of the

[9] As seen earlier, this term of union bound is precise since all codewords are drawn independently of each other. This follows by standard arguments like Chebyshev's inequality as in [41].

wrong codeword, that is,

$$\sum_{x,z,y} Q_{XZ}(x,z)Q_{Y|XZ}(y|xz)\log\frac{W_{Y|X}(y|x)}{W_{Y|X}(y|z)} \leq 0 \quad \Leftrightarrow \quad \text{error} \tag{2.48}$$

$$\text{i.e.} \quad \mathrm{E}_{Q_{XZ}Q_{Y|XZ}}\left[L(Y|X,Z)\right] \leq 0 \quad \Leftrightarrow \quad \text{error} \tag{2.49}$$

where $L(y|x,z)$ is a shorthand for the log-likelihood ratio $\log\frac{W_{Y|X}(y|x)}{W_{Y|X}(y|z)}$. Thus calculating $E_x(Q_{XZ})$ involves minimizing a weighted average of $D\left(Q_{Y|XZ}(\cdot|xz)\|W_{Y|X}(\cdot|x)\right)$ in (2.47) under the constraint (2.49) on the weighted average of log-likelihood ratio.

In this optimization for $E_x(Q_{XZ})$, I-projection implies that optimum conditional type $Q_{Y|XZ}(\cdot|xz)$ for any pair (x,z) lies on the exponential family $\{P^{(t)}_{Y|XZ} : t \in [0,1]\}$ connecting the channel from correct input $W_{Y|X}(\cdot|x)$ to the channel from incorrect input $W_{Y|X}(\cdot|z)$:

$$P^{(t)}_{Y|XZ}(\cdot|xz) \propto W^{1-t}_{Y|X}(\cdot|x)\cdot W^{t}_{Y|X}(\cdot|z) \quad \forall\, x,z$$

Similar to Remark 1, the exponential parameter t is the same for each pair $(x,z) \in \mathcal{X}\times\mathcal{X}$. Hence finding $E(Q_{XZ})$ only needs a scalar optimization:

$$E(Q_{XZ}) = \min_{\hat{t}:\text{error}} \sum_{x,z} Q_{XZ}(x,z)D\left(P^{(\hat{t})}_{Y|XZ}(\cdot|xz)\|W_{Y|X}(\cdot|x)\right) \tag{2.50}$$

where error happens for $\hat{t}$, i.e., $\mathrm{E}_{Q_{XZ}P^{(\hat{t})}_{Y|XZ}(Y|XZ)}\left[L(Y|X,Z)\right] \leq 0$.

Let us compare this approach to Gallager's analysis of the expurgated ensemble in [38]. There the error probability for codeword pair (x^n, z^n) is bounded as follows:

$$\begin{aligned}\Pr(\text{error from } x^n \text{ to } z^n) &= \sum_{y^n:\ W_{Y|X}(y^n|z^n)\geq W_{Y|X}(y^n|x^n)} W_{Y|X}(y^n|x^n) \\ &\leq \sum_{y^n} W_{Y|X}(y^n|x^n)\sqrt{\frac{W_{Y|X}(y^n|z^n)}{W_{Y|X}(y^n|x^n)}}\end{aligned}$$

where $W_{Y|X}(y^n|x^n)$ is used as a shorthand for $\Pi_i W_{Y|X}(y_i|x_i)$, the conditional proba-

bility for this channel $W_{Y|X}$.

An exercise in [4] also starts with the same square-root trick and gets the same error exponent as in [38]. This trick is equivalent to substituting $\hat{t} = \frac{1}{2}$ as the minima in (2.50). It is not clear why the minimum should always be attained at $\frac{1}{2}$.

However, at $R = 0$, the expurgation constraint $D(Q_{XZ}\|P \otimes P) \leq 0$ implies $Q_{XZ} = P \otimes P$, which is a symmetric distribution in (x, z). It is easy to see that $\hat{t} = \frac{1}{2}$ should attain the minimum in (2.50). It follows since for a symmetric Q_{XZ}, $\frac{1}{2}$ is the smallest $\hat{t}$ where the error constraint in (2.49) is satisfied (with equality). For the minimization in (2.50), we should choose the smallest possible $\hat{t}$ as $W_{Y|X}(\cdot|x)$ corresponds to $t = 0$.

For $R > 0$ however, it is not as obvious. Nonetheless, even for $R > 0$, it can be shown that the minimum must be attained at $\frac{1}{2}$. This follows by noticing that the overall optimization problem for $E_{ex}(R)$ over Q_{XZ} and $Q_{Y|XZ}$:

$$E_{ex}(R) = \min_{Q_{XZ}: D(Q_{XZ}\|P\otimes P)\leq R} \; \min_{Q_{Y|XZ}:\ \text{error}} \left(E_x(Q_{XZ}) + D(Q_{XZ}\|P \otimes P) - R\right)$$

$$\textit{where,} \quad \text{error} \equiv \mathrm{E}_{Q_{XZ}Q_{Y|XZ}}\left[L(Y|X,Z)\right] \leq 0$$

Note that for a fixed $Q_{Y|XZ}$, this problem is a convex minimization in Q_{XZ} and vice versa. Now fixing $\hat{t} = \frac{1}{2}$ means fixing $Q_{Y|XZ}$ to be $P^{(1/2)}_{Y|XZ}$. For this fixed choice of $Q_{Y|XZ}$, the above minimization over Q_{XZ} yields an optimal Q^*_{XZ} which is symmetric[10] in X and Z. As we discussed before, if we change the order of minimization by fixing this symmetric Q^*_{XZ} and then optimizing over $Q_{Y|XZ}$, we get $\hat{t} = \frac{1}{2}$ as the optimum choice.

The above discussion shows how $\hat{t} = \frac{1}{2}$, i.e., $Q_{Y|XZ} = P^{(1/2)}_{Y|XZ}$ and the corresponding optimal Q^*_{XZ} are a stationary pair for the minimization for $E_{ex}(R)$. Convexity of that minimization in its arguments implies the optimality of this pair. This concludes the argument for optimality of Gallager's square-root trick. This proves that his expurgated bound was tight for the expurgated ensemble and further strengthens the

[10]This follows since if an assymetric Q'_{XZ} is feasible, then interchanging X with Z in Q'_{XZ} also provides a feasible XZ distribution. Denote this flipped distribution by $\bar{Q}'_{XZ}$ and note that $\frac{Q'_{XZ}+\bar{Q}'_{XZ}}{2}$ is symmetric, feasibile, and gives smaller $E_{ex}(R)$ than Q'_{XZ} or $\bar{Q}'_{XZ}$ due to convexity of KL divergence.

conjecture in [3] about tightness of expurgated bound for an arbitrary code (at small enough R).

2.4 Concluding remarks

A similar analysis can be used for more high-dimensional problems such as expurgation for List-of-L decoding. It also clarifies what it means for a tuple of codewords to be "*too close*" to each other, i.e., which codeword-lists are likely to cause errors. Consider the List-of-2 decoding for example. If $(x^n, z^n, \tilde{z}^n)$ denotes the correct codeword and two incorrect codewords, we will expurgate all codeword triads for which the joint type $Q_{XZ\tilde{Z}}$ satisfies $D(Q_{XZ\tilde{Z}}\|P_X \otimes P_X \otimes P_X) \leq 2R$. This is on similar lines of the expurgation constraint $D(Q_{XZ}\|P_X \otimes P_X) \leq R$ in the last section. Again this expurgation step has no effect on the rate of this randomly generated code.

One interesting difference now is we need to consider the exponential family between three distributions as opposed to two. Intuitively, it corresponds to the "triangle" connecting these three distributions instead of the "straight line" connecting two distributions. There are some curious similarities between properties of this triangle with that of an Euclidean triangle. In particular, one can create an analogue of the elementary geometry theorem that 'all perpendicular bisectors of a triangle coincide at a point and its distance from all corners of the triangle equals the circum-radius'. Figuring out the details of this theorem's analog is left as an excercise[11]. This circum-radius also has an operational meaning. It is related to the error exponent for List-of-two decoding in a ternary Hypothesis-Testing problem, where under each hypothesis, the observed Y^n is distributed i.i.d. according to one of the corners of the triangle.

Now let us move to a rich class of error exponent problems that arise in unequal error protection scenarios.

[11]Hint: The the sides of the triangle are certain exponential families and their perpendicular bisectors are certain linear families.

Chapter 3

Unequal Error Protection near Capacity

This chapter discusses error exponents for UEP in a point-to-point channel without feedback. We will focus on situations where the data-rate is essentially at channel capacity[1]. Now exponential reliability becomes a luxury which is not available for all information. We will answer whether and how at least some special bits or some special messages can achieve exponential reliability.

We start by defining the channel model and some basic definitions in Section 3.1. Our results on UEP exponents for bit-wise UEP and message-wise UEP for block codes are discussed in Section 3.2. After discussing each theorem, we provide a brief description of the optimal strategy. Proof details can be found in Section 3.3.

3.1 Channel Model and Notation

3.1.1 Channel Model and Block Codes

We will consider a discrete memoryless channel $W_{Y|X}$, with input alphabet $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ and output alphabet $\mathcal{Y} = \{1, 2, \ldots, |\mathcal{Y}|\}$. As earlier, this means that the conditional distribution of output Y when the channel input X equals $x \in \mathcal{X}$ is

[1]This chapter and the next chapter is joint work with Baris Nakiboglu (in addition to Lizhong Zheng) and these results were reported first in [10, 11].

denoted by $W_{Y|X}(\cdot|x)$. We assume that all the entries of the channel transition matrix are non-zero, that is, every output letter is reachable from every input letter. This assumption is indeed a crucial one and many results will change when some channel transitions have zero probability.

A length n block code without feedback with message set $\mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\}$ consists of an encoder and decoder. The encoder assigns a length n codeword $\bar{x}^n(k) \equiv (\bar{x}_1(k), \bar{x}_2(k) \cdots, \bar{x}_n(k))$ for each $k \in \mathcal{M}$, where $\bar{x}_t(k)$ denotes its input at time t. At time zero, the transmitter is given the message M, which is chosen uniformly from $\mathcal{M}$. In the following n time units, it sends the corresponding codeword $\bar{x}^n(M)$. After observing Y^n, the receiver chooses the decoded message $\hat{M}(Y^n)$.

The average error probability P_e and rate R of the code is given by

$$P_e \equiv \Pr\left[\hat{M} \neq M\right] \qquad \text{and} \qquad R \equiv \frac{\log|\mathcal{M}|}{n}. \tag{3.1}$$

3.1.2 Different Kinds of Errors

In message-wise UEP, we will consider the conditional error probability for a particular message $i \in \mathcal{M}$:

$$\Pr\left[\hat{M} \neq i \middle| M = i\right]. \tag{3.2}$$

Recall that this is the same as the missed-detection probability for message i.

On the other hand when we are talking about bit-wise UEP, the overall message is composed of two components, $M = (M_1, M_2)$, where M_i is chosen uniformly from message set $\mathcal{M}_i$. For example, M_1 may correspond to the high-priority bits while M_2 corresponds to the low-priority bits. Note that now the message set $\mathcal{M}$ is equal to the Cartesian product $\mathcal{M}_1 \times \mathcal{M}_2$. The error probability of decoding M_j is given by

$$\Pr\left[\hat{M}_j \neq M_j\right] \qquad j = 1, 2 \tag{3.3}$$

Note that the overall message M is decoded incorrectly when either M_1 or M_2 or both are decoded incorrectly. The goal of bit-wise UEP is to achieve the best possible $\Pr\left[\hat{M}_1 \neq M_1\right]$ while still ensuring vanishingly small overall $P_e = \Pr\left[\hat{M} \neq M\right]$.

3.1.3 Reliable Code Sequences

For our discussion of error exponents while reliably approaching capacity, we use the notion of code-sequences to simplify our discussion without losing its rigor.

A sequence of codes indexed by code-length is called *reliable* if and only if

$$\lim_{n\to\infty} {P_e}^{(n)} = 0 \tag{3.4}$$

For any reliable code-sequence $\mathcal{Q}$, its rate $R_{\mathcal{Q}}$ is given by

$$R_{\mathcal{Q}} \equiv \lim_{n\to\infty} \frac{\log|\mathcal{M}^{(n)}|}{n} \tag{3.5}$$

Throughout this thesis, rates will be defined in terms of $\lim_{n\to\infty}$ as above and we will solely focus on $\mathcal{Q}$ for which this limit exists. The (conventional) error exponent of a reliable sequence is defined as

$$E_{\mathcal{Q}} \equiv \liminf_{n\to\infty} \frac{-\log {P_e}^{(n)}}{n} \tag{3.6}$$

Thus the number of messages in $\mathcal{Q}$ grows as $\doteq e^{nR_{\mathcal{Q}}}$ and its average error probability decays as ${P_e}^{(n)} \doteq e^{-nE_{\mathcal{Q}}}$, where $\doteq$ denotes equality in the exponential sense. More precisely, for a sequence $a^{(n)}$

$$a^{(n)} \doteq e^{nF} \Leftrightarrow F = \liminf_{n\to\infty} \frac{\log a^{(n)}}{n}. \tag{3.7}$$

For cleaner expressions, we will omit the sequence index (n) in the superscript whenever possible. Now let us define error exponent $E(R)$ in the conventional sense, as in [2],[3],[4],[5],[7].

Definition 3 *For any $R \leq$ C the error exponent $E(R)$ is defined as*

$$E(R) \equiv \sup_{\mathcal{Q}:R_{\mathcal{Q}}\geq R} E_{\mathcal{Q}} \tag{3.8}$$

As mentioned previously, we are interested here in UEP when operating at capacity.

We already know that $E(C) = 0$, [3], i.e. the overall error probability cannot decay exponentially at capacity. In the following sections, we will show how certain parts of information can still achieve a positive exponent at capacity. In doing that, we will be solely focusing on the reliable $\mathcal{Q}$'s whose rates are equal to C. We will call such reliable code sequences *capacity-achieving sequences.* This definition is especially useful since capacity is defined as the supremum of all achievable rates and the notion of operating *at* capacity needs to be formalized carefully. With the above definition of a capacity-achieving sequence, we can transmit reliably *at* capacity[2].

Adhering to earlier notation, the KL divergence between two distributions $\alpha_X(\cdot)$ and $\beta_X(\cdot)$ is denoted by $D\left(\alpha_X(\cdot)\|\,\beta_X(\cdot)\right)$ or $D\left(\alpha_X\|\,\beta_X\right)$ in short.

$$D\left(\alpha_X(\cdot)\|\,\beta_X(\cdot)\right) = \sum_{x\in\mathcal{X}} \alpha_X(x) \log \tfrac{\alpha_X(x)}{\beta_X(x)}$$

The conditional KL divergence between $W_{Y|X}(\cdot|\cdot)$ and $V_{Y|X}(\cdot|\cdot)$ under $P_X(\cdot)$ will be denoted by $D\left(W_{Y|X}(\cdot|X)\middle\|\,V_{Y|X}(\cdot|X)\middle|\,P_X\right)$. This is the same as the divergence between the joint XY-distributions $P_X W_{Y|X}$ and $P_X V_{Y|X}$.

$$\begin{aligned} D\left(W_{Y|X}(\cdot|X)\middle\|\,V_{Y|X}(\cdot|X)\middle|\,P_X\right) &\equiv \sum_{x\in\mathcal{X}} P_X(x) D\left(W_{Y|X}(\cdot|x)\middle\|\,V_{Y|X}(\cdot|x)\right) \\ &= \mathrm{E}_{P_X}\left[D\left(W_{Y|X}(\cdot|X)\middle\|\,V_{Y|X}(\cdot|X)\right)\right] \end{aligned}$$

An input distribution that achieves capacity will be denoted by P_X^*. The corresponding output distribution will be denoted by P_Y^*.

[2]Recall that in conventional definitions [2], the message set grows as $|\mathcal{M}^{(n)}| = \lceil \exp(nR) \rceil$ at data rate R. In that setup, reliable communication at $R = C$ may not be possible. However, with our definition, the message set of a capacity achieving $\mathcal{Q}$ could grow as $|\mathcal{M}^{(n)}| = \lceil \frac{\exp(nR)}{n^2} \rceil$ for example. This decrease in $|\mathcal{M}^{(n)}|$ by a sub-exponential factor allows reliable communication at capacity for our definition.

3.2 UEP Exponents for Block Codes

3.2.1 Special bit

We first address the situation where one particular information bit (say the first) out of the total $\log_2 |\mathcal{M}|$ information bits is a special bit—it needs a much better error protection than the overall information. For example, this special bit may be thought as the shortest possible packet header. If this first bit is denoted as b_1 and its decoded value is denoted by $\hat{b}_1$, we require the error probability for b_1 to decay exponentially while still ensuring reliable communication at capacity for the remaining bits.

In the Cartesian-product terminology, the single special bit scenario is equivalent to defining $\mathcal{M}_1 = \{0, 1\}$, from which M_1 is chosen uniformly—it denotes the special bit b_1. The overall message equals $M = (M_1, M_2)$, where M_2 is independent of M_1 and chosen uniformly from $\mathcal{M}_2$. The optimal error exponent E_b for the special bit is be defined as follows[3].

Definition 4 *For a capacity-achieving sequence $\mathcal{Q}$ with message sets $\mathcal{M}^{(n)} = \mathcal{M}_1 \times \mathcal{M}_2^{(n)}$ where $\mathcal{M}_1 = \{0, 1\}$, the special bit error exponent is defined as*

$$E_{b,\mathcal{Q}} \equiv \liminf_{n\to\infty} \frac{-\log \Pr\left[\hat{M}_1 \neq M_1\right]}{n} \tag{3.9}$$

Then E_b is defined as $E_b \equiv \sup_{\mathcal{Q}} E_{b,\mathcal{Q}}$.

Thus if $\Pr\left[\hat{b}_1 \neq b_1\right] \doteq \exp(-nE_{\mathrm{b},\mathcal{Q}})$ for a reliable sequence $\mathcal{Q}$, then E_b is the supremum of $E_{\mathrm{b},\mathcal{Q}}$ over all capacity-achieving $\mathcal{Q}$.

Since $E(C) = 0$, it is clear that the entire information cannot achieve any positive error exponent at capacity. However, it is not clear whether a single special bit can steal a positive error exponent E_b at capacity.

Theorem 5 $E_b = 0$

[3]Appendix B discusses a different but equivalent type of definition and shows its equivalence to this one. In that definition, we first define the best single bit exponent achievable while communicating reliably at a data-rate below capacity, $R < C$. Then the single bit exponent at capacity is defined as the infimum of these single bit exponents over $R \in [0, C)$. These two types of definitions are equivalent for all the UEP exponents we discuss.

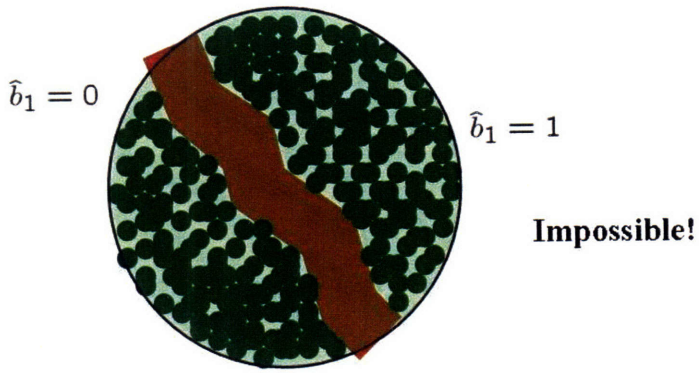


Figure 3-1: Splitting the output space into 2 distant enough clusters.

This implies that if we are aiming to protect a single bit with exponential reliability, the data-rate must be strictly less than capacity. Section 3.3 contains a detailed proof of this theorem.

Intuitive Interpretation: Let the shaded balls in Fig. 3-1 denote the smallest decoding regions of the $\doteq e^{nC}$ messages for ensuring reliable communication. These decoding regions essentially denote the typical noise-balls [9] around codewords, which ensure reliable communication.

The decoding regions on the left of the thick line corresponds to $\hat{b}_1 = 1$ and those on the right correspond to the same when $\hat{b}_1 = 0$. Each of these halves includes half of the decoding regions.

For achieving a positive error exponent for the special bit, the codewords in the two halves should be sufficiently separated from each other as seen in Fig. 3-1. Such separation is necessary to ensure exponentially small probability of landing in the wrong half. However, the above theorem indicates that such a thick patch takes too much volume, and is impossible when we have to fill $\doteq e^{nC}$ typical noise balls in this output space.

Remark: This theorem for a single special bit immediately implies that no positive exponent is possible for multiple special bits when the data-rate is approaching capacity. This is helpful for characterizing the effectiveness of codes when the packet header is very short compared to the total packet length. In many communication

protocols, such short headers are protected by simply using some low-rate code. Such a low-rate header code can achieve a good (positive) error exponent in terms of its own length. However, often the header code is simply concatenated with the code for ordinary bits and the length of the header code is negligible compared to the code for ordinary bits. Hence if the code for ordinary bits is approaching capacity, the concatenated code also approaches capacity. Since the length of the header code becomes a negligible fraction of the overall codelength, the special bits in the header cannot achieve any error exponent in terms of the overall codelength. Theorem 5 implies that no other scheme can achieve a better error exponent for the special bits[4] in terms of the overall codelength.

3.2.2 Special message

Now consider situations where one particular message (say $M = 1$) out of the $\doteq e^{nC}$ total messages is a special message—it needs a superior error protection. The missed-detection probability for this 'emergency' message needs to be minimized. A popular approach in many network protocol is to simply add a flag bit to indicate this special message. This flag bit is 1 when the special message is to be sent and 0 otherwise. The special message is then protected by protecting this special flag bit better than other bits. However, this approach cannot provide a positive exponent for the special message due to the previous negative result for a special bit, $E_{\mathrm{b}} = 0$. Hence some different approach needs to be used. Let us now define the best missed-detection exponent E_{md} on similar lines of E_{b}.

Definition 6 *For a capacity-achieving sequence $\mathcal{Q}$, the missed-detection exponent is defined as*

$$E_{md,\mathcal{Q}} \equiv \liminf_{n\to\infty} \frac{-\log \Pr\left[\hat{M}\neq 1 \middle| M=1\right]}{n}. \tag{3.10}$$

Now define $E_{md} = \sup_{\mathcal{Q}} E_{md,\mathcal{Q}}$.

[4]However, this simple concatenation scheme may not be optimal when the overall data-rate is not approaching capacity. The UEP exponents for the case of rates below capacity are discussed in Chapter 5.

Compare this with the situation where we aim to protect all the messages uniformly well. If all the messages demand equally good missed-detection exponents, then no positive exponent is achievable at capacity. This follows from the earlier discussion about $E(C) = 0$. The theorem below shows the improvement in this exponent if we only demand it for a single message instead of all.

Definition 7 *The parameter* $\mathsf{E}_{\mathsf{Red}}$ *is called*[5] *the* Red-Alert Exponent *of a channel.*

$$\mathsf{E}_{\mathsf{Red}} \equiv \max_{x \in \mathcal{X}} D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|x)\right) \tag{3.11}$$

We will denote an input letter achieving above maximum by x_r.

Theorem 8 $E_{md} = \mathsf{E}_{\mathsf{Red}}$.

Notice the relation between $\mathsf{E}_{\mathsf{Red}}$ and C: the arguments to KL divergence are flipped. It is because Karush-Kuhn-Tucker (KKT) conditions for achieving capacity imply the following expression for C [4].

$$C = \max_{x \in \mathcal{X}} D\left(W_{Y|X}(\cdot|x) \| P_Y^*(\cdot)\right) \tag{3.12}$$

Capacity C represents the best possible data-rate for a channel, then Red-Alert Exponent $\mathsf{E}_{\mathsf{Red}}$ represents its best possible protection for a message while achieving capacity.

It is worth mentioning here the "very noisy" channel in [2]. As discussed later in Chapter 6, the KL divergence is symmetric for very noisy channels. This means $D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|i)\right) \approx D\left(W_{Y|X}(\cdot|i) \| P_Y^*(\cdot)\right)$, which implies the Red-Alert Exponent and the capacity are essentially equal. For a symmetric channel like the BSC, every input $x \in \mathcal{X}$ achieves the maximum in Eq. (3.11) and hence can be chosen as x_r. Since P_Y^* is the uniform distribution for these channels, $\mathsf{E}_{\mathsf{Red}} = D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|x)\right)$ for any input letter x. This also happens to be the sphere-packing exponent $E_{\mathrm{sp}}(0)$ of this channel [3] at rate 0.

[5] Thanks to Krishnan Eswaran of UC Berkeley for suggesting this name.

Optimal strategy: The codeword for the special message can be optimally chosen as a repetition sequence of the input x_r. Its decoding region $\mathcal{G}(1)$ contains every output sequence whose empirical distribution, i.e., type is not (approximately) P_Y^*. For ordinary messages, codewords of a capacity achieving code will be used. The receiver uses maximum-likelihood (ML) decoding over the ordinary codewords for output sequences outside $\mathcal{G}(1)$.

Intuitive Interpretation: The missed-detection exponent for the special message corresponds to having a large decoding region $\mathcal{G}(1)$ for the special message. This ensures that when the special message is transmitted, the probability of landing outside $\mathcal{G}(1)$ is exponentially small. In a sense, E_{md} indicates how large $\mathcal{G}(1)$ could be made, while still filling $\doteq e^{nC}$ typical noise balls in the remaining space. The red region in Fig. 3-2 denotes such a large region. Note that the actual decoding region $\mathcal{G}(1)$ is much larger than this illustration, because it consists of all output types except P_Y^*, whereas the ordinary decoding regions only contain the output types close[6] to P_Y^*.

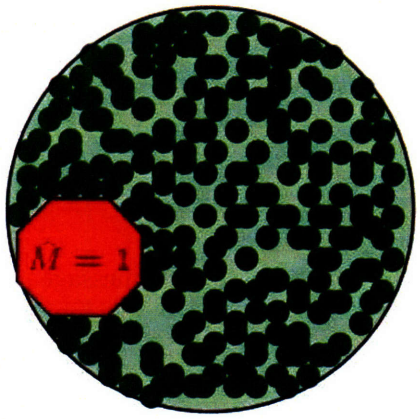


Figure 3-2: Avoiding missed-detection

The utility of this result is two fold: first, the optimality of such a simple scheme

[6]To be precise, this is the following set of output types Q

$$\{\mathsf{Q}: \max_{y\in\mathcal{Y}} \|\mathsf{Q}(y) - P_Y^*(y)\| \leq \sqrt[4]{1/n}\}$$

This set of Q close to P_Y^* is denoted by $[P_Y^*]$.

was not obvious before; second, protecting a single special message can be a key building block for many other problems when some feedback is available, for example[7] [30, 31]. We also use this building block in Chapter 4.

BSC Case Study: Consider a binary symmetric channel of capacity .9 bits/symbol. Its crossover probability is around 0.012, which implies its red alert exponent equals 1.48. It means even for modest codelength like 2000, as in some Ethernet standards, the missed-detection probability of the special message is around 10^{-1250}! Thus missed-detection probability for the special message is orders of magnitude[8] smaller than the overall error probability at these codelengths, which is around 10^{-15} at best. This example also demonstrates how a positive exponent translates to extremely small error probabilities. Although these calculations provide good guidelines for system design, they should taken with a pinch of salt. It is because these error probabilities in practice may be dominated by other reasons such as outages.

Also remember that the scheme to achieve this error probability was simple—just append your favorite capacity-achieving code (say an LDPC code) with the special codeword $(x_r, x_r, \cdots, x_r)$. The decoder applies its favorite decoding rule between ordinary messages (say message-passing) if the output sequence is typical and chooses the special message otherwise.

3.2.3 Many special messages

Now consider that instead of a single special message, exponentially many of the total $\doteq e^{nC}$ messages are special. Let $\mathcal{M}_s^{(n)} \subseteq \mathcal{M}^{(n)}$ denote this set of special messages with a given rate r:

$$\lim_{n\to\infty} \frac{\log|\mathcal{M}_s^{(n)}|}{n} = r \tag{3.13}$$

Thus the number of special messages is $\doteq e^{nr}$. For example, $\mathcal{M}_s^{(n)}$ may be equal to $\{1, 2, \ldots, e^{nr}\}$. The best missed-detection exponent, achievable simultaneously for all

[7]Thanks to Anant Sahai, who recently pointed us to these works.

[8]Even assuming a symbol per pico-second, this probability means one missed-detection in several centuries.

these special messages, is denoted by $E_{\rm md}(r)$.

Definition 9 *For a capacity-achieving sequence $\mathcal{Q}$, the missed-detection exponent for special messages in $\mathcal{M}_s^{(n)} \subseteq \mathcal{M}^{(n)}$ is defined as:*

$$E_{md,\mathcal{Q}} \equiv \liminf_{n\to\infty} \frac{-\log \max_{i\in\mathcal{M}_s^{(n)}} \Pr\left[\hat{M}\neq i|M=i\right]}{n}.$$

where $|\mathcal{M}_s^{(n)}| \doteq e^{nr}$ as in (3.13). We define $E_{md}(r) \equiv \sup_{\mathcal{Q}} E_{md,\mathcal{Q}}$.

Essentially, $E_{\rm md}(r)$ is the best value for which the missed-detection probability of every special message is $\doteq \exp(-nE_{\rm md}(r))$ or smaller. Note that if the only messages in the code are these $\doteq e^{nr}$ special messages (instead of $|\mathcal{M}^{(n)}| \doteq e^{nC}$ total messages), their best missed-detection exponent equals the classical exponent $E(r)$.

Theorem 10 $E_{md}(r) = E(r) \quad \forall\, r \in [0, \mathrm{C})$.

Thus we can communicate reliably at capacity and still protect the special messages as if we are only communicating the special messages. Note that the classical error exponent $E(r)$ is yet unknown for rates below critical rate (except zero rate). Nonetheless, this theorem says that whatever $E(r)$ can be achieved for only $\doteq e^{nr}$ messages, can still be achieved when there are $\doteq e^{nC}$ additional ordinary messages requiring reliable communication. Thus two points on the optimal error exponent curve $E(R)$ can be achieved simultaneously. The first point is at capacity corresponding to the ordinary messages and the second point is at rate r corresponding to the special messages. Using a different approach, Csiszár had shown a closely related result in[9] [35]. It showed that multiple points on the random coding exponent curve can be achieved simultaneously.

Optimal strategy: Start with an optimal code-book for $\doteq e^{nr}$ messages which achieves error exponent $E(r)$. These codewords are used for the special messages. Now the ordinary codewords are added using random coding. The ordinary codewords

[9]This paper with a somewhat unrelated name was recently pointed out to us by Pulkit Grover of UC Berkeley.

which land close to a special codeword may be discarded without essentially any effect on the rate of communication. At the decoder, a two-stage decoding rule is employed. The first stage decides that some special codeword was sent if at least one of the special codewords is 'close enough' to the received sequence. Otherwise, the first stage decides that an ordinary codeword was sent. Depending on the first stage decision, the second stage ignores all codewords of one kind and applies ML decoding to the rest.

The overall missed-detection exponent $E_{\text{md}}(r)$ is bottle-necked by the second stage errors. This is because the first-stage error exponent is essentially the sphere-packing exponent $E_{\text{sp}}(r)$, which is never smaller than the second stage error exponent $E(r)$.

Intuitive Interpretation: This means that we can start with a code of $\doteq e^{nr}$ messages, where the decoding regions are large enough to provide a missed-detection exponent of $E(r)$. Consider the balls around each codeword with sphere-packing radius (see Fig. 3-3(a)). For each message, the probability of going outside its ball decays exponentially with the sphere-packing exponent.

Although, these $\doteq e^{nr}$ balls fill up most of the output space, there are still some cavities left between them. These small cavities can still accommodate $\doteq e^{nC}$ typical noise balls for the ordinary messages (see Fig. 3-3(b)), which are much smaller than the original $\doteq e^{nr}$ balls. This is analogous to filling sand particles in a box full of large boulders. This theorem is like saying that the number of sand particles remains unaffected (exponentially) in spite of the large boulders although the boulders fill up all the space exponentially.

Remark for $r = 0$: It is worth commenting on this case, which corresponds to sub-exponentially many special messages. The best missed-detection exponent $E_{\text{md}}(0) = E(0)$ is smaller than the best missed-detection exponent $E_{\text{md}} = \mathsf{E}_{\mathsf{Red}}$ for a single special message. Consider the BSC for example where $\mathsf{E}_{\mathsf{Red}}$ equals the sphere-packing bound at rate 0 and is strictly larger than $E(0)$, which equals [3] the expurgated bound at rate 0. To understand this discontinuous behavior of missed-detection exponents, note that missed detections in case of sub-exponentially many special messages are

(a) Exponent optimal code (b) Achieving capacity

Figure 3-3: "There is always room for capacity!"

dominated by errors between different special codewords. Obviously, these errors are absent for the case of a single special message.

3.2.4 Allowing erasures

In some situations, A decoder may be allowed to declare an erasure when it is not sure about the transmitted message. These erasure events are not counted as errors and are usually followed by a retransmission using a decision feedback protocol like Hybrid-ARQ. This subsection extends the earlier result for $E_{\text{md}}(r)$ when such erasures are allowed.

In decoding with erasures, in addition to the message set $\mathcal{M}$, the decoder can map the received sequence Y^n to a virtual message called "erasure". Let P_{erasure} denote the average erasure probability of a code.

$$P_{\text{erasure}} = \Pr\left[\hat{M} = \text{erasure}\right]$$

Previously when there were no erasures, errors were not detected. For errors and erasures decoding, erasures will be detected errors, and the remaining errors will be undetected errors. P_e will denote the undetected error probability. Thus the undetected error probability (averaged over messages or conditioned on a message) is

respectively given by

$$P_e = \Pr\left[\hat{M} \neq M, \hat{M} \neq \text{erasure}\right] \quad \text{and} \quad P_e(i) = \Pr\left[\hat{M} \neq M, \hat{M} \neq \text{erasure} \middle| M = i\right]$$

An infinite sequence $\mathcal{Q}$ of block codes with errors and erasures decoding is called *reliable*, if its average error probability and average erasure probability, both vanish with n.

$$\lim_{n\to\infty} P_e^{(n)} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} P_{\text{erasure}}^{(n)} = 0 \tag{3.14}$$

If the erasure probability is small, then the average number of retransmissions needed is also small. Hence this condition of vanishingly small $P_{\text{erasure}}^{(n)}$ ensures that the effective data-rate of a decision feedback protocol remains unchanged in spite of retransmissions. We again restrict to reliable $\mathcal{Q}$ whose rate $R_{\mathcal{Q}}$ equals C.

For such decision-feedback (DF) scenarios, we could now redefine all previous exponents for reliable codes with erasure decoding. For example, on similar lines with $E_{\text{md}}(r)$, let us define $E_{\text{md},\mathcal{Q}}^{\text{df}}(r)$ as the best missed-detection exponent achievable uniformly over the special messages. Intuitively, we want to have large "gray regions" around the special codewords when an erasure is declared. In contrast to a large decoding region, landing in this large gray region causes an erasure but avoids undetected errors.

Definition 11 *For a given $r < \mathrm{C}$, let $E_{md,\mathcal{Q}}^{\text{df}}(r)$ denote the missed-detection exponent of a capacity-achieving sequence $\mathcal{Q}$ which is achieved uniformly over messages in $\mathcal{M}_s^{(n)} \subseteq \mathcal{M}^{(n)}$.*

$$E_{md,\mathcal{Q}}^{\text{df}}(r) = \liminf_{n\to\infty} \frac{-\log \max_{i\in\mathcal{M}_s^{(n)}} \Pr\left[\hat{M}\neq i, \hat{M}\neq erasure | M=i\right]}{n}. \tag{3.15}$$

where $|\mathcal{M}_s^{(n)}| \doteq e^{nr}$. Now define $E_{md}^{\text{df}}(r) = \sup_{\mathcal{Q}} E_{md,\mathcal{Q}}^{\text{df}}(r)$.

The next theorem shows that compared to $E_{\text{md}}(r)$ in the no-erasure case, allowing erasures increases the missed-detection exponent for r below critical rate[10].

[10]In all the previous problems, the provision of erasures with vanishing probability does not improve the achievable exponents. This implies that decision feedback protocols such as Hybrid-ARQ cannot improve E_{b} and E_{md} by allowing erasures.

Theorem 12

$$E_{md}^{\mathsf{df}}(r) \geq E_{sp}(r) \quad \forall\, r \in [0, \mathrm{C}).$$

The coding strategy here is similar to the no-erasure case. We first start with an erasure code in [8] for special messages. Then we add randomly generated ordinary codewords. Again a two-stage decoding is performed where the first stage decides between ordinary and special codewords using a threshold distance. If this first stage chooses special codewords, the second stage applies the errors-and-erasures decoding rule in [8] amongst special codewords. Otherwise, the second stage chooses the ML ordinary codeword.

The overall missed-detection exponent $E_{\mathrm{md}}^{\mathsf{df}}(r)$ is bottle-necked by the first stage errors. This is because the first-stage error exponent $E_{\mathrm{sp}}(r)$ is smaller than the second stage error exponent $E_{\mathrm{sp}}(r) + C - r$. This is in contrast to the case without erasures.

3.3 UEP Exponents for Block Codes: Proofs

Let us first prove that $E_{\mathrm{b}} = 0$ for even a single bit.

3.3.1 Proof of Theorem 5

Proof In order to prove that $E_{\mathrm{b}} = 0$, we will first show that any capacity-achieving sequence $\mathcal{Q}$ with $E_{\mathrm{b},\mathcal{Q}}$ can be used to construct a capacity-achieving sequence $\mathcal{Q}'$, whose elements are all fixed composition codes with $E_{\mathrm{b},\mathcal{Q}'} = \frac{E_{\mathrm{b},\mathcal{Q}}}{2}$. In the second part, we complete the proof by showing that $E_{\mathrm{b},\mathcal{Q}'} = 0$ for any capacity-achieving sequence which contains only fixed composition codes.

The proof of the second part is heavy in calculations, but the main idea is the "blowing up lemma" [4]. Conventionally, this lemma is used for strong converses for various capacity theorems. It is also worth mentioning that the conventional converse techniques like Fano's inequality are not sufficient to prove this result. Intuitively, the blowing up lemma implies that if we try to add slight extra thickness to the left cluster in Fig. 3-1, it blows up to occupy almost all the output space. This strange

phenomenon in high dimensional spaces leaves no room for the right cluster to fit. The infeasibility of adding even a slight extra thickness implies zero error exponent for the special bit.

Conversion to fixed composition codes: Consider a general capacity-achieving sequence $\mathcal{Q}$. Let the codebooks be arranged such that the top-half codewords represent messages where the special bit is 0, i.e., messages of the form $M = (0, M_2)$. Similarly, the bottom half represents messages of the form $M = (0, M_2)$.

For the length n code from this code sequence, let $\mathcal{A}$ denote the set of the top $7/8$ of the codewords according to increasing conditional error probability. The conditional error probability for codewords in $\mathcal{A}$ will be at most $8P_e{}^{(n)}$, where $P_e{}^{(n)}$ denotes the overall average error probability of the original code.

Now let $\mathcal{B}$ denote the set of the top $7/8$ of the codewords according to increasing error probability for the special bit. Again, the special bit error probability for every codeword in $\mathcal{B}$ will be at most $8\Pr\left[\hat{M}_1 \neq M_1\right]$, where $\Pr\left[\hat{M}_1 \neq M_1\right]$ denotes the average error probability for the special bit.

The number of codewords in the set $\mathcal{A} \cap \mathcal{B}$ will be at least $6/8$ of the original code. Hence from the original code, at least half the top half and and half the bottom half is contained in $\mathcal{A} \cap \mathcal{B}$. We will keep only half the codewords in each half which were contained in $\mathcal{A} \cap \mathcal{B}$ and expurgate all other codewords from the original code. This also ensures the size of each half is the same after expurgation.

This expurgated code will achieve capacity since its size is $1/4$ of the original code—insignificant in the exponential scale. Moreover, every codeword in it will achieve special bit error probability exponent $E_{\mathrm{b},\mathcal{Q}}$. Let us define the message set of this expurgated code as $\mathcal{M}^{(n)} = \{0,1\} \times \mathcal{M}_2^{(n)}$.

In this expurgated code, if we group the codewords for the messages of the form $M = (0, M_2)$ according to their empirical distribution at least one of the groups will have more than $\frac{|\mathcal{M}_2^{(n)}|}{(n+1)^{|\mathcal{X}|}}$ messages. This is because the number of different empirical distributions for elements of $\mathcal{X}^n$ is less than $(n+1)^{|\mathcal{X}|}$. Let us choose the first $\frac{|\mathcal{M}_2^{(n)}|}{(n+1)^{|\mathcal{X}|}}$ of the codewords of the this most crowded type of the first half and denote them by $\bar{x}_A^n(\cdot)$ and throw away all the other codewords. We can do the same for the messages

of the form $M = (1, M_2)$ and denote the corresponding codewords by $\bar{x}_B^n(\cdot)$.

Now let us consider the following length $2n$ code with message set $\mathcal{M}^{(2n)} = \{0,1\} \times \mathcal{M}_2^{(n)} \times \mathcal{M}_3^{(n)}$ where $\mathcal{M}_2^{(n)} = \mathcal{M}_3^{(n)} = \left\{1, 2, \ldots, \frac{|\mathcal{M}_2^{(n)}|}{(n+1)^{|\mathcal{X}|}}\right\}$. If $M = (0, M_2, M_3)$ then $\bar{x}(M) = \bar{x}_A^n(M_2)\bar{x}_B^n(M_3)$. That is, concatenate a codeword in the top half with a codeword in the bottom half of the original code. These codewords are used when the special bit equals 0. Similarly if $M = (1, M_2, M_3)$ then $\bar{x}(M) = \bar{x}_B^n(M_2)\bar{x}_A^n(M_3)$. That is, concatenate a codeword in the bottom half with a codeword in the top half of the original code. These codewords are used when the special bit equals 0.

This construction is analogous to a length two code of two codewords, where codeword "01" conveys message 0 and codeword "10" conveys message 1. Our construction replaces the "0" and "1" (respectively) by codewords in the top half and the bottom half of our original code.

The decoder of this new code of length $2n$ will separately decode y^n and y_{n+1}^{2n} using two copies of a decoder for the original length n code. If the concatenation of these decoded codewords (of length n each) corresponds to a valid codeword for some $i \in \mathcal{M}^{(2n)}$, then $\hat{M} = i$. Otherwise, assume an error and decode to an arbitrary message. This happens when the two decoded codewords of length n correspond to the same half in the original code.

By the union bound, the overall error probability of this length $2n$ code is at most twice the overall error probability of the original code. Furthermore, bit error exponent of the new code is half the bit error exponent of the original code. This is because now block length is twice the old block length.

Thus using these codes one can obtain a capacity-achieving sequence $\mathcal{Q}'$ whose every element is a fixed composition code and $E_{\mathrm{b},\mathcal{Q}'} = \frac{E_{\mathrm{b},\mathcal{Q}}}{2}$. In the following discussion we will focus on capacity achieving $\mathcal{Q}$'s whose members are fixed composition codes and show that $E_{\mathrm{b},\mathcal{Q}} = 0$ for any such code. The discussion above will then imply $E_{\mathrm{b}} = 0$.

$E_b = 0$ for fixed composition codes: We will call the empirical conditional distribution of a given output sequence y^n, given the codeword $\bar{x}^n(i)$, the conditional type of y^n given message i and denote it by $\mathsf{V}(y^n, i)$. Furthermore we will call the set

of y^n's whose conditional type with message i is V as the V-shell of i and denote it by $\mathsf{T}_V(i)$. Similarly we will denote the set of output sequences y^n with the empirical distribution Q_Y, by T_{Q_Y}.

For codelength n, we denote the fixed empirical distribution of the codewords by $P_X^{(n)}$ and the corresponding output distribution by $P_Y^{(n)}$, i.e.

$$P_Y^{(n)}(\cdot) = \sum_{x \in \mathcal{X}} W_{Y|X}(\cdot|x) P_X^{(n)}(x).$$

Whenever the value of n is unambiguous from the context, we simply use P_X and P_Y. Furthermore $\mathbb{P}_Y^n(\cdot)$ will denote the probability measure on $\mathcal{Y}^n$ such that

$$\mathbb{P}_Y^n(y^n) = \prod_{t=1}^{n} P_Y(y_t).$$

We will denote the set of y^n such that $\hat{M}_1 = 0$ and $\mathsf{V}(y^n, \hat{M}(y^n)) = V$ by $\mathcal{S}_{0,V}^{(n)}$.

$$\mathcal{S}_{0,V}^{(n)} \equiv \{y^n : \mathsf{V}(y^n, \hat{M}(y^n)) = V \text{ and } \hat{M}(y^n) = (0, j) \text{ for some } j \in \mathcal{M}_2\} \qquad (3.16)$$

Note that since for each y^n there is a unique $\hat{M}(y^n)$ and for each y^n and message $i \in \mathcal{M}$ there is unique $\mathsf{V}(y^n, i)$; each y^n belongs to a unique $\mathcal{S}_{0,V}^{(n)}$ or $\mathcal{S}_{1,V}^{(n)}$, i.e. $\mathcal{S}_{0,V}^{(n)}$'s and $\mathcal{S}_{1,V}^{(n)}$'s are disjoint sets that collectively cover the set $\mathcal{Y}^n$.

