Local to Global Geometric Methods in Information Theory

by

Emmanuel Auguste Abbe

B.S., M.S., Mathematics Department, Ecole Polytechnique Fédérale de Lausanne

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

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

© Massachusetts Institute of Technology 2008. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
6, 2008

Certified by... .....
Lizhong Zheng
Associate Professor
Thesis Supervisor

Certified by.. .....
Emre Telatar
Professor
Thesis Supervisor

Accepted by ....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 0 1 2008
LIBRARIES

ARCHIVES

Local to Global Geometric Methods in Information Theory

by

Emmanuel Auguste Abbe

Submitted to the Department of Electrical Engineering and Computer Science
on June 6, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

This thesis treats several information theoretic problems with a unified geometric approach. The development of this approach was motivated by the challenges encountered while working on these problems, and in turn, the testing of the initial tools to these problems suggested numerous refinements and improvements on the geometric methods.

In ergodic probabilistic settings, Sanov's theorem gives asymptotic estimates on the probabilities of very rare events. The theorem also characterizes the exponential decay of the probabilities, as the sample size grows, and the exponential rate is given by the minimization of a certain divergence expression. In his seminal paper, *A Mathematical Theory of Communication*, Shannon introduced two influential ideas to simplify the complex task of evaluating the performance of a coding scheme: the asymptotic perspective (in the number of channel uses) and the random coding argument. In this setting, Sanov's theorem can be used to analyze ergodic information theoretic problems, and the performance of a coding scheme can be estimated by expressions involving the divergence. One would then like to use a geometric intuition to solve these problems, but the divergence is not a distance and our naive geometric intuition may lead to incorrect conclusions. In information geometry, a specific differential geometric structure is introduced by means of "dual affine connections". The approach we take in this thesis is slightly different and is based on introducing additional asymptotic regimes to analyze the divergence expressions. The following two properties play an important role. The divergence may not be a distance, but locally (i.e., when its arguments are "close to each other"), the divergence behaves like a squared distance. Moreover, globally (i.e., when its arguments have no local restriction), it also preserves certain properties satisfied by squared distances.

Therefore, we develop the Very Noisy and Hermite transformations, as techniques to map our global information theoretic problems in local ones. Through this localization, our global divergence expressions reduce in the limit to expressions defined in an inner product space. This provides us with a valuable geometric insight to the global problems, as well as a strong tool to find counter-examples. Finally, in certain cases, we have been able to "lift" results proven locally to results proven globally.

We consider the following three problems. First, we address the problem of finding good linear decoders (maximizing additive metrics) for compound discrete memoryless channels. Known universal decoders are not linear and most of them heavily depend on the finite alphabet assumption. We show that by using a finite number of additive metrics, we can construct decoders that are universal (capacity achieving) on most compound sets. We then consider additive Gaussian noise channels. For a given perturbation of a Gaussian input distribution, we define an operator that measures how much variation is induced in the output entropy. We found that the singular functions of this operator are the Hermite polynomials, and the singular values are the powers of a signal to noise ratio. We show, in particular, how to use this structure on a Gaussian interference channel to characterize a regime where interference should not be treated as noise. Finally, we consider multi-input multi-output channels and discuss the properties of the optimal input distributions, for various random fading matrix ensembles. In particular, we prove Telatar's conjecture on the covariance structure minimizing the outage probability for output dimension one and input dimensions less than one hundred.

Thesis Supervisor: Lizhong Zheng
Title: Associate Professor

Thesis Supervisor: Emre Telatar
Title: Professor

Acknowledgments

I would first and foremost like to thank my family. Je remercie mes parents, Maria et Gérard, merci de tout coeur pour m'avoir sans cesse soutenu, pour votre grande patience et pour tout votre amour; mon travail vous est dédié. Je remercie en particulier ma chère soeur Vanessa, qui chaque jour me soutient avec tant d'affection et gaieté, ainsi que mon beau-frère Fabien et ma petite nièce adorée Malena. Ces remerciements vont également à ma grand-mère Assunta, ma tante Lina, mon oncle Pietro, mes cousins et cousines, ainsi qu'à toute la famille.

I would next like to thank my advisors, Professor Lizhong Zheng and Emre Telatar, for their invaluable guidance throughout my graduate studies. I am very grateful for the opportunity to learn from both of you, in so many rich and diverse ways. You have been very close to me (sometimes discussing research late at night over a Big & Tasty) and I will not forget your immense generosity. I would also like to thank my thesis committee members, Professor Sanjoy Mitter and Daniel Stroock, for teaching me many important subjects in an enlightening manner, for their helpful comments on my research, and for mentoring me during my graduate education. I feel very fortunate to have had the chance to interact closely with my whole committee throughout my Ph.D. studies. I would also like to thank Professor Robert Gallager for his precious advice on my presentations and research, as well as Doctor Olivier Lévêque and Professor Gérard Ben Arous for their valuable comments and suggestions at various stages of my studies. Finally, I would like to thank Professor Andrea Montanari and Devavrat Shah for introducing me to different fields of discrete probability, Professor Imre Csiszár for his stimulating discussions, Pascal Marmier (Swissnex) for inviting me to many interesting events and Doris Inslee for her attention to my administrative questions.

I would also like to express my recognition to all the persons who have been close to me over these years. My time at MIT would not have been the same without my close friends and our many adventures. I would like to specially thank Stephanie Gil for her care, her encouragement and help during my thesis redaction, Amir Khandani,

Shahriar Khushrushahi (BJIJ) and Kayvan Zainabadi (BJIS) for their candid support and friendship, as well as Alexander Tsankov (BJIF) and Hector Mobine; I also thank Mukul Agarwal, Amir Ali Ahmadi, Shashi Borade, Chung Chan, Sheng Jing and Baris Nakiboglu for making LIDS such a fun and interesting place to be. In addition to my friends in Boston, I would also like to thank all my friends in Switzerland, in particular Sebastian Aeschbach, Patrick Polikar and Steve Salom, as well as the members of the Information Processing Group in EPFL.

Contents

List of Figures

Chapter 1

Introduction

1.1 Place Your Bets

Here is a game that you can play for only one thousand swiss francs. We will toss one million fair coins a thousand times, and if any coin comes out head more than 750 times, you win two thousand swiss francs. Otherwise you are welcome to play the game again. Would you like to play?

When tossing one fair coin a large number of times, we expect the average number of heads to be close to a half. Observing an unusually high number of heads would be unexpected, but when millions of coins are tossed, this could conceivably happen. We may be tempted to agree with these statements at first glance, however, if asked to justify more specifically the validity of such claims, we would run into trouble quite quickly. There are many ambiguities in these statements: how large should the number of tosses be? What is an unusually high number of heads? And how strong should our expectations be? However, our attempt to justify these statements is actually becoming more confusing by asking these questions. In fact, these questions are not making much sense individually; the number of coin tosses, the notion of "unusually high" and the quantification of the word "unexpected" are all interconnected.

There is however, a way to make sense of such statements while circumventing the questions that caused ambiguities to begin with, and this way is by introducing an asymptotic perspective. Large Deviations theory provides a framework to define and

analyze very rare events, and the *rate* at which the probability of such events decays to zero is characterized.

Let $\{X_i\}_{i=1}^n$ be a set of mutually independent random variables defined on a finite set $\mathcal{Z}$ with probability distribution Q, and let $\hat{Q}_n$ denote their empirical distribution (which is consequently a random distribution). Let Π be a subset of probability distributions such that $Q \notin \text{cl}(\Pi)$. Sanov's theorem tells us that[1]

$$\mathbb{P}\{\hat{Q}_n \in \Pi\} \doteq e^{-n \inf_{P \in \Pi} D(P||Q)}.$$

One can use this theorem to conclude that some particular events have asymptotically negligible probabilities, or more precisely, have probabilities vanishing exponentially fast with the number of observations; in order to draw this conclusion, what matters is that the constant

$$\inf_{P \in \Pi} D(P||Q)$$

is strictly positive, i.e., that $Q \notin \text{cl}(\Pi)$. However, the rate at which the exponential decay occurs is important in many applications and this is precisely the case in information theory, where Sanov's theorem can be used to estimate the error probability of a specific ergodic communication scheme. Let us go back to the coin tossing problem for now. We can interpret the X_i's introduced earlier as being the outcome of the ith coin toss, i.e., X_i is a head (H) with probability $1/2$ and is a tail (T) with probability $1/2$. Let us consider the event that the average number of heads is larger than $3/4$, which is equivalent to requiring that the empirical distribution of the sequence belongs to the set $\Pi = \{P : P(H) \geq 3/4, P(T) \leq 1/4\}$. Hence Π is not containing (in its closure) the uniform distribution U that assigns probabilities $(1/2, 1/2)$ to head and tail. Applying Sanov's theorem, this event is a very rare event whose probability decays exponentially fast with n and with exponential rate

$$C = \inf_{P \in \Pi} D(P||U).$$

[1] a precise formulation of this statement, defining $\doteq$, will be presented in section 2.1.1

If we now toss more than one coin, is the probability of observing one sequence with an average number of heads larger than 3/4 still very rare? Of course this is more likely to happen now, but, as long as the number of coins is not growing with n, this is still a very rare event. However, if the number of coins *is* growing with n, we become less confident about such a claim. If it grows sub-exponentially or exponentially fast with a rate strictly less than C, this is still a very rare event; but if we increase the exponential rate a tiny bit above C, then this becomes a typical event. Hence, if the number of coin tosses grows exponentially, there is a phase transition happening around this critical value of C, where the model undergoes drastic behavioral changes. It is then crucial to be aware of where such a transition happens. But what is the value of C? How do we find the minimizer in Π of $D(P||U)$? If the divergence were to behave like a distance or a squared distance, the minimizer would be the distribution assigning probabilities $(3/4, 1/4)$, which would give $C \approx 0.13$. Although the divergence is not behaving as such in general, this conclusion is still true (and easy to check in the current simple setting). With this analysis in mind, the reader may make a more informed decision of whether to pay for the game or not.

The phenomenon illustrated in this example is the crux of many information theory problems. The number of coins to be tossed becomes the number of messages to be sent. If the house must ensure that having one winning coin is a very rare event, the communicators must ensure that one error in the messages detection is a very rare event. The more coins, or the more messages, the more risky. In both cases, an ergodic setting allows the use of repeated schemes, coin tosses or channel uses, to counter randomness. The more repetitions, the more randomness cancellations. Then, with large deviation techniques, critical phase transitions are pointed out. Of course, the communication problem is much more complex and it is mostly when considering non-binary channels that the real challenges of understanding the divergence geometry appear. But let us introduce for now more information theoretic perspectives through the following simple example of a binary channel. Roberto wants to communicate messages to Alicia, however, Alicia's dad is unhappy with Roberto talking to his daughter. The pair found a way around this. They decide to encode all Italian

words they need to communicate with sequences of 0's and 1's, all sequences having a constant length of 60. For example, "midnight" is encoded by the sequence of sixty 1's, "church" is encoded by the sequence of sixty 0's, whereas "fountain" is encoded by the sequence of fifty five 0's followed by five 1's. Then, every night at nine sharp, Roberto goes on his balcony which faces Alicia's window in the other end of the village and turns on or off his reading lamp every second, where off corresponds to the signal 0 and on corresponds to the signal 1. Of course, although Roberto's imagination when he talks to Alicia is boundless, the young boy does not need 2^{60} words to express his messages. Even if they had to encode all possible italian words, sequences of length 20 would be sufficiently long, since 2^{20} is over a million. So why do they need such extended sequences for only a few hundred words they may use? Once in a while, other lights may turn on and off in the proximity of Roberto's balcony, causing Alicia to receive an erroneous signal. Hence the pair purposely added extra signals in their encoding, hoping that by doing so, even if some signals have been corrupted in the sequence, Alicia could still be able to figure out Roberto's message. One night, Alicia faced the following dilemma: Roberto was sending her a message encoding the location at which they would secretly meet at midnight, but that night, Alicia received a sequence of fifty seven 0's and three 1's. This was close to the "fountain" as well as "the church"; she guessed the message was conveying the word "fountain". At the clock bell's twelfth toll, Roberto did not see Alicia at the church. After this, Roberto never returned to his balcony. What did the young people do wrong? Was their encoding inefficient to ensure enough reliability? Was Alicia's decoding inaccurate?

Assuming that Roberto's neighbors are switching their lights on and off in an ergodic manner, Alicia and Roberto may not have chosen their code book and decoding rules in the most efficient way. For example, when Alicia declares that the received message ending with 001111 comes from the transmitted message ending with 11111 instead of 00000, she may assume that the channel is flipping 0 to 1 and 1 to 0 with the same probability, which is less than a half. But if the channel instead very rarely flips the 1's into 0's (if she sees no light at all, there are few chances that Roberto had his light

on), her decoding rule was mismatched and suboptimal. Moreover, if the encoding was avoiding sequences that were as close as the code words for church and fountain, they may also have better prevented their mistake.

As we will see in the next chapters, when communication takes place over a discrete memoryless channel, and when codewords are randomly generated from a distribution $P_{\mathcal{X}}$, the probability that a code word, which has not been sent, achieves a score of at least γ for a score function F, is at most

$$e^{-n \inf_{P \text{ s.t. } F(P) \geq \gamma} D(P||P_{\mathcal{X}} \times P_{\mathcal{Y}})}. \tag{1.1}$$

The formal definition and interpretation of these expressions will be given in chapter 3. Roughly speaking, if we denote by $P_{\mathcal{Y}}$ the marginal distribution of the output signals (which is uniquely determined by the input and channel distribution), a code word drawn under $P_{\mathcal{X}}$ which has not been sent is independent from the received sequence, hence its joint distribution is the product distribution $P_{\mathcal{X}} \times P_{\mathcal{Y}}$. Therefore, previous bound can be obtained in a analogue way as the bound obtained for the introductory coin tossing problem. It is important to notice that now, the set Π and the reference measure Q depend on quantities to be designed by the receiver and transmitter, namely the decoder F and the encoding distribution $P_{\mathcal{X}}$. From the bound given in (1.1), we would like to choose F and $P_{\mathcal{X}}$ in order to maximize the exponent, i.e., to make sure that the probability that a code word which has not been sent, receives a score larger than gamma, is as small as possible. This allows us to get a faster exponential decay in the error probability, hence, having in mind the coin tossing problem, a higher possible data rate for the messages to be sent reliably, or a more reliable communication scheme for Alicia and Roberto[2]. Of course we cannot approach the whole coding problem over a discrete memoryless channel by simply looking at the bound (1.1). However, it gives quite an accurate idea of what kind of mathematical expressions are describing the performance of communication systems.

More generally, it is not so much of a restrictive point of view to claim that

[2]the rigorous definition of reliable communication will be given in chapter 3

most results in information theory can be stated in terms of optimizing a divergence expression under some set of constrained probability distributions. Moreover, the use of the divergence in order to "measure the distance" between probability distributions is present in many more applications of probability and statistics than just information theory, and may not always originate from a large deviation principle. Other examples are: statistical physics, quantum information theory, hypothesis testing, bayesian updating, EM algorithms and more.

1.2 Problems and Results Description

Motivations

Throughout this thesis we have two complementary motivations. The original motivation is to determine good coding schemes on the specific information theoretic problems that we considered, and that are described in this section. Coming up against familiar difficulties and challenges that have been well documented in the literature regarding these problems, one may have the impression that in order to make headway on these subjects, a new perspective may be required. Previous sections motivated how important the role of the divergence is in asymptotic results and information theory, hence how important it is to understand its behavior. Via the presentation of these toy problems, and furthermore, via the problems described below, we wish to underline how appealing and helpful a geometrical perspective would be. *This is the main thrust and second motivation of the work in this thesis; the development of geometrical methods in information theory.*

Therefore our two motivations have been feeding into each other throughout this work; introducing a geometrical perspective allowed us to make advancements in solving our problems, and in turn, progression in the problems brought to light some important geometric principles and techniques. These will be described in the next section, we now briefly present the problems.

Problems

1. Universality:

   Compound memoryless channels model communication over a memoryless channel whose law is unknown but remains fixed throughout a transmission. The transmitter and receiver, however, know that the channel law belongs to a given set. In [5], a generalized notion of capacity is defined for such compound discrete memoryless channels. The random coding arguments must be reexamined carefully for such problems, but a major difficulty arising in compound channels is regarding decoding strategies. The optimal decoding rule for a memoryless channel with known law (and equiprobable messages) is the maximum likelihood (ML) decoding rule. However, on a compound channel, the use of ML or any decoding rule using a notion of typicality are obviously ruled out, since the decoder must be defined without knowledge of the channel. It may be suspected that without making use of the channel law, a decoding rule could hardly perform as well as a decoder which can make use of the channel law. Yet, Goppa defined a decoding rule called the maximum mutual information (MMI), which performed equivalently well with or without the channel knowledge, and other decoding rules having this property have been introduced in [14], [21], [19]. Although MMI is theoretically ideal, it has a few drawbacks. Firstly, it is highly impractical (and so are all universal decoding rule) and secondly, it is highly dependendant on the discrete nature of the alphabets. The maximum likelihood is initially hard to implement as well, but its linear (additive) structure allows the use of algorithms such as belief propagation that simplify drastically its complexity (when code words have an algebraic structure). In the final discussion of the survey paper "Reliable Communication Under Channel Uncertainty", [19], the authors wrote the following conclusion: "the task of finding universal decoders of manageable complexity constitutes a challenging research direction". "The maximum likelihood decoder is generally much simpler to implement than a universal decoder (e.g. MMI), particularly if the codes

being used have a strong algebraic structure". The reason for this is that the maximum likelihood decoders have a linear structure, i.e., the maximum likelihood decoders maximize a score function which is linear over the block length, since $\log W^n(y|x) = \sum_{i=1}^n \log W(y_i|x_i)$. However, none of the known universal decoders[3] are linear.

Can a single decoder embody the property of linearity and universality?

In chapter 5, we raised the problem of finding good linear decoders over compound discrete memoryless channels.

2. Multi-user information theory:

   In his celebrated paper [29], Shannon established the capacity of the additive white Gaussian noise channel, whose performance is limited by thermal noise. In mutli-user communication schemes, interference caused by other users also perturb the transmitted signals. However, interference is fundamentally different than noise; because it is transmitted by other users, it has a definite structure. When should we or should we not treat interference as noise? This a central question raised in the Interference Channel, whose capacity region is unknown to date. In chapter 6, we consider symmetric Gaussian interference channel with two users and perform a *local analyzes*. It has been shown in [3] that for low interference, the optimal scheme for the sum-capacity is to treat interference as noise and use independent Gaussian code books. Say that we are now allowed to move in different directions around independent Gaussian distributions and want to maximize the sum-rate; how would we move to get a higher sum-rate? How does the value of the interference coefficient modify the optimal input distribution? In this chapter, we aim to quantify how to perturb independent Gaussian distribution in order to hurt or help each of the two users mutual informations, in particular, we want to identify for which values of the interference coefficient should we treat interference as noise or not.

3. MIMO channels:

[3] achieving capacity or optimal error exponents

If we consider the previous problem but now allow the transmitters and receivers to cooperate at any time of the communication, we are dealing with a Multiple Input Multiple Output channel. These channels model in particular, communication between a receiver and a transmitter having several antennas available for use. In fully scattered environments, independent structures on the fading matrix are assumed and it has been shown in [32] that the achievable rates can be greatly improved with the number of antennas (namely linearly increased with the minimum between the number of transmitting and receiving antennas). What kind of independent structures can support such claims? What are the optimal input distribution (in the ergodic coherent setting) when the fading matrix distribution has weaker symmetric structures than the one assumed in [32], or when correlations are present between the fading matrix entries? How does this change when we consider non-ergodic settings? In the non-ergodic setting, Telatar's conjecture describes the optimal input covariance matrix in the i.i.d Gaussian setting. This conjecture can be stated as follows in the case when the output dimension is one: let us consider the metrics on $\mathbb{C}^t$ induced by all possibles positive definite matrices of trace 1 ($\|h\|_A^2 = h^t A h$ with $A \geq 0$ and $\mathrm{tr}A = 1$). Which metric should we chose in order to minimize the probability that a vector drawn from an i.i.d complex (circularly symmetric) Gaussian distribution has a length shorter than x? Conjecture: for any $x \in \mathbb{R}$, there exists $k = k(x) \in \mathbb{Z}_+$, such that the optimal matrices are all contained in the unitary orbit of the diagonal matrix with k times the value $1/k$.

Results

- On a discrete memoryless channel and for compound sets having a finite union of one-sided components[4], we found a decoding rule that maximizes a finite number of linear metrics and achieves the compound capacity (cf. theorem 9). Practically, this gives a linear universal[5] decoding rule for most compound

[4] one-sided set are defined in definition 24, the reader may think for now of union of convex sets
[5] universality here is only concerned with achieving the same rate as an optimal decoder

sets. MMI can be seen as a generalized maximum a posteriori (MAP) decoder maximizing all possible metrics induced by any DMC. Hence, our result is telling that we do not need to take all DMC metrics in order to achieve the capacity on a given compound set S. It also tells us which metrics are the important ones. By extracting the one-sided components of S, and taking the MAP metrics induced by the worst channel of these components, we get a capacity achieving decoding rule. When S has a finite number of one-sided components, this decoding rule is generalized linear. We give a geometric interpretation of this result.

- Let g_v denote the Gaussian density with mean zero and variance v, and let $T : L \rightsquigarrow \frac{\sqrt{g_s} L \star g_v}{\sqrt{g_{s+v}}}$, where g_v is the Gaussian density of mean 0 and variance v. We found that the singular functions of this operator are given by the Hermite polynomials in $L_2(g_s, \mathbb{R})$ (multiplied by $\sqrt{g_s}$), and the singular values are powers of $\frac{s}{s+v}$ (which represents a signal to signal plus noise ratio). We show that for an additive Gaussian noise channel, and for Gaussian inputs, the operator $\|T(L)\|_{L_2}$ measures how much variation in the output entropy is induced by the input perturbation $g_s(1 + \varepsilon L)$. With this novel tool, we can prove that the optimal input distribution for the sum-rate (unit power for each users) undergoes a regime transition, if $a < 0.68$ the i.i.d. Gaussian distribution is a local maxima of the sum-rate and otherwise it is not a local maxima. This tells us that for $\alpha > 0.68$, interference should not be treated as noise and that the recent sum-capacity expression found in [3] cannot be tight for $\alpha > 0.68$. The numerical values given here are expressed as the roots of some polynomials given in chapter 6.

- For ergodic, coherent, MIMO channels, if the fading matrix and input covariance constraint set are invariant with respect to a subgroup G of unitary matrices, the optimal covariance matrix must commute with G. For the Kronecker fading model, we characterize a martini-filling optimal power allocation which preserves, although smoothens, the water-filling characteristics. We prove that for non-ergodic MIMO channels, Telatar's conjecture, concerning the structure

of the input distribution minimizing the outage probability, is verified in the MISO case for input dimensions[6] $n \leq 100$.

- Other side results included in this thesis are the followings. In section 4.2.2, a notion of very clear channels, representing the other extreme case of very noisy channels, is defined. Results concerning capacity and error exponents on very clear channels are presented. In section 4.3, channels that are getting noisier with the number of channel uses are introduced (abstracting a model with limited energy supply or dense interference network, for example). We show that although the Shannon capacity is zero for such channels, we can still use codes growing at a sub-exponential scale and whose adapted notion of rate is bounded by a modified notion of capacity, given by the Fisher information.

1.2.1 Geometric Approach

Geometrical approaches to information theoretic expressions and statistics problems have been investigated in different aspects. In [8], several geometric properties (pythagorean theorems) of the divergence are described. In [2],[26] (and in a work by N. N. Cencov) differential geometric techniques are applied to families of probability distributions, and to statistical models. The Fisher metric is used as a Riemannian metric, but instead of considering a connection which is Riemannian, dual connections, satisfying a generalized metric connection condition, are employed. The divergence is then defined through those dual connections, in agreement with its non symmetric behavior, and is shown to satisfy the geometrical results presented in [8] (independently of the differential structure). The reason for which this geometrical setting is somehow peculiar, is precisely to take into account the non symmetric nature of the divergence, which at times behaves just like a squared distance, but in general is not even symmetric, nor has its symmetric sum $(\frac{1}{2}(D(p||q) + D(q||p)))$ satisfying the triangle inequality.

[6] the value 100 is symbolic and expresses the fact that as long as the dimension is given to us, we could conclude the last step of the proof, which asks to satisfy the increasing property of some confluent hypergeometric functions. We do not have a general argument to conclude the last step for generic values of n, due to the complexity of the expressions to manipulate

The approach we take in this thesis is in slightly different. The divergence is not a distance, but it behaves locally like a squared distance. We use the term "local", when the probability distributions considered are assumed to be close to each other. The rigorous meaning of "close to each other" is given in 2.2; in words, we want to express a setting for which the divergence is well approximated by a squared distance, since for $Q \in M_1(\mathcal{Z})$ and $L \in M_0(P)$, we have

$$D(Q(1+\varepsilon L)||Q) = \varepsilon^2 \frac{1}{2} \sum_{z \in Z} L^2(z) Q(z) + o(\varepsilon^2)$$

Hence, "global" is simply referring to the case where no local assumptions are made. But in addition to behave locally like a distance square, the divergence satisfies several properties of squared distances in the global setting (cf. section 2.1). In chapter 5, we will see how the divergence expressions appearing in information theory problems can be cumbersome and hard to manipulate (cf. (5.8)), this is why we develop the VN and Hermite transformation, to map global problems into local ones. The VN transformation maps global discrete memoryless channels into very noisy channels. Very noisy channels have been used since 1963 (by Reiffen) in different contexts, but here, we are investigating their geometrical properties. Mathematically, this transformation maps an arbitrary stochastic matrix into stochastic matrices which are perturbations of a constant column matrix. The Hermite transformation is instead used to analyze input distribution for the Gaussian interference channel, by perturbing input distribution in the hermite polynomials directions. As we will see in chapter 4 and 6, these two transformations will precisely map our respective global problems in the local setting presented just above. In both cases, original information theoretic quantities are expressed as objects in an inner product space, providing us with a significant geometrical insight. But the localization is not only providing intuition or counter-examples on the original problem, it also preserves in some cases most of the global problem's essence. This allowed us in some cases to lift the results found locally to global results, such as in the problem of linear universal decoding, where we could establish a global linear universal decoding rule for most compound sets.

This thesis summarizes some techniques that made significant breakthroughs in our problems and we believe, would also successfully apply to many other problems in information theory.

1.3 Thesis Outline

Outline:

The thesis is divided into three major parts. The first part includes chapter 2 to 4 and set the main global and local geometric ideas in an abstract setting. Chapter 2 is a generic introduction large deviations and the divergence. We interpret the divergence has a "distance" governing the geometry of rare events, and precisely because it is not a formal distance, we spend some time in section 2.1 to understand its global geometric properties. Section 2.2 introduces the local behavior of the divergence and of the I-projection. In chapter 3, we deal with the information theoretic setting of discrete memoryless channels and use the ideas of chapter 2 to understand the global geometric properties of those channels. In chapter 4 we consider very noisy channels to introduce the local setting developed in section 2.2 in discrete memoryless channels and understand their local geometry. The second part of the thesis is the application of these techniques to the concrete information theoretic problems described in the previous section. Chapter 5, deals with the first problem on linear universal decoding, and chapter 6 with the second problem on the interference channel (this chapter is also containing work in progress). Finally, chapter 7 and 8 are dealing with the third problem on MIMO channels. Those two chapters are considered to belong to a third part of the thesis, since they do not use the local to global geometric techniques but directly consider the global problem.

Chapter 2

Local and Global Settings

We denote by $M_1(\mathcal{Z})$ the set of probability distributions on $\mathcal{Z}$, where $\mathcal{Z}$ is a finite set. We denote by $M_0(P)$ all real functions on $\mathcal{Z}$ which are integrating to zero with respect to P, i.e., $M_0(P) = \{L : \mathcal{Z} \to \mathbb{R} | \sum_{z\in\mathcal{Z}} L(z)P(z) = 0\}$. Roughly speaking, we use the term "local" to describe a given problem setting, when the probability distributions considered are assumed to be close to each other. The rigorous meaning of "close to each other" is given in 2.2; in words, we want to express a setting for which the divergence is well approximated by a squared distance, since for $Q \in M_1(\mathcal{Z})$ and $L \in M_0(P)$, we have

$$D(Q(1+\varepsilon L)||Q) = \varepsilon^2 \frac{1}{2} \|L\|_Q^2 + o(\varepsilon^2),$$

where

$$\|L\|_Q^2 = \sum_{z\in Z} L^2(z)Q(z).$$

"Global" is then simply referring to the case where no assumptions are made on the considered probability distributions regarding their "distances", measured with the divergence. In chapter 5, we will see how the divergence expressions appearing from asymptotic results in information theoretic problems can be cumbersome and hard to manipulate (cf. (5.8)). However, working locally allows us to reduce divergence expressions into objects defined in an inner product space, giving us a better intuition

on how to picture these expressions. Hence, the problems expressed in the local setting become more tractable and concrete solutions can be found. An important idea presented in the current chapter is the following. Not only does the divergence locally behave like a squared distance but also, globally, it satisfies certain properties of squared distances; with this, the localization will turn out to be a accurate reduction of the considered problem, allowing us in certain cases to extend local results, to the global setting.

2.1 Global Geometric Properties of the Divergence

Formally speaking, the divergence is not a distance. Although it is always positive and vanishes only when its two arguments are identical, it is not symmetric and it does not satisfy the triangle inequality. Also, its symmetric sum, $\frac{1}{2}(D(p||q)+D(q||p))$, does not satisfy the triangle inequality. However, as will be illustrated in the following results, the divergence satisfies a few properties that are characteristic of squared distances.

The set of probability distributions (over a finite set) is not a space; it can be identified with the simplex of corresponding dimension. If one considers only distributions having non-zero probabilities, i.e., the open simplex, we have a clear differentiable manifold structure and tangent planes are easily defined (when borders are included, see [30] for expressions of the tangent planes). In [2], [26] a differential geometric framework is introduced, proposing an interpretation of the divergence by the means of dual connections. The Fisher metric is used as a Riemannian metric, but instead of considering a connection which is Riemannian, dual connections, satisfying a generalized metric connection equation, are employed. The divergence is then defined through those dual connections, agreeing with its non symmetric behavior, but leaving us with an unusual Riemannian geometry. We will not focus on this setting here, we will simply present a few geometric properties that the divergence satisfies, helping us to build our first geometric intuition. We start by introducing how the divergence originate in our problems.

2.1.1 Large Deviations and Induced Geometry

Let $\{X_i\}_{i=1}^n$ be a set of mutually independent random variables defined on a finite set $\mathcal{Z}$ with probability distribution $Q \in M_1(\mathcal{Z})$, and let $\hat{Q}_n$ denote their empirical distribution (which is consequently a random distribution). Let $\Pi \subset M_1(\mathcal{Z})$ be a set whose closure, $\mathrm{cl}(\Pi)$, is equal to the closure of its interior and such that $Q \notin \mathrm{cl}(\Pi)$ (though without latter assumption the forthcoming results become trivial). We now state the Sanov's theorem in this particular setting, although the theorem can be stated in much greater generality.

Theorem 1. (Sanov)

$$-\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\{\hat{Q}_n \in \Pi\} = \inf_{P\in\Pi} D(P||Q).$$

The functional D is called the Kullback-Leibler or Information divergence, as well as the relative entropy (also denoted $h(P|Q)$) and we will simply call it the divergence:

$$D(P||Q) = \sum_{z\in\mathcal{Z}} P(z) \log \frac{P(z)}{Q(z)}, \quad \text{if } \mathrm{Supp}(Q) \subseteq \mathrm{Supp}(P),$$

and infinity otherwise.

We adopt the following notation to express previous statement in a more handy way:

$$\mathbb{P}\{\hat{Q}_n \in \Pi\} \doteq e^{-n \inf_{P\in\Pi} D(P||Q)},$$

where $a(n) \doteq b(n)$ means $-\lim_{n\to+\infty} \frac{1}{n} \log a(n) = -\lim_{n\to+\infty} \frac{1}{n} \log b(n)$.

Definition 1. Let Π be a closed convex non-empty subset of $M_1(Z)$ and let $Q \in M_1(\mathcal{Z})$ with $Q > 0$ (i.e., $Q(z) > 0, \forall z \in \mathcal{Z}$). The I-projection of Q onto Π is

$$P_0 = \arg\min_{P\in\Pi} D(P||Q).$$

Note that since $Q > 0$, the function $P \mapsto D(P||Q)$ is continuous and strictly convex in P, which implies the existence and uniqueness of P_0. The assumption made on Π and Q in the definition can certainly be relaxed in order to get meaningful

variant of this definition, but since we will not need a more general setting in this chapter, we will content ourselves with this definition.

In what follows, several properties of the I-projection will be investigated. We will see that the I-projection behaves in several respects like an analogue of the Euclidean projection defined on $\mathbb{R}^N$ by

$$p_0 = \arg\min_{p \in S} ||p - q||^2,$$

where $p_0 \in \mathbb{R}^N$ and S is a closed subset of $\mathbb{R}^N$.

2.1.2 Pythagorean Theorems

In the Euclidean geometry of $\mathbb{R}^N$, an hyper-plane is described by all points x satisfying a set of $1 \leq i \leq N$ linear equations of the form $\langle f_i, x \rangle_{\text{eucl}} = f_i^T \cdot x = \alpha_i$, with $f_i \in \mathbb{R}^N$ and $\alpha_i \in \mathbb{R}$. We refer to the f_i's as being the normal directions, since for any point x in the hyper-plane, the other hyper-plane given by all points y satisfying $y = x + \sum_i \lambda_i f_i$ for some λ_i's in $\mathbb{R}$, is orthogonal (with respect to the Euclidean inner product) to the first one. Moreover, the projection of a point onto an hyper-plane belongs to the intersection with the normal hyper-plane.

We now consider $M_1(\mathcal{Z})$ instead of $\mathbb{R}^N$.

Definition 2. Let $k \geq 1$, $i \in \{1, \ldots, k\}$, $f_i : \mathcal{Z} \to \mathbb{R}$ (*normal directions*) and $\alpha_i \in \mathbb{R}$ (*shifts*). A linear family $\mathcal{L}_{\{f_i, \alpha_i\}}$ in $M_1(\mathcal{Z})$ is defined by

$$\mathcal{L}_{\{f_i, \alpha_i\}} = \{P \in M_1(\mathcal{Z}) | \forall 1 \leq i \leq k, \mathbb{E}_P f_i = \alpha_i\}.$$

Definition 3. Let $k \geq 1$, $i \in \{1, \ldots, k\}$, $f_i : \mathcal{Z} \to \mathbb{R}$ (*directions*) and $P_0 \in M_1(\mathcal{Z})$. An exponential family $\mathcal{E}_{P_0, \{f_i\}}$ in $M_1(\mathcal{Z})$ is defined by

$$\mathcal{E}_{P_0, \{f_i\}} = \{P \in M_1(\mathcal{Z}) | \exists \lambda \in \mathbb{R}^k \text{ s.t. } P = P_0 e^{\sum_{i=1}^k \lambda_i f_i} c(\lambda)\},$$

where $c(\lambda) = (\sum_{z \in \mathcal{Z}} P_0(z) e^{\sum_{i=1}^k \lambda_i f_i(z)})^{-1}$.

The linear families will be pictured in a similar way as the hyper-planes in the Euclidean geometry, the f_i's can also be interpreted as normal directions, not with respect to another linear family, but with respect to an exponential family. Let $\mathcal{L}_{f,\alpha}$ be a linear family passing through a point P_0 and $\mathcal{E}_{P_0,f}$ its "normal" exponential family passing through P_0. As the following theorem shows, we then have similar properties as in the Euclidean setting, involving the divergence instead of the Euclidean squared norm. The proofs of this results can be found in [8].

Theorem 2. *For any $Q \in \mathcal{E}_{P_0,f}$, we have*

$$\arg \min_{P \in \mathcal{L}_{f,\mathbb{E}_{P_0} f}} D(P||Q) = P_0.$$

In the Euclidean setting, the projection p_0 of a vector q on a linear subspace S is characterized by the orthogonality principle, or equivalently by the Pythagorean theorem $||p - q||^2 = ||p - p_0||^2 + ||p_0 - q||^2$, $\forall p \in S$. In the probability setting, the following similar result holds, which encompasses the previous theorem.

Theorem 3. *Let $Q \in M_1(\mathcal{Z})$ and*

$$P_0 = \arg \min_{P \in \mathcal{L}_{\{f_i,\alpha_i\}}} D(P||Q),$$

then

$$D(P||Q) = D(P||P_0) + D(P_0||Q)$$

and if $\text{Supp}(\mathcal{L}_{\{f_i,\alpha_i\}}) = \mathcal{Z}$, *we have*

$$\mathcal{L}_{\{f_i,\alpha_i\}} \cap \mathcal{E}_{Q,\{f_i\}} = \{P_0\}.$$

If $\text{Supp}(\mathcal{L}_{\{f_i,\alpha_i\}}) \neq \mathcal{Z}$, *last statement holds when $\mathcal{E}_{Q,\{f_i\}}$ is replaced with* $\text{cl}(\mathcal{E}_{Q,\{f_i\}})$.

This theorem is illustrated in figure 2-1.

Corollary 1. *For any directions $f_1, \ldots, f_k$ and shifts $\alpha_1, \ldots, \alpha_k$, if $Q_1, Q_2 \in \mathcal{E}_{Q,f_1}$*

$$\arg \min_{P \in \mathcal{L}_{\{f_i,\alpha_i\}}} D(P||Q_1) = \arg \min_{P \in \mathcal{L}_{\{f_i,\alpha_i\}}} D(P||Q_2).$$

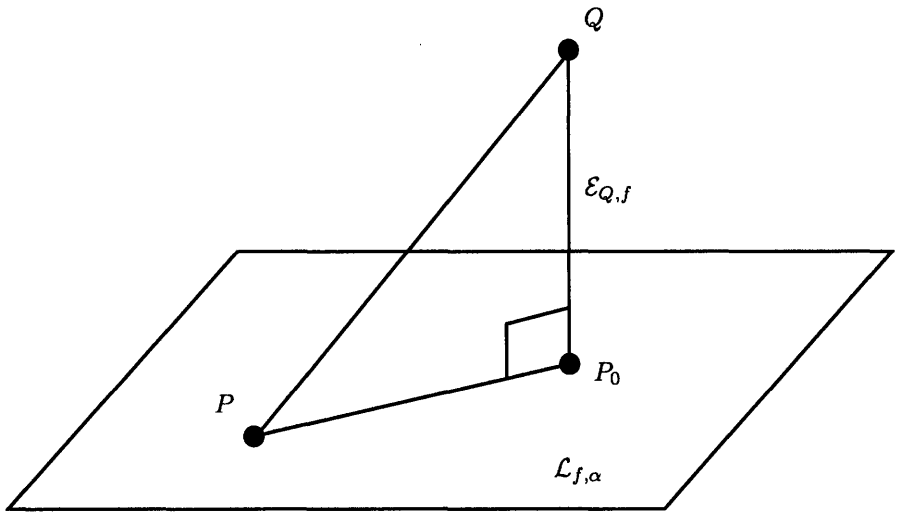

Figure 2-1: I-projection over a linear family and Pythagorean equality

The statement of this corollary is illustrated in figure 2-2.

We now consider constraint sets that are not necessarily linear, but just convex. In the Euclidean setting, when minimizing the Euclidean distance from a reference point to a convex closed set C:

$$p_0 = \min_{p \in C} ||q - p||^2,$$

we clearly have the following inequality

$$||p - q||^2 \geq ||p - p_0||^2 + ||p_0 - q||^2.$$

Again, a similar result hold for the I-projection.

Theorem 4. *Let C a convex set, then* $\text{Supp}(P_0) = \text{Supp}(C)$ *and*

$$D(P||Q) \geq D(P||P_0) + D(P_0||Q).$$

Where the support of a convex set C is defined by the support of the element of C

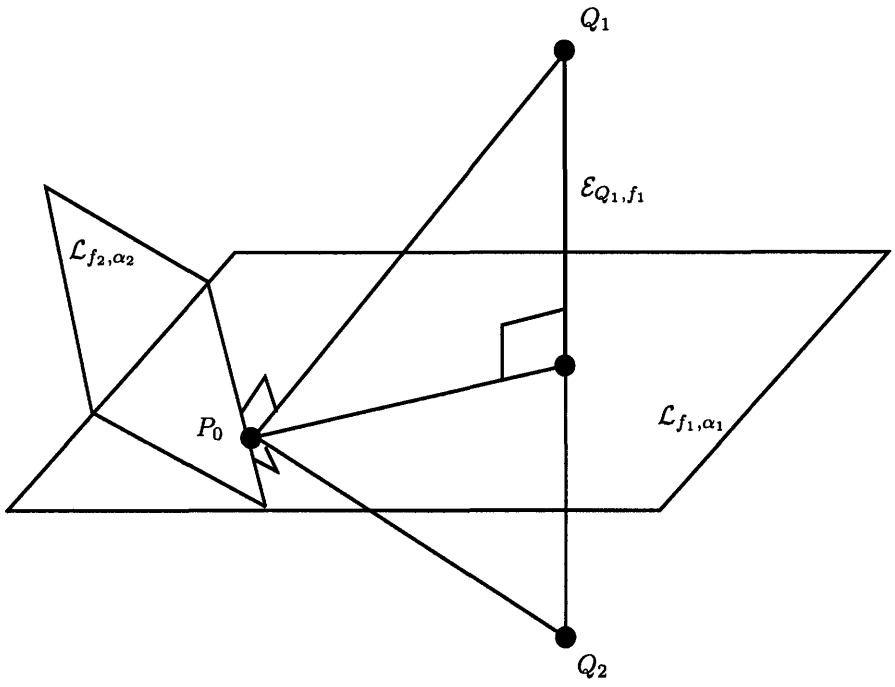

Figure 2-2: I-projections from a common exponential family onto several linear families

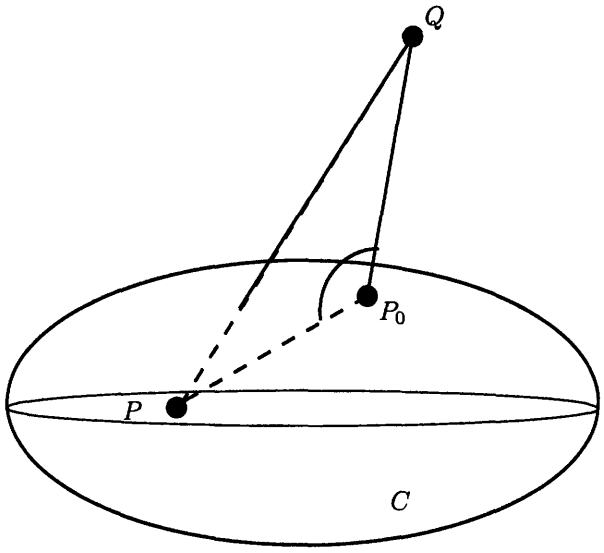

Figure 2-3: Minimization of the divergence over a convex family and obtuse principle

that contains all other elements support.

This theorem is illustrated in figure 2-3.

This result also suggests that $D(P||P_0) + D(P_0||Q) - D(P||Q)$ could perhaps be related to a notion of angle between the distributions P, P_0 and Q at P_0, having in mind the analogy with the Euclidean distance where

$$||p - p_0||^2 + ||p_0 - q||^2 - ||p - q||^2 = 2\langle p - p_0, q - p_0\rangle_{\text{eucl}},$$

which is zero for the orthogonality principle of the projection on linear families, and negative on convex sets, since the angle must be obtuse. So perhaps we could define an inner product on the set of distributions such that

$$\langle P - P_0, Q - P_0\rangle \propto (D(P||P_0) + D(P_0||Q) - D(P||Q)). \tag{2.1}$$

Of course, a few abuses of the analogy with the Euclidean setting have been made

here. Since we are not working in a space, subtracting distributions takes us out of the simplex. If one sees $M_1^o(Z)$ as a manifold, to be equipped with a Riemmanien inner product, the element of the tangent planes will indeed be measures integrating to 0, just like the tangent of the curve $tP+(1-t)P_0 \in M_1(Z)$ with $t \in [0,1]$, which is $P - P_0$. However, the non symmetric behavior of the Divergence, implies that P and Q cannot be exchanged in the right hand side of (2.1), and as consequence, such an inner product (which has to be symmetric) would not take the tangent vector in the symmetric way suggested by (2.1). Note that when proving the Pythagorean theorem of the I-projectoin on a linear family, P and P_0 are both distributions belonging to the linear family, hence the curve $\gamma_l(t) = tP + (1-t)P_0 \in M_1(Z)$, with $t \in [0,1]$, is included in this linear family and its tangent vector at P_0 (or at any other points) is

$$\frac{\partial}{\partial t}\Big|_{t=0} \gamma_l(t) = P - P_0.$$

Note that γ_l is a 1-dimensional linear family embedded in any linear family containing containing P and P_0. But Q and P_0 both belong to the exponential family (not the linear family) orthogonal to the linear family containing P and P_0, and $\gamma_e(t) = P^t P_0^{1-t} c(t)$ also belongs to that family, with tangent vector at P_0 given by

$$\frac{\partial}{\partial t}\Big|_{t=0} \gamma_e(t) = P_0(\log \frac{Q}{P_0} + D(P_0||Q)).$$

Consider now the Fisher inner product of those curves at $t = 0$, i.e., $\gamma_l(0) = \gamma_e(0) = P_0$, we get

$$\begin{aligned}
\langle \gamma_l, \gamma_e \rangle_{\text{Fisher},t=0} &= \mathbb{E}_{\gamma_l(0)} \frac{\partial}{\partial t}\Big|_{t=0} \log \gamma_l \frac{\partial}{\partial t}\Big|_{t=0} \log \gamma_e && (2.2)\\
&= \mathbb{E}_{P_0}(P - P_0)(P_0 \log \frac{Q}{P_0} + P_0 D(P_0||Q)) \frac{1}{P_0^2} && (2.3)\\
&= \sum_z (P(z) - P_0(z)) \log \frac{Q_(z)}{P_0(z)} && (2.4)\\
&= D(P||P_0) + D(P_0||Q) - D(P||Q). && (2.5)
\end{aligned}$$

The inner product defined in (2.2) can be expressed in a simpler way by

$$\langle X, Y \rangle_{Fisher,P_0} = \mathbb{E}_{P_0} \frac{XY}{P_0^2}, \quad \forall X, Y \in M_0(\mathcal{Z}).$$

This is an ad hoc way of introducing an inner product structure using those specific curves. A more rigorous treatment of those ideas can be found in [2], where two connections are introduced, ∇ and ∇^*, for which the linear and exponential families γ_1 and γ_2 are respective geodesics, and that are dual to each other for the Fisher metric g in the sense that they satisfy the following generalized metric connection property:

$$Z\langle X, Y \rangle_g = \langle \nabla_Z X, Y \rangle_g + \langle X, \nabla_Z^* Y \rangle_g.$$

Although the latter definitions may embed the expansions presented in this section in a more formal differential geometric setting, they also introduce an unusual geometry, since the connection used is not metric.

More results regarding the geometrical properties of the divergence can be found in [8], the ones that we presented here are the more fundamental ones, and also the ones we will use in our problems.

2.2 Local Properties of the Divergence

Although the divergence is not a squared distance, we showed in previous section that for certain properties, the divergence behaves like a squared distance. In this section, we will show that when the probability distributions are close to each other, the divergence does indeed collapse to a squared distance, giving up its non-symmetric behavior.

Let $Q \in M_1(\mathcal{Z})$ and $L \in M_0(P)$. The following identity is the main ingredient of this section:

$$D(Q(1 + \varepsilon L)||Q) = \varepsilon^2 \frac{1}{2} \sum_{z \in Z} L^2(z) Q(z) + o(\varepsilon^2) \tag{2.6}$$

We will use the following terminology to talk about expressions such as $Q(1+\varepsilon L)$, we refer to Q as the limiting distribution and to L as the direction of the perturbation.

