Linear Universal Decoder for Compound Discrete Memoryless Channels

Rethnakaran Pulikkoonattu

Supervisor : Dr. Emmanuel Abbe
Professor : Prof. Bixio Rimoldi

Start date : 03 March 2009
End date : 03 July 2009
Defended on : 16 July 2009

Submitted for the degree of Master of Science

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

July 16, 2009

Abstract

Shannon in his seminal work [1] formalized the framework on the problem of digital communication of information and storage. He quantified the fundamental limits of compression and transmission rates. The quantity *channel capacity* was coined to give an operational meaning on the ultimate limit of information transfer between two entities, transmitter and receiver. By using a random coding argument, he showed the achievability of reliable communication on any noisy channel as long as the rate is less than the channel capacity. More precisely, he argued that, using a fixed composition random codebook at the sender and a maximum likelihood decoder in the receiver, one can achieve almost zero error communication of information.

Shannon's result was a promise on the achievability on the amount of data one can transfer over a medium. Achieving those promises has spurred scientific interest since 1948. Significant effort has been spent to find practical codes which gets close to the Shannon limits. For some class of channels, codes which are capacity achieving has been found since then. Low Density Parity Check Codes (LDPC) is one such capacity achieving family of codes for the class of binary erasure channels. Recently, Arikan proposed a family of codes known as Polar codes which are found to be capacity achieving for binary input memoryless channels. Slowly but steadily the longstanding problem of achieving Shannon limit is thus getting settled.

In order to realize a practical system which can achieve the Shannon limits, two important things are necessary. One of them is an efficient capacity achieving code sequence and the other is an implementable decoding rule. Both these require the knowledge of the probabilistic law of the channel through which communication take place. At the transmitter one needs to know the channel capacity whereas at the receiver a perfect knowledge of the channel is necessitated to build optimum decoder. The question of whether a decoder can be designed without knowing the underlying channel, yet we can do a reliable communication arises here. This is the subject of discussion in this report.

Consider a situation where a communication system is to be designed without explicit knowledge about the channel. Here, neither the transmitter nor the receiver knows the exact channel law. Suppose, the possible list of channels is made available upfront to both entities (transmitter and receiver). Can

we devise a reliable communication scheme, irrespective of the actual channel picked without both transmitter and receiver being aware of it. This is the central theme in what is known[1] as *coding for compound channels*[5].

Designing a reliable communication system for compound channels essentially involve building a codebook at the transmitter and a decoding rule at the receiver. The encoder-decoder pair together must guarantee reliable communication over a set of channels. In the general setting, no feedback is assumed from the receiver to the transmitter. Thus, a single coding strategy (codebook) must be devised (and fixed) upfront prior to transmission. At the receiver end, the decoding strategy must be devised independent of any knowledge of the channel. Of course, the codebook devised at the sender is perfectly known to the receiver.

The answer to the question on whether such a communication system can be build is on the affirmative. In order to talk about the existence of reliable communication strategies, one must first talk about the maximum possible rate, the capacity. The highest achievable rate in a compound setting is known as the *compound capacity* or capacity of the family. This is analogous to the famous Shannon capacity being the ultimate limit on achievable rate for any single channel. The capacity of compound set of memoryless channels has been studied by Blackwell et al in [6] and that of linear dispersive channels is investigated by Root and Varaiya in [8]. Lapidoth and Telatar looked at the compound capacity for families of channels with memory, more specifically the finite state channels [29]. As a special case, they have also derived the compound channel capacity of a class of Gilbert-Elliot channels.

A decoder which operates without the knowledge of the channel in this setup is called a universal decoder. It is known that Maximum Mutual Information (MMI) decoders proposed by Goppa are universal. MMI decoders compute the empirical mutual between a received codeword against the codebook and find the best matching word as the true estimate. The complexity of MMI decoder remain fixed even if we were to find structured codes. This motivate us to ask the question whether we can build a universal decoder which offer better structure. Decoding rules which brings additive nature were considered in the literature as a potential scheme. Our work in this has been driven by this line of thoughts.

In this work, we focus on class of discrete memoryless channels (DMC) and more specifically binary memoryless channels. We show that it is possible to build a linear universal decoder for any compound binary memoryless channels.

The recent introduction of Polar codes motivated us to look into their suitability to the compound channel setup. We have carried out a preliminary investigation in this direction. While it is not clear whether Polar codes are universal under the optimum universal decoding, we find that they are universal for certain restricted classes of compound BMC sets.

[1] Also referred as *coding for class of channels* as discussed in [6], [7].

Acknowledgements

One of the most gratifying thing about my stay in EPFL over the last two years was the opportunity came in my way, to meet, interact and learn from some of the smartest and finest minds in the world.

First of all, I sincerely thank Emmanuel Abbe for being a such a great adviser during this thesis work and more importantly (still) staying as a very good friend. I cherish the many informal discussions we had over lunch and coffee on various topics related to information theory and others. I express my utmost regards to Emmanuel for his patience while explaining various difficult concepts and also for the willingness to correct and connect my (often) vague ideas and thought processes.

I like to thank Professor Bixio Rimoldi for agreeing to be my thesis advisor and more importantly for giving many ideas and hints on practical scenarios where our results could be used. It really helped me to understand a perspective on many things conceived in theory, which I wouldn't have otherwise imagined off. His support and suggestions have further helped to improve the quality of this thesis.

Emre Telatar has been a tremendous source of inspiration in many ways during the last two years. I am thankful to him for sparing time to have many fruitful discussions and also for providing valuable insights, both during the thesis work and otherwise. Above all, his greatness as a teacher as well as the human side of helping students is something I would like to remember him for.

Dr. Juha K. Laurila of Nokia research center, kindly agreed to be there in the thesis committee and I thank him for taking time off to review my thesis, offer suggestions and also being present during my oral defense.

To Ruediger Urbanke, I owe many thanks. First, for teaching many fascinating analytical tools in combinatorics and coding theory. The project during the doctoral course on coding theory is one of the most satisfying experience I ever had in my academic life. I truly have enjoyed every day of that, week long open book examination, while thinking and playing with the stochastic processes on graphs. In my book, he stands in as one of the most easily approachable professors around. I also thank him for the financial support and also being so very friendly and generous.

Olivier Leveque and Professor Pascal Frossard were generous enough to

provide some partial funding at different stages of my stay in EPFL. Besides, the financial gain, scribing the lecture notes of the *Random matrices in communication* course by Olivier, in a great deal helped me to better understand many useful random matrix theory concepts. A big thank you to Nikos and Pascal for our collaborative work in distributed streaming.

Two other people with who I had the pleasure of working together are Etienne Perron and Professor Suhas Diggavi. The Xitip project we did in 2007 is something very satisfying. I am also grateful to Nicolas Macris for sharing some technical wisdom and also for showing interest in my research.

One of the striking thing I will remember about EPFL is the list of great teachers gifted with supreme teaching skills. I am truly fortunate to have had the chance to attend the lectures of easily the best boys in that category. I cannot say enough of these fine teachers like Emre Telatar, Martin Vetterli, Janos Pach and Arjen Lenstra. Twenty years later, I will still have echoes of their words of wisdom flowing around the lecture halls. This list may be too short, but this is what I sampled from my experience here.

I have made several friends during the last two years in EPFL. I gladly take and remember the friendship offered by a lot of good people in EPFL and in particular the warmly atmosphere in IPG lab. Without naming anyone, I like to express my warm regards to all the IPG friends. To Damir being for all the IT support and to the triplet Muriel, Franoise and Yvonne for helping me with the adminstrativia. Special thanks to my current and past office friends Murielle Ange, Vaneet Aggarwal,Hamed Hasani,Mattias Hesse,Jad Khalife,Tiodara Kostic, Pooya Pakzad and Boutaleb Reda for the umpteen number of fun discussions,tea chats and gossips. I also like to thank Adrian Tarniceiru for the numerous tennis games we played in the sports center and for being such a kind friend.To all the cricket lausanne friends, let me express my gratitude for providing so much fun during my stay in Lausanne.

My several well wishers and friends for all the support extended throughout. My close relatives and friends gave me more than I could hope, in terms of taking care of things back home. I am indebted to all of you. To my parents I cant say enough thanks, but this much I have to tell. Without you, literally and figuratively, I simply do not exist. Last, but not least there is no better discovery and happiness than my world around you Maya and I will always place that above all.

Contents

1 Introduction

We are truly living in a digital information age. Every tick of a second see enormous amount of information being exchanged across the space around us. Information flows through the wires and cables from our homes to even the distant islands, at times crossing the oceans and the free space. An email message sent from Bangalore reaches Lausanne in matter of seconds. The gadget called cellular phone sends and receives volumes of digital data. The notion of doing all in *digital* transformed these phenomena as an exchange of '0's and '1's.

What paved the way in realizing today's digital communication revolution, among other enabling factors [2] is the seminal work of Shannon [1], who formulated a mathematical model for efficient representation and reliable transmission of information. Information theory as a scientific discipline was born, largely out of that single event in 1948. This new discipline stays (even today) as the fundamental theoretical framework behind any communication system.

1.1 Digital communication model

Shannon asserted that the communication process is essentially stochastic in nature. He claimed that, the semantic meaning of information is not important in the theory. What matters (in a communication setup) instead is the *surprise* element present in a message. He coined the term *entropy* as a quantifiable measure of this information.

The basic model of Communication proposed by Shannon is shown in Figure.1.1. It consist of a source which generate information, a destination called sink which receives the information. The channel represents the physical medium through which information transfer takes place (from source to

sink).