Let us define the typical neighborhood of $W_{Y|X}$ as $[W]$

$$[W] \equiv \{V_{Y|X} : |V_{Y|X}(y|x) P_X^{(n)}(x) - W_{Y|X}(y|x) P_X^{(n)}(x)| \leq \sqrt[4]{1/n} \quad \forall x, y\} \qquad (3.17)$$

Let us denote the union of all $\mathcal{S}_{0,V}^{(n)}$'s for typical V's by $\mathcal{S}_0^{(n)} = \bigcup_{V \in [W]} \mathcal{S}_{0,V}^{(n)}$. We will establish the following inequality (3.18) later. Let us assume for the moment that it holds.

$$\begin{aligned} & \mathbb{P}_Y^n\left(\mathcal{S}_0^{(n)}\right) \geq e^{n(R^{(n)} - C)} \left(\tfrac{1}{2} - \tfrac{|\mathcal{X}||\mathcal{Y}|}{8\sqrt{n}} - P_e\right) \\ \Rightarrow \quad & \tfrac{-\log \mathbb{P}_Y^n\left(\mathcal{S}_0^{(n)}\right)}{n} \leq C - R^{(n)} - \tfrac{1}{n} \log\left(\tfrac{1}{2} - \tfrac{|\mathcal{X}||\mathcal{Y}|}{8\sqrt{n}} - P_e\right) \end{aligned} \qquad (3.18)$$

Since $P_e^{(n)} \to 0$ and $\frac{|\mathcal{X}||\mathcal{Y}|}{8\sqrt{n}} \to 0$, the last term on the RHS above tends to $\frac{1}{n}\log(\frac{1}{2})$ which vanishes with n. Hence the RHS above vanishes with n because $R^{(n)}$ tends to C for a capacity-achieving sequence $\mathcal{Q}$. This shows that $\mathbb{P}_Y^n\left(\mathcal{S}_0^{(n)}\right)$ does not approach 0 exponentially. This fact allows us to apply blowing up lemma [4, Ch. 1, Lemma 5.4, page 92] to $\mathcal{S}_0^{(n)}$. It implies that for any capacity-achieving sequence $\mathcal{Q}$, there exists a sequence of (ℓ_n, η_n) pairs satisfying $\lim_{n\to\infty}\eta_n = 1$ and $\lim_{n\to\infty}\frac{\ell_n}{n} = 0$ such that

$$\mathbb{P}_Y^n\left(\Gamma^{\ell_n}(\mathcal{S}_0^{(n)})\right) \geq \eta_n \tag{3.19}$$

where $\Gamma^{\ell_n}(A)$ is the set of all y^n's which differs from one or more elements of A in at most ℓ_n places. Clearly one can repeat the same argument for $\Gamma^{\ell_n}(\mathcal{S}_1^{(n)})$ and thus

Note that if $y^n \in \Gamma^{\ell_n}(\mathcal{S}_1^{(n)})$, then there exist a $\tilde{y}^n \in \mathsf{T}_{P_Y}$ that differs from y^n in at most $(|\mathcal{Y}||\mathcal{X}|n^{3/4} + \ell_n)$ places[11]—the $|\mathcal{Y}||\mathcal{X}|n^{3/4}$ term arises due to our definition of the typical neighborhood $[W]$ and the ℓ_n term arises due to the blowing up lemma. Thus we can upper bound its probability by,

$$y^n \in \Gamma^{\ell_n}(\mathcal{S}_1^{(n)}) \Rightarrow \mathbb{P}_Y^n(y^n) \leq e^{-nH(P_Y)-(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda} \tag{3.20}$$

where $\lambda = \min_{x,y} W_{Y|X}(y|x)$. Thus we have

$$|\Gamma^{\ell_n}(\mathcal{S}_0^{(n)}) \bigcap \Gamma^{\ell_n}(\mathcal{S}_1^{(n)})| \geq (2\eta_n - 1)e^{nH(P_Y)+(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda} \tag{3.21}$$

Note that we also know that, if $y^n \in \Gamma^{\ell_n}(\mathcal{S}_0^{(n)}) \bigcap \Gamma^{\ell_n}(\mathcal{S}_1^{(n)})$ then there exist an $\tilde{y}^n \in \mathsf{T}_W(i)$ for a i of the form $i = (0, M_2)$ which differs from y^n in at most $(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)$ places[12]. Thus we can lower bound the probability of y^n under the hypothesis $M_1 = 0$ as follows

$$\Pr[y^n | M_1 = 0] \geq e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda}$$

[11] A small caveat here is due to the integer constraint, T_{P_Y} might actually be an empty set. If so we can make a similar argument for the U_Y^* which minimizes $\sum_j |U_Y(j) - P_Y(j)|$. However this technicality is inconsequential.

[12] The integer constraint here is also inconsequential.

Clearly the same holds for $M_1 = 1$ too and thus

$$\Pr\left[y^n \middle| M_1 = 1\right] \geq e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda}$$

Consequently

$$\begin{aligned}
\Pr\left[\hat{M}_1 \neq M_1\right] &\geq \sum_{y^n} \tfrac{1}{2}\min(\Pr\left[y^n \middle| M_1 = 0\right], \Pr\left[y^n \middle| M_1 = 0\right]) \\
&\geq \tfrac{1}{2} \sum_{y^n \in \Gamma^{\ell_n}(\mathcal{S}_1^{(n)}) \bigcap \Gamma^{\ell_n}(\mathcal{S}_1^{(n)})} e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda} \\
&\geq \tfrac{1}{2}(2\eta_n - 1)e^{nH(P_Y)+(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda} e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda} \\
&= (\eta_n - \tfrac{1}{2})e^{n(I(P_X,W)-R^{(n)})+2(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda} \\
&\geq (\eta_n - \tfrac{1}{2})\tfrac{1}{2}e^{-nR^{(n)}P_e{}^{(n)}+2(|\mathcal{Y}||\mathcal{X}|n^{3/4}+\ell_n)\log\lambda}
\end{aligned}$$

where in the last step we have used Fano's inequality. Thus we have

$$\lim_{n\to\infty} \frac{-\log\Pr\left[\hat{M}_1 \neq M_1\right]}{n} = 0$$

The only thing remaining is to establish inequality (3.18). One can write the error probability of the n^{th} code of $\mathcal{Q}$ as

$$\begin{aligned}
P_e{}^{(n)} &= \sum_{i\in\mathcal{M}^{(n)}} \tfrac{1}{\mathcal{M}} \sum_{y^n\in\mathcal{Y}^n} (1 - \mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}}) \Pr\left[y^n \middle| M = i\right] \\
&= \sum_{i\in\mathcal{M}} \sum_{V} \sum_{y^n\in\mathsf{T}_V(i)} (1 - \mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}}) e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)+R^{(n)})} \\
&= \sum_{V} e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)+R^{(n)})} \sum_{i\in\mathcal{M}} \sum_{y^n\in\mathsf{T}_V(i)} (1 - \mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}}) \\
&= \sum_{V} e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)+R^{(n)})} (Q_{0,V} + Q_{1,V}) \qquad (3.22)
\end{aligned}$$

where $Q_{k,V} = \sum_{\substack{i=(k,j)\\ j\in\mathcal{M}_2}} \sum_{y^n\in\mathsf{T}_V(i)} (1 - \mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}})$ for $k = 0, 1$.

Note that $Q_{k,V}$ is the sum, over the messages i for which $M_1 = k$, of the number

of the elements in $\mathsf{T}_V(i)$ that are not decoded to message i. In a sense it is a measure of the overlap in the V-shells of different codewords. We will use equation (3.22) to establish lower bounds on $\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right)$'s.

Let us denote $\sum_x P_X(x)V_{Y|X}(\cdot|x)$ by $(PV)_Y(\cdot)$, then all elements of $\mathcal{S}_{0,V}^{(n)}$ have the same probability under $\mathbb{P}_Y^n(\cdot)$. Thus

$$\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right) = |\mathcal{S}_{0,V}^{(n)}|e^{-n(D((PV)_Y(\cdot)\|P_Y(\cdot))+H((PV)_Y))} \tag{3.23}$$

As a result of convexity of the KL divergence we get

$$\begin{aligned}
\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right) &\geq |\mathcal{S}_{0,V}^{(n)}|e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H((PV)_Y))} \\
&= |\mathcal{S}_{0,V}^{(n)}|e^{-nI(P_X,V_{Y|X})}e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))} \\
&\geq |\mathcal{S}_{0,V}^{(n)}|e^{-nC}e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))}
\end{aligned}$$

Note that

$$|\mathcal{S}_{0,V}^{(n)}| = |\mathcal{M}_2^{(n)}| \cdot |\mathsf{T}_V(i)| - Q_{0,V} = \tfrac{1}{2}|\mathsf{T}_V(i)|e^{nR} - Q_{0,V} \tag{3.24}$$

Using $|\mathcal{M}_2^{(n)}| = \frac{e^{nR^{(n)}}}{2}$ we get,

$$\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right) \geq e^{-nC}\left(\tfrac{1}{2}|\mathsf{T}_V(i)|e^{nR^{(n)}} - Q_{0,V}\right)e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))} \tag{3.25}$$

Recall that $\mathcal{S}_{0,V}^{(n)}$'s are disjoint and consequently the inequality (3.25) implies,

$$\begin{aligned}
\mathbb{P}_Y^n\left(\mathcal{S}_0^{(n)}\right) &\geq \sum_{V\in[W]} e^{-nC}\left(\tfrac{1}{2}|\mathsf{T}_V(i)|e^{nR^{(n)}} - Q_{0,V}\right)e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))} \\
&\geq e^{n(R^{(n)}-C)}\left(\sum_{V\in[W]_{\delta_n}} \tfrac{1}{2}|\mathsf{T}_V(i)|e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))} - P_e\right) \\
&\geq e^{n(R^{(n)}-C)}\left(\tfrac{1}{2}(\sum_{V\in[W]_{\delta_n}}\sum_{y^n\in\mathsf{T}_V(i)} \Pr[y^n|\,M=i]) - P_e\right) \\
&\geq e^{n(R^{(n)}-C)}\left(\tfrac{1}{2} - \tfrac{|\mathcal{X}||\mathcal{Y}|}{8\sqrt{n}} - P_e\right)
\end{aligned}$$

where the last inequality follows from the Chebyshev's inequality. •

3.3.2 Proof of Theorem 8

Achievability: $E_{\mathbf{md}} \geq \mathsf{E}_{\mathsf{Red}}$

Proof For each block-length n, the special message is sent with the length-n repetition sequence $\bar{x}^n(1) = (x_r, x_r \cdots, x_r)$ where x_r is an input letter satisfying

$$D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|x_r)\right) = \max_{x \in \mathcal{X}} D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|x)\right) = \mathsf{E}_{\mathsf{Red}}.$$

The remaining $|\mathcal{M}^{(n)}| - 1$ ordinary codewords are generated randomly and independently of each other using a capacity achieving input distribution P_X^* i.i.d. over time.

Let us denote the empirical distribution of a particular output sequence y^n by $\mathsf{Q}_y(y^n)$. The receiver will decide that the special message was sent only when the output distribution is not close to P_Y^*. More precisely,

$$\mathcal{G}(1) = \{y^n : \|\mathsf{Q}_y(y^n)(i) - P_Y^*(i)\| \geq \sqrt[4]{1/n} \text{ for some } \forall i \in \mathcal{Y}\}$$

Let us denote the set of output sequences close to P_Y^* by $[P_Y^*]$. Since there are at most $(n+1)^{|\mathcal{Y}|}$ different empirical output distribution for elements of $\mathcal{Y}^n$ we get,

$$\Pr\left[y^n \notin \mathcal{G}(1) | \, M = 1\right] \leq (n+1)^{|\mathcal{Y}|} e^{-n \min_{Q_Y \in [P_Y^*]} D\left(Q_Y(\cdot) \| W_{Y|X}(\cdot|x_r)\right)}$$

Thus $\lim_{n\to\infty} \frac{-\log \Pr\left[y^n \notin \mathcal{G}(1) | M=1\right]}{n} = D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|x_r)\right) = \mathsf{E}_{\mathsf{Red}}$.

Now the only thing we are left to prove is that we can have low enough probability for the remaining messages. For doing that we will first calculate the average error probability of the following random code ensemble. Each entry of the code-book will be generated by using a capacity achieving input distribution P_X^*, independent of all other entries of the codebook. Thus the (conditional) error probability will be the same for all $i \neq 1$ in $\mathcal{M}^{(n)}$. Hence without loss of generality, let us calculate the error

probability of the message $M = 2$.

Assuming that the second message was transmitted, $\Pr[y^n \in \mathcal{G}(1) | M = 2]$ is vanishingly small. This is because the output distribution for the random ensemble for ordinary codewords is i.i.d. P_Y^*. Chebyshev's inequality guarantees that the probability of the output type being outside a $\sqrt[4]{1/n}$ ball around P_Y^*, i.e., outside $[P_Y^*]$, is of the order $\sqrt{1/n}$.

If the second message was transmitted, $\Pr[y^n \in \cup_{i>2}\mathcal{G}(i) | M = 2]$ is vanishingly small by the standard random coding argument for achieving capacity [1].

Thus for any $P_e > 0$, for all large enough n, the average error probability of the code ensemble is smaller than P_e. Hence we will have at least one code with that P_e. For that code, the error probability of at least half of the codewords will be less then $2P_e$. •

Converse: $E_{\mathbf{md}} \leq \mathsf{E}_{\mathsf{Red}}$

In the next chapter, we will prove that even with feedback and variable decoding time, the missed-detection exponent of a single special message is at most $\mathsf{E}_{\mathsf{Red}}$. Since feedback can only help, $E_{\mathrm{md}} \leq \mathsf{E}_{\mathsf{Red}}$.

3.3.3 Proof of Theorem 10

Achievability: $E_{\mathbf{md}} \geq E(r)$

Now we prove the achievability part of missed-detection exponents for e^{nr} special messages. The case of a general DMC is considered here, although, some readers may prefer to first read the proof for a BSC in Appendix C. On similar lines to [5], that analysis is based on Hamming distances, which could make it easier to visualize. The general DMC considered here essentially replaces those Hamming distances with KL divergences.

Proof Let us start by choosing the codewords for special messages.

Special codewords: At any given block length n, we start with a optimum code-

book (say $\mathcal{C}_{special}$) for $|\mathcal{M}_s^{(n)}|$ messages. Such an optimum code-book achieves error exponent $E(r)$ for every message.

$$\Pr\left[\hat{M} \neq i | M = i\right] \doteq e^{-nE(r)} \qquad \forall i \in \mathcal{M}_s$$

Since there are at most $(n+1)^{|\mathcal{X}|}$ different types, there is at least one type, say P_X, which has $\frac{|\mathcal{M}_s^{(n)}|}{(1+n)^{|\mathcal{Y}|}}$ or more codewords. Throw away all other codewords from $\mathcal{C}_{special}$ and denote call the remaining fixed composition code-book as $\hat{\mathcal{C}}_{special}$. Code-book $\hat{\mathcal{C}}_{special}$ is used for transmitting the special messages. Let the message set corresponding to these codewords be denoted as $\hat{\mathcal{M}}_s$

As shown in Fig. 3-3(a), let the ball for special message i be denoted by $\mathcal{B}_i$. These balls need not be disjoint. Now let $\mathcal{B}$ denote the union of these balls of all special messages.

$$\mathcal{B} = \bigcup_{i \in \hat{\mathcal{M}}_s} \mathcal{B}_i$$

If the output sequence Y^n lies in $\mathcal{B}$, the first stage of the decoder decides a special message was transmitted. The second stage then chooses the ML candidate amongst the messages in $\hat{\mathcal{M}}_s$.

Let us define $\mathcal{B}_i$ precisely now.

$$\mathcal{B}_i = \{y^n : \mathsf{V}(y^n, i) \in \mathcal{W}(r + \epsilon_n, P_X)\}$$

where $\mathcal{W}(r+\epsilon_n, P_X) = \{V_{Y|X} : D\left(V_{Y|X}(\cdot|X) \middle\| W_{Y|X}(\cdot|X) \middle| P_X\right) \leq E_{\text{sp}}(r+\epsilon_n; P_X)\}$ is the set of channel types within divergence $r+\epsilon_n$ from the actual channel $W_{Y|X}$. Here ϵ_n is a non-negative sequence which vanishes to zero with n. Recall that the sphere-packing exponent for input type P_X at rate r, $E_{\text{sp}}(r; P_X)$ is given by,

$$E_{\text{sp}}(r; P_X) = \min_{V_{Y|X}: D\left(V_{Y|X}(\cdot|X) \middle\| (PV)_Y(\cdot)|P_X\right) \leq r} D\left(V_{Y|X}(\cdot|X) \middle\| W_{Y|X}(\cdot|X) \middle| P_X\right)$$

The constraint for optimization above is that the mutual information of channel $V_{Y|X}$ under input distribution P_X is than r, since $(PV)_Y$ denotes the output distribution

of channel $V_{Y|X}(\cdot|\cdot)$ for input distribution P_X.

Ordinary codewords: The ordinary codewords will be chosen by random coding, i.i.d. P_X^* over time, where P_X^*. This is the same as Shannon's construction for achieving capacity. The random coding construction provides a simple way to show that in the cavity $\mathcal{B}^c$ (complement of $\mathcal{B}$), we can essentially fit enough typical noise-balls to achieve capacity. This will avoid the complicated task of carefully choosing the ordinary codewords and their decoding regions in the cavity space $\mathcal{B}^c$.

If the output sequence Y^n lies in the cavity $\mathcal{B}^c$, the first stage of the decoder decides an ordinary message was transmitted. The second stage then chooses the ML candidate from ordinary codewords.

Error analysis: First, consider the case that a special codeword $\bar{x}^n(i)$ is transmitted. By Stein's lemma and definition of $\mathcal{B}_i$, the probability of $Y^n \notin \mathcal{B}_i$ has exponent $E_{\text{sp}}(r+\epsilon_n; P_X)$. Hence the first stage error exponent is at least $E_{\text{sp}}(r+\epsilon_n; P_X)$.

Assuming correct first stage decoding, the second stage error exponent for special messages equals $E(r)$. Hence the effective error exponent for special messages is

$$\min\{E(r), E_{\text{sp}}(r+\epsilon_n; P_X)\}$$

Since $E(r)$ is at most [6] the sphere-packing exponent $E_{\text{sp}}(r; P_X)$ and ϵ_n vanishes to 0, the missed-detection exponent of each special message equals $E(r)$.

Now consider the situation of a uniformly chosen ordinary codeword being transmitted. We have to make sure the error probability is vanishingly small now. In this case, the output sequence distribution is i.i.d. P_Y^* for the random coding ensemble. The first stage decoding error happens y^n lies in $\bigcup \mathcal{B}_i$. Again by Stein's lemma, this

exponent for any particular $\mathcal{B}_i$ equals E_{o}:

$$\begin{aligned}
E_{\text{o}} &= \min_{V_{Y|X} \in \mathcal{W}} D\left(V_{Y|X}(\cdot|X) \middle\| P_Y^*(\cdot) \middle| P_X\right) \\
&= \min_{V_{Y|X} \in \mathcal{W}} D\left(V_{Y|X}(\cdot|X) \middle\| (PV)_Y(\cdot) \middle| P_X\right) + D\left((PV)_Y(\cdot) \| P_Y^*(\cdot)\right) \\
&\geq (r + \epsilon_n) + D\left((PV)_Y(\cdot) \| P_Y^*(\cdot)\right) \\
&\geq r + \epsilon_n
\end{aligned}$$

The first step follows since the actual output distribution in this case is i.i.d. P_Y^*. The second step follows by multiplying and dividing by $(PV)_y(\cdot)$ in the log terms in the summed to get $D\left(V_{Y|X}(\cdot|X) \middle\| P_Y^*(\cdot) \middle| P_X\right)$ The third step follows from the definition of $E_{\text{sp}}(r + \epsilon_n; q_X)$. The forth step is simply the non-negativity of the KL divergence.

Applying the union bound over the special messages, the probability of a first-stage decoding error after sending an ordinary message is at most $\doteq \exp(nr - nE_{\text{o}})$. We have already shown that $E_{\text{o}} \geq r + \epsilon_n$, which ensures that probability of first stage decoding error for ordinary messages is at most $\doteq e^{-n\epsilon_n}$ for the random coding ensemble. Recall that for the random coding ensemble, average error probability of the second-stage decoding also vanishes below capacity. To summarize, we have shown these two properties of the random coding ensemble:

1. Error probability of first stage decoding vanishes as $a^{(n)} \doteq \exp(-n\epsilon_n)$ with n when a uniformly chosen ordinary message is transmitted.

2. Error probability of second stage decoding (say $b^{(n)}$) vanishes with n when a uniformly chosen ordinary message is transmitted.

Since the first error probability is at most $4a^{(n)}$ for 75% of the codes in the random ensemble, and the second error probability is also at most $4b^{(n)}$ for some 75% of the codes, there exists a particular code which satisfies both these properties. The overall error probability for ordinary messages is at most $4(a^{(n)}+b^{(n)})$, which vanishes with n. We will use this particular code for the ordinary codewords. This de-randomization completes our construction of a reliable code for ordinary messages to be combined

with the code $\mathcal{C}_{special}$ for special messages. •

Converse: $E_{\mathbf{md}} \leq E(r)$

The converse argument for this result is obvious. Removing the ordinary messages from the code can only improve the error probability of the special messages. Even then, (by definition) the best missed-detection exponent for the special messages equals $E(r)$.

3.3.4 Proof of Theorem 12

Let us now address the case with erasures. In this achievability result, the first stage of decoding remains unchanged from the no-erasure case.

Proof We use essentially the same strategy as before. Let us start with a good code for $|\mathcal{M}_s^{(n)}|$ messages allowing erasure decoding. Forney had shown in [8] that an error exponent equal to $E_{\text{sp}}(r) + C - r$ is achievable while ensuring that the erasure probability vanishes with n. We can use that code for the special codewords. As before, for $Y^n \in \bigcup_i \mathcal{B}_i$, the first stage decides a special codeword was sent. Then the second stage applies the erasure decoding method in [8] amongst the special codewords.

With this decoding rule, when a special message is transmitted, the error probability of the two-stage decoding is bottle-necked by the first stage: its error exponent $E_{\text{sp}}(r+\epsilon_n)$ is smaller than that of the second stage $(E_{\text{sp}}(r)+C-r)$. Since ϵ_n vanishes to 0, the special messages can achieve $E_{\text{sp}}(r)$ as their missed-detection exponent.

The ordinary codewords are again generated i.i.d. P_X^*. If the first stage decides in favor of the ordinary messages, choose the ML ordinary codeword. If an ordinary message was transmitted, we can ensure a vanishing error probability as before by repeating earlier arguments for no-erasure case. •

Chapter 4

Unequal Error Protection Near Capacity: Feedback Case

In the last chapter, we analyzed UEP problems for fixed length block codes without feedback. In this chapter, we will revisit the same problems for variable-length block codes with perfect feedback. Again the focus is on situations where the data-rate is essentially at channel capacity. We will answer whether and how feedback can improve the achievable exponents in various UEP situations seen earlier.

We start by explaining variable-length block codes for channels with feedback in Section 4.1. Our results on UEP exponents are discussed in Section 4.2. Then Section 4.3 addresses message-wise UEP situations where special messages (like the reformat-disk command) demand protection against false-alarms instead of missed-detections. We first address false alarms with no feedback and then with full feedback. This discussion for false-alarms was postponed to the end of this chapter to avoid confusion with earlier results on missed-detection.

As in the previous chapter, after discussing each theorem, we briefly describe the optimal strategy. Further proof details can be found in Section 4.4, which contains proofs of the results in Section 4.2. Section 4.5 contains proofs for the false-alarm results in Section 4.3.

4.1 Variable-Length Block Codes with Feedback

A variable-length block code with feedback is composed of a coding algorithm and a decoding rule. The decoding rule determines the decoding time and the message that will be decoded then. Possible observations of the receiver can be seen as the leaves of a $|\mathcal{Y}|$-ary tree, as in [28]. In this tree, all nodes at length 1 from the root denote all possible $|\mathcal{Y}|$ outputs at time $t = 1$. All non-leaf nodes among these will split into further $|\mathcal{Y}|$ branches at the next time $t = 2$ and the branching of the non-leaf nodes will continue like this ever after. Each node of depth t in this tree corresponds to a particular sequence, y^t, i.e. a history of outputs until time t. The parent of node y^t is its prefix y^{t-1}. Leaves of this tree are like a prefix free source code, because the decision to stop and complete decoding should to be a causal decision. In other words the event $\{\tau = t\}$ will be measurable in the σ-field generated by Y^t. In addition, we have $\Pr[\tau < \infty] = 1$ and thus τ will be a Markov stopping time with respect to the received outputs.

The encoding algorithm on the other hand will assign an input letter, $X_{t+1}(Y^t; i)$, to each message, $i \in \mathcal{M}$, at each non-leaf node, Y^t, of this tree. The encoder stops transmission of a message when a leaf has been reached and the decoding is complete.

The codes we consider are block codes in the sense that transmission of each message (packet) will start only after the transmission of the previous one has ended. The error probability and rate of the code will be defined by

$$P_e = \Pr\left[\hat{M} \neq M\right] \qquad \text{and,} \qquad R = \frac{\log|\mathcal{M}|}{E[\tau]} \tag{4.1}$$

where $E[\tau]$ is the average decoding time when the messages are chosen uniformly from $\mathcal{M}$. A more thorough discussion of variable-length block codes with feedback can be found in [27], [28].

A reliable sequence of variable decoding time codes with feedback, $\mathcal{Q}$, will be any countably infinite collection of codes indexed by integers, such that

$$\lim_{\kappa \to \infty} P_e^{(\kappa)} = 0 \tag{4.2}$$

For defining the rate $R_{\mathcal{Q}}$ and exponent $E_{\mathcal{Q}}$ of such a sequence, simply replace block-length n for the no-feedback case by the average decoding time $E\left[\tau^{(\kappa)}\right]$ for the κ'th element of this code sequence.

$$R_{\mathcal{Q}} \equiv \lim_{\kappa\to\infty} \frac{\log|\mathcal{M}^{(\kappa)}|}{E\left[\tau^{(\kappa)}\right]} \quad \text{and} \quad E_{\mathcal{Q}} \equiv \liminf_{\kappa\to\infty} \frac{-\log P_e^{(\kappa)}}{E\left[\tau^{(\kappa)}\right]}$$

A *capacity-achieving sequence with feedback* will mean a reliable sequence of variable-length block codes with feedback whose rate equals C

It is worth noting the importance of our assumption that all the entries of the transition probability matrix $W_{Y|X}$ are positive. For any channel with a $W_{Y|X}$ which has one or more zero probability transitions, it is possible to have *error-free* codes that achieve capacity [27]. Then all the exponents discussed below become trivially infinite.

4.2 UEP at Capacity: Variable-Length Block Codes with Feedback

4.2.1 Special bit

Let us consider a capacity-achieving sequence $\mathcal{Q}$ whose message sets are of the form $\mathcal{M}^{(\kappa)} = \mathcal{M}_1 \times \mathcal{M}_2^{(\kappa)}$ where $\mathcal{M}_1 = \{0,1\}$. Then the error exponent of M_1, *i.e.*, the initial bit b_1, is defined as follows.

Definition 13 *Let a capacity-achieving sequence $\mathcal{Q}$ with feedback have message sets $\mathcal{M}^{(\kappa)}$ of the form $\mathcal{M}^{(\kappa)} = \mathcal{M}_1 \times \mathcal{M}_2^{(\kappa)}$ where $\mathcal{M}_1 = \{0,1\}$. The special bit error exponent is defined as*

$$E_{b,\mathcal{Q}}^{\mathsf{f}} = \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}_1 \neq M_1\right]}{E\left[\tau^{(\kappa)}\right]} \tag{4.3}$$

Then $E_b^{\mathsf{f}} = \sup_{\mathcal{Q}} E_{b,\mathcal{Q}}^{\mathsf{f}}$

Theorem 14 $E_b^{\mathsf{f}} = \mathsf{E}_{\mathsf{Red}}$.

Section 4.4 contains a detailed proof of this theorem. Recall that without feedback, the single bit could not achieve any positive error exponent at capacity, i.e., $E_{\mathrm{b}} = 0$. The following strategy shows how feedback connects message-wise UEP with bit-wise UEP: the strategy for protecting a special message becomes useful for protecting special bits. This special message is used to indicate incorrect decisions at the receiver (on similar lines of [30, 31]).

Optimal strategy: We achieve this exponent using the missed-detection exponent of $\mathsf{E}_{\mathsf{Red}}$ for a special message (see Fig. 4-1). This special message notifies the receiver when its tentative estimate of M_1 is incorrect. More specifically, we use a code of length $\kappa + \sqrt{\kappa}$ followed by decoding with erasures. The transmitter first transmits M_1 using a short repetition code of length $\sqrt{\kappa}$. If the temporary decision about M_1, called $\tilde{M}_1$, is correct after this repetition code, the transmitter transmits M_2 with a capacity-achieving code of length κ. If $\tilde{M}_1$ is incorrect after the repetition code, the transmitter will transmit the symbol x_r for κ symbols where x_r is the input letter i maximizing the $D\left(W_{Y|X}(\cdot|i) \middle\| P_Y^*(\cdot)\right)$. This is a special buzzer codeword to indicate an error in $\tilde{M}_1$.

If the decoder detects a buzzer over these κ symbols, an erasure is declared. The same message is retransmitted by repeating the same strategy afresh. If no buzzer is detected, the receiver uses an ML decoder to chose $\hat{M}_2$ and declares $\hat{M} = (\tilde{M}_1, \hat{M}_2)$.

The erasure probability is vanishingly small, which ensures $E[\tau]$ is essentially equal to $\kappa + \sqrt{\kappa}$. Hence the effective rate of communication still approaches capacity in spite of such retransmissions. A decoding error for M_1 happens only when the buzzer message is not detected. This probability is $\doteq \exp(-\kappa \mathsf{E}_{\mathsf{Red}}) \doteq \exp(-E[\tau]\, \mathsf{E}_{\mathsf{Red}})$.

Remark: Note that when the receiver detects a buzzer indicating incorrect $\tilde{M}_1$, the whole block is retransmitted instead of simply flipping the tentative $\tilde{M}_1$. This is because if we simply flip the earlier decision, the special bit will be in error even when $\tilde{M}_1$ is correct but the receiver erroneously detects a buzzer and flips $\tilde{M}_1$. Our retransmission based strategy avoids this drawback.

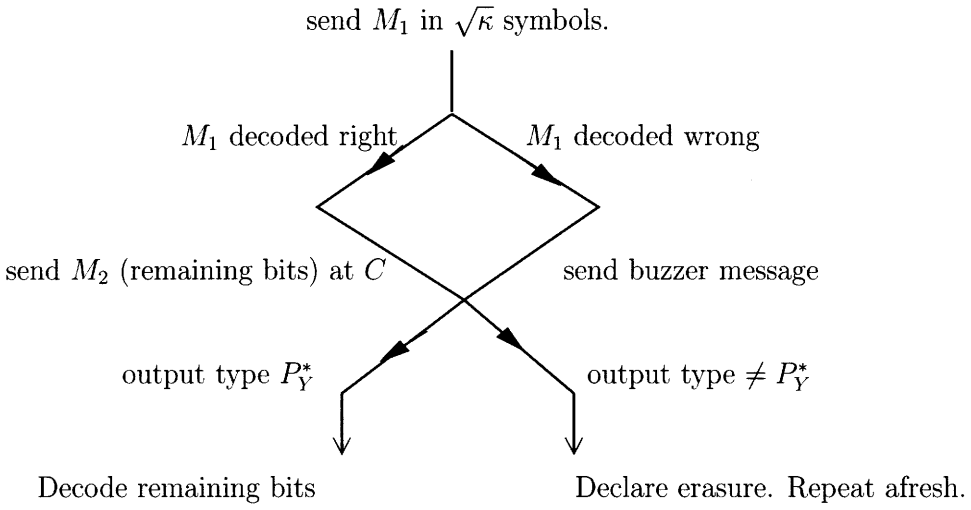


Figure 4-1: Sending a special bit using a special message

4.2.2 Many special bits

We now analyze the situation where instead of a single special bit, there are approximately $E[\tau]\, r/\ln 2$ special bits out of the total $E[\tau]\, C/\ln 2$ (approx.) bits. Here we again consider capacity-achieving sequences with feedback having message sets of the form $\mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)}$. Now unlike the previous subsection where the size of $\mathcal{M}_1^{(\kappa)}$ was fixed at 2, the size varies with the index of the code. We assume that the size of $\mathcal{M}_1^{(\kappa)}$ grows exponentially at rate r:

$$\lim_{\kappa\to\infty} \frac{\log|\mathcal{M}_1^{(\kappa)}|}{E\left[\tau^{(\kappa)}\right]} = r$$

This simply says that the rate of special bits equals r. It is worth noting at this point that even when the rate r of special bits is zero, the number of special bits might not be bounded, that is, $\lim_{\kappa\to\infty} |\mathcal{M}_1^{(\kappa)}|$ might be infinite. For example, $|\mathcal{M}_1^{(\kappa)}|$ could grow polynomially with $E\left[\tau^{(\kappa)}\right]$. The error exponent $E^{\mathrm{f}}_{\mathrm{bits},\mathcal{Q}}$ for the special bits at rate r is defined as follows,

Definition 15 *Let a capacity-achieving sequence $\mathcal{Q}$ with feedback have message sets*

$\mathcal{M}^{(\kappa)}$ *of the form* $\mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)}$, *where* $r_{\mathcal{Q}} = \lim_{\kappa\to\infty} \frac{\log|\mathcal{M}_1^{(\kappa)}|}{E\left[\tau^{(\kappa)}\right]}$. *Then*

$$E^{\mathsf{f}}_{bits,\mathcal{Q}} = \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}_1 \neq M_1\right]}{E\left[\tau^{(\kappa)}\right]} \tag{4.4}$$

Now define $E^{\mathsf{f}}_{bits}(r) = \sup_{\mathcal{Q}: r_{\mathcal{Q}} \geq r} E^{\mathsf{f}}_{bits,\mathcal{Q}}$

We show next that this exponent decays linearly with the rate r of the special bits.

Theorem 16

$$E^{\mathsf{f}}_{bits}(r) = \left(1 - \tfrac{r}{C}\right) \mathsf{E}_{\mathsf{Red}}$$

Notice that for $r = 0$, the same exponent, $\mathsf{E}_{\mathsf{Red}}$, as the single bit case in the previous subsection could be achieved, although here the number of bits can be growing to infinity with $E[\tau]$. This linear tradeoff between rate and reliability reminds us of Burnashev's result [27]. Of course, in contrast to Burnashev's result, here the special bits are achieving $E^{\mathsf{f}}_{\text{bits}}(r)$ in spite of the sum data-rate approaching capacity.

Optimal strategy: Like the single bit case, we use a fixed length erasure code, where erasures are used to initiate retransmissions. In the first phase, transmit M_1 using a capacity-achieving code of length $\frac{r}{C}\kappa$. If the temporary decision $\tilde{M}_1$ is correct after this transmission, the transmitter sends M_2 in the second phase using a capacity-achieving code of length $(1 - \frac{r}{C})\kappa$. Otherwise, the transmitter sends a buzzer in these $(1 - \frac{r}{C})\kappa$ symbols by repeating the symbol x_r.

If the decoder detects a buzzer in these last $(1 - \frac{r}{C})\kappa$ symbols, an erasure is declared. The same message is retransmitted by repeating the same strategy afresh. If no buzzer is detected, the receiver uses an ML decoder to chose $\hat{M}_2$ and declares $\hat{M} = (\tilde{M}_1, \hat{M}_2)$.

Similarly to the single bit case, the erasure probability remains vanishingly small and channel capacity is achieved in spite of retransmissions. A decoding error for M_1 happens only when an error happens after the first phase and the buzzer message sent in the second phase is not detected. The probability of the later event is $\doteq$

$\exp(-E[\tau](1-\frac{r}{C})\mathsf{E}_{\mathsf{Red}})$. The factor of $(1-\frac{r}{C})$ arises because the buzzer is only sent in that fraction of the entire block.

4.2.3 Multiple layers of priority

We can generalize this result to the case of multiple levels of priority, where the most important layer contains $E[\tau]\, r_1/\ln 2$ bits, the second-most important layer contains $E[\tau]\, r_2/\ln 2$ bits and so on. Hence for an L-layer situation, we consider the capacity-achieving code sequences with message sets of the form $\mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)} \times \cdots \times \mathcal{M}_L^{(\kappa)}$. We assume that the order of importance of the M_i's will be $M_1 \succ M_2 \succ \cdots \succ M_L$. Hence we require that $P_e{}^{M_1} \leq P_e{}^{M_2} \leq \cdots \leq P_e{}^{M_L}$.

For any L-layer capacity-achieving sequence with feedback, we can define the error exponent of the s^{th} layer as

$$E^{\mathsf{f}}_{\text{bits},s,\mathcal{Q}} = \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}_s \neq M_s\right]}{E\left[\tau^{(\kappa)}\right]} \tag{4.5}$$

The achievable error exponent region of L-layered capacity-achieving sequences with feedback is the set of all achievable exponent vectors $(E^{\mathsf{f}}_{\text{bits},1,\mathcal{Q}}, E^{\mathsf{f}}_{\text{bits},2,\mathcal{Q}}, \ldots, E^{\mathsf{f}}_{\text{bits},L-1,\mathcal{Q}})$. The following theorem determines that region.

Theorem 17 *Consider L-layered capacity-achieving sequences with feedback whose rate vectors $(r_1, r_2, \ldots, r_L)$ satisfy $\sum_{j=1}^{L} r_j = C$. Achievable error exponent region for the $L-1$ most important layers is composed of $(E_1, E_2, \ldots, E_{L-1})$ satisfying the following condition,*

$$E_i \leq \left(1 - \frac{\sum_{j=1}^{i} r_j}{\mathrm{C}}\right) \mathsf{E}_{\mathsf{Red}} \qquad \forall i \in \{1, 2, \ldots, (L-1)\} \tag{4.6}$$

Note that the least important layer cannot achieve a positive error exponent since we are communicating at capacity.

Optimal strategy: We first transmit the most important layer using a capacity-achieving code of length $\frac{r_1}{C}\kappa$. If it is decoded correctly, then transmit the next layer

with a capacity-achieving code of length $\frac{r_2}{C}\kappa$. Otherwise, start the 'buzzer' with input x_r till the end of this κ symbol block. Repeat the same strategy for all future layers too (except the last): start buzzer if wrong, next layer if right.

After this block of κ symbols is received at the decoder, it successively looks for the buzzer from the end of each sub-block to the end of the κ symbols. If a buzzer is detected in any of them, the entire transmission is repeated afresh.

This decoding strategy is equivalent to peeling an onion layer by layer and checking if it was rotten after peeling each layer—we decode layer after layer till a buzzer is detected after some layer. After decoding each layer (except the least important), the receiver decides whether a buzzer was sent after that layer. Each layer is decoded if no buzzer is detected after its transmission and an erasure is declared otherwise. Thus layer after layer is decoded till either a buzzer has been detected or all layers have been decoded. If a buzzer has been detected, the whole block is retransmitted from scratch.

Thus for each layer i, we can achieve the same exponent as if there were only two kinds of bits (as in Theorem 16):

- Bits in layer i and more important layers $k < i$ are special and
- bits in less important layers than layer i are ordinary.

Hence this could be considered as a successively refinable version of Theorem 16. Figure 4-2 shows these simultaneously achievable exponents across layers. This is a successively refinable version of the linear tradeoff $E_{\mathrm{md}}(r) = (1-r/C)\mathsf{E}_{\mathsf{Red}}$ in Theorem 16.

Note that the most important layer can achieve an exponent close to $\mathsf{E}_{\mathsf{Red}}$ if its rate is zero. As we move to layers with decreasing importance, the achievable error exponent decays gradually.

4.2.4 A special message

Now consider one particular message, say $M = 1$, which requires a small missed-detection probability. Similar to the no-feedback case, define $E_{\mathrm{md}}^{\mathrm{f}}$ as its missed-

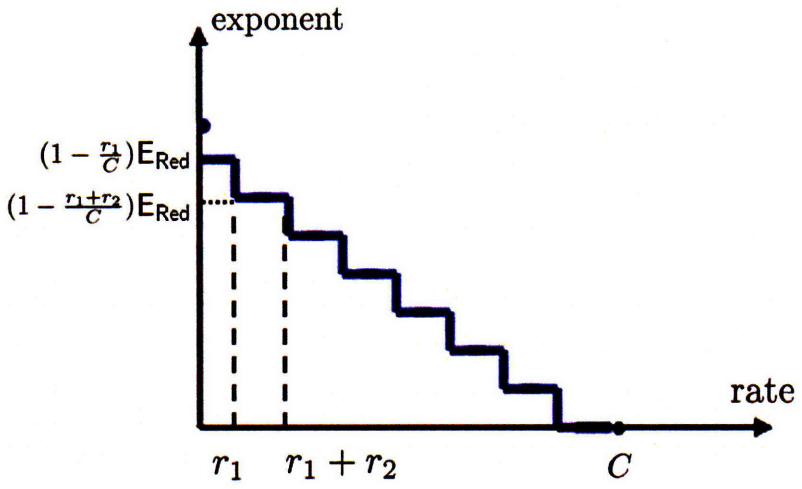


Figure 4-2: Successive refinability for multiple layers of priority

detection exponent at capacity.

Definition 18 *For a capacity-achieving sequence $\mathcal{Q}$ with feedback, the missed-detection exponent for the special message is defined as*

$$E^{\mathsf{f}}_{md,\mathcal{Q}} \equiv \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}\neq 1 \middle| M=1\right]}{E\left[\tau^{(\kappa)}\right]} \tag{4.7}$$

Now define $E^{\mathsf{f}}_{md} \equiv \sup_{\mathcal{Q}} E^{\mathsf{f}}_{md,\mathcal{Q}}$.

Theorem 19 *Feedback does not improve the missed-detection exponent of a single special message:* $E^{\mathsf{f}}_{md} = E_{md} = \mathsf{E}_{\mathsf{Red}}$.

Proof of this theorem is provided in Section 4.4. Since the Red-Alert Exponent without feedback is the best protection of a special message achievable at capacity without feedback, this result could be thought of as an analog to "feedback does not increase capacity" for the Red-Alert Exponent. Also note that with feedback, $E^{\mathsf{f}}_{\mathrm{md}}$ for the special message and $E^{\mathsf{f}}_{\mathrm{b}}$ for the special bit become equal.

4.2.5 Many special messages

Now let us consider the problem where an exponentially large subset $\mathcal{M}_s^{(n)} \subseteq \mathcal{M}^{(n)}$ of messages is special.

$$\lim_{n\to\infty} \frac{\log |\mathcal{M}_s^{(n)}|}{n} = r \tag{4.8}$$

Unlike the previous problems, now we also impose a uniform-expected-delay constraint as follows.

Definition 20 *For a reliable variable-length block code with feedback,*

$$\Gamma \equiv \frac{\max_{i\in\mathcal{M}} E[\tau|M=i]}{E[\tau]} \tag{4.9}$$

A reliable sequence with feedback $\mathcal{Q}$ is called a uniform-delay reliable sequence with feedback if $\lim_{\kappa\to\infty} \Gamma^{(\kappa)} = 1$.

This means that the average decoding time $E[\tau|\, M=i]$ for every message i is essentially equal to $E[\tau]$ (if not smaller). This uniformity constraint reflects a system requirement for ensuring a robust delay performance, which is invariant of the transmitted message[1]. Let us define the missed-detection exponent $E_{\text{md}}^{\text{f}}(r)$ under this uniform-delay constraint.

Definition 21 *For a uniform-delay capacity-achieving sequence $\mathcal{Q}$ with feedback, the missed-detection exponent for special messages in $\mathcal{M}_s^{(\kappa)}$ is defined as:*

$$E_{md,\mathcal{Q}}^{\text{f}} \equiv \liminf_{\kappa\to\infty} \frac{-\log \max_{i\in\mathcal{M}_s^{(\kappa)}} \Pr[\hat{M}\neq i|M=i]}{E\left[\tau^{(\kappa)}\right]}.$$

where $|\mathcal{M}_s^{(\kappa)}| \doteq \exp(E\left[\tau^{(\kappa)}\right] r)$ as in (4.8). We define $E_{md}^{\text{f}}(r) \equiv \sup_{\mathcal{Q}} E_{md,\mathcal{Q}}^{\text{f}}$.