2.2.1 Local Large Deviations

Let us go back to our probability framework, where $\{X_i\}_{i=1}^n$ are i.i.d. from a distribution $Q \in M_1^o(\mathcal{Z})$ and Π is as described in section 2.1.1. We now assume that Q is parameterized by

$$Q_\varepsilon = R(1+\varepsilon L_Q), \quad \varepsilon \ll 1 \tag{2.7}$$

where $R \in M_1^o(\mathcal{Z})$ and $L_Q \in M_0(R)$ and similarly

$$\Pi_\varepsilon = \{R(1+\varepsilon L) | L \in \Lambda\}, \tag{2.8}$$

where $\Lambda \subset M_0(R)$ is convex closed. Note that we expressed Q and Π in a parameterized form to start with, but we could have taken our original Q and Π and considered the following parametrized distributions:

$$Q_t = c(t)Q^t R^{1-t} = c(t)Re^{t\log\frac{Q}{R}} = R(1+t(\log\frac{Q}{R} + D(Q||R))) + o(t),$$

provided that Q and R are strictly positive distributions (and similarly for Π). Note that the addition of $D(Q||R) = -\mathbb{E}_R \log\frac{Q}{R}$ ensures that the direction is centered with respect to R, forcing the evolute to stay in the simplex at any time. The previous expansion gives a justification of why we use the letter L for the direction of our local perturbations, we can think of L as being the (centered) log-likelihood ratio between the target distribution and the limiting distribution where we want to perform our local analysis, i.e.,

$$L_Q = \log\frac{Q}{R} + D(Q||R)$$

We could define a linear evolute too:

$$Q_t = tQ + (1-t)R = R + t(Q-R) = R(1 + t\frac{Q-R}{R}),$$

where last equality holds only if R is a strictly positive distribution. The meaning of these specific choices of paths, to get to a limiting distribution, is not investigated further in this work, but as mentioned in section 2.1, exponential and linear curves (1-dimensional families) can be interpreted as geodesics with respect to Amari's (e) and (m) connections. But for the purpose of this section, we only need to care about the description of these paths near the limiting distributions, which is, as shown in previous paragraph, described by linear perturbation in the infinitesimal parameter. The next section will provide an example on how to define and use these local perturbations in information theoretic settings, for now we will we use the setting defined in (2.7) and (2.8).

From Sanov's theorem, we have for any $\varepsilon > 0$

$$-\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\{\hat{Q}_n \in \Pi_\varepsilon\} = \inf_{P\in\Pi_\varepsilon} D(P||Q_\varepsilon)$$

and obviously

$$\lim_{\varepsilon\to 0} \inf_{P\in\Pi_\varepsilon} D(P||Q_\varepsilon) = 0,$$

but what we are interested in is the behavior of these expressions for small ε.

Proposition 1. *Let $R \in M_1(\mathcal{Z})$ with $R > 0$, $L_Q \in M_0(R)$ and $\Lambda \subset M_0(R)$ convex compact. Then,*

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon^2} \inf_{P\in\Pi_\varepsilon} D(P||Q_\varepsilon) = \frac{1}{2} \inf_{L\in\Lambda} ||L - L_Q||_R^2,$$

where $||\cdot||_R$ is the L_2 norm with weight measure R.

Proof. Let $M = \max_{a\in\mathcal{X}, b\in\mathcal{Y}} L(a,b) \vee L_Q(a,b)$ and $T = 1/M > 0$. Then

$$f : \varepsilon \mapsto D(R(1+\varepsilon L)||R(1+\varepsilon L_Q))$$

is analytic on $(-T, T)$ and since $f(0) = 0$ and f is positive, the first two terms in the Taylor expansion of f are vanishing and computing the second derivative at zero we get

$$D(R(1+\varepsilon L)||R(1+\varepsilon L_Q)) = \frac{1}{2}||L - L_Q||_R^2 \varepsilon^2 + o(\varepsilon^2).$$

But

$$\inf_{P \in \Pi_\varepsilon} D(P||Q_\varepsilon) = \inf_{L \in \Lambda} D(R(1+\varepsilon L)||R(1+\varepsilon L_Q)),$$

hence

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \inf_{P \in \Pi_\varepsilon} D(P||Q_\varepsilon) &= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \left[\inf_{L \in \Lambda} \frac{1}{2} ||L - L_Q||_R^2 \varepsilon^2 + o(\varepsilon^2) \right] \\ &= \frac{1}{2} \inf_{L \in \Lambda} ||L - L_Q||_R^2. \end{aligned}$$

□

This result tells us that if we work locally around a distribution R, the I-projection behaves like a norm-space projection defined in $M_0(R)$, the space of functions having zero mean under R (i.e., $\sum_z L(z)R(z) = 0$), with the L_2 norm weighted by R. This is illustrated in figure 2-4

Note that this result strengthens the choice of the inner product defined in section 2.1.1, as illustrated in figure 2-5, since the tangent vector at P_0 of the exponential curve is given by $P_0(\log \frac{Q}{P_0} + D(Q||P_0))$ and the tangent vector at P_0 of the linear curve is given by $P - P_0 = P_0 \frac{P - P_0}{P_0}$, the Fisher inner product between between these two tangent vectors, as expressed in (2.3), is indeed equal to the L_2 inner product of the two directions $\log \frac{Q}{P_0} + D(Q||P_0)$ and $\frac{P-P_0}{P_0}$ with weight P_0, i.e.,

$$\mathbb{E}_{P_0}(P - P_0)(\log \frac{Q}{P_0} + D(Q||P_0))$$

agreeing with the local result.

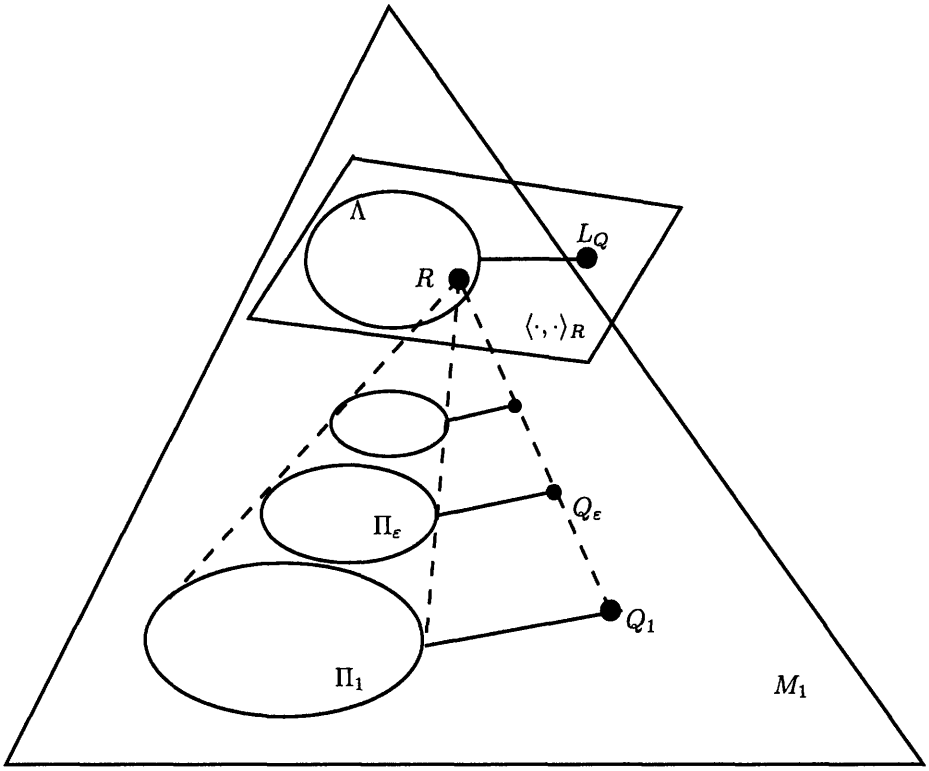

Figure 2-4: Local I-projection in proposition 2.2.1

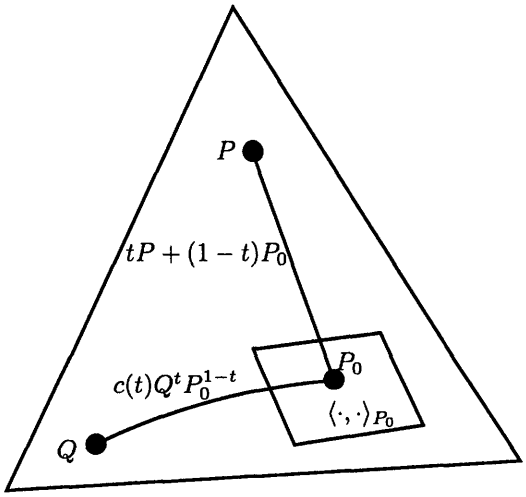

Figure 2-5: Fisher inner product

2.2.2 Moderate Deviations

The local setting of previous section has been defined by taking first the limit as n goes to infinity and then the limit as ε goes to zero. The current section investigates what happens if those two parameters are approaching their limits simultaneously.

We keep the same setting as in previous section (with $Q > 0$) and we consider $\varepsilon = \varepsilon(n)$ such that

$$\varepsilon(n) \to 0, \quad n\varepsilon(n)^2 \to \infty.$$

Theorem 5. *For $\Gamma \subset M_0(\mathcal{Z})$, we have*

$$\mathbb{P}\{\frac{1}{\varepsilon(n)}(\hat{Q}_n - Q) \in \Gamma\} \doteq e^{-n\varepsilon(n)^2 \inf_{\nu\in\Gamma} \frac{1}{2}\sum_{z\in\mathcal{Z}} \frac{\nu(z)^2}{Q(z)}}$$

or equivalently, for $\Lambda \in M_0(Q)$

$$\mathbb{P}\{\hat{Q}_n \in Q(1+\varepsilon(n)\Lambda)\} \doteq e^{-n\varepsilon(n)^2 \inf_{L\in\Lambda} \frac{1}{2}\|L\|_Q^2}$$

Corollary 1. *Let $Q_{\varepsilon(n)}$ such that $\lim_{n\to\infty} Q_{\varepsilon(n)} = Q$, then*

$$\mathbb{P}\{\frac{1}{\varepsilon(n)}(\hat{Q}_n - Q_{\varepsilon(n)}) \in \Gamma\} \doteq e^{-n\varepsilon(n)^2 \inf_{\nu\in\Gamma} \frac{1}{2}\sum_{z\in\mathcal{Z}} \frac{\nu(z)^2}{Q(z)}}$$

Heuristic Proof: we know, from previous section, that

$$\mathbb{P}\{\hat{Q}_n \in Q(1+\varepsilon\Lambda)\} \doteq e^{-n\left(\varepsilon^2 \inf_{L\in\Lambda} \frac{1}{2}\|L\|_Q^2 + o(\varepsilon^2)\right)}.$$

Therefore, if $\varepsilon = \varepsilon(n)$ decreases in a way that $n\varepsilon(n)^2$ tends to infinity, we could expect the result to hold. In order to prove this rigorously, the Gartner-Ellis theorem can be used. (A proof of this theorem can be found in [12],[1] in a more general framework). Note that if $\varepsilon(n)$ tends to zero too fast, e.g. if $\varepsilon(n) = 1/\sqrt{n}$, which implies $n\varepsilon(n)^2 = 1$, we then hit the central limit theorem regime, and the events measured in theorem 5 are no longer rare, i.e. their probabilities are no longer vanishing. So this result shows that before hitting the central limit theorem regime, a window can be opened where some rare events see their probabilities decaying slower than the large

deviations events, namely at a sub-exponential rate, yet, decaying to zero.

This concludes the section on local geometric properties of the divergence. In the next chapters, discrete memoryless channels are introduced, giving an information theory meaning to the objects treated here. The local approach will then be carefully investigated on different problems, and a crucial point to remember from the current chapter, is that, not only is the divergence locally like a squared distance but also globally it satisfies certain properties of squared distances.

Chapter 3

Discrete Memoryless Channels and Global Geometry

3.1 Channel Model

We denote by $\mathcal{X}$ and $\mathcal{Y}$ the input and output alphabets, which are two finite sets. A communication scheme is defined as follows: at time 1, the transmitter sends an input symbol $x(1) \in \mathcal{X}$, over a channel that randomly generates an output $y(1) \in \mathcal{Y}$, which is observed by the receiver. At time n, the input $x(n)$ is sent, and the output $y(n)$ is received through the same communication channel. We assume the channel to be memoryless (homogenous) and without feedback, i.e., the probability of receiving the output sequence $y = (y(1), \ldots, y(n))$ when the input sequence $x = (x(1), \ldots, x(n))$ has been sent through the channel is given by:

$$\mathbb{P}(y|x) = \prod_{i=1}^{n} W(y(i)|x(i)),$$

where W is a probability transition matrix, i.e., a stochastic matrix of size $\mathcal{X} \times \mathcal{Y}$, whose entries $(W)_{i,j}$ is denoted by $W(j|i)$ and represents the probability of observing the j-th element of $\mathcal{Y}$ when the i-th element of $\mathcal{X}$ is sent (hence the rows of W add up to 1). The length of the sequences, which here can be thought as time, the delay or the number of channel uses, will be called the block length or sometimes simply

the length of the sequences.

This defines a discrete memoryless channel (DMC). Note that a DMC is entirely characterized by its probability transition matrix W, and we will from now on identify DMC's with their respective stochastic matrices. Indeed, we will employ the terminologies, channel, transition probability matrix and stochastic matrix to talk about the same object.

From a communication point of view, DMC's are an abstraction (and simplification) of a sequence of communication layers. Symbols from a finite alphabet are not to be physically transmitted, they are first converted into waveforms by a modulator and then sent through a waveform channel representing a communication link (such as mobile radio) where randomness is added. After this, the demodulator maps the received waveforms into symbols again. Therefore, discrete channels represent the black box embodying those different communication layers. The memoryless assumption is conceivable when "flat fading" assumptions are made on the delay spread mechanism of the communication model, avoiding intersymbol interferences. In this work, we are primarily interested in the design of the encoder (encoding the source symbols into the discrete symbols to be modulated) and decoder (decoding the symbols after demodulation into the sink). Hence we will focus on this discrete channels and consider their ergodic theory. For more details regarding the communication models, cf. [16],[17].

We are interested in sending information reliably through such a DMC, i.e., we want to build an encoding and decoding scheme that ensures a probability of wrong recovery of each message to be sent, as small as desired. Let us assume that only two messages have to be transmitted. First, encode them in the input alphabet language of the channel. In order to have a probability of wrong recovery as small as desired, it is necessary to add redundancies in the encoding (unless some inputs can never be confused by the channel, it will not be possible to just encode each message with

one input symbol and ensure reliable transmission). So the dimension of the input sequences, which we called the block length, has to be exploited. For a finite number of messages, it is then easy to imagine arbitrarily long block length encoding that will ensure reliable communication, e.g. encode each message by repeating a symbol as many times as necessary. One could ask the question of finding the optimal (in the sense of minimizing the error probability) encoding and decoding strategies for a finite number of messages, but another problem is to analyze the optimal scaling between the number of messages that can reliably be sent with respect to the block length they use. We now investigate more formally the second question.

Definition 4. An encoder of block length n and rate R is defined as a mapping E_n such that

$$E_n : m \in \mathcal{M} = \{1, \ldots, M = \lfloor e^{nR} \rfloor\} \mapsto x_m \in \mathcal{X}^n.$$

The image of the encoder defines the code book, denoted by $\mathcal{C}_n = \{x_m\}_{m=1}^{M}$. A decoder is defined as a mapping D_n such that

$$D_n : y \in \mathcal{Y}^n \mapsto m \in \mathcal{M},$$

that is allowed to (and should) depend on the encoder. Note that an encoder (a code book) and a decoder are implicitly defined for a given rate R. Finally, an n-code of rate R is a pair of encoder and decoders of block length n and rate R.

Assumption: we assume that for each block length and rate, the messages $\mathcal{M}$ to be transmitted are equiprobable.

Definition 5. We define the average probability of error for a given block length n, rate R, encoder E_n, decoder D_n and channel W as

$$P_e(E_n, D_n, W) = \frac{1}{M} \sum_{m \in \mathcal{M}} P_{e,m}(E_n, D_n, W)$$

where

$$P_{e,m}(E_n, D_n, W) = \sum_{y:\, D_n(y) \neq m} W^n(y | E_n(m)).$$

We may sometimes replace E_n by $\mathcal{C}_n$ and, when there is no ambiguity, we may replace D_n by a term representing a decoding rule: for example, if ML is used as the argument for the decoder, we mean that the decoder used for a block length n and code book $\mathcal{C}_n$, is the induced maximum likelihood decoder. Equivalently, MMI stands for the maximum mutual information decoding rule. Those specific decoding rules will be investigated in section 3.2.2.

Instead of the average probability of error, the maximal probability of error can be defined by $\max_{m=1}^{M} P_{e,m}(E_n, D_n, W)$. In most problems, good estimates obtained on the average probability of error directly lead to a good estimate for maximal probability of error (i.e., up to some constant factor). For the problem treated here, it is sufficient to work with the average probability of error. Finally, average probability of error and error probability are synonyms.

Definition 6. We say that a rate R is achievable for the channel W if for any $\epsilon > 0$, there exists a block length n, an encoder E_n and decoder D_n of rate at least R such that $P_e(E_n, D_n, W) < \epsilon$.

When the messages are equiprobable, the decoding rule minimizing the average probability of error for a given code book $\{x_m\}_{m=1}^{M}$ of length n and a channel W, is the maximum likelihood decoding, defined as follows.

Definition 7. For a given output sequence y of length n, the maximum likelihood (ML) decoder is defined through the mapping

$$y \mapsto \hat{x}_{\mathrm{ML}}(y) = \arg \max_{x_m,\, m=1,\ldots,M} W^n(y|x_m),$$

and if the maximizer is not unique, an error is declared. This is the convention we use, since ties can be resolved arbitrarily without affecting the forthcoming results. This convention is however convenient, since it preserves the code words symmetry in the problem, which will simplify the error probability analysis. Formally, we should introduce a new symbol in the output alphabet (which represents an error) and is declared when a tie occurs.

Definition 8. Let $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$. We define the joint empirical distribution $P_{x,y}$ of (x, y) by

$$P_{x,y}(a, b) = \frac{|\{i : x(i) = a, y(i) = b\}|}{n}.$$

For a code x_m and a received output y, we sometimes use the notations $P_m = P_{x_m,y}$. In general, the notation P_z where z is a vector denotes the empirical distribution of z, when z is pair of vectors, it denotes the joint empirical distribution, when Z is a random vector, P_Z denotes the empirical distribution of the random vector, which is a random empirical distribution (and similarly for a pair of random vectors).

Note that

$$W^n(y|x_m) = \prod_{i=1}^{n} W(y(i)|x_m(i)) = \prod_{a \in X, b \in \mathcal{Y}} W(b|a)^{nP_{x_m,y}(a,b)} = e^{n\mathbb{E}_{P_{x_m,y}} \log W},$$

where $P_{x_m,y}$ denotes the joint empirical distribution of the vectors x_m and y. Hence

$$\hat{x}_{ML}(y) = \arg \max_{x_m, m=1,\ldots,M} \mathbb{E}_{P_{x_m,y}} \log W. \tag{3.1}$$

Definition 9. For a given output sequence y of length n, the maximum mutual information (MMI) decoder is defined through the mapping

$$y \mapsto \hat{x}_{\text{MMI}}(y) = \arg \max_{x_m, m=1,\ldots,M} I(P_{x_m,y}),$$

and if the maximizer is not unique, an error is declared.

This decoder was first introduced by Goppa.

3.1.1 Random Coding

We start with an informal introduction. Let us assume that there exists a rate $R > 0$ that can be achieved on a given channel which is not a permutation of the identity. Different codes achieving R may then achieve different exponents in the probability of error. Let us also assume that achievable rates are bounded (this will be proved soon,

but it can be expected, as one can show that code books growing faster than exponentially, i.e., $\lim_{n\to\infty} \frac{\log M(n)}{n} = \infty$, must have a probability of error tending to one). We then expect that for the rate R, there could be an optimal error exponent. If so, one may think that in order to achieve the best error exponent for a rate R, the best structure for the code book should be found. What is a good code book structure? A rigorous answer to this question would start by exposing what coding theory is and no general comprehensive answer can probably be found. Informally, a good code should contain code words that are as spread as possible, i.e., as far from each other as the rate R allows. It turns out that trying to formulate this problem rigorously, with a notion of distance, or any explicit mathematical (geometric) construction (which does not proceed by exhaustive search, such as maximal codes) has not been able to get to rates as high as what Shannon's results predicted using randomly generated code books. By drawing the code words randomly under a well chosen distribution, and proving results for the averaged performance, Shannon, Gallager and Berlekamp have been able to show the existence of code books achieving error exponents, that still today, no deterministic "construction" could achieve (notions of "construction" and "complexity" should probably be defined for a more formal discussion). The random coding argument can be seen as an application of what is called the probabilistic method to our problem, although, the argument of Shannon came around the same time as the seminal work of Erdös introducing the probabilistic method ideas.

Definition 10. An iid random code book of distribution $P_{\mathcal{X}} \in M_1(\mathcal{X})$, length n, and rate R (i.e., $M = \lfloor e^{nR} \rfloor$) is defined by the distribution $P_{\mathcal{X}}^{Mn}$, with

$$P_{\mathcal{X}}^{Mn}(x_1, \ldots, x_M) = \prod_{m=1}^{M} P_{\mathcal{X}}^{n}(x_m)$$

where

$$P_{\mathcal{X}}^{n}(x_m) = \prod_{i=1}^{n} P_{\mathcal{X}}(x_m(i)).$$

Definition 11. We denote by $T^n(P_{\mathcal{X}})$ the set of sequences in $\mathcal{X}^n$ whose empirical

distribution is $P_{\mathcal{X}}$ (which is a non empty set only if $P_{\mathcal{X}}$ is such that $nP_{\mathcal{X}}(a) \in \mathbb{Z}_+$, $\forall a \in \mathcal{X}$).

Lemma 1.

$$|T^n(P_{\mathcal{X}})| \doteq e^{nH(P_{\mathcal{X}})}.$$

Definition 12. A code book of length n is said to have a fix composition $P_{\mathcal{X}}$ if all its code words are elements of $T^n(P_{\mathcal{X}})$ (assuming this set is non empty).

Definition 13. A random code book of length n, rate R, and fixed composition $P_{\mathcal{X}}$, is defined by drawing independently and uniformly at random $M = \lfloor e^{nR} \rfloor$ codewords $\{X_m\}_{m=1}^M$ in $T^n(P_{\mathcal{X}})$ (assuming this set is non empty), i.e., by the probability distribution $P_{\mathcal{X}}^{\mathrm{M(T)}}$

$$P_{\mathcal{X}}^{\mathrm{M(T)}}(x_1, \ldots, x_M) = \prod_{m=1}^{M} P_{\mathcal{X}}^{(\mathrm{T})}(x_m) = (P_{\mathcal{X}}^{(\mathrm{T})}(x_1))^M,$$

where

$$P_{\mathcal{X}}^{(\mathrm{T})}(x_1) = |T^n(P_{\mathcal{X}})|^{-1}$$

if $x_1, \ldots, x_M \in T^n(P_{\mathcal{X}})$, and zero otherwise.

Let $\{X_m\}_{m=1}^M$ be an iid random code book with distribution $P_{\mathcal{X}}$, length n and rate R. The error probability is now a random variable $P_e(\{X_m\}_{m=1}^M, D_n, W)$, whose value is $P_e(\{x_m\}_{m=1}^M, D_n, W)$, with probability $P_{\mathcal{X}}^{Mn}(\{x_m\}_{m=1}^M)$ defined in definition 10. Therefore, the expectation of the error probability over this random coding ensemble is given by

$$\mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_e(\{X_m\}_{m=1}^M, D_n, W) = \mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_{e,1}(\{X_m\}_{m=1}^M, D_n, W),$$

which is the expectation of the error probability when transmitting the first codeword (w.l.o.g. we pick the first codeword, but any codewords could have been considered in the right hand side of above equality).

Random coding argument: if for a given n and D_n we have

$$\mathbb{E}_{P_{\mathcal{X}}^{M^n}} P_e(\{X_m\}_{m=1}^M, D_n, W) < \epsilon,$$

then there exists at least one realization of the random code book, $X_1 = x_1, \ldots, X_m = x_m$, that satisfies $P_e(\{x_m\}_{m=1}^M, D_n, W) < \epsilon$. (In fact, many code books are expected to be good.)

The following theorem gives a lower bound on the largest error exponent that can be achieved at rate R on a channel W (i.e. an upper bound on the smallest error probability). This lower bound is proved in [9] using the maximum mutual information (MMI) decoder and in a modified form in [16] using the ML decoder.

Theorem 6. *Let $\{X_m\}_{m=1}^M$ be an iid code book of length n, rate R and distribution $P_{\mathcal{X}}$. We then have*

$$P_e(\{X_m\}_{m=1}^M, ML, W) \doteq e^{-n\left[\inf_{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}}} D(\mu||\mu^j)+\left|R-D(\mu^j||\mu^p)\right|^+\right]}.$$

Hence, for any $\varepsilon > 0$, if $R < C(W)$, where

$$C(W) = \max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} I(P_{\mathcal{X}}, W),$$

there exists a code book $\{x_m\}_{m=1}^M$ of length n, with $M = \lfloor e^{nR} \rfloor$, such that $P(\{x_m\}_{m=1}^M, ML, W) < \varepsilon$. In other words, any R such that $R < C(W)$ is achievable.

These results imply that the highest achievable rate is upper bounded by $C(W)$. One can prove with a converse result (cf. [9],[16]) that the highest achievable rate cannot exceed $C(W)$, which is called the capacity of the channel.

We present a different proof than the one mentioned earlier, that uses the same technique introduced by D. Forney and A. Montanari in [25].

3.2 Error Exponents Estimates

We begin this section by evaluating the averaged error probability of an iid random code book with ML-decoder. Note that using (3.1)

$$\mathbb{E}_{P_{\mathcal{X}}^{M^n}} P_{e,1}(\{X_m\}_{m=1}^M, \mathrm{ML}, W) = \mathbb{P}\{\cup_{m\neq 1}\{\mathbb{E}_{P_{X_1,Y}} \log W \leq \mathbb{E}_{P_{X_m,Y}} \log W\}\}$$

where

$$\mathbb{P}\{X_1 = x_1, \ldots, X_M = x_m, Y = y\} \;=\; W^n(y|x_1) \prod_{m=1}^{M} P_{\mathcal{X}}^n(x_m),$$

and in particular

$$\mathbb{P}\{X_1 = x_1, Y = y\} = \prod_{i=1}^{n} P_{\mathcal{X}}(x_1(i))W(y(i)|x_1(i)), \tag{3.2}$$

$$\mathbb{P}\{X_m = x_m, Y = y\} = \prod_{i=1}^{n} P_{\mathcal{X}}(x_1(i))P_{\mathcal{Y}}(y(i)), \quad \forall m \neq 1 \tag{3.3}$$

with $P_{\mathcal{Y}}$ given through the following definitions

$$\mu^j(a,b) \;=\; P_{\mathcal{X}}(a)W(b|a), \quad \forall a \in X, b \in \mathcal{Y} \tag{3.4}$$

$$P_{\mathcal{Y}}(b) \;=\; \sum_{a\in\mathcal{X}} \mu^j(a,b), \quad \forall b \in \mathcal{Y} \tag{3.5}$$

$$\mu^p(a,b) \;=\; P_{\mathcal{X}}(a)P_{\mathcal{Y}}(b), \quad \forall a \in X, b \in \mathcal{Y}. \tag{3.6}$$

In words, μ^j is the joint distribution between the codeword which is sent, in our case X_1, and the received output, whereas $X_2, \ldots, X_M$ have not been sent, hence are independent of both X_1 and Y. So from the memoryless assumption, we have that with probability one, the following limits hold [1]

$$P_{X_1,Y} \;\overset{n\to\infty}{\longrightarrow}\; \mu^j$$

$$P_{X_m,Y} \;\overset{n\to\infty}{\longrightarrow}\; \mu^p.$$

[1] since we work with finite sets, the limits hold with the topology induced by $\mathbb{R}$; in the more general setting, such as in proposition 11, these limits hold with the weak topology

Remark:

If we work with a constant composition (instead of iid) code book $\{\widetilde{X}_m\}_{m=1}^M$ of distribution $P_{\mathcal{X}}$, we have

$$\begin{aligned}\mathbb{P}_{\text{fix}}\{\widetilde{X}_1 = x_1, \ldots, \widetilde{X}_M = x_m, Y = y\} &= W^n(y|x_1)P_{\mathcal{X}}^{(\mathrm{T})}(x_1) \\ &\doteq \prod_{i=1}^{n} W(y(i)|x_1(i))e^{-nH(P_{\mathcal{X}})} \\ &= e^{-n\mathbb{E}_{P_{X_1,Y}} \log W \circ P_{\mathcal{X}}} \\ &\doteq \mathbb{P}\{X_1 = x_1, \ldots, X_M = x_m, Y = y\},\end{aligned}$$

and in the exponential asymptotic, the estimate we will get for iid or constant composition code book are equivalent.

Note that

$$\{\mathbb{E}_{P_{X_1,Y}} \log W \leq \mathbb{E}_{P_{X_m,Y}} \log W\} \equiv \{\mathbb{E}_{P_{X_1,Y}} \log \frac{\mu^j}{\mu^p} \leq \mathbb{E}_{P_{X_m,Y}} \log \frac{\mu^j}{\mu^p}\}$$

hence

$$\mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_{e,1}(\{X_m\}_{m=1}^M, \mathrm{ML}, W) = \mathbb{P}\{\cup_{m\neq 1}\{\mathbb{E}_{P_{X_1,Y}} \log\log \frac{\mu^j}{\mu^p} \leq \mathbb{E}_{P_{X_m,Y}} \log\log \frac{\mu^j}{\mu^p}\}\}.$$

Let $F : M_1(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be any function and let the random variables

$$F_m = F(P_{X_m,Y}).$$

Using the decoder

$$\hat{x}_F(y) = \arg \max_{x_m, m=1,\ldots,M} F(P_{x_m,y}),$$

we then express the expected error probability as

$$\mathbb{P}\{\cup_{m\neq 1}\{F_1 \leq F_m\}\},$$

recovering the result of the ML-decoder by setting $F(\mu) = L(\mu) = \mathbb{E}_\mu \log \frac{\mu^j}{\mu^p}$.

Proof of (6)*:* We have

$$\begin{aligned} & \mathbb{P}\{\cup_{m\neq 1}\{L_1 \leq L_m\}\} && (3.7)\\ = \; & \mathbb{P}\{\cup_{\gamma}\{L_1 < \gamma \cap \cup_{m\neq 1}\{L_m \geq \gamma\}\}\} && (3.8)\\ \doteq \; & \sup_{\gamma\in\mathbb{R}} \mathbb{P}\{L_1 < \gamma \cap \cup_{m\neq 1}\{L_m \geq \gamma\}\} \end{aligned}$$

The last asymptotic equality results from the following argument. Recall that L_m is the random variable $L(P_{X_m,Y})$, where $P_{X_m,Y}$ is the random empirical distribution of the joint n-dimensional joint random vector (X_m, Y). So any realization of $P_{X_m,Y}$ is an empirical distribution belonging to $M_1^{(n)}(\mathcal{X}\times\mathcal{Y}) = \{P \in M_1(\mathcal{X}\times\mathcal{Y}) | nP(a,b) \in \mathbb{Z}_+, \forall a \in \mathcal{X}, b \in \mathcal{Y}\}$. But

$$|M_1^{(n)}(\mathcal{X}\times\mathcal{Y})| = \binom{n+|\mathcal{X}||\mathcal{Y}|-1}{|\mathcal{X}||\mathcal{Y}|-1},$$

which grows sub exponentially, therefore we can use the union bound to take out the union of γ as a supremum over γ and be tight in the exponential scale. The same argument is used in the asymptotic equality below.

$$\begin{aligned} & \mathbb{P}\{\cup_{m\neq 1}\{L_1 \leq L_m\}\}\\ \doteq \; & \sup_{\gamma\in\mathbb{R}, Q_{\mathcal{Y}}M_1^{(n)}(\mathcal{Y})} \mathbb{P}\{L_1 < \gamma \cap \cup_{m\neq 1}\{L_m \geq \gamma\}, P_Y = Q_{\mathcal{Y}}\} \end{aligned}$$

For any $\varepsilon > 0$, the event $\{P_{X_1} \notin B(P_{\mathcal{X}}, \varepsilon)\}$, where $B(P_{\mathcal{X}}, \varepsilon)$ is a neighborhood of $P_{\mathcal{X}}$, say, the closure of the norm ball of radius ε for the L_1-norm on $M_1(\mathcal{X})$, is vanishing

exponentially fast. Hence, for any $\varepsilon > 0$

$$
\begin{aligned}
& \mathbb{P}\left\{\cup_{m\neq 1}\{L_1 \leq L_m\}\right\} \\
\doteq & \sup_{\gamma, Q_{\mathcal{Y}}} \mathbb{P}\left\{L_1 < \gamma \cap \cup_{m\neq 1}\{L_m \geq \gamma\}, P_Y = Q_{\mathcal{Y}}, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon)\right\} \\
\doteq & \sup_{\gamma, Q_{\mathcal{Y}}} \mathbb{P}\left\{L_1 < \gamma, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon), P_Y = Q_{\mathcal{Y}}\right\} \\
& \quad \cdot \mathbb{P}\left\{\cup_{m\neq 1}\{L_m \geq \gamma\}, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon), P_Y = Q_{\mathcal{Y}}\right\} \mathbb{P}\left\{P_{\mathcal{Y}} = Q_{\mathcal{Y}}\right\}^{-1} \\
\leq & \sup_{\gamma, Q_{\mathcal{Y}}} \mathbb{P}\left\{L_1 < \gamma, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon), P_Y = Q_{\mathcal{Y}}\right\} \\
& \quad \cdot \min(M\mathbb{P}\left\{L_2 \geq \gamma, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon), P_Y = Q_{\mathcal{Y}}\right\}, 1)\mathbb{P}\left\{P_{\mathcal{Y}} = Q_{\mathcal{Y}}\right\}^{-1}
\end{aligned}
$$

where the second equality above uses the independence of the X_m's and the memoryless assumption (i.e., knowing Y or P_Y is equivalent). Using Sanov's theorem, we then get

$$\sup_{\gamma, Q_{\mathcal{Y}}} \mathbb{P}\left\{L_1 < \gamma, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon), P_Y = Q_{\mathcal{Y}}\right\} \tag{3.9}$$

$$\cdot \min(M\mathbb{P}\left\{L_2 \geq \gamma, P_{X_1} \in B(P_{\mathcal{X}}, \varepsilon), P_Y = Q_{\mathcal{Y}}\right\}, 1)\mathbb{P}\left\{P_{\mathcal{Y}} = Q_{\mathcal{Y}}\right\}^{-1} \tag{3.10}$$

$$\doteq \exp(-n \inf_{\gamma, Q_{\mathcal{Y}}} [\inf_{\substack{\mu:\, \mu_{\mathcal{X}} \in B(P_{\mathcal{X}}, \varepsilon), \mu_{\mathcal{Y}} = Q_{\mathcal{Y}} \\ L(\mu) < \gamma}} D(\mu || \mu^j) \tag{3.11}$$

$$+ \left|R - \inf_{\substack{\mu:\, \mu_{\mathcal{X}} \in B(P_{\mathcal{X}}, \varepsilon), \mu_{\mathcal{Y}} = Q_{\mathcal{Y}} \\ L(\mu) \geq \gamma}} D(\mu || \mu^p)\right|^+ - D(Q_Y || P_{\mathcal{Y}})]) \tag{3.12}$$

We now argue that in above expression, both infimums taken over the distribution μ are achieved for the same distribution. This requires two checks, which we outline here. Note that using continuity arguments, we can think of ε as being 0, another way of avoiding to have the ε neighborhood is to work with fixed composition instead of iid random codes. Let $B = \{\mu \text{ s.t. } \mu_{\mathcal{X}} = P_{\mathcal{X}}, \mu_{\mathcal{Y}} = Q_{\mathcal{Y}}, L(\mu) < \gamma\}$. The optimal γ is such that $\mu^j \notin B$ and $\mu^p \notin B^c$, hence, we can replace B and B^c by ∂B in both infimums. We now deal with two I-projections onto the intersections of the same linear families, the two marginal constraints and ∂B, a linear family of direction L and shift γ. Note that μ^j and μ^p both belong to the same exponential family or-

thogonal to the linear family of direction $\log \frac{\mu^j}{\mu^p}$, which is precisely the direction of L, since $L(\mu) = \mathbb{E}_\mu \log \frac{\mu^j}{\mu^p}$. Therefore, from corollary 1, the two I-projections appearing in equation (3.12) must be the same distribution.

Note: the last step uses the specific structure of the ML-decoder, i.e., the fact that $F(\mu) = \mathbb{E}_\mu \log \frac{\mu^j}{\mu^p}$ is a linear family precisely orthogonal to the exponential family connecting μ^j to μ^p. Any decoder that does not necessarily have this strong orthogonality property, but that still has the same I-projections for both infimums, would achieve the same exponent of the ML-decoder, which is given by

$$\exp\left(-n\left[\inf_{\mu \text{ s.t. } \mu_{\mathcal{X}}=P_{\mathcal{X}}} D(\mu||\mu^j) + \left|R - D(\mu^j||\mu^p)\right|^+\right]\right).$$

3.2.1 Exponent at Capacity

In this section, we derive an upper bound to the averaged error probability which is in general loser than the one of previous section, except for rates close to capacity. As opposed to (3.8), we will use the following upper bound on the error probability: we first pick a $\gamma \in \mathbb{R}$ and notice that

$$\begin{aligned} & \mathbb{P}\left\{\cup_{m\neq 1}\{L_1 \leq L_m\}\right\} \\ \leq\ & \mathbb{P}\left\{\{L_1 < \gamma\} \cup \{\cup_{m\neq 1}\{L_m \geq \gamma\}\}\right\} \end{aligned}$$

and this is true for any $\gamma \in \mathbb{R}$. This upper bound is equivalent to the expression of the probability of error when using the suboptimal decoding rule that declares the code word whose likelihood function is more than a threshold given by γ, and if there are more than one such code word, declares an error. Using the union bound, we get

$$\begin{aligned} & \mathbb{P}\left\{\cup_{m\neq 1}\{L_1 \leq L_m\}\right\} \\ \leq\ & \mathbb{P}\{L_1 < \gamma\} + \min(M\mathbb{P}\{L_2 \geq \gamma\}, 1) \end{aligned}$$

and from Sanov's theorem, this gives the following exponent.

$$\sup_{\gamma\in\mathbb{R}} \inf_{\substack{\mu:\,\mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}}\\ L(\mu)<\gamma}} D(\mu||\mu^j) \wedge \inf_{\substack{\mu:\,\mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}}\\ L(\mu)\geq\gamma}} \left|R - D(\mu||\mu^p)\right|^+. \tag{3.13}$$

We now show that this upper bound becomes tight when considering R close to $D(\mu^j||\mu^p) = I(P_{\mathcal{X}}, W)$: we know that $\inf_{\substack{\mu:\,\mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}}\\ L(\mu)\geq D(\mu^j||\mu^p)}} D(\mu||\mu^p)$ is achieved at μ^j, this is a consequence of the theorem 5.30. When $R = D(\mu^j||\mu^p) - \varepsilon$, we can take $\gamma = R$ and the exponent given in (3.13) becomes

$$o_+(1) \wedge [D(\mu^j||\mu^p) - o_+(1)],$$

where $o_+(1) > 0$ and $\lim_{\varepsilon\searrow 0} o_+(1) = 0$.

3.2.2 Global Geometry of Decoders at Capacity

We now consider decoders that maximize score functions that are not necessarily the log likelihood.

Definition 14. Let

$$F:\ M_1(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R},$$

a decoder is said to maximize the score function F if it is of the form

$$\hat{x}_F(y) = \arg\max_{x_m,\, m=1,\ldots,M} F(P_{x_m,y}),$$

and if the maximizer is not unique, an error is declared.

We keep the convention that x_1 is sent and y is received. For an iid code book of distribution $P_{\mathcal{X}}$, we define the random variables

$$F_m = F(P_{X_m,Y}),$$

the error probability (averaged over the random code book) is then given by

$$\mathbb{P}\{\cup_{m\neq 1}\{F_1 \leq F_m\}\}.$$

When $F(P_{x_m,y}) = \mathbb{E}_{P_{x_m,y}} \log \frac{\mu^j}{\mu^p}$, we saw two ways to get upper bounds on the error probability (see previous section) and in this section we are interested in the highest achievable rates only, and hence the second technique is sufficient.

Proposition 2. *The exponent of the error probability averaged over an iid code book of distribution $P_{\mathcal{X}}$ and with decoder maximizing the score function F is lower bounded by*

$$\sup_{\gamma\in\mathbb{R}} \inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}} \\ F(\mu)<\gamma}} D(\mu||\mu^j) \wedge \inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}} \\ F(\mu)\geq\gamma}} \left|R - D(\mu||\mu^p)\right|^+. \tag{3.14}$$

Therefore the capacity is lower bounded by

$$\sup_{P_{\mathcal{X}}\in M_1(P_{\mathcal{X}})} \inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}} \\ F(\mu)\geq F(\mu^j)}} D(\mu||\mu^p). \tag{3.15}$$

Proof. The first part follows from the fact that (3.13) did not depend on the fact that L was the ML-decoder. For a fixed $P_{\mathcal{X}}$, if

$$R < \inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},(\mu)_{\mathcal{Y}}=P_{\mathcal{Y}} \\ F(\mu)\geq F(\mu^j)}} D(\mu||\mu^p),$$

there exists $\varepsilon > 0$, such that by choosing $\gamma = L(\mu^j) - \varepsilon$, the second term in (3.14) will be strictly positive, and by the definition of γ, the first term too. □

Definition 15. We define the mismatched mutual information of an input P_X, a channel W, and decoding metric F, by

$$\sup_{P_{\mathcal{X}}\in M_1(P_{\mathcal{X}})} \inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},\mu_{\mathcal{Y}}=P_{\mathcal{Y}} \\ \mathbb{E}_\mu F\geq \mathbb{E}_{\mu^j} F}} D(\mu||\mu^p). \tag{3.16}$$

From proposition 2, previous expression represents an achievable rate, obtained

from a random code book of distribution $P_{\mathcal{X}}$ (constant composition or iid), when the true channel is W and the decoder maximizes the metric F. It has been shown that this bound is tight (for the achievable rates), when restricted to code books drawn from a random ensemble but otherwise it can be loose, cf. [11],[23]. It is always tight, it determines the capacity, for binary inputs, cf. [4]. Figure 3-1 represents this expression. The region delimitated by the vertical plane represents the constraint region appearing in the I-projection for the capacity of a maximum likelihood decoder tuned to the channel. As illustrated, the I-projection of μ^p onto this region is μ^j the distance is $D(\mu^j||\mu^p) = I(\mu^j)$. This arises since both distributions are contained in the exponential family of direction $L = \log \frac{\mu^j}{\mu^p}$. If the maximum likelihood decoder is mismatched to the channel of communication, the constraint set appearing in the I-projection is not perpendicular to the exponential family passing though μ^j and μ^p and the projection's distance is less than $D(\mu^j||\mu^p)$, as illustrated.

This proposition gives the following sufficient condition to ensure that a decoder achieves same highest rate as the optimal ML-decoder.

Proposition 3. *If a decoder maximizes a score function F and*

$$\sup_{P_{\mathcal{X}} \in M_1(P_{\mathcal{X}})} \inf_{\substack{\mu: \mu_{\mathcal{X}} = P_{\mathcal{X}}, \mu_{\mathcal{Y}} = P_{\mathcal{Y}} \\ F(\mu) \geq F(\mu^j)}} D(\mu||\mu^p) = D(\mu^j||\mu^p), \tag{3.17}$$

the decoder is optimal.

Equivalently, we can rewrite previous propositions as follows. Let $B_D(\mu^p, D(\mu^j||\mu^p))$ be the set of all distributions on $\mathcal{X} \times \mathcal{Y}$ for which $D(\mu||\mu^p) < D(\mu^j||\mu^p)$, we then have the following result.

Proposition 4. *Any decoder maximizing a score function F which satisfies*

$$\{F(\mu) \geq F(\mu^j)\} \subseteq B_D(\mu^p, D(\mu^j||\mu^p))^c$$

is optimal.

Example: $F(\mu) = I(\mu) = D(\mu||\mu^p)$.