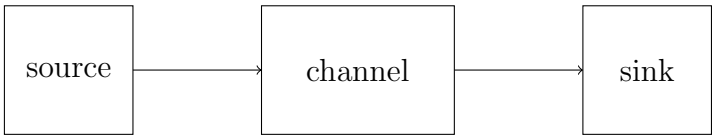


Figure 1.1: Basic communication model

While the model appear extremely simple, it is worth mentioning that, the very same systematic partition holds good for any digital communication or storage system. As previously mentioned, the framework is probabilistic in nature. The source is represented as a realization of a random process $\{X(n)\}$.

The channel model is represented by a conditional probability distribution $W(y|x)$. If $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output alphabet sizes respectively, then the channel is simply a map $W : \mathcal{X} \to \mathcal{Y}$. Shannon showed that, despite the randomness imposed by the channel W, by intelligently introducing redundancy, the intended message can be reproduced at the receiver with high probability. He ntroduced a term known as the *channel capacity* $C(W)$ of a channel W. The operational meaning of this term is that, as long as the transmission is carried out at a rate R less than $C(W)$, there exists ways to transmit information from source to sink, with vanishing error probability.

One of the significant feature of the model that Shannon proposed is the source channel separability. Remarkably, he showed that, the problems of representation and transmission can be dealt separately, without any loss. This inturn enabled the source coding and the channel coding problems to be treated independently. Figure. 1.2 shows the model for point to point communication, illustrating the source-channel separation theorem.

With such a mathematical model, Shannon went on to prove the existence of suitable representation as well as reliable transmission of information. The main arguments in Shannon's claim are outlined in the next section.

1.2 Shannon's promise

The two main theorems in Shannon's communication framework are the *source coding theorem* and the *channel coding theorem*s.

As mentioned earlier, the output of the source is modeled as a stochastic process. The source encoder's goal is to find a minimal (least number of bits per emitted source symbol) representation of the source. Shannon's source coding theorem asserts that, for a given source and a distortion measure, there exists a minimum rate $R = R(d)$ (bits per emitted source symbol) which is necessary and sufficient to describe the source with distortion at most equal to d.

Source encoder is followed by a channel coding module. Shannon's channel coding theorem asserts the existence of an upper limit on the rate at which

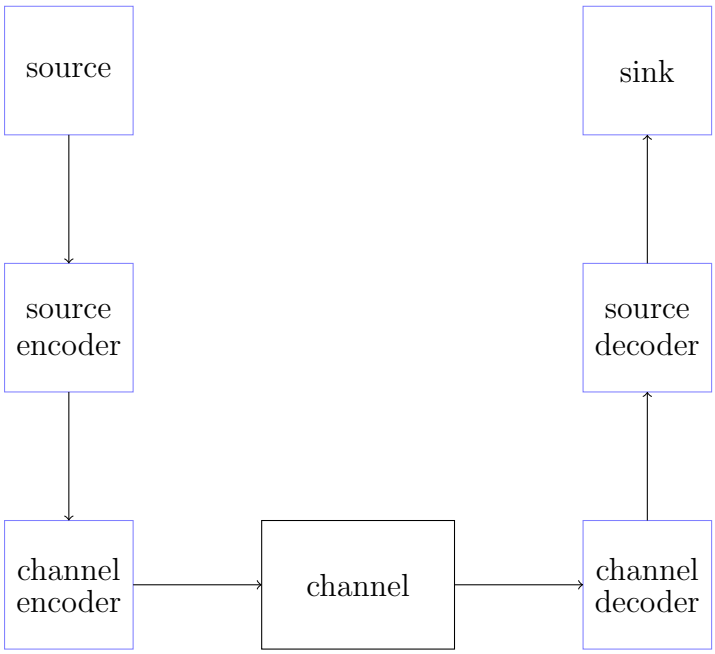


Figure 1.2: Shannon's model of communication:Point to point communication model with source channel separation theorem

information can be reliably transmitted over a probabilistic channel. Here reliability refers to the guarantee that, the probability of error can be brought below any (arbitrarily small) chosen value, for the given channel. This maximum rate is called the capacity (denoted as $C = C(W)$) of the channel W. The way to achieve rate close to capacity is by devising an encoding rule (channel encoder) where redundant bits are systematically added to the source output bit stream.

At the receiver the channel decoder recovers the transmitted information from the received bits. This decoded bits is used by the source decoder to reconstruct the original source data. Shannon's source-channel separation theorem asserts that, the source can be reconstructed at the receiver, with a distortion not exceeding d as long as $R(d) < C(W)$. Moreover, no scheme can do better than this.

In this thesis, we restrict our attention exclusively to the channel coding problem (We assume that, the source coding problem is already solved optimally). We shall assume that, the source emits independent identically distributed (i.i.d) bit sequences. From here onwards(in this thesis), the term encoder and decoder refer to channel encoder and channel decoder respectively.

1.3 From promise to practice

In order to achieve reliable communication over a channel, appropriate encoder (encoding rule) and decoder (decoding rule) are to be designed. A pair of

encoding and decoding map is defined as an *error correcting code* or simply as *code*. The encoding map is applied to the information sequence and produce an encoded message, which is transmitted through the channel. The decoding map is applied to the (corrupted aka noisy) channel output.

The proof of existence of code is based on a random code ensemble argument(random coding argument as it is known). The idea,when roughly stated is as follows [18]: An ensemble of codes $\mathcal{C}$ is constructed using a random process and one proves that *good* codes[1] exist with probability close to 1.

Consider a random ensemble $\mathcal{C}\left(n, 2^{nR}\right)$ of codes of length n and cardinality $M = 2^{nR}$ (where R is the rate). We can think of each code $C\left(n, 2^{nR}\right)$ as an ordered list of M tuples of length n.Using the random coding argument, it is proved that there exist codes $C\left(n, 2^{nR}\right) \in \mathcal{C}\left(n, 2^{nR}\right)$ suitable for reliable communication at rates close to capacity at low error probability, provided that the block length n is sufficiently large.

The random coding argument and the proof of existence however do not consider the description, the encoding and the decoding complexity. If there is no restriction on the code structure, it is easy to see that, the description of the code becomes impractical since the codebook size 2^{nR} grows exponentially with n (So we require $n2^{nR}$).

Achieving reliability with affordable complexity of encoder and decoder has been a goal for communication engineers since 1948.The early research in algebraic codes showed tremendous promise in building structured codes. Even though not quite capacity achieving, they were amenable to building encoders in practical systems. Significant step in the direction of practical coding came with the invention of Convolutional codes by Elias [3] and its low complexity decoding rule by Viterbi [4]. Other important class of codes which came in the early 1950s and yet found significant practical appeal is the Reed Solomon codes. And there were many others too. None of these schemes however achieved the Shannon limit on any channels.

It actually took nearly 40 years before Shannons original promise to be fully realized. A key step towards approaching the Shannon Limit occurred in 1962 Gallager discovered Low Density Parity-Check (LDPC) codes.These codes have since been proved to approach the Shannon Limit for the class of binary erasure channels (BEC) [18]. Largely because of the meager computational capabilities of those times, they remained eluded public attention, until re discovered [19] in the 1990s. The invention of turbo codes [20] in fact spurred the activity on the question of achieving Shannon capacity.

Very recently, Arikan came out with a new class of codes named polar codes [51], which are provably capacity achieving for binary input memoryless channels. Since then, it is shown to have capacity achieving property on many classes of channels [50]. polar codes also have a significant appeal because of the low encoding and decoding complexity. In a sense the arrival of polar

[1]The notion of *goodness* can be thought of as the condition of lowest probability of error in decoded output

codes certainly settled the longstanding search for capacity achieving schemes for a wide class of channels.

During the above discission on the codes, we did not put much emphasis on the decoding aspects. However, an important point to stress here is the decoding rule's dependency on the channel law. The optimal decoder is based on the maximum likelihood (ML) estimate where in the codeword (from the codebook) with the highest likelihood (i.e., the conditional probability of the codeword upon the condition of receiving the given received word) is declared as the likely transmitted word. Here likelihood is related to the channel law. When the code has additional structure (such as algebraic or tree codes),simplification of optimal rules often results to reduced complexity algorithms.

On the other hand, without the knowledge of the channel at the decoder, there is no guarantee on relaible communication of information. From a practical point of view, the knowledge of the exact channel law is a strong assumption. An optimal decoder designed for a certian channel law need not be the one of interest under a different channel. From and engineering point of view, a code (encoder and decoder) which work well for a set of channels is of significant interest. The next section addresses this problem.

1.4 Dealing with a class of channels

Consider a situation where a communication system is to be designed without explicit knowledge of the channel. Here, neither the transmitter nor the receiver knows the exact channel law. Suppose, the possible list of channels (the actual channel of communication will be one from this list) is made available upfront to both entities (transmitter and receiver). Can we devise a reliable communication scheme, when both transmitter and receiver remain ignorant about the exact channel through which communication have taken place? This is the central theme in what is known[2] as *coding for compound channels* [5].

Designing a reliable communication system for compound channels essentially involve building a code book at the transmitter and a decoding rule at the receiver. The encoder-decoder pair together must guarantee reliable communication over a set of channels. In the general setting, no feedback is assumed from the receiver to the transmitter. Thus, a single coding strategy (codebook) must be devised (and fixed) upfront prior to transmission. At the receiver end, the decoding strategy must be devised independent of any knowledge of the channel. Of course, the codebook devised at the sender is perfectly known to the receiver.

A compound channel problem is illustrated in Figure. 1.3. The source here represents (with a slight mix up of notation) the message together with the (universal) source encoder. The output of the source is (already compressed) fed to the encoder. The channel encoder outputs the codeword and send it

[2]Also referred as *coding for class of channels* as discussed in [6], [7].