The following theorem shows that the special messages could achieve the minimum of the Red-Alert Exponent and the Burnashev's exponent at rate r.

[1]Optimal exponents in all the previous problems (i.e., $E_{\text{b}}^{\text{f}}, E_{\text{bits}}^{\text{f}}(r)$, and E_{md}^{f}) remain unchanged irrespective of this uniform-delay constraint.

Theorem 22

$$E_{md}^{\mathsf{f}}(r) \;\equiv\; \min\left\{\mathsf{E}_{\mathsf{Red}}, (1-\tfrac{r}{\mathrm{C}})\mathrm{D}_{\max}\right\}, \quad \forall\, r < \mathrm{C}.$$

where $\mathrm{D}_{\max} \;\equiv\; \max_{x_1,x_2} D\left(W_{Y|X}(\cdot|x_1)\middle\| W_{Y|X}(\cdot|x_2)\right)$.

For r at which $\mathsf{E}_{\mathsf{Red}} \leq (1-\frac{r}{C})D_{\max}$, all $\lceil e^{E[\tau]r}\rceil$ special messages achieve the best missed-detection exponent $\mathsf{E}_{\mathsf{Red}}$ for a single special message. Since $\mathsf{E}_{\mathsf{Red}}$ denotes the best missed-detection exponent for a single special message, no better exponent could be achieved for multiple special messages. For larger r where $\mathsf{E}_{\mathsf{Red}} > (1-\frac{r}{C})D_{\max}$, the special messages achieve Burnashev's exponent as if the ordinary messages were absent. Since Burnashev's exponent is the best error exponent when there are only $\lceil\exp(E[\tau]r)\rceil$ messages, no better exponent could be achieved when the additional $\doteq \exp(E[\tau]C)$ ordinary messages are also present.

The optimal strategy is based on transmitting a special bit first. It again shows how feedback connects bit-wise UEP with message-wise UEP. In the optimal strategy for bit-wise UEP with many bits a special message was used, whereas now in message wise UEP with many messages a special bit is used. The roles of bits and messages, in two optimal strategies are simply swapped between two cases.

Optimal strategy: We combine the strategy for achieving $\mathsf{E}_{\mathsf{Red}}$ for a special bit and the Yamamoto-Itoh strategy for achieving Burnashev's exponent [29]. In the first phase, a special bit b is sent with a repetition code of $\sqrt{\kappa}$ symbols. This is an indicator bit for special messages: it is 1 when a special message is to be sent and 0 when an ordinary message is to be sent.

If b is decoded incorrectly as $\hat{b} = 0$, the buzzer is sent with input x_r for the remaining κ symbols. If it is decoded correctly as $\hat{b} = 0$, then the ordinary message is sent using a capacity-achieving code of length κ. If the receiver detects no buzzer in these κ symbols, it chooses the ML ordinary message. Otherwise, an erasure is declared and the entire block of length $\kappa + \sqrt{\kappa}$ is retransmitted afresh.

If b is decoded correctly as $\hat{b} = 1$, then a two-phase scheme of length κ is used to convey the particular special message. This scheme is exactly the same as the

Yamamoto and Itoh scheme in [29].

Yamamato-Itoh Scheme: First, the communication phase of this scheme takes place. A length $\frac{r}{C}\kappa$ capacity-achieving code (without feedback) is used to send the special message. Then the control phase of this scheme starts. An accept letter x_a is repeated for $(1-\frac{r}{C})\kappa$ symbols if the special message was decoded correctly after the communication phase. Otherwise, a reject letter x_d is repeated for $(1-\frac{r}{C})\kappa$ symbols. If the empirical distribution in the control phase has type $W_{Y|X}(\cdot|x_a)$, then the special message decoded at the end of the communication phase is finalized as $\hat{M}$. Otherwise, an erasure is declared and the entire block of length $\kappa+\sqrt{\kappa}$ is retransmitted afresh. The accept letter x_a and reject letter x_d are chosen such that $D_{\max} = D\left(W_{Y|X}(\cdot|x_a)\middle\| W_{Y|X}(\cdot|x_d)\right)$.

4.3 Avoiding False Alarms

4.3.1 Block Codes without Feedback

We now study the scenario where the false-alarm of a special message is a critical event. The false-alarm probability $\Pr\left[\hat{M}=1|M=j\right]$ for this message should be minimized for $j \neq 1$. In classical error exponent analysis [2], the error probability for a given message usually means its missed-detection probability. However, examples such as "reboot" and "format" necessitate this notion of false-alarm probability.

Definition 23 *For a capacity-achieving sequence $\mathcal{Q}$, the false-alarm exponent is defined as*

$$E_{fa,\mathcal{Q}} \equiv \liminf_{\kappa\to\infty} \frac{-\log \max_{j\neq 1} \Pr\left[\hat{M}=1\middle|M=j\right]}{E[\tau]}.$$

Then E_{fa} is defined as $E_{fa} \equiv \sup_{\mathcal{Q}} E_{fa,\mathcal{Q}}$. The supremum is taken over all $\mathcal{Q}$ which ensure vanishingly small conditional error probability for the special message: $\lim_{n\to\infty} P_e^{(n)}(1) = 0$.

The last clause in the definition is to ensure that whenever the special message is sent, it is recovered reliably. Without such a reliability constraint, the definition of

E_{fa} is not very sensible.

Theorem 24

$$E_{fa}^{l} \leq E_{fa} \leq E_{fa}^{u} \tag{4.10}$$

The upper and lower bound to the false-alarm exponent are given by

$$E_{fa}^{l} \equiv \max_{\hat{x}\in\mathcal{X}} \min_{\substack{V_{Y|X}: \\ \sum_x P_X^*(x)V_{Y|X}(\cdot|x)=W_{Y|X}(\cdot|\hat{x})}} D\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right) \tag{4.11}$$

$$E_{fa}^{u} \equiv \max_{\hat{x}\in\mathcal{X}} D\left(W_{Y|X}(\cdot|\hat{x})\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right). \tag{4.12}$$

For conciseness, define $\mathcal{V}_x$ to be the set of channels $V_{Y|X}$ in (4.11) for which input distribution P_X^* induces output distribution $W_{Y|X}(\cdot|x)$. Now if the input letters achieving the maximum in the above optimizations are denoted by x_l and x_u (respectively),

$$E_{\text{fa}}^{\text{l}} = \min_{V_{Y|X}:\mathcal{V}_{x_l}} D\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right) \tag{4.13}$$

$$E_{\text{fa}}^{\text{u}} = D\left(V_{Y|X}(\cdot|x_u)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right). \tag{4.14}$$

The strategy for achieving this lower bound is briefly described here and the upper bound is proved in Section 4.5. Section 4.5 contains detailed proofs of these bounds.

Achieving the Lower Bound: The codeword for the special message $M = 1$ is a repetition sequence of input letter x_l. Its decoding region $\mathcal{G}(1)$ is the typical 'noise ball' around it, i.e., the output sequences of type close to $W_{Y|X}(\cdot|x_l)$. For the ordinary messages, we use a capacity-achieving code-book where all codewords have the same empirical distribution (approx.) P_X^*. Then for $y^n \notin \mathcal{G}(1)$, the receiver uses ML decoding amongst ordinary codewords.

Note the contrast between this strategy for achieving E_{fa}^{l} and the optimal strategy for achieving E_{md}. For achieving E_{md}, output sequences of any type other than P_Y^* were assigned to $\mathcal{G}(1)$, whereas for achieving E_{fa} only the output sequences of type close to $W_{Y|X}(\cdot|x_l)$ are in $\mathcal{G}(1)$.

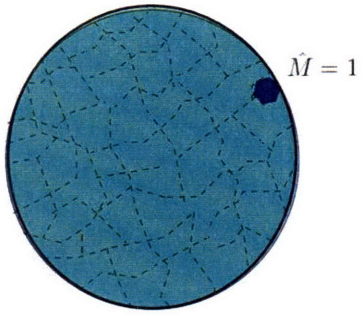


Figure 4-3: Avoiding false-alarm

Intuitive Interpretation: A false-alarm exponent for the special message corresponds to having the smallest possible decoding region $\mathcal{G}(1)$ for the special message subject to its missed-detection probability ${P_e}^{(n)}(1)$ vanishing to 0. This ensures that when some ordinary message is transmitted, the probability of landing in $\mathcal{G}(1)$ is exponentially small. We cannot make $\mathcal{G}(1)$ too small though, because when the special message is transmitted, the probability of landing outside it should be small too. Hence $\mathcal{G}(1)$ must contain essentially the typical noise ball around the special codeword. The blue region in Fig. 4-3 denotes such a region.

Note that E_{fa}^{l} is strictly larger than channel capacity C due to the convexity of KL divergence.

$$
\begin{aligned}
E_{\text{fa}}^{\text{l}} &= \max_{x \in \mathcal{X}} \min_{V_{Y|X} \in \mathcal{V}_x} D\left(V_{Y|X}(\cdot|X) \middle\| W_{Y|X}(\cdot|X) \middle| P_X^*\right) \\
&> \max_{x \in \mathcal{X}} \min_{V_{Y|X} \in \mathcal{V}_x} D\left(\sum_k P_X^*(k) V_{Y|X}(\cdot|k) \middle\| \sum_{k'} P_X^*(k') W_{Y|X}(\cdot|k')\right) \\
&= \max_{x \in \mathcal{X}} D\left(W_{Y|X}(\cdot|x) \middle\| P_Y^*(\cdot)\right) \qquad \text{(definition of } \mathcal{V}_x\text{)} \\
&= C \qquad \text{(KKT conditions for achieving capacity [2])}
\end{aligned}
$$

Now we can compare our result for a special message with the similar result for the classical situation where all messages are treated equally. It turns out that if every message in a capacity-achieving code demands equally good false-alarm exponent, then this uniform exponent cannot be larger[2] than C. This result seems to be directly

[2]This curious fact follows simply by writing the overall average error probability P_e in terms of

connected with the problem of identification via channels [26]. We can prove the achievability part of their capacity theorem using an extension of the achievability part of E_{fa}. Perhaps a new converse of their result is also possible using such results. Furthermore we see that reducing the demand of false-alarm exponent to only one message, instead of all, enhances it to at least E_{fa}^{l}.

4.3.2 Variable-Length Block Codes with Feedback

Recall that feedback did not improve the missed-detection exponent for a special message. On the contrary, we will see that the false-alarm exponent for a special message can be improved when feedback is available. We again restrict to uniform-delay capacity-achieving sequences with feedback, i.e., capacity-achieving sequences satisfying $\lim_{\kappa\to\infty} \Gamma^{(\kappa)} = 1$ for $\Gamma^{(\kappa)}$ in (4.9).

Definition 25 *For a uniform-delay capacity-achieving sequence $\mathcal{Q}$ with feedback, the false-alarm exponent is defined as*

$$E_{fa,\mathcal{Q}}^{\text{f}} \equiv \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}=1 \middle| M\neq 1\right]}{E[\tau^{\kappa}]}.$$

Then E_{fa}^{f} is defined as $E_{fa}^{\text{f}} \equiv \sup_{\mathcal{Q}} E_{fa,\mathcal{Q}}^{\text{f}}$. The supremum is taken over all $\mathcal{Q}$ which ensure vanishingly small conditional error probability for the special message: $\lim_{\kappa\to\infty} P_e^{(\kappa)}(1) = 0$.

Theorem 26 $E_{fa}^{\text{f}} = \text{D}_{\max} \equiv \max_{x_1,x_2} D\left(W_{Y|X}(\cdot|x_1) \middle\| W_{Y|X}(\cdot|x_2)\right)$.

Since $D_{\max} > E_{\text{fa}}^{\text{u}} \geq E_{\text{fa}}$, feedback strictly improves false-alarm exponent, $E_{\text{fa}}^{\text{f}} > E_{\text{fa}}$. Note that $D_{\max}$ also equals the best exponent for binary hypothesis testing with (or without) feedback [28].

Optimal strategy: We use the strategy employed in proving Theorem 22 in subsection 4.2.5. In the first phase, a length $\sqrt{\kappa}$ code is used to convey whether $M = 1$ or

the sum of false-alarm probabilities of all messages. A larger than C false-alarm exponent for every message would imply a positive classical exponent at capacity.

not. As before, we use a special bit b which is 1 if $M = 1$ and 0 otherwise. After the first phase:

- If $\hat{b}$ is decoded correctly as $\hat{b} = 0$, use a length κ capacity-achieving code to convey the ordinary codeword. If $\hat{b}$ is decoded incorrectly as $\hat{b} = 0$, send a length κ buzzer by repeating input x_r.

- If $\hat{b} = 1$, use a length κ code having two codewords: The "Accept" codeword $(x_a, x_a, \ldots, x_a)$ and the "Reject" codeword $(x_d, x_d, \ldots, x_d)$. The receiver finalizes decoding to the special message if the output sequence in the second phase has type $W_{Y|X}(\cdot|x_a)$. This is the same as the confirmation phase in the Yamamato-Itoh strategy described earlier.

If a buzzer is detected after $\hat{b} = 0$ or if a "reject" codeword is detected after $\hat{b} = 1$, an erasure is declared and retransmission starts afresh. Notice that this strategy will simultaneously achieve the best missed-detection exponent $\mathsf{E}_{\mathsf{Red}}$ and the best false-alarm exponent $D_{\max}$ for this special message.

4.4 Variable-Length Block Codes with Feedback: Proofs

This section presents detailed proofs of the results in Section 4.2, that is, Theorems 14, 16, 17, 19 and 22.

4.4.1 Proof of Theorem 14

Achievability: $E_{\mathsf{b}}^{\mathsf{f}} \geq \mathsf{E}_{\mathsf{Red}}$

As mentioned earlier, this single bit exponent is achieved using the missed-detection exponent of a single special message, indicating a decoding error for the special bit. The decoding error for the bit goes unnoticed when this special buzzer message is not detected.

Proof We prove $E_{\mathrm{b}}^{\mathrm{f}} \geq \mathsf{E}_{\mathsf{Red}}$ by constructing a capacity-achieving sequence with feedback, $\mathcal{Q}$, such that $E_{\mathrm{b},\mathcal{Q}}^{\mathrm{f}} = \mathsf{E}_{\mathsf{Red}}$. Let $\mathcal{Q}'$ be a capacity-achieving sequence of block codes without feedback which achieves the Red-Alert Exponent $E_{\mathrm{md},\mathcal{Q}'} = \mathsf{E}_{\mathsf{Red}}$. The message sets for $\mathcal{Q}'$ are denoted as $\mathcal{M}'^{(\kappa)}$. We first construct a two phase fixed length block code with feedback and erasures. Then $\mathcal{Q}$ is created from this fixed length block code by allowing retransmissions.

In the first phase, the transmitter uses a length $\lceil\sqrt{\kappa}\rceil$ code of two messages for sending the special bit M_1. At the end of this phase, the receiver forms a temporary decision, $\tilde{M}_1$. Note that as a result of [2, Theorem 5.7.1, page 153]

$$\Pr\left[\tilde{M}_1 \neq M_1\right] \leq e^{-\sqrt{\kappa}E_{ex}\left(\frac{\ln 8}{\sqrt{\kappa}}\right)} \tag{4.15}$$

where $E_{ex}(\cdot)$ stands for the expurgated exponent. Thus the error probability of the temporary decision is vanishingly small.

In the second phase, the transmitter uses a length κ code in $\mathcal{Q}'$. The message in the second phase, ϑ, will be determined by M_2 based on whether M_1 equals $\tilde{M}_1$ or not.

$$\begin{aligned} \tilde{M}_1 \neq M_1 &\Rightarrow \vartheta = 0 \qquad \text{(special message)} \\ \tilde{M}_1 = M_1 \text{ and } M_2 = i &\Rightarrow \vartheta = i \qquad \forall i \in \mathcal{M}_2^{\kappa} \end{aligned}$$

At the end of the second phase, the receiver decodes ϑ using the decoder of $\mathcal{Q}'$.

If it decodes the special message $\hat{\vartheta} = 0$, an erasure is declared. Otherwise, the temporary decision $\tilde{M}_1$ for the special bit is finalized ($\hat{M}_1 = \tilde{M}_1$) and the ordinary bits are decoded as $\hat{M}_2 = \hat{\vartheta}$.

Note that the erasure probability for the two-phase fixed-length block code is upper bounded as

$$\begin{aligned} \Pr\left[\hat{\vartheta} = 0\right] &\leq \Pr\left[\tilde{M}_1 \neq M_1\right] + \Pr\left[\vartheta = 0 | \, \vartheta \neq 0\right] \\ &\leq e^{-\sqrt{\kappa}E_{ex}\left(\frac{\ln 4}{\sqrt{\kappa}}\right)} + \frac{|\mathcal{M}'^{(\kappa)}|}{|\mathcal{M}'^{(\kappa)}|-1} P_e'^{(\kappa)} \end{aligned} \tag{4.16}$$

where $P_e{}'^{(\kappa)}$ is the error probability of the κ^{th} element of $\mathcal{Q}'$. Hence the overall erasure probability vanishes with κ.

Similarly we can upper bound the (undetected) error probabilities of the two-phase fixed-length block code as follows:

$$\Pr\left[\hat{M}_1 \neq M_1,\ \hat{\vartheta} \neq 0\right] \leq P_e{}'^{(\kappa)}(0) \tag{4.17}$$

$$\Pr\left[\hat{M} \neq M,\ \hat{\vartheta} \neq 0\right] \leq \frac{\mathcal{M}'^{(\kappa)}}{\mathcal{M}'^{(\kappa)}-1} P_e{}'^{(\kappa)} + P_e{}'^{(\kappa)}(1) \tag{4.18}$$

where $P_e{}'^{(\kappa)}(0)$ is the conditional error probability of the special message, i.e., the message 0 in the κ^{th} code in $\mathcal{Q}'$.

If there is an erasure the transmitter and the receiver will repeat what they have done again, until they get $\hat{\vartheta} \neq 0$. If we sum all the error probabilities in each step of repetition we get;

$$\Pr\left[\hat{M}_1 \neq M_1\right] \leq \frac{\Pr\left[\hat{M}_1 \neq M_1,\ \hat{\vartheta} \neq 0\right]}{1-\Pr\left[\hat{\vartheta}=0\right]} \tag{4.19}$$

$$\Pr\left[\hat{M} \neq M\right] \leq \frac{\Pr\left[\hat{M} \neq M,\ \hat{\vartheta} \neq 0\right]}{1-\Pr\left[\hat{\vartheta}=0\right]} \tag{4.20}$$

Since the number of retransmissions is a geometric process, the expected decoding time of the code equals

$$E\left[\tau\right] \leq \frac{\kappa+\lceil\sqrt{\kappa}\rceil}{1-\Pr\left[\hat{\vartheta}=0\right]} \tag{4.21}$$

Using equations (4.16), (4.17), (4.18), (4.19), (4.20) and (4.21) one can conclude that the resulting sequence $\mathcal{Q}$ of variable-length block codes with feedback is a capacity-achieving sequence and $E^{\mathsf{f}}_{\mathrm{b},\mathcal{Q}} = \mathsf{E}_{\mathsf{Red}}$. •

Converse: $E^{\mathsf{f}}_{\mathbf{b}} \leq \mathsf{E}_{\mathsf{Red}}$

We will use a converse result we have not proved yet, namely the converse in Theorem 19 for the single special message exponent $E^{\mathsf{f}}_{\mathrm{md}}$ with feedback. We now show that error exponent for a special bit cannot be larger than that for a special message. This is done by converting a code for a single special bit into a code for a single special

message.

Proof Consider $\mathcal{Q}$, a capacity-achieving sequence with feedback whose message sets are of the form $\mathcal{M}^{(\kappa)} = \{0,1\} \times \mathcal{M}_2^{(\kappa)}$. Using $\mathcal{Q}$ we construct another capacity-achieving sequence with feedback $\bar{\mathcal{Q}}$, which has a special message 0. Message sets for $\bar{\mathcal{Q}}$ are of the form $\bar{\mathcal{M}}^{(\kappa)} = \{0\} \cup \mathcal{M}_2^{(\kappa)}$, which is essentially half the size of the message sets for $\mathcal{Q}$. We will show that this special message in $\bar{\mathcal{Q}}$ can achieve the same missed-detection exponent as the special bit in $\mathcal{Q}$: $E^{\mathrm{f}}_{\mathrm{md},\bar{\mathcal{Q}}} \geq E^{\mathrm{f}}_{\mathrm{b},\mathcal{Q}}$. Consequently $E^{\mathrm{f}}_{\mathrm{md}} \geq E^{\mathrm{f}}_{\mathrm{b}}$. Since $E^{\mathrm{f}}_{\mathrm{md}} \leq \mathsf{E}_{\mathsf{Red}}$ from Theorem 19, $E^{\mathrm{f}}_{\mathrm{b}} \leq \mathsf{E}_{\mathsf{Red}}$.

To avoid confusion, let us denote message in $\mathcal{Q}$ by M and message in $\bar{\mathcal{Q}}$ by ϑ. The κ^{th} code in $\bar{\mathcal{Q}}$ works as follows. If the message of $\bar{\mathcal{Q}}$ is not 0, i.e. $\vartheta \neq 0$ then the transmitter uses the codeword for $M = (1, \vartheta)$ to convey ϑ. If $\vartheta = 0$ transmitter picks a (dummy) M_2 with uniform distribution on $\mathcal{M}_2^{(\kappa)}$ and uses the codeword for $M = (0, M_2)$ to convey that $\vartheta = 0$. Receiver makes decoding using the decoder of $\mathcal{Q}$. If $\hat{M} = (1, k)$ it declares $\hat{\vartheta} = k$. If $\hat{M} = (0, k)$, it declares the special message $\hat{\vartheta} = 0$.

Essentially, the special bit in $\mathcal{Q}$ is being used to convey the special message in $\bar{\mathcal{Q}}$: this special bit is 0 when special message is to be sent and 1 when some ordinary message is to be sent. This converse argument is essentially the reverse of the optimal strategy for achieving $E^{\mathrm{f}}_{\mathrm{b}} = \mathsf{E}_{\mathsf{Red}}$, where a special message was used to send a special bit. This further emphasizes how feedback connects message-wise and bit-wise UEP.

Let $E\left[\tau^{(\kappa)}\right]$ denote the average decoding time for $\mathcal{Q}$ assuming uniformly chosen messages from $\{0,1\} \times \mathcal{M}_2^{(\kappa)}$. Then by definition of $E^{\mathrm{f}}_{\mathrm{b},\mathcal{Q}}$), the probability of decoding $\hat{M}_1 = 1$ when $M_1 = 0$ has exponent $\doteq \exp(-E\left[\tau^{(\kappa)}\right] E^{\mathrm{f}}_{\mathrm{b},\mathcal{Q}})$, which is the same as the missed-detection probability of the special message in $\bar{\mathcal{Q}}$. •

4.4.2 Proof of Theorem 16

Achievability: $E^{\mathrm{f}}_{\mathbf{bits}}(r) \geq \left(1 - \frac{r}{C}\right) \mathsf{E}_{\mathsf{Red}}$

Proof We construct the capacity-achieving sequence with feedback $\mathcal{Q}$ using a capacity-achieving sequence (without feedback) $\mathcal{Q}'$, which achieves $E_{\mathrm{md},\mathcal{Q}'} = \mathsf{E}_{\mathsf{Red}}$. Again $\mathcal{M}'^{(\kappa)}$

denotes the message sets for this code sequence $\mathcal{Q}'$. This strategy is similar to our achievability proof for Theorem 14 for a single special bit. Each code of in $\mathcal{Q}$ uses 2 codes from $\mathcal{Q}'$: one of length $\lceil r\kappa/C \rceil$ and the other of length $\lceil (1-r/C)\kappa \rceil$.

We again construct a two phase block code with feedback and erasures. Consider the method for creating the κ^{th} element of the code sequence $\mathcal{Q}$. In the first phase, the transmitter uses the length $\lceil r\kappa/C \rceil$ code in $\mathcal{Q}'$ to convey M_1. At the end of this phase, the receiver makes a temporary decision, $\tilde{M}_1$. In the second phase, the transmitter uses the length $\lceil (1-r/C)\kappa \rceil$ code in $\mathcal{Q}'$ to convey M_2 with a special message if $\tilde{M}_1 \neq M_1$. It uses a mapping similar to the one in the proof of Theorem 14. As before, let ϑ denote the message in the second phase.

$$\begin{aligned} \tilde{M}_1 \neq M_1 &\Rightarrow \vartheta = 0 \qquad \text{(special message)} \\ \tilde{M}_1 = M_1 \text{ and } M_2 = i &\Rightarrow \vartheta = i \quad \forall i \in \mathcal{M}_2^\kappa \end{aligned}$$

This construction gives

$$|\mathcal{M}_1^\kappa| = |\mathcal{M}'^{(\lceil r\kappa/C \rceil)}| \text{ and } |\mathcal{M}_2^\kappa| = |\mathcal{M}'^{(\lceil (1-r/C)\kappa \rceil)}| - 1$$

for the code sequence $\mathcal{Q}$. Applying a similar decoding algorithm as in the proof of Theorem 14 and repeating essentially the same analysis, we get that $\mathcal{Q}$ is a capacity-achieving sequence with $E^{\mathsf{f}}_{\text{bits},\mathcal{Q}} = \left(1 - \frac{r}{C}\right) \mathsf{E}_{\mathsf{Red}}$ and $r_{\mathcal{Q}} = r$ as the rate of special bits. •

Converse: $E^{\mathsf{f}}_{\mathbf{bits}}(r) \leq \left(1 - \frac{r}{C}\right) \mathsf{E}_{\mathsf{Red}}$

This converse is based on a technique previously used in [28] and a lemma proved later in the converse part of Theorem 8. For this proof as well as other converse proofs in this chapter, we need to analyze the conditional entropy of messages given the receiver's observations. In such analysis, we use the following notation for conditional

entropy and conditional mutual information,

$$\begin{aligned}
\mathcal{H}(M|y^n) &= -\sum_{i\in\mathcal{M}} \Pr\left[M=i|\, y^n\right] \ln \Pr\left[M=i|\, y^n\right] \\
\mathcal{I}\left(M; Y_{n+1}\,|y^n\right) &= \mathcal{H}(M|y^n) - \sum_{y_{n+1}\in\mathcal{Y}} \Pr\left[Y_{n+1}=y_{n+1}|\, y^n\right] \mathcal{H}(M|y^n, y_{n+1}) \\
&= \mathcal{H}(Y_{n+1}|y^n) - \sum_{i\in\mathcal{M}} \Pr\left[M=i|\, y^n\right] \mathcal{H}(Y_{n+1}|y^n, M=i) \quad \text{and} \\
\mathcal{I}\left(X_{n+1}; Y_{n+1}\,|y^n\right) &= \mathcal{H}(Y_{n+1}|y^n) - \sum_{x_{n+1}\in\mathcal{X}} \Pr\left[X_{n+1}=x_{n+1}|\, y^n\right] \mathcal{H}(Y_{n+1}|y^n, x_{n+1}).
\end{aligned}$$

These quantities should be thought as random variables and as functions of the history Y^n. It is worth noting that this notation is different than the widely used notation, which includes a further expectation over the conditioned variable. The term "$H(M|Y^n = y^n)$" in conventional notation means just $\mathcal{H}(M|y^n)$ in our notation and "$H(M|Y^n)$" means $E\left[\mathcal{H}(M|Y^n)\right]$, where the expectation is taken over Y^n. Note that applying memoryless property of the channel to the last two equations above implies

$$\mathcal{I}\left(M; Y_{n+1}\,|y^n\right) = \mathcal{I}\left(X_{n+1}; Y_{n+1}\,|y^n\right).$$

Proof Consider any variable-length block code with feedback whose message set $\mathcal{M}$ is of the form $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$. Let t_δ be the first time instance that an $i \in \mathcal{M}_1$ becomes more likely than $(1-\delta)$ given the observations so far and let $\tau_\delta = \min\{t_\delta, \tau\}$ where τ is the decoding time.

For each possible y^{τ_δ} we divide the message set $\mathcal{M}$ into $|\mathcal{M}_2|+1$ subsets called $\mathcal{F}_0(y^{\tau_\delta}), \mathcal{F}_1(y^{\tau_\delta}), \cdots, \mathcal{F}_{|\mathcal{M}_2|}(y^{\tau_\delta})$. For $\ell = 1$ to $|\mathcal{M}_2|$, subset $\mathcal{F}_\ell$ is composed of the message $(\tilde{M}_1(Y^{\tau_\delta}), \ell)$, where $\tilde{M}_1(Y^{\tau_\delta})$ is the most likely message given Y^{τ_δ}. The remaining subset $\mathcal{F}_0$ is composed of the rest of the messages, i.e., all messages of the form (i,j) where $i \neq \tilde{M}_1(Y^{\tau_\delta})$.

The index ℓ for which $\mathcal{F}_\ell(Y^{\tau_\delta})$ contains the transmitted message M is defined as the axillary-message $\vartheta(Y^{\tau_\delta})$. Essentially, it denotes a quantization of the ordinal message M. The index k for which $\mathcal{F}_k(Y^{\tau_\delta})$ contains the final decoded message $\hat{M}(Y^\tau)$

is defined as the decoded $\vartheta(Y^\tau)$, i.e.,

$$\hat{\vartheta}(Y^\tau) = k \;\Leftrightarrow\; \hat{M}(Y^\tau) \in \mathcal{F}_k(Y^{\tau_\delta}) \tag{4.22}$$

With these definition we get

$$\Pr\left[\hat{M}(Y^\tau) \neq M \middle| Y^{\tau_\delta}\right] \geq \Pr\left[\hat{\vartheta}(Y^\tau) \neq \vartheta(Y^{\tau_\delta}) \middle| Y^{\tau_\delta}\right] \tag{4.23}$$

$$\Pr\left[\hat{M}_1(Y^\tau) \neq M_1 \middle| Y^{\tau_\delta}\right] \geq \Pr\left[\hat{\vartheta}(Y^\tau) \neq 0 \middle| \vartheta(Y^{\tau_\delta}) = 0, Y^{\tau_\delta}\right] \Pr\left[\vartheta(Y^{\tau_\delta}) = 0 \middle| Y^{\tau_\delta}\right] \tag{4.24}$$

Now, we apply Lemma 27, which will be proved in the the converse proof of Theorem 19. For the ease of notation, define the shorthand:

$$\begin{aligned} P_e^{\vartheta}(Y^{\tau_\delta}) &= \Pr\left[\hat{\vartheta}(Y^\tau) \neq \vartheta(Y^{\tau_\delta}) \middle| Y^{\tau_\delta}\right] \\ P_e^{\vartheta}(0, Y^{\tau_\delta}) &= \Pr\left[\hat{\vartheta}(Y^\tau) \neq 0 \middle| \vartheta(Y^{\tau_\delta}) = 0, Y^{\tau_\delta}\right] \\ \xi(Y^{\tau_\delta}) &= \Pr\left[\vartheta(Y^{\tau_\delta}) = 0 \middle| Y^{\tau_\delta}\right] \end{aligned}$$

From Lemma 27, for each realization of Y^{τ_δ} such that $\tau_\delta < \tau$,

$$(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta})) \ln \tfrac{1}{P_e^{\vartheta}(0,Y^{\tau_\delta})} \leq \ln 2 + E\left[\tau - \tau_\delta \middle| Y^{\tau_\delta}\right] \mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}(Y^{\tau_\delta}) \ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)$$

where $\mathcal{J}(R)$ denotes the best missed-detection exponent for a special message while ensuring reliable communication are rate $R \leq C$. It is a generalization of the Red-Alert Exponent $\mathsf{E}_{\mathsf{Red}}$ for rates below capacity (see Chapter 5). More details and its precise formula can also be found in Section 4.4.4 ahead. At this moment, we only need the fact that it is a decreasing concave function.

If we multiply both sides of this inequality by the indicator function $\mathbb{I}_{\{\tau_\delta<\tau\}}$, the expression we get holds for all Y^{τ_δ}. Thus

$$\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta})) \ln \tfrac{1}{P_e^{\vartheta}(0,Y^{\tau_\delta})} \leq \mathbb{I}_{\{\tau_\delta<\tau\}}\left[\ln 2 + E\left[\tau - \tau_\delta \middle| Y^{\tau_\delta}\right] \mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}(Y^{\tau_\delta}) \ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)\right]$$

Now take the expectation of both sides over Y^{τ_δ}

$$\begin{aligned}
L.H.S. &= E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))\ln\tfrac{1}{P_e^{\vartheta}(0,Y^{\tau_\delta})}\right] \\
&\overset{(a)}{\geq} E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))\right]\ln\frac{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))\right]}{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))P_e^{\vartheta}(0,Y^{\tau_\delta})\right]} \\
&\overset{(b)}{\geq} (1-\tfrac{P_e}{\lambda\delta}-\delta)\ln\frac{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))\right]}{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))P_e^{\vartheta}(0,Y^{\tau_\delta})\right]} \\
&\overset{(c)}{\geq} (1-\tfrac{P_e}{\lambda\delta}-\delta)\ln\frac{(1-\frac{P_e}{\lambda\delta}-\delta)}{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))P_e^{\vartheta}(0,Y^{\tau_\delta})\right]}
\end{aligned}$$

where $\lambda \equiv \min_{i,j} W_{Y|X}(i|j)$. Step (a) follows from log sum inequality. Steps (b) and (c) follow because the log term in (b) is positive and

$$E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}(Y^{\tau_\delta}))\right] \geq (1-\tfrac{P_e}{\lambda\delta}-\delta).$$

Now note that

$$E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta})-P_e^{\vartheta}((\,)Y^{\tau_\delta}))P_e^{\vartheta}(0,Y^{\tau_\delta})\right] \leq \tfrac{P_e M_1}{\delta\lambda}$$

Thus

$$L.H.S. \geq -\ln 2-(1-\tfrac{P_e}{\lambda\delta}-\delta)\ln\tfrac{P_e M_1}{\lambda\delta} \tag{4.25}$$

For the right hand side expectation, we use the fact that $\mathcal{J}(R)$ is a decreasing concave function of R to get,

$$\begin{aligned}
R.H.S. &= E\left[\left(\ln 2+E\left[\tau-\tau_\delta|\,Y^{\tau_\delta}\right]\mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta})-\ln 2-P_e^{\vartheta}(Y^{\tau_\delta})\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)\right)\mathbb{I}_{\{\tau_\delta<\tau\}}\right] \\
&\leq \ln 2+E\left[E\left[\tau-\tau_\delta|\,Y^{\tau_\delta}\right]\mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta})-\ln 2-P_e^{\vartheta}(Y^{\tau_\delta})\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)\mathbb{I}_{\{\tau_\delta<\tau\}}\right] \\
&\leq \ln 2+E\left[\tau-\tau_\delta\right]\mathcal{J}\left(E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta})-\ln 2-P_e^{\vartheta}(Y^{\tau_\delta})\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta]}\right]\right) \\
&\leq \ln 2+E\left[\tau-\tau_\delta\right]\mathcal{J}\left(\tfrac{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\mathcal{H}(\vartheta|Y^{\tau_\delta})\right]-\ln 2-P_e\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta]}\right)
\end{aligned} \tag{4.26}$$

Now lower bound $E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\mathcal{H}(\vartheta|Y^{\tau_\delta})\right]$ in terms of $E\left[\mathcal{H}(M|Y^{\tau_\delta})\right]$. Note that for any

realization of $\mathcal{Y}^{\tau_\delta}$ we have

$$\begin{aligned}\mathcal{H}(M|Y^{\tau_\delta}) &= \mathcal{H}(\vartheta|Y^{\tau_\delta}) + \Pr\left[M_1 \neq \tilde{M}_1(Y^{\tau_\delta})\middle| Y^{\tau_\delta}\right]\mathcal{H}(M|M_1 \neq \tilde{M}_1(Y^{\tau_\delta}), Y^{\tau_\delta}) \\ &\leq \mathcal{H}(\vartheta|Y^{\tau_\delta}) + \Pr\left[M_1 \neq \tilde{M}_1(Y^{\tau_\delta})\middle| Y^{\tau_\delta}\right]\ln(|\mathcal{M}_1||\mathcal{M}_2|) \qquad (4.27)\end{aligned}$$

Furthermore for all Y^{τ_δ} such that $\tau > \tau_\delta$, we have $\Pr\left[\tilde{M}_1(Y^{\tau_\delta} = M_1)\middle| Y^{\tau_\delta}\right] \geq (1-\delta)$. This gives,

$$\begin{aligned}E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\mathcal{H}(\vartheta|Y^{\tau_\delta})\right] &\geq E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(\mathcal{H}(M|Y^{\tau_\delta}) - \delta\ln|\mathcal{M}_1||\mathcal{M}_2|)\right] \\ &= E\left[(1-\mathbb{I}_{\{\tau_\delta=\tau\}})\mathcal{H}(M|Y^{\tau_\delta})\right] - \delta\ln|\mathcal{M}_1||\mathcal{M}_2| \\ &\geq E\left[\mathcal{H}(M|Y^{\tau_\delta})\right] - \Pr\left[\tau_\delta = \tau\right]\ln|\mathcal{M}_1||\mathcal{M}_2| - \delta\ln(|\mathcal{M}_1||\mathcal{M}_2|) \qquad (4.28)\end{aligned}$$

Note that $\Pr\left[\tau_\delta = \tau\right] \leq \frac{P_e}{\lambda\delta}$. Inserting this together with equation (4.28) in the inequality given in (4.26).

$$\begin{aligned}R.H.S. &\leq \ln 2 + E\left[\tau - \tau_\delta\right]\mathcal{J}\left(\frac{E[\mathcal{H}(M|Y^{\tau_\delta})] - \left(\frac{P_e}{\lambda\delta}+\delta\right)\ln|\mathcal{M}_1||\mathcal{M}_2| - \ln 2 - P_e\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta]}\right) \\ &\leq \ln 2 + E\left[\tau - \tau_\delta\right]\mathcal{J}\left(\frac{\ln|\mathcal{M}_1||\mathcal{M}_2|\left(1-\frac{P_e}{\lambda\delta}-\delta-P_e\right) - \frac{E\left[\mathcal{H}(M|Y^0)-\mathcal{H}(M|Y^{\tau_\delta})\right]}{E[\tau_\delta]}E[\tau_\delta] - \ln 2}{E[\tau-\tau_\delta]}\right)\end{aligned}$$

Now using a result from [28],

$$\frac{E\left[\mathcal{H}(M|Y^0)-\mathcal{H}(M|Y^{\tau_\delta})\right]}{E[\tau_\delta]} \leq C \qquad (4.29)$$

Since $\mathcal{J}(\cdot)$ is a decreasing function we get,

$$R.H.S. \leq \ln 2 + E\left[\tau - \tau_\delta\right]\mathcal{J}\left(\frac{\ln|\mathcal{M}_1||\mathcal{M}_2|\left(1-\frac{P_e}{\lambda\delta}-\delta-P_e\right) - E[\tau_\delta]C - \ln 2}{E[\tau-\tau_\delta]}\right) \qquad (4.30)$$

Note that $\forall a > 0, b > 0, C > 0$,

$$\frac{d}{dx}(b-x)\mathcal{J}\left(\frac{a-Cx}{b-x}\right)|_{x=x_0} = -\mathcal{J}\left(\frac{a-Cx_0}{b-x_0}\right) - \left(C - \frac{a-Cx_0}{b-x_0}\right)\frac{d}{dx}\mathcal{J}(x)|_{x=\frac{a-Cx_0}{b-x_0}} \leq 0$$

Thus replacing $E[\tau_\delta]$ with a term lower than $E[\tau_\delta]$ itself will increase the value of the expression given in (4.30). Now using (4.29) and the fact that $\mathcal{H}(M_1|Y^{\tau_\delta}) \leq \ln 2 + (\delta + \frac{P_e}{\lambda\delta})\ln|\mathcal{M}_1|$ in (4.30),

$$R.H.S. \leq \ln 2 + \left(E[\tau] - (1-\delta\tfrac{P_e}{\lambda\delta})\tfrac{\ln|\mathcal{M}_1|}{C}\right)\mathcal{J}\left(\frac{\left(1-\frac{P_e}{\lambda\delta}-\delta-P_e\right)\ln|\mathcal{M}_2| - P_e\ln|\mathcal{M}_1| - 2\ln 2}{E[\tau]-(1-\delta-\frac{P_e}{\lambda\delta})\frac{\ln|\mathcal{M}_1|}{C}}\right) \tag{4.31}$$

Using this with (4.25) and choosing $\delta = \sqrt{P_e}$ implies $E^{\mathsf{f}}_{\text{bits},\mathcal{Q}} \leq \left(1 - \frac{r_{\mathcal{Q}}}{C}\right)\mathcal{J}(C)$. Since $\mathcal{J}(C) = \mathsf{E}_{\mathsf{Red}}$ this proves $E^{\mathsf{f}}_{\text{bits}}(r) \leq \left(1 - \frac{r}{C}\right)\mathsf{E}_{\mathsf{Red}}$. •

4.4.3 Proof of of Theorem 17

Achievability

Proof This argument is almost identical to the achievability proof for Theorem 16. The capacity-achieving $\mathcal{Q}$ with feedback is created from a capacity-achieving sequence (without feedback) $\mathcal{Q}'$ for which $E^{\mathsf{f}}_{\text{b},\mathcal{Q}'} = \mathsf{E}_{\mathsf{Red}}$. Each code of capacity-achieving sequence with feedback $\mathcal{Q}$ uses L codes from $\mathcal{Q}'$ as follows.

For the κ^{th} element of $\mathcal{Q}$, transmitter uses the length $\lceil \kappa r_1/C \rceil$ code in $\mathcal{Q}'$ to send the first part of the message M_1. Later, every phase $l \geq 2$ uses the length $\lceil \kappa r_l/C \rceil$ code in $\mathcal{Q}'$. The special message in phase l denotes an error event in previous phases.

$$(\tilde{M}_1, \ldots, \tilde{M}_{(l-1)}) \neq (M_1, \ldots, M_{(l-1)}) \Rightarrow \vartheta_l = 0 \quad \text{(special message)}$$
$$(\tilde{M}_1, \ldots, \tilde{M}_{(l-1)}) = (M_1, \ldots, M_{(l-1)}) \Rightarrow \vartheta_l = M_l$$

Thus $|\mathcal{M}_1^{(\kappa)}| = |\mathcal{M}'^{(\lceil r_1\kappa/C\rceil)}|$ and $|\mathcal{M}_l^{(\kappa)}| = |\mathcal{M}'^{(\lceil r_l\kappa/C\rceil)}| - 1$. If $\hat{\vartheta}_l \neq 0$ for all $l \in \{2, \ldots, L\}$, the receiver finalizes temporary decisions in all the phases. Otherwise an erasure is declared and a retransmission is initiated. We skip the error analysis because it is essentially identical to that of Theorem 16. •

Converse

Proof This simple converse is proved by contradiction. Since we have,

$$\begin{aligned} \max\{P_e{}^{M_1}, P_e{}^{M_2}, \ldots, P_e{}^{M_i}\} &\leq P_e{}^{M_1,M_2,\ldots,M_i} \leq P_e{}^{M_1} + P_e{}^{M_2} + \cdots + P_e{}^{M_i} \\ \Rightarrow P_e{}^{M_i} &\leq P_e{}^{M_1,M_2,\ldots,M_j} \leq iP_e{}^{M_i} \quad (\text{since } Pe^{M_1} \leq P_e{}^{M_2} \leq \ldots \leq P_e{}^{M_i}) \end{aligned}$$

due to assumption of better error probability for more important layers. Now if we think of two super messages as follows,

$$\vartheta_1 = (M_1, M_2, \ldots, M_i) \quad \text{and} \quad \vartheta_2 = (M_{i+1}, M_{i+2}, \ldots, M_l) \tag{4.32}$$

Then ϑ_1 achieves error exponent E_i and has rate $\sum_{j=1}^{i} r_j$. Now if there exists a scheme that can reach an error exponent vector outside the region given in Theorem 17, there will be at least one $E_i \geq (1 - \frac{\sum_{j=1}^{i} r_j}{C})\mathsf{E}_{\mathsf{Red}}$. This contradicts Theorem 16 for two layers where ϑ_1 corresponds to special bits and ϑ_2 corresponds to ordinary bits. •

4.4.4 Proof of Theorem 19

Achievability: $E_{\mathbf{md}}^{\mathsf{f}} \geq \mathsf{E}_{\mathsf{Red}}$

Since a fixed length block code without feedback is a special case of variable-length block codes with feedback, we get $E_{\mathrm{md}}^{\mathsf{f}} \geq E_{\mathrm{md}}$. Using the capacity-achieving sequence $\mathcal{Q}'$ without feedback that achieved the Red-Alert Exponent proves $E_{\mathrm{md}}^{\mathsf{f}} \geq \mathsf{E}_{\mathsf{Red}}$.