This result is illustrated in figure 3-2. As opposed to the case of a mismatched

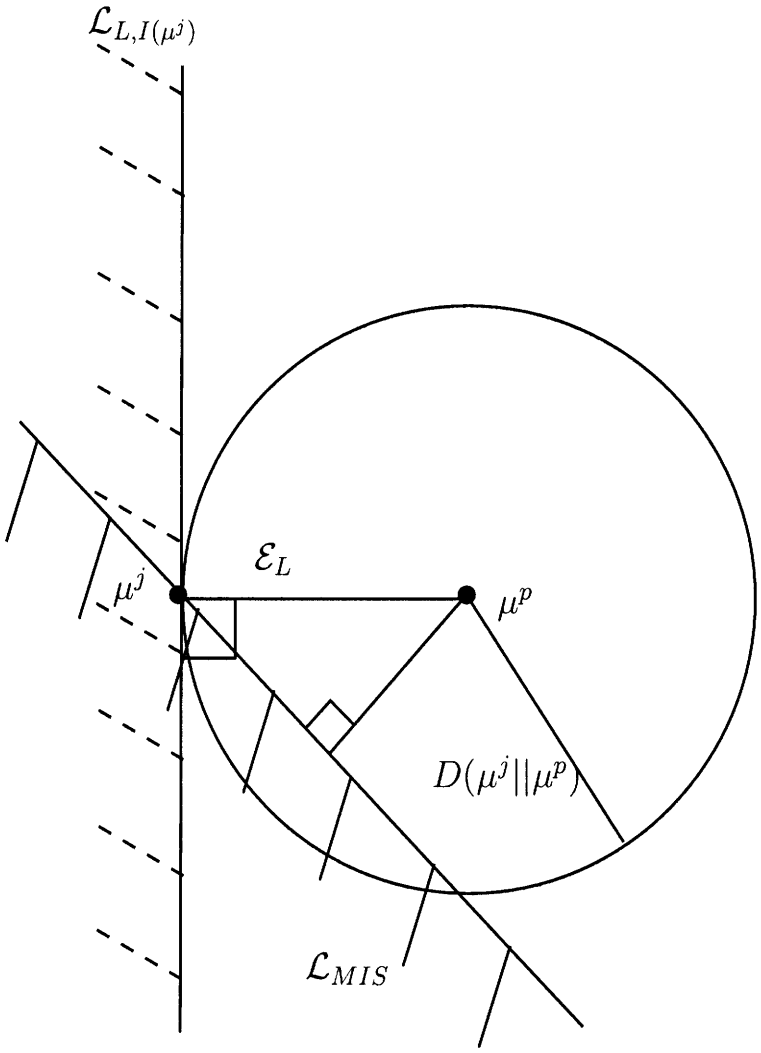

Figure 3-1: Mismatched mutual information

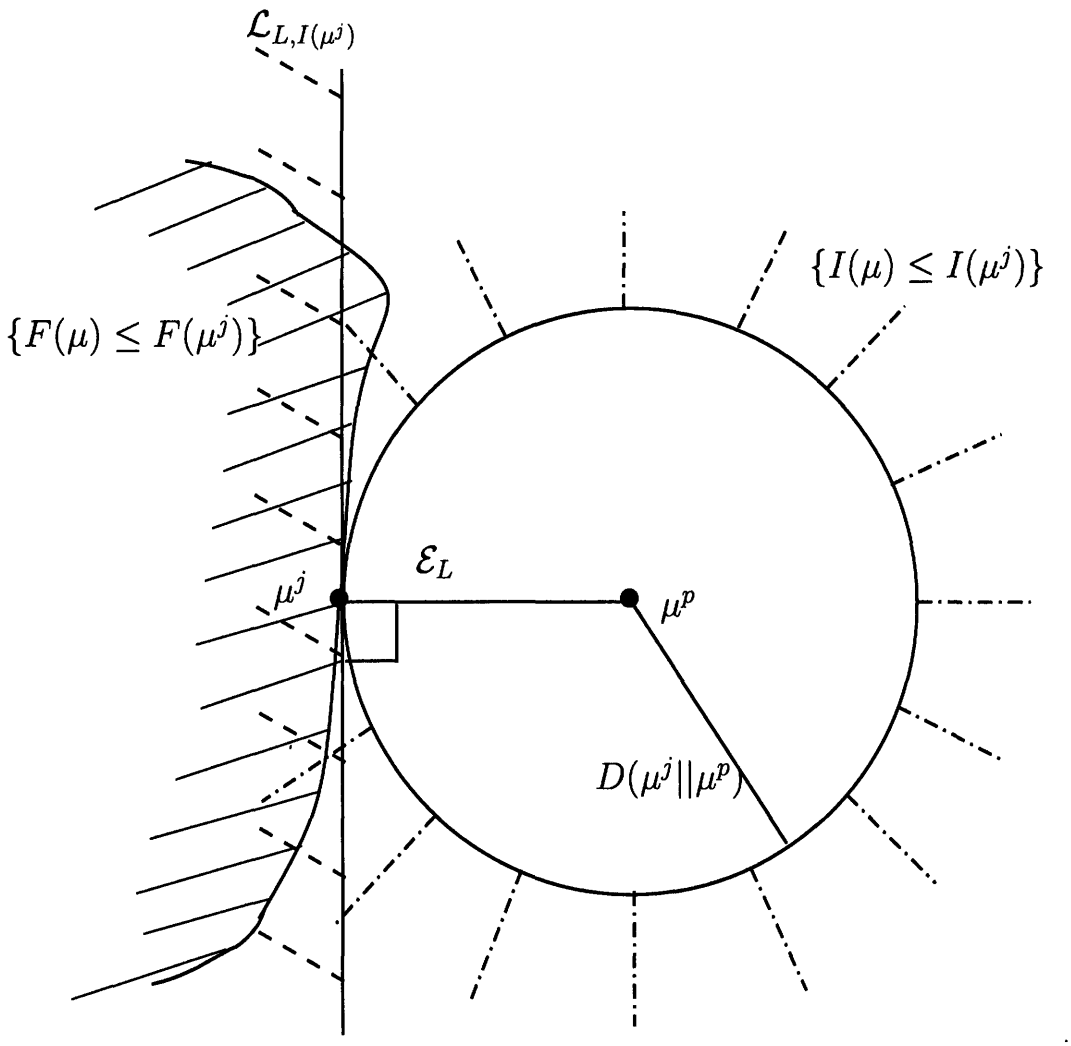

Figure 3-2: Optimal decoders at capacity

decoder illustrated in figure 3-1, the three I-projecton regions illustrated in this picture represents decoder that are achieving capacity, since the I-projection of μ^p onto each regions is at μ^j. This is always the case if the regions are excluding the sphere of radius $D(\mu^j||\mu^p) = I(\mu^j)$ centered at μ^p. Indeed, the maximum mutual information region is equal to the complement of this sphere.

Chapter 4

Very Noisy Transformation and Local Channel Geometry

In this section, we use the ideas developed in section 2.2 to analyze Discrete Memoryless Channels. We saw in the previous chapter that the performance of the considered communication schemes are evaluated through the optimization (alternated minimizations and maximizations) of divergence expressions under some constrained (often linearly constrained) probability distributions. In this chapter, we consider very noisy channels, and as we will see, this setting will bring μ^j and μ^p close to each other no matter what the input distribution is and will allow us to use the local results presented in section 2.2. The intuitive geometry described for the local setting will hence come into the picture. This approach will be particularly useful to design good decoders (cf. chapter 5) since most decoders, e.g. ML, are "functions of the channel". In chapter 6 we will perform a local analysis of input distributions that in turn will be useful to design encoders. In both cases, the same technical ideas of section 2.2 will be applied. In section 4.1.1,the different information theoretic expressions encountered till now, such as mutual information and mismatched mutual information, are analyzed in the very noisy setting. We also mention how the very noisy channels act on other kinds of channels such as compound and broadcast channels. The problem of universal decoding over a compound channel will then be investigated in details in chapter 5, where the very noisy channels will help us getting to general results.

4.1 Very Noisy Channels

Roughly speaking, we want to consider channels which are weakly depending on the input that is sent. If the transition probabilities of observing any output does not depend on the input, i.e., the transition probability matrix has constant columns, we have a "pure noise" channel. So a very noisy channel should be somehow close to such a pure noise channel. Although we will use the very noisy results to inspire the proof of global results, we will always give formal proofs of the global results, when they were achieved.

Definition 16. *Terminology*
We say that W_ε is a very noisy channel with limiting distribution $P_\mathcal{Y}$ and direction L if

$$W_\varepsilon(y|x) = P_\mathcal{Y}(y)(1 + \varepsilon L(x,y)), \quad \varepsilon \ll 1$$

where $P_\mathcal{Y} \in M_1(\mathcal{Y})$ and $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfies for any $x \in \mathcal{X}$

$$\sum_{y \in \mathcal{Y}} L(x,y) P_\mathcal{Y}(y) = 0. \tag{4.1}$$

For a given direction and limiting distribution, we refer to the *VN transformation* of W by the mapping $W \rightsquigarrow W_\varepsilon(P_\mathcal{Y}, L)$. For a given limiting distribution and several directions, the VN transformation of an expression containing several channels is obtained by taking the VN transformation of each channel for the common limiting distribution[1] and their respective directions. If $E(W_1, \ldots, W_k)$ denotes an expression depending on k channels, we define its very noisy limit by

$$\lim_{\varepsilon \searrow 0} \frac{2}{\varepsilon^2} E(W_{1,\varepsilon}(P_\mathcal{Y}, L), \ldots, W_{k,\varepsilon}(P_\mathcal{Y}, L))$$

[1] in the present work, we restricted ourself to consider common limiting distributions, however different limiting distributions can be considered too

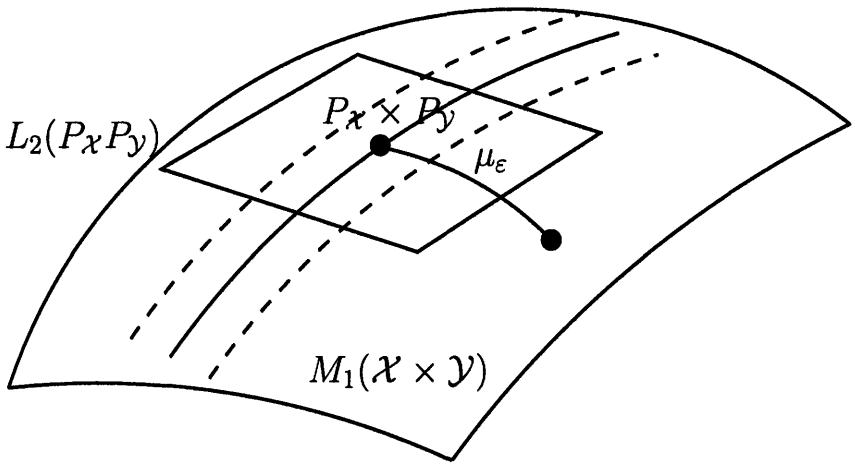

Figure 4-1: Very noisy channels and neighborhoods

and we use the notation $\xrightarrow{VN}$ to denote

$$E(W_{1,\varepsilon}(P_{\mathcal{Y}}, L), \ldots, W_{k,\varepsilon}(P_{\mathcal{Y}}, L)) \xrightarrow{VN} \lim_{\varepsilon \searrow 0} \frac{2}{\varepsilon^2} E(W_{1,\varepsilon}(P_{\mathcal{Y}}, L), \ldots, W_{k,\varepsilon}(P_{\mathcal{Y}}, L)).$$

In informal discussions, we will often talk about VN transformations without explicitly mentioning the directions. If the expression considered through the VN transformation has a name assigned to it, e.g. mutual information or mismatched mutual information, the very noisy limit is then named by preceding the original name with "very noisy", such as very noisy mutual information and very noisy mismatched mutual information. Finally the initials VN refer to "very noisy".

In terms of stochastic matrices, $P_{\mathcal{Y}}$ leads to a probability transition matrix that does not depend on the value of x, i.e., a stochastic matrix with constant columns and W_ε is a perturbation of this specific matrix in a specific direction. Hence, very noisy channels are living in the neighborhoods of stochastic matrices with constant columns.

If the input distribution is $P_{\mathcal{X}}$, the induced output distribution through such a very noisy channel at any ε is given by $P_{\mathcal{Y}}(y)(1 + \varepsilon \bar{L}(y))$, where $\bar{L}(y) = \sum_x L(x, y) P_{\mathcal{X}}(x)$.

Therefore, the joint distribution induced by the input distribution $P_{\mathcal{X}}$ and the channel W_ε is

$$\mu^j(x,y) = P_{\mathcal{X}}(x)W_\varepsilon(y|x) = P_{\mathcal{X}}(x)P_{\mathcal{Y}}(y)(1+\varepsilon L(x,y)) \tag{4.2}$$

and the product measure between the input and output marginals is

$$\mu^p(x,y) = P_{\mathcal{X}}(x)W_\varepsilon(y|x) = P_{\mathcal{X}}(x)P_{\mathcal{Y}}(y)(1+\varepsilon \bar{L}(y)). \tag{4.3}$$

As expected, both distributions are local perturbation of the distribution $P_{\mathcal{X}} \times P_{\mathcal{Y}}$. Hence, as illustrated in figure ??, for a given input distribution, the very noisy channels set the induced joint and product distributions in neighborhoods of the following subset of $M_1(\mathcal{X} \times \mathcal{Y})$ containing the product measures

$$\{\mu \in M_1(\mathcal{X} \times \mathcal{Y}) | \mu = P_{\mathcal{X}} \times P_{\mathcal{Y}}, P_{\mathcal{X}} \in M_1(\mathcal{X}), P_{\mathcal{Y}} \in M_1(\mathcal{Y})\},$$

which is in matrix notations parametrized by $\mu = \text{diag}(P_{\mathcal{X}})\mathbb{1}\,\text{diag}(P_Z)$. With this remark, we are ready to use our results developed in section 2.2.

4.1.1 Very Noisy Information Theoretic Expressions

Very noisy Mutual Information

Let us start by analyzing how the mutual information of such channels behave. From (2.6) using the distribution (4.2) and (4.3), we get the following fact.

Fact: For any $P_{\mathcal{X}} \in M_1(\mathcal{X})$, $P_{\mathcal{Y}} \in M_1(\mathcal{Y})$ and L satisfying (4.1), we have

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2} I(P_{\mathcal{X}}, P_\varepsilon) = I_{VN}(P_{\mathcal{X}}, P_{\mathcal{Y}}, L),$$

where

$$I_{VN}(P_{\mathcal{X}}, P_{\mathcal{Y}}, L)$$
$$= \frac{1}{2}\sum_{a,b}(L(a,b) - \sum_{c} L(c,b)P_{\mathcal{X}}(c))^2 P_{\mathcal{X}}(a)P_{\mathcal{Y}}(b),$$

which is strictly positive as long as L is not independent of the $\mathcal{X}$-component. We thus have

$$I(P_{\mathcal{X}}, P_{\varepsilon}) = I_{VN}(P_{\mathcal{X}}, P_{\mathcal{Y}}, L)\varepsilon^2 + o(\varepsilon^2).$$

Previous expansions have been known for long (cf. [20],[16] and references therein). We now introduce different ways to express the very noisy mutual information. We give three expressions for I_{VN}, which will all tell us something different. We first need some notation.

Notations: for $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we define $\bar{L} : y \to \sum_x L(x,y)P_{\mathcal{X}}(x)$, $\tilde{L} = L - \bar{L}$, $L_x : y \to L(x,y)$ and $L_y : x \to L(x,y)$.

Fact:

$$I_{VN}(P_{\mathcal{X}}, P_{\mathcal{Y}}, L) = \frac{1}{2}||\tilde{L}||^2_{P_{\mathcal{X}}P_{\mathcal{Y}}} \tag{4.4}$$
$$= \frac{1}{2}\sum_x P_{\mathcal{X}}(x)||L_x - \bar{L}||^2_{P_{\mathcal{Y}}} \tag{4.5}$$

- The first expression relates the VN mutual information to the squared norm (under the product measure $P_{\mathcal{X}} \times P_{\mathcal{Y}}$) of the centered direction $\tilde{L}$, which belongs to $M_0(P_{\mathcal{X}} \times P_{\mathcal{Y}})$. Hence the VN mutual information is the energy of an element in

  $$L_2(M_0(P_{\mathcal{X}}, P_{\mathcal{Y}}), P_{\mathcal{X}} \times P_{\mathcal{Y}})$$

  where

  $$M_0(P_{\mathcal{X}}, P_{\mathcal{Y}}) =$$
  $$\{v \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}} | \sum_x v(x,y)P_{\mathcal{X}}(x) = \sum_y v(x,y)P_{\mathcal{Y}}(y) = 0\},$$

and the inner product is

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{P_{\mathcal{X}} \times P_{\mathcal{Y}}}.$$

The simple fact of recognizing this mathematical structure and expressing very noisy objects by means of this inner product, will greatly simplify the VN limit expressions and introduce a geometrical framework for our problems. The next section will illustrate this further.

- The second expression gives us some intuition on what is happening when we are optimizing the VN mutual information on the input distributions to get the capacity. From the KT conditions, the optimal input distribution should produce a $\bar{L}^* = \sum_x L_x P^*_{\mathcal{X}}(x)$ such that when $P^*_{\mathcal{X}}(x) \neq 0$, the distances from $\bar{L}^*$ to the L_x's are balanced, and otherwise the distances are smaller. In other words, $\bar{L}^*$ is the center of the smallest sphere containing all the L_x's, and its radius is the capacity, where we now work with the geometry of $\mathcal{L}_2(M_0(P_{\mathcal{Y}}), P_{\mathcal{Y}}))$ (see figure 4-2). This directly gives us an equivalent way of expressing the very noisy capacity (VNCa):

$$C_{VN} = \frac{1}{2} \min_{P_{\mathcal{X}}} \max_x ||L_x - \bar{L}||_{P_{\mathcal{Y}}}. \tag{4.6}$$

Note that these geometrical results have an equivalent formulation in the general setting with divergences. Of course, since the divergence is not symmetric, it now matters how the arguments are evaluated. In particular, the KT-conditions for the input distribution maximizing the mutual information are:

$$D(P_{Y|X=x}||P^*_{\mathcal{Y}}) = \gamma, \quad \text{when } P^*_{\mathcal{X}}(x) \neq 0 \tag{4.7}$$

$$< \gamma, \quad \text{otherwise.} \tag{4.8}$$

And we also have:

$$C = \min_{P_{\mathcal{Y}}} \max_{x \in \mathcal{X}} D(P_{Y|X=x}||P_{\mathcal{Y}}).$$

With the next results we will see that the previous VN structure is not specific to

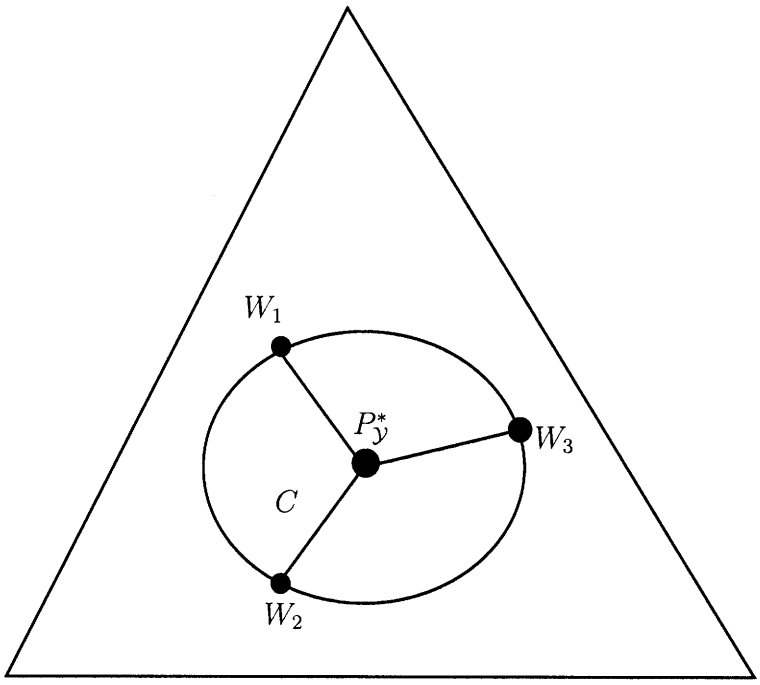

Figure 4-2: Optimal input distribution

just the mutual information of VN channels, and in fact, a more general structure is present for several information theoretic VN quantities.

Mismatched Mutual Information

We saw in section 2.2 that locally, the I-projection turns into a norm-space projection. In this section, we are interested in the specific I-projection given by the mismatched mutual information in (3.15).

Proposition 5. *Let $W_{0,\epsilon} = P_{\mathcal{Y}}(1+\epsilon L_0)$ be the VN transformation of W_0 and let us decode with the metric $d = \log W_{1,\epsilon}$, where $W_{1,\epsilon} = P_{\mathcal{Y}}(1+\epsilon L_1)$. Then, taking then the VN transformation of the mismatched mutual information, we get the following very noisy limit*

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{1}{\epsilon^2}&\inf_{\substack{\mu:\,\mu_{\mathcal{X}}=P_{\mathcal{X}},\,\mu_{\mathcal{Y}}=(\mu_0)_{Y,\epsilon}\\ \mathbb{E}_\mu \log W_{1,\epsilon}\geq \mathbb{E}_{\mu_{0,\epsilon}}\log W_{1,\epsilon}}} D(\mu||\mu^p_{0,\epsilon})\\ &=\frac{1}{2}\inf_{\substack{V:\,\bar{\tilde{V}}=\bar{\tilde{V}}=0\\ \langle V,\tilde{L}_1\rangle\geq\langle\tilde{L}_0,\tilde{L}_1\rangle}}||V||^2=\frac{1}{2}\frac{\langle\tilde{L}_0,\tilde{L}_1\rangle^2}{||\tilde{L}_1||^2}\end{aligned}$$

Proof. For each ε, the minimizer μ_ε can be expressed as

$$\mu_\varepsilon = P_{\mathcal{X}}P_{\mathcal{Y}}(1+\varepsilon L)$$

with $L \in M_0(P_{\mathcal{Y}})$, i.e.,

$$\bar{\bar{L}} = 0, \tag{4.9}$$

which ensures that the first constraint, $\mu_{\mathcal{X}} = P_{\mathcal{X}}$, is satisfied. Moreover,

$$\mu_{\mathcal{Y},\varepsilon} = P_{\mathcal{Y}}(1+\varepsilon\bar{L}),$$

hence, the second constraint becomes

$$\bar{L} = \bar{L}_0. \tag{4.10}$$

Finally, the third constraint is given by

$$\begin{aligned} & \sum_{a\in\mathcal{X},b\in\mathcal{Y}} P_{\mathcal{X}}(a)P_{\mathcal{Y}}(b)(1+\varepsilon L(a,b))[\log P_{\mathcal{Y}} + \log(1+\varepsilon L_1(x,y))] \\ \geq & \sum_{a\in\mathcal{X},b\in\mathcal{Y}} P_{\mathcal{X}}(a)P_{\mathcal{Y}}(b)(1+\varepsilon L_0(a,b))[\log P_{\mathcal{Y}} + \log(1+\varepsilon L_1(x,y))]. \end{aligned}$$

Using $\bar{L} = \bar{L}_0$, the terms with $\log P_{\mathcal{Y}}$ cancel each other and by taking ε small enough, we have

$$\begin{aligned} \log(1+\varepsilon L_0(x,y)) &= \varepsilon L_0(x,y) - \varepsilon^2 \frac{L_0(x,y)^2}{2} + o(\varepsilon^2), \\ \log(1+\varepsilon L_1(x,y)) &= \varepsilon L_1(x,y) - \varepsilon^2 \frac{L_1(x,y)^2}{2} + o(\varepsilon^2). \end{aligned}$$

so that the third constraint is

$$\langle L, L_1\rangle \geq \langle L_0, L_1\rangle + o(1), \tag{4.11}$$

where we used the fact that $\bar{\bar{L}}_0 = \bar{\bar{L}}_1 = 0$. Therefore, from (4.9), (4.10) and (4.11), the overall limiting optimization is given by

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon^2} \inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},\, \mu_{\mathcal{Y}}=(\mu_0)_{Y,\epsilon} \\ \mathbb{E}_\mu \log W_{1,\epsilon} \geq \mathbb{E}_{\mu_{0,\epsilon}} \log W_{1,\epsilon}}} D(\mu||\mu^p_{0,\epsilon}) = \frac{1}{2} \inf_{\substack{L:\, \bar{\bar{L}}=0, \bar{L}=\bar{L}_0 \\ \langle L, L_1\rangle \geq \langle L_0, L_1\rangle}} \left\| L - \bar{L}_0 \right\|^2.$$

Setting $V = L - \bar{L}_0$, the third constraint becomes

$$\langle V, L_1\rangle \geq \langle L_0, L_1\rangle - \langle \bar{L}_0, \bar{L}_1\rangle,$$

but

$$\langle L_0, L_1\rangle - \langle \bar{L}_0, \bar{L}_1\rangle = \langle \widetilde{L}_0, \widetilde{L}_1\rangle$$

and, since $\bar{V} = \bar{\bar{V}} = 0$,

$$\langle V, L_1\rangle = \langle V, \widetilde{L}_1\rangle.$$

This proves the first equality of the proposition and the second one is trivial, since

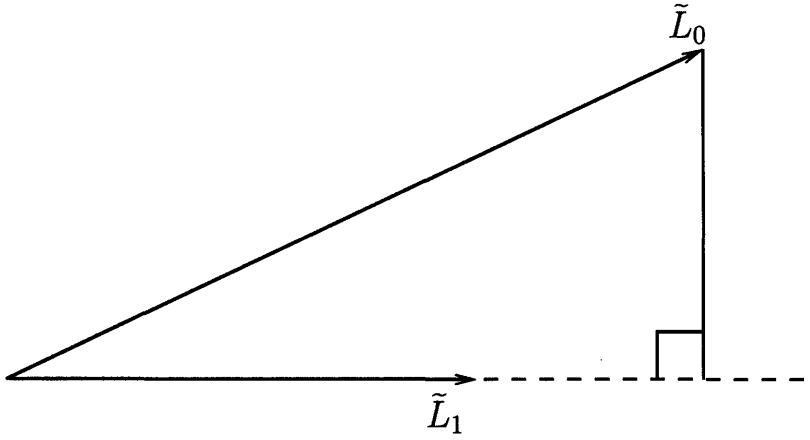

Figure 4-3: Very noisy mismatched mutual information

we now have V and $\widetilde{L}_1$ in the same space; the minimization leads to the projection given by

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_1 \rangle^2}{||\widetilde{L}_1||^2}.$$

□

This result says that the mismatched mutual information obtained when decoding with the mismatched metric $\log W_{1,\epsilon}$, whereas the true channel is $W_{0,\epsilon}$, is approximately the projection (squared norm) of the true channel direction $\widetilde{L}_0$ onto the mismatched direction $\widetilde{L}_1$. This result gives an intuitive picture of the mismatched mutual information (cf. figure 4-3). As expected, if the decoder is matched, i.e., $\widetilde{L}_0 = \widetilde{L}_1$, the projection squared norm is $||\widetilde{L}_0||^2$, which is the very noisy mutual information of $\widetilde{L}_0$, and the more orthogonal $\widetilde{L}_1$ is to $\widetilde{L}_0$, the worse the mismatched decoding rule is.

By choosing $N = 1$ and $W_1 = W_0$, we recover the VN mutual information.

Corollary 2.

$$\lim_{\epsilon \searrow 0} \frac{1}{\epsilon^2} I(P_{\mathcal{X}}, P_\epsilon) = \frac{1}{2} ||\tilde{L}_0||^2. \tag{4.12}$$

Channels Concatenation

Let $P_\epsilon = P_\epsilon(P_\mathcal{Y}, L)$ be a very noisy channel with limiting distribution $P_\mathcal{Y}$ and direction L.

Proposition 6.

- *If* $Q_\epsilon = P_\epsilon M$, *then* $Q_\epsilon = Q_\epsilon(P_Z, K)$, *where* $P_Z(z) = \sum_y M(z|y)P_\mathcal{Y}(y)$ *and* $K(z|y) = P_Z(z)^{-1}\sum_y P_\mathcal{Y}(y)M(z|y)L(y|x)$.
- *If* $R_\epsilon = NP_\epsilon$, *then* $R_\epsilon = R_\epsilon(P_\mathcal{Y}, NL)$.

This result is useful when considering multi-user information theory problems, such as Broadcast channels. If one considers the following situation:

$$U \xrightarrow{N} X \to Y$$

where $X \to Y$ is a VN channel, and $U \to X$ is a pre-encoding channel and U is an auxiliary random variable. Note that the distributions of the pre-encoder must satisfy

$$P_U N = P_X.$$

The overall channel $U \to Y$ is then VN as well, and from previous proposition, its limiting distribution is $P_\mathcal{Y}$ and the directions is NL.

4.1.2 Inner Product Space Structure

In this section, we discuss in an informal way the result presented in previous sections. Roughly speaking, what we observed in previous sections can be summarized with the following schematic mappings. When dealing with a fixed input distribution, the different information theoretic quantities investigated earlier, which are divergence optimizations over the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$, become, in the VN limit, quantities defined in the inner product space mentioned below and illustrated

in figure 4-1

$$(M_1(\mathcal{X} \times \mathcal{Y}), D(\cdot||\cdot)) \xrightarrow{VN} L_2(M_0(P_{\mathcal{X}} \times P_{\mathcal{Y}}), \langle \cdot, \cdot \rangle_{P_{\mathcal{X}} \times P_{\mathcal{Y}}}). \tag{4.13}$$

We saw rigorous statement of such situation in the specific cases of section 4.1.1. We also saw that when the input optimization is carried out, such as when optimizing the input distribution for the mutual information to get the capacity, the VN limit mapped the problem into a geometrical problem defined in the space

$$L_2(M_0(P_{\mathcal{Y}}^*), \langle \cdot, \cdot \rangle_{P_{\mathcal{Y}}^*}),$$

where $P_{\mathcal{Y}}^*$ denotes the induced optimal output distribution. For example, the capacity achieving input distribution is found by looking for the smallest circle containing all the L_x's (cf. (4.6)).

It is of course possible to construct expressions whose very noisy limit does not live in the space given in (4.13). However, in the problems of section 5.1, we will see that sitting in this inner product space is actually of further importance, namely, in the problem considered, it is only when the VN limit of the quantities of interest ends up in the space described by 4.13, that the solution turns out to be optimal.

Section 4.2 discusses how the VN transformation can be useful.

4.2 Use of the Very Noisy Transformation

It is common, in Information theory, to analyze problems by considering their limiting regime with respect to some specific parameter (example: high/low signal to noise ratio, blocklength). In that regards, the VN transformation can be seen as a limiting analysis, which is tight when the channel is very noisy. Shannon's results for channel coding say that the error probability can be made as small as desired by taking the block length large enough, and the rate of decay is shown to be exponential with n. In a similar fashion, we saw that the very noisy capacity is tight in a quadratic scale of the parameter ϵ. In both cases, the discussion of determining how large

the block length should be, or how noisy the channel should be in order to trust our estimates, is of a different kind and not of interest here. However, these two situations are also different. If no constraint is imposed, the block length is a parameter that can be increased as much as desired by the transmitter, whereas the channel noise structure is given by nature. But there is another point of view on how to use the very noisy analysis. We are not interested in the very noisy regime per se: what we want to acquire through it, is a better intuition. We can see it as simplification of a considered problem, where a nice geometrical insight can be acquired, and that can lead to suggestions for the solution structure of the general problem. This is in fact the main attribute used in this work and chapter 5 will show how powerful the VN transformation is and can lead to the solution of hard problems.

This discussion also raises the following points:

- Are there cases of channels which are given to be very noisy (where the very noisy transformation is not used as a tool)? And if the capacity of a channel is zero (e.g. the channel is too noisy), can we still provide information at a rate other than the exponential rate? Section 4.3 investigates this point.

- Is there also a structure present in the other extreme case, i.e., when the channel is very clear? How can we trust what the VN structure tells us? Can we use our very noisy geometry to understand better or answer questions dealing with non-very noisy channels? Clearly, it can help to find counter-examples, i.e., if a statement is denied in the VN limit, it will not hold in the general case. But if a statement holds in the VN limit, can we translate it in a true statement for the general setting? Sections 4.2.2 and 4.2.1 investigate these points.

4.2.1 Very Noisy Inverse Transformation: Lifting

The VN transformation has the advantage of simplifying the geometry, by setting the expressions in a inner product space, where we can use our intuition to better understand or possibly solve the problems. But it is not clear how much of the original problem's essence has been lost through this reduction. Do we have any guarantee

that a claim proved for very noisy limits must hold in general? No, in fact, one can find examples desproving this. However, in the problem we have approached in this work (and other ones not presented here), we will see how statements that were satisfied in the VN limit found corresponding statements that are true in the global setting. So could we know *when* we can trust the VN limit? We do not have an answer to that question. Nevertheless, we will see in chapter 5 that for a statement holding in the VN limit, there is a good way to guess what statement should be claimed in the general setting, and roughly speaking statements that can be expressed as inequalities between norms expressions seem to successfully lift to general statements. To do this, one has to figure out what is the expression (defined in $(M_1(\mathcal{X}\times\mathcal{Y}), D(\cdot||\cdot))$) that has a VN limit corresponding to the quantity of interest. We call this procedure "lifting". We have seen in previous sections that the divergence of two arguments close to each other maps to a norm in the VN limit, more precisely

$$D(P_Z(1+\varepsilon L_1)||P_Z(1+\varepsilon L_2)) \xrightarrow{VN} ||L_1 - L_2||^2_{P_Z}, \tag{4.14}$$

hence, each statement in the VN limit that can be expressed in terms of norms can be lifted to a general statement dealing with divergences. Each steps required to prove the statement can actually be lifted, but that any lifted proof's step holds is not guaranteed, in particular, the order in which arguments are place for the divergences will be crucial. Let us consider the following examples.

Let W_1, W_2 be two channels and $P_{\mathcal{X}}$ an input distribution. We define μ and μ^p the joint and product induced distributions respectively. We consider the VN transformations of W_1 and W_2 around $P_{\mathcal{Y}}$, in the respective directions L_1 and L_2. Using (4.14), we have

$$D(\mu_1||\mu_2) \xrightarrow{VN} ||L_1 - L_2||^2$$

$$D(\mu_1^p||\mu_2^p) \xrightarrow{VN} ||\bar{L}_1 - \bar{L}_2||^2$$

hence

$$D(\mu_1||\mu_2) - D(\mu_1^p||\mu_2^p) \xrightarrow{VN} ||L_1 - L_2||^2 - ||\bar{L}_1 - \bar{L}_2||^2 = ||\widetilde{L}_1 - \widetilde{L}_2||^2$$

where last equality simply uses the projection principle, i.e., that the projection of L onto centered directions $\widetilde{L} = L - \bar{L}$ is orthogonal to the projection's height $\bar{L}$. Therefore $||L_1 - L_2||^2 \geq ||\bar{L}_1 - \bar{L}_2||^2$. Having this in mind, one is tempted to claim:

$$D(\mu_1||\mu_2) \overset{?}{\geq} D(\mu_1^p||\mu_2^p), \tag{4.15}$$

which turns out to be true by the log-sum inequality. Now, with the projection picture in mind, the fact that last inequality holds is not surprising. But the point is that initially, without having this geometrical picture in mind, it may not be obvious, when trying to prove a claim, to see the divergence expressions in a geometric way and know that (4.15) holds. In that respect, the lifting of VN proofs can guide us in writing down the steps to be proved in the general setting. We conclude this section with the following comment about the local to global lifting. Most of the inequalities in information theory are using the concavity of the logarithm; when taking the VN limit, we are roughly replacing the logarithm expressions with quadratic expressions. But the function $x \mapsto -x^2$ is also concave, hence, in the respect of convex inequalities, the local and global problems share a common behavior (as illustrated with previous example).

4.2.2 Very Clear Channels

In previous sections, we analyzed the behavior of channels tending to a pure noise channel. It is then tempting to look at the other extreme case, i.e., when the family of channels is tending to a noiseless channel, where W is called a noiseless channel if each row of W contains a 1 and not all the rows are the same. In particular, assuming

$|\mathcal{X}| = |\mathcal{Y}| = A$, let us define

$$W_\varepsilon = I + \varepsilon(W - I), \quad \varepsilon \ll 1,$$

where I is the identity channel and W is an arbitrary channel. Note that, as opposed to the very noisy case when the limiting distribution has full support, we cannot consider all directions V for which

$$\sum_{y \in \mathcal{Y}} V(x, y) = 0, \quad \forall x \in \mathcal{X},$$

since the limiting channel is I, we need to ensure that $V(i, j) \geq 0$ whenever $i \neq j$. We will call such channels very clear channels. Surprisingly enough, we could not find references in the literature regarding such kinds of channels.

For an input distribution $P_\mathcal{X} \in M_1(\mathcal{X})$, the joint distribution induced by such channels is given by

$$\mu_\varepsilon^j = \mathrm{diag}(P_\mathcal{X}) + \varepsilon(P_\mathcal{X} \circ W - \mathrm{diag}(P_\mathcal{X})),$$

hence

$$(\mu_\varepsilon^j)_\mathcal{Y} = P_\mathcal{X} + \varepsilon(P_\mathcal{Y} - P_\mathcal{X})$$

and

$$\mu_\varepsilon^p = P_\mathcal{X} \times (P_\mathcal{X} + \varepsilon(P_\mathcal{Y} - P_\mathcal{X})).$$

Proposition 7.

$$I(P_X, W_\varepsilon) = H(P_\mathcal{X}) + \mathrm{tr}V\varepsilon \log \frac{1}{\varepsilon} + o(\varepsilon \log \frac{1}{\varepsilon}),$$

where

$$\mathrm{tr}V = \mathrm{tr}(P_\mathcal{X} \circ W) - 1.$$

The optimal input distribution is then clearly the uniform distribution and the

capacity scales as

$$C_\varepsilon = \log A + \mathrm{tr} V \varepsilon \log \frac{1}{\varepsilon} + o(\varepsilon \log \frac{1}{\varepsilon}).$$

The main feature attributed to very noisy channels, when very noisy channels were introduced (cf. [16] and references therein), was the fact that "their error exponents are known", i.e., the random coding and sphere packing exponents are tight in the limit. Therefore, before aiming to analyze the geometrical behavior of very clear channels, we analyze their error exponents. Recall that the random coding and sphere packing exponents for a rate R, channel W and input distribution $P_{\mathcal{X}}$ are given by

$$\begin{aligned} E_r(R, P_{\mathcal{X}}, W) &= \inf_{\mu : \mu_{\mathcal{X}} = P_{\mathcal{X}}} D(\mu || P_{\mathcal{X}} \circ W) + |I(\mu) - R|_+, \\ E_{sp}(R, P_{\mathcal{X}}, W) &= \inf_{\substack{\mu : \mu_{\mathcal{X}} = P_{\mathcal{X}} \\ I(\mu) < R}} D(\mu || P_{\mathcal{X}} \circ W). \end{aligned}$$

Let us consider a noiseless channel, i.e., $W = I$ and $P_{\mathcal{X}} \circ W = \mathrm{diag}(P_{\mathcal{X}})$. Then, by taking $\mu = \mathrm{diag}(P_{\mathcal{X}})$ we find an upper bound on $E_r(R, P_{\mathcal{X}}, I)$ given by $|H(P_{\mathcal{X}}) - R|_+$. However, for the sphere packing bound, taking μ diagonal is excluded by the constraint, and the error exponent is infinite below $H(P_{\mathcal{X}})$. This implies that the tightness of these two bounds in the very noisy limit is not going to occur in the very clear limit. This may seem contradictory with the fact that the random coding and sphere packing exponents are equal above the cut-off rate, but it simply means that in the noiseless case, the cut off rate is at the capacity. For very clear channel, we claim that the cut off rate and the capacity are tending to $H(P_{\mathcal{X}})$. It is then interesting to analyze the structure of the error exponent for rates scaling like capacity. If one considers fixed rates below capacity, the sphere packing error exponent blows up as

$$C(R) \log 1/\varepsilon,$$

where C depends on V. For example, if $V = 1/(A-1)I^c - I$, we have

$$C = f^{-1}(R),$$

where

$$f:\ s \in [0, \frac{A-1}{A}] \mapsto f(s) = D(B(s)||B(\frac{A-1}{A}))$$

and $B(s)$ is the Bernoulli probability measure with $\mathbb{P}\{0\} = s$. If

$$R_\varepsilon = C_\varepsilon + R\varepsilon \log \frac{1}{\varepsilon}$$

with $R \leq 0$. We have the following result.

Proposition 8.

$$\lim_{\varepsilon \to 0} \frac{E_{\text{sp}}(R_\varepsilon, P_\mathcal{X}, W_\varepsilon)}{\varepsilon} \leq R - \text{tr}V - R \log \frac{R}{\text{tr}V},$$

with equality for certain channels, and in fact we conjecture that this is the exponent for all channels.

We proved that equality holds for symmetric channels W. It would also be interesting to check where the cut-off rate is in term of R.

A similar result holds for universal source coding:

Proposition 9. *Let X_ϵ be an iid very clear source, i.e. $Q_\epsilon = \delta_i + \epsilon(Q - \delta_i)$ for some i and Q with $1 - Q_i < r$ (this implies that $h(Q_\varepsilon) \leq r\epsilon \log 1/\epsilon$). We can then universally encode this source at the rate R, such that if $i = 1$ and $Q = P$, the error exponent is given by*

$$E(r\epsilon \log 1/\epsilon, X_\epsilon)/\epsilon \to r \log \frac{r}{1 - P_1} - r + 1 - P_1,$$

with $1 - P_1 < r$.

Remark: Note that for both results found in the very clear setting, i.e., propositions 8 and 9, the exponent structure is of the form[2]

$$d_e(R,T) - d_d(R,T),$$

where $d_e(R,T) = R - T$, $d_d(R,T) = R \log \frac{R}{T}$, $T = \text{tr}(P_X \circ W) - 1$ in the channel case

[2]with different signs whether it is the channel or source case

and $T = 1 - P_1$ in the source case.

For other quantities investigated in the very noisy section, the problems becomes much more combinatorial for very clear channels. For example, even for a perfect channel, the mismatched mutual information is not easy to characterize. We have raised the point of analyzing very clear channels to examine how channels behave in the "other extreme" case of very noisy channels. One one hand it would be very interesting to understand both extreme cases and possibly have an homotopic view of the problem, on the other hand, any global claim eventually needs a rigorous proof and the very noisy setting appeared to preserved most of the global problem's essence in many situations.

4.3 Sub-Exponential Scaling

The main results of information theory makes heavy uses of the fact that for large block length n, the statistics of a codeword and its output through a channel will be enough distinguishable enough from the statistics of an independent codewords which has not been sent and the received output. By choosing appropriate code books and decoding rules, the receiver would guess a wrong codeword with a probability decaying to zero exponentially fast with n. The question is how fast can the number of messages M grow, in order to keep this scenario permissible, and showing it can grow at most exponentially, the rate $R = \lim_{n\to\infty} \frac{\log M}{n}$ is the quantity to maximize, leading to the capacity C. For certain channels, or for certain constraints on the input, the capacity is zero. No matter how large n is, the error probability cannot be made as small as desired; actually this statement is not quite true, although C can be zero, it does not mean that the error probability cannot be made as small as desired, it means it does not decay exponentially fast to zero. As we will see in what follows, even for zero capacity channels, one can still allow a number of messages growing to infinity with an error probability decaying to zero, *but the notion of rate will have to be adapted to the channel.*

4.3.1 Degrading Channels

We investigate discrete channels that are "memoryless" but not homogeneous, hence strictly speaking they are not memoryless according to the definition of *memoryless* given in chapter 3. We wish to analyze channels that are penalized with the increase of the block length. In this scenario, what was our main tour de force to fight the randomness for discrete memoryless channels, i.e., increasing block length, must be reexamined carefully.

We introduce a channel model that gets noisier with the block length (i.e., the ε parameter of the VN transformation is a function of the block length n that vanishes when n grows). Of course, because of this dependence $\varepsilon = \varepsilon(n)$, one has to be careful on how the limits are taken, and this model requires more than just the VN analysis, namely it requires the use of moderate deviations introduced in 2.2.2. The channel model is defined as follows: the probability of receiving the sequence y when the sequence x has been sent, with $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$, is given by

$$W^n(y|x) = \prod_{i=1}^{n} W^{(n)}(y_i|x_i)$$

and

$$W^{(n)}(y_i|x_i) \to P_{\mathcal{Y}}(y_i).$$

The speed at which the convergence happens and the direction are relevant to the problem, we assume that

$$W^{(n)}(y_i|x_i) = P_{\mathcal{Y}}(y_i)(1 + n^{-\alpha}L(x_i, y_i)). \tag{4.16}$$

For such block length penalty channels, with $\alpha \neq 0$, the usual notion of capacity is zero. No matter how you construct your code book, if one considers a number of code words growing like e^{nR}, then for any $R > 0$, the error probability cannot be made as small as desired with n, but this might be corrected if the number of codewords increases sub-exponentially.

Proposition 10. *For a channel defined as in* (4.16), *with* $0 < \alpha < 1/2$, *the maximal speed at which code books can grow in order to have a probability of error decaying to zero with the block length n is*

$$e^{n^{1-2\alpha}\widetilde{R}},$$

and the coefficient $\widetilde{R}$ *must satisfy*

$$\widetilde{R} < \widetilde{C} = \max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} \left\| \widetilde{L} \right\|_{P_{\mathcal{X}} P_{\mathcal{Y}}},$$

where $\widetilde{L}(a,b) = L(a,b) - \sum_{c \in \mathcal{X}} L(c,b) P_{\mathcal{X}}(c)$.

Proof. The probability of receiving an output $y_i \in \mathcal{Y}$ when $x_i \in \mathcal{X}$ is sent is

$$W^{(n)}(y_i|x_i) = P_{\mathcal{Y}}(y_i)(1 + n^{-\alpha} L(x_i, y_i)). \tag{4.17}$$

We now generate a iid code book $\{X_m\}_{m=1}^{M}$ of rate R, length n and distribution $P_{\mathcal{X}} \in M_1(\mathcal{X})$, Since we deal with equiprobable messages, let us assume that X_1 is sent. The joint distribution of X_1 and the received output sequence Y has then independent components and each component is distributed according to

$$\mu_n^j(a,b) = P_{\mathcal{X}}(a) P_{\mathcal{Y}}(b)(1 + n^{-\alpha} L(a,b)),$$

which induces an output marginal distribution given by

$$(\mu_n^j)_{\mathcal{Y}}(b) = P_{\mathcal{Y}}(b)(1 + n^{-\alpha} \bar{L}(b)),$$

hence a product distribution is

$$\mu_n^p(a,b) = P_{\mathcal{X}}(a) P_{\mathcal{Y}}(b)(1 + n^{-\alpha} \bar{L}(b)).$$

The decoding rule we will use, is the equivalent of the decoding rule described in section 3.2.2. If there exists a unique element $\hat{x}$ in the code book $\{x_m\}_{m=1}^{M}$ such that the joint empirical distribution of $\hat{x}$ and the received output y is within the norm ball

$B_{\|\cdot\|}(\mu_n^j, \delta n^{-\beta})$, where the norm $\|\cdot\|$ does not really matter, and say, is the L_2-norm. If there is more than one such element declare an error. We could have chosen to use the maximum likelihood decoding with a threshold test as described in section 3.2.1, however, the decoding rule chosen here is somehow easily pictured and since we only care about achievable rates and not exponent, it is sufficiently good. We can then upper bound the error probability by

$$\begin{aligned}\mathbb{P}\{\hat{X} \neq X_1\} &= \mathbb{P}\{P_{X_1,Y} \notin B(\mu_n^j, \delta n^{-\beta}) \cup \cup_{m\neq 1} P_{X_m,Y} \in B(\mu_n^j, \delta n^{-\beta})\} \\ &\leq \mathbb{P}\{P_{X_1,Y} \notin B(\mu_n^j, \delta n^{-\beta})\} + 1 \wedge M\mathbb{P}\{P_{X_2,Y} \in B(\mu_n^j, \delta n^{-\beta})\}\end{aligned}$$

But

$$\mathbb{P}\{P_{X_1,Y} \notin B(\mu_n^j, \delta n^{-\beta})\} = \mathbb{P}\{n^\beta(P_{X_1,Y} - \mu_n^j) \notin B(0,\delta)\},$$

which, from corollary 1, is decaying exponentially fast as long as $\delta > 0$. Moreover,

$$\mathbb{P}\{P_{X_2,Y} \in B(\mu_n^j, \delta n^{-\beta})\} = \mathbb{P}\{n^\beta(P_{X_2,Y} - \mu_n^p) \in B(n^\beta(\mu_n^j - \mu_n^p), \delta)\}$$

and

$$\mu_n^j(a,b) - \mu_n^p(a,b) = n^{-\alpha} P_{\mathcal{X}}(a) P_{\mathcal{Y}}(b)(L(a,b) - \bar{L}(b)). \tag{4.18}$$

Therfore,

$$B(n^\beta(\mu_n^j - \mu_n^p), \delta) = B(P_{\mathcal{X}} P_{\mathcal{Y}} \widetilde{L}, \delta) \tag{4.19}$$

and again, from corollary 1

$$\begin{aligned}\mathbb{P}\{P_{X_2,Y} \in B(\mu_n^j, \delta n^{-\beta})\} &= \mathbb{P}\{n^\beta(P_{X_2,Y} - \mu_n^p) \in B(P_{\mathcal{X}} P_{\mathcal{Y}} \widetilde{L}, \delta)\} \\ &\doteq e^{-n^{1-2\beta} |\inf_{\nu \in B(P_{\mathcal{X}} P_{\mathcal{Y}} \widetilde{L}, \delta)} \frac{1}{2} \sum_{a\in\mathcal{X}, b\in\mathcal{Y}} \frac{\nu^2(a,b)}{p_{\mathcal{X}}(a) p_{\mathcal{Y}}(b)} - R|^+}. \end{aligned} \tag{4.20}$$

Therefore, if $1 - 2\beta < 1$ and

$$R < \frac{1}{2} \left\| \widetilde{L} \right\|^2_{P_{\mathcal{X}} P_{\mathcal{Y}}},$$

we can choose $\delta > 0$ such that the exponent in (4.20) is strictly positive. □

This proof also convinces us that if $\beta \geq 1/2$, any code book having a number of messages growing to infinity with n, at any scale, cannot have a probability of error decaying to zero with n.

4.3.2 Fisher Information as Capacity

Previous result considers channels that become noisier with the block length, and it shows that as long as the speed of convergence to the pure noise channel is slow enough, we can still ensure reliable communication with a new definition of rate and capacity. The capacity took the form that we expected from the very noisy capacity expression of chapter 4. In this section we investigate continuous time channels, although we have not dealt with such channels till now, we skip for now the introduction to continuous alphabet channels and directly present the result. An introduction to continuous alphabet channel is given in chapter 6.