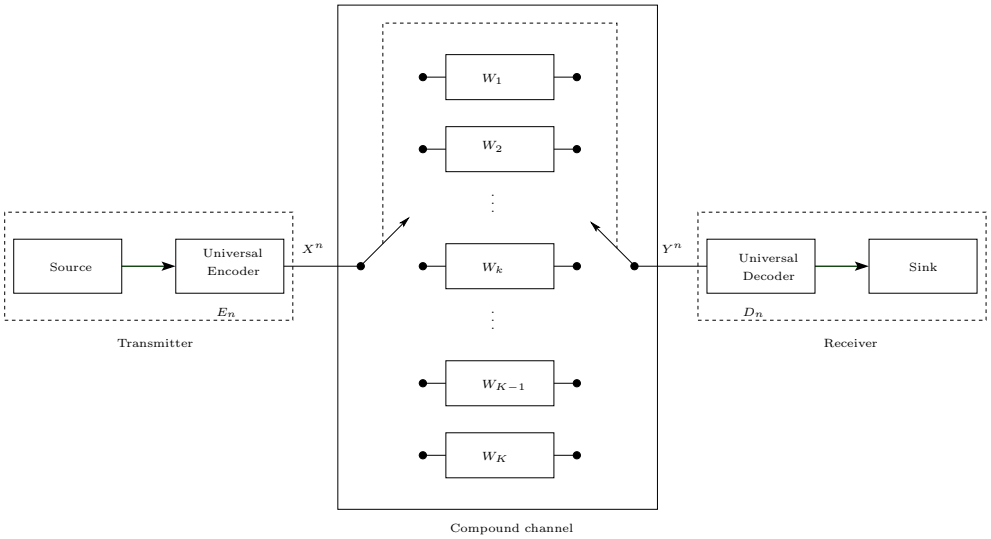


Figure 1.3: Compound channel setup:

to the channel. The exact channel is unknown but it is one among the K possible channels $\{W_1, W_2, \ldots, W_K\}$. The decoder takes the (noisy) channel output and produce the information word. In general K can be infinite, but in this thesis we shall assume it to be finite.

An encoder and a decoder designed for such a setup is known as universal (channel) encoder and universal (channel) decoder [3]. Such a pair of encoder and decoder is defined as a universal code.

Extensive research have been done on the compound channels and universal coding.

1.5 Linearity and universality

From a practical point of view, in order to have a realizable universal decoder, it is highly important that, the decoding algorithm is less to moderately complex. It is well known that Maximum Mutual Information (MMI) decoders are indeed universal for any compound set of memoryless channels. However, the MMI complexity is exponential in the blocklength. Unfortunately it remain to be pessimistic even when the optimal universal capacity achieving code is structured, say for instance tree code. The question then is whether we can design a universal decoding algorithm different from MMI, with a possibility to have a less complexity approximation. The success of iterative decoding strategies using belief propagation (BP) algorithm for Maximum Aposteriori (MAP) decoding on certain practical problems of interest prompt thoughts in

[3]There is similar notion of universality in source coding. The famous Ziv-Lempel compression scheme [9] is considered to be a successful example of such a universal source coder. Our interest here is restricted to the channel coding problem and hence the terms universal encoder and universal decoders are referred to the latter.

these direction. The well known Viterbi algorithm as a moderate alternative to the exponential computation of a maximum likelihood sequence estimation is another example were viable practical alternatives came in place of optimal rule.

The lesser complex alternatives or approximations of MAP and ML decoding algorithms are facilitated by the inherent additive nature of these rules. It is a natural extension to think about a universal decoder which possess additive structure as well.

1.6 Universal codes

Like the famous Shannon's framework, the achievability claims and the existence proofs of universal decoders are based on a random coding argument. Random codes, while serving as powerful tools to prove the existence are not something practical. An interesting question then is, whether a structured universal code exist. If the existence of such a universal structured code exist, that will make the idea of designing efficient receivers without knowledge of the channel of communication a reality.

1.7 Contribution to this Thesis

1.7.1 Linear decoders for compound binary DMC

The universal coding and decoding problems is a longstanding one. The Maximum mutual information (MMI) decoder proposed by Goppa is the first known universal decoder which work for any class of channels. MMI decoders has fixed complexity in the exponential order of the codelength. The complexity does not reduce even if a structured universal code be found. The quest for finding an additive metric induced decoder was considered to alleviate this complexity concern. Abbe and Zheng has established a sufficient condition for a compound set in order to realize an additive (linear) universal decoder. We have considered the case of binary memoryless channels (BMC) and their universal decoding. We prove that, for the class of BMCs, linear decoders exists even beyond the one sidedness. In other words, for any arbitrary binary compound set, it is possible to achieve compound symmetric capacity with a linear decoder. Thus the MMI performance can be achieved by an additive decoder when operating with binary memoryless channels.

1.7.2 Universal polar codes

polar codes caught considerable attention recently because of their capacity achieving properties on binary input memoryless channels. We have looked into the feasibility of polar codes as a universal capacity achieving code. Some recent and emerging results in this arena [49] suggest that these codes cannot

achieve the compound capacity for arbitrary compound set using the successive decoding strategy. Successive decoding rule is adopted in polar codes because of its appealing computational complexity. We are investigating the performance of polar codes with Generalized MAP decoder in the case of binary memoryless compound channels. The subset of good code indices for a degraded channel is proved to be a super set of the mother channel [49]. Using this property, we have characterized the compound sets of binary memoryless channels for which polar codes serve as universal codes. For a given compound rate, we can thus establish a sufficient condition on the set so as to use polar codes as universal. We are still investigating the universal property of polar codes with GMAP decoder for compound BMCs.

Compound Channels and Universal Decoding

2

In this chapter we outline the compound channel problem and the concept of universal decoder.These are the central themes discussed in this thesis. Before addressing these topics, we define the broader class of channels considered in our investigation. We shall restrict the study to the class of discrete alphabet memoryless channels (simply known as the discrete memoryless channels (DMC)). A brief summary of the single channel communication problem is also presented before introducing the subject of compound channels.

2.1 Channel model

Broadly speaking, the channel models considered in this thesis are the discrete alphabet memoryless channels.

Definition 2.1 (discrete channel)**.** *A discrete channel is a system* $(\mathcal{X}, W, \mathcal{Y})$*, with input alphabet* $\mathcal{X}$ *output alphabet* $\mathcal{Y}$ *and transition probability distribution* W*.*

Let X, Y *be finite sets and* $W = \{W(y|x), x \in X, y \in Y\}$ *be a stochastic matrix. A discrete channel* W *with input alphabet* $\mathcal{X}$ *and output alphabet* $\mathcal{Y}$ *is defined by a stochastic matrix of transition probabilities*

$$W : X \to Y. \tag{2.1}$$

An element $W(y|x)$ *of the matrix is a conditional probability of receiving the symbol* $y \in \mathcal{Y}$ *on the channel output if the symbol* $X \in \mathcal{X}$ *is transmitted from the input.*

2.1.1 Discrete memoryless channel

A discrete channel is said to be memoryless if the probability distribution is independent of previous input symbols. A precise definition follows:

Definition 2.2 (discrete memoryless channel (DMC))**.** *Let*

$$X^n \triangleq (X_1, X_2, \ldots, X_n), X_i \in \mathcal{X}$$

and

$$Y^n \triangleq (Y_1, Y_2, \ldots, Y_n), Y_i \in \mathcal{Y}$$

be sequence of n input and output symbols respectively. Let

$$W^n(Y^n|X^n) \triangleq W^n(Y_1, Y_2, \ldots, Y_n|X_1, X_2, \ldots, X_n)$$

be the transition probability $\mathbb{P}(Y_1, Y_2, \ldots, Y_n|X_1, X_2, \ldots, X_n)$. If

$$W^n(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i) \tag{2.2}$$

then W is called a discrete memoryless channel (DMC).

As the name implies, DMC has no memory. The output of the channel at any given instant is independent of the previous (and future) symbols input to the channel.

2.2 Single channel communication

For a single point to point communication link, Shannon's well known result asserts that reliable communication is possible using appropriate coding schemes. We can recall the famous channel coding theorem, when stated it exclusively[1] for the DMC [38].

Theorem 2.3. *Let W be the channel probability corresponding to a discrete memoryless channel (DMC) and let $C(W)$ denote its Shannon capacity. For rate $R < C(W)$ there exists sequence of codes $\mathcal{C}_n$ of increasing length n and rate $R_n \to R$ such that the block error probability $P_B^{(n)} \to 0$ as $n \to \infty$. Conversely, for any $R > C(W)$, the block error probability for any code with rate at least R is bounded away from zero.*

The theorem simply claims the existence of at least one code using which one can achieve an arbitrarily small error probability as long as the rate is less than the channel capacity. The converse statement implies that, when the rate exceeds channel capacity, the error probability cannot be brought down to zero, no matter how sophisticated the code is.

The proof of this famous theorem is based on the random code ensemble idea. We shall assume binary codes for simpler explanation. A code with block length n and rate R has 2^{nR} codewords[2]. Each of the codeword is a

[1] the theorem applies in general to any channel, not restricted to DMC

[2] Strictly speaking it should be $\lceil 2^{nR} \rceil$, but illustrating the idea we skip these details

point in the n-dimensional hypercube F_2^n where each element is an n-tuple. Among 2^n possible $n-$ elements construct a code $\mathcal{C}$ by drawing tuples based on a distribution (for example, uniformly at random in the simple case) 2^{nR} times. Repeat this process many times to produce an ensemble of codes. Once an ensemble is defined, one show that the block error probability averaged over the code ensemble vanishes in the block length n when $R < C(W)$ and then conclude that there exist at least one code in the ensemble with performance as good (if not better) as the average.

We have omitted the exact proof by merely referring to [1], [18] or [17] for details. We now focus attention to the optimum decoding rules.