Converse: $E_{\mathrm{md}}^{\mathrm{f}} \leq \mathsf{E}_{\mathsf{Red}}$

Now we prove that even with feedback and variable decoding time, the best missed-detection exponent of single special message is at most $\mathsf{E}_{\mathsf{Red}}$. Since feedback can only help, this also implies $E_{\mathrm{md}} \leq \mathsf{E}_{\mathsf{Red}}$ for the no-feedback case and proves the converse for Theorem 8 in the previous chapter.

Before proving this converse part of Theorem 19, we prove the following lemma.

Lemma 27 *For any variable-length block code with feedback with $|\mathcal{M}|$ messages with initial entropy $\mathcal{H}(M|Y^0) \equiv \log|\mathcal{M}|$ (since Y^0 means no observations) and with average error probability P_e, the conditional error probability of each message is lower bounded as follows,*

$$\Pr\left[\hat{M} \neq i \middle| M = i\right] \geq e^{-\frac{1}{1-\Pr[M=i]-P_e}\left(\mathcal{J}\left(\frac{\log|\mathcal{M}|-h(P_e)-P_e\log(|\mathcal{M}|-1)}{E[\tau]}\right)E[\tau]+\log 2\right)} \qquad \forall i \tag{4.33}$$

where $h(\cdot)$ denotes the binary entropy function and $\mathcal{J}(R)$ is given by

$$\mathcal{J}(R) = \max_{\substack{\alpha_X, P_X^1, P_X^2, \ldots, P_X^{|\mathcal{X}|}: \\ \sum_{j\in\mathcal{X}} \alpha_X(j) I(P_X^j, W_{Y|X}) \geq R}} \sum_{j\in\mathcal{X}} \alpha_X(j) D\left((P_X^j W_{Y|X})_Y(\cdot) \middle\| W_{Y|X}(\cdot|j)\right)$$

where α_X and $\{P_X^j\}$ are distributions over $\mathcal{X}$. It is worthwhile remembering the notation we established previously:

$$\begin{aligned}(P_X^j W_{Y|X})_Y(\cdot) &= \sum_{x\in\mathcal{X}} P_X^j(x) W_{Y|X}(\cdot|x) \quad \text{and} \\ I(P_X^j, W_{Y|X}) &= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_X^j(x) W_{Y|X}(y|x) \log \frac{W_{Y|X}(y|x)}{(P_X^x W_{Y|X})_Y(y)}\end{aligned}$$

Note that $\mathcal{J}(R)$ is concave and strictly decreasing in R for $R \leq C$. Considering the achievability proof of Theorem 8 and the following converse proof, one can see that $\mathcal{J}(R)$ is the best exponent of a special message in a reliable code sequence of rate R. However, we will only need that fact for $R = C$, i.e., $\mathcal{J}(C) = \mathsf{E}_{\mathsf{Red}}$.

Proof of Lemma 27: For upper bounding the error probability of the special

message, let us consider the following stochastic sequence which is a function of the output sequence Y^n.

$$S_n \equiv S_n(Y^n) = \ln \frac{\Pr[Y^n]}{\Pr[Y^n|M=i]} - \sum_{t=1}^{n} E\left[\ln \frac{\Pr[Y_t|Y^{t-1}]}{\Pr[Y_t|M=i,Y^{t-1}]}\middle| Y^{t-1}\right] \tag{4.34}$$

where the expectation is taken over Y_t for a fixed Y^{t-1}. Note that $E[S_{n+1}|Y^n] = S_n$, which implies S_n is a Martingale process. Now recalling $\lambda \equiv \min W_{Y|X}(i|j)$ implies $|S_{n+1} - S_n| \leq 2\ln\frac{1}{\lambda}$. Furthermore since $E[\tau] < \infty$, we can use [32, Theorem 2 p 487], to get

$$E[S_\tau] = S_0 = 0. \tag{4.35}$$

This is essentially Doob' optional stopping time Theorem. Thus

$$E\left[\ln \frac{\Pr[Y^\tau]}{\Pr[Y^\tau|M=i]}\right] = E\left[\sum_{t=1}^{\tau} E\left[\ln \frac{\Pr[Y_t|Y^{t-1}]}{\Pr[Y_t|M=i,Y^{t-1}]}\middle| Y^{t-1}\right]\right] \tag{4.36}$$

$$\leq E\left[\sum_{t=1}^{\tau} \mathcal{J}\left(\mathcal{I}\left(X_t; Y_t\,|Y^{t-1}\right)\right)\right] \tag{4.37}$$

Note that the inner conditional expectation in the RHS of (4.36) equals

$$D(\Pr\left[Y_t|\,Y^{t-1}\right] \|\Pr\left[Y_t|\,M=i, Y^{t-1}\right]|Y^{t-1}).$$

If $\bar{x}_t(i)$ denotes the input at time t for message i and output history Y^{t-1}, this KL divergence is equal to $D(\Pr[Y_t|\,Y^{t-1}]\,\|W_{Y|X}(Y_t|\bar{x}_t(i))|Y^{t-1})$. Now (4.4.4) follows by definition of $\mathcal{J}(\cdot)$ and noting that To see that, choose distribution α_X to have all its mass at $b \equiv \bar{x}_t(i)$, the t'th symbol of i'th codeword $\bar{x}^n(i)$. Now the second argument of the KL divergence term equals $\Pr[Y_t|\,M=i,Y^{t-1}] = W_{Y|X}(\cdot|b)$. Choosing $P_X^b = \Pr[X_t]\,Y^{t-1}$ completes the argument for step .

Let $\mathcal{G}(i) = \{y^\tau : \hat{M}(y^\tau) = i\}$ denote decoding region for $\hat{M} = i$ and let $\overline{\mathcal{G}(i)}$ denote

its complement. As a result of the data processing inequality for KL divergence,

$$E\left[\ln\frac{\Pr[Y^\tau]}{\Pr[Y^\tau|M=i]}\right] \geq \Pr\left[\mathcal{G}(i)\right]\ln\frac{\Pr[\mathcal{G}(i)]}{\Pr[\mathcal{G}(i)|M=i]} + \Pr\left[\overline{\mathcal{G}(i)}\right]\ln\frac{\Pr\left[\overline{\mathcal{G}(i)}\right]}{\Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]}$$
$$\geq -h(\Pr\left[\mathcal{G}(i)\right]) + \Pr\left[\overline{\mathcal{G}(i)}\right]\ln\frac{1}{\Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]}$$

Using $h(\Pr[\mathcal{G}(i)]) \leq \ln 2$ and equation (4.4.4) we get

$$\Pr\left[\overline{\mathcal{G}(i)}\right]\ln\frac{1}{\Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]} \leq \ln 2 + E\left[\ln\frac{\Pr[Y^\tau]}{\Pr[Y^\tau|M=i]}\right]$$
$$\leq \ln 2 + E\left[\sum_{t=1}^{\tau}\mathcal{J}\left(\mathcal{I}\left(X_t;Y_t\middle|Y^{t-1}\right)\right)\right] \tag{4.38}$$

Note that

$$\Pr\left[\overline{\mathcal{G}(i)}\right] = \Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]\Pr\left[M=i\right] + \Pr\left[\overline{\mathcal{G}(i)}\middle|M\neq i\right]\Pr\left[M\neq i\right]$$
$$\geq (1 - P_e - \Pr\left[M=i\right]) \tag{4.39}$$

Using concavity of $\mathcal{J}(\cdot)$ along with (4.38) and (4.39) gives

$$\Pr\left[\hat{M}\neq i\middle|M=i\right] \geq e^{-\frac{1}{1-P_e-\Pr[M=i]}\left(\mathcal{J}\left(\frac{E\left[\sum_{t=1}^{\tau}\mathcal{I}\left(X_t;Y_t\middle|Y^{t-1}\right)\right]}{E[\tau]}\right)E[\tau]+\ln 2\right)} \tag{4.40}$$

Since $\mathcal{J}(R)$ is decreasing in R, only thing left to be shown is

$$E\left[\sum_{t=1}^{\tau}\mathcal{I}\left(X_t;Y_t\middle|Y^{t-1}\right)\right] \geq \mathcal{H}(M|Y^0) - h(P_e) - P_e\ln(|\mathcal{M}|-1) \tag{4.41}$$

For that consider the stochastic sequence,

$$V_n = \mathcal{H}(M|Y^n) + \sum_{t=1}^{n}\mathcal{I}\left(X_t;Y_t\middle|Y^{t-1}\right).$$

Clearly $E\left[V_{n+1}|Y^n\right] = V_n$ and $E\left[|V_n|\right] < \infty$, thus $\{V_n\}$ is a martingale. Furthermore $|V_{n+1} - V_n| \leq K$ for some finite K and $E[\tau] < \infty$. Now using a version of Doob's

optional stopping theorem [32, Theorem 2 p 487],

$$V_0 \;=\; E\left[V_\tau\right] \;= = E\left[\mathcal{H}(M|Y^\tau)\right] + E\left[\sum_{t=1}^{\tau} \mathcal{I}\left(X_t; Y_t \left| Y^{t-1}\right.\right)\right] \tag{4.42}$$

One can write Fano's inequality for every Y^τ as follows,

$$\mathcal{H}(M|Y^\tau) \leq h\left(\Pr\left[\hat{M}(Y^\tau) \neq M \middle| Y^\tau\right]\right) + \Pr\left[\hat{M}(Y^\tau) \neq M \middle| Y^\tau\right] \ln(|\mathcal{M}| - 1)$$

Taking average over Y^τ,

$$E\left[\mathcal{H}(M|Y^\tau)\right] \leq E\left[h\left(\Pr\left[\hat{M}(Y^\tau) \neq M \middle| Y^\tau\right]\right)\right] + E\left[\Pr\left[\hat{M}(Y^\tau) \neq M \middle| Y^\tau\right]\right] \ln(|\mathcal{M}| - 1)$$

Using convexity of binary entropy,

$$E\left[\mathcal{H}(M|Y^\tau)\right] \leq h(P_e) + P_e \ln(|\mathcal{M}| - 1) \tag{4.43}$$

Using (4.42) and (4.43), we get the desired condition in (4.41). •

Having proved Lemma 27, we are ready to complete the converse for Theorem 19.

Converse part of Theorem 19: To prove $E_{\text{md}}^{\text{f}} \geq \mathsf{E}_{\mathsf{Red}}$, first recall our assumption of uniformly chosen messages from $\mathcal{M}^{(\kappa)}$, i.e., $\Pr\left[M = i\right] = \frac{1}{|\mathcal{M}^{(\kappa)}|}$. The error probability $P_e(i) = \Pr\left[\hat{M} \neq i | M = i\right]$ for any i can be bounded as

$$-\frac{\ln P_e(i)}{E\left[\tau^{(\kappa)}\right]} \leq \frac{1}{1 - P_e{}^{(\kappa)} - \frac{1}{|\mathcal{M}^{(\kappa)}|}} \left(\mathcal{J}\left(\frac{\ln|\mathcal{M}^{(\kappa)}| - h(P_e(\kappa)) - P_e{}^{(\kappa)} \ln(|\mathcal{M}(\kappa)| - 1)}{E\left[\tau^{(\kappa)}\right]}\right) + \frac{\ln 2}{E\left[\tau^{(\kappa)}\right]}\right) \tag{4.44}$$

Thus for any message i in a capacity-achieving sequence with feedback

$$\lim_{\kappa \to \infty} -\frac{\ln P_e(i)}{E\left[\tau^{(\kappa)}\right]} \leq \mathcal{J}\left(C\right) = \mathsf{E}_{\mathsf{Red}} \tag{4.45}$$

4.4.5 Proof of Theorem 22

In this subsection, we show how the strategy for sending a special bit can be combined with the Yamamoto-Itoh strategy when many special messages demand a

missed-detection exponent. However unlike previous theorems, we add the additional uniform-delay constraint now[3].

Clearly capacity-achieving sequences in general need not be uniform delay. Many messages can have an expected delay, $E[\tau|\, M = i]$ much larger than the average delay, $E[\tau]$. This in return can decrease the error probability of these messages. The potential drawback of such codes is that their average delay is sensitive to the assumption of uniformly chosen messages. The expected decoding time, $E[\tau]$, can increase a lot if messages are not chosen uniformly.

It is worth emphasizing that all previously discussed exponents (single message exponent E_{md}^{f}, single bit exponent E_{b}^{f}, many bits exponent $E_{\text{b}}^{\text{f}}(r)$ and achievable multi-layer exponent regions) remain unchanged whether or not this uniform-delay constraint is imposed. Thus the flexibility to provide different expected delays to different messages does not improve these exponents.

However, this is not true for message-wise UEP. Removing the uniform-delay constraint can considerably enhance the protection of special messages at rate higher than $(1 - \frac{\mathsf{E}_{\mathsf{Red}}}{D_{\max}})C$. In fact, all special messages can achieve $\mathsf{E}_{\mathsf{Red}}$ then. The flexibility of providing more resources (decoding delay) to special messages achieves this gain. However, we do not discuss those cases here and stick to uniform-delay codes.

Achievability: $E_{\mathbf{md}}^{\text{f}}(r) \geq \min\{\mathsf{E}_{\mathsf{Red}}, (1 - \frac{r}{C})\boldsymbol{D}_{\max}\}$

The optimal scheme here reverses the trick for achieving E_{b}^{f}: now a special bit tells the receiver whether the message being transmitted is special or not. This further emphasizes how feedback connects bit-wise and message-wise UEP.

Proof Like all the previous achievability results, we will construct a capacity-achieving sequence $\mathcal{Q}$. To create the κ^{th} code in $\mathcal{Q}$, a multi-phase fixed-length code with erasure decoding will be used. The first phase uses a length $\lceil\sqrt{\kappa}\rceil$ non-feedback code with two codewords, to tell whether $M \in \mathcal{M}_s^{(\kappa)}$ or not. Let $b = \mathbb{I}_{\left\{M \in \mathcal{M}_s^{(\kappa)}\right\}}$ denote the

[3]Recall that for any reliable variable-length block code with feedback $\Gamma = \frac{\max_{i \in \mathcal{M}} E[\tau|M=i]}{E[\tau]}$ and uniform-delay codes are those which satisfy $\lim_{\kappa\to\infty} \Gamma_{\mathcal{Q}}^{(\kappa)} = 1$.

indicator bit for special messages. As a result of [2, Theorem 5.7.1, page 153]

$$\Pr\left[\hat{b} \neq 1 \middle| b = 1\right] = \Pr\left[\hat{b} \neq 0 \middle| b = 0\right] \leq e^{-\sqrt{\kappa} E_{ex}\left(\frac{\ln 8}{\sqrt{\kappa}}\right)} \tag{4.46}$$

where $E_{ex}(\cdot)$ stands for the expurgated exponent. This probability vanishes with κ. In the second phase, one of two codes is used depending on $\hat{b}$.

- If $\hat{b} = 0$ after the first phase, the transmitter will use the length κ code in a capacity-achieving sequence (without feedback) $\mathcal{Q}'$ such that $E_{\mathrm{md},\mathcal{Q}'} = \mathsf{E}_{\mathsf{Red}}$. The message, ϑ for $\mathcal{Q}'$ will be decided according to the following mapping:

$$\begin{aligned} M \in \mathcal{M}_s^{(\kappa)} &\Rightarrow \vartheta = 0 \qquad \text{(buzzer message)} \\ M \notin \mathcal{M}_s^{(\kappa)} &\Rightarrow \vartheta = M - |\mathcal{M}_s^{(\kappa)}| \end{aligned}$$

  The receiver decodes $\hat{\vartheta}$ at the end of the second phase. If $\hat{\vartheta} = 0$, then an erasure is declared. If $\vartheta \neq 0$, then $\hat{M} = \hat{\vartheta} + |\mathcal{M}_s^{(\kappa)}|$. Since $\mathcal{Q}'$ achieved $\mathsf{E}_{\mathsf{Red}}$, we get

$$\Pr\left[\hat{M} \notin \mathcal{M}_s^{(\kappa)} \middle| M \in \mathcal{M}_s^{(\kappa)}\right] \doteq \exp(-\kappa \mathsf{E}_{\mathsf{Red}}) \tag{4.47}$$

- If $\hat{b} = 1$ after the first phase, the transmitter uses a two phase code of length κ, as in the Yamamoto-Itoh scheme [29]. The two phases of this code are called communication and control phases, respectively. First, the length $\lceil r\kappa/C \rceil$ code in $\mathcal{Q}'$ is used to convey the particular special message if $M \in \mathcal{M}_s^{(\kappa)}$. The last codeword in this code is used when $M \notin \mathcal{M}_s^{(\kappa)}$. Thus the message ϑ in the second phase is given by

$$\begin{aligned} M \in \mathcal{M}_s &\Rightarrow \vartheta = M \\ M \notin \mathcal{M}_s &\Rightarrow \vartheta = |\mathcal{M}_s^{(\kappa)}|+1 \end{aligned}$$

  At the end of the communication phase, the receiver forms a temporary decision

$\tilde{\vartheta}$ using ML decoding. Using [2, Corollary , page 140] implies

$$\Pr\left[\tilde{\vartheta} \neq \vartheta \middle| \vartheta = i\right] \leq 4e^{-\kappa \frac{r}{C} E_r(\frac{r_\kappa}{r} C)} \qquad \forall 1 \leq i \leq |\mathcal{M}_s^{(\kappa)|+1} \quad (4.48)$$

where $r_\kappa \equiv \frac{\log(|\mathcal{M}_s^{(\kappa)|}+1|)}{\lceil r\kappa/C \rceil}$ denotes the exact rate of the codebook for ϑ. In the control phase, temporary decision $\tilde{\vartheta} = \vartheta$, is confirmed by sending accept symbol x_a for $\kappa - \lceil \frac{r}{C}\kappa \rceil$ time units, and is rejected by sending reject symbol x_d instead. If $b_c = \mathbb{I}_{\{\tilde{\vartheta} \neq \vartheta\}}$, then confirmation codeword corresponds to $b_c = 0$ and the rejection codeword corresponds to $b_c = 1$. The receiver decodes $\hat{b}_c = 0$ if the output type of the confirmation phase is $W_{Y|X}(\cdot|x_a)$ and $\hat{b}_c = 1$ otherwise. This ensures

$$\Pr\left[\hat{b}_c = 0 \middle| b_c = 1\right] \doteq \exp(-\kappa(1 - r/C)D_{\max}) \quad (4.49)$$

If $\hat{b}_c = 0$, then receiver finalizes $\hat{\vartheta} = \tilde{\vartheta}$ and otherwise an erasure is declared. If $\hat{\vartheta} = |\mathcal{M}_s| + 1$ or an erasure is declared for ϑ, we declare an erasure for the whole block and initiate retransmission afresh. Otherwise, the final message is decoded as $\hat{M} = \hat{\vartheta}$.

Thus a special message M is decoded wrongly to another special message when $\hat{b}_c = 0$. Hence the exponent of erring to between two special messages equals $(1 - r/C)D_{\max}$.

Eq. (4.47) and (4.49) show that each special message achieves a missed-detection exponent of $\min\{\mathsf{E}_{\mathsf{Red}}, (1 - r/C)D_{\max}\}$. Moreover, the overall erasure probability of this scheme also vanishes, and hence capacity is achieved in spite of retransmissions. •

Converse: $E_{\mathbf{md}}^{\mathsf{f}}(r) \leq \min\{\mathsf{E}_{\mathsf{Red}}, (1 - \frac{r}{C})\boldsymbol{D}_{\max}\}$

Proof Consider any uniform-delay capacity-achieving sequence with feedback $\mathcal{Q}$. Excluding all ordinary messages from the original κ'th code in $\mathcal{Q}$ gives a code with $|\mathcal{M}_s^{(\kappa)}| = \lceil e^{E[\tau^{(\kappa)}]r} \rceil$ messages. Its average error probability $P_e{}'^{(\kappa)}$ and average delay

$E\left[\tau'^{(\kappa)}\right]$ satisfy,

$$P_e' \leq \Pr\left[\hat{M} \neq M | M \in \mathcal{M}_s^{(\kappa)}\right]$$
$$E\left[\tau'^{(\kappa)}\right] \leq \Gamma^{(\kappa)} E\left[\tau^{(\kappa)}\right]$$

Consequently,

$$\frac{-\ln \Pr\left[\hat{M} \neq M \middle| M \in \mathcal{M}_s^{(\kappa)}\right]}{E\left[\tau^{(\kappa)}\right]} \leq -\left(\frac{\ln P_e'^{(\kappa)}}{E\left[\tau'^{(\kappa)}\right]}\right)\Gamma^{(\kappa)} \tag{4.50}$$

Remember that by the uniform-delay assumption, $\Gamma^{(\kappa)}$ tends to 1 when κ goes infinity. Hence the right hand side above tends to error exponent of a feedback code of rate r. This exponent is at most [27] the Burnashev exponent $(1 - \frac{r}{C})D_{\max}$. Since limit of left hand side is at most $E_{\text{md}}^{\text{f}}(r)$ by definition, we get $E_{\text{md}}^{\text{f}}(r) \leq (1 - \frac{r}{C})D_{\max}$. Similarly by excluding all but one of special messages in $\mathcal{M}_s$, we get $E_{\text{md}}^{\text{f}}(r) \leq \mathsf{E}_{\mathsf{Red}}$. •

4.5 Avoiding False Alarms: Proofs

4.5.1 Block Codes without Feedback: Proof of Theorem 24

Lower Bound: $E_{\mathbf{fa}} \geq E_{\mathbf{fa}}^{\mathbf{l}}$

Proof As a result of the coding theorem [4, Ch. 2 Corollary 1.3, page 102] we know that there exits a capacity-achieving sequence $\mathcal{Q}'$ of fixed composition codes (without feedback). Let $P_X^{(n)}$ denote the composition of the length n code in this sequence, which satisfies

$$\sum_{x \in \mathcal{X}} |P_X^{(n)}(x) - P_X^*(x)| \leq \sqrt[4]{\frac{1}{n}}.$$

We use the codewords in this length n code in $\mathcal{Q}'$ as the codewords for ordinary messages in the length n code in $\mathcal{Q}$. For the special message, we use the length n repetition sequence $\bar{x}^n(1) = (x_l, x_l, \cdots, x_l)$.

The decoding region $\mathcal{G}(1)$ for the special message will be essentially the bare

minimum. We include the typical channel outputs within the decoding region of the special message to ensure small missed-detection probability for the special message, but we do not include any other output sequence.

$$\mathcal{G}(1) = \{y^n : \sum_{i \in \mathcal{Y}} |\mathsf{Q}_y(y^n)(i) - W_{Y|X}(i|x_l)| \leq \sqrt[4]{1/n}\}$$

Note that this definition of $\mathcal{G}(1)$ itself ensures that the special message is transmitted reliably when it is sent, $P_e^{(n)}(1) \to 0$.

The decoding regions of an ordinary message $j \in \{2, 3, \ldots, |\mathcal{M}^{(n)}|\}$ is simply its corresponding decoding region in $\mathcal{Q}'$ which lies outside of $\mathcal{G}(1)$. Thus the fact that $\mathcal{Q}'$ is a reliable sequence will imply,

$$\lim_{n\to\infty} \Pr\left[y^n \in \bigcup_{j \notin \{1,i\}} \mathcal{G}(j) \middle| M = i\right] = 0.$$

Consequently the only thing we are left to prove is that decay rate of the $\Pr\left[\hat{M} = 1 \middle| M \neq 1\right]$ is fast enough. Note the probability of a V-shell of a message i is equal to given by,

$$\Pr\left[\mathsf{T}_V(i) | M = i\right] = e^{-nD\left(V_{Y|X}(\cdot|X) \| W_{Y|X}(\cdot|X) | P_X^{(n)}\right)}$$

Note that also that $\mathcal{G}(1)$ can be written as the union of V-shells of a message i as,

$$\mathcal{G}(1) = \bigcup_{V_{Y|X} \in \mathcal{V}^{(n)}} \mathsf{T}_V(i) \qquad \forall i \neq 1$$

where $\mathcal{V}^{(n)}$ is the set of channel types $V_{Y|X}$ whose corresponding marginal output distribution $(P_X^{(n)} V_{Y|X})_Y$ is close to $W_{Y|X}(\cdot|x_l)$.

$$\mathcal{V}^{(n)} = \{V_{Y|X} : \sum_j |\sum_k V_{Y|X}(j|k) P_X^n(k) - W_{Y|X}(j|x_l)| \leq \sqrt[4]{1/n}\}.$$

Note that since there are at most $(1+n)^{|\mathcal{X}||\mathcal{Y}|}$ different conditional types.

$$\Pr\left[\mathcal{G}(1)\middle|\, M=i\right] \leq (1+n)^{|\mathcal{X}||\mathcal{Y}|} \max_{V_{Y|X}\in\mathcal{V}^{(n)}} \Pr\left[\mathsf{T}_V\left(i\right)\middle|\, M=i\right]$$

Thus for all $i \neq 1$,

$$\lim_{n\to\infty} \frac{-\log \Pr\left[\mathcal{G}(1)|M=i\right]}{n} = \min_{V_{Y|X}:\ \sum_j P_X^*(j) V_{Y|X}(\cdot|j)=W_{Y|X}(\cdot|x_l)} D\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle|\, P_X^*\right)$$

•

Upper Bound: $E_{\mathbf{fa}} \leq E_{\mathbf{fa}}^{\mathbf{u}}$

Proof The data processing inequality for KL divergence implies

$$\begin{aligned}\sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle|\, M=1\right] \log \frac{\Pr[y^n|M=1]}{\Pr[y^n|M\neq 1]} &\geq \Pr\left[\mathcal{G}(1)\middle|\, M=1\right] \log \frac{\Pr[\mathcal{G}(1)|M=1]}{\Pr[\mathcal{G}(1)|M\neq 1]} \\ &\quad + \Pr\left[\overline{\mathcal{G}(1)}\middle|\, M=1\right] \log \frac{\Pr\left[\overline{\mathcal{G}(1)}\middle|M=1\right]}{\Pr\left[\overline{\mathcal{G}(1)}\middle|M\neq 1\right]} \qquad (4.51)\\ &\geq -\ln 2 - \Pr\left[\mathcal{G}(1)\middle|\, M=1\right] \log \Pr\left[\mathcal{G}(1)\middle|\, M\neq 1\right] \qquad (4.52)\end{aligned}$$

Using the convexity of the KL divergence we we get

$$\begin{aligned}\sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle|\, M=1\right] \log \frac{\Pr[y^n|M=1]}{\Pr[y^n|M\neq 1]} &\leq \sum_{i=2}^{|\mathcal{M}|} \frac{1}{|\mathcal{M}|-1} \sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle|\, M=1\right] \log \frac{\Pr[y^n|M=1]}{\Pr[y^n|M=i]} \\ &= \sum_{i=2}^{|\mathcal{M}|} \frac{1}{|\mathcal{M}|-1} \sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle|\, M=1\right] \sum_{k=1}^{n} \log \frac{\Pr\left[y_k|M=1,y^{k-1}\right]}{\Pr\left[y_k|M=i,y^{k-1}\right]} \\ &= \sum_{k=1}^{n} \sum_{i=2}^{|\mathcal{M}|} \frac{1}{|\mathcal{M}|-1} D\left(W_{Y|X}(\cdot|\bar{x}_k(1))\middle\| W_{Y|X}(\cdot|\bar{x}_k(i))\right)\end{aligned} \qquad (4.53)$$

where $\bar{x}_k(i)$ denotes the input letter for codeword of message i, at time k.

Let P_{X_k} denote the empirical distribution of $\bar{x}_k(i)$ for a fixed time k.

$$P_{X_k}(x) \ \equiv = \frac{\sum_{i\in\mathcal{M}} \mathbb{I}_{\{\bar{x}_k(i)=x\}}}{|\mathcal{M}|} \quad \forall x \in \mathcal{X}$$

It is the fraction of codewords whose input at time k equals i. Using equation (4.52) and (4.53) we get

$$\Pr\left[\mathcal{G}(1)|\, M \neq 1\right] \geq e^{-\frac{1}{\Pr[\mathcal{G}(1)|M=1]}\left(\frac{|\mathcal{M}|}{|\mathcal{M}|-1}\sum_k D\left(W_{Y|X}(\cdot|\bar{x}_k(1))\middle\|W_{Y|X}(\cdot|X_k)\middle|P_{X_k}\right)-\ln 2\right)} \tag{4.54}$$

We show below that for all capacity-achieving codes, P_{X_k} for almost all the k's is essentially equal to P_X^*. For that purpose, let us first define the set $\mathcal{P}_\epsilon$ and $\delta(\epsilon)$.

$$\mathcal{P}_\epsilon \equiv \{P_X : I(P_X, W_{Y|X}) \geq C - \epsilon\} \quad \text{and} \quad \delta(\epsilon) \equiv \max_{P_X \in \mathcal{P}_\epsilon} \sum_i |P_X(i) - P_X^*(i)|$$

Note that $\lim_{\epsilon \to 0} \delta(\epsilon) = 0$.

Let us now show that P_{X_k} is essentially P_X^* for almost all of the k's. First, note that as a result of Fano's inequality we get,

$$I(M, Y^n) \geq nR^{(n)}(1 - P_e) + \ln 2 \tag{4.55}$$

On the other hand using standard manipulations on mutual information we get

$$\begin{aligned} I(M; Y^n) &\leq \sum_{k=1}^{n} I(X_k, Y_k) \\ &= \sum_{k=1}^{n} I(P_{X_k}, W_{Y|X}) \\ &\leq nC - \epsilon \sum_{k=1}^{n} \mathbb{I}_{\{P_{X_k} \in \mathcal{P}_\epsilon\}} \end{aligned}$$

Inserting this in to equation (4.55) we get,

$$\sum_{k=1}^{n} \mathbb{I}_{\{P_{X_k} \notin \mathcal{P}_\epsilon\}} \leq n \frac{(C - R^{(n)}(1 - P_e) - \ln 2/n)}{\epsilon} \tag{4.56}$$

Thus the fraction of time-indices at which P_{X_k} is not $\mathcal{P}_{\epsilon^{(n)}}$ goes to 0 as $\epsilon^{(n)}$.

Let us chose $\epsilon^{(n)} = \sqrt{(C - R^{(n)}(1 - P_e) - \ln 2/n)}$. Then for any capacity-achieving

sequence $\lim_{n\to\infty} \epsilon^{(n)} = 0$. Moreover,

$$\sum_{k=1}^{n} \mathbb{I}_{\left\{P_{X_k} \notin \mathcal{P}_{\epsilon^{(n)}}\right\}} \leq n\epsilon^{(n)} \tag{4.57}$$

Note for any $P_X \in \mathcal{P}_{\epsilon^{(n)}}$ we have

$$\begin{aligned} D\left(W_{Y|X}(\cdot|\bar{x}_k(1))\middle\| W_{Y|X}(\cdot|X_k)\middle| P_X\right) &\leq D\left(W_{Y|X}(\cdot|x_k(1))\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right) + \delta(\epsilon^{(n)})D_{\max} \\ &\leq E_{\text{fa}}^{\text{u}} + \delta(\epsilon^{(n)})D_{\max} \end{aligned} \tag{4.58}$$

where $E_{\text{fa}}^{\text{u}} = \max_{\hat{x}\in\mathcal{X}} D\left(W_{Y|X}(\cdot|\hat{x})\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right)$.

Using equations (4.57) and (4.58)

$$\sum_k D\left(W_{Y|X}(\cdot|x_k(1))\middle\| W_{Y|X}(\cdot|X_k)\middle| P_{X_k}\right) \leq n(E_{\text{fa}}^{\text{u}} + \delta(\epsilon^{(n)})D_{\max} + \epsilon^{(n)}D_{\max}) \tag{4.59}$$

Inserting this in equation (4.54) gives

$$\lim_{n\to\infty} \frac{-\log \Pr\left[\mathcal{G}(1)|M\neq 1\right]}{n} \leq E_{\text{fa}}^{\text{u}} \tag{4.60}$$

because $\Pr\left[\mathcal{G}(1)|M=1\right]$ tends to 1 to ensure vanishing conditional error probability for the special message. It remains to show how this upper bounds E_{fa} which is defined in terms of the worst case false-alarm probability ordinary messages. By Baye's rule and assuming uniformly chosen messages from $\mathcal{M}$,

4.5.2 Variable-Length Block Codes with Feedback: Proof of Theorem 26

Achievability: $E_{\text{fa}}^{\text{f}} \geq \boldsymbol{D}_{\max}$

Proof We construct a capacity-achieving sequence with feedback $\mathcal{Q}$, by using a construction similar to the one achieving $E_{\text{md}}^{\text{f}}(r)$. In fact with this scheme, the special

message achieves the false-alarm exponent $D_{\max}$ simultaneously with the best missed-detection exponent $\mathsf{E}_{\mathsf{Red}}$. The κ^{th} element of the code is formed as follows.

First, use a fixed length two code with erasure decoding. In the first phase of the code, a length $\lceil \kappa \rceil$ code of two messages is used to convey whether $M = 1$ or not. Let $b = \mathbb{I}_{\{M=1\}}$ denote a special bit which is an indicator of the special message. Using [2, Theorem 5.7.1, page 153],

$$\Pr\left[\hat{b} \neq 1 \middle| b = 1\right] = \Pr\left[\hat{b} \neq 0 \middle| b = 0\right] \leq e^{-\sqrt{\kappa} E_{ex}\left(\frac{\ln 8}{\sqrt{\kappa}}\right)} \tag{4.61}$$

where $E_{ex}(\cdot)$ stands for the expurgated exponent. This error probability vanishes with κ.

In the second phase, one of two codes will be used depending on $\hat{b}$.

- If $\hat{b} = 0$, transmitter uses the length κ code in capacity-achieving sequence $\mathcal{Q}'$ such that $E_{\mathrm{md},\mathcal{Q}'} = \mathsf{E}_{\mathsf{Red}}$. This code conveys the particular ordinary message to transmitted or a buzzer message indicating missed-detection of the special message. If ϑ denotes the message for this code in $\mathcal{Q}'$,

$$\begin{aligned} \hat{b} \neq b = 1 &\Rightarrow \vartheta = 0 \qquad \text{(buzzer message in } \mathcal{Q}') \\ \hat{b} = b = 0 &\Rightarrow \vartheta = M \end{aligned}$$

  If a buzzer is detected after the second phase, $\hat{\vartheta} = 0$, an erasure is declared and a retransmission is initiated afresh. Otherwise, it declares M as $\hat{M} = \hat{\vartheta}$. This strategy achieves $\mathsf{E}_{\mathsf{Red}}$ as the missed-detection exponent for the special message.

$$\Pr\left[\hat{M} \neq 1 \middle| M = 1\right] \doteq \exp(-\kappa \mathsf{E}_{\mathsf{Red}}).$$

- If $\hat{b} = 1$ after the first phase, transmitter uses a length κ repetition code to confirm whether $M = 1$ or not. If $M = 1$, transmitter sends the codeword $(x_a, x_a, \ldots, x_a)$. If $M \neq 1$, transmitter sends the codeword $(x_d, x_d, \ldots, x_d)$. Confirming $M = 1$ only if the output type of the second phase is $W_{Y|X}(\cdot|x_a)$

ensures

$$\Pr\left[\hat{M}=1\middle|\, M\neq 1\right] \doteq \exp(-\kappa D(W_{Y|X}(\cdot|x_a)\|W_{Y|X}(\cdot|x_d))) = \exp(-\kappa D_{\max}).$$

An erasure is declared for all other types of the second phase and a retransmission is initiated afresh.

As in earlier proofs, the overall erasure probability vanishes and $\mathcal{Q}$ achieves capacity in spite of the the retransmissions. •

Converse: $E_{\mathbf{fa}}^{\mathrm{f}} \leq \boldsymbol{D}_{\max}$

Proof Let $\mathcal{G}(i)$ denote the decoding region of each message i:

$$\mathcal{G}(i) = \{y^\tau : \hat{M} = i\}$$

As result of convexity of KL divergence,

$$\begin{aligned} E\left[\log\tfrac{\Pr[Y^\tau|M=1]}{\Pr[Y^\tau|M\neq 1]}\middle|\, M=1\right] &\geq \Pr\left[\mathcal{G}(1)|\, M=1\right]\log\tfrac{\Pr[\mathcal{G}(1)|M=1]}{\Pr[\mathcal{G}(1)|M\neq 1]} + \Pr\left[\overline{\mathcal{G}(1)}\middle|\, M=1\right]\log\tfrac{\Pr\left[\overline{\mathcal{G}(1)}\middle|M=1\right]}{\Pr\left[\overline{\mathcal{G}(1)}\middle|M\neq 1\right]} \\ &\geq -\ln 2 + \Pr\left[\mathcal{G}(1)|\, M=1\right]\log\tfrac{1}{\Pr[\mathcal{G}(1)|M\neq 1]} \end{aligned} \tag{4.62}$$

It has already been proved in [28] that,

$$E\left[\log\tfrac{\Pr[Y^\tau|M=1]}{\Pr[Y^\tau|M\neq 1]}\middle|\, M=1\right] \leq D_{\max} E\left[\tau|\, M=1\right] \tag{4.63}$$

As a result of definition of Γ we have $E\left[\tau|\, M=1\right] \leq E\left[\tau\right]\Gamma$. Combining this with (4.62) and (4.63) gives,

$$\Pr\left[\mathcal{G}(1)|\, M\neq 1\right] \geq e^{-\frac{\ln 2+\Gamma D_{\max} E[\tau]}{\Pr[\mathcal{G}(1)|M=1]}} \tag{4.64}$$

Since Γ tends to 1 due to uniform constraint on $\mathcal{Q}$, we get $E_{\mathrm{fa},\mathcal{Q}}^{\mathrm{f}} \leq D_{\max}$. •

Chapter 5

Unequal Error Protection at Rates Below Capacity

In this chapter, we relax the constraint of overall data-rate approaching capacity. Sacrificing the data-rate from capacity should provide additional reliability. Now we are interested in optimal tradeoffs between reliability of the crucial parts of information vs. reliability of its ordinary parts. We formulate this question in two ways. This chapter addresses the first formulation, which is based on error exponents as in Chapters 3 and 4. The second formulation is based on network information theory and is postponed till the next chapter.

For data rates strictly below capacity, even the ordinary parts of information can achieve a positive error exponent. One can tradeoff this exponent for a better exponent for special information. To understand this tradeoff, we analyze the optimal region of exponent pairs $(E_{\text{special}}, E_{\text{ordinary}})$, where E_{special} is the exponent for special information and E_{ordinary} is the same for ordinary information. Contrast this with the case of operating at capacity, where we only focused on finding the maximum E_{special} exponent for the special parts because E_{ordinary} was always zero. Since even ordinary information can achieve a positive exponent at rates below capacity, the number of UEP problems becomes much bigger now. We will analyze only a representative sample of this rich set of UEP problems. In particular, we will revisit a few UEP scenarios from Chapters 3 and 4.

Throughout this chapter, we assume that the overall data rate equals $R < C$. In Section 5.1, we discuss the case of block codes without feedback. We first consider a bit-wise UEP scenario where the special bits have rate r_1 and the ordinary bits have rate $r_2 \equiv R - r_1$. We obtain upper bounds on achievable exponents which are similar to sphere-packing bounds for classical error exponents. We also obtain lower bounds using random-coding arguments. The upper and lower bounds match in the high rate region for symmetric channels like the BSC and very noisy channels[1]. We then consider the message-wise problem of a single special message and calculate the generalization of the Red-Alert Exponent for rates below capacity—the best missed-detection exponent for a special message while ensuring reliable communication at rate R.

In Section 5.2, we discuss the case of variable-length block codes with feedback. We revisit the bit-wise UEP scenario where the special bits have rate r_1 and the ordinary bits have rate $r_2 \equiv R - r_1$. An achievable bound on error exponents is provided using a simple retransmission based protocol. Section 5.3 contains proof details for all results in this chapter.

5.1 UEP Exponents for Block Codes

5.1.1 Many special bits

We first analyze the situation where out of the total $nR/\ln 2$ (approx.) bits, approximately $nr_1/\ln 2$ bits are special. For that purpose, consider a reliable code sequence $\mathcal{Q}$ with message sets of the form $\mathcal{M}^{(n)} = \mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)}$, where the cardinality of $\mathcal{M}_1^{(n)}$ and $\mathcal{M}_2^{(n)}$ grows exponentially at rate r_1 and $r_2 \equiv R - r_1$, respectively. This simply says that the rate of special bits and ordinary bits equals r_1 and r_2, respectively. The error exponent pair $(\bar{E}_{\text{bits},1,\mathcal{Q}}, \bar{E}_{\text{bits},2,\mathcal{Q}})$ for such a code-sequence $\mathcal{Q}$ is defined as follows.

Definition 28 *Consider a reliable code sequence $\mathcal{Q}$ with message sets of the form*

[1]These bounds on UEP exponents do not match at low rates, but remember that even the classical error exponents are still unknown at low rates.

$\mathcal{M}^{(n)} = \mathcal{M}_1^{(n)} \times \mathcal{M}_2^{(n)}$, *where*

$$r_{1,\mathcal{Q}} \equiv \lim_{n\to\infty} \frac{\log|\mathcal{M}_1^{(n)}|}{n}, \qquad r_{2,\mathcal{Q}} \equiv \lim_{n\to\infty} \frac{\log|\mathcal{M}_2^{(n)}|}{n}.$$

Then the pair of exponents for special bits and ordinary bits is defined as,

$$\bar{E}_{bits,1,\mathcal{Q}} = \liminf_{n\to\infty} \frac{-\log \Pr\left[\hat{M}_1 \neq M_1\right]}{n}, \qquad \bar{E}_{bits,2,\mathcal{Q}} = \liminf_{n\to\infty} \frac{-\log \Pr\left[\hat{M}_2 \neq M_2\right]}{n} \tag{5.1}$$

Now the achievable error exponent region at rate-pair (r_1, r_2) *is defined as*[2] *the set of all achievable exponent pairs:*

$$\bar{\mathcal{E}}(r_1, r_2) \equiv \{(\bar{E}_{bits,1,\mathcal{Q}}, \bar{E}_{bits,2,\mathcal{Q}}) \; \exists \; \mathcal{Q} \; s.t. \; r_{1,\mathcal{Q}} \geq r_1 \; and \; r_{2,\mathcal{Q}} \geq r_2\}.$$

In the following, we first discuss a converse result based on sphere-packing arguments. These bit-wise UEP exponents are stated assuming our channel $W_{Y|X}$ to be a binary symmetric channel of crossover probability p. Similar results can be stated for any symmetric channel $W_{Y|X}$ for which $\mathcal{X} = \mathcal{Y}$, all off-diagonal entries of $W_{Y|X}$ are equal to each other, and all diagonal entries are equal to each other. The reason for focusing on these channel classes is their simplicity—such channels are completely described by a single scalar parameter denoting their noise-level.

After discussing the BSC case, we describe a similar converse result for very noisy channels as in [2] and Chapter 6. The expressions for UEP exponents are particularly simple for such channels. These expressions generalize the classical sphere-packing exponents for very noisy channels in [2]. These expressions also provide an intuitive interpretation of UEP exponents where the pie of channel resources is first divided between the special bits and ordinary bits. Each type of bits can then separately decide the utilization of its piece of pie for its rate and error exponent.

After discussing the converse results for the BSC and the very noisy channels, we discuss their achievability results based on superposition coding. For these channels,

[2]All exponents and exponent regions in this chapter will have a bar on top (e.g., $\bar{E}_{\text{bits},1,\mathcal{Q}}$ and $\bar{\mathcal{E}}(r_1, r_2)$) to emphasize the rate below capacity.

these achievability results match the converse results in the high rate region. This is similar to classical error exponents where the sphere-packing upper bound matches the random coding lower bound at high rates. For more general channels, similar converse and achievability results can be stated. However, they need not match even in the high rate region.