We consider an additive noise channel:

$$Y_i = u_i + Z_i, \quad i = 1, \ldots, n$$

where n is the block length and Z_i's are i.i.d. random variables with a differentiable density (not depending on the u_i's), with the following type of constraint on the inputs u_i's:

$$n^{-\alpha} \sum_{i=1}^{n} u_i^2 \leq P,$$

for some fixed values $0 \leq \alpha \leq 1$ and $P > 0$. This means that the available energy is not scaling with the channel uses (the higher the block length, the lower the power), with the extreme case of having finite energy if $\alpha = 0$.

By defining $x_i = n^{\frac{1-\alpha}{2}} u_i$, this problem is then equivalent to the following one:

Channel:

$$Y_i = n^{-\beta} x_i + Z_i, \quad i = 1, \ldots, n$$

constraint:

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P,$$

where in this example, $0 \leq \beta \leq 1/2$, as we consider a total power constraint, but with an initial constraint of the form $n^{-\alpha} \sum_{i=1}^{n} |u_i|^\rho \leq P$, the range of the parameter β would then be $0 \leq \beta \leq 1/\rho$.

A similar situation could arise in a communication network where the interference of other users is treated as noise. Let us assume that for specific transmitter and receiver in the network, the channel of use is modeled as

$$Y_i = n^{-\beta} x_i + Z_i + \sum_{k \in N} X_i^k, \quad i = 1, \ldots, n$$

where X_i^k are iid Gaussian (standard) random variables, representing the interference of a number N of neighboring users. The inputs are constrained with an average power constraint. Depending on how the number of users in the network grows with respect to the available network volume, the number of neighbors N can be increasing with the number of users. The capacity per users may then go to zero, i.e., a number of messages increasing exponentially fast with the block length cannot be reliably communicated. However, if the number of neighbors grows slow enough with respect to the block length, we may still be able to communicate reliably some information at another scale. Considering a scaling between the number of neighbors and the block length leads to a channel similar as the one mentioned earlier.

Proposition 11. *Let a channel be such that for a block length n,*

$$Y_i = n^{-\beta} x_i + Z_i, \quad i = 1, \ldots, n$$

where the Z_i's are i.i.d. random variables with variance σ^2 and differentiable density,

and the x_i's are constrained by

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P.$$

- *If $\beta = 0$, we are in the usual setting and to ensure reliable communication, the number of messages can grow at most exponentially fast, as long as the rate R satisfies*

$$R < C = \sup_{X:EX^2 \leq P} I(X,Y).$$

- *If $0 < \beta < 1/2$, the shannon capacity is zero. Nevertheless, if the number of messages $M(n)$ increases sub-exponentially with n, at most like $M(n) = e^{n^{1-2\beta}\tilde{R}}$, we can still communicate reliably, as long as the coefficient $\tilde{R}$ satisfies*

$$\tilde{R} < \tilde{C} = \sup_{X:EX^2 \leq P} \frac{1}{2}\mathrm{Var}(X)J(Z) = \frac{P}{2}J(Z),$$

where $J(Z)$ is the fisher information of the noise, i.e., $J(Z) = \int_{\mathbb{R}} \frac{p_Z(x)'^2(x)}{p_Z(x)} dx$.

- *If $\beta \geq 1/2$, then for a number of messages increasing with n , i.e., $M(n) \to \infty$, no reliable communication is possible.*

Proof. We now have $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$. Let $p_{Y|X=x}$ be the probability density of an output of length n when the input $x \in \mathbb{R}^n$ is sent through the channel, i.e.,

$$p_{Y|X=x}(y|x) = \prod_{i=1}^{n} p^{(n)}(y_i|x_i),$$

where

$$p^{(n)}(y_i|x_i) = p_Z(y_i - n^{-\beta}x_i)$$

But

$$p_Z(y_i - n^{-\beta}x_i) = p_Z(y_i) - n^{-\beta}x_i p_Z'(y_i) + o(n^{-\beta}x_i),$$

hence, let us work for now with the channel transition density given by

$$q_{Y|X=x}(y|x) = \prod_{i=1}^{n} q^{(n)}(y_i|x_i),$$

where

$$q^{(n)}(y_i|x_i) = p_Z(y_i) - n^{-\beta}x_i p_Z'(y_i)$$

and we will deal with the approximation afterwards.

We now generate a random code book as follows: we pick a probability density $p_{\mathcal{X}}$ on $\mathbb{R}$, and we draw independently $\{X_m\}_{m=1}^M$, with

$$X_m(1), \ldots, X_m(n) \overset{\text{iid}}{\sim} p_{\mathcal{X}},$$

this is the natural extension of what we defined to be a iid random code book of length n and distribution $P_{\mathcal{X}}$ in the discrete setting. However, we also need to ensure that the code book we generated satisfies the average power inequality. We require then $p_{\mathcal{X}}$ to have a bounded variance: $\text{Var}(p_{\mathcal{X}}) < P - \varepsilon$, for some $\varepsilon > 0$, where $\text{Var}(p_{\mathcal{X}}) = \int_{\mathbb{R}}(a - m(p_{\mathcal{X}}))^2 p_{\mathcal{X}}(a)da$ and $m(p_{\mathcal{X}}) = \int_{\mathbb{R}} a p_{\mathcal{X}}(a)da$. We can then proceed as usual using Sanov's theorem to show that code books that are not satisfying the power constraint have probabilities decaying exponentially fast and use a continuity argument to get rid of ε.

Since we deal with equiprobable messages, let us assume that X_1 is sent. The joint distribution of X_1 and the received output sequence Y then has independent components and each component is distributed according to

$$\mu_n^j(a,b) = q^{(n)}(b|a)p_{\mathcal{X}}(a) = p_{\mathcal{X}}(a)p_Z(b) - n^{-\beta}b p_{\mathcal{X}}(a)p_Z'(b),$$

which induces an output marginal distribution given by

$$(\mu_n^j)_{\mathcal{Y}}(b) = p_Z(b) - n^{-\beta}m(p_{\mathcal{X}})p_Z'(b),$$

hence a product distribution

$$\mu_n^p(a,b) = p_{\mathcal{X}}(a)(\mu_n^j)_{\mathcal{Y}}(b) = p_{\mathcal{X}}(a)p_Z(b) - n^{-\beta}m(p_{\mathcal{X}})p_{\mathcal{X}}(a)p_Z'(b).$$

Note that

$$\mu_n^j(a,b) - \mu_n^p(a,b) = -n^{-\beta}(b - m(p_{\mathcal{X}}))p_{\mathcal{X}}(a)p_Z'(b). \tag{4.21}$$

For two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we define the joint empirical distribution by

$$p_{x,y} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x(i),y(i)},$$

where δ is the Dirac delta distribution. The decoding rule we will use, is the equivalent of the decoding rule described in section 3.2.2. If there exists a unique element $\hat{x}$ in the code book $\{x_m\}_{m=1}^M$ such that the joint empirical distribution $P_{\hat{x},y}$ of $\hat{x}$ and the received output y, is within the variational norm ball $B(\mu_n^j, \delta n^{-\beta})$, declare $\hat{x}$, otherwise declare an error.

$$\begin{aligned}
\mathbb{P}\{\hat{X} \neq X_1\} &= \mathbb{P}\{p_{X_1,Y} \notin B(\mu_n^j, \delta n^{-\beta}) \cup \cup_{m\neq 1} p_{X_m,Y} \in B(\mu_n^j, \delta n^{-\beta})\} \\
&\leq \mathbb{P}\{p_{X_1,Y} \notin B(\mu_n^j, \delta n^{-\beta})\} + 1 \wedge M\mathbb{P}\{p_{X_2,Y} \in B(\mu_n^j, \delta n^{-\beta})\}
\end{aligned}$$

But

$$\mathbb{P}\{p_{X_1,Y} \notin B(\mu_n^j, \delta n^{-\beta})\} = \mathbb{P}\{n^{\beta}(p_{X_1,Y} - \mu_n^j) \notin B(0,\delta)\},$$

which, from the generalization of theorem 5 in [1], is decaying exponentially fast as long as $\delta > 0$. Moreover,

$$\mathbb{P}\{p_{X_2,Y} \in B(\mu_n^j, \delta n^{-\beta})\} = \mathbb{P}\{n^{\beta}(p_{X_2,Y} - \mu_n^p) \in B(n^{\beta}(\mu_n^j - \mu_n^p), \delta)\}.$$

and recalling that from (4.22),

$$\mu_n^j(a,b) - \mu_n^p(a,b) = n^{-\beta}(a - m(p_{\mathcal{X}}))p_{\mathcal{X}}(a)p_Z'(b). \tag{4.22}$$

we have

$$B(n^\beta(\mu_n^j - \mu_n^p), \delta) = B((\mathrm{id}_\mathcal{X} - m(p_\mathcal{X}))p_\mathcal{X}p_Z^{'}, \delta) \tag{4.23}$$

and again, from the generalization of theorem 5 in [1]

$$\begin{aligned}\mathbb{P}\{p_{X_2,Y} \in B(\mu_n^j, \delta n^{-\beta})\} &= \mathbb{P}\{n^\beta(p_{X_2,Y} - \mu_n^p) \in B((\mathrm{id}_\mathcal{X} - m(p_\mathcal{X}))p_\mathcal{X}p_Z^{'}, \delta)\} \\ &\doteq e^{-n^{1-2\beta}|\inf_{\nu \in B((\mathrm{id}_\mathcal{X} - m(p_\mathcal{X}))p_\mathcal{X}p_Z^{'}, \delta)} \frac{1}{2}\int_{\mathbb{R}^2} \frac{\nu^2(a,b)}{p_\mathcal{X}(a)p_Z(b)}dadb - R|^+}\end{aligned} \tag{4.24}$$

Therefore, if $1 - 2\beta < 1$ and

$$R < \frac{1}{2}\int_{\mathbb{R}^2} \frac{(a - m(p_\mathcal{X}))p_\mathcal{X}(a)p_Z^{'}(b)}{p_\mathcal{X}(a)p_Z(b)}dadb,$$

from a continuity argument, we can choose $\delta > 0$ such that the exponent in (4.24) is strictly positive. Finally, observe that

$$\int_{\mathbb{R}^2} \frac{(a - m(p_\mathcal{X}))^2 p_\mathcal{X}^2(a)(p_Z^{'})^2(b)}{p_\mathcal{X}(a)p_Z(b)}dadb = \mathrm{Var}(p_\mathcal{X})J(p_Z),$$

where $\mathrm{Var}(p_\mathcal{X}) = \int_\mathbb{R}(a - m(p_\mathcal{X}))^2 p_\mathcal{X}(a)da$ and $J(p_Z) = \int_\mathbb{R} \frac{p_Z(b)^{'2}(b)}{p_Z(b)}db$. To get rid of the initial approximation, we sould restrict ourself to distributions $p_\mathcal{X}$ that are decaying fast enough, in order to have a finite variance; but this has been taken care of by the power constraint. □

Chapter 5

Linear Universal Decoding

In memoryless settings, the maximum likelihood (ML) decoder is not only optimal, but it is also linear, i.e., it maximizes a score function that is additive over the code length. This linear structure affords several nice properties. In particular, it allows ML to be considered in a practical setting. In [11], the authors mention that the class of linear decoders "itself affords many interesting problems" and "may further enhance the interplay of information theory and combinatorics". They also mention that "consideration of complexity" may provide a primary reason for the use of these decoders. However, the optimality of ML is contingent upon knowledge of the channel law, and in general the channel law is unknown to the transmitter and receiver, such as for compound channels. In order to account for the user's ignorance of the channel law, several universal decoders have been proposed ([9], [14], [19], [21]). Although theoretically optimal, none of these decoders is linear (hence practical) and some depend heavily on the discrete alphabet assumption.

Are universality and linearity two properties that cannot be embodied by a single decoder?

In this chapter, we address the problem of finding good linear decoders over compound discrete memoryless channels. We will prove that under minor concessions, linear universal decoders[1] exist. Indeed, we will construct such decoders.

[1] In this work, universal decoders are asked to be capacity achieving; formal definitions of these terms are given in section 5.2.1 and 5.2.2

5.1 Compound Channels

In this chapter, we use the same setting as in chapter 3. We consider a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The receiver and transmitter do not know what the exact channel is, i.e., what the exact probability transition matrix is, all they know is that it belongs to a set S of possible channels. This is known as a compound discrete memoryless channel. As briefly exposed in chapter 3, DMC's can be seen as (simplified) models for wireless communication channels, resulting from the combinations of modulators, waveform transmissions channel and demodulators, under flat fading assumptions. Assuming that the channel's law is exactly known at the receiver and transmitter is convenient to analyze DMC's, but cannot be part of the communication model assumption. Compound DMC's are appropriate models when the fading is flat and slowly varying. If T_n denotes the transmission time of a codeword of block length n and T_c the channel coherence time (all units in seconds), slow fading means that $T_n \ll T_c$, so that the channel law remains effectively unchanged during the transmission of a codeword. Employing training sequences could be attractive if the channel remains unchanged over multiple transmissions. However, a drawback of this approach is an effective reduction of code rates. Moreover, any estimators achieved through training sequences would still leave us with a certain amount of unknowns in the channel knowledge.

A rate-R, block length-n encoder E_n and decoder D_n are defined in the same way we have defined them for DMC's, in the definition 4 of chapter 3. We denote by $\{x_m\}_{m=1}^{M}$ the codebook whose codewords have length n, and by y a received sequence of length n. Since the channel is memoryless, the probability of observing the output y when x_m is transmitted is given by

$$W^n(y|x_m) = \prod_{i=1}^{n} W(y(i)|x_m(i)),$$

where the channel W can be any channels in S. If the set S contains only one channel, then we are back to the usual definition of DMC.

We assume that the messages to be transmitted are equiprobable. As defined in chapter 3, we denote by $P_e(E_n, D_n, W)$ the average probability of error for a given block length n, rate R, encoder E_n, decoder D_n and channel W.

Definition 17. We say that a rate R is achievable on the compound set S if for any $\epsilon > 0$, there exists a block length n, an encoder E_n and decoder D_n of rate at least R, such that for all $W \in S$, we have $P_e(E_n, D_n, W) < \epsilon$.

The following theorem is due to D. Blackwell, L. Breiman and A. Thomasian (cf. [5]).

Theorem 7. *[5]*

The supremum of all achievable rates, on a compound set S, is given by

$$C(S) = \max_{P_X \in M_1(\mathcal{X})} \inf_{W \in S} I(P_X, W). \tag{5.1}$$

We call (5.1) the compound capacity. This result generalizes Shannon's basic theorem on the capacity of a single DMC to a set of DMC's. In the basic setting, an optimal decoder such as maximum likelihood, or any decoders using notions of typicality, can be used at the receiver, since the channel of communication is known. In the compound setting this is no longer possible, the encoder and decoder must then be built efficiently for all possible channels in the set, without knowing what the exact channel is. Moreover, in the basic (single channel) setting, the random coding argument proceeds by computing the expected probability of error of a randomly chosen code and allows us to conclude that a code exists with a probability of error of at most this expected value. In the compound setting, even if we figure out how to construct good decoders not depending on the channel knowledge, and even if we show that the expected error probability of a randomly chosen code is small, this is not enough to conclude the existence of a code with small error probability for all channels in S; the expectation may be small because different codes have small error probability for different channels.

If S is a finite set, it is then easy to check that the random coding argument ensures the existence of a sufficient number of good codebooks for each channel so that the the intersection of those good codes is not empty: let $|S| = K$ and let $\{X_m\}_{m=1}^M$ be a iid codebook of length n and distribution $P_{\mathcal{X}}$. As in chapter 3, we define the induced random probability of error when MMI decoding is used over the channel W to be $P_e(\{X_m\}, \text{MMI}, W)$. We know from section 3.2.2, that for any $\varepsilon > 0$, there exists a block length n such that for any $W \in S$,

$$\mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_e(\{X_m\}, \text{MMI}, W) < \frac{\varepsilon}{K},$$

hence, using Markov's inequality, we have that for any $W \in S$

$$\mathbb{P}\{P_e(\{X_m\}, \text{MMI}, W) > \varepsilon\} < \frac{1}{2K}.$$

Using the union bound, we have

$$\begin{aligned} \mathbb{P}\{\max_{W \in S} P_e(\{X_m\}, \text{MMI}, W) > \varepsilon\} &= \mathbb{P}\{\cup_{W \in S}\{P_e(\{X_m\}, \text{MMI}, W) > \varepsilon\}\} \\ &< \frac{1}{2}, \end{aligned}$$

which implies

$$\mathbb{P}\{\max_{W \in S} P_e(\{X_m\}_{m=1}^M, \text{MMI}, W) < \varepsilon\} > 0,$$

showing that there exists a realization of the codebook $\{X_m\}_{m=1}^M$ for which the error probability is less than ε for all W in S.

But for arbitrary sets, more work is required to show that this conclusion is still valid. In their original proof, the authors in [5] use a decoder that maximizes a uniform mixture of likelihoods over a family of channels, which is not finite but growth only as a polynomial in n. Ideally, we would like to take the family of channels to be the whole set S, but the polynomial grows is necessary to show the existence of a good codeword from the random coding argument. We now present a proof which is slightly different from the original proof of [5], but that will be useful to later prove

theorem 8.

Proof. Using the MMI decoder, which is universal as shown in section 3.2.2, we have for $R < I(P_{\mathcal{X}}, W)$,

$$\mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_e(\{X_m\}, \text{MMI}, W) \leq e^{-nE_r(R, P_{\mathcal{X}}, W)},$$

where $E_r(R, P_{\mathcal{X}}, W) > 0$ is given in section 3.2. Let us consider $R < \inf_{W \in S} I(P_{\mathcal{X}}, W) - \delta$, for some $\delta > 0$, we then have $E_r(R, P_{\mathcal{X}}, W) > 0$ for any $W \in S$. We now proceed to the approximation of S by the polynomial subset.

Lemma 2. *[5]*
Let S be a set of $A \times B$ stochastic matrices, with $A, B \in \mathbb{Z}_+$. For any $C \in \mathbb{Z}_+$ such that $C \geq 2B^2$, there exists a set of $A \times B$ stochastic matrices $S^{(C)}$ with $|S^{(C)}| < (C+1)^{AB}$, such that for any $W \in S$, there exists $W^{(C)} \in S^{(C)}$ satisfying

$$|W(b|a) - W^{(C)}(b|a)| < B/C, \quad \forall 1 \leq a \leq A, 1 \leq b \leq B$$

and

$$W(b|a) < e^{\frac{2B^2}{A}} W^{(C)}(b|a), \quad \forall 1 \leq a \leq A, 1 \leq b \leq B.$$

By choosing $C = n^2$, we get the following result.

Lemma 3. *[5]*
There exists a set $\hat{S}$ of $|\mathcal{X}| \times |\mathcal{Y}|$ stochastic matrices with

$$|\hat{S}| \leq (1 + n^2)^{|\mathcal{X}||\mathcal{Y}|},$$

such that for any $|\mathcal{X}| \times |\mathcal{Y}|$ stochastic matrix W, there exists $\hat{W} \in \hat{S}$ satisfying for all $x \in \mathcal{X}^n$, $D_n \subset \mathcal{Y}^n$:

$$|W(y|x) - \hat{W}(y|x)| < |\mathcal{Y}|/n^2 \tag{5.2}$$

and

$$\sum_{y \in D_n} W(y|x) \leq e^{2|\mathcal{Y}|^2/n^2} \sum_{y \in D_n} \hat{W}(y|x). \tag{5.3}$$

From (5.2), we deduce that by taking n large enough, we can ensure $R < I(P_{\mathcal{X}}, \hat{W}) - \delta/2$ for any $\hat{W} \in \hat{S}$ which is an approximation of $W \in S$, hence

$$\mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_e(\{X_m\}, \mathrm{MMI}, \hat{W}) \leq e^{-nE_r(R, P_{\mathcal{X}}, \hat{W})}.$$

From Markov's inequality, we then have

$$\mathbb{P}\{P_e(\{X_m\}, \mathrm{MMI}, \hat{W}) > \varepsilon\} < \frac{1}{\varepsilon} e^{-nE_r(R, P_{\mathcal{X}}, \hat{W})}$$

and

$$\begin{aligned} \mathbb{P}\{\cup_{\hat{W} \in \hat{S}} \{P_e(\{X_m\}, \mathrm{MMI}, \hat{W}) > \varepsilon\}\} &< \frac{1}{\varepsilon} \sum_{\hat{W} \in \hat{S}} e^{-nE_r(R, P_{\mathcal{X}}, \hat{W})} \\ &\leq \frac{1}{\varepsilon} (1 + n^2)^{|\mathcal{X}||\mathcal{Y}|} e^{-n \inf_{\hat{W} \in \hat{S}} E_r(R, P_{\mathcal{X}}, \hat{W})} \\ &\doteq \frac{1}{\varepsilon} e^{-n \inf_{\hat{W} \in \hat{S}} E_r(R, P_{\mathcal{X}}, \hat{W})}. \end{aligned}$$

Moreover, using (5.3) and taking n large enough, we have

$$\begin{aligned} \mathbb{P}\{\cup_{W \in S} \{P_e(\{X_m\}, \mathrm{MMI}, \hat{W}) > \varepsilon\}\} &\leq \mathbb{P}\{\cup_{\hat{W} \in \hat{S}} \{P_e(\{X_m\}, \mathrm{MMI}, \hat{W}) > \varepsilon/2\}\} \\ &\dot{<} \frac{2}{\varepsilon} e^{-n \inf_{\hat{W} \in \hat{S}} E_r(R, P_{\mathcal{X}}, \hat{W})} \end{aligned}$$

But for all $\hat{W} \in \hat{S}$, we have $R < I(P_{\mathcal{X}}, \hat{W}) - \delta/2$, hence $\inf_{\hat{W} \in \hat{S}} E_r(R, P_{\mathcal{X}}, \hat{W}) > 0$. Finally, note that there exists $C(|\mathcal{X}|, |\mathcal{Y}|) > 0$ such that

$$\frac{\partial E_r(R, P_{\mathcal{X}}, W)}{\partial R}\Big|_{R = I(P_{\mathcal{X}}, W)} \geq C(|\mathcal{X}|, |\mathcal{Y}|),$$

therefore, we can take the limit of δ goes to 0 and as long as $R < \inf_{W \in S} I(P_X, W)$, we ensure the existence of codebooks having arbitrarily small error probability for any $W \in S$. □

5.2 Linearity and Universality

5.2.1 Universal Decoding

The notion of universal decoding, although commonly used in the literature, may sometimes appear a little bit confusing. First, a decoder is in general making sense only when considered jointly with its encoder, and talking about a decoder which is capacity achieving must then implicitly require the existence the complementary encoder. Moreover, all those definitions are depending on the block length, and the achievability is an asymptotic notion; so should we talk about a universal sequence of decoder? The next definitions aim to clarify these points, avoiding superfluous formalism when possible.

Definition 18. We say that a sequence of encoders and decoders is universal for the compound set S if it achieves the compound capacity, i.e., if for any $R < C(S)$, $\epsilon > 0$, there exists n, E_n and D_n from the sequence with rate at least R, such that for any $W \in S$, we have $P_e(E_n, D_n, W) < \epsilon$.

Note that this definition of universality is weaker than the one defined in [19] and references therein, where a decoder is declared to be universal if it achieves the same random-coding error exponent as the ML decoder tuned to the true channel of communication.

Definition 19. (Informal)
We say that a decoder is induced by a "decoding rule" if the mappings D_n can be defined generically for all R, n and E_n.
Examples: ML with respect to any distribution, MMI [9], any α and β decoders as de-

fined in [10], LZ-based algorithm [21], merged likelihood [14], Generalized Likelihood Ratio Test (GLRT) [19].

Definition 20. We say that a decoding rule is universal for a family of sets, if for any compound set S in the family, there exists a sequence of encoders for which the generated sequence of encoders and decoders is universal for the compound set S.
A decoding rule is universal if it is universal for all subset of DMC's.
Example: MMI, LZ-based algorithm, merged likelihood are universal decoding rule.
ML is a universal decoding rule for singleton sets.

Note that a decoding rule such as MMI decoding does not even require the knowledge of the compound set S, whereas our definition of a universal decoding rule allows us to use the knowledge of the compound set. But since encoders and decoders must anyway cooperate and agree on a codebook before the communication takes place, and since the encoder must know the compound set S (to figure out which rates can be used), this feature of MMI has no real advantage in this setting.

5.2.2 Linear Decoding

Definition 21. We say that a decoding rule is additive, or linear, and induced by a metric d (though it may not be a metric in the formal sense), if it is given by

$$D_n(y) = \arg\max_m d^n(x_m, y), \tag{5.4}$$

where

$$d^n(x_m, y) = \frac{1}{n}\sum_{i=1}^{n} d(x_m(i), y(i)) = \mathbb{E}_{P_m} d$$

and d ("the metric") is any real function on $\mathcal{X} \times \mathcal{Y}$.
If the maximizer is not unique, an error is declared.

Example: the maximum likelihood decoder, or more precisely, the corresponding maximum log-likelihood decoder, with respect to any channel W, is additive and its metric is given by $\log W$.

Lemma 4. *Let n, $P_{\mathcal{X}} \in M_1(\mathcal{X})$ and $\{x_m\}_{m=1}^M \subset \mathcal{X}^n$ such that $P_{x_m} = P_{\mathcal{X}}$, $\forall 1 \leq m \leq M$. Then, for any $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $\alpha : \mathcal{X} \to \mathbb{R}$ and $\beta : \mathcal{Y} \to \mathbb{R}$, the additive decoding rule induced by the metric d is equivalent to the additive decoding rule induced by the metric $d + \alpha + \beta$.*
In particular, there exists $W : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ with $\sum_{b \in \mathcal{Y}} W(a,b) = 1$, $\forall a \in \mathcal{X}$, such that the additive decoding rule induced by the metric d is equivalent to the additive decoding rule induced by the metric $\log W$.

Proof. The first part of the lemma is trivial since the codewords $\{x_m\}_{m=1}^M$ have constant composition. For the second part, let

$$A : a \in \mathcal{X} \mapsto \sum_{b \in \mathcal{Y}} e^{d(a,b)} > 0$$

and

$$W(a,b) = e^{d(a,b) - \log A(a)} \geq 0, \quad \forall a \in \mathcal{X}, b \in \mathcal{Y}.$$

By construction, W satisfies the required hypotheses and since $\log W(a,b) = d(a,b) - \log A(a)$, we can use the first part of the lemma to conclude the proof. □

An additive decoding rule has certain advantages with respect to a non-additive one. First, it can be much simpler to implement, in particular if the codes being used have a trellis structure (with bounded width), the additive structure will allow the use of algorithms such as Viterbi's algorithm, keeping track of a finite number of candidates, as opposed to the initial exponential number. If codes such as convolutional codes are used, the additive structure also allows to decode successively without having to wait for the full block length, as would be the case when decoding with MMI. Other algorithms, such as Belief propagation, also require the additive structure to be implemented. However, we shall recall here that the notion of universality we use for decoding rules, requires the existence of a sequence of encoders such that the generated sequence of codes achieves any rate below capacity. This implies that if the encoders employed to prove the universality of an additive decoding rule are not of the required algebraic type (as the ones mentioned previously), the con-

sidered additive decoding rule may not have the claimed complexity. If the random coding argument is used to show the universality of an additive decoding rule, with a random codebook of fix optimal type $P_{\mathcal{X}}$, the algebraic structure should be somehow well represented in the good realization of the random codebook to ensure linear complexity. However, this discussion is beyond the scope of this work, our goal in this chapter is to analyze the behavior of those linear decoders over unknown channels through randomly generated codebooks. This will show limiting performances of the considered schemes, but the task of making the schemes practical is a different one. As opposed to decoding rules such as MMI, if we achieve a certain rate using a linear decoder over a compound set, we at least have the hope to achieve this rate with the claimed complexities of algebraic codes. Theoretically, additive decoders do not only seem to naturally suit a memoryless setting, but as we saw in section 2.1, much more is known when the constraint under which a divergence is minimized (in its first argument) is linear; we can then use the geometrical properties presented in section 2.1 to facilitate the analysis of the considered coding schemes. Hence, the additive framework will allow us to understand better the geometry of decoders. Finally, a universal decoder such as MMI can hardly be generalized to continuous alphabets.

Definition 22. We say that a decoding rule is generalized linear and induced by the metrics $\{d_k\}_{k=1}^K$, if it is given by

$$D_n(y) = \arg\max_m \vee_{k=1}^K d_k^n(x_m, y) = \arg\max_m \vee_{k=1}^K \mathbb{E}_{P_m} d_k,$$

where d_k^n is an additive decoding rules induced by the metrics d_k and $K = K(S) < +\infty$ does not depend on n.

We talk about "decoding with the metrics $\{d_k\}_{k=1}^K$" when such a decoding rule is used.

Example: GLRT with respect to any finite set of distributions.

Remarks:

- Note that formally speaking, the mapping $\mu \mapsto \vee_{k=1}^K \mathbb{E}_\mu d_k$ is not linear. However, it is equivalent to performing finitely many linear decoding rules in parallel and

doing one comparison of finitely many real numbers at the end. *Therefore, since K is finite (not growing with n), all above attributes associated with additive decoding rules still hold with linear generalized decoding rules.* In the following, we will then often omit the term "generalized" when referring to such decoding rules.

- We do not allow K to vary with the rate R (indeed, this is voided by the definition of decoding rules). Therefore, with this definition, a linear and universal decoding rule is such that for any set S, any rate R with $R < C$ can be achieved with the same $K = K(S)$ and the same metrics $\{d_k\}_{k=1}^K$. One can define a weaker notion of linear universal decoder by requiring that for any set S and rate $R < C$, there exists a constant $K = K(S, R)$, which is not depending on n, and a set of metrics $\{d_k(R)\}_{k=1}^K$, such that decoding with these metrics can achieve R.

- Lemma 4 does not generalize to linear decoding rules, unless the functions α and β are constant. Therefore, metrics which are not the logarithm of a stochastic matrix may and will be relevant.

Problem: we are interested in finding "good" linear decoding rules for compound discrete memoryless channels. Ideally, we would like to find universal linear decoding rules (on any families of compound sets). However, it is not clear that this goal can be achieved; here the meaning of "good" has to be understood in terms of rate achievability.

5.2.3 Linearity VS Universality

To get started, we need an estimate of what rates can be achieved with an arbitrary linear decoding rule. We first introduce some notations:

- In the proof of following results, we use constant composition random codes from a distribution $P_{\mathcal{X}}$. The existence of good encoders are then deduced from the random coding argument, while decoders are explicitly constructed from the

considered decoding rules. The input distribution $P_{\mathcal{X}}$ will be kept fixed through all the chapter. We will not have results depending on what input distribution is considered (or that are true only for the optimal input distribution), but we can always think of dealing with the optimal input distribution, which is defined for a set S to be

$$\arg \max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} \inf_{W \in S} I(P_{\mathcal{X}}, W),$$

and if the maximizer is not unique, we define $P_{\mathcal{X}}$ to be one of the maximizers arbitrarily.

- $W_0 \in S$ denotes the true channel of communication
- We define the worse channel of a set S by $W_S = \min_{W \in \mathrm{cl}(S)} I(P_{\mathcal{X}}, W)$, which may not belong to S if S is open, but this will not represent a problem.
- μ denotes a joint distribution on $\mathcal{X} \times \mathcal{Y}$
- $\mu_{\mathcal{X}}$, $\mu_{\mathcal{Y}}$ denote the respective marginal distributions in $\mathcal{X}$, $\mathcal{Y}$ and the induced product distribution is denoted by $\mu^p = \mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$
- When a channel is indexed with a subscript k, the joint and product distribution generated with first marginal $P_{\mathcal{X}}$ are denoted as follows: $W_k \xrightarrow{P_{\mathcal{X}}} \mu_k = P_{\mathcal{X}} \circ W_k \xrightarrow{\text{marginals}} \mu_k^p = P_{\mathcal{X}} \times (\mu_k)_{\mathcal{Y}}$,
  where $k = 0$ is reserved for the true channel and S instead of k for the worse channel.

Theorem 8. *Using the linear decoding rule induced by the metrics $\{d_k\}_{k=1}^K$, we can achieve the rate*

$$\max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} \inf_{W_0 \in S} \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \vee_{k=1}^K \mathbb{E}_\mu d_k \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} d_k}} D(\mu || \mu_0^p). \tag{5.5}$$

Moreover, observe that:

$$(5.5) = \inf_{W_0 \in S} \wedge_{k=1}^K \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu d_k \geq \vee_{l=1}^K \mathbb{E}_{\mu_0} d_l}} D(\mu || \mu_0^p). \tag{5.6}$$

Proof. From (3.15), decoding with the metrics $\{d_k\}_{k=1}^K$ when the true channel is W_0, can achieve rates as high as

$$\underline{C}(P_{\mathcal{X}}, W_0) = \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \vee_{k=1}^K \mathbb{E}_\mu d_k \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} d_k}} D(\mu || \mu_0^p) \tag{5.7}$$

Therefore, if $\underline{C}(P_{\mathcal{X}}, W_0) > 0$ and if $R < \underline{C}(P_{\mathcal{X}}, W_0)$, there exists a function $(R, W_0) \mapsto \widetilde{E}_r(R, P_{\mathcal{X}}, W_0) > 0$ such that

$$\mathbb{E}_{P_{\mathcal{X}}^{Mn}} P_e(\{X_m\}, \{d_k\}, W_0) \leq e^{-n\widetilde{E}_r(R, P_{\mathcal{X}}, W_0)},$$

where $P_e(\{X_m\}, \{d_k\}, W_0)$ is the random error probability, for a constant composition random code $\{X_m\}_{m=1}^M$ of distribution $P_{\mathcal{X}}$, decoding with the metrics $\{d_k\}_{k=1}^K$, when the true channel is W_0. Taking now the proof written for theorem 7, replacing the MMI decoder by the linear decoder induced by the metrics $\{d_k\}_{k=1}^K$ and replacing $E_r(R, P_{\mathcal{X}}, \cdot)$ by $\widetilde{E}_r(R, P_{\mathcal{X}}, \cdot)$, we obtain a proof of theorem 8. □

5.2.4 Problem Formulation

With this theorem, our problem can be addressed by choosing K, $\{d_k\}_{k=1}^K$ in order to solve

$$\sup_{K \in \mathbb{Z}_+} \sup_{\{d_k\}_{k=1}^K} \sup_{P_{\mathcal{X}} \in M_1(\mathcal{X})} \inf_{W_0 \in S} \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \vee_{k=1}^K \mathbb{E}_\mu d_k \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} d_k}} D(\mu || \mu_0^p) \tag{5.8}$$

and we are interested in achieving (5.8) for a finite K. We know that (5.8) is upper bounded by $\max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} \inf_{W \in S} I(P_{\mathcal{X}}, W)$, we can then ask if there are sets S, for which we have existence of $K = K(S) < +\infty$ and $\{d_k\}_{k=1}^K$ such that

$$\inf_{W_0 \in S} \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \vee_{k=1}^K \mathbb{E}_\mu d_k \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} d_k}} D(\mu || \mu_0^p) = \inf_{W_0 \in S} I(P_{\mathcal{X}}, W_0), \text{ for any or some } P_{\mathcal{X}}$$

and if so, we would like to find the most general characterization of such sets and an explicit construction of the metrics $\{d_k\}_{k=1}^K$.

5.3 Very Noisy Case

In this section, we express the problem mentioned above through the VN transformation around a common limiting distribution $P_{\mathcal{Y}} \in M_1(\mathcal{Y})$, and $P_{\mathcal{X}} \in M_1(\mathcal{X})$ is fixed.

- The VN transformation of the set S is denoted by

$$S_\varepsilon = \{P_{\mathcal{Y}}(1+\varepsilon L) | L \in \mathcal{S}\}$$

where $\mathcal{S} \subset M_0(P_{\mathcal{Y}})$, so that any elements $W \in S$ has the VN transformation $W_\varepsilon = P_{\mathcal{Y}}(1+\varepsilon L)$, $L \in \mathcal{S}$. For any ε, the set S_ε is convex, respectively compact, if and only if the set $\mathcal{S}$ is convex, respectively compact.

The joint distribution μ of any $W \in S$ with marginal $P_{\mathcal{X}}$ will then have the VN transformation

$$\mu_\varepsilon = P_{\mathcal{X}} P_{\mathcal{Y}}(1+\varepsilon L), \quad L \in \mathcal{S}$$

- When a subscript k is used to denote a channel W_k, the VN transformation is denoted by

$$W_{k,\varepsilon} = P_{\mathcal{Y}}(1+\varepsilon L_k).$$

- We keep the subscript $k = 0$ for the true channel $W_0 \in S$, whose VN transformation is denoted by

$$W_{0,\varepsilon} = P_Z(1+\varepsilon L_0), \quad L_0 \in \mathcal{S}.$$

- If one considers the metrics to be the log of some channels, i.e., $d_k = \log W_k$, the VN transformations are denoted by

$$d_{k,\varepsilon} = \log W_{k,\varepsilon} = \log(P_{\mathcal{Y}}) + \log(1+\varepsilon L_k),$$

where $L_k \in M_0(P_{\mathcal{Y}})$ may not have to be in $\mathcal{S}$.

In the following

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{P_{\mathcal{X}} \times P_{\mathcal{Y}}}.$$

Using 4.4, we have the following result.

Proposition 12. *For $\mathcal{S}$ compact, we have*

$$\min_{W \in S_\epsilon} I(P_{\mathcal{X}}, W) = \frac{\varepsilon^2}{2} \min_{L \in \mathcal{S}} ||\widetilde{L}||^2 + o(\varepsilon^2),$$

where

$$\min_{L \in \mathcal{S}} ||\widetilde{L}||^2 = ||\widetilde{L}_{\mathcal{S}}||^2$$

is called the VN compound mutual information on $\mathcal{S}$.

5.3.1 One-sided Sets

Let us start by considering $K = 1$, i.e., when only one metric is used. We recall here proposition 5.

Proposition 13. *Let $W_{0,\epsilon} = P_{\mathcal{Y}}(1 + \epsilon L_0)$ and decode with the metric $d = \log W_{1,\epsilon}$, where $W_{1,\epsilon} = P_{\mathcal{Y}}(1 + \epsilon L_1)$. Then,*

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{Y,\epsilon} \\ \mathbb{E}_\mu \log W_{1,\epsilon} \geq \mathbb{E}_{\mu_{0,\epsilon}} \log W_{1,\epsilon}}} D(\mu || \mu^p_{0,\epsilon})$$

$$= \frac{1}{2} \inf_{\substack{V:\, \tilde{V} = \tilde{V} = 0 \\ \langle V, \widetilde{L}_1 \rangle \geq \langle \widetilde{L}_0, \widetilde{L}_1 \rangle}} ||V||^2 = \frac{1}{2} \frac{\langle \widetilde{L}_0, \widetilde{L}_1 \rangle^2}{||\widetilde{L}_1||^2}$$

As explained earlier, this result says that the mismatched mutual information obtained when decoding with the mismatched metric $\log W_{1,\epsilon}$, whereas the true channel is $W_{0,\epsilon}$, is approximately the projection (squared norm) of the true channel direction $\widetilde{L}_0$ onto the mismatched direction $\widetilde{L}_1$. This result gives an intuitive picture of the mismatched mutual information. As expected, if the decoder is matched, i.e., $\widetilde{L}_0 = \widetilde{L}_1$, the projection squared norm is $||\widetilde{L}_0||^2$, which is the very noisy mutual information of $\widetilde{L}_0$, and the more orthogonal $\widetilde{L}_1$ is to $\widetilde{L}_0$, the worse the mismatched decoding rule is.

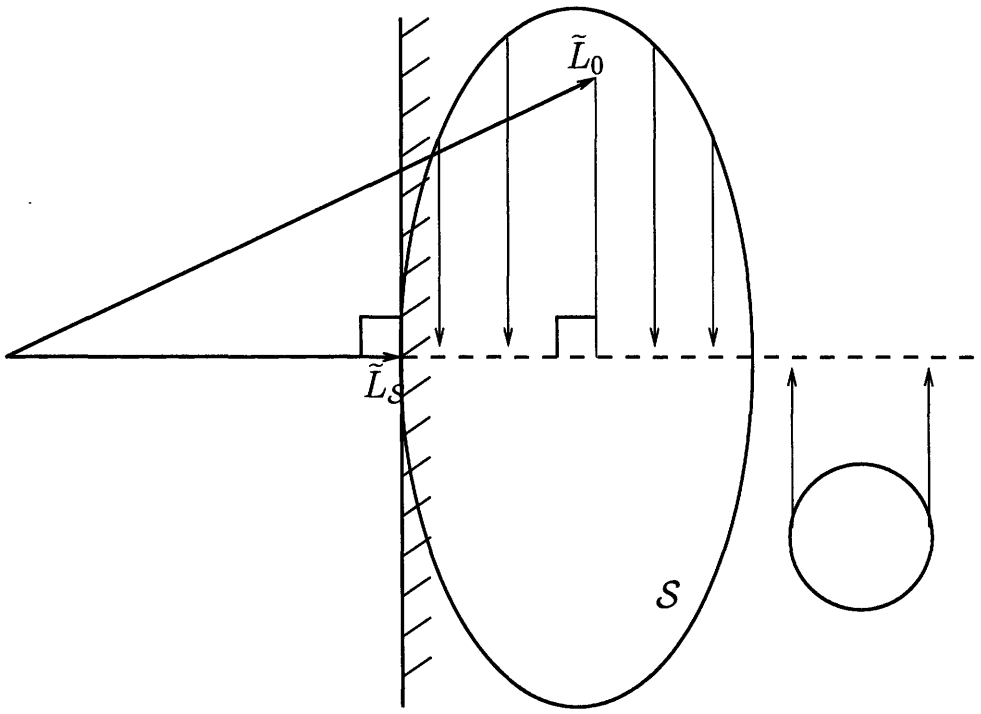

Figure 5-1: Very noisy one-sided compound set

This picture of the mismatched mutual information directly suggests a first result. Let $\widetilde{\mathcal{S}} = \{L - \bar{L} | S \in \mathcal{S}\}$. Assume $\mathcal{S}$, hence $\widetilde{\mathcal{S}}$, to be convex. By using the worse channel to be the only decoding metric, it is then clear that the VN compound capacity can be achieved. In fact, no matter what the true channel $\widetilde{L}_0 \in \widetilde{\mathcal{S}}$ is, the mismatched mutual information given by the projection of $\widetilde{L}_0$ onto $\widetilde{L}_{\mathcal{S}}$ cannot be shorter than $||\widetilde{L}_{\mathcal{S}}||$, which is the very noisy compound capacity of $\mathcal{S}$. This is illustrated in figure 5-1. Moreover, the notion of convexity is not necessary. As long as the compound set is such that its projection in the direction of the minimal vector stays on one side, i.e., if the compound set is entirely contained in the half space delimited by the normal plan to the minimal vector:

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_{\mathcal{S}} \rangle^2}{||\widetilde{L}_{\mathcal{S}}||^2} \geq ||\widetilde{L}_{\mathcal{S}}||^2, \quad \forall L_0 \in \mathcal{S},$$

we will be universal with the worse channel metric (cf. figure 5-1 where S can include

the circle set without violating above condition). We call such sets one-sided sets and by rewriting the property in terms of norms, we get the following definition.

Definition 23. A compact set $\mathcal{S} \subset M_0(P_{\mathcal{Y}})$ is one-sided if

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_{\mathcal{S}} \rangle^2}{||\widetilde{L}_{\mathcal{S}}||^2} \geq ||\widetilde{L}_{\mathcal{S}}||^2, \quad \forall L_0 \in \mathcal{S},$$

where it is sufficient to ensure this property for the $\mathcal{X}$-marginal given by

$$P_{\mathcal{X}} = \arg \max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} \min_{L \in \mathcal{S}} ||\widetilde{L}||^2.$$

Proposition 14. *In the VN setting, decoding with the worst channel metric is universal for one-sided sets - and is linear.*

It is also clear that for a non-onsided set, decoding with one metric gives a mismatched random coding capacity which must be less than the compound capacity. However, the mismatched random coding capacity is known to be tight (i.e., the mismatched capacity) only for binary channels, hence only for binary channels, we can state that decoding with one metric on non-onesided sets cannot achieve compound capacity for very noisy channels, and the same is to be expected in general and for non-binary channels.

5.3.2 Finite Sets

Let us consider a simple case of non one-sided set, namely when S contains only two channels not satisfying the one-sided property, say

$$S = \{W_1, W_2\}.$$

A first idea would be to use the metrics $d_1 = \log W_1$ and $d_2 = \log W_2$, i.e., decoding with the GLRT test using both channels

$$\arg \max_{x_m} W_1^n(y|x_m) \vee W_2^n(y|x_m).$$

Let us assume w.l.o.g. that $W_0 = W_1$. In this situation, one of the two test will be an optimal ML test with the true channel, but the other test is a ML test with a channel that has nothing to do with the true channel, and we need to estimate how probable it is that a codeword which has not been sent looks too typical under that other channel (i.e., an error event). In other words, we need to compare the values of the two projections that formula (5.6) provides. We already know that one of the projection's norm is the mutual information of W_0. We need to check if the other projection's norm can be smaller or must always be greater. We check this here in the very noisy setting. Using the same notations as previously, we have that the second projection becomes, in the VN setting,

$$\inf_{\substack{L:\, \bar{\tilde{L}}=0,\, \bar{L}=\bar{L}_0 \\ \langle L, L_1\rangle \geq \frac{1}{2}(||L_0||^2+||L_1||^2)}} ||L-\bar{L}_0||^2 \geq ||\widetilde{L}_0||^2 \wedge ||\widetilde{L}_1||^2 \tag{5.9}$$

or equivalently

$$\inf_{\substack{V:\, \bar{V}=\bar{\tilde{V}}=0 \\ \langle V, \bar{L}_1\rangle \geq \frac{1}{2}(||L_0||^2+||L_1||^2)-\langle \bar{L}_0, \bar{L}_1\rangle}} ||V||^2 \geq ||\widetilde{L}_0||^2 \wedge ||\widetilde{L}_1||^2 \tag{5.10}$$

But

$$\begin{aligned} &\frac{1}{2}(||L_0||^2+||L_1||^2) - \langle \bar{L}_0, \bar{L}_1\rangle \\ &= \frac{1}{2}(||\widetilde{L}_0||^2+||\widetilde{L}_1||^2+||\bar{L}_0-\bar{L}_1||^2), \end{aligned}$$

so we need to show

$$\frac{1}{2}(||\widetilde{L}_0||^2+||\widetilde{L}_1||^2+||\bar{L}_0-\bar{L}_1||^2) \geq ||\widetilde{L}_0||\,||\widetilde{L}_1|| \wedge ||\widetilde{L}_1||^2,$$

which clearly holds.

This can be directly generalized to any finite sets and we have the following result.