2.2.1 Optimum decoder

Consider the transmission scheme over a channel W at rate R using a code $C\left(n, M = 2^{nR}\right) = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\right\}$. The transmitter chooses a codeword $X \in C(n, M)$ with probability $P_X(x)$. This codeword is transmitted over a channel with transition probability $W(y|x)$ to produce and observation Y at the output. The decoding task is to map this Y to a valid codeword such that the error in such a decision is minimum. If we decode Y to $\hat{x}(Y) \in C$, then the error is $1 - P_{X|Y}\left(\hat{x}(Y)|y\right)$. The rule which maximizes $P_{X|Y}\left(\hat{x}(Y)|y\right)$ is called the *maximum a posteriori* (MAP) decoding rule and it is

$$\begin{aligned}
\hat{x}_{MAP}(y) &= \arg\max_{x \in C} P_{X|Y}(x|y) \\
&= \arg\max_{x \in C} P_{Y|X}(y|x)\frac{P_X(x)}{P_Y(y)} \\
&= \arg\max_{x \in C} P_{Y|X}(y|x)P_X(x) \\
&= \arg\max_{x \in C} W(y|x)P_X(x)
\end{aligned}$$

when all the codewords are equally likely (i.e., $P_X(x)$ is uniform), then $\hat{x}_{MAP}(y) = \arg\max_{x \in C} W(y|x)$ and it is simply the *maximum likelihood* (ML) decoder.

$$\hat{x}_{ML}(y) = \arg\max_{x \in C} W(y|x) \tag{2.3}$$

As discussed, the ML decoder realizes the least decoding probability of error when the codewords are transmitted equal likely. Uniform distribution of codewords is a valid assumption in digital transmission setup and hence, ML rule is an optimum decoder when decoding probability of error is chosen as the optimality criteria.

Although, ML decoder is optimum, from a practical point of view two important hurdles comes along with it. Firstly, the decoder needs to know the channel rule and secondly, the huge computational burden of searching through the codebook to find the optimum codeword. The latter bottleneck can be brought down when the code is structured. Codes such as convolutional

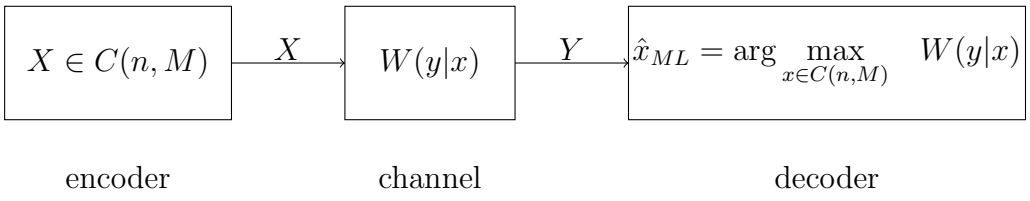


Figure 2.1: ML decoder

codes which assumes interesting structural properties, significant reduction in decoding complexity can be achieved. Finding suitable codes which admit less complex decoder alternatives is a difficult task and that is beyond the scope of this work.

On the other hand, the dependency of the decoder on the channel will demand that the channel rule be be known at the receiver. This is often an unrealistic assumption in practice. Using training symbols (pilot sequences as referred in many communication standards) one can hope to estimate the channel and use this estimate subsequently to decode the codeword. Clearly, such a scheme will have to compromise on the rate and hence the capacity achieving scheme will be jeopardized when the receiver a priori do not the channel. Does it mean that all decoders operating without knowledge of the underlying channel perform poorly?

Curiously enough, a decoder proposed by Goppa can indeed achieve capacity for discrete memoryless channels. The decoder, referred to as the maximum mutual information (MMI) decoder, selects an input message that maximizes the empirical mutual information with the given output message (channel output vector). Goppa has shown that for discrete memoryless channels (DMC),a receiver when employed the MMI decoder, which is independent of the unknown channel statistics, the channel capacity is achievable.

Csiszir and Komer later sharpened Goppa's result and proved the existence of a deterministic universal fixed composition block code, which achieves the random coding error exponent for the given channel, using the MMI decoder.

The formal definition and a subsequent discussion of MMI decoder is present later in the chapter. Since MMI rule do not need the channel information, such a scheme can serve as a decoder for the compound channel setup. We now present a brief introduction to the compound channel problem and the concept of universal decoding. The game is now about information transmission carried over a set of channels instead of a single channel.

2.3 Communication over family of (unknown) channels

Consider a situation where a communication system is to be designed without explicit knowledge of the channel. Here, neither the transmitter nor the

receiver knows the exact channel law. Suppose, the possible list of channels is made available upfront to both entities (transmitter and receiver). Can we devise a reliable communication scheme, irrespective of the actual channel picked without both transmitter and receiver being aware of it. This is the central theme in what is known[3] as *coding for compound channels*[5].

Designing a reliable communication system for compound channels essentially involve building a codebook at the transmitter and a decoding rule at the receiver. The encoder-decoder pair together must guarantee reliable communication over a set of channels. In the general setting, no feedback is assumed from the receiver to the transmitter. Thus, a single coding strategy (codebook) must be devised (and fixed) upfront prior to transmission. At the receiver end, the decoding strategy must be devised independent of any knowledge of the channel. Of course, the codebook devised at the sender is perfectly known to the receiver.

The answer to the question on whether such a communication system can be build is on the affirmative. In order to talk about the existence of reliable communication strategies, one must first talk about the maximum possible rate, the capacity. The highest achievable rate in a compound setting is known as the *compound capacity* or capacity of the family. This is analogous to the famous Shannon capacity being the ultimate limit on achievable rate for any single channel. The capacity of compound set of memoryless channels has been studied by Blackwell et al in [6] and that of linear dispersive channels is investigated by Root and Varaiya in [8]. Lapidoth and Telatar looked at the compound capacity for families of channels with memory, more specifically the finite state channels [29]. As a special case, they have also derived the compound channel capacity of a class of Gilbert-Elliot channels.

In this work, we restrict our attention to class of discrete memoryless channels (DMC) and more specifically binary memoryless channels.

2.4 Compound channels

The term *compound channels* refer to a set of channels. The exact channel is hidden from both transmitter and receiver, but some general information on the broader type (say for instance all channels in the set are DMCs) is known to everyone. The set of channels is also known as compound set (of channels). All the channels in the compound set have the same input alphabets as well as output alphabets[4].

A compound discrete memoryless channel (compound DMC) is a compound set of discrete memoryless channels.

[3]Also referred as *coding for class of channels* as discussed in [6], [7].

[4]We consider only discrete channels in this thesis.

Compound DMC

Consider a DMC with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The transmitter and receiver do not know the exact channel transition matrix W.But W belong to a set $\mathcal{S}$ of $K = |\mathcal{S}|$ DMCs. That is., $W \in \{W_1, W_2, \ldots W_K\}$. Such a setup is referred to as compound DMC.

2.5 Compound channel capacity

The fundamental limit on the highest rate of reliable transmission over a compound set is given by its compound capacity. We present the formulation of the compound channel capacity of a compound DMC.

Theorem 2.4 (Channel capacity of a compound DMC)**.** *Consider a DMC with fixed input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The compound channel capacity of a set $\mathcal{S}$ of DMCs is given by,*

$$C_{\mathcal{S}} = \max_{P_X} \inf_{W \in \mathcal{S}} I\left(P_X, W\right) \tag{2.4}$$

It is implicit that in the set $\mathcal{S}$, each of the constituent member DMC is defined over fixed input alphabets $\mathcal{X}$ and output alphabets $\mathcal{Y}$. The infimum is performed over the set of channels (in the set $\mathcal{S}$) and the maximization is carried out over all possible input distributions. $I\left(P_X, W\right)$ is the mutual information[5] between the input and output, defined as.

$$I\left(P_X, W\right) = \sum_x P_X(x) \log \left(\frac{P_X(x) W(y|x)}{P_X(x) \sum_x P_X(x) W(y|x)} \right) \tag{2.5}$$

It is worth noting that the compound channel capacity in general, is not equal to the infimum of the individual capacities of the constituent channels in the set. The optimum source distribution $P_X(x)$ which achieve capacity may be different for different channels in the set. We can thus state an upper bound on the compound capacity as the infimum of member channel capacities. This is summarized in the following proposition.

Proposition 2.5. *The compound channel capacity $C_{\mathcal{S}}$ of a set of DMCs is at most equal to the minimum of the member channel capacities. That is,*

$$C_{\mathcal{S}} \leq \inf_{W \in \mathcal{S}} C(W) = \inf_{W \in \mathcal{S}} \max_{P_X(x)} I\left(P_X(x), W(y|x)\right). \tag{2.6}$$

The equality holds when the capacity achieving distributions of all channels in the set is identical.

[5]The usual notation of mutual information $I(X;Y)$ is the same as $I\left(P_X, W\right)$. Since mutual information is a function of distributions, we have adopted this notation.

The above proposition states that, exchange of infimum and maxima in 2.4 are not equivalent.

To show the achievability of reliable communication in the compound setting, we must then show the existence of codes such that neither the codebook nor the decoder rely on the actual channel being used.

The classical coding theorems uses optimum decoding rule such as maximum likelihood (ML) or decoders using joint typicality arguments in order to prove the existence of codes which can asymptotically achieve reliable communication with rate less than capacity. In the compound setting this cannot be used since the channel law is unknown. Both ML and joint typicality arguments exploit the channel knowledge at the decoder. The other important tool used in the classical case is the random coding argument. For a random ensemble of codes, it computes the average probability of error and argues that at least one code in the ensemble will perform as good as this average. Since the codebook chosen from the random ensemble is usually dependent on the channel, for the compound setting, one must also show the existence of a codebook which is simultaneously good for all the channels in the set.