Binary Symmetric Channel: Converse

Instead of describing the best error exponent pairs achievable for a given (r_1, r_2), we will describe the best possible (r_1, r_2) for a given pair $(\bar{E}_1, \bar{E}_2)$ of error exponents. Here $\bar{E}_1$ and $\bar{E}_2$ are non-negative numbers and denote the exponent required for special bits and ordinary bits, respectively. To describe this converse, let us define the achievable rate region $\bar{\mathcal{R}}(\bar{E}_1, \bar{E}_2)$ for given exponent pair $(\bar{E}_1, \bar{E}_2)$. This is defined on similar lines of the exponent region $\bar{\mathcal{E}}(r_1, r_2)$ for given rate pair (r_1, r_2).

$$\bar{\mathcal{R}}(\bar{E}_1, \bar{E}_2) \equiv \{(r_{1,\mathcal{Q}}, r_{2,\mathcal{Q}}) : \ \mathcal{Q} \text{ s.t. } \bar{E}_{\text{bits},1,\mathcal{Q}} \geq \bar{E}_1 \text{ and } \bar{E}_{\text{bits},2,\mathcal{Q}} \geq \bar{E}_2\}.$$

We will upper bound this rate region in terms of the capacity region of a certain BSC broadcast channel[3]. This broadcast channel has two users and the channel for user $j \in \{1, 2\}$ is a BSC with crossover probability $p_j \geq p$. This crossover probability p_j is defined as follows in terms of $\bar{E}_j$.

$$D_{\text{b}}(p_1 \| p) \equiv \bar{E}_1, \qquad D_{\text{b}}(p_2 \| p) \equiv \bar{E}_2 \qquad (\text{s.t. } p \leq p_1, p_2) \tag{5.2}$$

where $D_{\text{b}}(h \| g)$ denotes the KL divergence between two Bernoulli distributions of parameters h and g.

By Stein's lemma, $D_{\text{b}}(p_1 \| p)$ equals the exponent of the observed crossover probability being p_1 or more when p is the crossover probability of the actual channel $W_{Y|X}$. Since ensuring better exponent for special bits implies $\bar{E}_1 \geq \bar{E}_2$, we have $p \leq p_2 \leq p_1$.

[3]Such a connection between the UEP problem and the broadcast channel was also used in [18] to analyze the rate-region of priority encoded transmission [17] over erasure channels.

Theorem 29 *Consider a broadcast channel where channel for user i is a BSC of crossover probability p_i. If $\mathcal{C}_{broadcast}(p_1, p_2)$ denotes the capacity region of this channel,*

$$\bar{\mathcal{R}}(\bar{E}_1, \bar{E}_2) \subseteq \mathcal{C}_{broadcast}(p_1, p_2).$$

Equivalently, if a rate-pair (r_1, r_2) is not achievable over this broadcast channel, it is also not achievable in our bit-wise UEP problem. This converse is a generalization of the sphere-packing bound for classical error exponents [3]. It follows because the classical sphere-packing bound can be stated as an upper bound on the achievable rate when the (classical) error exponent should be at least $\bar{E}$. To be precise, define $\tilde{p} \geq p$ as the crossover probability which satisfies $D_{\mathrm{b}}(\tilde{p}\|p) \equiv \bar{E}$. If $C_{\mathrm{b}}(\cdot)$ denotes the capacity of a BSC as a function of its crossover probability, the achievable rate with error exponent $\bar{E}$ is at most $C_{\mathrm{b}}(\tilde{p}) \equiv r_{\mathrm{sp}}(\bar{E})$. This result can be recovered by substituting $\bar{E}_1 = \bar{E}_2 = \bar{E}$ in Theorem 29.

Now we describe a weaker analogue of the $E_{\mathrm{b}} = 0$ result at capacity in Chapter 3. Assume $r_1, r_2 > 0$ and say that even the ordinary bits demand the sphere packing exponent at sum rate R. Equivalently, the sum rate R equals $r_{\mathrm{sp}}(\bar{E}_2)$.

$$\bar{E}_2 = E_{\mathrm{sp}}(R) \qquad \Leftrightarrow \qquad R = r_{\mathrm{sp}}(\bar{E}_2)$$

This is the best $\bar{E}_2$ at sum rate R since the exponent for special bits is at least $\bar{E}_2$. Under this constraint on $\bar{E}_2$, Theorem 29 implies that $\bar{E}_1$ for special bits cannot be larger than $\bar{E}_2$ for any $r_1 > 0$. Thus Theorem 29 essentially generalizes the single special bit result in Chapter 3 for the case of many special bits (of rate r_1). However, it does not imply that even a single special bit cannot achieve any higher exponent than $\bar{E}_2 = E_{\mathrm{sp}}(R) = E_{\mathrm{sp}}(r_2)$.

Very Noisy Channel: Converse

In a very noisy channel [2], the conditional output distributions $\{W_{Y|X}(\cdot|x) : x \in \mathcal{X}\}$ are 'very' close to each other. This implies that every $W_{Y|X}(\cdot|x)$ is also 'very'

close to the capacity achieving distribution $Q_Y^*(\cdot)$, which is a convex combination of $\{W_{Y|X}(\cdot|x) : x \in \mathcal{X}\}$. To be precise, a channel $W_{Y|X}$ is very noisy if it has the characterization below and ϵ tends to 0.

$$\begin{aligned} W_{Y|X}(\cdot|x) &= Q_Y^*(\cdot) + \epsilon\Theta_{Y|X}(\cdot|x) \quad \text{where} \quad \sum_y \Theta_{Y|X}(y|x) = 0 \\ W_{Y|X} &= \mathbf{1}'Q_Y^* + \epsilon\Theta_{Y|X} \quad \text{where} \quad \Theta_{Y|X}\mathbf{1}' = \mathbf{0}'. \end{aligned} \tag{5.3}$$

Here $\Theta_{Y|X}$ denotes the perturbation matrix, $\mathbf{1}'$ denotes a column vector of all ones, and $\mathbf{1}'Q_Y^*$ denotes the matrix whose every row equals Q_Y^*. Thus each $W_{Y|X}(\cdot|x)$ is obtained by perturbing $Q_Y^*(\cdot)$ along direction $\Theta_{Y|X}(\cdot|x)$. Such very noisy channels are also discussed in Chapter 6 in more detail.

Theorem 30 *Consider a very noisy channel of capacity C and two given exponents $\bar{E}_1, \bar{E}_2 < C$. In the very noisy limit, every achievable rate pair $(r_1, r_2) \in \bar{\mathcal{R}}(\bar{E}_1, \bar{E}_2)$ must satisfy*

$$\left(\frac{\sqrt{r_1}}{\sqrt{C} - \sqrt{\bar{E}_1}}\right)^2 + \left(\frac{\sqrt{r_2}}{\sqrt{C} - \sqrt{\bar{E}_2}}\right)^2 \leq 1.$$

Thus the converse bound for a very noisy channel only depends on its channel capacity. This bound generalizes the classical sphere-packing bound [2] for a very noisy channel:

$$\left(\frac{\sqrt{r}}{\sqrt{C} - \sqrt{\bar{E}}}\right)^2 \leq 1,$$

where r denotes rate and $\bar{E} < C$ denotes the classical error exponent. In the upper bound for UEP in Theorem 30, we had one such term for the special bits and another such term for ordinary bits.

Pretending this sphere-packing bound to be tight, Theorem 30 can be interpreted as splitting the unit resource of the channel in two parts: α (say) for the special bits and $(1-\alpha)$ for the ordinary bits. The special bits can then subdivide their piece into r_1 and $\bar{E}_1$ provided $\bar{E}_2 \leq C$ and

$$\left(\frac{\sqrt{r_1}}{\sqrt{C} - \sqrt{\bar{E}_1}}\right)^2 \leq \alpha.$$

Similarly, the ordinary bits can sub-divide their piece in into r_2 and $\bar{E}_2$ provided $\bar{E}_2 \leq C$ and

$$\left(\frac{\sqrt{r_2}}{\sqrt{C}-\sqrt{\bar{E}_2}}\right)^2 \leq 1-\alpha.$$

Binary Symmetric Channel: Achievability

We use a superposition coding strategy similar to [48] for encoding and use successive cancelation for decoding. One key difference in our strategy is we use random fixed composition codes as opposed to i.i.d. random codes in [48]. This difference is crucial for achieving the sphere-packing bound in the high-rates region, which cannot be achieved with i.i.d. random codes in [48].

Superposition Coding: First choose an auxiliary variable U satisfying the Markov chain $U-X-Y$. For the BSC, it suffices to for U to be binary uniform and $P_{X|U}$ to be another BSC of some crossover probability t. The corresponding X distribution is also uniform. The channel from U to Y is now a cascade of two binary symmetric channels of crossover probabilities p and t. Let $t \odot p \equiv (1-t)p + (1-p)t$ denote the effective crossover probability from U to Y. The capacity of this BSC equals $C_{\mathrm{b}}(t \odot p) = \ln 2 - h(t \odot p)$, where $h(\cdot)$ denotes the binary entropy function.

Message M_1 (i.e., the special bits) are conveyed by the U-codeword (cloud center) and M_2 (i.e., ordinary bits) are conveyed by the X-codewords (satellites) around this U-codeword. Hence the number of total U-codewords equals $|\mathcal{M}_1|$ and the number of X-codewords around each U-codeword equals $|\mathcal{M}_2|$. At the receiver, M_1 is decoded first by ML decoding over the U-codewords—choose the U-codeword nearest to the received sequence. Then M_2 is decoded by successive cancelation, that is, by choosing the ML candidate from the satellite X-codewords around the decoded U-codeword.

By fixed composition codes, we mean that all $|\mathcal{M}_1|$ U-codewords have the same type P_U. Moreover, in a particular cloud around a U-codeword, all $|\mathcal{M}_2|$ X-codewords have the same conditional type $P_{X|U}$. That is all U-codewords and all X-codewords have type $(1/2, 1/2)$. Moreover, all X-codewords around a particular U-codeword differ from that U codeword in t fraction of the code-length.

The purpose of choosing fixed composition codes was to avoid large deviations in the artificial 'channel' from U to X. This only leaves us with large deviations in the actual channel from X to Y, which are beyond our control. Choosing i.i.d. codes in superposition codes causes large deviations in both these channels, U to X and X to Y. The unnecessary large deviations of the artificial channel from U to X makes it impossible at low rates to achieve the sphere-packing bound with i.i.d. codes.

Theorem 31 *For rate-pair (r_1, r_2), the following $(\bar{E}_1, \bar{E}_2)$ is achievable, i.e., contained in $\bar{\mathcal{E}}(r_1, r_2)$ for any given $t \in [0, 1/2]$.*

$$\begin{aligned}
\bar{E}_1 &= \min_{\theta \in [p,1/2]} D_b(\theta \| p) + [C_b(t \odot \theta) - r_1]^+ \\
&\equiv \min_{\theta \in [p,1/2]} D_b(\theta \| p) + [I_\theta(U;Y) - r_1]^+ \\
\textit{and} \quad \bar{E}_2 &= \min\{\bar{E}_1, \tilde{E}_2\} \qquad \textit{where,} \\
\tilde{E}_2 &= \min_{\theta \in [p,1/2]} D_b(\theta \| p) + [h(t \odot \theta) - h(\theta) - r_2]^+ \\
&\equiv \min_{\theta \in [p,1/2]} D_b(\theta \| p) + [I_\theta(X;Y|U) - r_2]^+
\end{aligned}$$

In this scalar optimization, parameter θ denotes the empirical crossover probability from X to Y. Note that $I_\theta(U;Y)$ and $I_\theta(X;Y|U)$ denote mutual information between U and Y when the empirical crossover probability from X to Y equals θ. The conditional mutual information $I_\theta(X;Y|U)$ is defined similarly.

It is easy to verify that for high enough (r_1, r_2), these achievable exponents match the sphere-packing bound for BSC seen earlier. It follows since at high rates, the $[\cdot]^+$ terms are equal to 0. Assuming $\bar{E}_1 \geq \tilde{E}_2$, it implies

$$\begin{aligned}
\bar{E}_1 = D_{\mathrm{b}}(\theta_1 \| p) \quad &\text{for } \theta_1 \text{ s.t.} \quad I_{\theta_1}(U;Y) = r_1 \\
\text{and} \qquad \bar{E}_2 = \tilde{E}_2 = D_{\mathrm{b}}(\theta_2 \| p) \quad &\text{for } \theta_1 \text{ s.t.} \quad I_{\theta_2}(X;Y|U) = r_2
\end{aligned}$$

This precisely matches the sphere packing bound in Theorem 29, because the above pair of (r_1, r_2) lies on the capacity region of broadcast over BSC channels of crossover probabilities θ_1 and θ_2.

Remark: For some (r_1, r_2), the achievable exponents in Theorem 31 could be im-

proved by replacing successive cancelation decoding by its modification in [43] based on joint decoding of (M_1, M_2).

Very Noisy Channel: Achievabiltiy

Consider the very noisy channel $W_{Y|X} = \mathbf{1}'Q_Y^* + \epsilon\Theta_{Y|X}$ seen earlier, where Q_Y^* is the capacity achieving output distribution. Assume that we are operating at rate pair $(r_1, r_2) = \epsilon^2(\hat{r}_1, \hat{r}_2)$. Let the capacity of this channel be $C = \epsilon^2\hat{C}$ for ϵ tending to 0 and let P_X^* denote the capacity achieving input distribution. We will again use random fixed composition coding for this result.

Superposition Coding: Similar fixed composition coding is used as seen earlier for BSC. First, we need to choose auxiliary variable U satisfying $U - X - Y$. For the auxiliary variable U, we choose the same alphabet $\mathcal{X}$ as the input X. Moreover, we choose the marginal distribution for both U and X is chosen to be the capacity achieving input distribution, $P_U = P_X^*$. The channel $P_{X|U}$ is chosen such that the effective channel from U to Y, which is the cascade of $P_{X|U}$ with $W_{Y|X}$, equals

$$W_{Y|X}^{\beta} \equiv \mathbf{1}'Q_Y^* + \epsilon\beta\Theta_{Y|X} \quad \text{for some} \quad 0 < \beta < 1. \tag{5.4}$$

This channel from U to Y is obtained by 'shrinking' (see Fig. 5-1) the original channel $W_{Y|X}$ by a factor of β. This shrinking is performed by treating Q_Y^* as the origin.

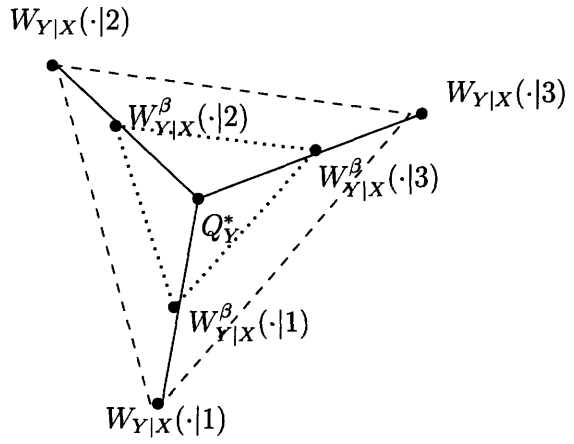


Figure 5-1: Channel $W_{Y|X}^{\beta}$ is obtained by shrinking original $W_{Y|X}$ by factor β with Q_Y^* as the origin.

In the very noisy limit, capacity of $W^{\beta}_{Y|X}$ is simply $\beta^2 C = \beta^2\epsilon^2\hat{C}$ (see Chapter 6). The capacity achieving input distribution for $W^{\beta}_{Y|X}$ is the same as P^*_X for the original $W_{Y|X}$.

Now we are ready to state the achievable error exponents.

Theorem 32 *For rate-pair* $(r_1, r_2) = \epsilon^2(\hat{r}_1, \hat{r}_2)$, *the following* $(\bar{E}_1, \bar{E}_2)$ *is achievable, i.e., contained in* $\bar{\mathcal{E}}(r_1, r_2)$ *for any given* $\beta \in [0, 1]$.

$$\begin{aligned} \lim_{\epsilon\to 0}\frac{\bar{E}_1}{\epsilon^2} &= \min_{\gamma\in[0,1]}(1-\gamma)^2\hat{C} + [\gamma^2\beta^2\hat{C} - \hat{r}_1]^+ \\ \textit{and} \quad \bar{E}_2 &= \min\{\bar{E}_1,\ \tilde{E}_2\} \qquad \textit{where,} \\ \lim_{\epsilon\to 0}\frac{\tilde{E}_2}{\epsilon^2} &= \min_{\gamma\in[0,1]}(1-\gamma)^2\hat{C} + [\gamma^2(1-\beta^2)\hat{C} - \hat{r}_2]^+ \end{aligned}$$

In this simple quadratic optimization, parameter γ denotes the empirically observed shrinking factor, that is, $W^{\gamma}_{Y|X}$ denotes the empirically observed channel type from X to Y. Smaller the value of γ, noisier is the observed channel type.

Also note that $\gamma^2\beta^2\hat{C}$ and $\gamma^2(1-\beta^2)\hat{C}$ in above theorem respectively correspond to $I(U;Y)$ and $I(X;Y|U)$ when the empirical shrinking factor equals γ. It is because the empirical channel from U to Y is a cascade of $W^{\beta}_{Y|X}$ with $W^{\gamma}_{Y|X}$, the effective shrinking factor from U to Y is $\beta\gamma$. Hence the mutual information $I(U;Y)$ for this empirical channel equals $\epsilon^2(\gamma^2\beta^2\hat{C})$. Similarly, for this empirical channel, $I(X;Y|U) = I(X;Y) - I(U;Y)$ equals $\epsilon^2(\gamma^2\hat{C} - \gamma^2\beta^2\hat{C})$.

It is easy to verify that for high enough $(\hat{r}_1, \hat{r}_2)$, these achievable exponents match the sphere-packing bound seen earlier. This is because in the optimization for $\bar{E}_1$ and $\bar{E}_2$, the arguments of $[\cdot]^+$ terms in Theorem 32 are equal to 0 for high enough rates. Hence for high enough $(\hat{r}_1, \hat{r}_2)$,

$$\begin{aligned} \lim_{\epsilon\to 0}\frac{\bar{E}_1}{\epsilon^2} = \left(1-\sqrt{\frac{\hat{r}_1}{\beta^2\hat{C}}}\right)^2\hat{C} \quad \text{and} \quad \lim_{\epsilon\to 0}\frac{\bar{E}_2}{\epsilon^2} = \left(1-\sqrt{\frac{\hat{r}_2}{(1-\beta^2)\hat{C}}}\right)^2\hat{C} \\ \Rightarrow \qquad \bar{E}_1 \approx \left(1-\sqrt{\frac{r_1}{\beta^2 C}}\right)^2 C \quad \text{and} \quad \bar{E}_2 \approx \left(1-\sqrt{\frac{r_2}{(1-\beta^2)C}}\right)^2 C \end{aligned}$$

Recall that $a \approx b$ denotes $\lim_{\epsilon \to 0} \frac{a}{b} = 1$. Eliminating β^2 from these equations, in the limit of vanishing ϵ we get:

$$\left(\frac{\sqrt{r_1}}{\sqrt{C} - \sqrt{\bar{E}_1}}\right)^2 + \left(\frac{\sqrt{r_2}}{\sqrt{C} - \sqrt{\bar{E}_2}}\right)^2 = 1.$$

This is exactly the sphere-packing bound seen earlier.

5.1.2 Single special message

Now consider the missed-detection exponent for one special message (say $M = 1$) while reliably communicating the remaining $\doteq e^{nR}$ ordinary messages. Recall that reliable communication means overall error probability tends to 0 with increasing codelength. Now let us define the best missed-detection exponent $\bar{E}_{\text{md}}(R)$ of the special message for date rate $R \leq C$.

Definition 33 *For a reliable code sequence $\mathcal{Q}$ of rate $R_{\mathcal{Q}}$, the missed-detection exponent is defined as*

$$\bar{E}_{md,\mathcal{Q}} \equiv \liminf_{n \to \infty} \frac{-\log \Pr\left[\hat{M} \neq 1 \middle| M = 1\right]}{n}. \tag{5.5}$$

Then define $\bar{E}_{md}(R) = \sup_{\mathcal{Q}: R_{\mathcal{Q}} \geq R} \bar{E}_{md,\mathcal{Q}}$.

Now we recall the function $\mathcal{J}(R)$ defined in Chapter 3 for reading convenience.

$$\mathcal{J}(R) = \max_{\substack{\alpha_X, P_X^1, P_X^2, \ldots, P_X^{|\mathcal{X}|}: \\ \sum_{j \in \mathcal{X}} \alpha_X(j) I(P_X^j, W_{Y|X}) \geq R}} \sum_{j \in \mathcal{X}} \alpha_X(j) D\left((P_X^j W_{Y|X})_Y(\cdot) \middle\| W_{Y|X}(\cdot|j)\right) \tag{5.6}$$

where α_X and $\{P_X^j\}$ are distributions over $\mathcal{X}$.

Theorem 34 [4] $\bar{E}_{md}(R) = \mathcal{J}(R)$

For $R = C$, this implies $\bar{E}_{\text{md}}(C) = \mathsf{E}_{\mathsf{Red}}$. Thus this theorem generalizes the notion of Red-Alert Exponent for data rates strictly below capacity. It will interesting to extend this result when all the ordinary messages also demand a positive error

[4] This theorem and Theorem 36 in the next section is joint work with Barış Nakiboğlu.

exponent E_{ordinary}. The above theorem only addresses the special case of this problem where $E_{\text{ordinary}} = 0$. This is because all we require is reliable communication for the ordinary messages. The machinery of Renyi-Divergence [36, 37] instead of KL divergence could be useful for this purpose.

Optimal strategy: Consider the α_X, $P_X^1, P_X^2, \ldots, P_X^{|\mathcal{X}|}$ which achieve the maximization in the definition of $\mathcal{J}(R)$. Now the codelength n is divided into $|\mathcal{X}|$ contiguous blocks. The first block consists of the first $\lceil n\alpha_X(1) \rceil$ symbols, the second block consists of the next $\lceil n\alpha_X(2) \rceil$ symbols and so on. The length of the last block is chosen so that total codelength equals n.

The special codeword is simply obtained by repeating input letter $j \in \mathcal{X}$ in block j. For ordinary codewords, we use blockwise i.i.d. random coding as follows. For every ordinary codeword, choose the symbols in the j'th block by i.i.d. P_X^j distribution.

At the decoder, a two stage decision rule is employed. If output type in any block j is not $(P_X^j W_{Y|X})_Y$, the special message is chosen: $\hat{M} = 1$. Otherwise, the ML candidate amongst the ordinary codewords is chosen as $\hat{M}$.

5.2 Variable-Length Block Codes with Feedback

This section discusses an achievable result for a bit-wise UEP scenario without proving its optimality. For other scenarios of message-wise and bit-wise UEP with feedback, similar achievable results can be provided.

5.2.1 Many special bits

We now address the situation where out of the total $E[\tau] R/\ln 2$ (approx.) bits, approximately $E[\tau] r_1/\ln 2$ bits are special. Again we consider reliable code sequences with feedback with message sets of the form $\mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)}$, where the cardinality of $\mathcal{M}_1^{(\kappa)}$ and $\mathcal{M}_2^{(\kappa)}$ grows exponentially at rate r_1 and $r_2 \equiv R - r_1$, respectively. This simply says that the rate of special bits and ordinary bits equals r_1 and r_2 respectively. We will focus on the case when the ordinary bits require the best possible exponent at sum rate R, i.e., the Burnashev exponent at rate R. Under this

requirement, the best error exponent for special bits, $\bar{E}^{\mathsf{f}}_{\text{bits},1}$ is defined as follows:

Definition 35 *Consider a reliable sequence $\mathcal{Q}$ with feedback, with message sets $\mathcal{M}^{(\kappa)}$ of the form $\mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)}$, where*

$$r_{1,\mathcal{Q}} \equiv \lim_{\kappa\to\infty} \frac{\log|\mathcal{M}_1^{(\kappa)}|}{E\left[\tau^{(\kappa)}\right]}, \qquad r_{2,\mathcal{Q}} \equiv \lim_{\kappa\to\infty} \frac{\log|\mathcal{M}_2^{(\kappa)}|}{E\left[\tau^{(\kappa)}\right]}.$$

Then the exponents for special bits and ordinary bits are defined as,

$$\bar{E}^{\mathsf{f}}_{bits,1,\mathcal{Q}} = \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}_1 \neq M_1\right]}{E\left[\tau^{(\kappa)}\right]}, \qquad \bar{E}^{\mathsf{f}}_{bits,2,\mathcal{Q}} = \liminf_{\kappa\to\infty} \frac{-\log \Pr\left[\hat{M}_2 \neq M_2\right]}{E\left[\tau^{(\kappa)}\right]}$$

Now define $\bar{E}^{\mathsf{f}}_{bits,1} = \sup_{\mathcal{Q}\in\mathcal{F}} \bar{E}^{\mathsf{f}}_{bits,1,\mathcal{Q}}$, where $\mathcal{F}$ is set of reliable code sequences $\mathcal{Q}$ for which $r_{1,\mathcal{Q}} \geq r_1$, $r_{2,\mathcal{Q}} \geq r_2$ and $\bar{E}^{\mathsf{f}}_{bits,2,\mathcal{Q}} = (1 - R/C)\mathrm{D}_{\max}$.

Theorem 36

$$E^{\mathsf{f}}_{bits1} = \left(1 - \tfrac{R}{C}\right)\mathrm{D}_{\max} + \tfrac{r_2}{C}\mathsf{E}_{\mathsf{Red}}$$

Thus while ensuring that ordinary bits achieve the Burnashev exponent, the exponent for special bits can exceed the Burnashev exponent by $\frac{r_2}{C}\mathsf{E}_{\mathsf{Red}}$. Contrast this with the no-feedback case, where we saw that special bits could not achieve any better exponent that the ordinary bits if ordinary bits were to achieve the sphere-packing exponent at sum rate R.

Optimal strategy: We concatenate the buzzer strategy for many special bits from Chapter 4 with the Yamamoto-Itoh strategy for achieving Burnashev's exponent [29]. As in earlier feedback schemes, we use a fixed length erasure code where erasures are used to initiate retransmissions.

This strategy works in three phases: first stage is of length $\frac{r_1}{C}\kappa$, second phase is of length $\frac{r_2}{C}\kappa$ and third phase is of length $(1 - \frac{r_1+r_2}{C})\kappa$. In the first phase, transmit M_1 using a capacity achieving code of length $\frac{r_1}{C}\kappa$. If the temporary decision $\tilde{M}_1$ is correct after this transmission, the transmitter sends M_2 in the second phase using a

capacity achieving code of length $\frac{r_2}{C}\kappa$. Otherwise, the transmitter sends a buzzer in these $\frac{r_2}{C}\kappa$ symbols by repeating the symbol x_r.

In the third phase of length $(1-\frac{r_1+r_2}{C})\kappa$ symbols, Yamamato-Itoh scheme is used for accept or reject previous temporary decisions. If both the temporary decisions $\tilde{M}_1$ and $\tilde{M}_2$ in first two phases were correct, repeat accept letter x_a. If either $\tilde{M}_1$ or $\tilde{M}_1$ was wrong, repeat reject letter x_d.

An erasure is declared and retransmission is initiated if the decoder detects a buzzer in the second phase $(1-\frac{r}{C})\kappa$ symbols or a reject codeword in the third phase. Otherwise the receiver finalizes its temporary decision $(\tilde{M}_1, \tilde{M}_2)$ based on ML decoding over first two phases.

The Yamamato-Itoh scheme in the third phase ensures that Burnashev exponent is achieved for M_2. For M_1, an error happens only if the receiver misses the buzzer in the second phase *and* the reject message in the third phase. The exponent for the first event equals $\frac{r_2}{C}\mathsf{E}_{\mathsf{Red}}$ and that for the second event equals $(1-\frac{r_1+r_2}{C})D_{\max}$. Hence the error exponent for M_1 equals $(1-\frac{r_1+r_2}{C})D_{\max}+\frac{r_2}{C}\mathsf{E}_{\mathsf{Red}}$. We omit other simple proof details of this achievability result because they are very similar to the achievability proofs in Chapter 4.

5.3 Rates Below Capacity: Proofs

5.3.1 Proof of Theorem 29

We use a genie to prove this converse. Assume that for all channel types besides two, a genie provides the exact M_1 and M_2 to the receiver. The two channel types for which the genie does not help are binary symmetric channels with crossover probability p_1 and p_2 satisfying:

$$D_{\mathrm{b}}(p_1\|p) \equiv \bar{E}_1, D_{\mathrm{b}}(p_2\|p) \equiv \bar{E}_2 \qquad (\text{s.t. } p \leq p_1, p_2) \tag{5.7}$$

Thanks to the genie, we can assume that the channel type can only be a BSC of crossover probability p_1 or p_2.

Hence we can consider the corresponding broadcast channel where channel to user $j \in \{1,2\}$ is a BSC of crossover probability p_j. We need to ensure that M_2 is reliably decoded at user 2 and M_1 is reliably decoded by the (degraded) user 1. Since $p_1 \geq p_2$, user 2 can also decode message M_1 for the degraded user 1.

If a rate pair (r_1, r_2) is outside the capacity region of this broadcast channel, then in spite of the genie above, that pair is infeasible for the UEP problem.

5.3.2 Proof of Theorem 30

The proof for the very noisy channel is very similar to the BSC. Again we assume a genie which provides exact M_1 and M_2 to the receiver for all channel types besides two. The two channel types for which the genie does not help are shrunk versions (see Fig. 5-1 and Eq. (5.4)) of the original channel, $W_{Y|X}^{\gamma_1}$ and $W_{Y|X}^{\gamma_2}$. Parameter $\gamma_1 \leq \gamma_2$ are fixed so that

$$\begin{aligned} (1-\gamma_1)^2 C = \bar{E}_1 \quad &\text{and} \quad (1-\gamma_2)^2 C = \bar{E}_2 \\ \gamma_1 = 1 - \sqrt{\bar{E}_1}/C \quad &\text{and} \quad \gamma_2 = 1 - \sqrt{\bar{E}_2}/C \end{aligned} \tag{5.9}$$

Recalling $W_{Y|X}^{\gamma_i} = \mathbf{1}' Q_Y^* + \epsilon\gamma_i \Theta_{Y|X}$ implies $W_{Y|X}^{\gamma_1}$ is a degraded version of $W_{Y|X}^{\gamma_2}$. That is, we can find a channel Ω such that channel $W_{Y|X}^{\gamma_1}$ could be thought as channel $W_{Y|X}^{\gamma_2}$ cascaded by channel Ω:

$$W_{Y|X}^{\gamma_1} = W_{Y|X}^{\gamma_2} \Omega \qquad \text{for} \qquad \Omega = \tfrac{\gamma_1}{\gamma_2} I + (1 - \tfrac{\gamma_1}{\gamma_2}) \cdot \mathbf{1}' Q_Y^*$$

where I denotes the identity matrix. This Ω is mixture of a noiseless channel with probability $\frac{\gamma_1}{\gamma_2}$ and a trivial channel $\mathbf{1}' Q_Y^*$ with probability $(1 - \frac{\gamma_1}{\gamma_2})$.

Capacity achieving distribution for both $W_{Y|X}^{\gamma_1}$ and $W_{Y|X}^{\gamma_2}$ is the same in the very noisy limit. For such degraded broadcast channels, [44] showed that time sharing between its two users achieves the entire capacity region. Since capacity of $W_{Y|X}^{\gamma_i}$ equals $\gamma_i^2 C$ in the very noisy limit, this capacity region is given by all (r_1, r_2) pairs satisfying

$$\frac{r_1}{\gamma_1^2 C} + \frac{r_2}{\gamma_2^2 C} \leq 1$$

Substituting γ_1 and γ_2 from (5.8) completes the proof of this sphere packing bound.

5.3.3 Proof of Theorem 31

Consider the set of all binary sequences of empirical distribution[5] P_U. For the BSC, we choose P_U to be the uniform binary distribution. Each of the e^{nr_1} U-codewords is chosen uniformly from this set of fixed composition sequences. Let the i'th U-codeword be denoted by $\bar{u}^n(i)$.

For each $i \in \mathcal{M}_1 = \{1, 2, \ldots, e^{nr_1}\}$, consider the set of X-sequences whose empirical channel from $\bar{u}^n(i)$ is $P_{X|U}$. For the BSC, remember that our choice of $P_{X|U}$ was a binary symmetric channel of crossover probability t. Each of the e^{nr_2} X-codewords (satellites) around $\bar{u}^n(i)$ (cloud center) is chosen uniformly from this set of X-sequences. Notice that empirical distribution of every X-codeword will also be uniform binary.

Let the X-codewords around $\bar{u}^n(i)$ be denoted by $\bar{x}^n(i,j)$, where $j \in \mathcal{M}_2 = \{1, 2, \ldots, e^{nr_2}\}$ denotes the ordinary message. For sending $M = (M_1, M_2)$, the encoder transmits $\bar{x}^n(M_1, M_2)$. The decoder first decodes $\hat{M}_1$ as the nearest U-codeword assuming a memoryless channel $P_{Y|U}$ from U to Y. This $P_{Y|U}$ is the cascade of $P_{X|U}$ followed by $W_{Y|X}$. For our BSC case, this rule for choosing $\hat{M}_1$ is equivalent to nearest neighbor decoding over $\bar{u}^n(i)$ codewords.

$$\hat{M}_1 = \arg\min_{i \in \mathcal{M}_1} |Y^n \oplus \bar{u}^n(i)|_H$$

[5] This means that each input letter $u \in \mathcal{U}$ except the last appears $\lceil nP_U(u) \rceil$ times. The last letter appears for the remaining number of symbols out of n.

where $\oplus$ denotes element-wise XOR of two binary sequences and $|\cdot|_H$ denotes the Hamming weight of a sequence.

After decoding $\hat{M}_1$, it performs successive cancelation for decoding M_2. That is, it chooses $\hat{M}_2$ as the ML candidate from $\{\bar{x}^n(\hat{M}_1, j) : j \in \mathcal{M}_2\}$, i.e., the cloud of codewords around $\bar{u}^n(\hat{M}_1)$. For our BSC case, this rule for choosing $\hat{M}_2$ is again equivalent to nearest neighbor decoding over $\bar{x}^n(\hat{M}_1, j)$ codewords.

$$\hat{M}_2 = \arg\min_{j \in \mathcal{M}_2} |Y^n \oplus \bar{x}^n(\hat{M}_1, j)|_H.$$

We first prove the achievability of $\bar{E}_1$ for special bits. Without loss of generality, we can assume $M_1 = 1$. For simplicity of analysis, we can pretend that $\bar{u}^n(1) = (0, 0, \cdots, 0)$. Now in our fixed composition construction, every satellite $\bar{x}^n(1, j)$ around $\bar{u}^n(1)$ will be chosen uniformly from the sequences of Hamming weight nt.

If the empirical channel from X to Y acts like a BSC of crossover probability θ, every $\bar{x}^n(1, j)$ leads to an output sequence of Hamming weight $n(t \odot \theta)$. By Stein's lemma, the exponent of observing such a BSC of crossover probability θ equals $D_{\mathrm{b}}(\theta \| p)$. Now an error happens in decoding M_1 if some other $\bar{u}^n(i)$ for $i \neq 1$ is closer to Y^n than $n(t \odot \theta)$. Since the $\bar{u}^n(i)$ sequences are chosen uniformly from binary sequences of weight $n/2$, for any given output sequence Y^n, $Y^n \oplus \bar{u}^n(i)$ is also distributed uniformly over binary sequences of weight $n/2$. Hence the exponent of $Y^n \oplus \bar{u}^n(i)$ having Hamming weight less than $n(t \odot \theta)$ equals

$$D_{\mathrm{b}}(t \odot \theta \| \tfrac{1}{2}) = C_{\mathrm{b}}(t \odot \theta)$$

By union bound, the exponent of this happening for at least one $i \neq 1$ is given by $[C_{\mathrm{b}}(t \odot \theta) - r_1]^+$. Thus the overall error exponent for M_1 by observing crossover probability θ equals:

$$D_{\mathrm{b}}(\theta \| p) + [C_{\mathrm{b}}(t \odot \theta) - r_1]^+$$

Since the number of possible empirical crossover probabilities at length n is just $n+1$,

the overall error exponent for M_1 is obtained by minimizing the above expression over all $\theta \in [p, 1/2]$. This completes the achievability of $\bar{E}_1$.

Now for achievability of $\bar{E}_2$, we first show that assuming $\hat{M}_1$ was decoded correctly, the exponent of for decoding $\hat{M}_2$ equals $\tilde{E}_2$. Then (pessimistically) assuming that an incorrect $\hat{M}_1$ leads to incorrect $\hat{M}_2$ in our successive cancelation decoder, M_2 can achieve $E_2 = \min\{E_1, \tilde{E}_2\}$.

Now we only need to prove that $\tilde{E}_2$ is error exponent for decoding M_2 when $\hat{M}_1$ is correct.

Again without loss of generality, assume that $(M_1, M_2) = (1, 1)$ was transmitted. If the empirical crossover probability of the channel is θ, the Hamming weight of Y^n is $n(t \odot \theta)$ and the Hamming distance of Y^n from $\bar{x}^n(1,1)$ is $n\theta$. Now M_2 is decoded incorrectly if some incorrect $\bar{x}^n(1,j)$ lies within distance $n\theta$ of Y^n.

Recall that each $\bar{x}^n(1,j)$ is chosen uniformly from binary sequences of weight nt. The number of sequences x^n of Hamming weight nt which are also within distance $n\theta$ from given Y^n is $\doteq \exp(nh(t) + nh(\theta) - nh(t \odot \theta))$. To see this, consider the set $\mathcal{K}$ defined as all (x^n, y^n) sequences where the hamming weight of x^n is nt and the empirical channel from X to Y has crossover probability θ. Size of this $\mathcal{K}$ is $\doteq e^{nh(t)+nh(\theta)}$. Since the number y^n sequences of Hamming weight $n(t \odot \theta)$ is $\doteq e^{nh(t \odot \theta)}$, restricting the y^n part to observed Y^n gives $\doteq e^{nh(t)+nh(\theta)-nh(t \odot \theta)}$ elements in $\mathcal{K}$. The x^n part of each such element is the possible set of x^n of Hamming weight nt and is within distance $n\theta$ from observed Y^n.

Now since any wrong $\bar{x}^n(1,j)$ for $j \neq 1$ is chosen uniformly from the set of sequences of Hamming weight nt, the probability any weight nt to be chosen as $\bar{x}^n(1,j)$ is $\approx e^{-nh(t)}$. Hence the probability that a wrong $\bar{x}^n(1,j)$ lies within distance $n\theta$ of Y^n is

$$\doteq \frac{e^{nh(t)+nh(\theta)-nh(t \odot \theta)}}{e^{nh(t)}} = e^{nh(\theta)-nh(t \odot \theta)}$$

Thus the exponent of a particular $\bar{x}^n(1,j)$ being within $n\theta$ of Y^n equals $h(t \odot \theta) - h(\theta)$. By union bound, the exponent of this event for some $j \neq 1$ is given by $[h(t \odot \theta) - h(\theta) - r_2]^+ = [I_\theta(X;Y|U) - r_2]^+$. Again recalling that exponent of the empirical

crossover probability θ is $D_{\mathrm{b}}(\theta\|p)$ and minimizing over θ implies that

$$\tilde{E}_2 = \min_{\theta\in[p,1/2]} D_{\mathrm{b}}(\theta\|p) + [I_\theta(X;Y|U) - r_2]^+$$

equals the error exponent for M_2 when M_1 is decoded correctly.

5.3.4 Proof of Theorem 32

For a general DMC $W_{Y|X}$, we can repeat the same fixed composition argument as for the BSC and achieve the following exponents for any given choice of P_U and $P_{X|U}$ used for superposition coding.

$$\bar{E}_1 = \min_{V_{Y|X}} D(V_{Y|X}\|W_{Y|X}|P_X) + [I_{V_{Y|X}}(U;Y) - r_1]^+ \tag{5.10}$$

$$\text{and} \quad \bar{E}_2 = \min\{\bar{E}_1,\ \tilde{E}_2\} \qquad \text{where,} \tag{5.11}$$

$$\tilde{E}_2 = D(V_{Y|X}\|W_{Y|X}|P_X) + [I_{V_{Y|X}}(X;Y|U) - r_2]^+ \tag{5.12}$$

Here $I_{V_{Y|X}}(U;Y)$ denotes the mutual information between U and Y when the channel from X to Y is $V_{Y|X}$. This implies the channel from U to Y is a cascade of $P_{X|U}$ with $V_{Y|X}$. Similarly, $I_{V_{Y|X}}(X;Y|U)$ denotes the conditional mutual information between X and Y for the channel $V_{Y|X}$.

In the remaining proof of the theorem, we will analyze $\bar{E}_1$ and $\tilde{E}_2$ in (5.10) (5.11) and show their equivalence to their corresponding expressions in Theorem 32.

Let us first focus on $\bar{E}_1$ in (5.10). Recall our choice of $P_{X|U}$, which ensures the channel from U to Y is a shrunk version of the channel from X to Y by a factor of β. As we will see in Chapter 6, in the very noisy limit, the KL divergence between two nearby distributions, say P and Q, behaves as

$$D(P\|Q) \approx \sum_i \frac{(Q(i) - P(i))^2}{2P(i)}. \tag{5.13}$$

See 6.1 in Chapter 6 for more details. Here P and Q are related as $P(\cdot) = Q(\cdot) + \epsilon T(\cdot)$ for some direction T along the probability simplex. Since mutual information can

be expressed as a convex combination of KL divergences between conditional and marginal distributions, this quadratic approximation of KL divergence implies that mutual information of the shrunk channel from U to Y is β^2 times mutual information of the channel from X to Y.

Hence for our fixed composition codes having $P_X = P_X^*$, $\bar{E}_1$ in (5.10) can be written as

$$\begin{aligned} \bar{E}_1 &\approx \min_{V_{Y|X}} D(V_{Y|X}\|W_{Y|X}|P_X^*) + [\beta^2 I_{V_{Y|X}}(X;Y) - r_1]^+ \\ &= \min_{V_{Y|X}} D(V_{Y|X}\|W_{Y|X}|P_X^*) + [\beta^2 D(V_{Y|X}\|(P_X^* V_{Y|X})_Y|P_X^*) - r_1]^+ \end{aligned} \quad (5.14)$$

where the second step follows by writing $I_{V_{Y|X}}(X;Y)$ in terms of KL divergences.

From the discussions in Chapter 2, optimum $V_{Y|X}$ for such optimizations lies on the exponential family connecting the true channel $W_{Y|X}$ to the trivial channel $\mathbf{1}'Q_Y^*$. Hence the optimum $V_{Y|X}(\cdot|x)$ are of the form:

$$\forall x, \qquad V_{Y|X}(y|x) = \frac{(W_{Y|X}(y|x))^\gamma (Q_Y^*(y|x))^{1-\gamma}}{k_x(\gamma)}$$

where γ is the exponential parameter and $k_x(\gamma)$ is the normalization factor to ensure a valid probability distribution.

Now we will use the very noisy assumption to show that this exponential family of channels is the same as the shrunk versions of $W_{Y|X}$. Since $W_{Y|X}(y|x) = Q_Y^*(y) + \epsilon\Theta_{Y|X}(y|x)$, Taylor's approximation implies

$$\begin{aligned} (W_{Y|X}(y|x))^\gamma &= (Q_Y^*(y) + \epsilon\Theta_{Y|X}(y|x))^\gamma \\ &= (Q_Y^*(y))^\gamma \left(1 + \epsilon\frac{\Theta_{Y|X}(y|x)}{Q_Y^*(y)}\right)^\gamma \\ &\approx (Q_Y^*(y))^\gamma \left(1 + \epsilon\gamma\frac{\Theta_{Y|X}(y|x)}{Q_Y^*(y)}\right) \end{aligned}$$

Substitute this in the exponential family of $V_{Y|X}$ to get

$$\forall x, \qquad V_{Y|X}(y|x) \approx \frac{Q_Y^*(y) + \epsilon\gamma\Theta_{Y|X}(y|x)}{k_x(\gamma)}$$

The normalization factor $k_x(\gamma)$ is simply one because the numerator sums up to 1 anyway over y. Thus $V_{Y|X} = W_{Y|X}^{\gamma}$ and the exponential family is the same as the family of shrunk versions of $W_{Y|X}$. As discussed earlier, for shrunk $V_{Y|X} = W_{Y|X}^{\gamma}$ mutual information $I_{V_{Y|X}}(X;Y) \approx \gamma^2 I_{W_{Y|X}}(X;Y)$. Since we are using the capacity achieving input distribution, $I_{W_{Y|X}}(X;Y) = C = \epsilon^2\hat{C}$ implying

$$I_{V_{Y|X}}(X;Y) \approx \epsilon^2\gamma^2\hat{C}.$$

Let us now simplify $D(V_{Y|X}\|W_{Y|X}|P_X^*) = D(W_{Y|X}^{\gamma}\|W_{Y|X}|P_X^*)$ in (5.10) using the quadratic approximation of KL divergence in (6.1). Using the definition of $W_{Y|X}^{\gamma}$,

$$\begin{aligned} W_{Y|X} - W_{Y|X}^{\gamma} &= (1-\gamma)\epsilon\Theta_{Y|X} = (1-\gamma)(W_{Y|X} - W_{Y|X}^{0}) \\ \Rightarrow \quad D(W_{Y|X}^{\gamma}\|W_{Y|X}|P_X^*) &\approx (1-\gamma)^2 D(W_{Y|X}^{0}\|W_{Y|X}|P_X^*) \\ &= (1-\gamma)^2 D(Q_Y^*\|W_{Y|X}|P_X^*) \end{aligned}$$

where $W_{Y|X}^{0}$ is the completely shrunk channel $\mathbf{1}'Q_Y^*$. In the very noisy limit, KL divergence becomes symmetric in its arguments (see Chapter 6) and hence the RHS above equals $(1-\gamma)^2 D(W_{Y|X}\|Q_Y^*|P_X^*)$. By definition of mutual information, this equals $(1-\gamma)^2 I_{W_{Y|X}}(X;Y) = (1-\gamma)^2 C$. This shows that the divergence term in (5.10) equals $(1-\gamma)^2\epsilon^2\hat{C}$ in the very noisy limit. Substitute this along with $I_{V_{Y|X}}(X;Y) \approx \epsilon^2\gamma^2\hat{C}$ in (5.14) to complete the proof for $\bar{E}_1$ in Theorem 32.