Proposition 15. *In the VN setting, GLRT with all channels in the set is universal for finite compound sets - and linear.*

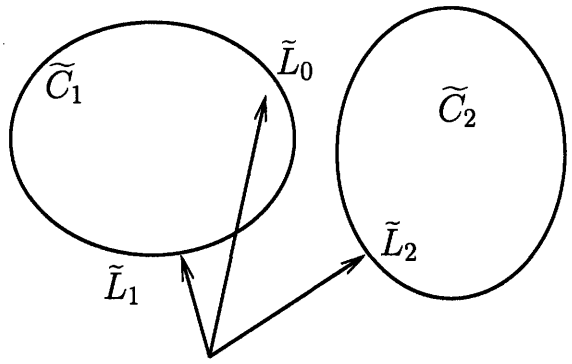

Figure 5-2: Union of two one-sided components

5.3.3 Finite Union of One-sided Sets

Using ML Metrics

We found a linear universal decoding rule for one-sided sets and for finite sets. Thus, the next sets that we consider are finite unions of one-sided sets. Assume

$$S = C_1 \cup C_2,$$

where C_1 and C_2 are one-sided. Let $W_1 = W_{C_1}$ and $W_2 = W_{C_2}$ (cf. figure 5-2). A legitimate candidate for a linear universal decoder would be to use the GLRT with metrics W_1 and W_2, hoping that a combination of earlier results for finite and one-sided sets will make this decoding rule capacity achieving.
Say w.l.o.g. that $W_0 \in C_1$ and using (5.6), let us try to verify the following

$$\begin{aligned} &\wedge_{k=1}^{2} \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu \log W_k \geq \vee_{l=1}^{2} \mathbb{E}_{\mu_0} \log W_l}} D(\mu || \mu_0^p) \\ &\stackrel{?}{=} I(P_{\mathcal{X}}, W_1) \wedge I(P_{\mathcal{X}}, W_2). \end{aligned}$$

As opposed to the finite compound set case, we cannot decide in general which one of the threshold tests $\mathbb{E}_{\mu_0} \log W_1$ or $\mathbb{E}_{\mu_0} \log W_2$, is the maximum. But no matter what

this maximum is, the result of the $k = 1$ projection must be greater than $I(P_{\mathcal{X}}, W_1)$, since when $\vee_{l=1}^{2} \mathbb{E}_{\mu_0} \log W_l = \mathbb{E}_{\mu_0} \log W_1$, we can use the one-sided result (because we assumed that $W_0 \in C_1$), and if $\vee_{l=1}^{2} \mathbb{E}_{\mu_0} \log W_l = \mathbb{E}_{\mu_0} \log W_2$, the resulting projection can only be larger than the one in the one-sided case. So all we need to check is that the second projection does not get too short, i.e.,

$$\inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu \log W_2 \geq \vee_{l=1}^{2} \mathbb{E}_{\mu_0} \log W_l}} D(\mu || \mu_0^p) \overset{?}{\geq} I(P_{\mathcal{X}}, W_1) \wedge I(P_{\mathcal{X}}, W_2). \tag{5.11}$$

Let us first understand what is the meaning of $\mathbb{E}_{\mu_0} \log W_1 \leq \mathbb{E}_{\mu_0} \log W_2$ in the very noisy setting. With usual notations, we get

$$\mathbb{E}_{\mu_0} \log W_1 \leq \mathbb{E}_{\mu_0} \log W_2 \overset{\text{VN}}{\longrightarrow} ||L_0 - L_2|| \leq ||L_0 - L_1||.$$

Let us assume that we are in a situation where $||L_0 - L_2|| = ||L_0 - L_1||$, then we know that the expression (5.11) in the very noisy setting is

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_2 \rangle}{||\widetilde{L}_2||} \overset{?}{\geq} ||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||,$$

and say that $||\widetilde{L}_1|| = ||\widetilde{L}_2||$. Since $W_0 \in C_1$, we do not have any reason to believe that

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_2 \rangle}{||\widetilde{L}_2||} \geq ||\widetilde{L}_2||,$$

since we do not have the one-sided property for W_0 and C_2. However, we can still hope that this holds **when** we impose $||L_0 - L_2|| = ||L_0 - L_1||$. We can write

$$\begin{aligned}
&||\widetilde{L}_0||^2 - ||\widetilde{L}_2||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2 \\
&= ||\widetilde{L}_0||^2 - ||\widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_1||^2 \\
&\quad + ||\widetilde{L}_0 - \widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2 \\
&= ||\widetilde{L}_0||^2 - ||\widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_1||^2 \\
&\quad + ||\bar{L}_0 - \bar{L}_2||^2 - ||\bar{L}_0 - \bar{L}_1||^2,
\end{aligned}$$

but even with $||L_0 - L_2|| = ||L_0 - L_1||$, we cannot control $||\bar{L}_0 - \bar{L}_2||^2 - ||\bar{L}_0 - \bar{L}_1||^2$, and this could indeed be negative. This suggests that the result may not hold. In fact, we have the following.

Proposition 16. *In the VN setting and for compound sets having a finite number of one-sided components, GLRT with the worse channel of each component is not universal.*

Counter-example: Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $P_{\mathcal{X}} = P_{\mathcal{Y}} = \{1/2, 1/2\}$,

$$L_0 = \begin{pmatrix} -2 & 2 \\ -7 & 7 \end{pmatrix}, L_1 = \begin{pmatrix} 2 & -2 \\ 0 & 0 \end{pmatrix} \text{ and } L_2 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Note that $||\widetilde{L}_0||^2 - ||\widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_1||^2 \geq 0$, hence $C_1 = \{W_0, W_1\}$ is a one-sided component and $\widetilde{L}_2$ does not belong to that component, hence we define $C_2 = \{\widetilde{L}_2\}$. We chose the direction such that $||L_0 - L_1|| = ||L_0 - L_2||$ and $||\widetilde{L}_1||^2 = ||\widetilde{L}_2||^2$, to simplify the counter-example (but it is not necessary). Finally, we have

$$||\widetilde{L}_0||^2 - ||\widetilde{L}_2||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2 < 0,$$

which means that we are loosing rate compared to the compound capacity. This counter-examples is picture in figure 5-3. In this picture, the two dimensional plane of the background contains the non-tilde vectors and the line with a negative slope contains the tilde vectors. Each time a vector is projected into this line, the height of the projection is the bar-component. We can see that having this extra degrees in the plan for non-tilde vectors, we can pick L_1 and L_2 to be equidistant from L_0 (they are indeed on the same circle centered at L_0), in a way that their projection on the tilde-vectror space are very different, namely, $\widetilde{L}_0$ is opposite to $\widetilde{L}_2$, violating the one-sided property.

A counter-example in the very noisy setting is of course sufficient to prove the proposition.

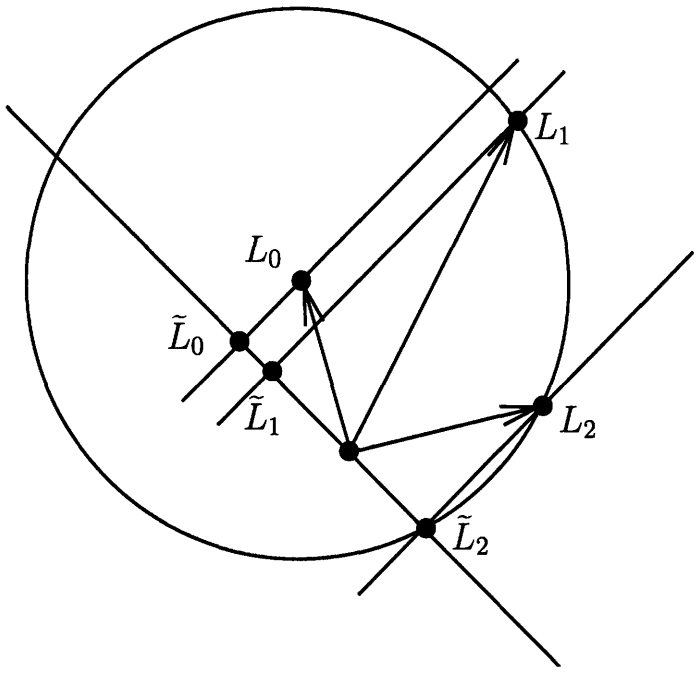

Figure 5-3: Counter-example for ML metrics

Using MAP Metrics

We now use different metrics than the one used in previous section, instead of the ML metrics given by $\log W_k$, we use the metrics

$$\log \frac{W_k}{(\mu_k)_Y}, \tag{5.12}$$

which we call the MAP metrics for maximum a posteriori (although they should be called the "a posteriori" metrics) and which are also sometimes referred to the Fano metrics in the literature.

As before, let us consider W_0, W_1 and W_2 such that W_1 and W_2 are the worst channels of two one-sided components C_1 and C_2, and W_0 belongs to C_1. We want

to show that

$$\inf_{\substack{\mu:\,\mu_X=P_{\mathcal{X}},\,\mu_Y=(\mu_0)_Y \\ \mathbb{E}_\mu \log \frac{W_2}{(\mu_2)_Y} \geq \vee_{k=1}^2 \mathbb{E}_{\mu_0} \log \frac{W_k}{(\mu_k)_Y}}} D(\mu||\mu_0^p) \geq \wedge_{k=1}^K I(P_{\mathcal{X}}, W_k), \tag{5.13}$$

since the other projection in the $\log \frac{W_1}{(\mu_1)_Y}$ direction is known to be greater than $I(P_{\mathcal{X}}, W_1)$ from the one-sided property.

Note that

$$\mathbb{E}_\mu \log \frac{W_2}{(\mu_2)_Y} \xrightarrow{VN} \langle \widetilde{L}, \widetilde{L}_2 \rangle - \frac{1}{2}||\widetilde{L}_2||^2 = \frac{1}{2}(||\widetilde{L}||^2 - ||\widetilde{L} - \widetilde{L}_2||^2).$$

Therefore, the VN transformation of (5.13)

$$\inf_{\substack{L:\,\bar{\bar{L}}=0,\,\bar{L}=\bar{L}_0 \\ \langle \bar{L}, \bar{L}_2 \rangle \geq \frac{1}{2}(||\bar{L}_0||^2 + ||\bar{L}_2||^2 - ||\bar{L}_0 - \bar{L}_1||^2 \wedge ||\bar{L}_0 - \bar{L}_2||^2)}} ||L - \bar{L}_0||^2 \geq ||\widetilde{L}_1||^2 \wedge ||\widetilde{L}_2||^2. \tag{5.14}$$

So the two cases to check are

$$\frac{\frac{1}{2}(||\widetilde{L}_0||^2 + ||\widetilde{L}_2||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2)}{||\widetilde{L}_2||} = \frac{\langle \widetilde{L}_0, \widetilde{L}_2 \rangle}{||\widetilde{L}_2||} \geq ||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||,$$

if $||\widetilde{L}_0 - \widetilde{L}_2||^2 \leq ||\widetilde{L}_0 - \widetilde{L}_1||^2$; and

$$\frac{\frac{1}{2}(||\widetilde{L}_0||^2 + ||\widetilde{L}_2||^2 - ||\widetilde{L}_0 - \widetilde{L}_1||^2)}{||\widetilde{L}_2||} \geq ||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||,$$

if $||\widetilde{L}_0 - \widetilde{L}_1||^2 \leq ||\widetilde{L}_0 - \widetilde{L}_2||^2$.

In fact we will check that both inequalities hold with $||\widetilde{L}_1||$ instead of $||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||$ on the right hand side.

Let us investigate the first inequality. As it was the case for the GLRT decoding rule, the following inequality

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_2 \rangle}{||\widetilde{L}_2||} \stackrel{?}{\geq} ||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||,$$

does not hold by assumption, since we assume that $W_0 \in C_1$, and not $W_0 \in C_2$. But

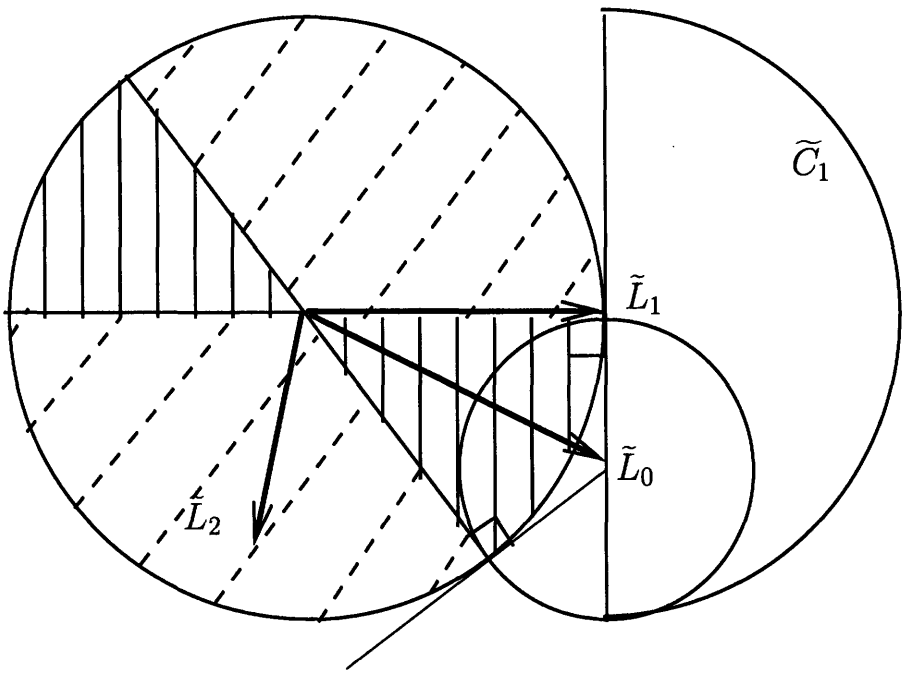

Figure 5-4: Bad projection regions

as opposed to the GLRT case, we ask that this inequality holds **if** $||\widetilde{L}_0 - \widetilde{L}_1||^2 \leq ||\widetilde{L}_0 - \widetilde{L}_2||^2$. As illustrated in figure 5-4, if $\widetilde{L}_2$ is crossing the dashed region, the projection of $\widetilde{L}_0$ onto $\widetilde{L}_2$ can be smaller than $||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||$,

but because we assume that $||\widetilde{L}_0 - \widetilde{L}_1||^2 \leq ||\widetilde{L}_0 - \widetilde{L}_2||^2$, $\widetilde{L}_2$ must be in the small circle, thus, it cannot belong to the dashed region and the projection must be greater than $||\widetilde{L}_1||$, as illustrated in figure 5-5.

One can also check this analytically. For the first case:

$$\begin{aligned}
& \frac{\frac{1}{2}(||\widetilde{L}_0||^2 + ||\widetilde{L}_2||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2)}{||\widetilde{L}_2||} - ||\widetilde{L}_1|| \\
= & \frac{\frac{1}{2}(||\widetilde{L}_0||^2 + ||\widetilde{L}_2||^2 - 2||\widetilde{L}_1||||\widetilde{L}_2|| - ||\widetilde{L}_0 - \widetilde{L}_2||^2)}{||\widetilde{L}_2||} \\
= & \frac{\frac{1}{2}((||\widetilde{L}_1|| - ||\widetilde{L}_2||)^2 + ||\widetilde{L}_0||^2 - ||\widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2)}{||\widetilde{L}_2||} \\
\geq & \frac{\frac{1}{2}((||\widetilde{L}_1|| - ||\widetilde{L}_2||)^2 + ||\widetilde{L}_0||^2 - ||\widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_1||^2)}{||\widetilde{L}_2||}
\end{aligned}$$

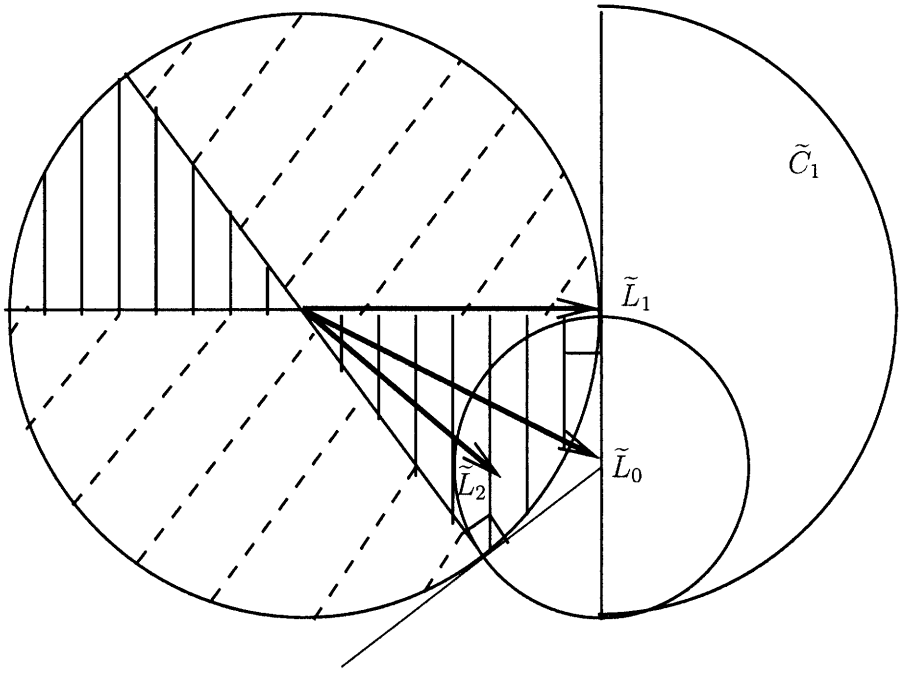

Figure 5-5: Good projection regions

where last inequality uses the assumption $||\widetilde{L}_0 - \widetilde{L}_2||^2 \leq ||\widetilde{L}_0 - \widetilde{L}_1||^2$. Finally, using the one-sided property, last expression is clearly positive.
For the second case, the same expansion as before gets us directly to

$$\frac{\frac{1}{2}((||\widetilde{L}_1|| - ||\widetilde{L}_2||)^2 + ||\widetilde{L}_0||^2 - ||\widetilde{L}_1||^2 - ||\widetilde{L}_0 - \widetilde{L}_1||^2)}{||\widetilde{L}_2||}.$$

Hence both cases are satisfied.

Uniqueness of MAP Normalization

Roughly speaking, we saw that combining the worse ML metrics fails to be canonical for union of one-sided components, but that normalizing the ML metrics properly, i.e., considering MAP metrics, clears the problem. Is the MAP normalization the only one that works? Why do we have to use this normalization as opposed to any other one? We discuss these questions for the very noisy setting.

Let us assume that, instead of normalizing the ML metrics with $(\mu_k)_Y$, we normalize it with an arbitrary function around $P_{\mathcal{Y}}$, i.e., $P_{\mathcal{Y}}(1 + \epsilon M_k)$, where M_k is a function of y only. We then have

$$\begin{aligned}
\mathbb{E}_{\mu_0} \log \frac{P_{\mathcal{Y}}(1+\epsilon L_2)}{P_{\mathcal{Y}}(1+\epsilon M_2)} &= \sum P_{\mathcal{X}} P_{\mathcal{Y}}(1+\epsilon L_0)(\epsilon(L_2 - M_2) - \frac{\epsilon^2}{2}(L_2^2 - M_2^2)) \\
&\xrightarrow{VN} \langle L_0, L_2 - M_2\rangle - \frac{1}{2}(||L_2||^2 - ||M_2||^2) \\
&= \langle \widetilde{L}_0, \widetilde{L}_2\rangle - \frac{1}{2}||\widetilde{L}_2||^2 + \langle \bar{L}_0, \bar{L}_2\rangle \\
&\quad -\frac{1}{2}||\bar{L}_2||^2 - \langle \bar{L}_0, M_2\rangle + \frac{1}{2}||M_2||^2
\end{aligned}$$

Hence, the first projection inequality that we checked in previous section, i.e., the projection of $\widetilde{L}_0$ onto $\widetilde{L}_2$ is greater than $||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||$, must now hold **if**

$$||\widetilde{L}_0 - \widetilde{L}_2||^2 + ||\bar{L}_0 - \bar{L}_2||^2 - ||\bar{L}_0 - M_2||^2 \tag{5.15}$$

$$< \; ||\widetilde{L}_0 - \widetilde{L}_1||^2 + ||\bar{L}_0 - \bar{L}_1||^2 - ||\bar{L}_0 - M_1||^2. \tag{5.16}$$

Let us define

$$\delta = ||\bar{L}_0 - \bar{L}_1||^2 - ||\bar{L}_0 - M_1||^2 - (||\bar{L}_0 - \bar{L}_2||^2 - ||\bar{L}_0 - M_2||^2)$$

and let us consider binary channels with $P_{\mathcal{X}} = P_{\mathcal{Y}} = (1/2, 1/2)$. In this case, the tilde vectors ($\widetilde{L}$) are all co-linear, on a same line:

$$\lambda \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \lambda \in \mathbb{R}.$$

The non-tilde (L) vectors are

$$\begin{pmatrix} a & -a \\ -b & b \end{pmatrix}, \quad a, b \in \mathbb{R}.$$

We can represent this case as pictured in figure 5-3, where the two dimensional plane of the background contains the non-tilde vectors and the line with a negative slope contains the tilde vectors. Each time a vector is projected into this line, the height of the projection is the bar-component. Say we are in the case where (5.16) holds with equality (this can be assumed). We want to get that

$$||\widetilde{L}_0||^2 - ||\widetilde{L}_2||^2 - ||\widetilde{L}_0 - \widetilde{L}_2||^2 < 0,$$

which is to say that the first projection inequality is not satisfied, and we need do this without violating (5.16). But, unless $\delta = 0$, we can always play with the heights of the bar-components of the L_i's to ensure that (5.16) holds and make $\widetilde{L}_2$ be on the other side of the origin of $\widetilde{L}_0$ and $\widetilde{L}_1$ (which gives a "bad" projection). So allowing an extra degree of freedom with the bar-vectors gives us an extra dimension with which we can play to violate the projection inequality.

Also, if we use other kinds of normalization, such as $P_{\mathcal{Y}}^2(1 + \epsilon \bar{L}_k)^2$, then we need

to ensure that the projection of $\widetilde{L}_0$ onto $\widetilde{L}_2$ is greater than $||\widetilde{L}_1|| \wedge ||\widetilde{L}_2||$, **if**

$$||\widetilde{L}_0 - \widetilde{L}_2||^2 - ||\bar{L}_0 - \bar{L}_2||^2 < ||\widetilde{L}_0 - \widetilde{L}_1||^2 - ||\bar{L}_0 - \bar{L}_1||^2,$$

so we can now define

$$\delta = ||\bar{L}_0 - \bar{L}_2||^2 - ||\bar{L}_0 - \bar{L}_1||^2$$

and the issue is raised. It is only when the condition is

$$||\widetilde{L}_0 - \widetilde{L}_2||^2 < ||\widetilde{L}_0 - \widetilde{L}_1||^2$$

that the projection inequality is always satisfied. And this is achieved "only" when the normalization is the MAP one.

5.4 General Case

5.4.1 The Results

The previous section gives us a series of results regarding linear decoders on different kinds of compound sets, in the very noisy setting. In this section, our goal is to verify which one of those results can be generalized to the non very noisy setting; we will see that all of them can actually be generalized. We already know that the negative results, i.e., statements not holding in the very noisy setting, will not hold as well in the general setting. In this section we list all the general results and in the next section, we illustrate how we can lift the very noisy results to achieve the general ones. The formal proofs are given in section 5.4.3.

Recall: We define the optimal input distribution of a set S by

$$P_{\mathcal{X}} = \arg \max_{P \in M_1(\mathcal{X})} \min_{W \in S} I(P; W),$$

and if the maximizers are not unique, we define $P_{\mathcal{X}}$ to be any arbitrary maximizer.

Definition 24. A set S is one-sided, if

$$D(\mu_0||\mu_S^p) \geq D(\mu_0||\mu_S) + D(\mu_S||\mu_S^p), \quad \forall W_0 \in S. \tag{5.17}$$

where

$$W_S = \arg \min_{W \in \text{cl}(S)} I(P_X, W). \tag{5.18}$$

Remark: (5.17) cannot hold if the minimizer in (5.18) is not unique.

Proposition 17. *For one-sided sets S, decoding with the metric $d = \log W_S$ is universal - and linear.*

This result is proved in [11] for convex sets.

Lemma 5. *Convex sets are one-sided and there are one-sided sets that are non-convex.*

Proposition 18. *For any set S, decoding with the metrics $\{\log W\}_{W \in S}$, i.e., maximizing $D^n = \max_{W \in S} \log W^n$, is universal, but generalized linear only if S is finite.*

Proposition 19. *For $S = \cup_{k=1}^K C_k$, where $\{C_k\}_{k=1}^K$ are one-sided sets, decoding with the metrics $d_k = \log W_{C_k}$, for $1 \leq k \leq K$, is not universal.*

Theorem 9. *For $S = \cup_{k=1}^K C_k$, where $\{C_k\}_{k=1}^K$ are one-sided sets, decoding with the metrics $d_k = \log \frac{W_{C_k}}{(\mu_{C_k})_{\mathcal{Y}}}$, for $1 \leq k \leq K$ is universal - and generalized linear.*

5.4.2 Lifting

In this section, we illustrate how the results and proofs obtained in the very noisy setting can be lifted to results and proofs in the general setting. We consider the case of one-sided sets, and we use the definitions made in section 5.3.1. In the very noisy setting, a one-sided set $\mathcal{S}$ is such that

$$\frac{\langle \widetilde{L}_0, \widetilde{L}_S \rangle^2}{||\widetilde{L}_S||^2} \geq ||\widetilde{L}_S||^2, \quad \forall L_0 \in \mathcal{S}.$$

The first step in order to lift this definition, is to write it with norms instead of inner products, i.e.,

Definition 25. A compact set $\mathcal{S} \subset M_0(P_\mathcal{Y})$ is one-sided if

$$||\widetilde{L}_0||^2 - ||\widetilde{L}_\mathcal{S}||^2 - ||\widetilde{L}_0 - \widetilde{L}_\mathcal{S}||^2 \geq 0, \quad \forall L_0 \in \mathcal{S}, \tag{5.19}$$

where it is sufficient to ensure this property for the $\mathcal{X}$-marginal given by

$$P_\mathcal{X} = \arg \max_{P_\mathcal{X} \in M_1(\mathcal{X})} \min_{L \in \mathcal{S}} ||\widetilde{L}||^2.$$

Now that we understand the concept of one-sided sets in terms of inequalities on "norms", we will lift the definition to the general case. Recall that

$$D(\mu_0||\mu_0^p) \xrightarrow{VN} ||L_0 - \bar{L}_0||^2 = ||\widetilde{L}_0||^2$$

and

$$D(\mu_\mathcal{S}||\mu_\mathcal{S}^p) \xrightarrow{VN} ||L_\mathcal{S} - \bar{L}_\mathcal{S}||^2 = ||\widetilde{L}_\mathcal{S}||^2,$$

therefore, we have candidates for the first two norms appearing in (5.19) and we would like to lift $||\widetilde{L}_0 - \widetilde{L}_\mathcal{S}||^2$. We know that

$$D(\mu_0||\mu_\mathcal{S}) \xrightarrow{VN} ||L_0 - L_\mathcal{S}||^2$$

and

$$D(\mu_0^p||\mu_\mathcal{S}^p) \xrightarrow{VN} ||\bar{L}_0 - \bar{L}_\mathcal{S}||^2,$$

hence

$$D(\mu_0||\mu_\mathcal{S}) - D(\mu_0^p||\mu_\mathcal{S}^p) \xrightarrow{VN} ||L_0 - L_\mathcal{S}||^2 - ||\bar{L}_0 - \bar{L}_\mathcal{S}||^2 = ||\widetilde{L}_0 - \widetilde{L}_\mathcal{S}||^2 \geq 0 \tag{5.20}$$

where the last equality simply uses the projection principle, i.e., that the projection of L onto centered directions $\widetilde{L} = L - \bar{L}$ is orthogonal to the projection's height $\bar{L}$,

implying

$$\|\widetilde{L}\|^2 = \|L\|^2 - \|\bar{L}\|^2.$$

Therefore,

$$D(\mu_0||\mu_0^p) - D(\mu_S||\mu_S^p) - (D(\mu_0||\mu_S) - D(\mu_0^p||\mu_S^p)) \geq 0 \tag{5.21}$$

is a lifting of (5.19). Hence, let us assume that S is satisfying (5.21) and let us use the metric $\log W_S$; we now want to see if this is still a capacity achieving decoding rule on such set S. In order to lift the proof, let us understand it in terms of norms. The VN mismatched mutual information is given by

$$\inf_{\substack{V:\, \bar{V}=\tilde{V}=0 \\ \langle V, \widetilde{L}_S\rangle \geq \langle \widetilde{L}_0, \widetilde{L}_S\rangle}} ||V||^2 \tag{5.22}$$

$$= \inf_{\substack{V:\, \bar{V}=\tilde{V}=0 \\ ||V||^2 - ||V-\widetilde{L}_S||^2 \geq ||\widetilde{L}_0||^2 - ||\widetilde{L}_0 - \widetilde{L}_S||^2}} ||V||^2 \tag{5.23}$$

By looking at the constraints of last expression, and since

$$||V - \widetilde{L}_S||^2 \geq 0,$$

we clearly have

$$(5.23) \quad \geq ||\widetilde{L}_0||^2 - ||\widetilde{L}_0 - \widetilde{L}_S||^2. \tag{5.24}$$

In the general setting, the mismatched mutual information is given by

$$\inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu \log W_S \geq \mathbb{E}_{\mu_0} \log W_S}} D(\mu||\mu^p) \tag{5.25}$$

expressing the quantities of interest in terms of divergences, i.e., rewriting $\mathbb{E}_\mu \log W_S = \mathbb{E}_\mu \log W_S \frac{\mu}{\mu} \frac{\mu^p}{\mu^p}$ and using the fact that $\mu^p = \mu_0^p$ (the marginal constraints), last expres-

sion is equivalent to

$$\inf_{\substack{\mu:\,\mu_{\mathcal{X}}=P_{\mathcal{X}},\,\mu_{\mathcal{Y}}=(\mu_0)_{\mathcal{Y}} \\ D(\mu||\mu^p)-D(\mu||\mu_S)+D(\mu^p||\mu_S^p) \\ \geq D(\mu_0||\mu_0^p)-D(\mu_0||\mu_S)+D(\mu_0^p||\mu_S^p)}} D(\mu||\mu^p). \tag{5.26}$$

Therefore, we are replicating the very noisy proof in a very parallel way, and if

$$D(\mu||\mu_S) - D(\mu^p||\mu_S^p) \geq 0,$$

we conclude that

$$(5.26) \quad \geq \quad D(\mu_0||\mu_0^p) - D(\mu_0||\mu_S) + D(\mu_0^p||\mu_S^p), \tag{5.27}$$

and by definition of one-sided sets, we have

$$D(\mu_0||\mu_0^p) - D(\mu_0||\mu_S) + D(\mu_0^p||\mu_S^p) \geq D(\mu_S||\mu_S^p), \tag{5.28}$$

which is the compound capacity. So we need to show that

$$D(\mu||\mu_S) - D(\mu^p||\mu_S^p) \geq 0, \tag{5.29}$$

which was, as expressed in 5.20, clear for the very noisy setting since $||L_0 - L_S||^2 - ||\bar{L}_0 - \bar{L}_S||^2 = ||\widetilde{L}_0 - \widetilde{L}_S||^2 \geq 0$. We can rewrite (5.29) as

$$\mathbb{E}_\mu \log \frac{\mu_S}{\mu_S^p} \geq \mathbb{E}_{(\mu)_{\mathcal{Y}}} \log \frac{(\mu)_{\mathcal{Y}}}{(\mu_S)_{\mathcal{Y}}} = \mathbb{E}_{\sum_x \mu} \log \frac{\sum_x \mu_S}{\sum_x \mu_S^p},$$

which is simply the log-sum inequality (cf. [16]).

Therefore, the definition of one-sided sets through (5.21) is appropriate, and since

$$D(\mu_0||\mu_0^p) + D(\mu_0^p||\mu_S^p) = D(\mu_0||\mu_S^p)$$

it is equivalent to

$$D(\mu_0||\mu_S^p) \geq D(\mu_0||\mu_S) + D(\mu_S||\mu_S^p), \quad \forall W_0 \in S.$$

Of course, the general proof of the one-sided result can be shortened, by using just a few inequalities. But by doing it step by step with each divergence term, we illustrate how the local and global results are interacting.

5.4.3 The Proofs

Proof of Lemma 5: Let C a convex set, the for any $P_{\mathcal{X}} \in M_1(\mathcal{X})$ the set $D = \{\mu|\mu(a,b) = P_{\mathcal{X}}(a)W(b|a), W \in C\}$ is a convex set as well. For μ such that $\mu(a,b) = P_{\mathcal{X}}(a)W(b|a)$, we have

$$D(\mu||\mu_C^p) = I(P_{\mathcal{X}}, W) + D(\mu_{\mathcal{Y}}||(\mu_C)_{\mathcal{Y}}),$$

hence we obtain, by definition of W_C being the worse channel of $\mathrm{cl}(C)$,

$$\mu_C = \min_{\mu \in \mathrm{cl}(D)} D(\mu||\mu_C^p).$$

Therefore, we can use theorem 4 and for any $\mu_0 \in D$, we have the pythagorean inequality for convex sets

$$D(\mu_0||\mu_C^p) \geq D(\mu_0||\mu_C) + D(\mu_C||\mu_C^p). \tag{5.30}$$

This concludes the first claim of the lemma. For the second claim, there are many examples of non-convex sets which are one-sided, e.g., let $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $S =$

$$W_1 = \begin{pmatrix} 2/3 & 1/3 \\ 1/4 & 3/4 \end{pmatrix} \quad W_1 = \begin{pmatrix} 3/4 & 1/4 \\ 2/9 & 7/9 \end{pmatrix}.$$

We then have that

$$\arg \max_{P_{\mathcal{X}} \in M_1(\mathcal{X})} (I(P_{\mathcal{X}}, W_1) \wedge I(P_{\mathcal{X}}, W_2)) = (1/2, 1/2)$$

and

$$I((1/2, 1/2), W_1) < I((1/2, 1/2), W_2),$$

which means that W_1 is the worse channel, and W_2 satisfies the one sided property:

$$D(\mu_2||\mu_1^p) - D(\mu_2||\mu_1) - D(\mu_1||\mu_1^p) > 0.$$

Moreover, for any $P_{\mathcal{X}} \in M_1(P_{\mathcal{X}})$, we have

$$I(P_{\mathcal{X}}, W_1) \leq I(P_{\mathcal{X}}, W_2)$$

and

$$D(\mu_2||\mu_1^p) - D(\mu_2||\mu_1) - D(\mu_1||\mu_1^p) \geq 0,$$

where last two inequalities are strict unless $(P_{\mathcal{X}}(0), P_{\mathcal{X}}(1))$ is $(1, 0)$ or $(0, 1)$. Therefore, convex sets are not the only sets for which the one-sided property holds for any $P_{\mathcal{X}}$. $\square$

Proof of Proposition 17: this is done in section 5.4.2. For convex sets, the result is proved in [11]. $\square$

Proof of Proposition 18: note that the proof of theorem 8 works for decoding rules having an infinite number of metrics, so we need to show the following

$$\wedge_{W_1 \in S} \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu \log W_1 \geq \vee_{W \in S} \mathbb{E}_{\mu_0} \log W}} D(\mu||\mu_0^p) \geq \wedge_{W \in S} I(P_{\mathcal{X}}, W),$$

and we will see that the left hand side of previous inequality is equal to $I(P_{\mathcal{X}}, W_0)$.

Note that $\vee_{W \in S} \mathbb{E}_{\mu_0} \log W = \mathbb{E}_{\mu_0} \log W_0 = I(P_{\mathcal{X}}, W_0)$ so that the projection cor-

responding to $W_1 = W_0$, resulting from the metric of the true channel $W_0 \in S$, is $I(P_{\mathcal{X}}, W_0)$, as expected. So we need to verify that for any $W_1 \in S$,

$$\inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu \log W_1 \geq \mathbb{E}_{\mu_0} \log W_0}} D(\mu || \mu_0^p) \geq \wedge_{W \in S} I(P_{\mathcal{X}}, W). \tag{5.31}$$

But

$$\mathbb{E}_\mu \log W_1 \geq \mathbb{E}_{\mu_0} \log W_0 \;\Leftrightarrow\; D(\mu||\mu^p) - D(\mu||\mu_1) \geq D(\mu_0||\mu_0^p), \tag{5.32}$$

therefore,

$$\begin{aligned}
&\inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}} \\ \mathbb{E}_\mu \log W_1 \geq \mathbb{E}_{\mu_0} \log W_0}} D(\mu||\mu_0^p) \\
&= \inf_{\substack{\mu:\, \mu_{\mathcal{X}} = P_{\mathcal{X}},\, \mu_{\mathcal{Y}} = P_{\mathcal{Y}}^0 \\ D(\mu||\mu^p) - D(\mu||\mu_1) \geq D(\mu_0||\mu_0^p)}} D(\mu||\mu^p) \\
&\geq D(\mu_0||\mu_0^p) = I(P_{\mathcal{X}}, W_0).
\end{aligned}$$

Note that above inequality simply uses the fact that $D(\mu||\mu_1) \geq 0$. One could get a tighter lower bound by expressing (5.32) as

$$\begin{aligned}
&\mathbb{E}_\mu \log W_1 \geq \mathbb{E}_{\mu_0} \log W_0 \;\Leftrightarrow \\
&D(\mu||\mu^p) - (D(\mu||\mu_1) - D(\mu^p||\mu_1^p)) \geq D(\mu_0||\mu_0^p) + D(\mu_0^p||\mu_1^p),
\end{aligned}$$

and using the log-sum inequality to show that $D(\mu||\mu_1) - D(\mu^p||\mu_1^p) \geq 0$, (8.23) is lower bounded by

$$D(\mu_0||\mu_0^p) + D(\mu_0^p||\mu_1^p).$$

Figure 5-6 illustrates this gap. □

Proof of Proposition 19: we found a counter-example for the very noisy setting in section 5.3.3, therefore the negative statement holds in the general setting. □

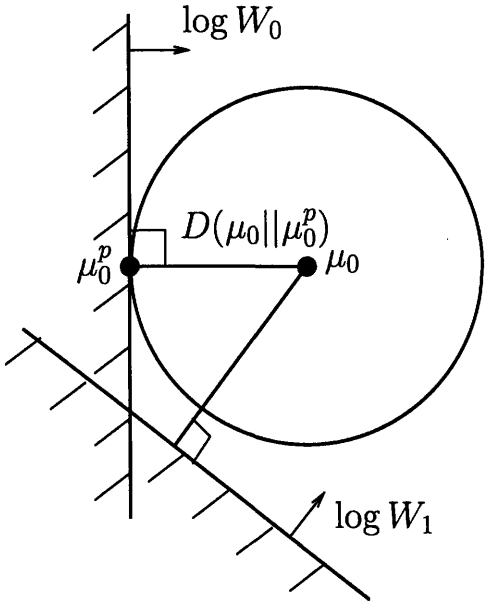

Figure 5-6: Projections of (5.6) in the case $S = \{W_0, W_1\}$

Proof of Proposition 9: using theorem 8, we need to show

$$\inf_{\substack{\mu:\, \mu_{\mathcal{X}}=P_{\mathcal{X}},\, \mu_{\mathcal{Y}}=(\mu_0)_{\mathcal{Y}} \\ \vee_{k=1}^K \mathbb{E}_\mu \log \frac{W_k}{(\mu_k)_{\mathcal{Y}}} \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} \log \frac{W_k}{(\mu_k)_{\mathcal{Y}}}}} D(\mu||\mu_0^p) \geq \wedge_{k=1}^K I(P_{\mathcal{X}}, W_k), \tag{5.33}$$

We can assume w.l.o.g. that $W_0 \in C_1$. Then, for any μ satisfying

$$\mu_{\mathcal{X}} = P_{\mathcal{X}}, \;\; \mu_{\mathcal{Y}} = (\mu_0)_{\mathcal{Y}}, \tag{5.34}$$

$$\vee_{k=1}^K \mathbb{E}_\mu \log \frac{W_k}{(\mu_k)_{\mathcal{Y}}} \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} \log \frac{W_k}{(\mu_k)_{\mathcal{Y}}}, \tag{5.35}$$

we get

$$\begin{aligned} D(\mu||\mu_0^p) &\overset{(A)}{=} D(\mu||\mu^p) \\ &\overset{(B)}{\geq} \vee_{k=1}^K \mathbb{E}_\mu \log \frac{W_k}{(\mu_k)_\mathcal{Y}} \\ &\overset{(C)}{\geq} \vee_{k=1}^K \mathbb{E}_{\mu_0} \log \frac{W_k}{(\mu_k)_\mathcal{Y}} \\ &\geq \mathbb{E}_{\mu_0} \log \frac{W_1}{(\mu_1)_\mathcal{Y}} \\ &\overset{(D)}{\geq} \mathbb{E}_{\mu_1} \log \frac{W_1}{(\mu_1)_\mathcal{Y}} \\ &= I(P_\mathcal{X}, W_1) \\ &\geq \wedge_{k=1}^K I(P_\mathcal{X}, W_k), \end{aligned}$$

where (A) uses (5.34), (B) uses the log-sum inequality:

$$\begin{aligned} \mathbb{E}_\mu \log \frac{W_k}{(\mu_k)_\mathcal{Y}} &= D(\mu||\mu^p) + \mathbb{E}_\mu \log \frac{W_k}{(\mu_k)_\mathcal{Y}} - D(\mu||\mu^p) \\ &= D(\mu||\mu^p) - \underbrace{(D(\mu||\mu_k) - D(\mu^p||\mu_k^p))}_{\geq 0}, \end{aligned}$$

(C) is simply (5.35) and (D) is the one sided property:

$$\begin{aligned} \mathbb{E}_{\mu_0} \log \frac{W_1}{(\mu_1)_\mathcal{Y}} - \mathbb{E}_{\mu_1} \log \frac{W_1}{(\mu_1)_\mathcal{Y}} &= D(\mu_0||\mu_0^p) - (D(\mu_0||\mu_1) - D(\mu_0^p||\mu_1^p)) \\ &\quad -D(\mu_1||\mu_1^p) \\ &= D(\mu_0||\mu_1^p) - D(\mu_0||\mu_1) - D(\mu_1||\mu_1^p) \\ &\geq 0. \end{aligned}$$

Hence

$$\begin{aligned} \inf_{\substack{\mu:\, \mu_\mathcal{X}=P_\mathcal{X},\, \mu_\mathcal{Y}=(\mu_0)_\mathcal{Y} \\ \vee_{k=1}^K \mathbb{E}_\mu \log \frac{W_k}{(\mu_k)_\mathcal{Y}} \geq \vee_{k=1}^K \mathbb{E}_{\mu_0} \log \frac{W_k}{(\mu_k)_\mathcal{Y}}}} & D(\mu||\mu_0^p) \\ \geq I(P_\mathcal{X}, W_1) \geq \wedge_{k=1}^K I(P_\mathcal{X}, W_k), & \end{aligned}$$

which proves (5.33) and concludes the proof. □

5.4.4 Generalized Worst A Posteriori (GWAP) Algorithm

Definition 26. For an arbitrary non empty set S, we define the worse channel set by

$$\{W_S\} = \arg \min_{W \in \mathrm{cl}(S)} I(P_{\mathcal{X}}, W). \tag{5.36}$$

Definition 27. *GWAP Algorithm*
Let S an arbitrary non empty set. If $|\{W_S\}|$ is finite, we define $C_0(S) = S$, and

$$C_1(S) = \{W \in S | D(\mu_0 || \mu_S^p) > D(\mu_0 || \mu_S) + D(\mu_S || \mu_S^p), \forall W_S \in \{W_S\}\},$$

otherwise $C_0(S) = *$ and we stop. Recursively for $k \geq 1$, if $C_k(S)$ is empty we define $C_k(S) = \emptyset$ and we stop. Otherwise, if $|\{W_{C_k(S)}\}|$ is finite, we define

$$C_{k+1}(S) = C_1(C_k(S)),$$

elsewhere, $C_{k+1}(S) = *$ and we stop.

If the algorithm returns $C_N(S) = \emptyset$ for some $N \geq 1$, define the metrics $\mathrm{WAP}(S) = \{d_k\}_{k=1}^K$ to be the MAP metrics of all worst channels encountered by the algorithm ($K = \sum_{l=0}^{N-1} |\{W_{C_l(S)}\}|$). The linear decoder induced by WAP(S) is denoted by GWAP(S), for General Worst A Posteriori decoding.

If S is a finite number of one-sided components, the sequence C_k never reaches $*$ and runs up to a finite number of iterations N, for which C_N is empty, as illustrated in figure 5-7. Hence, from theorem 9, GWAP(S) achieves compound capacity on S. If S is not a finite number of one-sided components, the algorithm may stop and return $*$, or may never stop. However, $M_1(\mathcal{X} \times \mathcal{Y})$ is a compact subset of $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$, so we claim that for any set S, there exists $\varepsilon > 0$ and a set S_ε which is one-sided and such that decoding with the metrics WAP(S_ε) can achieve all rates R with $R \leq C - \varepsilon$. This means that we can find a weak linear universal decoding rule for any compound

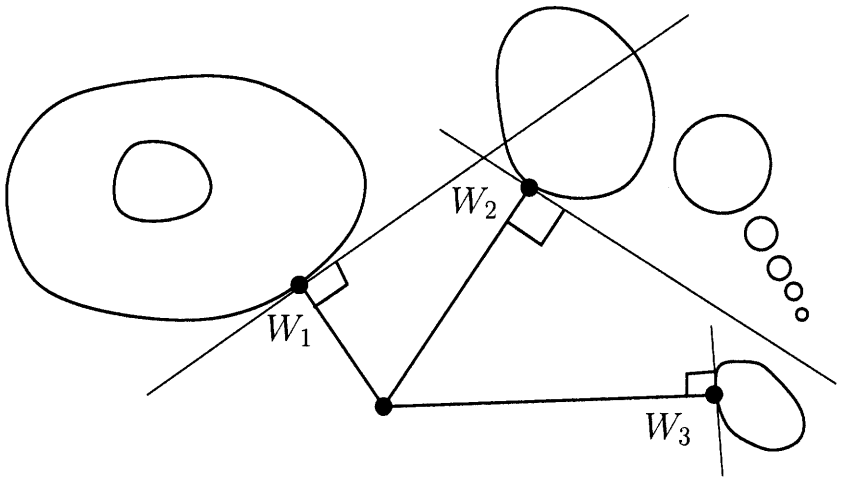

Figure 5-7: GWAP algorithm

set S, which is not the same as linear universal, since $\mathrm{WAP}(S_\varepsilon) \overset{\varepsilon \to 0}{\to} \infty$. The GWAP algorithm suggests how to construct $\mathrm{WAP}(S_\varepsilon)$: If for some $k \geq 0$, $|\{W_{C_k(S)}\}|$ is not finite, we pick $A_k = A_k(\varepsilon)$ and $W_{k,1}, \ldots, W_{k,A}$ within

$$\{W \in S : I(P_\mathcal{X}, W) = I(P_\mathcal{X}, W_{C_k(S)})\},$$

where $W_{C_k(S)}$ is an element of $\{W_{C_k(S)}\}$, in order to have

$$\vee_{l=1}^{A} I_{\mathrm{MIS}}(P_\mathcal{X}, W, W_{k,l}) \geq I(P_\mathcal{X}, W_{C_0(S)}) - \varepsilon.$$

We define $W_{k,l}$ to be the worst channels, just like when $|\{W_{C_k(S)}\}|$ is finite. We then move to $C_{k+1}(S)$. However, the algorithm may never stop, but because $M_1(\mathcal{X} \times \mathcal{Y})$ is a compact subset of $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$, we claim that there exists N such that

$$\vee_{k=1}^{N} \vee_{l=1}^{A_k} I_{\mathrm{MIS}}(P_\mathcal{X}, W, W_{k,l}) \geq I(P_\mathcal{X}, W_{C_0(S)}) - \varepsilon.$$

Polytope Decoder

The GWAP algorithm is universal for most sets (weakly universal in general) and, as the definition of universality allows it, it depends on the considered compound set. Since encoders and decoders must anyway cooperate before the communication takes place to agree on a code, and since the encoder must know the compound set S to identify which rates can be employed, a decoding rule which is universal without depending on the compound set may not have a real advantage. However, let us assume that the highest rate C at which reliable communication can be established is given to the transmitter and receiver without specifying the compound set S, we then want to construct a single decoder which will be used on different compound channel. The question is then which directions do we choose to build our decoding metrics. Without any knowledge of S, we would like to take the directions in a uniform way, shaping a regular polytope for which the sphere of radius C is an insphere (the direction are perpendicular to the polytopes faces). Different orientations of the regular polytope will achieve different rates on different compound sets. If the set is a finite union of one-sided components, we know that there exists an orientation of the directions for which the set S lies outside the polytope. Otherwise, infinitely many directions are required. However, if we are giving up on an ε-portion of the capacity, i.e., if we want to achieve any rate $R \leq C - \varepsilon$, there always exists a sets of directions shaping a polytope around the sphere of radius $C - \varepsilon$ out of which the set S is contained; indeed, any regular polytope having the sphere of radius $C - \varepsilon$ has a circumsphere that will give proper directions. One can then study the relationships between the dimension of the channel, i.e,. $\mathcal{X}$, $\mathcal{Y}$, the capacity value C, the tolerance ε and the number of directions required to achieve $C - \varepsilon$.