2.5.1 Achieving Compound Capacity

Let $P_e\left(C_n, D_n, W\right)$ be the probability of error when a codebook $C_n \in \mathcal{C}$ (A code C_n of block length n is chosen from the ensemble of codes $\mathcal{C}_n$ of same block length), decoding rule D_n and channel $W \in \mathcal{S}$ is used. Let us assume that, the decoding rule D_n is made independent of the actual channel W. The expectation of the probability of error over the ensemble of codes $\mathcal{C}$

$$P_e\left(\mathcal{C}_n, D_n, W\right) \doteq \underset{C_n \in \mathcal{C}_n}{\mathbb{E}}\left[P_e\left(C_n, D_n, W\right)\right] \leq \delta$$

does not necessarily imply the existence of a single code E_n such that

$$P_e\left(C_n, D_n, W\right) \leq \delta, \forall W \in \mathcal{S}$$

.

This is because, different codes can have small error probability and thus resulting in lower expected value of error probability. For the compound channel we need to prove the existence of (at least one) code which is simultaneously good for all the (member) channels in the set. The random coding arguments strengthened after incorporating the above requirements is used in [6] to prove the existence of a universal capacity achieving code for compound set of channels.

Theorem 2.6. *For a class $\mathcal{S}$ of DMCs, as long as the rate $R < \inf_{W \in \mathcal{S}} I\left(P_X, W\right)$ there exists a codebook rule E_n and decoding rule D_n, with both E_n and D_n independent of the channel law, such that for any arbitrary $\delta > 0$, the probability of decoding error can be made arbitrarily smaller than δ for any $W \in \mathcal{S}$.*

In the original work, the authors used a decoder D_n which maximizes a uniform mixture of likelihoods over a set of channels. When the compound set is finite, using the random coding argument, they prove that, the intersection of good codebooks for all channels is non empty. For arbitrary compound set, the existence of good codewords is established when the maximum of a uniform mixture of likelihoods grows only as *polynomial* in n, the codelength. We would like the reader to refer [6] for the original proof. A slightly different proof of the same theorem is provided by Abbe and Zheng in [21] using a different decoding rule namely, Maximum Mutual Information (MMI) decoder [10].

2.6 Maximum Mutual Information (MMI) Decoder

Maximum Mutual Information (MMI) decoder was proposed by Goppa in [10]. MMI decoder computes the empirical mutual information between the received vector y and all elements $x_m, m = 1, 2 \ldots, |\mathcal{C}|$ of the codebook $\mathcal{C}$. Once the empirical mutual information is computed, it chooses the x_m corresponding to the maximum mutual information as the estimate of the sent codeword. A formal definition of MMI decoder follows. First we define the joint empirical distribution of a pair of random vectors x and y.

A decoder refers to a decoding rule operating on a received vector $y \in \mathcal{Y}^n$ along with known codebook $\mathcal{C}_n$ and output a likely codeword that is sent.

Definition 2.7 (Joint Empirical distribution of (x, y))**.** *Let $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$. Joint empirical distribution $P_{x,y}$ of (x, y) is given by*

$$P_{x,y}(u, v) = \frac{f(u, v)}{\sum\limits_{\alpha \in |\mathcal{X}|, \beta \in |\mathcal{Y}|} f(\alpha, \beta)}$$

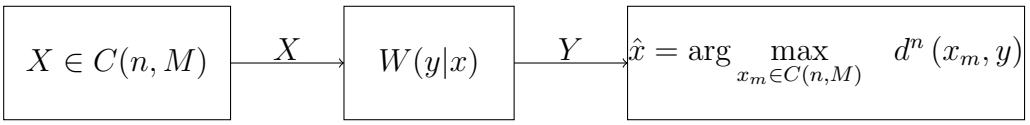
where

$$f(u, v) \triangleq |\{i : x(i) = u, y(i) = v\}|$$

Since $\sum\limits_{\alpha \in |\mathcal{X}|, \beta \in |\mathcal{Y}|} f(\alpha, \beta) = n$, we can simply express the joint distribution into the following form:

$$\hat{P}(u, v) \doteq P_{x,y}(u, v) = \frac{|\{i : x(i) = u, y(i) = v\}|}{n}$$

Before stating the MMI decoder formulation, let us bring in a (statistical) relation between the ML rule and joint empirical distribution we just discussed.



encoder channel decoder

$$d^n_{ML}(x_m, y) = W(y|x_m)$$
$$d^n_{MMI}(x_m, y) = I\left(\hat{P}(x_m, y)\right)$$

Figure 2.2: d-decoder

2.6.1 Maximum Likelihood Decoding

Maximum Likelihood (ML) rule is the optimum decoding strategy when the channel is known to the decoder. It is interesting to see that this optimum rule is connected to the empirical distribution.

$$W^n(y|x_m) = \prod_{u \in \mathcal{X}, v \in \mathcal{Y}} [W(v|u)]^{nP_{x_m,y}(u,v)} = 2^{n \underset{P_{x_m,y}}{\mathbb{E}}[\log W]} \tag{2.7}$$

$$\begin{aligned} x_{ML}(y) &= \arg\max_{x_m} \; W^n(y|x_m) \\ &= \arg\max_{x_m} \; 2^{n \underset{P_{x_m,y}}{\mathbb{E}}[\log W]} \\ &= \arg\max_{x_m} \; d^n_{MMI}(x_m, y) \end{aligned}$$

where

$$d^n_{ML}(x_m, y) = 2^{n\mathbb{E}_{\hat{P}_{x_m,y}}[\log W]}$$

Thus ML decoding maximizes the score (metric) $d^n_{ML}(x_m, y)$ for a given received vector $y \in \mathcal{Y}^n$ against all the elements $x_m, m = 1, 2, \ldots, |\mathcal{C}|$ of the codebook $\mathcal{C}$. The superscript n indicates the length of the codeword. Note that, the ML rule has the channel W incorporated in the metric. It is well known that ML rule is the optimum decoding rule when the channel law is available to the receiver.

Now, we are ready to define the maximum mutual information decoding principle.

2.6.2 MMI Decoding

Definition 2.8 (Maximum Mutual Information (MMI) rule)**.** *The maximum mutual information (MMI) decoding rule is given by,*

$$x_{MMI}(y) = \arg\max_{x_m} \; I\left(\hat{P}(x_m, y)\right)$$

$$= \arg\max_{x_m} \quad d^n_{MMI}\left(x_m, y\right)$$

Where $d^n_{MMI}\left(x_m, y\right)$ *is known as the MMI metric, given by*

$$d^n_{MMI}\left(x_m, y\right) = I\left(\hat{P}\left(x_m, y\right)\right),$$

and $I(\mu)$ *denote the mutual information of the joint distribution* μ *on* $\mathcal{X} \times \mathcal{Y}$. *Formally,*

$$\begin{aligned} I\left(\hat{P}\left(x_m, y\right)\right) &= \sum_{x_m \in C(n,M)} \hat{P}\left(x_m, y\right) \log \frac{\hat{P}\left(x_m, y\right)}{\hat{P}\left(x_m\right)\hat{P}\left(y\right)} \\ &= \sum_{x_m \in C(n,M)} \hat{P}\left(x_m, y\right) \log \frac{\hat{P}\left(x_m, y\right)}{\hat{P}\left(x_m\right) \sum\limits_{u \in C(n,M)} \hat{P}\left(u, y\right)} \end{aligned}$$

In words, MMI decoder computes the empirical mutual information (EMI) and perform a maximization over the codebook. The one codeword with the highest EMI is declared as the likely sent codeword. Figure 2.3 shows a pictorial representation of MMI decoder.

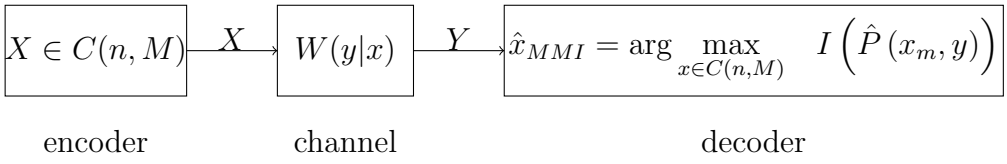


Figure 2.3: MMI decoder illustration

MMI decoder performance

Although, the Maximum Mutual Information (MMI) decoding rule is sub-optimal, for the case of DMC, it can achieve channel capacity. This result was proved by Goppa[**?**]. Subsequently, Csiszir and Komer proved the existence of a deterministic universal fixed composition block code which achieves the random coding error exponent for the given (DMC) channel, using the MMI decoder. These results substantiate the fact that a decoder without the knowledge of the channel law can indeed perform as good as the optimal decoder. In short, with MMI, knowing the (DMC) channel at the receiver (i.e. decoder) does not increase capacity!

There are no free lunches. While MMI decoder offer a strong lead in performance, it has some shortcomings when it comes to implementation. This is discussed in section 2.7.1. Before that, let us discuss this decoder in the context of universal decoding of compound DMCs.

2.7 MMI as universal decoder

As we have seen, the obvious difference between MMI and the optimal ML decoding rule is the non-dependency of the former to the channel (probability) law. MMI relies on the computed empirical joint distribution, which does not require to know the channel of communication. Because of this property, MMI immediately qualify as a potential candidate for universal decoding of compound channels. We have not yet explicitly stated (as a theorem) how well MMI perform in a compound setting, but it is a direct implication from the DMC claim (section 2.6.2) that such a decoder achieves compound capacity for DMCs. In fact, when used with optimal universal codebook, MMI achieves the compound channel capacity on any compound set.

As a universal decoding scheme, MMI decoder brings in other important advantages too. From a decoding point of view, it doesn't need any information about the compound set $\mathcal{S}$ (Not only that MMI doesn't need to know the channel, it also doesn't need to know anything about the set). Moreover MMI can achieve the random coding exponent for any $W \in \mathcal{S}$ [5]. The universal achievability of MMI decoder is summarized in the following theorem [21]. The reader can find a proof and discussion of this theorem in [21].