Let us now analyze $\tilde{E}_2$ in (5.12). First note that $I_{V_{Y|X}}(X;Y|U) = I_{V_{Y|X}}(X;Y) - I_{V_{Y|X}}(U;Y)$ by chain rule. Since mutual information between U and Y is β^2 times the mutual information between X and Y, we get $I_{V_{Y|X}}(X;Y|U) = (1-\beta^2) I_{V_{Y|X}}(X;Y)$. Substituting this in the optimization in (5.12) for $\tilde{E}_2$,

$$\tilde{E}_2 = \min_{V_{Y|X}} D(V_{Y|X}\|W_{Y|X}|P_X^*) + [(1-\beta^2) I_{V_{Y|X}}(X;Y) - r_2]^+$$

This is simply the expression in (5.14) for $\bar{E}_1$ but with β^2 is replaced by $(1-\beta^2)$. Thus all the arguments for $\bar{E}_1$ can be repeated to prove $\tilde{E}_2$ in Theorem 32.

5.3.5 Proof of Theorem 34

The achievability proof follows on similar lines of the achievability proof in Chapter 3 for $E_{\text{md}} \geq \mathsf{E}_{\mathsf{Red}}$ at capacity.

Achievability: $\bar{E}_{\mathbf{md}}(R) \geq \mathcal{J}(R)$

For each block-length n, the special message is sent by repeating input j in the j'th block of length $n\alpha_X(j)$. The remaining $|\mathcal{M}^{(n)}|-1$ ordinary codewords are generated with i.i.d. P_X^j symbols in block j.

Let the empirical output distribution in block j be denoted by Q_y^j. The receiver decides that the special message was sent only when no Q_y^j is close to $(P_X^j W_{Y|X})_Y$. To be precise, the decoding region of the special message is given by:

$$\mathcal{G}(1) = \{y^n : \exists j \text{ s.t. } |\mathsf{Q}_y^j(i) - (P_X^j W_{Y|X})_Y(i)| \geq \sqrt[4]{1/n} \text{ for some } \forall i \in \mathcal{Y}\}$$

Error happens after sending the special message if type Q_y^j in every block j is close to $(P_X^j W_{Y|X})_Y$ in the sense above. By Stein's lemma, the exponent of this event for block j equals $\alpha_X(j) D((P_X^j W_{Y|X})_Y(\cdot) \| W_{Y|X}(\cdot|j))$, where the $\alpha_X(j)$ factor arises due to length of block j. Memoryless property of the channel implies the missed-detection exponent for the special message is obtained by the summation

$$\sum_j \alpha_X(j) D((P_X^j W_{Y|X})_Y(\cdot) \| W_{Y|X}(\cdot|j)),$$

which by definition is equal to $\mathcal{J}(R)$.

Now the only thing left to prove is that we can have vanishing error probability for the ordinary messages at rate R. For that purpose, we first calculate the average error probability of their (blockwise) i.i.d. code ensemble described above. For that i.i.d. ensemble, the conditional error probability will be same for all $i \neq 1$ in $\mathcal{M}^{(n)}$.

Hence without loss of generality, let us calculate the error probability of the message $M = 2$.

Assuming that the second message was transmitted, $\Pr[y^n \in \mathcal{G}(1) | M = 2]$ is vanishingly small. It is because, the output distribution for the random ensemble for ordinary codewords is i.i.d. $(P_X^j W_{Y|X})_Y$ in block j. Chebyshev's inequality guarantees a vanishing probability for the output type Q_y^j of block j being outside a $\sqrt[4]{1/n}$ ball around $(P_X^j W_{Y|X})_Y$, i.e., outside $[P_Y^*]$. More precisely, this probability is of the order $\sqrt{1/n}$. By union bound, the probability of this event for some block j is of the order $|\mathcal{X}|/\sqrt{n}$, which still vanishes with n.

If the second message was transmitted, $\Pr[y^n \in \cup_{i>2}\mathcal{G}(i) | M = 2]$ is also vanishingly small by the standard random coding argument for rate R. This error probability of erring to another ordinary codeword is small because mutual information for this (blockwise) i.i.d. code ensemble exceeds R due to definition of $\mathcal{J}(R)$:

$$\sum_{i \in \mathcal{X}} \alpha_X(i) I(P_X^i, W_{Y|X}) \geq R$$

Thus data rate R can be achieved with vanishing error probability using this blockwise i.i.d. ensemble for ordinary codewords.

Converse: $\bar{E}_{\mathbf{md}}(R) \leq \mathcal{J}(R)$

This converse is a corollary of Lemma 27 proved earlier in Chapter 4. This proof simply follows by substituting $i = 1$ for the message index in Lemma 27. Note that in a length n block code without feedback, the average decoding time $E[\tau]$ is trivially equal to n. The converse is completed by noting that $\mathcal{H}(M|Y^0) = nR^{(n)}$, where $R^{(n)}$ of the code sequence $\mathcal{Q}$ tends to $R_{\mathcal{Q}} \geq R$.

Chapter 6

Rates Below Capacity: Network Information Theory Approach

Now we will address the case of rates below capacity using a different approach than all previous chapters. In previous chapters, we considered a point-to-point channel and UEP was used for extra protection for the crucial parts against large deviations of the channel noise. In other words, UEP was aimed at providing better error exponents for those crucial parts. Alternatively, UEP can be used in a broadcast network where the crucial parts should be protected against channel fading or movements of users. That is, the crucial parts should be received by all users, including those far from the base station or those experiencing bad fading. However, the better off users (which are near the base station or experiencing little fading) should be able to decode the ordinary parts as well.

This chapter primarily focuses on the bit-wise notion of UEP for networks. This problem is equivalent to the network information theory problem of broadcast with degraded message sets [42]. Let us describe this problem where transmitter X is broadcasting to K receivers denoted by $Y_0, Y_1, \cdots, Y_{K-1}$. The network is memoryless and completely characterized by the network transition matrix $W_{Y_0 Y_1 \cdots Y_{K-1}|X}$. The set of all K receivers is denoted by S_1. All users in S_1 want to decode M_1, which denotes the special (crucial) bits needed by everyone. There is a subset of users $S_0 \subset S_1$ and all users in S_0 want to also decode M_0, which denotes the ordinary bits. The overall

message at the transmitter is (M_0, M_1), the Cartesian product of ordinary bits and special bits.

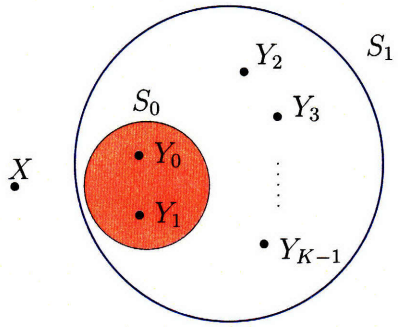


Figure 6-1: Broadcast with 2 degraded message sets: users in S_1 want message M_1 and those in S_0 want M_0 too.

Let $R_i,\ i \in \{0,1\}$ denote the rate of M_i. That is, M_i is chosen uniformly from the set $\{1, 2, \cdots, \lceil e^{nR_i} \rceil\}$ where n denotes the code-length. A rate pair (R_0, R_1) is said to be achievable if a sequence of codes exists for which users in S_1 can decode M_1 with vanishing error probability and users in S_0 can also decode M_0 with vanishing error probability. The capacity region of this broadcast network is defined as the closure of all achievable rate-pairs (R_0, R_1). This region captures the rate tradeoff between ordinary bits and special bits.

In general, there could be multiple speciality levels (say $L \geq 2$). This corresponds to L nested subsets of users $S_0 \subset S_1 \cdots S_{L-2} \subset S_{L-1}$, where S_{L-1} is the set of all users. Users in set $S_k,\ 0 \leq k \leq L-1$ demand to reliably decode messages $(M_k, \cdots, M_{L-1})$. Now the L-dimensional capacity region is the closure of all achievable rate-tuples $(R_0, R_2 \cdots, R_{L-1})$ with vanishing error probability.

Section 6.1 discusses the simplification with Euclidean geometry which applies to "very noisy" versions of this problem. For situations that need not be very noisy, Section 6.2 shows how the simplification via graphical models becomes useful. Graphical models provide a framework to systematically think about such broadcast situations. They enable us to solve some new classes of broadcast networks, which generalize many previously solved networks[1].

[1]Preliminary versions of results in Section 6.1 appeared in [13]. Results in Section 6.2 were first reported in [45] and are a joint work with Mitchell Trott (in addition to Lizhong Zheng).

Finally in Section 6.3, we briefly discuss the notion of message-wise UEP for networks, which is quite straightforward compared to bit-wise UEP for networks. Instead of degraded message sets, the message-wise UEP problem corresponds to the capacity of compound channels. We also discuss some relations between the network information theory formulation and the error exponent formulation.

6.1 Euclidean Information Theory

Many problems in information theory, including broadcasting with degraded message sets, involve optimizing the KL divergence between probability distributions. Understanding the structure of the optimum solution is helpful for characterizing the achievable regions as well as the converse bounds. For example, this could be helpful in converting a multi-letter characterization of a capacity region into a single-letter characterization. Moreover, even when a single-letter characterization is available, it is often implicit in the form of an optimization in the space of distributions. Knowledge of these optimum distributions gives additional insights into the capacity region and design of good codes.

However, there is no systematic approach for finding the optimum solutions in general. Since KL divergence is difficult to analyze, these optimizations are often intractable. A main source of difficulty is that the KL divergence is not a metric in the space of probability distributions. Another source of difficulty is that the dimension of these optimizations can be unbounded in multi-letter characterizations, which makes them even harder.

We simplify these problems by assuming the distributions of interest to be close to each other. After restricting our attention to a local neighborhood of distributions, the KL divergence behaves like a squared Euclidean distance and the manifold of distributions behaves like a Euclidean space. We demonstrate that, with this approach, new insights can be obtained on the structure of the optimal solutions for such optimizations. Specifically, this simplification completely solves the multi-letter

optimizations arising in broadcast with degraded message sets in very noisy networks[2].

First we obtain a simple upper bound on the KL divergence between two distributions P and Q of a discrete random variable Z, which takes values from a finite alphabet $\mathcal{Z}$. We think of P and Q as $|\mathcal{Z}|$-dimensional row-vectors and assume all their elements to be strictly positive, $P, Q > 0$. Using $\ln(1+t) \geq t - \frac{t^2}{2}$,

$$\begin{aligned} D(P\|Q) &= -\sum_{z\in\mathcal{Z}} P(z) \ln\left(1 + \frac{Q(z)-P(z)}{P(z)}\right) \\ &\leq -\sum_{z\in\mathcal{Z}} P(z)\left(\frac{Q(z)-P(z)}{P(z)} - \frac{(Q(z)-P(z))^2}{2P(z)^2}\right) \\ &= 0 + \sum_{z\in\mathcal{Z}} \frac{(Q(z)-P(z))^2}{2P(z)} \\ &\equiv \frac{1}{2}\|Q-P\|_P^2 \\ &= \frac{1}{2}\|[P^{-1/2}](Q-P)'\|^2 \end{aligned}$$

where $(Q-P)'$ denotes the transpose of row-vector $(Q-P)$ and $\|a\|_b^2$ denotes the squared norm of a weighted with b as its weight vector: $\|a\|_b^2 \equiv \sum_i \frac{a(i)^2}{b(i)}$ for $b > 0$. Matrix $[P^{-1/2}]$ is a diagonal matrix whose i'th diagonal entry equals $\sqrt{1/P(i)}$.

The above bound on $D(P\|Q)$ becomes tight when P and Q are close, i.e.,

$$P = Q + \epsilon T \quad \text{for } \epsilon \to 0$$

and the perturbation direction T is a fixed row-vector along the probability simplex satisfying $\sum_i T(i) = 0$. With this assumption, the divergence bound becomes $D(P\|Q) \leq \frac{\epsilon^2}{2}\|T\|_P^2$. Using $\ln(1+t) \leq t - \frac{t^2}{2} + \frac{t^3}{3}$, it turns out that difference between $D(P\|Q)$ and $\frac{\epsilon^2}{2}\|T\|_P^2$ is of the order ϵ^3.

$$|\frac{\epsilon^2}{2}\|T\|_P^2 - D(P\|Q)| \leq \frac{1}{3}\sum_z \frac{(P(z)-Q(z))^3}{P(z)^2} = \frac{\epsilon^3}{3}\sum_z \frac{T(z)^3}{P(z)^2} = O(\epsilon^3)$$

Hence up to order ϵ^2, we have $D(P\|Q) \approx \frac{\epsilon^2}{2}\|T\|_P^2$. The $\approx$ sign is used as a shorthand

[2]Additional applications of this Euclidean approach to other problems in source coding and not very noisy situations can be found in [13].

for equality up to ϵ^2 order:

$$f \approx g \quad \Leftrightarrow \quad \lim_{\epsilon \to 0} \frac{f}{\epsilon^2} = \lim_{\epsilon \to 0} \frac{g}{\epsilon^2} \quad \Leftrightarrow \quad f = g + o(\epsilon^2)$$

Note that the approximation $D(P\|Q) \approx \epsilon^2 \|T\|_P^2$ is valid even if the subscript P in $\|T\|_P^2$ (the weight vector for the squared norm) is changed to another nearby distribution $\hat{P}$ with strictly positive components[3].

$$\begin{aligned} \epsilon^2 \|T\|_P^2 &= \|Q - P\|_P^2 \approx \|Q - P\|_{\hat{P}}^2 \\ \text{where } \hat{P} &= P + \epsilon \hat{T} \quad \text{for } \sum \hat{T}(i) = 0. \end{aligned}$$

Thus we can view the weight vector as only dependent on the neighborhood of distributions. The divergence between any pair of distributions in this neighborhood has the same weight vector for its Euclidean approximation. In particular, this implies

$$D(P\|Q) \approx D(Q\|P) = \|P - Q\|_Q^2 \quad \text{when } P = Q + \epsilon T \tag{6.1}$$

With this simplification, we first visit the degraded broadcast channel problem [54, 48]. Our solution will shed some new light on this already solved problem. We will then move to the broadcast problem with degraded message sets [42] for two or more users. We obtain some new insights on the capacity region of such problems in terms of the singular value decomposition (SVD) of certain matrices, which depend on the channels involved.

6.1.1 Euclidean Approximation

Now let us write mutual information in this approximation. Let X, Y be a pair of discrete random variables with marginal distributions P_X, P_Y . Let $W_{Y|X}$ denote the probability transition matrix of a very noisy channel, which means that all conditional output distributions $W_{Y|X}(\cdot|x)$ are close to a fixed output distribution $Q_Y(\cdot)$. Think

[3]Note that since P is assumed to have strictly positive components, any $\hat{P}$ in a small enough neighborhood of P will also have strictly positive components.

of these distributions as nearby points on the probability simplex in the neighborhood of $Q_Y(\cdot)$.

$$W_{Y|X}(\cdot|x) = Q_Y(\cdot) + \epsilon\Theta_{Y|X}(\cdot|x) \quad \text{where} \quad \sum_y \Theta_{Y|X}(y|x) = 0 \tag{6.2}$$

$$W_{Y|X} = \mathbf{1}'Q_Y + \epsilon\Theta_{Y|X} \quad \text{where} \quad \Theta_{Y|X}\mathbf{1}' = \mathbf{0}'. \tag{6.3}$$

where $\Theta_{Y|X}$ denotes the perturbation matrix, $\mathbf{1}'$ denotes a column vector of all ones, and $\mathbf{1}'Q_Y$ denotes the matrix whose every row equals Q_Y. Thus each $W_{Y|X}(\cdot|x)$ is obtained by perturbing $Q_Y(\cdot)$ along direction $\Theta_{Y|X}(\cdot|x)$. The marginal output distribution $P_Y = P_X'W_{Y|X}$ is a convex combination of $\{W_{Y|X}(\cdot|x), x \in \mathcal{X}\}$, so it is also close to $Q_Y(\cdot)$. Using (6.1) and the definition of $I(X;Y)$,

$$I(X;Y) = \mathsf{E}_{P_X}\left[D(W_{Y|X}(\cdot|X)\|P_Y(\cdot))\right] \tag{6.4}$$

$$\approx \frac{1}{2}\mathsf{E}_{P_X}\left[\|W_{Y|X}(\cdot|X) - P_Y(\cdot)\|^2_{Q_Y}\right] \tag{6.5}$$

$$= \frac{1}{2}\mathsf{E}_{P_X}\left[\| \left[Q_Y^{-1/2}\right](W_{Y|X}(\cdot|X) - P_Y(\cdot))'\|^2\right] \tag{6.6}$$

Since $P_Y(\cdot)$ is the average of $W_{Y|X}(\cdot|X)$ under P_X, the above approximation for mutual information in (6.5) looks like (half of) the 'variance' of[4] the conditional distributions $\{W_{Y|X}(\cdot|x), x \in \mathcal{X}\}$. Although, remember that instead of the usual Euclidean norm, we are taking the weighted Euclidean norm according to Q_Y.

Remark: The capacity achieving output distribution $P_Y^*(\cdot)$ now has a very intuitive geometric interpretation now. Recall that for every input x used in the capacity achieving distribution, $D(W_{Y|X}(\cdot|x)\|P_Y^*(\cdot))$ equals capacity [2]. Under our simplification, this means that (half of) the weighted squared Euclidean norm $\|W_{Y|X}(\cdot|x) - P_Y^*(\cdot)\|^2_{Q_Y}$ equals capacity for all those inputs. Thus P_Y^* is the "circum-center" of the polygon formed by various $W_{Y|X}(\cdot|x)$ and the channel capacity equals half the squared "circum-radius" of this polygon. The quotation marks are to emphasize that instead of the standard Euclidean squared norm, we take a weighted squared norm according

[4]To physicists, this looks like the squared radius of gyration of $|\mathcal{X}|$ point masses in $|\mathcal{Y}|$ dimensional space, where point x is located at $W_{Y|X}(\cdot|x)$ and has mass $P_X(x)$.

to Q_Y. However, if Q_Y is the uniform distribution, then the weighted squared norm is simply a multiple of the standard (unweighted) Euclidean norm and the channel capacity is related to the standard Euclidean circum-radius of this polygon. Figure 6-2 shows this result for a channel with a ternary input alphabet $\mathcal{X} = \{1, 2, 3\}$.

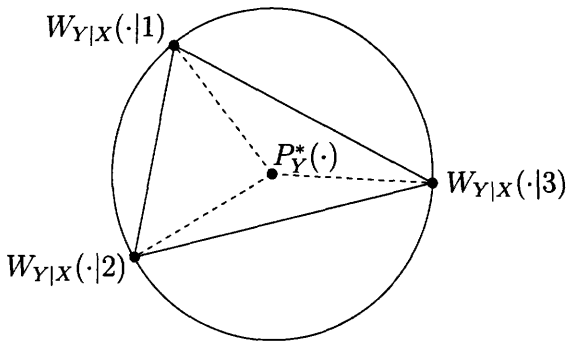


Figure 6-2: Geometric interpretation of channel capacity. Capacity achieving P_Y^* is "equidistant" from every conditional output distribution—it is the "circum-center" of the channel and half of its squared circum-radius equals the channel capacity.

6.1.2 Degraded broadcast channel

Now we consider the physically degraded broadcast channel (Fig. 6-3) from X to Y and Z. Let $W_{Y|X}$ and $\Omega_{Z|Y}$ denote the channels from X to Y and from Y to Z, respectively.

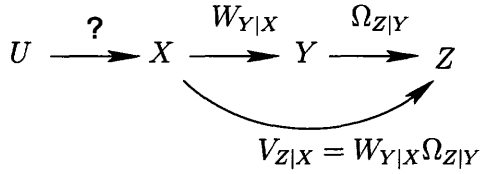


Figure 6-3: Physically degraded broadcast channel

It is known [48] that the achievable rates are $R_1 = I(U; Z)$ for common information (i.e., crucial bits) to both receivers and $R_0 = I(X; Y|U) = I(X; Y) - I(U; Y)$ for private information (i.e., ordinary bits) to Y, where U satisfies the Markov relation $U - X - Y - Z$. The capacity region is given by the following optimization over joint

(U, X) distributions.

$$R_1^* = \max_{P_{UX}:\ I(X;Y|U) \geq R_0} I(U;Z) \tag{6.7}$$

$$= \max_{P_{UX}:\ I(U;Y) \leq I(X;Y) - R_0} I(U;Z) \tag{6.8}$$

In the following, we fix the X distribution at P_X and only focus on optimizing the choice of U. Hence $I(X;Y) - R_0$ is a fixed constant (say γ) and $I(U;Y) \leq \gamma$.

To apply the Euclidean approach, we assume channel $W_{Y|X}$ to be very noisy. This implies that all Y distributions of interest are close to each other, in the neighborhood of Q_Y.

$$\begin{aligned} W_{Y|X} &= \mathbf{1}' Q_Y + \epsilon \Theta_{Y|X} \\ \text{and} \quad V_{Z|X} &= W_{Y|X} \Omega_{Z|Y} = \mathbf{1}' Q_Y \Omega_{Z|Y} + \epsilon \Theta_{Y|X} \Omega_{Z|Y} \\ &= \mathbf{1}' Q_Z + \epsilon (\Theta_{Y|X} \Omega_{Z|Y}) \qquad \text{where } Q_Z \equiv Q_Y \Omega_{Z|Y} \end{aligned}$$

Thus the very noisy nature of $W_{Y|X}$ implies that $V_{Z|X}$ is also very noisy and all Z distributions are in the neighborhood of Q_Z.

Now instead of directly calculating the optimum distribution (U, X), we first calculate the optimum (U, Y) distribution for a given P_X. The optimum (U, X) distribution can be easily obtained from this optimum (U, Y) distribution. The complete capacity region is obtained by repeating these steps for every P_X and taking the closure of all those solutions. We only focus on the optimization for a given P_X henceforth.

Define our optimization variables as

$$P_{Y|U}(\cdot|u) - P_Y(\cdot) \equiv \epsilon \theta_u$$

where P_Y is the fixed output distribution corresponding to given P_X. We only need to choose the optimum $P_{Y|U}$ for this P_Y. Equivalently, we only need to optimize the perturbations $\{\theta_u\}$. Now note that physical degradedness implies,

$$P_{Z|U}(\cdot|u) - P_Z(\cdot) = (P_{Y|U}(\cdot|u) - P_Y(\cdot)) \Omega_{Z|Y} = \epsilon \theta_u \Omega_{Z|Y}.$$

We can now rewrite the optimization in (6.8) using the Euclidean approximation in (6.6).

$$\max_{P_U,\theta_U:\ \mathsf{E}_{P_U}\left[\|\ [Q_Y^{-1/2}]\cdot\theta_U'\|^2\right]\le 2\gamma/\epsilon^2} \mathsf{E}_{P_U}\left[\|\ [Q_Z^{-1/2}]\cdot(\theta_U\Omega_{Z|Y})'\|^2\right]\cdot\frac{\epsilon^2}{2} \tag{6.9}$$

Performing a change of variables, define column vector $\phi_u \equiv [Q_Y^{-1/2}]\theta_u'$ and $\hat{\gamma} \equiv \gamma/\epsilon^2$. This converts (6.9) to the following optimization for R_1^*:

$$\frac{R_1^*}{\epsilon^2} = \frac{1}{2}\left(\max_{P_U,\phi_U:\ \mathsf{E}_{P_U}[\|\phi_U\|^2]\le 2\hat{\gamma}} \mathsf{E}_{P_U}\left[\|\ B\cdot\phi_U\|^2\right]\right) \tag{6.10}$$

$$\text{where}\quad B \equiv [Q_Z^{-1/2}]\cdot\Omega_{Z|Y}'\cdot[Q_Y^{1/2}] \quad \&\quad \hat{\gamma} = \frac{I(X;Y)-R_0}{\epsilon^2} \tag{6.11}$$

We call B the *divergence translation matrix*, since it transforms the divergence between Y distributions to that between Z distributions. If Q_Y and Q_Z are uniform distributions, then this matrix is equivalent to $\Omega_{Z|Y}'$, the (transposed) channel matrix itself. This optimization is a standard problem in linear algebra. Its solution depends on the SVD of B, which has the following property due to the data-processing theorem.

Lemma 37 *Let $\sigma_1, \sigma_2 \cdots$ denote the singular-values of B in descending order and the corresponding (right) singular vectors be $v_1, v_2 \cdots$. Then the largest singular-value σ_1 is 1 and $v_1 = {Q_Y'}^{1/2}$, which denotes the element-wise square-root of vector Q_Y.*

If there were no constraints on ϕ_u, the optimal choice of each ϕ_u should be along v_1, the singular vector with the largest singular-value. This is essentially like multi-antenna beamforming for maximum power gain—putting all the "power" along the largest eigenvector.

However, it turns out that v_1 is an infeasible direction for $\phi_u = [Q_Y^{-1/2}]\theta_u'$, where $\theta_u = P_{Y|U}(\cdot|u) - P_Y(\cdot)$. Note that θ_u lies along the probability simplex, so it satisfies

$$\begin{aligned}\sum\theta_u(\cdot) = 0 \quad &\Leftrightarrow\quad v_1'\cdot[Q_Y^{-1/2}]\theta_u' = v_1'\phi_u = 0\\ &\Rightarrow\quad v_1 \perp \phi_u \in \text{span}(v_2, v_3\cdots)\end{aligned}$$

This means that linear combinations of $\{v_2, v_3 \cdots\}$ correspond to all feasible θ_u directions along the simplex. Since $\phi_u \in \text{span}(v_2, v_3 \cdots)$, the optimal ϕ_u lies along v_2, the *feasible* direction with the largest singular value. This implies,

$$\frac{R_1^*}{\epsilon^2} \leq \sigma_2^2 \hat{\gamma} = \sigma_2^2 \frac{I(X;Y) - R_0}{\epsilon^2}$$

with equality achieved when R_0 is close to $I(X;Y)$ and ϕ_u can be chosen entirely along v_2.

Thus σ_2^2 equals the slope of the optimal R_1 vs. R_0 curve at the $R_1 = 0$ intercept. Note that $\sigma_2 \leq 1$ reflects the fact that the loss in R_0 for the better user is not completely compensated by the corresponding increase in R_1 for the degraded user. Perhaps it is surprising that this efficiency factor, i.e., the slope of the R_0 vs. R_1 curve depends on the degradation link $\Omega_{Z|Y}$ (its σ_2) but not at all on the channel $W_{Y|X}$ to the better user.

One can increase $R_1 = I(U;Z)$ by extending θ_u further along the direction corresponding to v_2 until $P_{Y|U}(\cdot|u)$ reaches the boundary of its feasible set. This feasible set is the convex hull of the channel $\{W_{Y|X}(\cdot|x), x \in \mathcal{X}\}$. Then to further increase R_1, the choice of θ_u will move along its boundary. The normalized rates $(\hat{R}_0, \hat{R}_1)$ are defined as

$$\hat{R}_0 \equiv \lim_{\epsilon \to 0} \frac{R_0}{\epsilon^2} \quad \text{and} \quad \hat{R}_1 \equiv \lim_{\epsilon \to 0} \frac{R_0}{\epsilon^2} \qquad \Leftrightarrow \qquad (\hat{R}_0, \hat{R}_1) \approx \epsilon^2 (R_0, R_1)$$

The resulting $(\hat{R}_0, \hat{R}_1)$ capacity region should hence be piecewise linear in shape. This reminds us of some results in [44] where the rate region was a triangle.

It is more interesting to consider the multi-letter problem. In the following, we study the 2-letter case and the general n-letter case follows on the same lines. Without loss of optimality, we can fix the distribution of input pair $(X_1, X_2) \equiv X_1^2$ to be i.i.d. P_X over time $P_{X_1^2} = P_X \otimes P_X$, which implies $P_{Y_1^2} = P_Y \otimes P_Y$, $P_{Z_1^2} = P_Z \otimes P_Z$, where $\otimes$ denotes the Kronecker product. Even if X_1^2 are not i.i.d, the difference between $P_{Y_1^2}$ and $P_Y \otimes P_Y$ will be of higher order than ϵ due to the very noisy nature of the

channel. It is because the two letter channel $W_{Y_1^2|X_1^2}$ equals

$$\begin{aligned} W_{Y_1^2|X_1^2} &= W_{Y|X} \otimes W_{Y|X} = (\mathbf{1}'Q_Y + \epsilon\Theta_{Y|X}) \otimes (\mathbf{1}'Q_Y + \epsilon\Theta_{Y|X}) \\ \Rightarrow \quad W_{Y_1^2|X_1^2} &= \mathbf{1}'Q_Y \otimes \mathbf{1}'Q_Y + \epsilon \mathbf{1}Q_Y \otimes \Theta_{Y|X} + \epsilon\Theta_{Y|X} \otimes \mathbf{1}'Q_Y + O(\epsilon^2) \text{ terms} \end{aligned}$$

Due to this property, any correlations between X_1 and X_2 are only reflected in the $O(\epsilon^2)$ terms of $P_{Y_1^2} = P_{X_1^2} W_{Y_1^2|X_1^2}$ and similarly of $P_{Z_1^2} = P_{X_1^2} V_{Z_1^2|X_1^2}$. In other words, if two distributions $P_{X_1^2}$ and $\bar{P}_{X_1^2}$ have the same marginal distributions of X_1 and X_2, then their corresponding $P_{Y_1^2}$ and $\bar{P}_{Y_1^2}$ will be the same up to order ϵ. Hence in our scaling of interest, we might as well assume i.i.d. X_i. Remember that if the difference between two distributions is $O(\epsilon^2)$, then the KL divergence between them is $O(\epsilon^4)$, which is negligible in the ϵ^2 scaling of our interest.

Also note that the memoryless property of the channel implies $P_{Z_1^2|Y_1^2} = \Omega_{Z|Y} \otimes \Omega_{Z|Y}$. For the 2-letter case, we need to consider the 2-letter version of the optimization in (6.8). This optimization is over joint distributions of (U, X_1^2) satisfying the Markov chain $U - X_1^2 - Y_1^2 - Z_1^2$.

$$\frac{1}{2} \max_{P_{UX_1^2}:\ I(U;Y_1^2) \le 2(I(X;Y) - R_0)} I(U; Z_1^2) \tag{6.12}$$

The initial factor of $1/2$ is to normalize the rate per channel use and also the factor of 2 in $2(I(X;Y) - R_0)$ arises due to two-channel uses.

Substitute $\epsilon\theta_u = P_{Y_1^2|U}(\cdot|u) - P_{Y_1^2}(\cdot)$ as before and denote the i.i.d. Q_Y distribution as $Q_{Y_1^2}$. Similarly define $Q_{Z_1^2}$. Again, with the local assumption on distributions and substituting $\phi_u = [Q_{Y_1^2}^{-1/2}]\theta_u'$, we can rewrite this optimization

$$\begin{aligned} & \frac{1}{2} \left(\max_{P_U, \phi_U:\ \mathsf{E}_{P_U}[\|\phi_U\|^2] \le 4\gamma} \mathsf{E}_{P_U} \left[\| B^{(2)} \cdot \phi_U \|^2 \right] \right) \\ \text{where } B^{(2)} &\equiv [Q_{Z_1^2}^{-1/2}] \cdot (\Omega_{Z|Y} \otimes \Omega_{Z|Y})' \cdot [Q_{Y_1^2}^{1/2}] = B \otimes B \end{aligned}$$

is the divergence translation matrix for this 2-letter case.

Lemma 38 *Let v_i and v_j denote two singular vectors of B with singular-values σ_i*

and σ_j. Then $v_i \otimes v_j$ is an singular vector of $B^{(2)}$ and its singular-value is $\sigma_i\sigma_j$.

Again, the largest singular-value equals 1 corresponding to the (right) singular vector $v_1 \otimes v_1$, which is an infeasible direction due to simplex constraint. The singular vectors $v_1 \otimes v_2$ and $v_2 \otimes v_1$ correspond to the second largest singular-value σ_2.

Recalling $v_1 = {Q'_Y}^{1/2}$ to notice that ϕ_u along $v_1 \otimes v_2$ translates to θ_u along $Q_Y \otimes \alpha$, where $\alpha = [Q_Y^{1/2}]v_2$. Similarly, $v_2 \otimes v_1$ translates to $\alpha \otimes Q_Y$. Thus the optimal $P_{Y_1^2|U}(\cdot|u)$ for any u looks like

$$\begin{aligned} P_{Y_1^2|U}(\cdot|u) &\equiv P_{Y_1^2} + \epsilon\theta_u \\ \text{(for some constant } c_1, c_2) &= P_Y \otimes P_Y + \epsilon c_1(Q_Y \otimes \alpha) + \epsilon c_2(\alpha \otimes Q_Y) \\ &= P_Y \otimes P_Y + \epsilon c_1(P_Y \otimes \alpha) + \epsilon c_2(\alpha \otimes P_Y) + O(\epsilon^2) \\ &\approx (P_Y + \epsilon c_1\alpha) \otimes (P_Y + \epsilon c_2\alpha) \end{aligned}$$

In the second last step, the error in replacing Q_Y by P_Y is $O(\epsilon^2)$ since the difference between P_Y and Q_Y is $O(\epsilon)$. The last step follows by adding $\epsilon^2 c_1 c_2(\alpha \otimes \alpha)$, which is of smaller order than the other terms and is negligible in our scale of interest. These $O(\epsilon^2)$ error terms can be ignored as mentioned earlier, since they cause a negligible $O(\epsilon^4)$ difference in the KL divergences.

This analysis says that any optimal conditional distribution $P_{Y_1^2|U}(\cdot|u)$ is independent over time in the scaling of our interest. The multi-letter this decomposes into a sum of single-letter optimizations and hence i.i.d. replication of the single letter solution becomes optimal. The final value to the 2-letter optimization is exactly the same as the single-letter optimization seen earlier. This fact is proved in [48] under general conditions using information theoretic technique of substitution of auxiliaries.

For any given number of letters n, the same results hold true when ϵ goes to zero and the multi-letter optimization becomes equivalent to the single-letter optimization. It is worth mentioning that we are fixing the number of letters n and let ϵ go to zero. This is different from fixing ϵ and letting n go to infinity, where the Euclidean approximations are not clearly justified. Unfortunately, for proving rigorous converses for these problems using Fano's inequality and related techniques, one needs to let n

go to infinity first. Resolving this issue of the order of taking limits is an important future direction. These comments also hold true for the next subsection.

6.1.3 Broadcast with degraded message sets

Now consider the situation when Z is not a degraded version of Y. For using our Euclidean framework, we assume that $W_{Y|X}$ and $V_{Z|X}$ are both very noisy. That is, for every input x, the conditional distributions $W_{Y|X}(\cdot|x)$ and $V_{Z|X}(\cdot|x)$ are close to some $Q_Y(\cdot)$ and $Q_Z(\cdot)$, respectively. Following our earlier notation,

$$W_{Y|X} = \mathbf{1}Q'_Y + \epsilon \cdot \Theta \quad \& \quad V_{Z|X} = \mathbf{1}Q'_Z + \epsilon \cdot \Phi \tag{6.13}$$

where $\mathbf{1}Q'_Z$ denotes a matrix whose every row equals Q_Z and Θ and Φ are fixed matrices whose every row sums to 0.

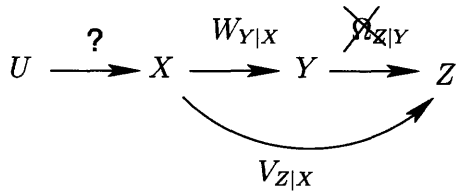


Figure 6-4: General two-user broadcast channel

Receiver Z wants to decode a common message (special bits) at rate R_1 and Y wants to decode this common message as well as a private message at rate R_0. The capacity region in this case is given by [42].

$$R_1^* = \max_{U-X-(YZ):\ I(X;Y|U)\geq R_0} \min\{I(U;Z), I(U;Y)\} \tag{6.14}$$

As before, we fix the X distribution at P_X and only focus on optimizing the choice of U, which means that $I(X;Y) - R_0$ is a fixed constraint (say γ) on $I(U;Y)$. Hence it is sufficient to solve the following optimization.

$$\max_{U-X-(YZ):\ I(U;Y)\leq\gamma} I(U;Z) \tag{6.15}$$

Under the very noisy assumptions, this is equivalent to solving

$$\frac{1}{2}\left(\max_{P_U,\phi_U:\ \mathsf{E}_{P_U}[\|\phi_U\|^2]\le 2\gamma} \mathsf{E}_{P_U}\left[\|\ B_{Y\to Z}\cdot\phi_U\|^2\right]\right) \tag{6.16}$$

$$\text{where}\quad B_{Y\to Z} = [Q_Z^{1/2}]\cdot(V'_{Z|X}W'^{-1}_{Y|X})\cdot[Q_Y^{1/2}] \tag{6.17}$$

As before, the divergence translation matrix $B_{Y\to Z}$ has a singular-vector $Q_Y^{1/2}$ with singular-value 1. However, since $V'_{Z|X}W'^{-1}_{Y|X}$ is not a (transposed) probability transition matrix like $\Omega'_{Z|Y}$ before, $B_{Y\to Z}$ could have singular-values larger than 1.

Assuming v_i is such a singular vector of $B_{Y\to Z}$ with singular-value $\sigma_i \ge 1$, we should choose ϕ_u along v_i for small enough γ. The optimum $I(U;Z)$ in this case satisfies $I(U;Y) \le \gamma \le I(U;Z)$, so the common information R_1 is bottlenecked at $R_1 = I(U;Y)$. Hence for small enough R_1, the R_1–R_0 tradeoff is $R_1 + R_0 = I(X;Y)$ as $R_0 = I(X;Y|U)$.

If all singular values of $B_{Y\to Z}$ are upper bounded by 1 with $\sigma_1 = 1$, then the common information is bottlenecked by Z and the slope of R_1–R_0 tradeoff for small R_1 equals σ_2^2, where σ_2 is second largest singular-value of $B_{Y\to Z}$. We should mention that even for a non-degraded broadcast channel, all singular values of $B_{Y\to Z}$ could upper bounded by 1. We will see soon see an example of this where $W_{Y|X}$ is a BSC and $V_{Z|X}$ is an asymmetric binary channel.

Multi-letter case

Lets define the divergence translation matrices from X to Y and from X to Z.

$$B_{X\to Y} = [Q_Y^{-1/2}]W'_{Y|X}[P_X^{1/2}] \quad \& \quad B_{X\to Z} = [Q_Z^{-1/2}]V'_{Z|X}[P_X^{1/2}]$$

and note that $B_{Y\to Z} = B_{X\to Z}B_{X\to Y}^{-1}$.

Let $\{\mu_1,\mu_2\cdots\}$ and $\{\nu_1,\nu_2\cdots\}$ denote the singular-values of $B_{X\to Y}$ and $B_{X\to Z}$ in descending order and let $\{g_1,g_2\cdots\}$ and $\{h_1,h_2\cdots\}$ denote the corresponding singular vectors (respectively). It can be verified that the largest singular values $\mu_1 = \nu_1 = 1$ and $g_1 = h_1 = P'^{1/2}_X$. Moreover, since $W_{Y|X}$ and $V_{Z|X}$ are very noisy as

in (6.13), all other singular values are of the order $O(\epsilon)$ and smaller.

For the 2-letter case, we can restrict to i.i.d. X_1^2 without loss of optimality as discussed earlier for the physically degraded situation. This implies $P_{X_1^2} = P_X \otimes P_X$ and $P_{Y_1^2} = P_Y \otimes P_Y$. Now using Lemma 45 for the 2-letter case, the singular-values of $B^{(2)}_{X \to Y}$ can be divided into three classes: $\mu_1\mu_1 = 1$, $\{\mu_1\mu_i = \mu_i = O(\epsilon) \text{ for } i \neq 1\}$, and $\{\mu_i\mu_j = O(\epsilon^2) \text{ for } i,j \neq 1\}$. The corresponding singular vectors are $g_1 \otimes g_1$, $\{g_i \otimes g_1 \ \& \ g_1 \otimes g_i\}$ and $\{g_i \otimes g_j\}$.

Choosing $\theta_u(\cdot) = P_{X_1^2|U}(\cdot|u) - P_{X_1^2}(\cdot)$ in the direction corresponding to $g_1 \otimes g_1$ is infeasible due to the simplex constraint. Choosing θ_u along the direction corresponding $g_1 \otimes g_i$ and $g_i \otimes g_1$ causes an $O(\epsilon)$ change in $P_{Y_1^2|U}(\cdot|u)$ and hence an $O(\epsilon^2)$ change in $I(U;Y)$. On the other hand, choosing θ_u along the direction corresponding $g_i \otimes g_j$causes an $O(\epsilon^2)$ change in $P_{Y_1^2|U}(\cdot|u)$ and hence an $O(\epsilon^4)$ change in $I(U;Y)$, which is negligible for us.

Now the only relevant singular vectors, $g_i \otimes g_1$ and $g_1 \otimes g_i$, correspond to θ_u of the form $Q_Y \otimes \alpha_i$ and $\alpha_i \otimes Q_Y$, where $\alpha_i \equiv [Q_Y^{1/2}]v_i$ for $i \neq 1$. Now as seen earlier for the physically degraded case, the optimal $P_{Y_1^2|U}(\cdot|u)$ looks like

$$\begin{aligned} P_{Y_1^2|U}(\cdot|u) &= P_{Y_1^2} + \epsilon c_1(Q_Y \otimes \hat{\alpha}) + \epsilon c_2(\tilde{\alpha} \otimes Q_Y) \\ &= P_Y \otimes P_Y + \epsilon c_1(P_Y \otimes \hat{\alpha}) + \epsilon c_2(\tilde{\alpha} \otimes P_Y) + O(\epsilon^2) \text{ terms} \\ &\approx (P_Y + \epsilon c_1\hat{\alpha}) \otimes (P_Y + \epsilon c_2\tilde{\alpha}) \end{aligned}$$

for some constant c_1, c_2. Here $\hat{\alpha}$ and $\tilde{\alpha}$ are some linear combinations of $\{\alpha_2, \alpha_3 \cdots\}$. As earlier, the second last step followed by replacing Q_Y by P_Y and last step followed by adding $\epsilon^2 c_1 c_2(\hat{\alpha} \otimes \tilde{\alpha})$, which is of smaller order than the other terms.

Thus the conditional distribution $P_{Y_1^2|U}(\cdot|u)$ is independent over time; hence a single letter solution is optimal even for this not physically degraded situation. The same analysis for 2-letter case can be repeated for any n-letter case. Notice that all this treatment is almost the same as the physically degraded situation, except that now some singular values of $B_{Y \to Z}$ could be larger than 1.

For a general 2-user broadcast channel, the optimality of single-letter character-

ization was proved in [42] using information theoretic techniques. Their technique does not extend to more than 2 receivers for broadcast with degraded message sets. However, our analysis can be applied for any number receivers as long as all channels are assumed to be very noisy. However, as mentioned earlier, the weakness of this technique is we cannot grow the code-length n to infinity first.

Three user example: Consider a broadcast channel where $W_{Y|X}, V_{Z_1|X}$, and $V_{Z_2|X}$ respectively denote channels to users Y, Z_1, and Z_2. Users Z_1 and Z_2 want the common message at rate R_1 and user Y wants the common message as well as a private message at rate R_0.