5.5 Discussion

The MMI decoding rule is remarkable for its theoretical properties. Observe that

$$I(P_m) = \sup_{W \in \text{ DMC}} \mathbb{E}_{P_m} \log \frac{W}{(P_{\mathcal{X}} \circ W)_{\mathcal{Y}}} \tag{5.37}$$

which means that the MMI is actually the GWAP decoders taking into account all DMC's:

$$\text{MMI} = \text{GWAP(DMC)}.$$

When communicating over a compound set S, MMI is achieving capacity, but our main result tells us that we do not need to take all DMC metrics in the supremum of (5.37) to achieve the capacity. By extracting the one-sided components of S, it is sufficient to take the worse channel of these components in the supremum of (5.37) and still achieve compound capacity. And when S has a finite number of one-sided components, the GWAP decoding rules is generalized linear. Hence, we are basically picturing the *one-sided components as equivalent classes* in the space of decoding metrics and the *MAP metrics as canonical metrics* under the max-operator.

Finally, how do we explain the fact that generalized ML metrics are not performing as well as generalized MAP metrics, after all, we also have

$$\arg\max_m I(P_m) = \arg\max_m \sup_{W \in \text{ DMC}} \mathbb{E}_{P_m} \log W,$$

which means that the MMI decoding rule is equivalent to the GLRT decoding rule with all channels, in the sense that decoding regions are the same for both decoding rules. However, this time we have

$$I(P_m) \neq \sup_{W \in \text{ DMC}} \mathbb{E}_{P_m} \log W,$$

and there is here a subtlety: a received y has different likeliness under different channels ($(\mu_k)_Y$ depends on k). Hence, to determine which input has been sent, without knowing if the channel that has transformed the input was W_1 or W_2, we need to measure how likely it is that x has generated y in agreement with how likely y is for the different channels. And the MAP metrics, as opposed to the ML metrics, takes this into account.

It is then surprising that we found in the literature many references discussing the use of GLRT tests, i.e., which is the usual name for generalized ML decoding,

whereas we could not find references discussing generalized a posteriori decoding.

Chapter 6

Gaussian Channels and Local Input Geometry

In last chapter, we introduced a local analysis of discrete memoryless channels, by fixing the input distribution and letting the channels become always noisier. This brought the joint and product distributions (i.e., the joint distribution of the sent input and received output, and the product distribution of any non-sent input and the received output) close to each other, allowing us to work in the local setting. This approach has been useful to construct good decoders in chapter 5. In this chapter, we look at localization of input distributions. By its nature, this localization will be useful for input optimization problems, to understand the structure of optimal input distributions (hence of optimal encoders), or any pre-encoding scheme used in multi-user information theoretic problems dealing with Gaussian noise. Instead of a DMC's, we now consider channels which are memoryless but with continuous alphabets. After analyzing different entropic properties of the operator consisting of convoluting with Gaussian densities, we consider a Gaussian interference channel problem.

6.1 Additive Gaussian Noise

The channels that are considered in this sections are discrete time continuous alphabets channel, where the continuous input and output alphabets are $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. The

additive white Gaussian noise (AWGN) channel is defined as follows. At each channel uses $i = 1, \ldots, n$, an input $x(i) \in \mathbb{R}$ is sent, and an output $Y(i)$ is received, resulting of the addition of mutually independent centered Gaussian random variables:

$$Y(i) = x(i) + Z(i), \quad i = 1, \ldots, n$$

where

$$\{Z(i)\}_{i=1}^{n} \overset{iid}{\sim} \mathcal{N}(0, v).$$

It would not be realistic to model an encoder on a continuous alphabet without having any constraint on the possible input magnitudes. A very common limitation imposed on the inputs is an average power constraint, which requires any n-sequences of input symbols to satisfy

$$\frac{1}{n} \sum_{i=1}^{n} x(i)^2 \leq P,$$

where $P \in \mathbb{R}_+$. For more details regarding the model, cf. [16],[17].

For continuous alphabets with average power constraint, a valid encoder is a mapping $E_n : \ m \in \{1, \ldots, M = \lfloor e^{nR} \rfloor\} \mapsto E_n(m) = x_m \in \mathbb{R}$, that must satisfy

$$\sum_{i=1}^{n} x_m(i)^2 \leq P, \quad \forall m \in \{1, \ldots, M\},$$

and a decoder is a mapping $D_n : \ y \in \mathbb{R} \mapsto m \in \{1, \ldots, M\}$.

In the same way we proved it for DMC's, one can show that the maximum of the mutual information between X and Y, over input distributions satisfying the power constraint, gives the AWGN channel capacity. The mutual information between X and Y is then given by

$$I(X, X + Z) = h(X + Z) - h(Z),$$

where, for a random variable W having a density p, $h(W)$ is given by

$$h(W) = h(p) = -\int_{\text{Supp}(p)} p(y)\log p(y)dy,$$

provided that the integral exists.

We now present several results concerning maximum entropy problems. In the following, maximums or minimums are taken over random variables, since we found this notation to be common in the literature. Hence arg max, respectively arg min, denotes the random variable[1] maximizing, respectively minimizing, the considered functional. However, in the next section, we will prefer to shift to a functional notation, working with densities.

Lemma 6. *Let $Z \sim \mathcal{N}(0, v)$, then*

$$\arg \max_{X:\,\text{Var}(X)\leq s} h(X+Z) \sim \mathcal{N}(0,s), \quad \forall s, v > 0.$$

Hence, previous lemma implies the following result.

Theorem 10. *The highest achievable rate on a AWGN channel with the noise having mean 0 and variance v is given by*

$$C = \sup_{X:\,\mathbb{E}X^2\leq P} I(X,Y) \tag{6.1}$$

$$= \log(1+\frac{P}{v}) \tag{6.2}$$

and the optimal input is a centered Gaussian with variance s.

The following result is the Entropy Power Inequality (first proved by Stam).

Lemma 7. *For any independent random variable X and Z,*

$$2^{2h(X+Z)} \geq 2^{2h(X)} + 2^{2h(Z)}.$$

[1] in general this would be a set of random variables, but in all the considered cases, the set contained only one element

When restricted to $Z \sim \mathcal{N}(0, v)$, the Entropy Power Inequality reduces to the following.

Lemma 8.

$$\arg \min_{X:\, h(X)=\frac{1}{2}\log 2\pi e s} h(X+Z) \sim \mathcal{N}(0,s), \quad \forall v, s > 0.$$

Lemmas 6 and 9 together imply

Lemma 9. *Let $X \sim \mathcal{N}(0,s)$, then*

$$\arg \min_{Z:\, \mathrm{Var}(Z)=v} h(X+Z) - h(Z) \sim \mathcal{N}(0,v), \quad \forall v, s > 0.$$

Which implies that for $X^g \sim \mathcal{N}(0,s)$,

$$\arg \min_{Z:\, \mathrm{Var}(Z)\leq v} I(X^g, X^g + Z) \sim \mathcal{N}(0,v), \tag{6.3}$$

i.e., the worse possible additive noise, when considering Gaussian inputs, is Gaussian as well. For additional treatments of these problems, cf. [7].

6.1.1 Motivation

We presented different variational properties of the entropy of a random variable under the addition of Gaussian noise. Those properties are important not only for the study of AWGN, but for many other kinds of channels having addition of Gaussian noise, such as for example, Gaussian interference channels or Gaussian broadcast channels. In these examples, more complex optimizations of entropic functions (with additive Gaussian noise) may appear, e.g. in the two users symmetric interference channel, treating interference as noise leads to a sum-rate lower bounded by

$$\sup_{X_1, X_2:\, \mathrm{Var}(X_1), Var(X_2)\leq P} I(X_1, X_1 + aX_2 + Z_1) + I(X_2, X_2 + aX_1 + Z_2),$$

where $a, P \geq 0$ and $Z_1, Z_2 \sim \mathcal{N}(0,v)$. In this situation, the basic principles presented earlier cannot be used directly to identify the optimal input distribution. Let $X_1^g \sim$

$\mathcal{N}(0,s)$, we know that

$$I(X_1^g, X_1^g + aX_2 + Z_1) = h(X_1^g + aX_2 + Z_1) - h(aX_2 + Z_1),$$

hence, choosing X_2 to be Gaussian would maximize the first term above, however it would penalize the second term. And of course, the value of a comes into the picture. If $a = 0$ the Gaussian distribution is optimal, and if a is very small, we expect this to be true as well. In the interference channel, these kinds of conflicting situations are at the heart of the problem and we would like to acquire a better understanding of them. Finding a way to quantify the interference, as noise, or as information, is a crucial point in the understanding of this channel. Analyzing the structure of the input distribution of the sum-rate under different assumptions is just a first cut analysis. It also serves as a good illustrative problem for more general problem encountered in the interference channel. The following questions are relevant for these considerations. Let $X^g \sim \mathcal{N}(0,s)$. If we add an independent Gaussian noise Z to X^g we get a Gaussian random variable, which has maximum entropy over all input distributions having variance s. Let us say that we can now perturb a little bit X_1^g, and let us denote by X_δ^g the perturbation. Which "direction" preserves a variance of s and makes $X_\delta^g + Z$ looks less or more Gaussian, i.e., minimizes or maximizes the entropy of $X_\delta^g + Z$? And if we now have to move from a fixed divergence distance, which direction makes $X_\delta^g + Z$ have maximal or minimal entropy?

6.2 Localization and Hermite Transformation

We start by adopting a different notation for the results of previous section.

Definition 28.

$$g_{a,s}(x) = \frac{1}{\sqrt{2\pi s}} e^{-\frac{(x-a)^2}{2s}}, \quad a, x \in \mathbb{R},\ s \in \mathbb{R}_+$$

and

$$g_s(x) = g_{0,s}(x).$$

We denote by $M_d(\mathbb{R})$ the set of densities on $\mathbb{R}$, i.e., positive integrable functions integrating to 1 on $\mathbb{R}$.

Definition 29. The Divergence between $p, q \in M_d(\mathbb{R})$ is defined by

$$D(p||q) = \int_{\mathbb{R}} p(x) \log.\frac{p(x)}{q(x)} dx,$$

if $\text{Supp}(p) \subseteq \text{Supp}(q)$, and infinity otherwise.

From now on, we adopt the convention $0 \log 0 = 1$ and we assume that in the following, $p \in M_d(\mathbb{R})$ is such that the quantities of interest are existing (which is true for the choices of p we will make later). We will use the notation

$$\text{Var}(p) = \int_{\mathbb{R}} p(x)(x - \int_{\mathbb{R}} yp(y)dy)^2 dx.$$

Note: for p with $\text{Var}(p) = s$,

$$h(p) = h(g_s) - D(p||g_s).$$

Hence lemma 6 is equivalent to

Theorem 11.

$$\arg \min_{p:\, \text{Var}(p)=s} D(p \star g_v || g_s \star g_v) \quad = \quad g_{0,s}, \quad \forall s, v > 0. \tag{6.4}$$

Definition 30. Let

$$\langle K, L \rangle_{g_s} = \int_{\mathbb{R}} K(x)L(x)g_s(x)dx$$

and let $L_2(g_s; \mathbb{R})$ be the set of real measurable functions L for which $\|L\|_{g_s}$ is finite (i.e., $L^2 g_s$ is Lebesgue integrable). Let $M_0(g_s) = \{L : \mathbb{R} \to \mathbb{R} | \int_{\mathbb{R}} L(x)g_s(x)dx = 0\}$ and

$$\hat{M}_0(g_s) = \{L : \mathbb{R} \to \mathbb{R} | \int_{\mathbb{R}} L(x)g_s(x)dx = 0, \inf_{x \in \mathbb{R}} L(x) > -\infty\},$$

and let $L_2(\hat{M}_0(g_s), g_s) = \hat{M}_0(g_s) \cap L_2(g_s; \mathbb{R})$. Finally, let $\mathbb{P}(g_s)$ be the set of all real

polynomials in $L_2(\hat{M}_0(g_s), g_s)$ and let

$$D(g_s) = L_2(\hat{M}_0(g_s), g_s) \cap \{L : \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2} \int_{D_\varepsilon^c} (L) g_s(x) L(x) dx = 0, \left\| L^{3/2} \right\| < \infty\}$$

where $D_\varepsilon(L) = \{x \in \mathbb{R} | -1/\varepsilon < L(x) \leq 1/\varepsilon\}$.

The reason for which we introduce these different sets is the following. For $\varepsilon \ll 1$, we have that $g_s(1 + \varepsilon L) \in M_d(\mathbb{R})$ as long as $L \in \hat{M}_0(g_s)$. Hence $\hat{M}_0(g_s)$ is our sets of possible directions. In the discrete setting, we did not have to worry about anything else to carry out our approximations, but here it is of course different. If we aim to make use of the approximation $D(g_s(1 + \varepsilon L)||g_s) = \frac{1}{2}\varepsilon^2 ||L||_{g_s}^2 + o(\varepsilon^2)$, we definitely need $L \in L_2(g_s; \mathbb{R})$. If L is also polynomial, we will see that no further assumptions are needed. More generally, $L \in D(g_s)$ allows the approximation. Moreover, we will see that for any $v \geq 0$, $\hat{M}_0(g_s)$ and $L_2(g_s; \mathbb{R})$ are closed under the mapping $L \rightsquigarrow g_s L \star g_v$, and so is $L_2(\hat{M}_0(g_s), g_s)$. The first constraint added in $D(g_s)$ is also closed under this mapping, however, we could not proved that the second constraint, i.e. $\left\| L^{3/2} \right\| < \infty$, is closed under Gaussian convolution. It may be that this constraint is not necessary, or is closed under this mapping. We actually did not try to check this, since for the application we have in mind, this is not relevant.

Example: $\sin(x)$, $x^2 - 1$ are example of valid directions in $D(g_s)$.

The following result is an analogue of the result in (2.6).

Lemma 10. *For $L \in D(g_s)$,*

$$D(g_s(1 + \varepsilon L)||g_s) = \frac{1}{2}\varepsilon^2 ||L||_{g_s}^2 + o(\varepsilon^2). \tag{6.5}$$

Proof.

$$D(g_s(1 + \varepsilon L)||g_s) = \int_{\mathbb{R}} g_s(x)(1 + \varepsilon L(x)) \log(1 + \varepsilon L(x)).$$

However, if $|L|$ is not bounded, not matter how small ε is, $\log(1 + \varepsilon L(x))$ may not be well approximated by its taylor expansion for all $x \in \mathbb{R}$. More precisely, it is only

when $x \in D_\varepsilon = \{x \in \mathbb{R} | -1/\varepsilon < L(x) \leq 1/\varepsilon\}$ that $\log(1+\varepsilon L(x)) = \varepsilon L(x) - \frac{\varepsilon^2 L(x)^2}{2} + o(\varepsilon^2 L(x)^2)$. Hence,

$$\begin{aligned} D(g_s(1+\varepsilon L)||g_s) &= \int_{D_\varepsilon} g_s(x)(1+\varepsilon L(x))\log(1+\varepsilon L(x))dx \\ &\quad + \int_{D_\varepsilon^c} g_s(x)(1+\varepsilon L(x))\log(1+\varepsilon L(x))dx \\ &= \int_{D_\varepsilon} g_s(x)[1+\varepsilon L(x))(\varepsilon L(x) - \varepsilon^2 L(x)^2/2 + o(\varepsilon^2 L(x)^2)]dx \\ &\quad + \int_{D_\varepsilon^c} g_s(x)(1+\varepsilon L(x))\log(1+\varepsilon L(x))dx \\ &= \int_{D_\varepsilon} g_s(x)[\varepsilon L(x) + \frac{1}{2}\varepsilon^2 L(x)^2 + o(\varepsilon^2 L(x)^2)]dx \\ &\quad + \int_{D_\varepsilon^c} g_s(x)(1+\varepsilon L(x))\log(1+\varepsilon L(x))dx \end{aligned}$$

By the Lebesgue dominated convergence theorem, since $||L||_{g_s} \wedge ||L^{3/2}||_{g_s} < \infty$, we have

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2}\frac{1}{2}\int_{\mathbb{R}} \mathbb{1}_{D_s}(x) g_s(x)\varepsilon^2 L(x)^2 dx = \frac{1}{2}||L||_{g_s}^2,$$

and

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2}\int_{\mathbb{R}} \mathbb{1}_{D_s}(x) g_s(x) o(\varepsilon^2 L(x)^2)dx = 0.$$

So we have to show

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon^2}\left[\int_{D_\varepsilon} g_s(x)\varepsilon L(x)dx + \int_{D_\varepsilon^c} g_s(x)(1+\varepsilon L(x))\log(1+\varepsilon L(x))dx\right] = 0$$

Since $L \in D(g_s)$, we have

$$\int_{D_\varepsilon} g_s(x)L(x)dx = -\int_{D_\varepsilon^c} g_s(x)L(x)dx,$$

which is a $o(\varepsilon^2)$ by assumption on L. Finally,

$$\int_{D_\varepsilon^c} g_s(x)(1+\varepsilon L(x))\log(1+\varepsilon L(x))dx \leq \int_{D_\varepsilon^c} g_s(x)(1+\varepsilon L(x))\varepsilon L(x)dx,$$

and using the assumptions on L, we have $\lim_{\varepsilon \searrow 0}\int_{D_\varepsilon^c} g_s(x)L^2(x)dx = 0$ and the last

term is a $o(\varepsilon^2)$ as well. $\square$

Lemma 11. *For $L \in D(g_s)$ such that $|| \left(\frac{g_s L \star g_v}{g_{s+v}}\right)^{3/2} ||^2_{g_{s+v}} < \infty$, we have*

$$D(g_s(1+\varepsilon L) \star g_v || g_s \star g_v) = \frac{1}{2}\varepsilon^2 || \frac{g_s L \star g_v}{g_{s+v}} ||^2_{g_{s+v}} + o(\varepsilon^2).$$

Proof. This lemma is a consequence of the first lemma if we can show that $\frac{g_s L \star g_v}{g_{s+v}}$ is in $D(g_{s+v})$ given that $L \in D(g_s)$. But

$$\begin{aligned} \int_{\mathbb{R}} \frac{(g_s L \star g_v)(x)}{g_{s+v}(x)} g_{s+v}(x)dx &= \int_{\mathbb{R}} (g_s L \star g_v)(x)dx \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} g_s(t)L(t)g_v(x-t)dtdx \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} g_v(x-t)dx\, g_s(t)L(t)dt \qquad (6.6) \\ &= \int_{\mathbb{R}} g_s(t)L(t)dt = 0 \end{aligned}$$

where equality (6.6) uses Fubini's theorem, since

$$|g_s(t)L(t)g_v(x-t)| \leq \frac{1}{2}(L(t)^2 g_s(t) + g_v(x-t)^2 g_s(t)),$$

and the right hand side of above inequality has a finite integral. This shows that $\frac{g_s L \star g_v}{g_{s+v}} \in M_0(g_{s+v})$ and it also belongs to $\hat{M}_0(g_{s+v})$ since

$$\frac{(g_s L \star g_v)(x)}{g_{s+v}(x)} \geq \inf_{y \in \mathbb{R}} L(y) \frac{(g_s \star g_v)(x)}{g_{s+v}(x)} = \inf_{y \in \mathbb{R}} L(y) > -\infty.$$

In order to prove that

$$|| \frac{g_s L \star g_v}{g_{s+v}} ||^2_{g_{s+v}} < \infty$$

given that

$$||L||^2_{g_s} < \infty,$$

we use the following result (for which the proof is provided in lemma 20 below): there

exists $\alpha_k \in \mathbb{R}$, $k \geq 0$, such that

$$L = \sum_k \alpha_k \bar{H}_k^{[s]}$$

where the $\bar{H}_k^{[s]}$ are orthonormal polynomials with respect to g_s. Moreover,

$$\frac{g_s \bar{H}_k^{[s]} \star g_v}{g_{s+v}} = \left(\frac{s}{s+v}\right)^{k/2} \bar{H}_k^{[s+v]}.$$

Therefore, since $\alpha_k^2 k! \frac{s^k}{(s+v)^k} \leq \alpha_k^2 k!$, for all $k \geq 0$, we have that

$$||\frac{g_s L \star g_v}{g_{s+v}}||_{g_{s+v}}^2 \leq ||L||_{g_s}^2 < \infty,$$

where the inequality is strict unless L is a constant. The condition on the third power is given as an assumption, so we do not have to check it. Finally, we need to show that

$$\int_{\{x|-1/\varepsilon < L(x) \leq 1/\varepsilon\}} (g_s L \star g_v)(x) dx = o(\varepsilon^2),$$

which holds in a similar fashion as for previous checks, using our assumptions on L, Fubini and the monotone convergence theorems. □

We know that for $p \in M_d(\mathbb{R})$ with $\mathrm{Var}(p) = s$, we have

$$h(p) = h(g_s) - D(p||g_s).$$

Therefore, statement (6.4) in theorem 11, can locally be expressed as

Corollary 3.

$$\arg \min_{\substack{L \in \bar{M}_0(\mathbb{R}) \text{ s.t.} \\ \langle 1, L \rangle_{g_s} = \langle x, L \rangle_{g_s} = \langle x^2, L \rangle_{g_s} = 0}} \left\| \frac{g_s L \star g_v}{g_{s+v}} \right\|_{g_{s+v}}^2 = 0, \tag{6.7}$$

where 0 means the constant function 0.

This is a corollary of (6.4), since it just means that the Gaussian distribution is a

local minimum of the divergence function that we are optimizing.

Concerning the entropy power inequality, we can separate the local problem in two stages and prove it locally with previous expansions.

If we express above quantities in terms of the Lebesgue measure λ instead of the Gaussian one, we have

$$\begin{aligned} D(g_s(1+\varepsilon L)||g_s) &= \|\sqrt{g_s}L\|_\lambda & (6.8)\\ \langle K, L\rangle_{g_s} &= \langle\sqrt{g_s}K, \sqrt{g_s}L\rangle_\lambda & (6.9)\\ \left\|\frac{g_s L \star g_v}{g_{s+v}}\right\|^2_{g_{s+v}} &= \left\|\frac{g_s L \star g_v}{\sqrt{g_{s+v}}}\right\|_\lambda & (6.10) \end{aligned}$$

Proposition 20. *Let*

$$T:\ L \in L_2(g_s;\mathbb{R}) \mapsto \frac{\sqrt{g_s}L \star g_v}{\sqrt{g_{s+v}}} \in L_2(g_s;\mathbb{R}),$$

then

$$T^t T L = \lambda L, \quad L \neq 0$$

holds for each pairs

$$(L,\lambda) \in \{(\bar{H}_k^{[s]}, (\frac{s}{s+v})^k)\}_{k\geq 0},$$

where

$$H_k(x) = (-1)^k e^{x^2/2}\frac{d^k}{dx^k}e^{x^2/2}, \quad k \geq 0, x \in \mathbb{R}$$

and

$$\bar{H}_k^{[s]}(x) = \frac{1}{\sqrt{k!}}H_k(x/\sqrt{s}).$$

These polynomials are the Hermite polynomials (for a Gaussian distribution having variance s).

This proposition is proved after the following two results.

Theorem 12. *For any $s > 0$, $\{\bar{H}_k^{[s]}\}_{k\geq 0}$ is an orthonormal basis of $L_2(g_s;\mathbb{R})$.*

A proof of this theorem can be found in [31]. We also refer to [15] for other properties of Hermite polynomials.

Since $\bar{H}_0^{[s]} = 1$, previous theorem implies in particular that $\bar{H}_k^{[s]} \in M_0(g_s)$, for any $k > 0$. Moreover, it is easy to check that $\bar{H}_1$, respectively $\bar{H}_2$ perturb a Gaussian distribution into another Gaussian distribution, with a different mean, respectively variance. It is for $k \geq 3$ that $\bar{H}_k$ perturbations are no longer Gaussian distributions.

Proposition 21.

$$\frac{g_s \bar{H}_k^{[s]} \star g_v}{g_{s+v}} = (\frac{s}{s+v})^{k/2} \bar{H}_k^{[s+v]}, \quad \forall s, v > 0, \, k \geq 0.$$

Important Fact: Propositions 21 is an important feature in addition to proposition 20. It tells us that the eigenfunctions of T acting on $L_2(g_s; \mathbb{R})$ are naturally mapped into the eigenfunctions of T acting on $L_2(g_{s+v}; \mathbb{R})$, since the kth eigenfunction is mapped to the kth eigenfunction contracted by the kth singular value. This is illustrated in figure 6-1.

Proof. We need to show

$$g_s \bar{H}_k^{[s]} \star g_v = (\frac{s}{s+v})^{k/2} g_{s+v} \bar{H}_k^{[s+v]}. \tag{6.11}$$

We prove this by induction. For $k = 0$ the statement is trivial since $\bar{H}_0^{[s]} = 1$. Let us assume that this is true for k; by taking the derivative on the left hand side, we have

$$\frac{\partial}{\partial x} \left[g_s \bar{H}_k^{[s]} \star g_v \right] = \left[\frac{\partial}{\partial x} (g_s \bar{H}_k^{[s]}) \right] \star g_v, \tag{6.12}$$

however, by definition we have $\bar{H}_k^{[s]} = \frac{1}{\sqrt{k!}} H_k(x/s)$ where

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{x^2/2},$$

hence we get

$$g_s \bar{H}_k^{[s]}(x) = \frac{1}{\sqrt{k!}} \frac{1}{2\pi\sqrt{s}} (-1)^k \frac{d^k}{dy^k} e^{y^2/2} \Big|_{y=\frac{x}{\sqrt{s}}}$$

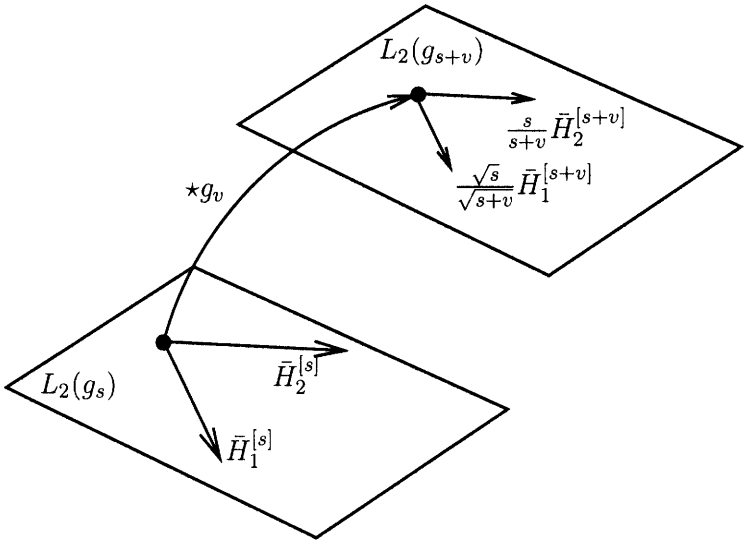

Figure 6-1: Hermite eigenfunctions correspondence

and

$$\frac{\partial}{\partial x} g_s \bar{H}_k^{[s]} = -\sqrt{\frac{k+1}{s}} g_s \bar{H}_{k+1}^{[s]}. \tag{6.13}$$

Therefore, the derivative of the left end side of 6.11 is

$$-\sqrt{\frac{k+1}{s}} g_s \bar{H}_{k+1}^{[s]} \star g_v$$

and using 6.13 again, the derivative of the right hand side of 6.11 is

$$-\sqrt{\frac{k+1}{s+v}} g_{s+v} \bar{H}_{k+1}^{[s+v]},$$

putting the equality back together, we proved the induction. □

This result is illustrated in figure 6-1.

Proof of proposition 20: using proposition 21, the claim of this proposition is

equivalent to:

$$g_v \star \bar{H}_k^{[s+v]} = \left(\frac{s}{s+v}\right)^k \bar{H}_k^{[s]}. \tag{6.14}$$

We proceed by induction, and again for $k = 0$ the result is trivial. Let us assume that the equality holds for k. With the arguments of previous proof, we can check that

$$\frac{\partial}{\partial x}\bar{H}_{k+1}^{[s]} = \frac{\sqrt{k+1}}{s}\bar{H}_k^{[s]},$$

using this identity in 6.14, we prove that the derivative of our claims holds for $k+1$ (and any x),

$$g_v \star \frac{\partial}{\partial x}\bar{H}_{k+1}^{[s+v]} = \left(\frac{s}{s+v}\right)^{k+1} \frac{\partial}{\partial x}\bar{H}_{k+1}^{[s]}.$$

and since, for any k, no constant can be added in 6.14, we can remove the derivative and the induction is proved. □

We will now perform a local analysis of the input distributions for the Gaussian interference channel; instead of using the VN transformation that brings the channel around very noisy distributions (defined in the discrete setting), we will bring the input distribution around Gaussian distributions. Using previous results regarding the Hermite basis, we consider the Hermite polynomials for the directions to perturb the Gaussian distribution. This process is referred to the Hermite transformation.

6.2.1 Optimal Input for Interference Sum-rate

We consider a memoryless additive white Gaussian noise symmetric interference channel, which is described by

$$Y_1(i) = x_1(i) + ax_2(i) + Z_1(i) \tag{6.15}$$

$$Y_2(i) = x_2(i) + ax_1(i) + Z_2(i) \tag{6.16}$$

where $i = 1, 2, \ldots$ denotes the channel uses. The inputs alphabet is the real line, i.e., $x_1(i), x_2(i) \in \mathbb{R}$ for any $i \geq 1$ and the inputs are subject to an average power constraint:

$$\frac{1}{n}\sum_{i=1}^{n} x_k(i)^2 \leq P_k, \quad k = 1, 2.$$

The random process $Z_k(\cdot)$ is (homogeneous) memoryless for $k = 1, 2$, with marginal distribution $Z_1(1), Z_2(1) \sim \mathcal{N}(0, 1)$ and $Z_1(\cdot)$ is independent of $Z_2(\cdot)$. The factor $a \in \mathbb{R}_+$ is called the interference coefficient.

The capacity region of the inference channel is an open problem. It has been solved in several particular cases. We know that for small values of the interference, treating the interference as noise and using the interference channel as two independent AWGN channels is optimal for the sum-rate (cf. [3]), and for an AWGN, Gaussian random code books are optimal, as we saw in previous section. We also know that for strong interference, i.e. $a \geq 1$, the optimal coding scheme requires the decoding of both users messages at each receivers, cf. [27], [28]. It is then tempting to believe that there are two transitions happening when a varies. For more details regarding the interference channel cf.[6],[18],[13] and references therein. The first regime should be when the other message is ignored and treated as noise, then there should be another regime where we want to partially decode the other message, and finally a regime where we completely decode the other message.

In this section, we would like to examine for which values of the interference are the independent Gaussian code books optimal or suboptimal for the sum-rate, i.e., we are interested in finding a threshold at which this switch happens. This will tell us where is the capacity expressions found in [3] no longer valid and it points out an interference value where "a transition" happens. This serves as a good illustrative problem to introduce the technique developed in previous section, but the result presented in this section for the single-letter case can generalize to the multi-letter case. Therefore, a transition where interference should not be treated has indeed been found with this technique. Other problems concerning the interference channel are currently being investigated with the Hermite transformation.

The sum-rate optimization for the single-letter case is given by

$$\max_{\substack{X_1 \perp\!\!\!\perp X_2 \text{ s.t.} \\ \mathrm{Var}X_1 = \mathrm{Var}X_2 = 1}} I(X_1; X_1 + aX_2 + W_1) + I(X_2; X_2 + aX_1 + W_2). \tag{6.17}$$

Proposition 22. *Let* $a_1 \approx 0.68$ *be the only real root of* $a^6(2+a^2)^3 - (1-a^3)^2(1+a^2)^3 = 0$ *and let* $a_2 = (\frac{\sqrt{5}-1}{2})^{\frac{1}{2}} \approx 0.79$. *If* $a \leq a_1$, $g_1 \times g_1$ *is a local maximizer of* (6.17), *otherwise* $g_1 \times g_1$ *is not a maximizer. If* $a_1 < a < a_2$, $g_1 \times g_1$ *is at a saddle point and if* $a \geq a_2$, $g_1 \times g_1$ *is a local minima of* (6.17).

This proposition can be generalize to the multi-letter case, but slightly more tedious calculations dealing with multivariate Gaussian distributions are required. However, the essence of the proof is contained in the single-letter case, hence, we focus here on this case.

Corollary 2. *If* $a > a_1$ *the sum-capacity* $C = \log(1 + \frac{1}{1+a^2})$ *(achieved by treating interference as noise) is not tight.*

It has been shown in [3] to be tight till $a_0 \approx 0.42$ solving $a_0 + a_0^3 = 1/2$.

Proof. Note that

$$I(X_1; X_1 + aX_2 + W_1) = h(X_1 + aX_2 + W_1) - h(aX_2 + W_1)$$

$$\begin{aligned} h(X_1 + aX_2 + W_1) &= h(X_1^g + aX_2^g + W_1) - D(X_1 + aX_2 + W_1 || X_1^g + aX_2^g + W_1) \\ h(aX_2 + W_1) &= h(aX_2^g + W_1) - D(aX_2 + W_1 || aX_2^g + W_1). \end{aligned}$$

We now proceed to an input localization. Let

$$p_i = g_1(1 + \varepsilon L_i), \quad i = 1, 2,$$

where

$$\langle 1, L_i \rangle_{g_1} = \langle x, L_i \rangle_{g_1} = \langle x^2, L_i \rangle_{g_1} = 0 \text{ and } \|L_i\|_{g_1} < \infty.$$

Then,

$$\begin{aligned} & D(X_1 + aX_2 + W_1 || X_1^g + aX_2^g + W_1) \\ = \; & D(g_1(1+\varepsilon L_1) \star g_{a^2}(1+\varepsilon L_2^{(a)}) \star g_1 || g_1 \star g_{a^2} \star g_1), \end{aligned}$$

where

$$L_i^{(a)}(x) = L_i(x/a), \quad \forall i = 1, 2, \, x \in \mathbb{R}.$$

But,

$$\begin{aligned} & g_1(1+\varepsilon L_1) \star g_{a^2}(1+\varepsilon L_2^{(a)}) \star g_1 \\ = \; & g_1 \star g_{a^2} \star g_1 + \varepsilon(g_1 L_1 \star g_{a^2} \star g_1 + g_1 \star g_{a^2} L_2^{(a)} \star g_1) + \varepsilon^2 g_1 L_1 \star g_{a^2} L_2^{(a)} \star g_1, \end{aligned}$$

hence,

$$D(g_1(1+\varepsilon L_1) \star g_{a^2}(1+\varepsilon L_2^{(a)}) \star g_1 || g_1 \star g_{a^2} \star g_1) \tag{6.18}$$

$$= \frac{\varepsilon^2}{2} \left\| \frac{g_1 L_1 \star g_{a^2} \star g_1 + g_1 \star g_{a^2} L_2^{(a)} \star g_1}{g_1 \star g_{a^2} \star g_1} \right\|^2_{g_1 \star g_{a^2} \star g_1} + o(\varepsilon^2). \tag{6.19}$$

Let us now consider

$$L_1 = \sum_{k \geq 3} b_k \bar{H}_k^{[1]}, \quad L_2 = \sum_{k \geq 3} c_k \bar{H}_k^{[1]},$$

where the largest non-zero coefficient must be even in both expansions and $\sum_{k\geq 3} b_k^2 \vee \sum_{k \geq 3} c_k^2 < \infty$. From lemma 21,

$$\frac{g_{\alpha^2} \bar{H}_k^{[\alpha^2]} \star g_{\beta^2}}{g_{\alpha^2+\beta^2}} = \left(\frac{\alpha^2}{\alpha^2+\beta^2}\right)^{k/2} \bar{H}_k^{[\alpha^2+\beta^2]}.$$

Therefore,

$$\frac{g_1 \bar{H}_k^{[1]} \star g_{a^2} \star g_1}{g_1 \star g_{a^2} \star g_1} = \left(\frac{1}{2+a^2}\right)^{k/2} \bar{H}_k^{[2+a^2]} \quad \frac{g_1 \star g_{a^2} \bar{H}_k^{[a^2]} \star g_1}{g_1 \star g_{a^2} \star g_1} = \left(\frac{a^2}{2+a^2}\right)^{k/2} \bar{H}_k^{[2+a^2]},$$

implying

$$\frac{g_1 L_1 \star g_{a^2} \star g_1}{g_1 \star g_{a^2} \star g_1} = \sum_{k\geq 3} b_k (\frac{1}{2+a^2})^{k/2} \bar{H}_k^{[2+a^2]} \tag{6.20}$$

and, since $L_2^{(a)} = \sum_{k\geq 3} c_k \bar{H}_k^{[a^2]}$,

$$\frac{g_1 \star g_{a^2} L_2^{(a)} \star g_1}{g_1 \star g_{a^2} \star g_1} = \sum_{k\geq 3} c_k (\frac{a^2}{2+a^2})^{k/2} \bar{H}_k^{[2+a^2]}. \tag{6.21}$$

Using (6.20), (6.21) and

$$\langle \bar{H}_k^{[\alpha^2]}, \bar{H}_l^{[\alpha^2]} \rangle_{g_{\alpha^2}} = \delta_{k,l}, \quad \forall k, l \geq 0,$$

we can express (6.19) as

$$\frac{\varepsilon^2}{2} \sum_{k\geq 3} (b_k + c_k a^k)^2 \frac{1}{(2+a^2)^k} + o(\varepsilon^2). \tag{6.22}$$

Similarly, we get

$$\begin{aligned} D(aX_2 + W_1 || aX_2^g + W_1) &= D(g_{a^2}(1 + \varepsilon L_2^{(a)}) \star g_1 || g_{a^2} \star g_1) \\ &= \frac{\varepsilon^2}{2} \left\| \frac{g_{a^2} L_2^{(a)} \star g_1}{g_{a^2} \star g_1} \right\|^2_{g_{a^2} \star g_1} + o(\varepsilon^2) \\ &= \frac{\varepsilon^2}{2} \sum_{k\geq 3} c_k^2 (\frac{a^2}{1+a^2})^k + o(\varepsilon^2). \end{aligned} \tag{6.23}$$

Finally, from (6.22) and (6.23), we have

$$\begin{aligned} & I(X_1; X_1 + aX_2 + W_1) - h(X_1^g + aX_2^g + W_1) + h(aX_2^g + W_1) \\ = \; & I(X_1; X_1 + aX_2 + W_1) - \frac{1}{2} \log(\frac{2+a^2}{1+a^2}) \\ = \; & D(aX_2 + W_1 || aX_2^g + W_1) - D(X_1 + aX_2 + W_1 || X_1^g + aX_2^g + W_1) \\ = \; & \frac{\varepsilon^2}{2} \sum_{k\geq 3} \left[c_k^2 \left(\frac{a^2}{1+a^2} \right)^k - \frac{(b_k + c_k a^k)^2}{(2+a^2)^k} \right] + o(\varepsilon^2) \end{aligned}$$

This gives us the following localized problem

$$\sup_{\{b_k\}_{k\geq 3},\{c_k\}_{k\geq 3}} \sum_{k\geq 3} \left[(b_k^2+c_k^2)\left(\frac{a^2}{1+a^2}\right)^k - \frac{(b_k+c_k a^k)^2+(c_k+b_k a^k)^2}{(2+a^2)^k} \right],$$

where the coefficients must satisfy the conditions imposed earlier. We have

$$\begin{aligned} & (b_k^2+c_k^2)\left(\frac{a^2}{1+a^2}\right)^k - \frac{(b_k+c_k a^k)^2+(c_k+b_k a^k)^2}{(2+a^2)^k} \\ = & \frac{(b_k^2+c_k^2)a^{2k}}{(1+a^2)^k} - \frac{4b_k c_k a^k + (b_k^2+c_k^2)(1+a^{2k})}{(2+a^2)^k} \\ = & \left[\left(\frac{a^2}{1+a^2}\right)^k - \frac{1+a^{2k}}{(2+a^2)^k}\right](b_k^2+c_k^2) - 4\left(\frac{a}{2+a^2}\right)^k b_k c_k \end{aligned}$$

Hence, we have the following optimization

$$\sup_{\{b_k\}_{k\geq 3},\{c_k\}_{k\geq 3}} \sum_{k\geq 3} \left\{ \left[\left(\frac{a^2}{1+a^2}\right)^k - \frac{1+a^{2k}}{(2+a^2)^k}\right](b_k^2+c_k^2) - 4\left(\frac{a}{2+a^2}\right)^k b_k c_k \right\}. \tag{6.24}$$

The quadratic function

$$(b,c)\in\mathbb{R}^2 \mapsto \gamma(b^2+c^2) - 2\delta bc,$$

with $\delta > 0$, is always 0 at $(0,0)$, positive if $\gamma \geq \delta$, negative if $\gamma \leq -\delta$ and is a saddle if $-\delta < \gamma < \delta$. One can check that

$$\left[\left(\frac{a^2}{1+a^2}\right)^k - \frac{1+a^{2k}}{(2+a^2)^k}\right] + 2\left(\frac{a}{2+a^2}\right)^k$$

are increasing functions on $[0,1]$ for each $k \geq 3$, with a single zero $a_0(k)$ which decreases, so that the smallest zero value is achieved for $k = 3$ at

$$a_0(3) = 0.6796410242,$$

which is the only real root of the polynomial

$$a^6(2+a^2)^3 - (1-a^3)^2(1+a^2)^3.$$

Successively, we have $a_0(k)$ as the only real root of

$$a^{2k}(2+a^2)^k - (1-a^k)^2(1+a^2)^k,$$

with

$$\lim_{k\to\infty} a_0(k) = (\frac{\sqrt{5}-1}{2})^{\frac{1}{2}} = 0.7861513770,$$

which is the largest root of

$$a^4 + a^2 - 1.$$

This means that when $a \leq 0.6796410242$, the product of Gaussian distributions, i.e., $L_1 = L_2 = 0$, is at least a local maxima (we know it is a global maxima till 0.42, from the paper [3]). When $a > 0.6796410242$, we can have terms in (6.24) that are strictly positive, the higher in this interval the more positive terms we can have. In order for this to happen, we can not take $b_k = c_k$, but $b_k = -c_k$. The first Hermite polynomial that leads to a positive value for $a_0(3) = 0.6796410242$ is $\bar{H}_3$. However, $\bar{H}_3$ is not a valid direction. But for a strictly larger than $a_0(3)$, we can add an arbitrarily small portion of $\bar{H}_4$ to $\bar{H}_3$, in order to get a valid direction that is achieving a higher sum-rate than the Gaussian distribution. Hence, the fact that certain directions are not allowed in order to satisfy the positive constraint of the perturbation is virtual in this proof, since we can always add a infinitely small portion of an even Hermite polynomial of higher degree to make the perturbation valid. Finally, when $a > 0.7861513770$, there exists a K such that no matter how we choose the b_k and c_k for $k \geq K$, the terms in (6.24) are positive. □

All results presented in this chapter admit generalizations to the multi-letter (vector) case. The results developed in this section also give simple "local proofs" (sometimes even tightened versions) of the entropy power inequality, data processing in-

equality, monotonicity of entropy and other similar results. It also allows us to approach problems having an additive noise which is slightly non-Gaussian. Finally, it provides a strong tool to find counter-examples, which is particularly useful for complex problems dealing with interference or broadcast Gaussian channels. This work is being pursued.

Chapter 7

Ergodic MIMO Channels

We consider ergodic, coherent, MIMO channels. We characterize the optimal input distribution achieving capacity for various fading matrix distributions. First, we describe how symmetries in the fading matrix distribution and the constraint set are preserved as symmetries in the optimal input covariance; this will allow us to characterize the structure of the optimal covariance matrix and in some cases, it will fully determine this matrix. We will see that group structures and the notion of commutant appear as key elements. Second, we investigate the Kronecker model, in this case we will show how an asymmetric structure in the problem is also preserved in the optimal input structure, leading to a new water-filling situation.

Notation:

We define the following subsets of the $n \times n$ complex matrices $M_n(\mathbb{C})$, $n \geq 1$:

$H(n)$: the hermitian matrices,

$H_+(n)$: the hermitian positive semidefinite matrices,

$H_+^*(n)$: the hermitian positive definite matrices,

$U(n)$: the unitary matrices,

$\Pi(n)$: the group of permutations matrices,

$C(n)$: the group of cyclic permutations matrices,

$\Sigma(n)$: the group of diagonal matrices with $\{-1, +1\}$ elements.

7.1 Channel Model and Capacity

We consider a channel in which a vector input $x \in \mathcal{X} = \mathbb{C}^t$, $t \geq 1$, is received as a vector output $y \in \mathcal{Y} = \mathbb{C}^r$, $r \geq 1$, under the following assumptions. At each use ($i \geq 1$) of the channel:

- an $r \times t$ matrix H_i is drawn from an ergodic process having marginal probability measure μ_H,
- an $r \times 1$ vector n_i is drawn i.i.d from a complex circularly-symmetric Gaussian (C.C.S.G.) random variable of covariance matrix K, independently from the H_i's,
- the transmitter, without knowing the H_i's and n_i's, sends x_i,
- the receiver gets $y_i = H_i x_i + n_i$ together with H_i (and hence the term "coherent").

Moreover, the inputs $\{x_i\}$ are constrained in the following way. If the receiver and transmitter agree on a code book $\mathcal{C} = \{c(1), \ldots, c(M)\} \subset \mathcal{X}^n$, $n \geq 1$, then the code words must satisfy: $\frac{1}{n}\sum_{i=1}^{n} c(m)_i c(m)_i^* \in \mathrm{D}_t$, $\forall 1 \leq m \leq M$, where $\mathrm{D}_t \subset H_+^*(t)$ is a given compact set (we use $H_+^*(n)$ to denote the set of hermitian positive definite matrices of size $n \times n$).

Let C be the capacity of this channel under this general constraint. Then, denoting by X a random vector (r.ve.) in $\mathbb{C}^t$, we know from standard information theoretic arguments that

$$C(\mu_H, \mathrm{C}_t) = \max_{X \in \mathrm{C}_t} I(X; Y, H)$$

where

1. $\mathrm{C}_t = \{X | \mathbb{E}XX^* \in \mathrm{D}_t\}$ and $\mathrm{D}_t \subset H_+^*(t)$ is a compact set
2. $N \overset{(d)}{\sim} \mathcal{N}_{\mathbb{C}^r}(K)$ with $K \in H_+^*(r)$
3. H is a $\mathbb{C}^{r \times t}$-random matrix with probability measure μ_H,

4. (X, H, N) are mutually independent,

5. $Y = HX + N$.

Because C_t is entirely determined by D_t, we will note from now on

$$C(\mu_H, \mathrm{C}_t) = C(\mu_H, \mathrm{D}_t).$$

A particular example of constraints set is when $\mathrm{D}_t = \{A \in H_+^*(t) | \mathrm{tr}A \leq P\}$, for a given $P \in \mathbb{R}$. This is equivalent to asking for $\mathbb{E}X^*X \leq P$ and is called the *total power constraint.* An *individual power constraint* can also be considered, i.e. when $\mathbb{E}|Xi|^2 \leq P_i$, for a given $P_i \in \mathbb{R}$, $1 \leq i \leq t$, then the set D_t would be $\{A \in H_+^*(t) | A_{ii} \leq P_i, 1 \leq i \leq t\}$.

When $t = 1$, we maximize the mutual information over random variable (r.v.) X having variance in a compact set of $\mathbb{R}_+$, with maximal value, say, $P \in \mathbb{R}_+$. In this case, the optimal input is known to be a C.C.S.G. r.v. with variance P, no matter what μ_H is. More generally, one can show that in the vector setting, the Gaussian distribution is still optimal, but an optimization remains to be done on the covariance matrices in D_t; the result of which may depend on the distribution μ_H. In the case where H has i.i.d. C.C.S.G. entries and D_t is the set of covariance matrices with trace bounded by a given value $P \in \mathbb{R}$ (total power constraint), it has been shown in [32] that the optimal covariance matrix is $\frac{P}{t} I_t$ and the capacity is linearly increasing with $\min(t, r)$.