Theorem 2.9 (MMI decoder is unversal)**.** *MMI decoder is universal. That is for* $R < I(P_X, W)$*,*

$$\mathop{\mathbb{E}}_{\mathcal{C}}\left[\mathbb{P}\left(\mathcal{C}, MMI, W\right)\right] \leq e^{-nE_r(R,P_X,W)} \tag{2.8}$$

where $E_r(R, P_X, W) > 0$

2.7.1 Discussion:Implementation complexity of MMI decoder

The machinery used in MMI decoder is rather simple. For any codebook that is compound capacity achieving, the decoder compute the empirical mutual information and decide in favour of the valid codeword which gives the highest mutual information to a received word. Even if the codebook identified (as capacity achieving) is structured (for example, tree code or algebraic codes), computation of empirical joint distribution calls for a run through procedure over the entire codebook which is exponential [6] in the codeword length n.

So far the existence claims of compound capacity achieving schemes are all based on random coding arguments. Explicit construction of structured capacity achieving code is an entirely difficult problem in itself, even for single known channels, let alone finding one for a class of channels. The field of coding theory is motivated to address challenging problems of this kind.

Now, suppose an explicit construction of structured compound capacity achieving code is discovered, in order to realize such a scheme in practice,

[6]The complexity order is $\mathcal{O}\left(M^2\right)$ where $M = 2^{nR}$ for binary codes

we still need to find a way to circumvent the fixed exponential complexity of MMI decoding. Can we hope to build a decoder which perform as well and at the same time manageable overall complexity? The next chapter addresses this problem by conceptualizing a universal linear decoder which exhibit an additive structure like the classical maximum likelihood decoder operating on a DMC.

3 Linear Universal Decoding

We have seen that MMI decoder serves as a universal decoder for any compound set. While this is a welcoming sign, it poses some concerns due to its rigid complexity even with a structured codebook. Since the capacity achieving codes are usually long length codewords, the exponential scaling of the complexity often serve as a sever bottleneck to realize such a decoder.

Then, a natural question arise is: whether there is a decoder which is significantly less complex than MMI and yet perform as good as MMI. While it is difficult to come out with an easy answer, to this problem, one line of thought is motivated by the success of many structured codes such as tree codes, convolutional codes and LDPC codes for which efficient low complexity approximate algorithms circumvent the implementation bottlenecks of optimum ML and MAP algorithms.

3.1 Linear Universal Decoder

3.1.1 Linear Decoder

Definition 3.1. *A linear decoder induced by a single metric d is a rule given by*

$$D_n(y) = \arg max_m d^n(x_m, y)$$

where

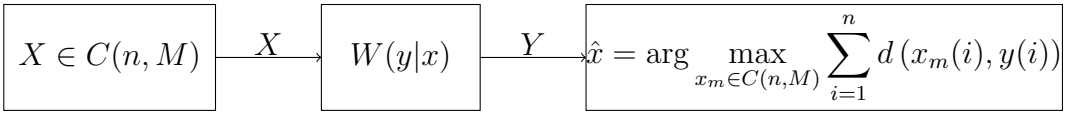
$$d^n(x_m, y) = \frac{1}{n}\sum_{i=1}^{n} d(x_m(i), y(i)) = \mathbb{E}_{\hat{P}(x_m,y)}[d]$$

3.1.2 Nomenclature: Linear Decoder

An obvious thing to notice in the above rule is that, the decoding metric $d^n(x_m, y)$ is a linear function of the empirical distribution $\hat{P}(x_m, y)$. Moreover, for a linear decoder, the decoding metric d^n for any codeword length n has an additive structure. It is simply the sum of n single letter metrics!

Example 3.2. *The ML rule when viewed as a log likelihood is a linear decoder. For a given channel W, the ML rule can be written as a linear decoder induced by*

$$d_{ML}(u, v) = \log W(v|u), \forall u \in \mathcal{X}, v \in \mathcal{Y}.$$



encoder channel decoder

$$d^n_{ML}(x_m, y) = \log W(y|x_m) = \sum_{i=1}^{n} \log W(y(i)|x_m(i))$$

Figure 3.1: Linear decoder

3.1.3 Advantages of linear decoders

The various merits of having linear decoders are discussed in the literature [**?**] [36] [34] [22]. The explicit advantages that comes with an additive decoder also depend on the nature and structure of the codebook used as well as on the class of channels. Under suitable conditions, by exploiting the particular code structure, the decoding rule can be simplified to within affordable complexity. A well known example is the classical Viterbi decoding algorithm for convolutional codes (a tree structured code) which perform the maximum likelihood sequence estimation within manageable computational steps, as compared to brute force ML algorithm which demand exponential complexity[53] [52]. Similarly, belief propagation (BP) algorithm is widely used as a reduced complexity iterative scheme in place of the optimum maximum aposteriory (MAP) rule in many applied problems in engineering. In some special cases, it has been proved that, MAP and BP are equivalent. A well known example of such equivalence is the capacity achieving problem in binary erasure channel (BEC) using Low Density parity Check (LDPC) codes [18].

Having said that, the mere exisitence of a linear decoding rule does not automatically yield a reduced complexity implementation. However in many cases, when the codebook has special structure one could hope to find an

equivalent or at the least an approximate algorithm as a practical alterative to the optimal scheme.

3.1.4 Performance of Linear decoder

The achievable rates achieved using linear decoder induced by a single metric is captured in the following lemma [6],[22].

Lemma 3.3. *For a DMC W_0, using a random codebook drawn from an input distribution P_X and a linear decoder induced by d, the following data rate can be achieved.*

$$R\left(P_X, W_0, d\right) = \inf_{\mu \in \mathcal{A}} \quad D\left(\mu \| \mu_0^p\right) \tag{3.1}$$

where

$$\mathcal{A} = \left\{\mu : \mu_X = P_X, \mu_Y = \left(\mu_0\right)_Y, \mathbb{E}_\mu[d] \geq \mathbb{E}_{\mu_0}[d]\right\}$$

3.2 Generalized Linear Decoder

Definition 3.4. *Let $d_1, d_2, \ldots, d_K$ be a finite number K of single letter metrics. The number K is finite and is assumed to be independent of n. A generalized linear decoder induced by this set of metrics is defined by the following map.*

$$\begin{aligned} D_n(y) &= \arg\max_m \quad \bigvee_{k=1}^{K} \sum_{i=1}^{n} d_k\left(x_m(i), y(i)\right) \\ &= \arg\max_m \quad \bigvee_{k=1}^{K} \mathbb{E}_{\hat{P}(x_m, y)}\left[d_k\right] \end{aligned}$$

where $\bigvee$ denotes the maximum of the set.

3.2.1 Performance of Generalized Linear Decoder

The achievable rates achieved by using an additive decoder is summarized in the following lemma [6],[22].

Lemma 3.5. *For a DMC W_0, using a random codebook drawn from an input distribution P_X and a linear decoder induced by a finite number K of single letter metric $\{d_k\}_{k=1}^K$, the following data rate can be achieved.*

$$R\left(P_X, W_0, \{d_k\}_{k=1}^K\right) = \inf_{\mu \in \mathcal{A}} \quad D\left(\mu \| \mu_0^p\right) \tag{3.2}$$

where

$$\mathcal{A} = \left\{\mu : \mu_X = P_X, \mu_Y = \left(\mu_0\right)_Y, \bigvee_{k=1}^{K} \mathbb{E}_\mu[d] \geq \bigvee_{k=1}^{K} \mathbb{E}_{\mu_0}[d]\right\}$$

3.3 Achieving Compound capacity

We know the achievable rates for a linear decoder for arbitrary metrics $\{d_k\}_{k=1}^K$. Our interest is motivated by the problem of achieving rates close to the compound channel capacity. For a given set $\mathcal{S}$ of DMCs with the compound capacity $C_{\mathcal{S}}$, how do we pick the metrics such that we can push the achievable rate $R\left(P_X, W_0, \{d_k\}_{k=1}^K\right)$ as close to $C_{\mathcal{S}}$ as possible? Does there exist a set of metrics any arbitrary compound set, which can provablly achieve the corresponding compound capacity of the set?

3.3.1 Sufficient conditions

For arbitrary compound sets, finding a finite set of metrics which can achieve compound capacity is still an open problem. Abbe and Zheng [22] have established a sufficient condition on compound DMC sets in order to realize a linear decoder which is capacity achieving. They introduced a new notion called *one-sided sets*. Informally speaking, they proved that for one sided compound DMCs, linear decoders which are compound capacity achieving can be designed. They have further extended the result to the case of compound sets which are union of finite number of one sided sets. Thus a sufficient condition for the existence of a linear universal decoder is to have a compound set which is either one sided or a union of finite one sided sets.We will formally state their main results. But first, we introduce the key concept of one sided sets.

Definition 3.6. *A set $\mathcal{S}$ is one sided if*

$$D\left(\mu_0 \| \mu_{\mathcal{S}}^p\right) \geq D\left(\mu_0 \| \mu_{\mathcal{S}}\right) + D\left(\mu_{\mathcal{S}} \| \mu_{\mathcal{S}}^p\right)$$

where,

$$W_{\mathcal{S}} = \arg \min_{W \in cl(S)} I\left(P_X, W\right)$$

and $\mu_0 = P_X \circ W_0, \mu_S = P_X \circ W_S$, are the joint distributions over the channel W_0 and W_S respectively.

Closure of a set $cl(S)$ is the smallest closed set containing S.

Proposition 3.7. *For one-sided sets $\mathcal{S}$, the linear decoder induced by the metric $d = \log W_{\mathcal{S}}$ is capacity achieving.*

Note that in [6], the same linear decoder is proved to be capacity achieving for the case where $\mathcal{S}$ is convex. Convex sets are one-sided and there exist one-sided sets that are not convex.