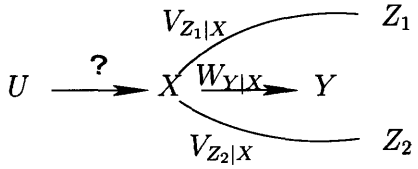


Figure 6-5: General three-user broadcast channel

The R_1–R_0 tradeoff achievable with superposition coding is obtained by solving,

$$\max_{U-X-(YZ_1Z_2):\ I(U;Y)\leq\gamma} \min\{I(U;Z_1), I(U;Z_2)\} \quad \text{where} \quad \gamma = I(X;Y) - R_0 \quad (6.18)$$

With Euclidean approximation, this simplifies to maximizing

$$\begin{aligned} \min\ \{\mathsf{E}_{P_U}\left[\|\ B_{Y\to Z_1}\cdot\phi_U\|^2\right],\quad \mathsf{E}_{P_U}\left[\|\ B_{Y\to Z_2}\cdot\phi_U\|^2\right]\} \\ \text{over the set: } \{P_U, \phi_U:\ \mathsf{E}_{P_U}\left[\|\phi_U\|^2\right] \leq 2\gamma\} \\ \text{where } \ B_{Y\to Z_i} = [Q_{Z_i}^{1/2}]\cdot(V'_{Z_i|X}W'^{-1}_{Y|X})\cdot[Q_Y^{1/2}] \end{aligned}$$

Here we have to choose a direction of ϕ_u to maximize the minimum of two quadratic forms. The solution need not be directly related to singular vectors of $B_{Y\to Z_1}$ and $B_{Y\to Z_2}$, nonetheless, it is just a quadratic optimization. It is particularly easy to solve when the singular-vectors of $B_{Y\to Z_1}$ and $B_{Y\to Z_2}$ are aligned.

Binary example: User Y wants both private and common information and channel to Y is a very noisy BSC: $W_{Y|X} = [1/2+\epsilon,\ 1/2-\epsilon;\ 1/2-\epsilon,\ 1/2+\epsilon]$. Let $V_{Z_1|X} =$

$[0.03+\epsilon/10,\ 0.97-\epsilon/10;\ 0.03-\epsilon/10,\ 0.97+\epsilon/10]$ be an asymmetric binary channel and let $V_{Z_2|X} = [0.97+\epsilon/10,\ 0.03-\epsilon/10;\ 0.97-\epsilon/10,\ 0.03+\epsilon/10]$ be its flipped version.

Note that neither $V_{Z_1|X}$ nor $V_{Z_2|X}$ are degraded versions of $W_{Y|X}$ because there is no stochastic matrix D which gives $W_{Y|X}D = V_{Z_i|X}$. However, for any P_X, all singular values of $B_{Y\to Z_1}$ or $B_{Y\to Z_2}$ are not greater than 1. Hence the bottlenecks for common information are $V_{Z_1|X}$ and $V_{Z_1|X}$, not $W_{Y|X}$.

By symmetry between $V_{Z_1|X}$ and $V_{Z_2|X}$, its clear that the optimum P_X is uniform binary, which is also the capacity achieving distribution for $W_{Y|X}$. The entire capacity region is obtained by superposition coding with binary uniform U where $P_{X|U}$ is a BSC with parameter t, which needs to be optimized.

If we convert $W_{Y|X}$ to a noisier BSC, $W_{Y|X} = [1/2+\epsilon/10,\ 1/2-\epsilon/10;\ 1/2-\epsilon/10,\ 1/2+\epsilon/10]$, then for any P_X, all eigenvalues of all singular values of $B_{Y\to Z_1}$ or $B_{Y\to Z_2}$ are not smaller than 1. Hence the bottlenecks for common information is Y itself, not Z_1 or Z_2. The capacity region for this problem is simply the sum rate constraint, $R_0 + R_1 = C$, where C is the channel capacity of $W_{Y|X}$.

6.2 Graphical Models: Multilevel Broadcast Networks

In this formulation, we represent the set of all receivers on a directed graph—called *degradation graph*—based on the quality of their observations. Each receiver corresponds to a node in this graph. If receiver Z is a physically degraded version of receiver Y, then node Z is a child of node Y in this graph which has a directed edge from Y to Z to indicate degradedness. We will obtain upper and lower bounds on the capacity region of such networks[5]. The upper bounds are based on auxiliary variables,

[5]Throughout this chapter, edges in a degradation graph denote *physical* degradation. However, the capacity regions for a degradation graph holds true even if its edges represent *stochastic* degradation instead of physical degradation. It is because the capacity region of a broadcast network only depends on the individual (marginalized) channels to the receivers—not the joint network channel. The physical degradedness assumption is a nice technical trick for proving converses.

whose structure is described by the mirror image of the channel's degradation graph. A packet broadcast network is considered as an example.

A simple manner in which outputs at various receivers could be related to each other is a Markov chain, which gives rise to the classical degraded broadcast channel[6].

$$X \to Y_0 \to Y_1 \to \cdots \to Y_{L-1} \tag{6.19}$$

For this channel with input X, amongst any two outputs Y_k and Y_m, one is a degraded version of the other, that is, there exists a directed path between any two outputs. To get rid of this limitation, we model the interdependence between various receivers using a "degradation tree". See the graph in Figure 6-6 for example, which denotes that the network channel $W_{Y_0Y_1Y_2Z_1Z_2|X}$ can be decomposed as follows:

$$W_{Y_0Y_1Y_2Z_1Z_2|X} = W_{Y_0|X} W_{Y_1Y_2|Y_0} W_{Z_1|Y_1} W_{Z_2|Y_2} \tag{6.20}$$

The root of this degradation tree denotes the input X, which is followed by a unique output Y_0. It is followed by its degraded versions Y_1 and Y_2, which in turn have their own degraded versions Z_1 and Z_2. Note that here for example Z_2 need not be a degraded version of W_1 since there is no directed path from Y_1 to Z_1.

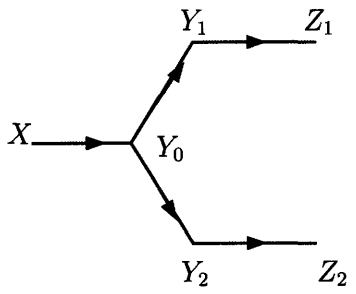


Figure 6-6: An example of degradation graph.

This degradation graph could have various physical interpretations. For example, it could represent users which are physically farther from the transmitter and hence

[6]With slight abuse of notation in earlier chapters, the subscript of Y_k denotes user k as opposed to output Y at time k.

receive noisier observations. It could also represent users which only listen during a fraction of the entire message transmission. In fact, even when there is only one physical user, the degradation graph could be used to represent the possible channel types due to large deviations of the channel.

One should be able to disseminate information such that even with partial observations, a user can receive some part of the transmitted information. Moreover, the information received should grow as the quality of observations improves. To formalize this notion of information dissemination, we can think of the degradation graph as a sequence of multiple layers, where layer k denotes the set of nodes at distance $k+1$ from the root. Outputs received at a user in layer $k+1$ is degraded versions of at least one user in layer k. Assuming total L layers (indexed from from 0 to $L-1$), let the total information be split into L parts. In our formulation, even the last layer users should get, the first part of this information (which is the most special part of information). The second last layer users should get the first as well as the second part and so on. Specifically, (last) layer $L-1$ should be able to decode message M_{L-1}, layer $L-2$ should decode M_{L-2} as well as M_{L-1}. In general[7], layer k should decode $M_k, M_{k+1} \cdots M_{L-1}$.

For example in Fig. 6-6, Z_1 and Z_2 are in layer 2, W_1 and W_2 are in layer 1 and Y is in layer 0. Hence both Z_1, Z_2 want message M_2; both Y_1, Y_2 want M_1 and M_2; and Y_0 wants M_0, M_1 and M_2.

Previous works addressing similar concept of information dissemination include multilevel diversity coding [50, 49] and priority encoded transmission [17], which can be modeled with specific degradation graphs.

We assume that messages $M_0, M_2 \cdots M_{L-1}$ are mutually independent. Message M_k $(0 \leq k \leq L-1)$ is chosen uniformly from $\{1, 2 \ldots, 2^{nR_k}\}$, where R_k denotes the rate of message k and n denotes the block length. We want to characterize the achievable rate region $(R_0, R_2 \ldots, R_{L-1})$.

It is worth mentioning that although the degradation graph in Fig. 6-6 was a

[7]This implies that the messages are indexed with increasing priority level: M_0 being least special to M_{L-1} being most special.

tree, it is not necessary. A degradation graph is a directed acyclic graph in general, which could have a node in layer k with multiple parents in layer $k-1$ if it is a noisier version of each parent. Thus a degradation graph has an edge directed from node A to node B whenever B is a noisier version of node A.

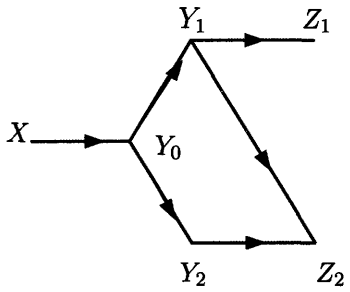


Figure 6-7: A degradation graph which is not tree.

For example in Fig. 6-7, Z_2 has two parents Y_1 and Y_2, thus Z_2 can be thought as a noisier version of Y_1 or Y_2. This generalization of allowing multiple parents is useful for modeling packet erasure networks as we will see later[8].

The remaining section is organized as follows. Subsection 6.2.1 studies the classical physically degraded broadcast channel. A converse was proved for the case of two users in [48] and also [53]. We first prove the converse for $L > 2$ users using a method similar to [48] for two users. As an example, we will consider broadcast over Binary Erasure Channels. A simple scheme of time-sharing between some binary linear codes can achieve the capacity region of this channel.

A channel with arbitrary degradation graph is studied in Section 6.2.2. An achievable rate region is presented and two converses are proved, which are based on mirror image of the degradation graph. This general converse follows on similar lines of the converse for classical physically degraded situation. This converse is not tight in general bit it converse gives the capacity region for a class of degradation graphs. This capacity region generalizes the results in [46] for a wider class of networks. We then study a packet erasure network as an example degradation graph, where our achievable rate region calculated in closed form. This example demonstrates how

[8]It is also useful in addressing the Gaussian MIMO situations. Thanks to Tie Liu of Texas A&M for pointing this out.

various problems such as priority encoded transmission and multilevel diversity coding ([50, 49] and [17]) can be analyzed in the degradation graph framework.

6.2.1 Classical degraded broadcast channel with multiple receivers

In a physically degraded broadcast channel to L users $(Y_0, Y_1 \ldots Y_{L-1})$, the network channel can be decomposed as

$$W_{Y_0 Y_1 \cdots Y_{L-1}|X} = W_{Y_0|X} W_{Y_1|Y_0} \cdots W_{Y_{L-1}|Y_{L-2}} \tag{6.21}$$

Thus this channel is fully described by the L probability transition functions $W_{Y_0|X}$, $W_{Y_1|Y_0} \cdots W_{Y_{L-1}|Y_{L-2}}$ for this channel. The following achievable region was proved in [54, 56] using superposition codes.

Consider a Markov chain

$$U_{L-1} - U_{L-2} \cdots - U_1 - X - Y_0 - Y_1 \cdots - Y_{L-1} \tag{6.22}$$

where $U_{L-1}, \cdots U_1$ are called auxiliary random variables. The joint distribution of all variables above is given by

$$(\bar{P}_{U_{L-1} U_{L-2} \cdots U_1 X}) \cdot (W_{Y_0 Y_1 \cdots Y_{L-1}|X})$$

where the second term is described by the channel as in (6.21). The first term denotes the joint distribution of input X and auxiliary variables. It is chosen by the code designer[9] and has a Markov structure:

$$\bar{P}_{U_{L-1} U_{L-2} \cdots U_1 X} = \bar{P}_{U_{L-1}} \left(\prod_{k=L-1}^{2} \bar{P}_{U_{k-1}|U_k} \right) \bar{P}_{X|U_1} \tag{6.23}$$

[9]Throughout this section, we will use $\bar{P}$ for distributions chosen by the code designer and P for channel transition probabilities.

Then rate tuples $(R_0, R_1 \cdots R_{L-1})$ satisfying following inequalities are achievable,

$$\begin{aligned} R_{L-1} &\leq I(U_{L-1}; Y_{L-1}) \\ R_k &\leq I(U_k; Y_k|U_{k+1}) \text{ for } k \in [1 : L-2] \\ R_0 &\leq I(X; Y_0|U_1) \end{aligned}$$

Let $\mathcal{R}_{\bar{P}}$ denote this rate region for a particular choice of $\bar{P}$ satisfying (6.22) and (6.23). The achievable region is given by

$$\mathcal{R}_{\text{Markov}} \equiv conv\left(\bigcup_{\bar{P} \in \text{Markov}} \mathcal{R}_{\bar{P}}\right), \tag{6.24}$$

where the union is taken over all joint distributions $\bar{P}$ with a Markov structure in (6.22,6.23) and $conv(\cdot)$ denotes the convex hull operation. Achievability of the convex hull follows by standard time-sharing arguments.

One may wonder why is it necessary that the auxiliary variables should have a Markov structure. The code designer could have chosen any joint distribution $\bar{P}_{U_{L-1}U_{L-2}\cdots U_1 X}$ which need not have a Markov structure. Using similar random coding construction as [56], one can show that rate-tuples obeying following equation are achievable with superposition.

$$\begin{aligned} R_{L-1} &\leq I(U_{L-1}; Y_{L-1}) \\ R_k &\leq I(U_k; Y_k|U_{k+1}U_{k+2}\cdots U_{L-1}) \ , \ k \in [1 : L-2] \\ R_0 &\leq I(X; Y_0|U_1 U_2 \cdots U_{L-1}) \end{aligned}$$

We need to slightly modify the coding scheme in [56] for achieving the rates above. Now the distribution of the U_k codebook for message M_k (for user Y_k) depends on all previous auxiliary codewords (in $U_{k+1}, U_{k+2}\cdots U_{L-1}$) corresponding to messages $M_{k+1}, M_{k+2}\cdots, M_{L-1}$ (for users $Y_{k+1}, Y_{k+2}\cdots Y_{L-1}$, respectively). In [56], this only depended on auxiliary U_{k+1} codeword corresponding to message M_{k+1} for user Y_{k+1}. Similarly, now the distribution of X codeword for user Y_0 depends on all auxiliary

codewords instead of just the auxiliary U_1 codeword corresponding to message M_1 for user Y_1.

Let the achievable region above be denoted by $\mathcal{R}^*_{\bar{P}}$, where $\bar{P}$ could be any joint distribution of auxiliary variables and the input variable. Since mutual information can decrease or increase by conditioning, it is unclear whether the convex hull $conv\left(\bigcup_{\bar{P}} \mathcal{R}^*_{\bar{P}}\right)$ over all joint distributions $\bar{P}$ remains unchanged when $\bar{P}$ is restricted to have a Markov structure. Next subsection clarifies this and proves the optimality of Markov structure.

Converse for Classical Degraded Broadcast Channel

Now let us prove optimality of this achievable region on similar lines of Gallager's proof for the two-user case[10] [48]. This proof is also helpful when proving the converse for general degradation graphs.

First, we define the following function over all non-negative (element-wise) vectors $\lambda \equiv (\lambda_1, \lambda_2, \cdots, \lambda_{L-1}) \geq 0$,

$$C_{\text{degraded}}(\lambda) = \sup_{\bar{P} \in \text{Markov}} \left(I(X; Y_0|U_1) + \sum_{k=1}^{L-1} \lambda_k I(U_k; Y_k|U_{k+1}) \right) \tag{6.25}$$

where the last term in the summation, $I(U_{L-1}; Y_{L-1}|U_L)$, is a shorthand for $I(U_{L-1}; Y_{L-1})$. The supremum is over all Markov chains $U_{L-1} - U_{L-2} \cdots U_1 - X$. Similar to [48], it can be shown that the supremum above is unchanged even if we restrict the cardinality of each U_i to that of X.

We will show that an achievable rate-tuple must satisfy

$$R(\lambda) \equiv R_0 + \sum_{k=1}^{L-1} \lambda_k R_k \leq C_{\text{degraded}}(\lambda), \quad \forall\ \lambda \geq 0 \tag{6.26}$$

More specifically, we show that if a rate-tuple disobeys the bound above for any $\lambda \geq 0$, then vanishing error probability cannot be achieved for all users. By con-

[10] It should be mentioned that we could not extend the converse proof in [53] for more than two users.

vex programming (theory of Lagrange multipliers), this is the same as showing that any point outside the achievable region $conv\left(\bigcup_{\bar{P}\in \text{Markov}} \mathcal{R}_{\bar{P}}\right)$ is not achievable, which proves the optimality of the achievable region in [56]. The proof mainly follows from the following lemma.

Lemma 39 *Let $Y_k^{1:n}$ denote a shorthand for all outputs at user Y_k from time 1 to n.*

$$\begin{aligned} nC(\lambda) &\geq I(M_0; Y_0^{1:n}|M_1 M_2 \cdots M_{L-1}) && (6.27)\\ &+ \sum_{k=1}^{L-2} \lambda_k I(M_k; Y_k^{1:n}|M_{k+1} M_{k+2} \cdots M_{L-1}) && (6.28)\\ &+ I(M_{L-1}; Y_{L-1}^{1:n}) && (6.29) \end{aligned}$$

Choice of auxiliaries: The proof of the lemma is in Appendix D but its main component is our substitution of auxiliary random variable U_k at time t. It is defined as the set of

1. Past symbols (up to time $t-1$) observed at user Y_k and
2. the messages for layer k and all further layers, i.e., $M_k M_{k+1} \cdots M_{L-1}$. These are all the messages that user Y_k can decode.

This method for choice of auxiliaries is also useful for converses of general degradation graphs. Now we state the precise statement of the converse.

Theorem 40 *For some $\lambda \geq 0$, $\epsilon > 0$, if a rate-tuple satisfies*

$$R(\lambda) \geq \epsilon + C(\lambda) \tag{6.30}$$

then error probability cannot vanish for all users, because error probabilities $(p_0, p_1 \cdots p_{L-1})$ of the L receivers satisfy

$$(1 + p_0 n R_0) + \sum_{k=1}^{L-1} \lambda_k (1 + p_k n R_k) \geq n\epsilon \tag{6.31}$$

Proof If (6.30) holds, then by Lemma 39,

$$\begin{aligned} nR(\lambda) &\geq n\epsilon + I(M_0; Y_0^{1:n} | M_1 M_2 \cdots M_{L-1}) && (6.32) \\ &+ \sum_{k=1}^{L-2} \lambda_k I(M_k; Y_k^{1:n} | M_{k+1} M_{k+2} \cdots M_{L-1}) && (6.33) \\ &+ I(M_{L-1}; Y_{L-1}^{1:n}) && (6.34) \end{aligned}$$

But each mutual information above can be bounded as:

$$\begin{aligned} I(M_k; Y_k^{1:n} | M_{k+1} M_{k+2} \cdots M_{L-1}) &\geq H(M_k | M_{k+1} M_{k+2} \cdots M_{L-1}) - H(M_k | Y_k^{1:n}) \\ &= nR_k - H(M_k | Y_k^{1:n}) \\ &\geq nR_k - (1 + p_k n R_k) \end{aligned}$$

where the first inequality follows since conditioning reduces entropy, the second equality is due to independence of messages and the last step is due to Fano's inequality. Rearranging this and substituting back (6.32) yields (6.31). •

Case study: Binary Erasure Channels

Consider this simple achievable scheme for L binary erasure channels with erasure probabilities $e_1, e_2 \cdots e_L$. The output block length n is divided into L separate blocks, where block k has length $n\alpha_k$. Receiver k's message of nR_k information bits is converted to $n\alpha_k$ coded bits. This can be done with a linear operation $r = Ab$, where b denotes a vector of the nR_k information bits, r is the vector of $n\alpha_k$ coded bits and A is a $n\alpha_k \times nR_k$ generator matrix.

We can choose each entry of A independently with uniform binary distribution. As block length n grows large, essentially any nR_k rows of this matrix will be linearly independent with high probability (i.e., probability tending to 1 with large n). Thus with high probability, receiving any nR_k elements of r is sufficient to decode b. Now note that essentially $(1 - e_k)n\alpha_k$ coded bits will reach unerased at receiver k. Hence if $\alpha_k(1 - e_k) \geq R_k$, then with high probability, receiver k can decode its message of

nR_k bits. Any receiver with smaller erasure probability will also decode this message. Since sum of α_k's is at most unity,

$$\sum_{k=1}^{L} R_k/(1-e_k) \leq \sum_{k=1}^{L} \alpha_k \leq 1$$

This indeed is the capacity region for this broadcast channel given by superposition coding in Eq. (6.24). The optimal structure of auxiliaries for this situation is a cascade of binary symmetric channels from U_L to X. Thus a simple scheme of dividing the block-length amongst random linear codes achieves the capacity region here.

However, such simple schemes are not optimal in general and superposition coding cannot be avoided.

6.2.2 Achievability and converse for general degradation graphs

Consider an arbitrary degradation graph (such as Fig. 6-6 or Fig. 6-7). Recall that layer k denotes all nodes at distance $k+1$ from the root node of input X. All nodes in layer k of this degradation graph want to decode $M_k, M_{k+1} \cdots M_{L-1}$, where L is the depth of the graph from the root node. The rate of message M_i is denoted by R_i.

We can use similar random coding construction as [56], which is based on superposition. Using standard typicality arguments, we can prove the following achievable rate region. For concreteness and clarity, we will state our the result for the particular degradation graph in Fig. 6-6. Similar rate-region can be written down for any degradation graph.

As the degradation graph in Fig. 6-6 has $L = 3$ layers, we choose $L - 1 = 2$ auxiliary random variables U_1, U_2 such that joint distribution of all auxiliaries, input X and all outputs has the following Markov structure.

Thus the joint distribution of all variables is given by $\bar{P}_{U_2U_1X} W_{Y_0Y_1Y_2Z_1Z_2|X}$, where the second term is determined by the network channel as in (6.20). The first term is chosen by the code designer and satisfies the Markov structure in above figure.

Theorem 41 *For every choice of $\bar{P}_{U_2U_1X}$ consistent with the Markov structure above,*

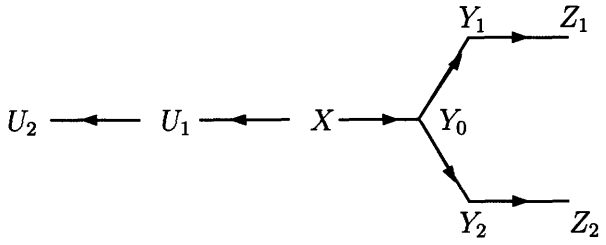


Figure 6-8: Markov structure for achievability

rate-tuples obeying following conditions are achievable:

$$\begin{aligned} R_2 &\leq \min\left(I(U_2;Z_1), I(U_2;Z_2)\right) \\ R_1 &\leq \min\left(I(U_1;W_1|U_2), I(U_1;W_2|U_2)\right) \\ R_0 &\leq I(X;Y|U_1) \end{aligned}$$

Let $\mathcal{R}_{\bar{P}}$ denote this region where $\bar{P}$ is joint distribution of (U_1, U_2, X). Then any rate-tuple in

$$\mathcal{R}_{Markov} \equiv \text{conv}\left(\bigcup_{\bar{P} \in Markov} \mathcal{R}_{\bar{P}}\right)$$

is achievable where $\bar{P}$ denotes all distributions of the Markov structure in Fig. 6-8.

We initially believed that similar to the classical degraded broadcast channel, $\mathcal{R}_{\text{Markov}}$ provides the entire capacity region for a general degradation graph. However, a neat counter-example has been provided recently in [47].

Similar to Subsection 6.2.1, one may wonder why restrict to a Markov chain for auxiliaries. One can indeed choose $\bar{P}_{U_2U_1X}$ which is not a Markov chain $U_2 - U_1 - X$. By similar arguments as in Section 6.2.1, we get the following achievable rate region, called as $\mathcal{R}^*_{\bar{P}}$, for every joint distribution $\bar{P}$ on (U_2, U_1, X).

$$\begin{aligned} R_2 &\leq \min\left(I(U_2;Z_1), I(U_2;Z_2)\right) \\ R_1 &\leq \min\left(I(U_1;W_1|U_2), I(U_1;W_2|U_2)\right) \\ R_0 &\leq I(X;Y|U_1U_2) \end{aligned}$$

That is, mutual information information between U_k and an output in layer k is

conditioned on $U_{k+1}U_{k+2}\cdots U_{L-1}$ instead of just on U_{k+1} as in the case of a Markov chain.

Theorem 42 *An achievable region is given by* $\text{conv}(\bigcup_{\bar{P}} \mathcal{R}^*_{\bar{P}})$, *where the union is over all joint distributions of auxiliaries and input* X.

Mirror Image Converse

This converse uses a similar *choice of auxiliaries* as discussed earlier for the classical broadcast channel. This choice of auxiliaries is described ahead precisely. For a user A in layer[11] $k \geq 1$, its corresponding auxiliary variable $\hat{A}$ at time t (denoted by $\hat{A}_t$) is defined as the set of

1. Past symbols (up to time $t-1$) observed at user A and
2. the messages for layer k and all further layers, i.e., $M_k M_{k+1}\cdots M_{L-1}$. These are all the messages that user Y_k can decode.

The proof of this converse is exactly analogous to the classical broadcast channel and hence omitted. The only additional point to be noted is this choice of auxiliaries satisfies the mirror image structure described ahead.

We consider the degradation graph in Fig. 6-7 for concreteness and converse for a general degradation graph can be expressed similarly. Given the degradation graph, its mirror image is created by placing a mirror between X and layer 0 (user Y_0 here) as shown in Fig. 6-9.

The auxiliary variables $(\hat{Y}_1, \hat{Y}_2, \hat{Z}_1, \hat{Z}_2)$ thus have the same degradation graph as the channel outputs (Y_1, Y_2, Z_1, Z_2). For example, $\hat{Z}_1$ is a degraded version of $\hat{Y}_1$, whereas $\hat{Z}_2$ is a degraded version of both $\hat{Y}_1$ and $\hat{Y}_2$. For a distribution $\bar{P}$ on $(X, \hat{Y}_1, \hat{Y}_2, \hat{Z}_1, \hat{Z}_2)$ satisfying the above mirror structure, define the following function over all non-negative (element-wise) vectors $\lambda \equiv (\lambda_{Y_1}, \lambda_{Y_2}, \lambda_{Z_1}, \lambda_{Z_2}) \geq 0$. This is on similar lines of the function $C_{\text{degraded}}(\cdot)$ for the classical degraded situation seen earlier.

[11] There are no auxiliary variables for users in layer 0 which is at distance 1 from input X in the degradation graph.

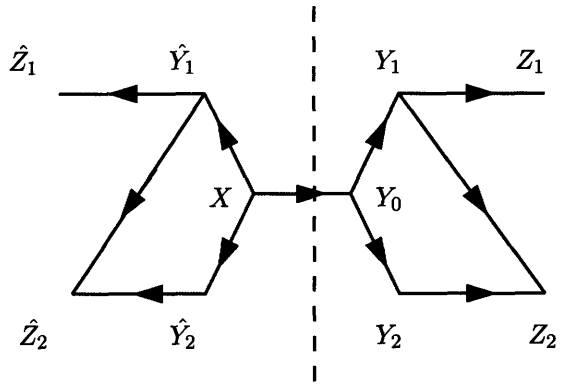


Figure 6-9: Upper bound with mirror image structure of auxiliaries

$$C_{\text{mirror}}(\lambda) = \sup_{\tilde{P}\in\text{mirror}} [I(X;Y_0|\hat{Y}_1\hat{Y}_2) + \lambda_{Y_1} I(\hat{Y}_1;Y_1|\hat{Z}_1,\hat{Z}_2) + \lambda_{Y_2} I(\hat{Y}_2;Y_2|\hat{Z}_2) + \lambda_{Z_1} I(\hat{Z}_1;Z_1) + \lambda_{Z_2} I(\hat{Z}_2;Z_2)]$$

The supremum is over all distributions of auxiliary variables and the input variable which satisfy the mirror image structure[12].

The general rule for defining $C_{\text{mirror}}(\lambda)$ is for each user A, there is a corresponding λ_A which is multiplied by the mutual information between user A's output and its mirror image auxiliary variable conditioned on all the children of the mirror auxiliary variable in the mirror structure[13]. The achievable rate region is upper bounded as follows in terms of $C_{\text{mirror}}(\lambda)$.

Theorem 43 *Every achievable rate-tuple satisfies*

$$R(\lambda) \equiv R_0 + (\lambda_{Y_1} + \lambda_{Y_2})R_1 + (\lambda_{Z_1} + \lambda_{Z_2})R_2 \leq C_{mirror}(\lambda), \quad \forall\ \lambda \geq 0 \qquad (6.35)$$

For defining $R(\lambda)$ in a general situation, each layer's rate is multiplied by the sum of λ coefficients of all users in that layer. Note that converse for the classical degraded broadcast channel in (6.26) is a special case of this converse because mirror image of a straight line is a straight line. This coincidence proved the optimality of the

[12]The cardinality of all auxiliary variables can be bounded within a finite size without loss of optimality (this standard argument can be seen in [4, 53]).

[13]The λ parameter for the best user (closest to X) is normalized to unity without loss of generality.

achievable region [56], where a Markov chain of auxiliaries was used for superposition coding.

Another Converse: Shifted Mirror Image

To construct the shifted mirror image, take the degradation graph and take its mirror image by placing the mirror at X. Now chop off the last layer of auxiliary variables in this mirror image. We demonstrate this in Fig. 6-10 for the degradation graph in Fig. 6-7 and converse for a general degradation graph can be expressed similarly.

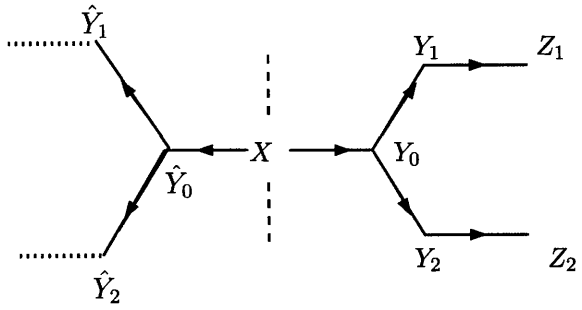


Figure 6-10: Upper bound with shifted mirror image structure of auxiliaries

The main difference of this structure with the mirror image seen earlier is now we kept the mirror at X instead of putting it between X and layer 0. Moreover, we chopped of the last layer of auxiliaries from the mirror image. This ensures the correct number of layers of auxiliary variables.

The auxiliary variables $(\hat{Y}_1, \hat{Y}_2, \hat{Y}_0)$ are mirror image of (Y_0, Y_1, Y_2). For a distribution $\bar{P}$ on (X, Y_0, Y_1, Y_2) satisfying the above mirror structure, define the following function over all non-negative (element-wise) vectors $\lambda \equiv (\lambda_{Y_1}, \lambda_{Y_2}, \lambda_{Z_2}) \geq 0$.

$$\begin{aligned} C'_{\text{mirror}}(\lambda) \;=\; \sup_{\bar{P} \in \text{mirror}} \; & [I(X; Y_0|\hat{Y}_0) + \lambda_{Y_1} I(\hat{Y}_0; Y_1|\hat{Y}_1) + \lambda_{Y_2} I(\hat{Y}_0; Y_2|\hat{Y}_2) \\ & + \lambda_{Z_1} I(\hat{Y}_1; Z_1) + \lambda_{Z_2} I(\hat{Y}_2; Z_2)] \end{aligned}$$

The supremum is over all distributions of auxiliary variables and the input variable which satisfy the shifted mirror image structure The general rule for defining $C'_{\text{mirror}}(\lambda)$ is for each user A, there is a corresponding λ_A which is multiplied by the mutual information between user A's output and the mirror image $\hat{B}$ of its immediate parent

node B and conditioned on its own mirror image auxiliary variable $\hat{A}$. Consider user Y_1 for example. We take the mutual information between Y_1 and mirror image $\hat{Y_0}$ of Y_1's immediate parent Y_0 and condition it on $\hat{Y_1}$. As before, the λ parameter for the best user (closest to X) is fixed at unity. The achievable rate region is upper bounded as follows in terms of $C'_{\text{mirror}}(\lambda)$.

Theorem 44 *Every achievable rate-tuple satisfies*

$$R(\lambda) \equiv R_0 + (\lambda_{Y_1} + \lambda_{Y_2})R_1 + (\lambda_{Z_1} + \lambda_{Z_2})R_2 \leq C'_{mirror}(\lambda) \quad \forall\ \lambda \geq 0 \qquad (6.36)$$

Proof This converse is based on a slight variation of previous choice of auxiliary variables. This modified choice is more similar to the choice in [48].

Alternate choice of auxiliaries: Consider user A in layer $k \geq 0$ and let B denote its its immediate parent[14]. The auxiliary random variable for user A at time t, denoted by $\hat{B}_t$, is defined as the set of

1. Past symbols (up to time $t-1$) observed at B

2. all the messages that user A can decode.

Thus the only difference with the earlier choice of auxiliaries is that now instead of using A's past symbols , we are using the B's past symbols. This difference changes the earlier auxiliary structure of mirror image to the shifted mirror image. An analog of Lemma 39 is stated next, which converts the multi-letter information theoretic problem to a single letter optimization in $C'_{\text{mirror}}(\lambda)$.

The remaining proof of this converse is exactly analogous to the classical broadcast channel and hence omitted. The only additional point to be noted is this choice of auxiliaries satisfies the shifted mirror image structure described ahead. •

Lemma 45 *As earlier, let $Z_1^{1:n}$ denote a shorthand for all outputs at user Z_1 from time 1 to n. Then we have,*

[14]If a user has multiple parents, choose any one of them and ignore others.

$$
\begin{aligned}
nC'_{mirror}(\lambda) &\geq I(M_0; Y_0^{1:n}|M_1M_2) + \lambda_{Y_1} I(M_1; Y_1^{1:n}|M_2) + \lambda_{Y_1} I(M_1; Y_2^{1:n}|M_2) \\
&+ \lambda_{Z_1} I(M_2; Z_1^{1:n}) + \lambda_{Z_2} I(M_2; Z_2^{1:n})
\end{aligned}
$$

Proof The proof is similar to the proof of Lemma 39. We only illustrate how $I(M_1; Y_1^{1:n}|M_2)$ can be bounded using the alternate choice of auxiliaries. The remaining mutual information terms in RHS can be bounded similarly.

$$
\begin{aligned}
I(M_1; Y_1^{1:n}|M_2) &\leq \sum_{t=1}^{n} I(M_1; Y_{1,t}|M_2, Y_1^{1:t-1}) \\
&\leq \sum_{t=1}^{n} \left(H(Y_{1,t}|M_2, Y_1^{1:t-1}) - H(Y_{1,t}|M_1, M_2, Y_1^{1:t-1}) \right) \\
&\leq \sum_{t=1}^{n} \left(H(Y_{1,t}|M_2, Y_1^{1:t-1}) - H(Y_{1,t}|M_1, M_2, Y_1^{1:t-1}, Y_0^{1:t-1}) \right) \\
&= \sum_{t=1}^{n} \left(H(Y_{1,t}|\hat{Y}_{1,t}) - H(Y_{1,t}|\hat{Y}_{1,t}\hat{Y}_{0,t}) \right)
\end{aligned}
$$

because conditioning reduces entropy and $\hat{Y}_{1,t} \equiv (M_2, Y_1^{1:t-1})$ and $\hat{Y}_{0,t} \equiv (M_1, M_2, Y_0^{1:t-1})$ in the alternate choice of auxiliaries. •

Complete Bipartite Networks

Consider a class of graphs is where all outputs in layer k are children of every node in previous layer (see Fig. 6-11 for example). In this case, edges between two adjacent layers form a complete bipartite graph. Network situations such as [52] fall into this category of degradation graphs.

For such networks, we can prove optimality of superposition coding with Markov auxiliaries. This generalizes results in [46], where the degradation graphs has only two layers and one of the two layers has only one user.

We illustrate the result for the network in Fig. 6-11 and the same arguments can be repeated for any complete bipartite network. Since this network has only two

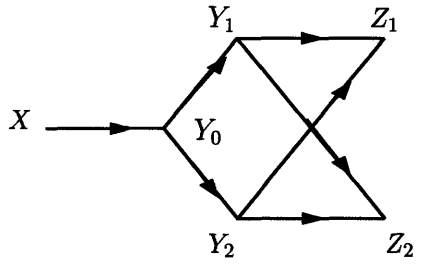


Figure 6-11: A degradation graph for which superposition coding is optimal.

layers, we will need only one auxiliary variable U behind X. First, on similar lines of C_{degraded}, C_{mirror} and C'_{mirror} define the C_{bipart} over all non-negative (element-wise) vectors $\lambda \equiv (\lambda_{Y_2}, \lambda_{Z_1}, \lambda_{Z_2}) \geq 0$.

$$C_{\text{bipart}}(\lambda) = \sup_{\bar{P} \in \text{Markov}} \left(I(X;Y_1|U) + \lambda_{Y_2} I(X;Y_2|U) + \lambda_{Z_1} I(U;Z_1) + \lambda_{Z_1} I(U;Z_2) \right)$$

The supremum is over all distributions of (U, X) satisfying the Markov condition $U - X - (Y_1, Y_2, Y_3, Y_4)$.

Theorem 46 *Every achievable rate-tuple* (R_0, R_1) *satisfies*

$$R(\lambda) \equiv R_0 + \lambda_{Y_2} R_0 + (\lambda_{Z_1} + \lambda_{Z_2}) R_1 \leq C_{bipart}(\lambda), \quad \forall\ \lambda \geq 0 \tag{6.37}$$

Thus capacity region of a complete bipartite network equals $\mathcal{R}_{Markov}$*, which is achievable by superposition coding.*

Proof Since achievability by superposition coding is straightforward, we only discuss the converse argument here. It is based on a simple choice of auxiliaries[15]: $U_t \equiv (M_1, Z_1^{1:t-1}, Z_2^{1:t-1})$. In terms of this auxiliary substitution, will upper bound $I(M_0; Y_1^{1:n}|M_1)$ and $I(M_0; Y_2^{1:n}|M_1)$ using conditioning reduces entropy.

[15]In general, the auxiliary variable for layer $k \geq 1$ is the set of messages decoded by that layer and the past of all the users in that layer.

$$
\begin{aligned}
I(M_1; Z_1^{1:n}) &\leq \sum_{t=1}^{n} I(M_1; Z_{1,t}|Z_1^{1:t-1}) \\
&\leq \sum_{t=1}^{n} H(Z_{1,t}) - H(Z_{1,t}|M_1, Z_1^{1:t-1}, Z_2^{1:t-1}) \\
&= \sum_{t=1}^{n} I(U_t; Z_{1,t}) \qquad \text{(definition of } U_t\text{)} \\
\text{Similarly, } I(M_1; Z_1^{1:n}) &\leq \sum_{t=1}^{n} I(U_t; Z_{2,t})
\end{aligned}
$$

Now let us upper bound $I(M_1; Z_1^{1:n})$ and $I(M_1; Z_2^{1:n})$ using physical degradedness.

$$
\begin{aligned}
I(M_0; Y_1^{1:n}|M_1) &\leq \sum_{t=1}^{n} I(M_0; Y_{1,t}|M_1, Y_1^{1:t-1}) \\
&= \sum_{t=1}^{n} H(Y_{1,t}|M_1, Y_1^{1:t-1}, Z_1^{1:t-1}, Z_2^{1:t-1}) - H(Y_{1,t}|M_0, M_1, Y_1^{1:t-1}) \\
&\leq \sum_{t=1}^{n} H(Y_{1,t}|M_1, Z_1^{1:t-1}, Z_2^{1:t-1}) - H(Y_{1,t}|M_0, M_1, Y_1^{1:t-1}, Y_2^{1:t-1}, X_t) \\
&= \sum_{t=1}^{n} H(Y_{1,t}|M_1, Z_1^{1:t-1}, Z_2^{1:t-1}) - H(Y_{1,t}|X_t) \qquad \text{(memorylessness)} \\
&= \sum_{t=1}^{n} I(X_t; Y_{1,t}|U_t) \qquad \text{(definition of } U_t\text{)}
\end{aligned}
$$

Similarly, one can prove $I(M_0; Y_2^{1:n}|M_1) \leq \sum_{t=1}^{n} I(X_t; Y_{2,t}|U_t)$. These bounds imply the following lemma.

Lemma 47

$$
nC_{bipart}(\lambda) \geq I(M_0; Y_1^{1:n}|M_1) + \lambda_{Y_2} I(M_0; Y_2^{1:n}|M_1) + \lambda_{Z_1} I(M_1; Z_1^{1:n}) + \lambda_{Z_2} I(M_1; Z_2^{1:n})
$$

The remaining proof follows as in earlier converses using Fano's inequality. •

The above theorem provides the entire capacity region in terms of an optimization over auxiliary variables. For discrete memoryless networks, these optimizations are

over finite size auxiliary variables. However, for Gaussian networks such as [52], these optimizations could be infinite dimensional and the optimality of Gaussian inputs (as in [57]) is not clear. For analyzing whether or not Gaussian superposition coding is optimal, techniques from [58] could be useful.

Case Study: Packet Broadcast Networks

As an application of the degradation graph framework, now let us consider a situation where transmitter emits a fixed number L of n-bit packets. Each of these packets can be either perfectly received or be erased completely. A user can receive any of the 2^L-1 non-empty subsets of these L packets. We want to ensure that any user who gets k number of packets can decode the k messages $(M_{L-k}, M_{L-k+1}, \cdots M_{L-1})$. Earlier problems of multilevel diversity coding [50, 49, 51] and priority encoded transmission [17] can be modeled this way. Degradation graphs provide a common framework to address them. In addition, this framework can be also used to model errors in packets in addition to erasures.

Let us assume $L = 3$ packets are transmitted for simplicity, although similar analysis can be performed for any L. The degradation graph in this situation is shown below. The actual transmission is denoted as (X_1, X_2, X_3) where each X_i represents a bit from packet i. Various receivers receive all seven subsets of this transmission. For better clarity, X_{123} is used to denote $(X_1X_2X_3)$ for example.

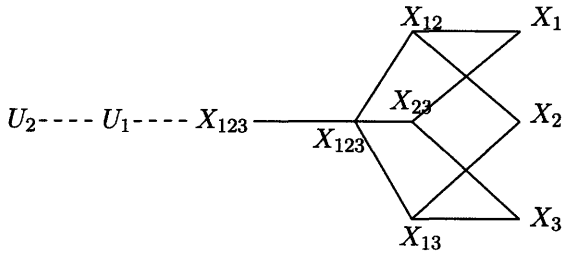


Figure 6-12: Solid lines show the degradation graph for packet erasure network. Dotted lines show the auxiliaries U_1, U_2 for superposition coding in Theorem 41.

Degradation graphs can be also used to model asymmetric situations, where only certain subsets of transmitted packets can be received. For example, there may not

be any user who gets packets 2 and 3. This can be modeled by simply removing X_{23} from the degradation graph. These asymmetric situations seem particularly relevant for distributed storage [51]. In a distributed storage system with L storage locations, each user may have access to a certain subset of these locations. In fact, now the effect of errors in stored data can also be analyzed using degradation graphs.

Now let us calculate the achievable region for Fig. 6-12 from Theorem 41. We show that this achievable rate-region is given by $R_2 + R_1/2 + R_0/3 \leq 1$, where R_k is the rate of message M_k for layer k. From Theorem 41,

$$R_2 \leq I(U_2; X_1),\ R_2 \leq I(U_2; X_2) \ \ \& \ R_2 \leq I(U_2; X_3) \tag{6.38}$$

$$R_1 \leq I(U_1; X_{12}|U_2) \tag{6.39}$$

$$R_1 \leq I(U_1; X_{13}|U_2) \tag{6.40}$$

$$R_1 \leq I(U_1; X_{23}|U_2) \tag{6.41}$$

$$R_0 \leq I(X_{123}; X_{123}|U_1) = H(X_{123}|U_1) \tag{6.42}$$

Performing (6.38) $\times$ 2 + (6.39) + (6.40) + (6.41) + 2 $\times$ (6.42) implies $6R_2 + 3R_1 + 2R_0$ is not greater than

$$\left[2\sum_{i=1}^{3} H(X_i)\right] - \left[2\sum_{i=1}^{3} H(X_i|U_2) - \sum_{i\neq j} H(X_{ij}|U_2)\right] + \left[2H(X_{123}|U_1) - \sum_{i\neq j} H(X_{ij}|U_1)\right]$$

First bracket above is at most 6, second bracket equals $\sum_{i\neq j} I(X_i; X_j|U_2)$ and hence non-negative. Third bracket equals $-I(X_3; X_1|X_2) - I(X_2; X_{13})$, which is non-positive. Thus $6R_2 + 3R_1 + 2R_0 \leq 6$, proving our achievable region.

For any L number of packets, similar analysis can be done to prove that achievable region in Theorem 41 equals

$$\sum_{i=0}^{L-1} R_i/(L-i) \leq 1$$

In fact, this achievable region was shown to be the entire capacity region in [49, 17]

which implies optimality of superposition coding for this scenario[16].