Questions:

1. The solution found when H has i.i.d C.C.S.G. entries is not surprising, in the sense that there are enough symmetries in the problem so that we expect a symmetric solution. But what does *enough symmetry* mean? What can we say when we have different *symmetric structures*, such as for example when we only have i.i.d entries? In other words, what are the relevant concepts of symmetry and how can we convert them into a specification of the solution?

2. What can we do when we have *asymmetric structures*?

We will develop some algebraic tools and present a result that gives an answer to the first question. We will see how group structure and notion of commutant comes into the picture as key features. This result is also applicable to other functionals than the capacity of MIMO channels. We will then investigate the Kronecker model (defined later) to get an understanding of the second question. Finally we will evaluate the capacity in several cases and show that it linearly increases with the dimension of the channel in several settings, independently of the law of the fading matrix entries.

Definition 31. We define the optimal inputs by

$$X_{\text{opt}}(\mu_H, \mathrm{D}_t) = \arg\max_{X \in \mathrm{C}_t} I(X; HX + N, H),$$

where $\arg\max_{X \in \mathrm{C}_t} f(X)$, for a real function f, denotes the set of the elements x satisfying $f(x) \geq f(y)$, $\forall y \in \mathrm{C}_t$.

We now use the assumptions we made on the channel to give a more specific expression for the capacity and the optimal inputs. The fact that the Gaussian distribution maximizes the entropy under a covariance constraint leads to the following result.

Proposition 23. *Let*

$$\psi : Q \in \mathrm{D}_t \mapsto \mathbb{E}^{\mu_H} \log\det(I + K^{-1} H Q H^*) \in \mathbb{R}, \tag{7.1}$$

which we call the mutual information function. Then, according to previous definitions and assumptions, we have

$$X_{\text{opt}}(\mu_H, \mathrm{D}_t) \sim \mathcal{N}_{\mathbb{C}^t}(Q_{\text{opt}}),$$

where

$$Q_{\text{opt}}(\mu_H, \mathrm{D}_t) = \arg\max_{Q \in \mathrm{D}_t} \psi(Q)$$

and

$$C(\mu_H, \mathrm{D}_t) = \max_{Q \in \mathrm{D}_t} \psi(Q).$$

7.2 Symmetries

7.2.1 Quantifying Symmetries

Assume that the channel has the same output distribution when sending any input X or a permuted version of it, say, PX, where P is a permutation matrix.

$$Y = HX + N \iff Y = H(PX) + N.$$

Then, we talk about a *symmetry of the channel* with respect to that transformation P. But from previous equivalence, this is to say that

$$HP \overset{(d)}{\simeq} H.$$

Remarks:

1. This type of invariance has a natural group structure: assume you have the invariance $HP_i \overset{(d)}{\simeq} H$ for a set of matrices P_i, then clearly $HP_iP_j \overset{(d)}{\simeq} H$ and if P_i is invertible $HP_i^{-1} \overset{(d)}{\simeq} H$. Thus this invariance still holds for the group generated by this set.

2. In order to compare X and PX, we need to ensure that PX is satisfying the considered contraint too, i.e. its covariance matrix $P(\mathbb{E}XX^*)P^*$ has to belong to D_t as well.

3. Groups other than the permutations might be of interest, for example if we want to consider situation where the symmetry is expressed by keeping the channel equivalent whether we send an input X or a modified version of it where some component's signs have been flipped, then the group of diagonal matrices with 1 and -1 is the appropriate group.

These remarks motivate the following definitions.

Definition 32. Let G be a group in $M_n(\mathbb{C})$.

1. A random matrix is G-invariant (on the right) if $Hg \overset{(d)}{\simeq} H$, $\forall g \in G$.

2. A set of matrices $D \subset M_n(\mathbb{C})$ is invariant in G-conjugation if $gQg^{-1} \in D$, $\forall Q \in D, g \in G$.

3. A function $\Psi : D \to \mathbb{R}$ is G-invariant if D is invariant in G-conjugation and if $\Psi(gQg^{-1}) = \Psi(Q)$, $\forall Q \in D, g \in G$.

Note that only subgroups of unitary matrices are of interest regarding our MIMO channel setting, because the mutual information function evaluated at Q depends on the distribution of HQH^*. Examples of functions which are invariant in G-conjugation for unitary subgroups are all functions of the form $x \rightsquigarrow \mathbb{E}f(MxM^*)$ where f is any measurable function and M is a random matrix that is G-invariant on the right. The reason for which a "conjugation" invariance for unitary subgroups is relevant in our MIMO settings is a consequence of the fact that we are working with a second order moment constraint, which implies that the mutual information has precisely the above described form (cf. (7.1)). Finally, examples of groups in $M_n(\mathbb{C})$ are $U(n)$, which is the largest group we will consider, and its subgroups $\Sigma(n)$ and $\Pi(n)$ (with the usual matrix multiplication), defined as:

1. $U(n)$: the unitary group of size $n \times n$,

2. $\Pi(n)$: the group of permutation matrices of size $n \times n$,

3. $\Sigma(n)$: the diagonal matrices group with 1 and -1 of size $n \times n$.

We now gave a definition to quantify symmetries in the problem, through *the group of invariance* of H and D, or equivalently of ψ, the question is then: how do we use this invariance in order to get knowledge on the optimal input? In the next section we will see that this is done through the commutant.

7.2.2 Invariant Structures

Definition 33. The commutant of G is defined by the algebra $\mathrm{Comm}(G) = \{A \in M_n(\mathbb{C}) | Ag = gA, \forall g \in G\} = \{A \in M_n(\mathbb{C}) | A = gAg^{-1}, \forall g \in G\}$.

We start with a trivial observation linking the commutant and G-invariant functions.

Lemma 12. *Let $G \subset M_n(\mathbb{C})$ be a group and $D \subset M_n(\mathbb{C})$. Let $\Psi : D \to \mathbb{R}$ a G-invariant function having a unique maximizer Q_{opt}. Then $Q_{\mathrm{opt}} \in \mathrm{Comm}(G) \cap D$.*

Proof. We have $\psi(Q_{\mathrm{opt}}) = \psi(gQ_{\mathrm{opt}}g^{-1})$, $\forall g \in G$. We conclude by the uniqueness of the maximizer. □

Note that the bigger the group, the smaller the commutant, which is what we expect in order to exploit symmetries.

Some inequalities can be achieved by requesting further hypotheses, namely if the group G is compact, the set D is convex, and the function Ψ is strictly concave, then, denoting by $\mathbb{G}$ a random variable with values in G and probability measure P_G on G, we have by Jensen's inequality

$$\mathbb{E}^{P_G}\Psi(\mathbb{G}d\mathbb{G}^{-1}) \leq \Psi(d^{P_G}),$$

where $d^{P_G} := \mathbb{E}^{P_G}\mathbb{G}D\mathbb{G}^{-1}$ and since $\Psi(GdG^{-1}) = \Psi(d)$, $\forall d \in D$, the last inequality becomes

$$\Psi(d) \leq \Psi(d^{P_G}).$$

Note that if Q_G denotes another probability measure on G, $(d^{P_G})^{Q_G} = d^{P_G \star Q_G}$, with $P_G \star Q_G = \int_G (\tau_h)_* Q_G P_G(dh)$, where $(\tau_h)_* Q_G(\Gamma) = Q_G(\Gamma h^{-1})$, for $\Gamma \in \mathcal{B}_G$. Furthermore, $P_G \star U_G = U_G \star P_G = U_G$, where U_G denotes the normalized Haar measure on the right on G. Therefore we have

$$\Psi(d) \leq \Psi(d^{P_G}) \leq \Psi((d^{P_G})^{U_G}) = \Psi(d^{U_G}),$$

which gives, by the last inequality, an estimate d^{U_G} belonging to $\mathrm{Comm}(G) \cap D$, agreeing with the previous lemma, as Ψ is strictly concave.

Invariant Structures in MIMO

We rewrite the previous observations in our MIMO channel context.

Proposition 24. *Let a MIMO channel be as defined in the introduction and let G be a subgroup of $U(t)$. If*

- *the constraint set D_t is invariant in G-conjugation,*
- *the fading matrix distribution μ_H is G-invariant,*

then

$$Q_{\mathrm{opt}} \in \mathrm{Comm(G)} \cap \mathrm{D_t}.$$

Proof. Observe that under these assumptions, the function ψ in (7.1) is G-invariant, moreover it is strictly concave on the set of positive definite matrices, thus lemma 12 applies. □

Also note that if G_1, G_2 are two groups in $U(t)$ and if D_t is invariant in G_1-conjugation whereas μ_H is G_2-invariant, then

$$Q_{\mathrm{opt}} \in \mathrm{Comm}(G_1 \cap G_2) \cap \mathrm{D}_t.$$

We will now see some specific applications of previous proposition. The cases that we will consider are dealing with the following commutants:

$$\mathrm{Comm}(\Sigma(n)) \text{ is the set of diagonal matrices in } M_n(\mathbb{C}),$$

$$\mathrm{Comm}(\Pi(n)) = \{\alpha I_n + \beta J_n | \alpha, \beta \in \mathbb{C}\}, \text{ where } J_n = 1^{n \times n},$$

$$\mathrm{Comm}(U(n)) = \{\alpha I_n | \alpha \in \mathbb{C}\}.$$

Corollary 4. *Total power constraint*

For a given $P \in \mathbb{R}_+$, we consider $Q \in \mathcal{D}_t = \{Q \in H_+^*(t) | \mathrm{tr}(Q) \leq P\}$. If μ_H is invariant in G-conjugation for a subgroup G of $U(t)$, then $Q_c \in \mathrm{Comm}(G) \cap \mathcal{D}_t$.

Simply observe that $\mathcal{D}_t$ is invariant in $U(t)$-conjugation. Two interesting cases of subgroups of $U(t)$ are $\Pi(t)$ and $\Sigma(t)$. From what we saw in the examples of the commutant, if we consider a distribution μ_H invariant under $\Sigma(t)$, then Q_c is diagonal and if it is invariant under $\Pi(t)$, then Q_c will have the same value for all components inside the diagonal ($\frac{P}{t}$ if one works in $\mathcal{D}_t$) and will also have the same value for all elements outside the diagonal, as long as it stays a positive definite matrix. Examples of $\Sigma(t)$-invariant random matrices are matrices with independent symmetric entries (symmetric means that $H_{ij} \overset{(d)}{\simeq} -H_{ij}$) and examples of $\Pi(t)$-invariant ones are matrices with i.i.d. entries or jointly Gaussian entries having a covariance matrix of the form $\alpha I_{rt} + \beta J_{rt}$.

Corollary 5. *Still considering $Q \in \mathcal{D}_t$, if H is $\Pi(t)\Sigma(t)$-invariant, which is for example the case when H_{ij} are i.i.d. and $H_{ij} \overset{(d)}{\simeq} -H_{ij}$, $\forall 1 \leq i \leq r, 1 \leq j \leq t$, then $Q_c = \frac{P}{t} I_t$.*

This is a particular case of corollary 1, where we consider the product group $\Pi(t)\Sigma(t) \subset U(t)$ containing all permutations matrices with $+1$ and -1. In this case we have that $\mathrm{Comm}(\Pi(t)\Sigma(t))$ contains only multiples of the identity and since $Q \in \mathcal{D}_t$ has normalized trace, the result follows. Note that we did not assume that the entries of H are Gaussian (which would be a particular case of this) in order to get $\frac{P}{t} I_t$ as a maximizer. Also note that the group $\Pi(t)$ could be replaced by $C(t)$, the group of cycling permutations, and we would get the same conclusion. Generally, this will be true as long as we have a group of invariance G such that $\mathrm{Comm}(G)$ is reduced to the multiple's of the identity.

Corollary 6. *Local power constraint*

If X is constrained by $\mathbb{E}|X_i|^2 \leq P_i$ for given $P_i \in \mathbb{R}_+$, $\forall 1 \leq i \leq t$, and if H is

$\Sigma(t)$-*invariant, then* $Q_c = \mathrm{diag}(P_1, \ldots, P_t)$.

The constraint $\mathbb{E}|X_i|^2 \leq P_i$ implies that $Q \in \widetilde{\mathcal{D}}_t = \{Q \in H_+^*(t) | Q_{ii} \leq P_i, \forall 1 \leq i \leq t\}$, now we no longer have that $\widetilde{\mathcal{D}}_t$ is invariant in $U(t)$-conjugation, but we still have, for example, invariance in $\Sigma(t)$-conjugation. Therefore, if H is $\Sigma(t)$-invariant, the optimal covariance matrix will be commuting with this group, which means it is diagonal and thus the optimal diagonal elements are the corresponding P_i's (we can increase ψ by increasing the trace).

Conclusion: As it has been illustrated in previous example, the problem of symmetries should be generally approached in the following way: first identify the invariance property of the domain D_t in which we are working (we saw examples of total and local power constrain (see corollaries 2 and 3), several intermediary cases are possible), then identify the invariance property of the fading matrix distribution μ_H, once we have these two groups of invariance, we know that we can restrict our search of Q_c to matrices commuting with these groups and staying in D_t. Which means that *the commutant is summarizing the information given by the symmetries in the problem.* We saw that in some cases (see corollary 2) this allows us to fully specify the optimal input covariance matrix, whereas in other cases, it only reduces the dimension of the optimization problem (such as for example in corollary 2, when we have a $\Sigma(t)$-invariance, we are left with t degrees of freedom for Q_c instead of $\frac{t^2+t}{2}$ at the beginning).

7.2.3 Asymptotic Capacity

In [32], it is shown that the capacity is linearly increasing with the dimension of the channel, more precisely with $\min(t, r)$. Although we showed that the covariance matrix $\frac{1}{t}I$ was still optimal in a more general setting than in the i.i.d. Gaussian fading one, we may now wonder whether the linear increase of the capacity can be lost if we drop this assumption. The following result confirms that in several settings, this

property is still preserved.

Definition 34. Let M be a matrix in $H_+(n)$, the set of hermitian positive semidefinite matrices of size $n \times n$. We define the EDF (empirical distribution function) of the eigenvalues of M, also called the spectral distribution of M, as

$$p^n_{\lambda(M)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(M)}.$$

Proposition 25. *Let $H = AW$, with W having $r \times t$ i.i.d. symmetric entries (variance 1) and $A \in H_+(r)$ admitting a limiting spectral distribution ν_A. Then, defining $m = r \wedge t$, $n = r \vee t$ and $\tau = n/m$, we have*

$$\lim_{m\to\infty} \frac{C(\mu_H, \mathcal{D}_m)}{m} = \gamma > 0,$$

where γ is a constant of the form $\int_{\mathbb{R}} \log(1 + Px) d\mu_{A,\tau}(x)$ and $\mu_{A,\tau}$ is a probability measure depending on ν_A and τ.

Proof. This is a consequence of corollary 2 and a theorem of Marchenko and Pastur (cf. [22]). □

In particular, if $r = t$ and $A = I$, then $\mu_{I,1} = \frac{1}{\pi}\sqrt{\frac{1}{x} - \frac{1}{4}}\mathbb{1}_{[0,4]}(x)dx$.
The capacity gain is not the only great feature of MIMO channels, different kinds of gains concerning MIMO channels are investigated in [33].

7.2.4 Bringing the Symmetries

In some situations, a symmetric structure is not clearly existing, but with an appropriate transformation one can bring some symmetries back into the problem. We now give two examples of how to carry out such a procedure, they are both based on the following simple observation.

Lemma 13. *Let $\Psi : D \to \mathbb{R}$ with a unique maximizer Q_{opt} and such that D is*

invariant in G-conjugation. Then, for any $M \in G$, we have

$$M^* Q_{\text{opt}} M = \arg\max_{Q \in D} \Psi(MQM^*).$$

We will use this simple "change of basis" in the following two sections.

The Kronecker model

We define the following specific MIMO channel.

Definition 35. The Kronecker model
We consider the constraint set $\mathcal{D}_t = \{Q \in H_+^*(t) | \text{tr}(Q) \leq P\}$, and $H = AWB$, where

- $A \in M_r(\mathbb{C})$ non-zero,
- $B \in M_t(\mathbb{C})$ non-zero,
- W is a $r \times t$ random matrix being $U(t)$-invariant on the right.

In this case, the mutual information function ψ is given by

$$Q \in \mathcal{D} \mapsto \psi(Q) = \mathbb{E} \log\det(I + K^{-1} AWBQB^*W^*A^*).$$

We now denote the SVD of B by $B = U_B \,\text{diag}(b) V_B^*$, where $U_B, V_B \in U(t)$ and $b \in \mathbb{R}_+^t$. Using our previous lemma, we can choose $M = V_B$, in order to get that

$$V_B^* Q_{\text{opt}} V_B \;=\; \arg\max_{Q \in D} \psi(V_B Q V_B^*) \tag{7.2}$$

The advantage of having to deal with the above maximization problem

$$Q \rightsquigarrow \psi(V_B Q V_B^*)$$

is a $\Sigma(t)$-invariant function and thus we can restrict our maximization to matrices being diagonal (with trace smaller than P). In other words, we showed the following observations:

Remark 1: the eigenvectors of Q_{opt} are the right-eigenvectors of B and its eigenvalues $q_{\text{opt}} = (q_1^{\text{opt}}, \ldots, q_t^{\text{opt}})$ are given by

$$\begin{aligned} q_{\text{opt}} &= \arg \max_{q \in \mathbb{R}_+ \text{ s.t. } \sum_{i=1}^t q_i \leq P} \mathbb{E} \log \det(I \\ &\quad + K^{-1} A W \operatorname{diag}(q_1 b_1^2, \ldots, q_t b_t^2)) W^* A^*). \end{aligned}$$

Note that if the additive noise N is assumed to have covariance $K = I$ and if W is $U(t)$-invariant on the left too, then one can equivalently consider A to be diagonal because $\det(I + MN) = \det(I + NM)$, no matter what the matrices M and N are (as long as the dimensions match).

So for this model, we have reduced the number of parameters from $t(t+1)/2$ to t by bringing a $Z_(t)$-invariance, or simply by changing our problem in the right basis at the transmitter and at the receiver. In the next section we will further investigate this model.

The Ricean model

We define the following specific MIMO channel.

Definition 36. We consider the constraint set $\mathrm{D}_t = \{Q \in H_+^*(t) | \operatorname{tr}(Q) \leq P\}$, and $H = A + W$, where

- $A \in M_{t \times t}(\mathbb{C})$ non-zero,
- W is a $r \times t$ random matrix being $U(t)$-invariant on the right and on the left.
- the covariance of the additive noise N is $K = I$

In this case, the mutual information function ψ is given by

$$Q \in \mathcal{D} \mapsto \psi(Q) = \mathbb{E} \log \det(I + (A + W) Q (A^* + W^*)).$$

We now denote the SVD of A by $A = U_A \operatorname{diag}^{r \wedge t}(a) V_A^*$ (a diagonal matrix of dimension $r \wedge t$ completed with 0 to dimension $r \times t$), where $U_A \in U(r)$, $V_A \in U(t)$ and

$a \in \mathbb{R}_+^{r\wedge t}$. Using our previous lemma with $M = U_A$ and the formula $\det(I + MN) = \det(I + NM)$, we can replace A by $\mathrm{diag}^{r\wedge t}(a)$ in the capacity expression. We now claim that

$$Q \rightsquigarrow \log\det(I + (\mathrm{diag}^{r\wedge t}(a) + W)Q(\mathrm{diag}^{r\wedge t}(a) + W^*))$$

is $\Sigma(t)$-invariant. In fact, although $\mathrm{diag}^{r\wedge t}(a)+W$ is not $\Sigma(t)$-invariant on the right, it is $\Sigma(t)$-invariant in conjugation: for any matrices of the form $Z_i = \mathrm{diag}(1,\dots,1,-1,1,\dots,1) \in M_t(\mathbb{C})$, where the -1 value appears at the i^{th} component, we can consider the matrix $Z_i^{r\wedge t} \in M_r(\mathbb{C})$ which is equal to Z_i completed with 1's if $r \wedge t = t$ and Z_i truncated if $r \wedge t = r$. We then have

$$Z_i^{r\wedge t}\,\mathrm{diag}(a)Z_i = \mathrm{diag}(a),\ \ \forall 1 \le i \le t$$

and since the matrices Z_i's generate the group $\Sigma(t)$, we get that H is invariant in $\Sigma(t)$-conjugation. Therefore, using the formula $\det(I + MN) = \det(I + NM)$ we can conclude for the $\Sigma(t)$-invariance of ψ. In conclusion, we showed the following observation:

Remark 2: the eigenvectors of Q_{opt} are the right-eigenvectors of A and its eigenvalues $q_{\mathrm{opt}} = (q_1^{\mathrm{opt}},\dots,q_t^{\mathrm{opt}})$ are given by

$$\begin{aligned} q_{\mathrm{opt}} &= \arg \max_{q\in\mathbb{R}_+ \text{ s.t. } \sum_{i=1}^t q_i \le P} \mathbb{E}\log\det(I \\ &\quad +(\mathrm{diag}(a) + W)\,\mathrm{diag}(q_1,\dots,q_t)(\mathrm{diag}(a) + W^*)). \end{aligned}$$

We conclude this section with the following result.

Proposition 26. *For the ricean model with $r = 1$, we have $Q^{\mathrm{opt}} = V_A\,\mathrm{diag}(q^{\mathrm{opt}})V_A^*$, where V_A is such that $AV_A = (\alpha, 0\dots,0)$, $\alpha = (\sum_{i=1}^t A_{1i}^2)^{\frac{1}{2}}$ and*

$$q^{\mathrm{opt}} = (\lambda, \frac{P-\lambda}{t-1},\dots,\frac{P-\lambda}{t-1})$$

where $\lambda = \arg\max_{0\le x\le P} \mathbb{E}\log(1 + x|\alpha + w_{11}|^2 + \frac{P-x}{t-1}\sum_{i=2}^t |w_{1i}|^2)$.

Proof. We only need to check that the eigenvalues can take only two different values

(the rest of the proposition is a direct consequence of previous expansions). The reason for this is that in this specific case, the mutual information function

$$\begin{aligned}\mathrm{diag}(q) \in \mathcal{D} \quad \mapsto \quad & \mathbb{E}\log(1 + ((\alpha, 0, \ldots, 0) + W))(\mathrm{diag}(q)) \\ & ((\alpha, 0, \ldots, 0)^T + W^*))\end{aligned}$$

is not only $\Sigma(t)$-invariant but also $\Pi(t)_{2:t}$-invariant, where $\Pi(t)_{2:t}$ denotes the group of permutations keeping the first component fixed (i.e. the first column is $(1, 0, \ldots, 0)$ for any of these matrices). We also use the fact that the maximum must be achieved for matrices having trace equal to P (as we can increase ψ by increasing the value of the trace). □

7.3 Asymmetries

As we saw in last section, there are not always enough symmetries in order to fully characterize the optimal input. For example, suppose that $H = WB$, where W has $r \times t$ i.i.d C.C.S.G entries and $B = \mathrm{diag}(b)$, with $b \in \mathbb{R}_+^t$. Then we know that the optimal covariance matrix is diagonal but we do not know the value of the diagonal elements. Now assume that $b_1 \leq \ldots \leq b_n$, can we then expect that the optimal covariance matrix should preserve this ordering in some sense?

We will investigate the Kronecker model with $H = AWB$ (cf. previous section) to analyze these kinds of questions. We will present two propositions that will help describe the optimal input for such a channel. If the random matrix H were replaced by the deterministic matrix B, we know that the optimal input covariance has eigenvalues q_i^{opt} given via "water-filing" on the singular values of B (cf. [32]). Two particular properties of the "water-filing" solution are the following.

1. **Monotonicity:** if $b_i \geq b_j$ then $q_i^{\mathrm{opt}} \geq q_j^{\mathrm{opt}}$ (with equality if $b_i = b_j$)

2. **On/Off threshold:** if b_{i+1} is sufficiently bigger than b_i, then the power of P may consequently be divided amongst the $t - i$ biggest components of b.

We will see in the next two propositions that these two properties are preserved in the Kronecker model.

7.3.1 Martini Filling

We start with the monotonicity result.

Proposition 27. *We have*

$$Q_{\text{opt}} = V_B \operatorname{diag}(q_{\text{opt}}) V_B^*$$

where q_{opt} *satisfies*

$$q_i^{\text{opt}} \geq 0, \ \textstyle\sum_{i=1}^t q_i^{\text{opt}} = P,$$
$$q_i^{\text{opt}} \geq q_j^{\text{opt}} \text{ if } b_i > b_j, \text{ and } q_i^{\text{opt}} = q_j^{\text{opt}} \text{ if } b_i = b_j.$$

Note: If $B = I_t$ *and* μ_W *is* G*-invariant on the right with* $G \leq U(t)$*, then* $Q_{\text{opt}} \in \operatorname{Comm}(G) \cap \mathcal{D}_t$.

Remark: This proposition says that the eigenvectors of Q_{opt} are the right-eigenvectors of B (which has been shown in previous section) and that its eigenvalues are monotonically distributed with respect to the singular values of B.

In order to prove this result, we need a preliminary lemma. Let $\lambda_1(M) \leq \ldots \leq \lambda_n(M)$ denote the ordered eigenvalues of any matrix $M \in H(n)$ — we use $H(n)$ to denote the set of hermitian matrices of size $n \times n$.

Lemma 14. *Let* $n \geq 1$, $P \in H_+^*(n)$ *and* $H \in H(n)$*. We then have,*

$$\lambda_k(H + P) > \lambda_k(H), \ \forall k = 1, \ldots, n.$$

Proof. This is a corollary of a *Weyl*'s theorem, which says that for any $H_1, H_2 \in H(n)$

and k=1,..., n,

$$\lambda_k(H_1) + \lambda_1(H_2) \leq \lambda_k(H_1 + H_2) \leq \lambda_k(H_1) + \lambda_n(H_2).$$

To prove the latter result, we use the *Courant-Fisher*'s theorem

$$\lambda_k(H) = \min_{s_1,\ldots,s_{n-k} \in \mathbb{C}^n} \max_{\substack{x \in \mathbb{C}^n \text{ s.t. } x^*x=1 \\ x \perp s_1,\ldots,s_n}} x^*Hx$$

and the fact that

$$\lambda_1(H) \leq x^*Hx \leq \lambda_n(H), \ \forall x \in \mathbb{C}^n \text{ s.t. } x^*x = 1,$$

this allows us to write

$$\begin{aligned}
\min_{s_1,\ldots,s_{n-k} \in \mathbb{C}^n} \max_{\substack{x \in \mathbb{C}^n \text{ s.t. } x^*x=1 \\ x \perp s_1,\ldots,s_n}} & x^*H_1x + \lambda_n(H_2) \geq \lambda_k(H_1 + H_2) \\
&= \min_{s_1,\ldots,s_{n-k} \in \mathbb{C}^n} \max_{\substack{x \in \mathbb{C}^n \text{ s.t. } x^*x=1 \\ x \perp s_1,\ldots,s_n}} (x^*H_1x + x^*H_2x) \\
&\geq \min_{s_1,\ldots,s_{n-k} \in \mathbb{C}^n} \max_{\substack{x \in \mathbb{C}^n \text{ s.t. } x^*x=1 \\ x \perp s_1,\ldots,s_n}} x^*H_1x + \lambda_1(H_2)
\end{aligned}$$

which proves the *Weyl*'s theorem. The left bound of this result and the fact that $\lambda_1(P) > 0$ proves the lemma. □

Proof of proposition 27. The initial expression of the mutual information function for this channel is

$$\psi(Q) = \mathbb{E} \log \det(I + K^{-1}AWBQB^*W^*A^*).$$

First note that A affects the function ψ in the same way as K^{-1}, in other words, we could consider one of these two matrices to be identity, for example, assume i.i.d. components for the noise and set $\tilde{A} = K^{-\frac{1}{2}}A$. If $B = I_r$ any invariance properties on the right for μ_W will be preserved for AW, thus the note after the proposition is a direct consequence of proposition 24.

The first part of the proposition is proved in the previous section, let us now look at

the eigenvalues. We have

$$q_{\text{opt}} = \arg \max_{q \in \mathbb{R}_+ \text{ s.t. } \sum_{i=1}^t q_i \leq P} \mathbb{E} \log \det(I + AW \operatorname{diag}(q_1 b_1^2, \ldots, q_t b_t^2)) W^* A^*).$$

Thus we will consider from now on

$$\psi : q \rightsquigarrow \mathbb{E} \log \det(I + AW \operatorname{diag}(q_1 b_1^2, \ldots, q_t b_t^2)) W^* A^*).$$

Now observe that if $b_i = b_j$ then ψ is $\Pi(t)_{ij}$-invariant, where $\Pi(t)_{ij}$ is the subgroup of permutations keeping the diagonal elements different than the i and j invariant (transposition), thus we get from proposition 24 that $q_i^{\text{opt}} = q_j^{\text{opt}}$.
Now, let $P' = P - \sum_{i=3}^t q_i^{\text{opt}}$, such that $q_1^{\text{opt}} + q_2^{\text{opt}} = P'$. We will show that if $b_1 > b_2$, then for any $0 \leq P' \leq P$,

$$\partial_{q_1} \psi(q)\big|_{(\frac{P'}{2}, \frac{P'}{2}, q_3^{\text{opt}}, \ldots, q_t^{\text{opt}})} > \partial_{q_2} \psi(q)\big|_{(\frac{P'}{2}, \frac{P'}{2}, q_3^{\text{opt}}, \ldots, q_t^{\text{opt}})},$$

which, by the concavity of ψ, implies that

$$q_1^{\text{opt}} > q_2^{\text{opt}}.$$

By symmetry of the problem, this clearly implies the result for any components i and j (other than 1 and 2).

We have

$$\psi(q) = \mathbb{E} \log \det(I + \sum_{i=1}^t q_i b_i^2 A w_i (A w_i)^*)$$

where w_i is the i-th column of W. For an invertible matrix M, we have the formula $\partial_{m_{ij}} \log \det(M) = (M^{-1})_{ji}$, therefore we have

$$\partial_{q_j} \psi(q) = b_j^2 \mathbb{E} \operatorname{tr}\big(I + \textstyle\sum_{i=1}^t q_i b_i^2 A w_i (A w_i)^*\big)^{-1} A w_j (A w_j)^*.$$

Let us denote $X_i = Aw_i(Aw_i)^*$, which are hermitian positive semidefinite matrices, as well as $\left(I + \sum_{i=1}^{t} q_i b_i^2 X_i\right)$ which is in addition positive definite and invertible. We define $Z = \sum_{i=3}^{t} q_i b_i^2 X_i$ and $Z_{\text{opt}} = \sum_{i=3}^{t} q_i^{\text{opt}} b_i^2 X_i$, we then rewrite

$$\begin{aligned} \partial_{q_1}\psi(q) &= b_1^2 \mathbb{E}\,\mathrm{tr}\big(I + q_1 b_1^2 X_1 + q_2 b_2^2 X_2 + Z\big)^{-1} X_1 \qquad (7.3) \\ \partial_{q_2}\psi(q) &= b_2^2 \mathbb{E}\,\mathrm{tr}\big(I + q_1 b_1^2 X_1 + q_2 b_2^2 X_2 + Z\big)^{-1} X_2 \\ &= b_2^2 \mathbb{E}\,\mathrm{tr}\big(I + q_1 b_1^2 X_2 + q_2 b_2^2 X_1 + Z\big)^{-1} X_1 \qquad (7.4) \end{aligned}$$

where in the last line we interchanged the random matrices X_1 and X_2, as W is $\Pi(t)$-invariant. To conclude the proof, we must show that if $b_1 > b_2$

$$\begin{aligned} & b_1^2 \mathbb{E}\,\mathrm{tr}\Big(I + \tfrac{P'}{2} b_1^2 X_1 + \tfrac{P'}{2} b_2^2 X_2 + Z_{\text{opt}}\Big)^{-1} X_1 \\ & > b_2^2 \mathbb{E}\,\mathrm{tr}\Big(I + \tfrac{P'}{2} b_1^2 X_2 + \tfrac{P'}{2} b_2^2 X_1 + Z_{\text{opt}}\Big)^{-1} X_1, \end{aligned}$$

for any $0 \leq P' \leq 1$. This is clearly satisfied in the scalar case ($r = 1$). In the matrix case, a few more steps (using the previous lemma) are required to show that the result hold. We now define

$$\chi_1, \chi_2 : \ [0,1] \ \rightarrow \ \mathbb{R}$$

by

$$\begin{aligned} \chi_1(\varepsilon) &= b_1^2 \mathrm{tr}\Big(I + \tfrac{P'}{2} b_1^2 X_1(\varepsilon) + \tfrac{P'}{2} b_2^2 X_2(\varepsilon) + Z_{\text{opt}}\Big)^{-1} X_1(\varepsilon) \\ \chi_2(\varepsilon) &= b_2^2 \mathrm{tr}\Big(I + \tfrac{P'}{2} b_1^2 X_2(\varepsilon) + \tfrac{P'}{2} b_2^2 X_1(\varepsilon) + Z_{\text{opt}}\Big)^{-1} X_1(\varepsilon) \end{aligned}$$

where $X_i(\varepsilon) = X_i + \varepsilon I_r$. Note that for $i = 1, 2$, χ_i are continuous functions. Therefore, $\lim_{\varepsilon \searrow 0} \chi_i(\varepsilon) = \chi_i(0)$. Moreover, from (7.3) and (7.4) we have

$$\mathbb{E}\chi_i(0) = \partial_{q_i}\psi(q)\big|_{(\frac{P'}{2}, \frac{P'}{2}, q_3^{\text{opt}}, \ldots, q_t^{\text{opt}})}, \quad i = 1, 2.$$

Let us now consider $\varepsilon \in (0, 1]$, we have that $X_i(\varepsilon)$ is in $H_+^*(r)$, and is thus invertible,

so we can write $\chi_i(\varepsilon) = \mathrm{tr}\left(\frac{P'}{2} I_r + M_i\right)^{-1}$ or equivalently

$$\chi_j(\varepsilon) = \sum_{i=1}^{r} \frac{1}{\frac{P'}{2} + \lambda_i(M_j)}, \quad j = 1, 2$$

where

$$M_1 := X_1^{-1}(\varepsilon) \left(b_1^{-2} I_r + \frac{P'}{2} b_1^{-2} b_2^2 X_2(\varepsilon) + b_1^{-2} Z_{\text{opt}} \right)$$

and

$$M_2 := X_1^{-1}(\varepsilon) \left(b_2^{-2} I_r + \frac{P'}{2} b_2^{-2} b_1^2 X_2(\varepsilon) + b_2^{-2} Z_{\text{opt}} \right).$$

If we try to directly insert $X_1^{-1}(\varepsilon)$ in the parenthesis of the above expressions, we will not be able to apply lemma 14 part *(ii)*, as $X_1^{-1}(\varepsilon) X_2(\varepsilon)$ may not be hermitian, even though $X_1^{-1}(\varepsilon) \in H_+^*(r)$ and $X_2(\varepsilon) \in H_+(n)$ (all of these affirmations are in the probabilistic "surely" sense). However, from lemma 14 part *(i)*, we have that the non-zero eigenvalues of M_1 are the same as the ones of

$$X_1^{-\frac{1}{2}}(\varepsilon) \left(b_1^{-2} I_r + \frac{P'}{2} b_1^{-2} b_2^2 X_2(\varepsilon) + b_1^{-2} Z_{\text{opt}} \right) X_1^{-\frac{1}{2}}(\varepsilon)$$

which is equal to

$$\begin{aligned} &b_1^{-2} X_1^{-1}(\varepsilon) + \tfrac{P'}{2} b_1^{-2} b_2^2 X_1^{-\frac{1}{2}}(\varepsilon) X_2(\varepsilon) X_1^{-\frac{1}{2}}(\varepsilon) \\ &+ b_1^{-2} X_1^{-\frac{1}{2}}(\varepsilon) Z_{\text{opt}} X_1^{-\frac{1}{2}}(\varepsilon) =: N_1 \end{aligned}$$

and that the non-zero eigenvalues of M_2 are the same as the ones of

$$X_1^{-\frac{1}{2}}(\varepsilon) \left(b_2^{-2} I_r + \frac{P'}{2} b_2^{-2} b_1^2 X_2(\varepsilon) + b_2^{-2} Z_{\text{opt}} \right) X_1^{-\frac{1}{2}}(\varepsilon)$$

which are equal to

$$\begin{aligned} &b_2^{-2} X_1^{-1}(\varepsilon) + \tfrac{P'}{2} b_2^{-2} b_1^2 X_1^{-\frac{1}{2}}(\varepsilon) X_2(\varepsilon) X_1^{-\frac{1}{2}}(\varepsilon) \\ &+ b_2^{-2} X_1^{-\frac{1}{2}}(\varepsilon) Z_{\text{opt}} X_1^{-\frac{1}{2}}(\varepsilon) =: N_2. \end{aligned}$$

And now we have

$$N_1 - N_2 \in H_+^*(r) \text{ surely},$$

therefore, we conclude from lemma 14 that

$$\chi_1(\varepsilon) > \chi_2(\varepsilon) \text{ surely, } \forall \varepsilon \in (0,1].$$

Thus, by the continuity of χ_i on $[0,1]$ and monotony of the expectation, we have

$$\chi_1(0) \geq \chi_2(0) \Longrightarrow q_1^{\text{opt}} \geq q_2^{\text{opt}}$$

and we conclude the proof. □

We now present an On/Off threshold result.

Proposition 28. *Let $b_1 \leq b_2 \leq \ldots \leq b_t$. We assume that $r = 1$, $w_{1j} \overset{i.i.d.}{\sim} \mathcal{N}_{\mathbb{C}}(1)$, $\forall 1 \leq j \leq t$. Then, for all $j = 1, \ldots, t$, there exists $\bar{b}(b_j) \geq 0$ such that*

$$\text{if } b_{j+1} > \bar{b}(b_j) \text{ then } q_i^{\text{opt}} = 0, \ \forall i = 1, \ldots, j.$$

Comments: We will see that one can take $\bar{b}(b_j) = \sqrt{\frac{\bar{a}(Pb_j^2)}{P}}$, where $\bar{a}$ is given by the reciprocal of the function $\frac{1}{F} - 1$, with $F(a) = \mathbb{E}\frac{1}{1+aX}$, which is also known as the Ei or exponential function. The previous result says the following, if there is a value b_{j+1} such that Pb_{j+1}^2 is bigger than $\bar{a}(Pb_j^2)$, we then know that the optimal q_i^{opt} are zero for $i = 1, \ldots, j$. In other words, if some of these "gains" (b_i's) are two small compared to some others, we switch off the corresponding antennas.

Proof. In this setting we have

$$\psi(q) = \mathbb{E}\log(1 + P\sum_{i=1}^{t} q_i d_i X_i),$$

with $b_i^2 = d_i$, $X_i \overset{i.i.d.}{\sim} \mathcal{E}(1)$, $\forall 1 \leq i \leq t$ and $q \in \Theta(t) = \{x \in \mathbb{R}_+^t | \sum_{i=1}^t x_i = 1\}$. Let

$Z_j = 1 + P\sum_{i\neq j,j+1} q_i X_i$. We then have

$$\partial_{q_j}\psi(q) = \mathbb{E}\frac{Pd_jX_j}{Z_j + Pd_jX_j + Pd_{j+1}X_{j+1}}.$$

Let $0 < T \leq 1$ and $p^{(j)}$ be a vector with $p^{(j)}_{j+1} = T$, $p^{(j)}_j = 0$ and thus $\sum_{i\neq j,j+1} p^{(j)}_i = 1 - T$. From the concavity of ψ, if

$$\partial_{q_j}\psi(q)|_{q=p^{(j)}} < \partial_{q_{j+1}}\psi(q)|_{q=p^{(j)}}, \ \ \forall 0 < T \leq 1, \tag{7.5}$$

then $q_i^{\text{opt}} = 0$, $\forall i = 1, \ldots, j$. Now, (7.5) becomes

$$\mathbb{E}\frac{Z + TPd_jX_j}{Z + TPd_{j+1}X_{j+1}} < 1, \ \ \forall 0 < T \leq 1,$$

so if for all $z \geq 1$ and $0 < T \leq 1$ we have

$$\mathbb{E}\frac{z + TPd_jX_j}{z + TPd_{j+1}X_{j+1}} = \mathbb{E}\frac{z/T + Pd_jX_j}{z/T + Pd_{j+1}X_{j+1}} < 1,$$

we are done. Last inequality is equivalent to

$$\mathbb{E}\frac{1}{z + Pd_{j+1}X_{j+1}} < \frac{1}{z + Pd_j}, \ \ \forall z \geq 1.$$

Let $F(a) = \mathbb{E}\frac{1}{1+aX}$, $a_{j+1} = Pd_{j+1}$ and $a_j = Pd_j$, we now wonder when

$$F(a_{j+1}/z) < \frac{1}{1 + a_j/z}, \ \ \forall z \geq 1.$$

For a given $\beta \in \mathbb{R}_+$, let $\alpha(\beta)$ be the smallest number satisfying $F(\alpha(\beta)) \leq \frac{1}{1+\beta}$. Then if for any possible a_j, $\bar{a}(a_j) = \sup_{z\geq 1} z\alpha(a_j/z) < +\infty$, we deduce that for $a_j > \bar{a}(a_j)$, we satisfy $F(a_{j+1}/z) < \frac{1}{1+a_j/z}$, $\forall z \geq 1$. Let us show that F is a convex function, in fact

$$\frac{d}{da}F(a) = \frac{\mathbb{E}(\frac{X}{(1+aX)^2})}{\mathbb{E}(\frac{1}{1+aX})^2},$$

so we need to verify that

$$A \mapsto \frac{\mathbb{E}\frac{X}{(A+X)^2}}{\mathbb{E}\frac{1}{(A+X)^2}}$$

is increasing, i.e. by derivation, we need to show that

$$\mathbb{E}\frac{1}{(A+X)^3}\mathbb{E}\frac{X}{(A+X)^2} \geq \mathbb{E}\frac{X}{(A+X)^3}\mathbb{E}\frac{1}{(A+X)^2}.$$

But, by defining $d\nu(x) \propto \frac{e^{-x}}{(A+x)^2}$, last inequality becomes

$$\mathbb{E}^\nu \frac{1}{A+X}\mathbb{E}^\nu X \geq \mathbb{E}^\nu \frac{X}{A+X}$$

or equivalently

$$\mathbb{E}^\nu X\mathbb{E}^\nu \frac{1}{A+X} + A\mathbb{E}^\nu \frac{1}{A+X} = \mathbb{E}^\nu (A+X)\mathbb{E}^\nu \frac{1}{A+X} \geq 1,$$

which is indeed satisfied by Jensen's inequality. Thus we get that α is convex and one can also check that it is a continuous increasing function with $\alpha(0) = 0$. Therefore

$$\alpha(a_j/z) = \alpha(a_j/z + 0(1-1/z)) \leq \alpha(a_j)/z + 0$$

and thus

$$z\alpha(a_j/z) \leq \alpha(a_j), \ \ \forall z \geq 1$$

which implies that $\bar{a}(a_j) = \sup_{z\geq 1} z\alpha(a_j/z) = \alpha(a_j)$. And we conclude by setting $\bar{b}(b_j) = \sqrt{\frac{\bar{a}(Pb_j^2)}{P}}$.

□

The function $\bar{b}^2(\cdot)$ is continuous convex and increasing with $\bar{b}(0) = 0$, a derivative of 1 at 0 and of 0 at infinity.
In the following figure, the inverse of the function $\bar{b}^2$ is plotted.

Let us now look at some numerical examples, we assume that we have $r = 6$ receiving antennas and $P = 1$. If for example B is such that its singular values

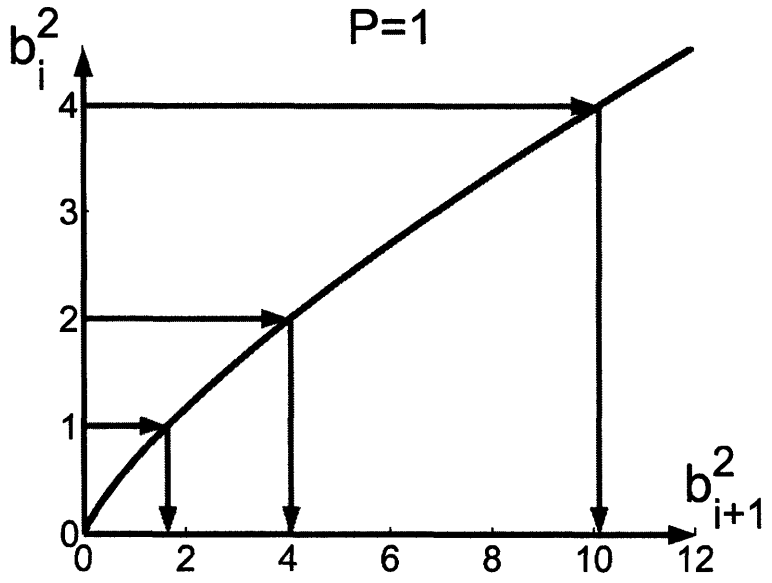

Figure 7-1: Inverse of $\bar{b}^2$, $P = 1$.

have squared given by $(b_1^2, \ldots, b_6^2) = (1, 1, 2, 3, 4, 11)$, it can be seen from figure 7-1 that b_6 exceeds the On/Off threshold of b_5 and thus the optimal power allocution is $(q_1^{\max}, \ldots, q_6^{\max}) = (0, \ldots, 0, 1)$, i.e. in this case we solved the problem. If we had $(b_1^2, \ldots, b_6^2) = (1, 1, 2, 5, 6, 8)$, then the previous situation does not hold anymore, but b_4 exceeds the On/Off threshold of b_3 and we are reduced to a half-dimensional optimization problem for the values of b_4, b_5 and b_6.

We now compare the on/off threshold for the new water-filling with respect to the on/off threshold for the usual water-filling. We consider $t = r = 2$, and denote the singular values of B by $b_1 \leq b_2$, we distinguish the deterministic fading case $H_{\mathrm{d}} = B$ and the random fading case $H_{\mathrm{r}} = WB$, where W has i.i.d. Gaussian entries. For a given power P, and a given value $b_2 \geq 0$, the on/off threshold T is defined as the maximal value such that for $b_1 \leq T$, the power allocation of the optimal covariance matrix eigenvalues is $(q_1^*, q_2^*) = (0, P)$. Denoting by T_{d} and T_{r} the respective thresholds for each case, we have

$$T_{\mathrm{d}} = (1/b_2 + P)^{-1}$$

and

$$T_{\mathrm{r}} = (2z^{-1} - 1)/P$$

where

$$z = 1 + \mathbb{E}(1 + Pb_2X)^{-1}$$

and

$$X \sim \chi^2_4.$$

i.e. X is a sum of two independent exponential random variables with mean 1. As it is shown in the following plot, the random fading threshold is bounded by the deterministic one, which is consistent with the idea that the random mixing of B with W (in the expression $H = WB$) smoothen the optimal power allocation as well.

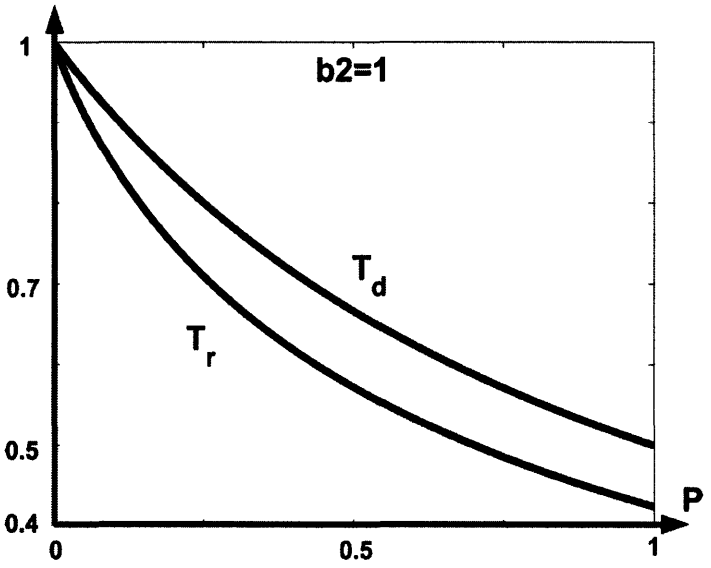

Figure 7-2: Comparison of T_{d} and T_{r}: water and martini thresholds

We now know how to deal with some cases in which the values of the b_i's are quite different. In what follows, we investigate a case in which the b_i's are close to each other (using first order taylor approximations).