Proposition 3.8. *For $\mathcal{S} = \bigcup_{k=1}^K \mathcal{S}_k$, where $\{\mathcal{S}_k\}_{k=1}^K$ are one-sided sets, the generalized linear decoder induced by the metrics*

$$d_k = \log W_k(y|x), k = 1, 2, \ldots, K.$$

is not necessarily capacity achieving.

Theorem 3.9. *For $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}_k$, where $\{\mathcal{S}_k\}_{k=1}^{K}$ are one-sided sets, the generalized linear decoder induced by the metrics*

$$d_k = \log \frac{W_k(y|x)}{\sum_{x \in \mathcal{X}} P_X(x) W_k(y|x)}, k = 1, 2, \ldots, K.$$

is capacity achieving.

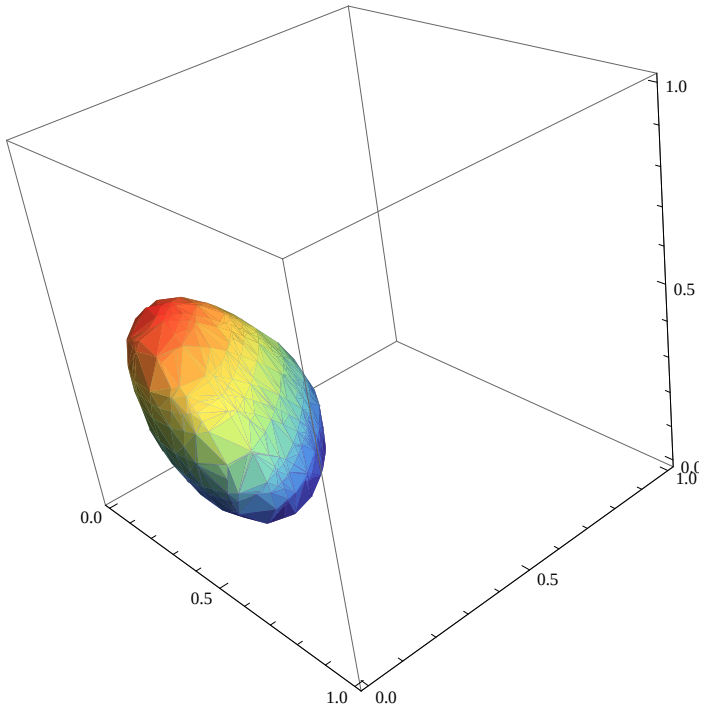


Figure 3.2: KL Ball in 3D (4 dimensions on a 3-simplex)

3.4 Information Geometry

A large number of problems in information theory can be expressed as optimization problems involving Kullback-Leibler divergences. Many channel capacity formulations, rate distortion functions turns out to be minimizations of KL divergences. Using the notion of geometry in a probability space, many useful insights can be obtained.Even though divergence is not strictly a metric, it has some properties of a metric which helps to exploit geometrical results, much in analogy to euclidean geometry. When two distributions are close by, the divergence behaves much like a Euclidean metric. Such a regime is alled Local geometry. When the distributions are away, the geometrical regime is

known as global geometry [21][22]. An important tool in information geometry is the notion of I-projection. This the equivalent of 'projection' in Euclidean geometry.

3.4.1 I-projections

The I-projection of a distribution $q(x)$ on a set of distributions $\mathcal{P}$ is defined as the distribution $p(x)$ in $\mathcal{P}$ which minimizes the KL-divergence $D(P\|Q)$.

$$P^\star = \arg\min_{P\in\mathcal{P}} D\left(P\|Q\right) \triangleq \mathcal{I}_p(Q,\mathcal{P})$$

We use the notation $\mathcal{I}_p(Q,\mathcal{P}) = P^\star$. The I-projection satisfies the triangle inequality property [16].

$$D(P\|Q) \leq D\left(P\|P^\star\right) + D\left(P^\star\|Q\right).$$

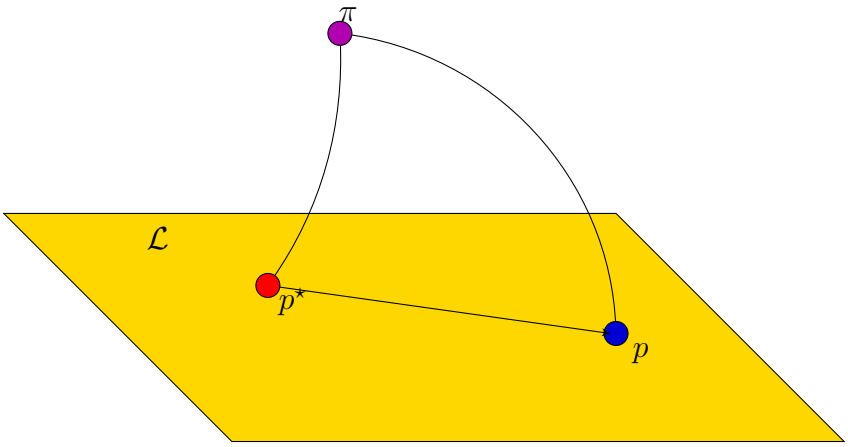


Figure 3.3: I-Projection: geometry

3.5 Geometry of divergence

Even though the KL divergence and Euclidean distance share corresponding notion of metric in Information geometry and Euclidean geometry respectively, the nature of their distance properties are strikingly dissimilar in general. For instance, the euclidean ball in 3 dimension is sphere, whereas the KL ball in a 3-simplex has an irregular shape as illustrated in Figure. 3.2. Similarly, the Euclidean equipotent surface in two dimension is a circle whereas, the trace of equal measure of KL divergence in Information geometry space namely the 2-simplex is rather different. This behaviourial comparison is depicted in Figure. 3.5.

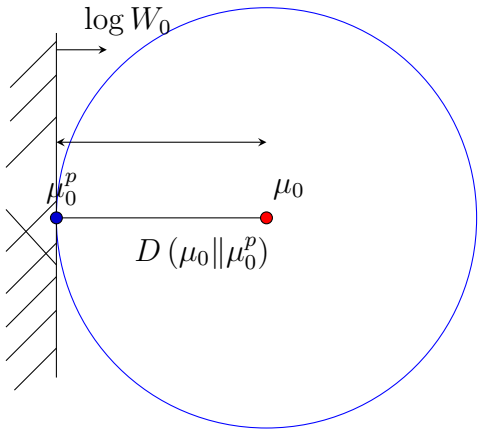


Figure 3.4: Geometry of divergence

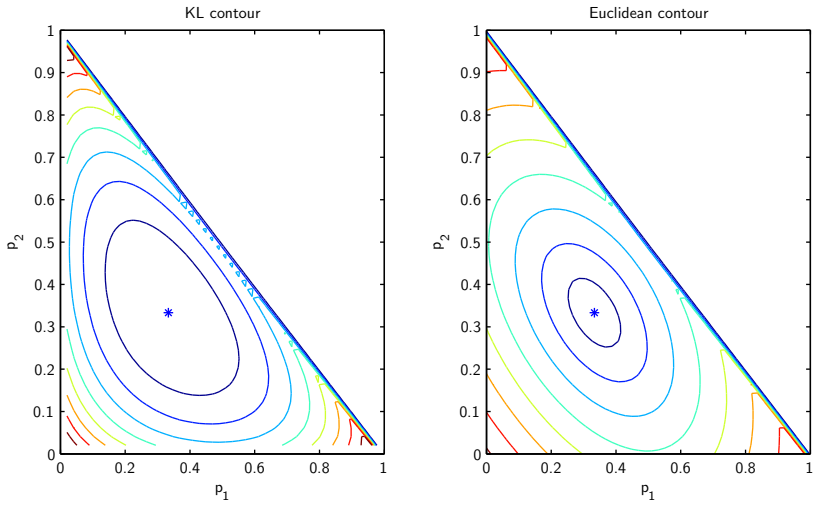


Figure 3.5: Contours of the two balls, Euclidean and KL ball

Linear Universal Decoder for Binary Memoryless Channels

4

We now focus attention to a special class of DMCs, the binary memoryless channels (BMC). These are channels which has input and output alphabet size equal to two. In the literature, the term *binary memoryless channel* is often referred to a slightly wider class of DMCs, where the input alphabet is binary, whereas the output alphabet can be of any arbitrary size [1], not necessarily equal to two.Our emphasis is however restricted to the case where both input and output alphabets are binary. A formal definition adopted for this work is as follows.

Definition 4.1. *A binary memoryless channel $BMC(a, b)$ is a Discrete memoryless channel (DMC) with input X drawn from the alphabet $\mathcal{X} = \{0, 1\}$,output Y drawn from the alphabet $\mathcal{Y} = \{0, 1\}$. The transition probabilities are denoted by a and b, where $\mathbb{P}(Y = 0|X = 0) = a$ and $\mathbb{P}(Y = 1|X = 1) = b$.*

A binary memoryless channel is illustrated in Figure. 4.1.

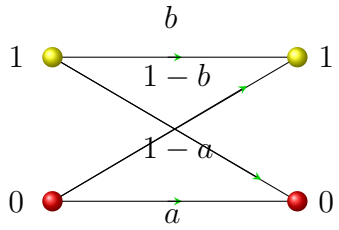


Figure 4.1: Binary channel

[1]For instance input alphabet 2 and output alphabet 3 is sometimes classified as a binary channel.

The well known binary symmetric channel (BSC) is a special case of BMC when the transition probabilities are symmetric with respect to the values 0 and 1. The classical $BSC(\epsilon)$ [16] is obtained by setting $a = b = 1-\epsilon$. That is, $BSC(\epsilon) = BMC(1-\epsilon, 1-\epsilon)$. Another special case of BMC is the Z-channel, which is obtained when one of the transition probabilities is set to 1 (That is, either $a = 1$ or $b = 1$).