6.3 Concluding remarks

6.3.1 Relations of Network Info. Theory to Error Exponents

The capacity region for broadcast with degraded message sets also provides insights for bit-wise UEP exponents of a point-to-point channel. We demonstrate with the example of the network in Fig. 6-1. Let the channel to user Y_k be denoted by $W_{Y_k|X}$. In a point-to-point channel $W_{Y|X}$ with input distribution P_X, the exponent of observing a channel type $W_{Y_k|X}$ equals the conditional divergence $D(W_{Y_k|X}\|W_{Y|X}|P_X)$ due to Sanov's theorem. Define this exponent of channel type being $W_{Y_k|X}$ as E_k. Let $\bar{E}_0$ denote the maximum exponent of observing a channel in layer 0.

$$\bar{E}_0 = \max_{k \in S_0} E_k$$

Similarly, let $\bar{E}_1$ denotes the maximum exponent of all E_k. Let $\mathcal{R}_{\text{capacity}}$ denotes the capacity region of this network for broadcast with degraded messages sets. In the error exponent formulation, no rate-pair outside this capacity region can be achieved if the special bits (which are demanded by all users) require at least an exponent E_1 and the ordinary bits (which are only demanded by users in layer 0) require at least an exponent of E_0. Otherwise, the capacity region for the broadcast problem will be violated.

Thus capacity region for the network formulation provides an easy upper bound for the achievable rates in the error exponent formulation. In fact, we already used such an argument in the previous chapter for the case of symmetric point-to-point channels. There we only needed the capacity region for the physically degraded broadcast channel. However, for general point-to-point channels, using the capacity region for broadcast with degraded message sets can be more powerful.

[16]In fact, as hinted by the linear nature of this region, much simple strategies based on time-sharing between MDS codes are optimal.

6.3.2 Message-wise UEP in networks

In the error exponent formulation in Chapter 3, we saw how the special messages can be protected optimally without sacrificing the overall data-rate from capacity. In particular, we saw that e^{nr} special messages can achieve the best exponent $E(r)$ at rate r as if the $\doteq e^{nC}$ ordinary messages were absent. There was no tradeoff for simultaneously transmitting the special messages and the ordinary messages. We will now see an analog of this result in the network formulation.

Although this discussion can be extended to more general degradation graphs, we consider the following simple two-user network for simplicity.

$$X - Y_0 - Y_1$$

Let C_i denote the capacity of the channel to user Y_i.

$$C_i = \max_{P_X} I(X; Y_i)$$

Adhering to the notation, let $\mathcal{M}$ denote the set of ordinary messages and let $\mathcal{M}_s \subset \mathcal{M}$ denote the set of special messages. If a special message from $\mathcal{M}_s$ is sent, both users should decode it correctly with high probability (tending to 1). In contrast, when an ordinary message is sent, only the better user Y_0 should decode it correctly. Assuming $|\mathcal{M}_s| \doteq e^{nr}$ special messages and $|\mathcal{M}| \doteq e^{nR}$, we want to calculate the achievable rate pairs (r, R).

This simple strategy shows how (C_1, C_0) is an achievable (r, R). Thus there is no tradeoff between the number os special messages and the number of ordinary messages, the best r and R can be achieved simultaneously.

Optimal strategy: The first $\sqrt{n}$ symbols of the code are reserved for indicating whether the message to be transmitted is special or ordinary. Using a simple repetition code, this single bit of information can be transmitted reliably in this first phase to both the users. When a special message is to be sent, send it in the remaining $n - \sqrt{n}$ symbols using a capacity achieving code for user Y_1. User Y_0 can also decode this

special message with high probability, since Y_1 is a degraded version of Y_0. When an ordinary message is to be sent, send it using in these $n-\sqrt{n}$ symbols using a capacity achieving code for user Y_0.

Thus we can send as many special messages to both users as if the ordinary messages are absent and as many ordinary message to Y_0 as if the special messages were absent. Note what when an ordinary message is sent, although Y_1 cannot decode which particular ordinary message it is, it correctly knows (with high high probability) that an ordinary message is being sent.

The same discussion of message-wise UEP can be extended for general networks with multiple speciality levels (say $L \geq 2$). This corresponds to L nested subsets of users $S_0 \subset S_1 \cdots S_{L-2} \subset S_{L-1}$ where S_{L-1} is the set of all users[17]. For this L layer situation, the messages are divided into L sets: $\mathcal{M}_0 \supset \mathcal{M}_1 \supset \cdots \supset \mathcal{M}_{L-1}$. Here messages in $\mathcal{M}_{L-1}$ are most special (need to be decoded correctly by everyone and messages in $\mathcal{M}_0$ are least special. In general, when a message in $\mathcal{M}_k$ is sent, all users in set S_k should be able to decode it. This message-wise UEP problem can be called as broadcast with *nested message sets* instead of broadcast with degraded message sets, which corresponds to bit-wise UEP.

For this general problem, the optimal strategy again uses two phases as before. In the first $\sqrt{n}$ symbols, the speciality class, $0 \leq k \leq L-1$, of the chosen message is conveyed reliably to everyone using some simple code of L codewords. In the remaining $n-\sqrt{n}$ symbols, the particular message in $\mathcal{M}_k$ is sent to all the users in S_k. Rate of each message set $\mathcal{M}_k$ can approach the capacity of a compound channel [4] composed of all users in S_k.

$$C_k = \max_{P_X} \min_{i \in S_k} I(X; Y_i)$$

This rate for message set $\mathcal{M}_k$ can be achieved with a capacity achieving code of this compound channel. For every $0 \leq k \leq L-1$, this strategy can simultaneously achieve rate C_k for message set S_k. Thus each message set achieves a rate as if all

[17]The users need not be related by any degradation graph in this general setup.

other message-sets are absent.

It is interesting to note how the notion of compound channel capacity becomes relevant for message-wise UEP in networks. In contrast for bit-wise UEP, the capacity region for broadcast with degraded message sets was relevant. Even more interesting is how message-wise UEP over general networks is completely solved due to simplicity of this compound channel capacity, whereas bit-wise UEP over general networks is wide open since capacity for broadcast with degraded message sets is very difficult.

Chapter 7

Summary and Future Directions

7.1 Summary

We saw that when available resources (such as delay, bandwidth, and power) are finite, information should be viewed as heterogeneous entity instead homogeneous. Homogeneous bits no longer suffice as the universal interface of information in this case. First, all bits need not be equally important; second, some messages could be more important instead of bits. This heterogeneous nature information can be leveraged for protecting these crucial parts (special bits or special messages) of information much better than its ordinary parts.

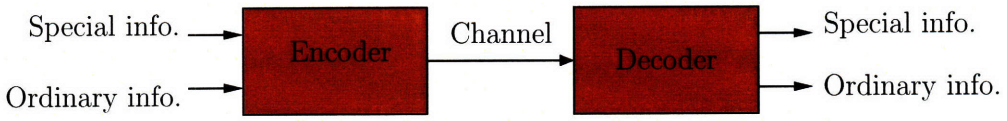


Figure 7-1: Architecture for Heterogeneous Information: The encoder jointly encodes the special and ordinary information on the same channel resource. This achieves better tradeoffs in general than sending the special and ordinary parts separately.

We also saw how error exponents can be used as a fundamental benchmark for understanding fundamental limits of unequal error protection. They allow us develop a general understanding of UEP which is not specific to the channel model or the coding strategy being used. This understanding provides engineering guidelines for

practical UEP designs—somewhat similar to how the notion of channel capacity provides guidelines and benchmarks for practical code designs. Here are some lessons we have learned,

- Exponential reliability can be achieved for some special parts of information without sacrificing capacity.
- Special messages are easier to protect than special bits in the no feedback case. In many current communication protocols, a special message is protected using a flag bit[1] for that message. This essentially converts a special message situation to a special bit situation. This artificial conversion is very inefficient for protecting the special message. Much better protection can be achieved for the special message if it is communicated as a special message, e.g., with a special (repetition) codeword.
- Feedback connects the bit-wise and message-wise notions of UEP in some fundamental ways. With feedback, schemes for one notion can be used for the other.
- Red-Alert Exponent is a fundamental channel parameter for transmission of heterogeneous information.
- Often optimal strategies for UEP require only some simple modifications to conventional coding mechanisms. This implies we can leverage the development of conventional code designs for UEP mechanisms too.

Fundamental limits of UEP were also analyzed in terms of some network information theory problems (e.g., broadcast with degraded message sets and compound channel capacity). We developed some new formulations for such problems using ideas from Euclidean geometry and graphical models. These formulations allowed us to simplify these difficult problems and also enabled us to think more systematically about network situations. This analysis gave rise to new canonical examples which provided fresh insights and also generalized some previous results.

[1]For example, the NACK message in network protocols.

7.2 Future Directions

Besides extending our results for discrete memoryless channels to other channel models such as Gaussian channels, this thesis leads to many interesting problems in areas ranging from coding theory to joint source-channel coding to network optimization.

7.2.1 Rates below capacity

Recall that only a few UEP scenarios were addressed in Chapters 5 and 6 for data-rates below capacity. Fundamental limits in many more UEP scenarios below capacity remain to be understood. Recall that even for data-rates at capacity, we had many UEP scenarios to study. Since even ordinary information can achieve a positive exponent for rates below capacity, the possible set of UEP problems becomes even richer now. We can trade off the exponent of ordinary information in favor of special information in a many ways.

7.2.2 Efficient coding

Designing practical codes which achieve our fundamental UEP limits gives rise to many new problems in coding theory. For low computational complexity, one can use the iterative coding approach as well as the algebraic coding approach.

In terms of fundamental limits too, almost every question for UEP error exponents has an analogue in coding theory where we are interested in Hamming distances instead of error exponents. These formulations demand better deterministic guarantees for special information (in terms of normalized Hamming distances) instead of better probabilistic guarantees (in terms of error exponents). For example, some such limits were obtained in [24, 25] for bit-wise UEP.

For example, consider the following analogue for sending a special bit at channel capacity. Given a code that achieves the Singleton bound, color half the codewords blue and the remaining half as red. The cluster of blue codewords corresponds to the special bit being 0 and that of red codewords corresponds to the special bit being 1. Is it possible to have a larger distance between the blue cluster and red cluster

compared to the minimum distance of the code given by the Singleton bound? In other words, we are again asking the feasibility of Fig. 7-2 seen earlier. Now however, the small green balls are Hamming balls (of the Singleton radius) around codewords instead of typical decoding balls.

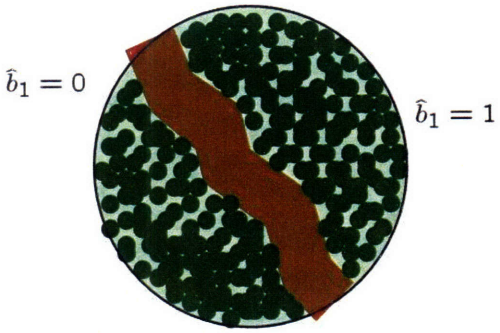


Figure 7-2: Splitting the output space into 2 distant enough clusters.

One can also ask a deterministic version of the boulders-and-sand result for special messages. In fact, an analogue of the error exponent result holds true and we have been able to prove that multiple points on the Gilbert-Varshamov curve can be simultaneously achieved. Thus more important messages can have larger Hamming balls around their codewords without affecting the size or number (in exponential scaling) of smaller Hamming balls.

Understanding the implications of allowing lists or erasures at the decoder is also an open area. We briefly analyzed some implications of allowing erasures in Chapter 3 but many questions remain unanswered. For homogeneous information, such questions were analyzed in [8] in terms of error exponents. For heterogeneous information, similar questions could be formulated again for various UEP situations. We can also analyze these issues in terms of Hamming distances.

7.2.3 Joint Source-Channel Coding and Data Compression

Throughout this thesis, we focused on the channel coding component of communication. However, often the final objective is to communicate a source within some distortion constraints. For heterogeneous information, some parts of the source may demand a smaller distortion than other parts. For example, images with faces are

more important than other images and may require better recovery across the noisy channel[2]. Understanding optimal methods for communicating such sources over noisy channels present many novel joint-source channel coding problems. It will be interesting if there are any separation theorems for UEP at the source level and UEP at the channel level.

Even for pure source coding, finding efficient compression methods for better recovery of important parts presents many new problems. Previous approaches to this problem are based on multiple-description coding [59] and successive refinement coding [60]. However, many other formulations yet remain to be analyzed. Another scintillating problem is finding out a source coding analogue for the notion of message-wise UEP in channel coding.

7.2.4 UEP in Networks

The notion of heterogeneous information becomes more exciting and even more prevalent for networks. We saw that a plethora of UEP problems had arisen even for point-to-point channels. For multiuser situations, such as multiaccess and two-way channels, the number of relevant UEP problems becomes much larger. Now each user can demand various protections for various pieces of information and one can study the achievable error exponent for these demands.

More importantly, we can actively use UEP in network protocols. For example, a relay can forward some partial information even if it cannot decode everything. This partial information could be characterized in terms of special bits as well as special messages. Another example is two-way communication, where UEP can be used for more reliable feedback and synchronization.

7.2.5 Coordination + Communication

In many scenarios, the final objective of communication is achieving some coordination between various agents [55]. We consider using the channel for a dual purpose—communicate data as well as achieve coordination. For concreteness, consider broad-

[2]Thanks to Professor Greg Wornell for suggesting this motivation.

casting to two mobile robots where usually we send normal data to them. However, once in a while a coordination instruction is to be sent to both robots to perform some crucial joint task. Assuming this joint task to be a crucial one, we need tp protect the coordination instruction much better. What are the tradeoffs between error exponents of this coordination and rates of the normal data to these users?

One can broadly think of UEP as a way to address a basic semantic aspect of information (its heterogeneity) because classical framework ignores all semantic aspects. These joint coordination-communication problems can be thought as addressing one more semantic aspect of information.

7.2.6 Network Optimization

Information theoretic understanding of UEP also gives rise to some network optimization problems. Essentially, the interface to physical layer is no longer bits as in Fig. 7.2.6. Instead, it is a basket of various levels of error protection as in Fig. 7-4. The achievable channel resources of reliability and rate need to be efficiently divided amongst these levels, which gives rise to many resource allocation problems. Information theory tells what baskets are achievable and optimization theory can tell which baskets should be used for maximizing some overall objective.

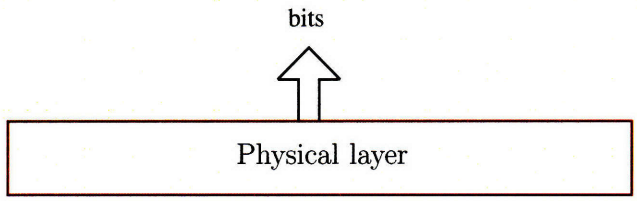


Figure 7-3: Homogeneous interface to physical layer

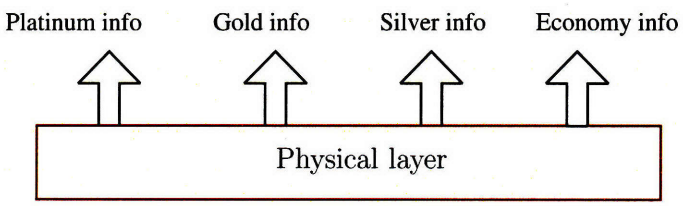


Figure 7-4: Heterogeneous interface to physical layer. Upper layers choose the priority levels for various parts of information.

Appendix A

Proof of Theorem 2 in Chapter 2

The optimum output type is given by Q_Y^* and the error dominating joint type is $Q_Y^* P_{X|Y}^{(t^*)}$. Using (2.35), the error exponent is given by

$$\begin{aligned} E_r(R) &= D(Q_Y^* \| P_Y) + E(R, Q_Y^*) \\ &= D(Q_Y^* \| P_Y) + D(Q_Y^* P_{X|Y}^{(t^*)} \| Q_Y^* P_{X|Y}) \\ &+ \left[D(Q_Y^* P_{X|Y}^{(t^*)} \| Q_Y^* P_X) - R \right]^+ \\ &= D(Q_Y^* P_{X|Y}^{(t^*)} \| P_{XY}) + \left[D(Q_Y^* P_{X|Y}^{(t^*)} \| Q_Y^* P_X) - R \right]^+ \end{aligned}$$

This is a non-decreasing function in $D(Q_Y^* P_{X|Y}^{(t^*)} \| P_{XY})$ and $D(Q_Y^* P_{X|Y}^{(t^*)} \| Q_Y^* P_X)$. Similar to previous subsection, we can show that a ML decoder is equivalent to a LLR decoder for $p_1 = P_{XY}$ to $p_0 = Q_Y^* P_X$. Now assume to the contrary that the dominating joint type $Q_Y^* P_{X|Y}^{(t^*)}$ does not lie on the exponential family joining p_1 and p_0. Now move $Q_Y^* P_{X|Y}^{(t^*)}$ to $Q_Y' P_{X|Y}^{(t')}$ which is on the exponential family and has the same expected LLR as $Q_Y^* P_{X|Y}^{(t^*)}$. Thus by I-projection theorem

$$\begin{aligned} & D(Q_Y' P_{X|Y}^{(t')} \| P_{XY}) < D(Q_Y^* P_{X|Y}^{(t^*)} \| P_{XY}) \\ \text{and} \quad & D(Q_Y' P_{X|Y}^{(t')} \| Q_Y^* P_X) < D(Q_Y^* P_{X|Y}^{(t^*)} \| Q_Y^* P_X) \end{aligned}$$

Moreover,

$$\begin{aligned} D(Q_Y' P_{X|Y}^{(t')} \| Q_Y^* P_X) &= D(Q_Y' P_{X|Y}^{(t')} \| Q_Y' P_X) + D(Q_Y' \| Q_Y^*) \\ &\geq D(Q_Y' P_{X|Y}^{(t')} \| Q_Y' P_X) \\ \Rightarrow \quad D(Q_Y' P_{X|Y}^{(t')} \| Q_Y' P_X) &< D(Q_Y^* P_{X|Y}^{(t^*)} \| Q_Y^* P_X) \end{aligned}$$

Thus replacing $Q_Y^* P_{X|Y}^{(t^*)}$ by $Q_Y' P_{X|Y}^{(t')}$ gives a smaller exponent, which contradicts the optimality of $Q_Y^* P_{X|Y}^{(t^*)}$. •

Appendix B

Equivalent definitions of UEP exponents

We could define all the UEP exponents without using the notion of capacity-achieving sequences. As an example, we will define the single-bit exponent in this alternate manner. This alternative first defines $\bar{E}_{\mathrm{b}}(R)$ as the best exponent for the special bit at a given data-rate R, and then minimizes $\bar{E}_{\mathrm{b}}(R)$ over all $R < C$ to obtain $\bar{E}_{\mathrm{b}}$.

Definition 48 *For a reliable code sequence $\mathcal{Q}$ of rate $R_{\mathcal{Q}}$, with message sets $\mathcal{M}^{(n)} = \mathcal{M}_1 \times \mathcal{M}_2^{(n)}$ where $\mathcal{M}_1 = \{0,1\}$, the exponent for the special bit error probability $\Pr^{(n)}\left[\hat{M}_1 \neq M_1\right]$ equals*

$$E_{b,\mathcal{Q}} = \liminf_{n\to\infty} \frac{-\log \Pr^{(n)}\left[\hat{M}_1 \neq M_1\right]}{n}. \tag{1}$$

Then define $\bar{E}_b(R) = \sup_{\mathcal{Q}:R_{\mathcal{Q}}\geq R} E_{b,\mathcal{Q}}$. Now the single bit exponent $\bar{E}_b$ is defined as

$$\bar{E}_b = \inf_{R<C} \bar{E}_b(R)$$

This definition of says that no matter how close the rate is to capacity, the special bit can achieve the exponent $\bar{E}_{\mathrm{b}}$. We now show briefly why this definition is equivalent to the earlier definition in terms of capacity-achieving sequences.

Lemma 49 $\bar{E}_b = E_b$

Proof of $E_{\mathbf{b}} \leq \bar{E}_{\mathbf{b}}$: By definition of E_{b}, for any given $\delta > 0$, there exists a capacity-achieving sequence $\mathcal{Q}$ whose single bit exponent $E_{\mathrm{b},\mathcal{Q}}$ satisfies,

$$E_{\mathrm{b},\mathcal{Q}} \geq E_{\mathrm{b}} - \delta.$$

We will use this capacity-achieving sequence $\mathcal{Q}$ to prove $\bar{E}_{\mathrm{b}} \geq E_{\mathrm{b},\mathcal{Q}} \geq E_{\mathrm{b}} - \delta$. This

is because rate of $\mathcal{Q}$ equals C by definition. Hence definition of $\bar{E}_{\mathrm{b}}(R)$ implies

$$\begin{aligned} E_{\mathrm{b},\mathcal{Q}} &\leq \bar{E}_{\mathrm{b}}(R) \text{ for any } R < C \\ (E_{\mathrm{b}} - \delta \leq) \quad E_{\mathrm{b},\mathcal{Q}} &\leq \bar{E}_{\mathrm{b}}. \end{aligned}$$

The proof follows by choosing arbitrarily small δ.

Proof of $\bar{E}_{\mathbf{b}} \geq E_{\mathbf{b}}$: First fix an arbitrarily small $\delta > 0$. In the table below, row k represents a reliable code-sequence $\bar{\mathcal{Q}}_k$ at rate $C - 1/k$, whose single-bit exponent

$$E_{\mathrm{b},\bar{\mathcal{Q}}_k} \geq \bar{E}_{\mathrm{b}}(R) - \delta$$

Let $\bar{\mathcal{Q}}_k(l)$ represent length-l code in this sequence. We construct a capacity-achieving sequence $\mathcal{Q}$ from this table as follows. This construction sequentially chooses elements from rows $1, 2, \cdots$.

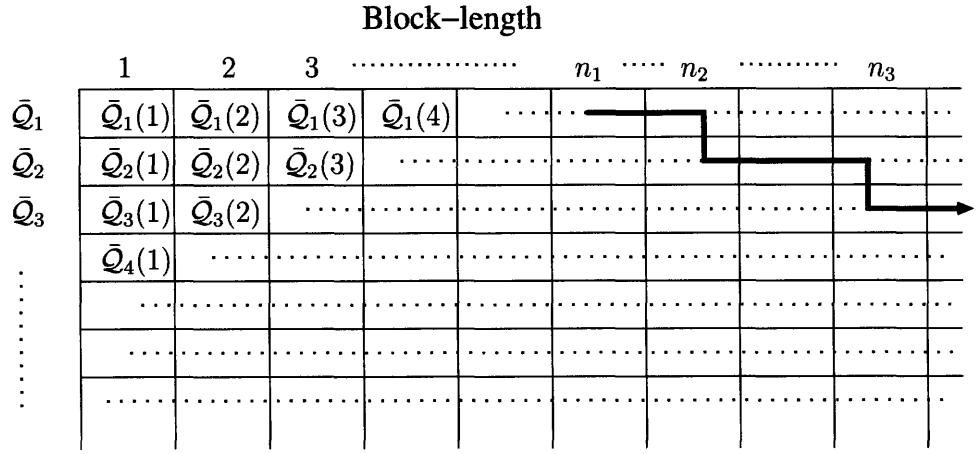


Figure: Row k denotes a reliable code sequence at rate $C - 1/k$. Bold path shows capacity-achieving sequence $\mathcal{Q}$.

Initialize: For sequence $\bar{\mathcal{Q}}_1$, let n_1 denote the smallest block length n at which the single bit error probability satisfies

$$\frac{-\log \Pr^{(n)}\left[\hat{M}_1 \neq M_1\right]}{n} \geq \bar{E}_{\mathrm{b}}(R) - 2\delta \quad \Leftrightarrow \quad \Pr^{(n)}\left[\hat{M}_1 \neq M_1\right] \leq \exp(-n(\bar{E}_{\mathrm{b}}(R) - 2\delta))$$

Iterate: For sequence $\bar{\mathcal{Q}}_{i+1}$, choose the smallest $n_{i+1} \geq n_i$ which satisfies above equation.

Given the sequence, $n_1, n_2, \cdots$, from each row i, we will choose codes of length n_i to $n_{i+1} - 1$, i.e.,

$$\left(\bar{\mathcal{Q}}_i(n_i), \bar{\mathcal{Q}}_i(n_i + 1) \cdots, \bar{\mathcal{Q}}_1(n_{i+1} - 1)\right)$$

as members of the capacity-achieving sequence $\mathcal{Q}$. Thus $\mathcal{Q}$ is a sampling of the code-table as shown by the bold path in this figure. Note that this choice of $\mathcal{Q}$ is a capacity-achieving sequence, moreover it will also achieve a single bit exponent

$$E_{\mathrm{b},\mathcal{Q}} = \inf_{R<C}\{\bar{E}_{\mathrm{b}}(R) - 2\delta\} = \bar{E}_{\mathrm{b}} - 2\delta$$

Choosing arbitrarily small δ proves $E_{\mathrm{b}} \geq \bar{E}_{\mathrm{b}}$. •

Appendix C

Proof of Theorem 10 for BSC

We will focus on a BSC with crossover probability p in this appendix. Before going further, we state the following lemma for binary hypothesis testing (see [5] for example). Consider binary random sequence Y^n of length n. Under hypothesis $H = 0$, it is i.i.d. over time with distribution Bern(p). Under hypothesis $H = 1$, it is i.i.d. over time with distribution Bern($\frac{1}{2}$). Here Bern(p) denotes the Bernoulli distribution with parameter p.

Lemma 50 *Let E_1 denote the exponent for missed-detection probability $\Pr(\hat{H} = 0|H = 1)$ and E_0 denote the exponent for false-alarm probability $\Pr(\hat{H} = 1|H = 0)$. The following implicit equation provides the optimal trade-off between these two exponents, where $D_b(h\|g)$ denotes the KL divergence between two Bernoulli distributions with parameters h and g.*

$$\textit{For some } \delta \leq 1/2, \textit{ let } E_1 = D_b(\delta\|\frac{1}{2}) \quad \Rightarrow E_0 \leq D_b(\delta\|p) \tag{2}$$

Moreover, this exponent pair is achievable by a threshold test on the Hamming weight of Y^n, which chooses $\hat{H} = 1$ if the Hamming weight of Y^n exceeds $n\delta$ and vice versa.

Note that if $E_1 = r$, then E_0 denotes the sphere-packing exponent at rate r and δ denotes the Gilbert-Varshamov distance for rate r. To emphasize the dependence on r, we will denote this Gilbert-Varshamov distance by $\delta_{\mathrm{GV}}(r)$. Now we are ready to prove the theorem.

Special codewords: At any given block length n, we start with a optimum code-book (say $\mathcal{C}_{special}$) for $\lceil e^{nr} \rceil$ messages. Such optimum code-book achieves error exponent $E(r)$ for every message in it.

$$\Pr\left[\hat{M} \neq i|M = i\right] \doteq \exp(-nE(r)) \qquad \forall i \in \mathcal{M}_s \equiv \{1, 2, \cdots, \lceil e^{nr} \rceil\}$$

This code-book is used for transmitting the special messages. At the decoder, let $\mathcal{B}_i$ denote the set of output sequences within Hamming distance $\Phi = n(\delta_{\mathrm{GV}}(r+\epsilon_n)$ from the i'th codeword $\bar{x}^n(i)$. Here ϵ_n is non-negative sequence which vanishes to 0 with increasing n.

$$\mathcal{B}_i = \{y^n : \ |y^n - \bar{x}^n(i)|_H \leq \Phi\}$$

Thus $\mathcal{B}_i$ is a ball of radius Φ around codeword i as shown in Fig. 3-3(a). The radius Φ is essentially the sphere-packing radius. Hence these balls will not be disjoint. Now let $\mathcal{B}$ denote the union of these balls around all special codewords.

$$\mathcal{B} = \bigcup_{i \in \mathcal{M}_s} \mathcal{B}_i$$

If the output sequence Y^n lies in $\mathcal{B}$, the first stage of the decoder decides a special message was transmitted. The second stage then chooses the ML candidate in $\mathcal{M}_s$, i.e., the nearest special codeword from Y^n.

Ordinary codewords: The ordinary codewords will be chosen by random coding: flipping a coin i.i.d. over time. This is the same as Shannon's construction for achieving capacity. The random coding construction provides a simple way to show that in the cavity space $\mathcal{B}^c$ (complement of $\mathcal{B}$), we can essentially fit enough typical noise-balls to achieve capacity. This will avoid the complicated task of carefully choosing the ordinary codewords and their decoding regions in the cavity space.

If the output sequence Y^n lies in $\mathcal{B}^c$, the first stage of the decoder decides an ordinary message was transmitted. The second stage then chooses the ML (nearest) candidate from ordinary codewords.

Error analysis: First, consider the case that a special codeword $\bar{x}^n(i)$ is transmitted. Note that $Y^n \in \mathcal{B}_i$ if and only if $Y^n \oplus \bar{x}^n(i)$ weighs less than Φ. Here $\oplus$ denotes element-wise XOR of two binary sequences. By Stein's lemma, the probability of $Y^n \notin \mathcal{B}_i$ has exponent $D_{\mathrm{b}}(\delta_{\mathrm{GV}}(r+\epsilon_n)\|p)$. It is because channel errors are i.i.d. Bern(p). Since first stage error cannot happen for $Y^n \in \mathcal{B}_i$, first stage error exponent is at least $D_{\mathrm{b}}(\delta_{\mathrm{GV}}(r+\epsilon_n)\|p) = E_{\mathrm{sp}}(r+\epsilon_n)$ when any special message is sent.

Assuming correct decoding in the first stage, the error exponent for the second stage of decoding between $\lceil e^{nr} \rceil$ codewords equals $E(r)$, which is at most the sphere-packing exponent $E_{\rm sp}(r)$ (see [3]). Since the first stage exponent equals $E_{\rm sp}(r+\epsilon_n)$, the effective error exponent for special messages equals

$$\min\{E(r), E_{\rm sp}(r+\epsilon_n)\}$$

Since ϵ_n vanishes, the above two-stage decoding ensures a missed-detection exponent of $E(r)$ for each special message.

Now consider the situation of a uniformly chosen ordinary codeword being transmitted. We have to make sure the error probability is vanishingly small now. In this case, the output sequence distribution is i.i.d. Bern($\frac{1}{2}$) for the random coding ensemble. The first stage decoding error happens if one of the error sequences weighs less (in Hamming weight) than the threshold Φ. Since the outputs are i.i.d. Bern($\frac{1}{2}$), error sequence $Y^n \oplus \bar{x}^n(j)$ corresponding to any special codeword $\bar{x}^n(j)$ is also i.i.d. Bern($\frac{1}{2}$). Since $Y^n \in \mathcal{B}_j$ if and only if $Y^n \oplus \bar{x}^n(j)$ weighs less than Φ, this probability is at most

$$\exp(-nD_{\rm b}(\delta_{\rm GV}(r+\epsilon_n)\|1/2)) = \exp(-n(r+\epsilon_n)).$$

Applying union bound, the probability of $Y^n \in \bigcup \mathcal{B}_i$ is at most $\exp(-n\epsilon_n)$. This probability of the first stage error hence vanishes for the random coding ensemble. Recall that for the random coding ensemble, average error probability of the second-stage decoding also vanishes below capacity. To summarize, we have shown these two properties of the random coding ensemble:

1. Error probability of first stage decoding vanishes as $a^{(n)} \doteq \exp(-n\epsilon_n)$ with n when a uniformly chosen ordinary message is transmitted.

2. Error probability of second stage decoding (say $b^{(n)}$) vanishes with n when a uniformly chosen ordinary message is transmitted.

Since the first error probability is at most $4a^{(n)}$ for some 75% fraction of the random ensemble, and the second error probability is at most $4b^{(n)}$ for some 75% fraction of the

random ensemble, there exists a particular code which satisfies both these properties. The overall error probability for ordinary messages is at most $4(a^{(n)} + b^{(n)})$, which vanishes with n. We will use this particular code for the ordinary codewords. This de-randomization completes our construction of a reliable code for ordinary messages to be combined with the code $\mathcal{C}_{special}$ for special messages.

For the special codewords, we had already shown that, probability of first stage decoding error decays exponentially with exponent $E_{\text{sp}}(r)$. This completes the achievability proof for the BSC.

Appendix D

Proof of Lemma 39 in Chapter 6

We use $Y_{k,t}$ to denote output at user Y_k at time t and $Y_k^{1:t}$ denotes its outputs from time i to t. Thus subscript of a letter denotes the user index and its superscript denotes the time index.

Lets first bound $I(M_{L-1}; Y_{L-1}^{1:n})$. By chain rule,

$$\begin{aligned} I(M_{L-1}; Y_{L-1}^{1:n})) &\leq \sum_{t=1}^{n} I(M_{L-1}; Y_{L-1,t}|Y_{L-1}^{1:t-1}) \\ &\leq \sum_{t=1}^{n} \big(H(Y_{L-1,t}) - H(Y_{L-1,t}|M_{L-1}Y_{L-1}^{1:t-1})\big) = \sum_{t=1}^{n} I(Y_{L-1,t}; U_{L-1,t}) \end{aligned}$$

where $U_{L-1,t}$ is defined as

$$U_{L-1,t} \equiv M_{L-1}Y_{L-1}^{1:t-1} \tag{3}$$

We now get an upper bound on $I(M_{L-2}; Y_{L-2}^{1:n}|M_{L-1})$. First note that $H(Y_{L-2}^{1:n}|M_{L-1}M_{L-2})$ equals

$$\sum_{t=1}^{n} H(Y_{L-2,t}|M_{L-1}M_{L-2}Y_{L-2}^{1:t-1}) \geq \sum_{t=1}^{n} H(Y_{L-2,t}|M_{L-1}M_{L-2}Y_{L-2}^{1:t-1}Y_{L-1}^{1:t-1}) \tag{4}$$

since conditioning reduces entropy.

Now invoking the physical degradedness implies $I(Y_{L-1}^{1:t-1}; Y_{L-2,t}|M_{L-1}Y_{L-2}^{1:t-1}) = 0$. That is, $Y_{L-1}^{1:t-1}$, which is a noisier version of $Y_{L-2}^{1:t-1}$ does not provide any additional information about $Y_{L-2,t}$. This implies

$$\begin{aligned} H(Y_{L-2,t}|M_{L-1}Y_{L-2}^{1:t-1}) &= H(Y_{L-2,t}|M_{L-1}Y_{L-1}^{1:t-1}Y_{L-2}^{1:t-1}) \qquad (5) \\ &\leq H(Y_{L-2,t}|M_{L-1}Y_{L-1}^{1:t-1}) \\ &\qquad (6) \end{aligned}$$

Combining (4) and (6) gives,

$$\begin{aligned} I(M_{L-2}; Y_{L-2}^{1:n}|M_{L-1}) &= \sum_{t=1}^{n} I(M_{L-2}; Y_{L-2,t}|M_{L-1}Y_{L-2}^{1:t-1}) \\ &\leq \sum_{t=1}^{n} H(Y_{L-2,t}|M_{L-1}Y_{L-1}^{1:t-1}) - \sum_{t=1}^{n} H(Y_{L-2,t}|M_{L-1}Y_{L-1}^{1:t-1}M_{L-2}Y_{L-2}^{1:t-1}) \end{aligned}$$

Recalling (3) and defining

$$U_{L-2,t} \equiv M_{L-2}M_{L-1}Y_{L-2}^{1:t-1} \tag{7}$$

this bound equals

$$I(M_{L-2}; Y_{L-2}^{1:n}|M_{L-1}) = \sum_{t=1}^{n} I(U_{L-2,t}; Y_{L-2,t}|U_{L-1,t})$$

Using similar substitution of auxiliary random variables, we get the following set of upper bounds

$$I(M_k; Y_k^{1:n}|M_{k+1}M_{k+2}\cdots M_{L-1}) \leq \sum_{t=1}^{n} I(U_{k,t}; Y_{k,t}|U_{k+1,t})$$

for any $1 \leq k \leq L-1$, where $U_{k,t}$ is defined as

$$U_{k,t} \equiv M_k M_{k+1}\cdots M_{L-1}Y_k^{1:t-1} \tag{8}$$

Choice of auxiliaries: Thus auxiliary random variable for certain user at time t is defined as the set of

1. Past observations (up to time $t-1$) of that user and

2. all the messages that user wants to decode.

Note that for any time t, our choice of auxiliary variables satisfies the Markov structure in (6.22) due to physical degraded nature of the channel

Combine all such upper bounds to upper bound the RHS of Lemma 39 by

$$\sum_{t=1}^{n} \left(I(X_t; Y_{0,t}|U_{1,t}) + \sum_{i=1}^{L-1} \lambda_i I(U_{i,t}; Y_{i,t}|U_{i+1,t}) \right)$$

which by definition is at most $nC(\lambda)$. •

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[2] R. Gallager, *Info. Theory and Reliable Communication*, Wiley, 1968.

[3] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels", *Information and Control*, pp. 65-103, December 1966.

[4] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[5] D. Forney, "On exponential error bounds for random codes on the BSC," unpublished manuscript.

[6] R. Gallager, "Fixed Composition Arguments and Lower Bounds to Error Probability," unpublished course notes in information theory, online at `http://web.mit.edu/gallager/www/pages/pubs.html`.

[7] A. Montanari and D. Forney, "On exponential error bounds for random codes on the DMC," unpublished manuscript.

[8] D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Info. Theory*, vol. 14, no. 2, pp. 206-220, Mar. 1968.

[9] T. Cover, *Elements of Information Theory*, Wiley-Interscience, 1991.

[10] S. Borade, B. Nakiboğlu, L. Zheng, "Some fundamental limits of unequal error protection," *IEEE International Symposium on Information Theory*, Toronto, July 2008.

[11] S. Borade, B. Nakiboğlu, L. Zheng, "Unequal error protection: some fundamental limits ," submitted to *IEEE Transactions on Information Theory.*

[12] S. Borade, B. Nakiboğlu, L. Zheng, "Fundamental limits of UEP: data-rates below capacity," in preparation.

[13] S. Borade and L. Zheng,"Euclidean information theory," Allerton Conference, Monticello, Sept. 2007.

[14] S. Borade and L. Zheng,"Geometry of error exponents," Allerton Conference, Monticello, Sept. 2006.

[15] E. Bedrosian, Weighted PCM, *IRE Trans. Inform. Theory*, vol. 4, pp. 45-49, March 1958.

[16] B. Masnick, J. Wolf, "On linear unequal error protection codes," *IEEE Trans. Inform. Theory*, vol. 13, no. 4, pp. 600-607, Oct. 1967.

[17] A. Albanese, J. Blomer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1737-1744, Nov. 1996.

[18] S. Boucheron, M. Salamatian, "About priority encoding transmission," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 699-705, Mar. 2000.

[19] J. Hagenauer, " Rate compatible punctured convolutional (RCPC) codes and their applications ," *IEEE Transactions on Communications*, vol. 36, no. 4, pp. 389-400, April 1988.

[20] A. R. Calderbank, N. Seshadri, "Multilevel codes for unequal error protection," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1234-1248, July 1993.

[21] M. Trott, "Unequal error protection codes: theory and practice," *Proc. IEEE Information Theory Workshop*, Haifa, June 1996.

[22] I. Csiszar, "Joint source-channel error exponent," *Problems of Control and Information Theory*, vol. 9 (5), pp. 315-328, 1980.

[23] S. Diggavi, D. Tse, "On successive refinement of diversity," Allerton Conference, Illinois, September 2004.

[24] L. Bassalygo, V. Zinoviev, V. Zyablov, M. Pinsker, G. Poltyrev, "Bounds for Codes with Unequal Protection of Two Sets of Messages", *Problems of Information Transmission*, vol. 15, no. 3, pp. 190-197, 1979.

[25] L. Bassalygo, M. Pinsker, "A Comment on the Paper of Kasami, Lin, Wei, and Yamamura: Coding for the Binary Symmetric Broadcast Channel with Two Receivers," *Problems of Information Transmission*, vol. 24, no. 3, pp. 253-257, 1988.

[26] R. Ahlswede, G. Dueck, "Identification via channels," *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 15-29, Jan. 1989.

[27] M. Burnashev, "Data transmission over a discrete channel with feedback, random trans. time", Problems Info. Trans., vol. 12, pp. 10-30, 1976.

[28] P. Berlin, B. Nakiboğlu, B. Rimoldi and E. Telatar, " A Simple Derivation of Burnashev's Reliability Function," preprint, `arXiv:cs/0610145v2`.

[29] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Info. Theory*, vol. 25, pp. 729733, November 1979.

[30] B. Kudryashov, "Message transmission over a discrete channel with noiseless feedback" *Problemy Peredachi Imformatsii*, 21(1):313, 1979.

[31] A. Sahai, S. Draper, "Beating the Burnashev bound using noisy feedback," Allerton Conference, Monticello, Sept. 2006.

[32] A. Shiriaev, *P*robability, Springer-Verlag Inc., New York, NY, USA, 1996.

[33] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for control over a noisy communication link: Part II: vector systems," submitted to *IEEE Trans. on Information Theory.* Online preprint: `arXiv:cs.IT/0610146`

[34] A. Sahai and S. Mitter, "Source coding and channel requirements for unstable processes," *IEEE Trans. on Information Theory.* Online preprint: `arXiv:cs.IT/0610143`

[35] I. Csiszar and P. Shields, *Information Theory and Statistics: a Tutorial*, Foundations and Trends in Communications and Information Theory, editor S. Verdu, vol. 1, issue 4, 2004.

[36] A. Renyi, "On measures of entropy and information," Proc. 4th Berkeley Symp. on Math. Statist. Probability, vol. 1, pp. 547-561, Berkeley, 1961.

[37] I. Csiszar, "Generalized cutoff rates and Renyi's information measures," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 26-34, Jan. 1995.

[38] R. Gallager,"A Simple Derivation of the Coding Theorem and Some Application", *IEEE Trans. on Information Theory*, Vol. 11, No. 1, pp. 3-18, Jan. 1965.

[39] D. Guo, S. Shamai and S. Verdu, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1261-1282, April 2005.

[40] M. Chiang, S. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Trans. Inform. Theory*, vol. 50, no. 2, pp. 245-258, Feb. 2004.

[41] A. El Gamal, E. van der Meulen "A proof of Marton's coding theorem for the discrete memoryless broadcast channel ," *IEEE Trans. Inform. Theory*, vol. 27, no. 1, pp. 120-122, Jan. 1981.

[42] K. Marton and J. Korner, "General broadcast channels with degraded message sets," *IEEE Trans. Inform. Theory*, vol. 23, no. 1, pp. 60-64, Jan. 1977.

[43] J. Korner, A. Sgarro, "Universally attainable error exponents for broadcast channels with degraded message sets," *IEEE Trans. Info. Theory*, vol. 23, pp. 670-679, Nov. 1980.

[44] A. Lapidoth, E. Telatar, R. Urbanke, On wideband broadcast channels, *IEEE Trans. Info. Theory*, vol. 49, no. 12, pp. 32503258, Dec. 2003.

[45] S. Borade, L. Zheng, M. Trott "Multilevel broadcast networks," *IEEE International Symposium on Information Theory*, Nice, June 2007.

[46] S. Diggavi and D. Tse, "On opportunistic codes and broadcast codes with degraded message sets," *IEEE Information theory workshop*, 2006.

[47] C. Nair and A. El Gamal, "The Capacity Region of a Class of 3-Receiver Broadcast Channels with Degraded Message Sets," *IEEE International Symposium on Information Theory*, Toronto, July 2008.

[48] R. Gallager, Capacity and coding for degraded broadcast channels, *Probl. Pered. Inform.*, vol. 10, no. 3, pp. 3-14, 1974.

[49] R. Yeung and Z. Zhang, "On symmetrical multilevel diversity coding," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 609-621, Mar. 1999.

[50] R. Yeung, "Multilevel diversity coding with distortion," *IEEE Trans. Inform. Theory*, vol. 41, no. 2, pp. 412-422, Mar. 1995.

[51] J. R. Roche, Distributed information storage, *Ph.D. dissertation*, Stanford University, Mar. 1992.

[52] P. Whiting, E. Yeh, "Optimal encoding over uncertain channels with decoding delayconstraints," *IEEE International Symposium on Information Theory*, Sorento, June 2000.

[53] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inform. Theory*, vol. 21, no. 6, pp. 629-637, Nov. 1975.

[54] T. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. 18, pp. 2-14, Jan 1972.

[55] T. Cover, H. Permuter, "Capacity of Coordinated Actions," *IEEE International Symposium on Information Theory*, Nice, June 2007.

[56] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. 19, no. 2, pp. 197-207, Mar. 1973.

[57] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 279-280, Mar. 1974.

[58] T. Liu and P. Viswanath, "An extremal inequality motivated by multiterminal information-theoretic problems," *IEEE Trans. Inform. Theory*, vol. 53, no. 5, pp. 1839-1851, May 2007.

[59] V. Goyal, "Multiple Description Coding: Compression Meets the Network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–94, Sept. 2001.

[60] W. Equitz, T. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269-275, Mar. 1991.