Proposition 29. *We assume that* $r = 1$, $w_{1j} \overset{i.i.d.}{\sim} \mathcal{N}_{\mathbb{C}}(1)$, $\forall 1 \leq j \leq t$. *Let* d *be*

a vector of dimension $t \geq 0$ *with* $d_i = b_i^2$ *and* M *a matrix of dimension* $t \times t$ *with components*

$$M_{ij} = \mathbb{E}\frac{X_i X_j}{\frac{1}{P} + \frac{\gamma}{t}\sum_{k=1}^{t} X_k}.$$

We then have $q = \hat{q} + o(||d - \gamma 1^t||)$ *where* $\gamma = \sum_{i=1}^{t} d_i/t$ *and*

$$\hat{q} = \frac{1}{t}1^t + \frac{1}{t\gamma}(I - \frac{1}{(1^t)^T M^{-1} 1^t} M^{-1} 1^t (1^t)^T)(d - \gamma 1^t).$$

Proof. We recall that we are dealing with the following mutual information function,

$$\psi : (q, d) \in \Theta(t) \times \mathbb{R}^t \mapsto \mathbb{E}\log(1 + P\sum_{i=1}^{t} q_i d_i X_i),$$

and we define its gradient with respect to the first component $\nabla : \Theta(t) \times \mathbb{R}^t \to \mathbb{R}^t$, by

$$\nabla_j : (q, d) \in \Theta(t) \times \mathbb{R}^t \mapsto \mathbb{E}\frac{d_j X_j}{1/P + \sum_{i=1}^{t} q_i d_i X_i}.$$

For a fixed d, we know, from the Khun-Tucker conditions, that at the optimal value q_{opt}, the gradient satisfies

$$\nabla(q_{\text{opt}}, d) = c1^t,$$

for some constant $c \in \mathbb{R}$. We also know from the previous section that for any $\gamma \in \mathbb{R}$,

$$\nabla(\frac{1}{t}1^t, \gamma 1^t) = c'1^t,$$

for some $c' \in \mathbb{R}$. By a Taylor's expansion, we have

$$\begin{aligned}\nabla(q, d) = c'1^t + H_1(q - \frac{1}{t}1^t) + H_2(d - \gamma 1^t) \\ + o(\max(||q - \frac{1}{t}1^t||, ||d - \gamma 1^t||))\end{aligned}$$

where $(H_1)_{ij} = \frac{\partial \nabla_i}{\partial q_j}(\frac{1}{t}1^t, \gamma 1^t)$ and $(H_2)_{ij} = \frac{\partial \nabla_i}{\partial d_j}(\frac{1}{t}1^t, \gamma 1^t)$ Thus, the approximative

solution $\hat{q}$ we are looking for should be in the simplex and should also satisfy:

$$\tilde{c}1^t = H_1(\hat{q} - \frac{1}{t}1^t) + H_2(d - \gamma 1^t)$$

or equivalently

$$\begin{pmatrix} H_1 & 1^t \\ (1^t)^T & 0 \end{pmatrix} \begin{pmatrix} \hat{q} - \frac{1}{t}1^t \\ -\tilde{c} \end{pmatrix} = \begin{pmatrix} v \\ 0 \end{pmatrix}$$

where

$$v = -H_2(d - \gamma 1^t).$$

In order to solve this linear equation, we need to compute the upper left block of the previous matrix inverse, which is given by (assuming H_1 to be invertible)

$$H_1^{-1} - \frac{1}{(1^t)^T 1^t} H_1^{-1} 1^t (1^t)^T H_1^{-1},$$

the solution of the linear equation is then given by

$$\hat{q} = \frac{1}{t}1^t - (H_1^{-1} - \frac{1}{(1^t)^T H_1^{-1} 1^t} H_1^{-1} 1^t (1^t)^T H_1^{-1}) v.$$

Given that the derivative of ψ with respect to the first or second component are similar, one can simplify the previous solution as done in the statement of the proposition. □

Conclusion: The optimal power allocation is not achieved in the same way as for the case of a deterministic fading matrix B, but it preserves the following same properties, the monotonicity and the on/off threshold.

Chapter 8

Non-Ergodic MIMO Channels

8.1 Channel Model and Outage Probability

In this section, we no longer assume that the process generating H is ergodic. We now assume that H is chosen randomly at the beginning of all time and is held fixed for all channel uses. As discussed in [32], the notion of capacity defined in previous chapter can no longer be employed. No matter how small the rate we attempt to communicate is, and no matter how large the blocklength is taken to be, there is a non-zero probability that no codebook and decoding rule allowing arbitrarily small error probability exist. In other words, the previous notion of capacity is zero in this case. On the other hand, one can try to minimize the probability that the channel will not support the rate at which one attempts to communicate, leading to the notion of outage probability as defined in [32].

Definition 37. The *outage probability* $P_{\text{out}}(R)$ is defined by

$$P_{\text{out}}(R, P) = \inf_{Q \in \mathcal{D}} \mathbb{P}\{\Phi(Q, H) < R\} \tag{8.1}$$

where

$$\mathcal{D} = \{Q \in \mathbb{C}^{t \times t} \mid Q \geq 0, \, \text{tr}Q \leq P\}$$

and

$$\Phi(Q, H) = \log\det(I_r + HQH^\dagger).$$

We also define the *outage probability function by* $P(\cdot, R) : Q \mapsto \mathbb{P}\{\Phi(Q, H) < R\}$.

8.2 Symmetries

According to definition 32.3., $P(\cdot, R)$ has the same symmetric properties as the capacity function Ψ for a given fading matrix H distribution, thus $P(\cdot, R)$ is still a $U(t)$-invariant function in our setting. Nevertheless, $P(\cdot, R)$ does not necessarily have a unique minimizer, in particular it is not convex, therefore lemma 12 does not apply here.

The symmetry properties tell us that we can restrict our search of optimizers to diagonal matrices, and that the order of the diagonal entries do not matter. If Q_0 is shown to be a minimizer of $P(\cdot, R)$, i.e. if $P(Q_0, R) \leq P(Q, R)$, $\forall Q \in \mathcal{D}$, then all elements in its orbit through unitary matrices, $\mathcal{U}Q_0 = \{UQ_0V | U, V \in U(t)\}$, will be minimizers.

8.2.1 Invariant Structure and Telatar's Conjecture

From now on, we assume that H has independent C.C.S.G. entries with variance 1. In [32], the following conjecture is stated.

Conjecture: the optimal Q's in (8.1) are given by

$$P\,\mathrm{diag}(\underbrace{\frac{1}{k}, \ldots, \frac{1}{k}}_{k}, 0, \ldots, 0)$$

and all multiplications of it by unitary matrices. The value of k depends on the rate: higher the rate R (i.e., higher the outage probability), smaller the k.

8.2.2 MISO Case and Gaussian Quadratic Forms

The Multi-Input Single-Output case refers to the same channel, but considering the number of receiving antennas to be one, i.e., $r = 1$. In the following, we also assume $P = 1$, which simplifies the expressions we are manipulating, but does not reduce the problem. The conjecture is then stated as follows.

Conjecture 1. *Let* $\xi(t) := \{Q \in \mathbb{C}^{t\times t} \mid Q \geq 0, \mathrm{tr}Q \leq 1\}$ *and* $(H_i)_{1\leq i\leq t} \overset{iid}{\sim} \mathcal{N}_{\mathbb{C}}(1)$. *For all* $x \in \mathbb{R}$, $\exists k \in \{1,\ldots,t\}$ *s.t.*

$$\arg\min_{Q\in\xi(t)} \mathbb{P}\{HQH^* \leq x\} = \mathcal{U}\,\mathrm{diag}(\underbrace{\frac{1}{k},\ldots,\frac{1}{k}}_{k},0,\ldots,0).$$

This conjecture has an interesting geometric interpretation. Say that you are given an random vector which is unitary invariant. You can pick a norm induced from a positive definite matrix, whose trace must be one. Which norm would you pick in order to minimize the probability of observing a short vector? Once this part of the conjecture is proved, the relation between k and R relies on properties of Gamma distributions. From the previous remark, the statement of this theorem is equivalent to the following one.

Conjecture 2. *Let* $\theta(t) := \{x \in \mathbb{R}_+^t \mid \sum_{i=1}^t x_i \leq 1\}$, $X = (X_1,\ldots,X_t)$ *with* $\{X_i\}_{1\leq i\leq t} \overset{iid}{\sim} \mathcal{E}(1)$ *and* $\langle q, X\rangle := \sum_{i=1}^t q_i X_i$. *For all* $x \in \mathbb{R}$, $\exists k \in \{1,\ldots,t\}$ *s.t.*

$$\arg\min_{q\in\theta(t)} \mathbb{P}\{\langle q, X\rangle \leq x\} = \Pi(\underbrace{\frac{1}{k},\ldots,\frac{1}{k}}_{k},0,\ldots,0).$$

We now present three lemmas required to prove the theorem. When a random varible Y admits a density function (which will be the case of all considered random variables), we will denote it by f_Y.

Lemma 15. *Let* $\widetilde{X}$ *be such that* $\widetilde{X} \overset{(d)}{\sim} X_1$ *and* $\widetilde{X}, X$ *are mutually independent. Then*

$\forall x \in \mathbb{R}$, $\forall q \in \mathbb{R}^n$ *and* $\forall k \in \{1, \ldots, t\}$,

$$\frac{\partial \mathbb{P}\{\langle q, X\rangle \leq x\}}{\partial q_k} = -f_{\langle q,X\rangle + q_k \tilde{X}}(x). \tag{8.2}$$

Proof. As $f_{\langle q,X\rangle} \in \mathrm{L}^1(\mathbb{R}) \cup \mathcal{C}^0(\mathbb{R}_+^*)$ (only when $t = 1$ there is a discontinuity at $x = 0$), we can use the Fourier transform to write:

$$f_{\langle q,X\rangle}(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \prod_{j=1}^{t} (1 + \omega i q_j)^{-1} e^{\omega i x} d\omega, \ \forall x \in \mathbb{R}_+^*,$$

therefore

$$\mathbb{P}\{\langle q, X\rangle \leq x\} = \frac{1}{2\pi} \int_{\mathbb{R}} \prod_{j=1}^{t} (1 + \omega i q_j)^{-1} \frac{1}{\omega i} (e^{\omega i x} - 1) d\omega, \ \forall x \in \mathbb{R}_+^* \tag{8.3}$$

and is zero for negative values of x. Thus we get

$$\begin{aligned} \frac{\partial \mathbb{P}\{\langle q, X\rangle \leq x\}}{\partial q_k} &= -\frac{1}{2\pi} \int_{\mathbb{R}} \prod_{j=1}^{t} (1 + \omega i q_j)^{-1} (1 + \omega i q_k)^{-1} e^{\omega i x} d\omega \\ &= -f_{\langle q,X\rangle + q_k \tilde{X}}(x). \end{aligned}$$

□

Lemma 16. *Let Y be a random variable independent of $X_1, X_2 \overset{iid}{\sim} \mathcal{E}(1)$, and let $x, q_1, q_2 \in \mathbb{R}$. We then have*

$$f_{Y+q_1 X_1}(x) - f_{Y+q_2 X_2}(x) = (q_2 - q_1) f'_{Y+q_1 X_1 + q_2 X_2}(x) \tag{8.4}$$

This is easily verified by using Fourier transform.

Lemma 17. *For all $t \geq 2$ and $q \in \theta(t)$, we have $f_{\langle q,X\rangle} \in \mathcal{C}^\infty(\mathbb{R}^*) \cap \mathcal{C}^{t-2}(\mathbb{R})$ and $\exists!$ $a \in \mathbb{R}_+^*$ s.t. $f'_{\langle q,X\rangle}(x) > 0$, $\forall 0 < x < a$, $f'_{\langle q,X\rangle}(a) = 0$ and $f'_{\langle q,X\rangle}(x) < 0$, $\forall x > a$.*

Proof. The fact that $f_{\langle q,X\rangle} \in \mathcal{C}^\infty(\mathbb{R}^*) \cap \mathcal{C}^{t-2}(\mathbb{R})$ can be verified by induction, knowing that the exponential density is in $\mathcal{C}^\infty(\mathbb{R}^*)$ and using properties of convolution and

differentiation.

Note that $\forall t \geq 2$ and for $q \in \theta(t)$ s.t. all q_i's are different, which can be written without loss of generality as $q_1 < q_2 < \cdots < q_t$, we have

$$f_{\langle q,X\rangle}(x) = \sum_{i=1}^{t} \prod_{\substack{j\in\{1,\ldots,t\} \\ \text{s.t.} i\neq j}} \frac{1}{q_i - q_j} {q_i}^{t-1} f_{q_i X_i}(x), \quad \forall x \in \mathbb{R}. \tag{8.5}$$

This can also be verified by induction. Moreover, we have that $\forall x \in \mathbb{R}, \forall t \geq 2$ the function

$$q \in \theta(t) \mapsto \sum_{i=1}^{t} \prod_{\substack{j\in\{1,\ldots,t\} \\ \text{s.t.} i\neq j}} \frac{1}{q_i - q_j} {q_i}^{t-1} f_{q_i X_i}(x) \in \mathbb{R}_+$$

is continuous (with the topology which $\theta(t)$ inherits as a subset of $\mathbb{R}^t$) and (8.5) converges when considering equal q_i's. So we can restrict ourself to prove the lemma for q's having all components different (and we will consider such q's in what follows).

For $t \geq 2$, we have $f^{(k)}_{\langle q,X\rangle}(0) = 0$, $\forall k = 0, \ldots, t-2$ (this is a consequence of the first statement in the lemma). Let us suppose that there exist $a, b > 0$ such that $a \neq b$ and

$$f'_{\langle q,X\rangle}(a) = f'_{\langle q,X\rangle}(b) = 0. \tag{8.6}$$

From (8.5), the assumption (8.6), in addition with $f^{(k)}_{\langle q,X\rangle}(0) = 0$ for $k = 0, \ldots, t-2$, implies that there exist $\alpha_1, \ldots, \alpha_t \in \mathbb{R}$ and $\beta_1, \ldots, \beta_t \in \mathbb{R}$, all different and non-zero, such that

$$\begin{pmatrix} 1 & \cdots & 1 \\ \beta_1 & \cdots & \beta_t \\ \vdots & & \vdots \\ \beta_1^{t-3} & \cdots & \beta_t^{t-3} \\ e^{a\beta_1} & \cdots & e^{a\beta_t} \\ e^{b\beta_1} & \cdots & e^{b\beta_t} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_t \end{pmatrix} = 0.$$

But this is to say that there exists $a_0, \ldots, a_{t-1} \in \mathbb{R}$, non all equal to zero, and $c \in \mathbb{R}_+^*$ s.t.

$$\sum_{i=0}^{t-3} a_i x^i + a_{t-2} e^x + a_{t-1} e^{cx} = 0, \ \forall x \in \{a\beta_1, \ldots, a\beta_t\}. \tag{8.7}$$

Now, if $a_{t-1} = 0$ and $a_{t-2} = 0$, we clearly need $a_0, \ldots, a_{t-3} = 0$ to ensure t solutions in (8.7), which lead to a contradiction. If $a_{t-1} = 0$ or $a_{t-2} = 0$, we are in the following situation

$$e^x = p(x), \tag{8.8}$$

where p is a real polynomial of degree $t-3$. But one can verify by induction (using differentiation and Rolle's theorem) that (8.8) can at most have $t-2$ different solutions. Hence we have a contradiction with (8.7). Otherwise, we have $a_{t-2}, a_{t-1} \neq 0$ and we are in the following situation

$$e^x + de^{cx} = \widetilde{p}(x), \tag{8.9}$$

where $d \in \mathbb{R}^*$ and $\widetilde{p}$ is a real polynomial of degree t. With the same argument as before, one can show that (8.9) has at most $t-1$ different solutions, hence we also have a contradiction and we cannot have $a \neq b$. The existence of a, as well as the sign of the derivatives around a are clearly justified. This concludes the proof of the lemma. □

Proof of conjecture 2 for $t \leq 100$:

Let $x \in \mathbb{R}$. From lemma 15, for any $0 \leq k \leq t$,

$$\frac{\partial \mathbb{P}\{\langle q, X \rangle \leq x\}}{\partial q_k} = -f_{\langle q,X \rangle + q_k \widetilde{X}_1}(x) \leq 0,$$

with $\widetilde{X}_1$ independent of X and $\widetilde{X}_1 \sim X_1$. We thus conclude that we can replace $\theta(t)$ by $\Theta(t) := \{q \in \mathbb{R}_+^t | \sum_{i=1}^t q_i = 1\}$.

Using the Kuhn-Tucker theorem, if $q^* \in \Theta(t)$ minimizes $\mathbb{P}\{\langle q, X\rangle \leq x\}$, then $\exists \lambda \in \mathbb{R}$ s.t.

$$\frac{\partial \mathbb{P}\{\langle q^*, X\rangle \leq x\}}{\partial q_k} \begin{cases} = \lambda, & \forall k \text{ s.t. } q_k^* > 0, \\ \geq \lambda, & \text{otherwise.} \end{cases} \tag{8.10}$$

By lemma 15 and 16,

$$\frac{\partial \mathbb{P}\{\langle q, X\rangle \leq x\}}{\partial q_k} = \frac{\partial \mathbb{P}\{\langle q, X\rangle \leq x\}}{\partial q_l}$$

is equivalent to

$$(q_k - q_l) f'_{\langle q, X\rangle + q_k \widetilde{X}_1 + q_l \widetilde{X}_2}(x) = 0$$

with $\widetilde{X}_1, \widetilde{X}_2, X$ mutually independent and $\widetilde{X}_2 \sim X_1$. Now, let us assume that $0 < q_1^* < q_2^*$ (this represent w.l.o.g. that at least two different non-zero values are in q^*). Then

$$f'_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_2^* \widetilde{X}_2}(x) = 0 \tag{8.11}$$

and using (8.4) (with $Y = \langle q^*, X\rangle + q_1^* \widetilde{X}_1$, $q_1 = 0$ and $q_2 = q_2^*$), we get

$$f_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_2^* \widetilde{X}_2}(x) = f_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1}(x). \tag{8.12}$$

We now assume that q_3^*, the third component of q^*, is non-zero. By successive use of (8.4) and by (8.12), we have

$$\begin{aligned} & f'_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_3^* \widetilde{X}_3}(x) \\ = \; & \frac{1}{q_3^*}\left(f_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1}(x) - f_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_3^* \widetilde{X}_3}(x)\right) \\ = \; & \frac{1}{q_3^*}\left(f_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_2^* \widetilde{X}_2}(x) - f_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_3^* \widetilde{X}_3}(x)\right) \\ = \; & \frac{q_3^* - q_2^*}{q_3^*} f'_{\langle q^*, X\rangle + q_1^* \widetilde{X}_1 + q_2^* \widetilde{X}_2 + q_3^* \widetilde{X}_3}(x) \end{aligned} \tag{8.13}$$

But from lemma 17 and (8.11), $f'_{\langle q^*,X\rangle+q_1^*\tilde{X}_1+q_2^*\tilde{X}_2}$ is strictly positive on $(0,x)$, thus

$$f'_{\langle q^*,X\rangle+q_1^*\tilde{X}_1+q_2^*\tilde{X}_2+q_3^*\tilde{X}_3}(x) = f'_{\langle q^*,X\rangle+q_1^*\tilde{X}_1+q_2^*\tilde{X}_2} \star f_{q_3^*\tilde{X}_3}(x) > 0 \qquad (8.14)$$

Therefore, if q_3^* is not equal to q_1^*, q_2^* or 0, we must have $f'_{\langle q^*,X\rangle+q_1^*\tilde{X}_1+q_3^*\tilde{X}_3}(x) = 0$, in order to satisfy the KT conditions, but this contradicts (8.13) and (8.14).

We have just shown that the KT conditions for minima can be satisfied only with points in $\Theta(n)$ that contains at most two different non-zero values, i.e. a minimizer has the following form

$$q^* = (\underbrace{p_1,\ldots,p_1}_{k},\underbrace{p_2,\ldots,p_2}_{l},0,\ldots,0),$$

with $k,l,k+l \in \{0,\ldots,t\}$, $p_1,p_2 \in [0,1]$, such that $kp_1+lp_2 = 1$. Let us assume that $k \geq 2$. We define $q_\delta^* := q^* + \delta e_1 - \delta e_2$, with $0 < \delta < p_1$ and $e_i \in \mathbb{R}^t$ s.t. $(e_i)_j = \delta_{ij}$. Since q^* is a minimizer, we have

$$f'_{\langle q^*,X\rangle+p_1\tilde{X}_1+p_2\tilde{X}_2}(x) = 0 \qquad (8.15)$$

and using lemmas 15 and 16, we get

$$\frac{\partial^2}{\partial\delta^2}\Big|_{\delta=0}\mathbb{P}\{\langle q_\delta^*,X\rangle \leq x\} = 2f'_{\langle q^*,X\rangle+p_1\tilde{X}_1+p_1\tilde{X}_2}(x). \qquad (8.16)$$

From the expansions in (8.13) and (8.14), if $p_1 < p_2$, we get $\frac{\partial^2}{\partial\delta^2}\Big|_{\delta=0}\mathbb{P}\{\langle q_\delta^*,X\rangle \leq x\} < 0$, and q^* cannot be a minimizer. Thus the minimal component in q^* has to appear only once. Say p_1 appears only once and p_2 appears l times ($1 \leq l \leq t-1$) and is greater than p_1, i.e. $p_2 = \frac{1-p_1}{l} > p_1$, which implies $p_1 < \frac{1}{l+1}$. At $p_1 = \frac{1}{l+1}$, all components of p^* are equal, and the function $p_1 \mapsto \mathbb{P}\{\langle q^*,X\rangle \leq x\}$ has an extremum at that point. Let us assume that there is at most one extremum within $(0,\frac{1}{l+1})$. A simple computation shows that $\frac{\partial}{\partial p_1}\Big|_{p_1=0}\mathbb{P}\{\langle q^*,X\rangle \leq x\} = -\frac{1}{l}f'_{\frac{1}{l}\sum_{i=1}^{l+1}X_i}(x)$, which is

strictly negative if and only if $x < 1$, and $\frac{\partial^2}{\partial p_1^2}\Big|_{p_1=\frac{1}{l+1}} \mathbb{P}\{\langle q^*, X\rangle \leq x\} = c(l)e^{-(l+1)x}(l + 2 - (l+1)x)x^{l+1}$, with $c(l) > 0$, leading to a strictly positive second derivative if and only if $x < \frac{l+2}{l+1}$. Therefore, no minimum can occur when p_1 belongs to $(0, \frac{1}{l+1})$.

So we want to show that $p_1 \mapsto \mathbb{P}\{\langle q^*, X\rangle \leq x\}$ has at most one extremum within $(0, \frac{1}{l+1})$, where $q^* = (p_1, \underbrace{\frac{1-p_1}{l}, \ldots, \frac{1-p_1}{l}}_{l \text{ times}}, 0, \ldots, 0)$. We know that

$$\frac{\partial^2}{\partial p_1^2}\mathbb{P}\{\langle q^*, X\rangle \leq x\} = (p_1 - \frac{1-p_1}{l}) f'_{\langle q^*, X\rangle + p_1\tilde{X}_1 + \frac{1-p_1}{l}\tilde{X}_2}(x).$$

We now use p instead of p_1 and k instead of l. Let us define

$$f_{p,I,J}(x) = f_{p\sum_{i=1}^{I} X_i + \frac{1-p}{k}\sum_{j=1}^{J} Y_i}(x), \quad x \in \mathbb{R},\, p \in (0,1),\, k, I, J \in \mathbb{Z}_+,$$

where $\{X_i\}_{1\leq i\leq I}, \{Y_j\}_{1\leq k\leq J} \overset{iid}{\sim} \mathcal{E}(1)$. We want to show a contradiction between the following assumptions:

$$k \geq 1 \tag{8.17}$$

$$p, q \in (0, \frac{1}{k+1}),\, p \neq q \tag{8.18}$$

$$f'_{p,2,k+1}(x) = f'_{q,2,k+1}(x) = 0,\, x \in \mathbb{R}_+ \tag{8.19}$$

Since we are now working with simpler combination of our exponential random variables, we can express above objects in the time domain with less complications. We get

$$\begin{aligned} f'_{p,2,k+1}(x) \;=\; & (k+1)\left(\frac{1}{p}\right)^2 \left(\frac{1}{\bar{p}}\right)^{k+1} e^{-x/\bar{p}} \left(\frac{-1}{\Delta}\right)^{k+1} \\ & \left[\left(\frac{\Delta p - \Delta x}{k+1} - 1\right)\left(e^{-\Delta x} - \sum_{l=0}^{k}\frac{(-\Delta x)^l}{l!}\right) + \frac{(-\Delta x)^{k+1}}{(k+1)!}\right] \end{aligned} \tag{8.20}$$

where

$$\Delta = \frac{1}{p} - \frac{1}{\bar{p}}, \quad \bar{p} = \frac{1-p}{k}.$$

Hence, $f'_{p,2,k+1}(x) = 0$ is equivalent to

$$\left(\frac{\Delta p - \Delta x}{k+1} - 1\right)\left(e^{-\Delta x} - \sum_{l=0}^{k} \frac{(-\Delta x)^l}{l!}\right) + \frac{(-\Delta x)^{k+1}}{(k+1)!} = 0 \tag{8.21}$$

Let $y = \Delta x$, $n = k+1$, $T_l(y) = \frac{(-y)^l}{l!}$, $S_{n-1}(y) = \sum_{l=0}^{n-1} T_l(y)$ and $f(\Delta) = \frac{\Delta p}{n} - 1$, then (8.21) is equivalent to

$$(e^{-y} - S_{n-1}(y))(f(\Delta) - \frac{y}{n}) + T_n(y) = 0. \tag{8.22}$$

By definition

$$\Delta = 1/p - (n-1)/(1-p),$$

and we can express p in terms of Δ as

$$p = \frac{\Delta + n - \sqrt{(\Delta+n)^2 - 4\Delta}}{2\Delta},$$

implying

$$f(\Delta) = \frac{\Delta + n - \sqrt{(\Delta+n)^2 - 4\Delta}}{2n} - 1.$$

This implies that $f(\Delta) = t$ is equivalent to

$$\Delta = nt(1 + \frac{1}{n(1+t) - 1}).$$

Let $y(\Delta)$ be the solution of 8.22. We want to show that the following cannot happen: for some $x \geq 0$ and $n \geq 2$ (where we think of x has being the slope of a linear function of Δ), there exists $\Delta_1 \neq \Delta_2$ with $\Delta_1, \Delta_2 > 0$, such that $y(\Delta_1) = x\Delta_1$ and $y(\Delta_2) = x\Delta_2$.

In order to show this, it is sufficient to check that

$$\forall n \geq 2, \quad \frac{f_n(t_n(y))}{y} \text{ is increasing in } y,$$

where

$$y \overset{t_n}{\mapsto} y/n - \frac{T_n(y)}{e^{-y} - S_{n-1}(y)} \tag{8.23}$$

$$t \overset{f_n}{\mapsto} nt(1 + \frac{1}{n(1+t) - 1}). \tag{8.24}$$

One can show that

$$\frac{f_n(t_n(y))}{y} = \frac{R_{n-2}}{R_{n-1}} \frac{nR_{n-1} + yR_{n-2}}{(n-1)R_{n-1} + yR_{n-2}},$$

where

$$R_n = R_n(y) = e^{-y} - S_n(y).$$

Therefore, by defining $x = -y$, $Q_n = n - x\frac{R_{n-2}}{R_{n-1}}(x)$, we want to show that

$$\forall n \geq 2, \quad \frac{Q_n(x)}{Q_{n-1}(x)} \text{ is increasing in } x,$$

or equivalently

$$\forall x < 0, \quad \frac{Q_n'(x)}{Q_n(x)} \text{ is decreasing in } n. \tag{8.25}$$

(Recall that increasing/decreasing refers to strictly increasing/decreasing). Let $\rho_n = \rho_n(x) = \frac{R_{n-2}}{R_{n-1}}(x)$, such that $Q_n = n - x\rho_n$. Note that

$$R_n' = R_{n-1} \tag{8.26}$$

$$R_n + T_n = R_{n-1} \tag{8.27}$$

$$T_n = \frac{x}{n} T_{n-1}. \tag{8.28}$$

Note: $R_0 < 0$ and $R_1' = R_0 < 0$ hence $R_0(0) - R_1(x) < 0$, $\forall x < 0$, implying $R_1 > 0$. By induction

$$(-1)^{n+1} R_n \geq 0,$$

implying

$$0 < \frac{R_n}{T_{n+1}} < 1.$$

We have

$$\begin{aligned} Q_n^{'} &= -\rho_n - x\rho_n^{'}, \\ \rho_n^{'} &= \frac{R_{n-2}^{'}R_{n-1} - R_{n-2}R_{n-1}^{'}}{R_{n-1}^2} \\ x\frac{R_{n-3}}{R_{n-1}} &= x\frac{R_{n-2}}{R_{n-1}} + x\frac{T_{n-2}}{R_{n-1}} \\ x\frac{T_{n-2}}{R_{n-1}} &= (n-1)(\frac{R_{n-2}}{R_{n-1}} - 1) \end{aligned}$$

which implies

$$Q_n^{'} = x\rho^2 - (x+n)\rho + (n-1),$$

and

$$\frac{Q_n^{'}(x)}{Q_n(x)} = 1 - \rho_n - \frac{1}{n - x\rho_n}. \tag{8.29}$$

Claim:

$$\forall x < 0, n \geq 2, \quad \rho_{n+1} < \rho_n \quad \Longleftrightarrow \quad n - x\rho_n > 0, \tag{8.30}$$

in fact: if n is even, $R_{n-1} > 0$, $R_{n-2} < 0$ and

$$\begin{aligned} n - x\rho_n > 0 \quad &\Longleftrightarrow \quad 1 > \frac{xR_{n-2}}{nR_{n-1}} && (8.31) \\ &\Longleftrightarrow \quad xR_{n-2} < nR_{n-1} && (8.32) \\ &\Longleftrightarrow \quad xR_{n-1} < nR_n && (8.33) \\ &\Longleftrightarrow \quad 1 > \frac{nR_n}{xR_{n-1}} && (8.34) \end{aligned}$$

implying

$$1 > \frac{xR_{n-2}}{nR_{n-1}}\frac{nR_n}{xR_{n-1}} = \frac{\rho_n}{\rho_{n+1}}.$$

If n is odd, (8.31) still holds and the next two inequalities are then inverted, but (8.34) holds again, getting to the same conclusion as that for n even.
To prove the other implication, we need the identity

$$\rho_{n+1} = \frac{n}{n + x - x\rho_n},$$

which is proved below. With it, we have that $1 > \frac{\rho_n}{\rho_{n+1}} = \frac{(n+x-x\rho_n)\rho_n}{n}$, or equivalently

$$-x\rho_n^2 + (n + x)\rho_n - n < 0.$$

But the roots of above polynomials in ρ_n are n/x and 1, and since ρ_n is negative, we get $\rho_n > n/x$, which proves the other implication.

Claim:

$$\rho_{n+1} = \frac{n}{n + x - x\rho_n}. \tag{8.35}$$

In fact:

$$\rho_n = 1 + \frac{T_{n-1}}{R_{n-1}} = 1 + \frac{n}{x}\frac{T_n}{R_{n-1}} = 1 + \frac{n}{x}\frac{T_n}{R_n + T_n},$$

which implies

$$\frac{R_n}{T_n} = \frac{n}{x}(\rho_n - 1)^{-1} - 1$$

hence

$$\rho_{n+1} = 1 + \frac{T_n}{R_n} = 1 + (\frac{n}{x}(\rho_n - 1)^{-1} - 1)^{-1},$$

which proves the claim.

Using the properties of confluent hypergeometric functions from [24], we have that (8.30) holds.

From (8.30) and (8.29), we can then express (8.25) as

$$(\rho_{n+1} - \rho_n)[(n - x\rho_n)(n + 1 - x\rho_{n+1}) + x] - 1 > 0, \quad \forall n \geq 2, x < 0. \tag{8.36}$$

Using (8.35), this is equivalent to

$$x^3r^4 - (x+3n)x^2r^3 + x[n(3(x+n)+1)-1]r^2 \tag{8.37}$$

$$-n[2nx + (x+n)(n+1)]r + n^2(n+1) < 0, \tag{8.38}$$

where $r = \rho_{n+1}$.

Since the above polynomial is of degree 4 in r, we can use Ferrari's solution to express its roots. Moreover, the coefficients signs are such that if $c_k > 0$, then $c_{k+1} > 0$, which implies from Descartes rule of signs, that there is at most one positive root and 0, 1 or 3 negative roots. Using symbolic computations in Maple, the roots can be computed in terms of x and n, confirming the number of 3 complex roots and one negative root for any values of n and x (which is a consequence of the general structure of the coefficients), and there is an analytic expression of the negative root, whose size exceeds the size of this page. The polynomial is then negative for values that are below this negative root. We can use Maple to check that for arbitrary n, we have $\rho_{n+1}(x) > r_0(x,n)$. An alternative way is the following. Since $\rho_{n+1} = 1 + \frac{T_n}{R_n}$, we can equivalently check that $\frac{T_n}{R_n}$ is in the negative region of a degree 4 polynomial. We know how to bound S_n and R_n arbitrarily close above and below by taking an arbitrarily large number of summands in its expression. In particular,

$$\frac{T_n}{R_n} = \frac{T_n}{e^x - S_n(x)} = \frac{T_n}{e^x(1 - e^{-x}S_n(x))}$$

Hence, if we take the example where n is even, we find that for any K which is odd

$$1 - e^{-x}S_n(x) < 1 - S_K(-x)S_n(x).$$

We then can write a Maple symbolic code that finds K large enough to lower bound the original expression by a function which is a weighted sum of exponentials, each weight being strictly negative.

Finally, [24] gives an approximation to the function $R_n(x)$ for x negative; we hope to avoid Maple symbolic computation using this paper.

8.3 Generalizations

Given a sequence of iid random variables, how do we construct a weighted sum of them (with a sum constraint) in order to minimize the probability of exceeding a given threshold. If in some examples the exponential distribution arises as the natural distribution (*e.g.* amplitude, waiting times), we may be interested in solving the problem for other distributions too (Gaussian in particular). In general, one can say several things regarding the moments of this sum, but it seems difficult to solve the minimization problem we expressed here in a general context. One may wonder for what kind of other distributions does the conjecture still hold? The proof we provided ($t \leq 100$) is very dependent on the exponential distribution and it is hard to think of a possible generalization. One can look at other examples. In the case of the Cauchy distribution, the function $q \mapsto \mathbb{P}\{\langle q, X\rangle \leq x\}$ is constant. In the case of the Gaussian distribution, the conjecture holds and k can easily be determined.

Proposition 30. *Let $n \in \mathbb{N}^*$, $q \in \Theta(n)$, $X = (X_1, \ldots, X_n)$ with $(X_i)_{1\leq i\leq n} \overset{iid}{\sim} \mathcal{N}(0,1)$. For $x > 0$,*

$$\arg\min_{q\in\theta(n)} \mathbb{P}\{\langle q, X\rangle \leq x\} \ni (1, 0, \ldots, 0),$$

for $x < 0$,

$$\arg\min_{q\in\theta(n)} \mathbb{P}\{\langle q, X\rangle \leq x\} = (1/n, \ldots, 1/n)$$

and for $x = 0$, $\mathbb{P}\{\langle q, X\rangle \leq x\} = 1/2$, $\forall q \in \theta(n)$.

The first part of the proof, which is covered by Lemmas 17, 15 and 16 could possibly be generalized for other distributions. By imposing conditions on the derivative of the Fourier transform of the sum (or of X_1), such that the KT conditions would only be satisfied under some symmetry of the q_i's. Infinite divisible laws may be a good point to start, having at hand, the Levy-Khinchine formula.

But without any conditions on the random variables $\{X_i\}$'s, except independent and identically distributed, the conjecture does not hold, i.e., a statement such as in theorem 2 is not universal. In fact, for $n = 2$, $X_1, X_2 \overset{iid}{\sim} \mathcal{E}(1)$, $x = 1.1$, the input that maximizes this "outage probability" is not of the form $(1, 0)$, $(0, 1)$ or $(1/2, 1/2)$ (it

is around (0.2,0.8)). Thus by choosing $Y_1, Y_2 \overset{iid}{\sim} -X_1$ and $y = -1.1$ we get a counter-example to the conjecture if stated for any independent real random variables. If stated for positive random variables, one can consider $Z_1, Z_2 \overset{iid}{\sim} L - X_1 \mathbb{1}_{[0,L]}$ for a large enough $L \in \mathbb{R}_+$, $z = 2L - 1.1$ and we get a counter-example for positive random variables.

However, all these counter-examples are not true generalization of the initial conjecture, in fact, the conjecture as stated in 2 is the reduction of theorem 1, so that $\langle q, X \rangle$ is a reduction of $\langle H, H \rangle_{\text{diag}(q)}$ in the case when H is unitary invariant and has iid entries, which implies it is iid CSCG distribution and hence X is iid exponential. Therefore, a generalization of 2 where $\langle H, H \rangle_{\text{diag}(q)}$ is considered for any unitary invariant vector H, may still hold.

Chapter 9

Conclusion

Linear Universal Decoding

We have raised the question whether it would be possible to have linearity and universality embodied by a single decoder. When a universal decoder is required to be capacity achieving, we showed that a generalized linear universal decoding rule for compound sets having a finite union of one-sided components exists. We defined it as follows: if $W_1, \ldots, W_K$ are the worst channels of each component, use the generalized linear decoding rule induced by the MAP metrics $\log \frac{W_1}{(\mu_1)_Y}, \ldots, \log \frac{W_K}{(\mu_K)_Y}$, i.e., decode with

$$\hat{x}(y) = \arg \max_{x_m, m=1,\ldots,M} \vee_{k=1}^{K} \mathbb{E}_{P_{x_m,y}} \log \frac{W_k}{(\mu_k)_Y},$$

where $\mu_k = P_X \circ W_k$ and the input distribution is the optimal input distribution on S. We denoted this decoding rule by GWAP$(W_1, \ldots, W_K)$. We also found that using the ML metrics, instead of the MAP metrics $W_1, \ldots, W_K$, i.e. GLRT$(W_1, \ldots, W_K)$, is not universal.

We saw that MMI is equivalent to GWAP decoding when all the DMC's MAP metrics are taken as the worst channels, i.e. MMI = GWAP(DMC). Therefore, our result tells us that we do not need to take all DMC metrics to achieve capacity, for a given compound set S, we can restrict ourself to selecting carefully a finite number of metrics and yet achieve the compound capacity. Those important metrics are found by extracting the one-sided components of S, and taking the MAP metrics induced

by the worst channel of these components. When S has a finite number of one-sided components, the GWAP decoding rule is generalized linear. This allows us to better understand the representation of the metrics space, with the equivalence relation of rate achievability. Further works may investigate the notion of universality that requires optimality in the exponent, as opposed to optimality in the achievable rates. Another problem, briefly introduced here, consists of using a decoding rule induced by a fixed number of metrics chosen without the knowledge of the compound set (most likely in a uniform manner); the relationship between the number of metrics and the performance would then be analyzed.

Gaussian Noise and Interference

For an additive white Gaussian noise channel, and for Gaussian inputs, we defined an operator that measures how much variation, a given perturbation of the input, induces in the output entropy. If g_s is a perturbation in the direction L, by an amount $g(1+\varepsilon L)$, then the "non-Gaussianess" of the perturbation is approximately $\frac{1}{2}\|L\|_{g_s}^2$ and the non Gaussianess of the output distribution is $\frac{1}{2}\|T(L)\|_{L_2(\mathbb{R})}^2$ where $T : L \rightsquigarrow \frac{\sqrt{g_s}L \star g_v}{\sqrt{g_{s+v}}}$. We found that the eigenfunctions of the T^tT operator are the Hermite polynomials $\bar{H}_k^{[s]}$ in $L_2(g_s;\mathbb{R})$ (multiplied by $\sqrt{g_s}$), and the eigenvalues are $\left(\frac{s}{s+v}\right)^k$. In addition, the eigenfunctions of this operator in $L_2(g_s;\mathbb{R})$ maps naturally to the eigenfunction in $L_2(g_{s+v};\mathbb{R})$, since $\frac{g_s \bar{H}_k^{[s]} \star g_v}{g_{s+v}} = (\frac{s}{s+v})^{k/2}\bar{H}_k^{[s+v]}$. This structure allows us to better understand the relationship between the interference coefficient a and the optimal input distributions in a Gaussian interference channel. In particular, with this structure we could show the optimal input distribution for the single letter sum-capacity undergoes two regime, where for if $a < 0.68$ (root of a degree twelve polynomial) the Gaussian distribution is a local maxima and elsewhere the Gaussian distribution is not a local maxima. This result can be generalized to the multi-letter case, hence, interference should not be treated as noise above the given threshold. The Hermite transformation introduced in this problem is a promising tool for approaching several multi-user information theory problems having Gaussian noise.

MIMO Channels

For ergodic coherent MIMO channels, we showed how symmetries (quantified by a group $G \subseteq U$) in the fading matrix distribution and input constraint set are transformed into symmetries in the optimal input distribution, specifying its structure (it has to belong to $\mathrm{Comm}(G)$). When the fading matrix is deterministic, the so-called water-filling power allocation is optimal. We show that when the fading matrix is a deterministic unitary matrix multiplied by random unitary matrices (the Kronecker model), the new power allocation is no longer the water-filling, but we characterized a martini-filling optimal power allocation which preserves, although smoothens, the water-filling characteristics. Finally, we saw that in a non-ergodic setting, although the symmetric properties of the outage probability are the same as for the mutual information, the symmetric structures of the minimizer are much more complex to analyze (since the outage probability is not convex). We could verify Telatar's conjecture in the MISO case for an input dimension t less than one hundred, where the value one hundred is symbolic and expresses the fact that as long as the dimension is given to us, we could conclude the last step of the proof, which requires to check that a certain confluent hypergeometric function is increasing. We do not have a general argument to conclude the last step for generic values of t, due to the complexity of the resulting expressions to be manipulated. This problem is equivalent to finding the positive definite matrices, with fixed trace, minimizing the probability that a vector's norm (using the norm induced by matrix) exceeds a given threshold. Expecting symmetric structures in the solutions when "rich" symmetric structures are present in the fading distribution, such as i.i.d. Gaussian, has been claimed in many ways, but no neat geometric arguments have been found to date.

Local to Global Geometric Method

Although the divergence may not be a squared distance, it behaves locally as such. Moreover, globally, it still preserves certain properties satisfied by squared distances (cf. [8], section 2.2). Therefore the localization provided an accurate and insightful

reduction of our problems. In chapter 4 and 6, we develop techniques to transform a global problem into a local one, through the VN and Hermite transformations. In both cases, global divergence expressions reduced to expressions defined in an inner product space, which we characterized. This brought geometrical insight to the problem. Additionally, in certain cases we have been able to "lift" results proven locally to results that we could prove globally. This technique has been used in chapter 5 to solve the problem stated for linear universal decoding. In chapter 6, the same technique has been used to find the eigenfunction structure described in the preceding section, which is a very promising structure to better understand a collection of multi-user problems. We believe that the local to global methods can be successfully employed on a large variety of information theory and related probability and statistics problems.

Bibliography

[1] A. De Acosta. Moderate deviations and associated Laplace approximations for sums of independent random vectors. *Trans. Amer. Math. Soc.*, 329(12):357–374, 1992.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry.* American Mathematical Society, 2001.

[3] V.S. Annapureddy and V.V. Veeravalli. Gaussian interference network: sum capacity in the low interference regime and new outer bounds on the capacity region. *Submitted IEEE Trans. Inform. Theory*, 2008.

[4] V.B. Balakirsky. A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels. *IEEE Trans. Inform. Theory*, 41(6):1889–1902, November 1995.

[5] D. Blackwell, L. Breiman, and A. Thomasian. The capacity of a class of channels. *The Annals of Mathematical Statistics*, 30(4):1229–1241, December 1959.

[6] H. F. Chong, M. Motani, H. K. Garg, and H. El Gamal. On the Han-Kobayashi region for the interference channel. *submitted to IEEE Trans. Inform. Theory*, August 2006.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications, 1991.

[8] I. Csiszar. Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4), 2004.

[9] I. Csiszar and J. Korner. *Information Theory: Coding Theorem for Discrete Memoryless Systems.* Akademiai Kiado, Budapest, 1986.

[10] I. Csiszar and J. Korner. Graph decomposition: A new key to coding theorems. *IEEE Trans. Inform. Theory*, 27(1):5–12, January 1981.

[11] I. Csiszar and P. Narayan. Channel capacity for a given decoding metric. *IEEE Trans. Inform. Theory*, 41(1):35–43, January 1995.

[12] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications.* Springer-Verlag, 1998.

[13] R. Etkin, D. Tse, and H. Wang. Gaussian interference channel capacity to within one bit: the symmetric case. *arXiv:cs.IT/0702045v2*, February 2007.

[14] M. Feder and A. Lapidoth. Universal decoding for channels with memory. *IEEE Trans. Inform. Theory*, 44(9):17261745, September 1998.

[15] G. B. Folland. *Harmonic Analysis in Phase Space.* Princeton University Press, 1989.

[16] Robert G. Gallager. *Information Theory and Reliable Communication.* John Wiley and Sons, 1968.

[17] Robert G. Gallager. *Principles of Digital Communication.* Cambridge University Press, 2008.

[18] T. Han and K. Kobayashi. A new achievable rate region for the interference channel. *IEEE Trans. Inform. Theory*, 27(1):49–60, January 1981.

[19] A. Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, 44(10):2148–2177, October 1998.

[20] A. Lapidoth, E. Telatar, and R. Urbanke. On wide-band broadcast channels. *IEEE Trans. Inform. Theory*, 49(12):3250–3258, December 2003.

[21] A. Lapidoth and J. Ziv. On the universality of the LZ-based decoding algorithm. *IEEE Trans. Inform. Theory*, 44(9):17461755, September 1998.

[22] V. Marchenko and L. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Math. USSR-Sbornik*, 72(4):507–536, 1967.

[23] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai. On information rates for mismatched decoders. *IEEE Trans. Inform. Theory*, 40(6):1953–1967, November 1994.

[24] M. Merkle. Inequalities for residuals of power expansions for the exponential function and completely monotone functions. *Journal of Math. Analysis and Appl.*, 212:126–134, 1997.

[25] A. Montanari and G. D. Forney. On exponential error bounds for random codes on the DMC. *manuscript*, 2001.

[26] M. K. Murray and J. W. Rice. *Differential Geometry and Statistics.* Chapman and Hall, 1993.

[27] H. Sato. The capacity of the Gaussian interference channel under strong interference. *IEEE Trans. Inform. Theory*, 27(11):786–788, November 1981.

[28] H. Sato. On the Gaussian interference channel. *IEEE Trans. Inform. Theory*, 31(9):607–615, September 1985.

[29] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423,623–656, 1948.

[30] D. W. Stroock. *Markov Processes from K. Ito's Perspective.* Princeton University Press, 2003.

[31] D. W. Stroock. *Probability Theory: an Analytic View.* Cambridge University Press, second edition, 2003.

[32] E. Telatar. Capacity of multi-antenna gaussian channels. *European Trans. on Telecom*, 10(11):585–595, November 1999.

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