4.0.1 Capacity of Binary asymmetric channel

The maximum rate R at which information can be reliably transmitted across a channel is upper bounded by its channel capacity. This value for BMC is computed below.

Let $I(a, b; u)$ denote the mutual information between $I(X;Y)$ the input X and output Y of a $BMC(a, b)$ with input distribution $\mathbb{P}(X = 0) = u$.The channel capacity is given by

$$C = \max_{\mathbb{P}_X} I(X;Y) = \max_{u} I(X;Y)$$

The mutual information $I(X;Y)$ is given by

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ H(Y) &= \mathbb{P}(Y=1)\log\frac{1}{\mathbb{P}(Y=0)} + \mathbb{P}(Y=1)\log\frac{1}{\mathbb{P}(Y=1)} \\ &= h(u(1-a) + (1-u)b) \\ H(Y|X) &= \mathbb{P}(X=0)H(Y|X=0) + \mathbb{P}(X=1)H(Y|X=1) \\ &= uh(a) + (1-u)h(b) \end{aligned}$$

where $h(u)$ is the binary entropy function given by,

$$h(u) = -u\log(u) - (1-u)\log(1-u), u \in [0,1].$$

Thus,

$$I(a, b; u) = h(u(1-a) + (1-u)p) - (uh(a) + (1-u)h(p))$$

We need to maximize this with respect to u to compute the channel capacity.

$$\begin{aligned} \frac{\partial I(a,b;u)}{\partial u} &= \frac{\partial}{\partial u} h(u(1-a) + (1-u)b) - (uh(a) + (1-u)h(b)) \\ &= h'(u(1-a) + (1-u)b)(1-b-a) - [h(a) - h(b)] \end{aligned}$$

Equating $\frac{\partial I(a,b;u)}{\partial u} = 0$ we get,

$$h'(u(1-a-b)+b) = \frac{h(b) - h(a)}{b + a - 1}$$

$$\begin{aligned} h'(x) &= \frac{d}{dx} - x\log(x) - (1-x)\log(1-x) \\ &= \log\left(\frac{1}{x} - 1\right) \end{aligned}$$

Hence,

$$h'(u(1-a-b)+b) = \log\left(\frac{1}{u(1-a-b)+b} - 1\right) = \frac{h(b)-h(a)}{b+a-1}$$

Upon simplification, we get the distribution $u^\star$which maximizes the mutual information (and this capacity achieving)

$$u^\star = \frac{b\left[1+e^{\frac{h(b)-h(a)}{b+a-1}}\right]-1}{(b+a-1)\left[1+e^{\frac{h(b)-h(a)}{b+a-1}}\right]}$$

and the corresponding mutual information is the channel capacity. That is., $C = I(a,b;u^\star)$. Note that, the optimum distribution (The optimum u which achieves the channel capacity) is a function of the channel parameters and hence is different for different channels. For binary symmetric channels $BSC(\epsilon), \epsilon \in [0,1]$, the optimum distribution is the uniform prior $u = \frac{1}{2}$ independent of the crossover transition probability ϵ. For other channels (other than BSCs), the capacity achieving prior is channel dependent. However, as we see will soon discover in section 4.1, for BMCs the optimum distribution is not too far from uniform distribution. This property helps in the design of an encoder for compound set of BMCs.

4.1 Compound Binary memoryless Channels

In a compound DMC setup, when the family of channels considered is further confined to be a class of binary channels (i.e.,BMCs), then we are referring to what is called Compound binary memoryless channels. The input and output alphabets are binary and the channel law can be anything chosen from the set $\mathcal{S} = \{W_k\}_1^K$ without transmitter and receiver knowing the selection. Each of the channels $\{W_k\}_1^K$ is a BMC. The cardinality K of the set $\mathcal{S}$ is assumed to be finite. As stated earlier, once picked at the beginning of communication, the channel law remain fixed for the entire duration of communication.

One of the main difficulty in designing the encoder and decoders in the compound setting is finding a single source distribution which work good for all channels in the set. While it may be easier to find the capacity achieving distributions for individual channels, identifying one which is universally good for all channels is significantly harder. We look into this problem for the compound set of binary channels in the following section. The objective is to

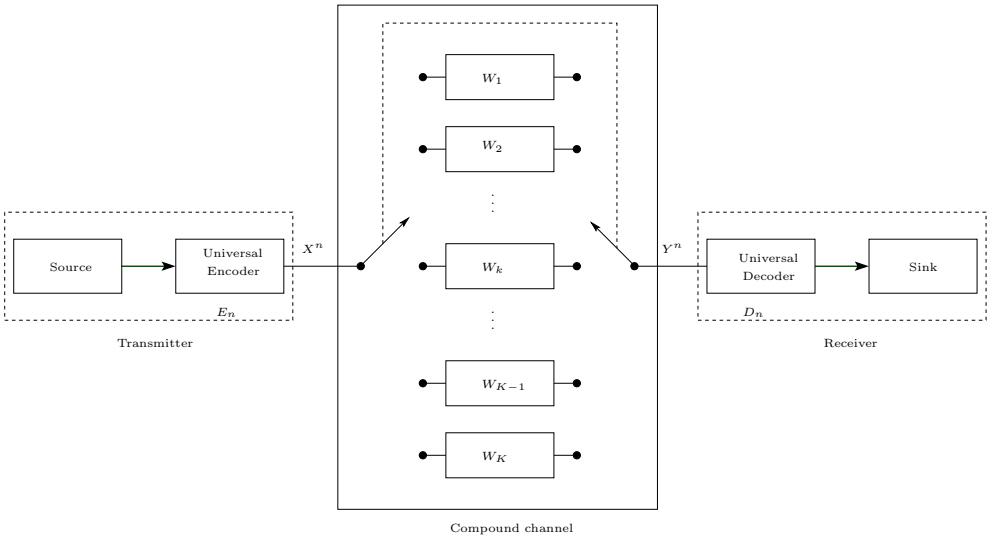


Figure 4.2: Compound channels: In the compound channel setting, the transmitter and receiver are ignorant about the actual channel through which communication take place. The only information available to the two entities is the possible list of candidate channels from the set of channels. This set is known as the compound set. Thus, the coding and decoding problem for such a setup involve designing a scheme independent of the underlying channel law.

find a single prior of probabilities that can be used for all binary channels in the set. The universal code design will then be done based on this obtained distribution.

4.1.1 Compound capacity achieving source distribution

The problem of finding the universal source distribution can be formulated as follows:

$$P_{opt} = \arg\max_{P} \inf_{W \in \mathcal{S}} \frac{I(P, W)}{C(W)}. \tag{4.1}$$

The optimum distribution is one which maximizes the maximum achievable rate for the worst possible channel. Here worst possible channel is referred to that channel which result in the least channel capacity. If such a P_{opt} exists, then mutual information of the worst channel under this source pair indeed provide the compound capacity of the set. Before solving the optimization problem of estimating the universal prior in the compound setting, it is worth considering other interesting criteria as a measure of *optimality*. An immediate choice is by considering the gap to capacity as an optimization cost. That is,

$$P_{opt} = \arg\min_{P} \sup_{W \in \mathcal{S}} I(P, W) - C(W).$$

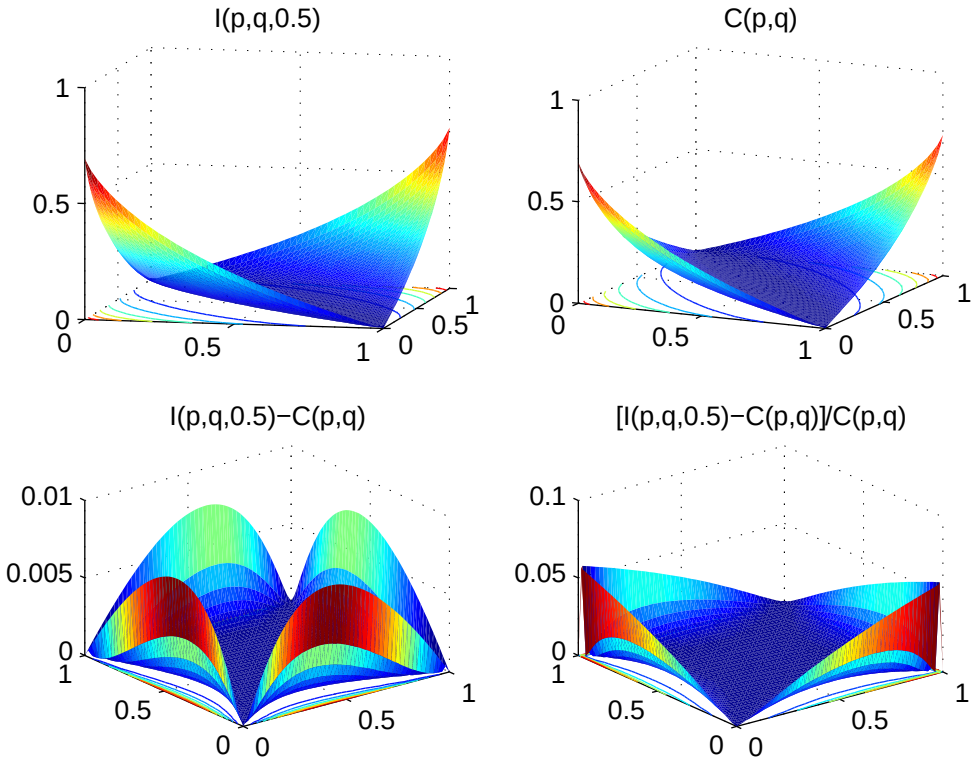


Figure 4.3: Binary channel: The comparative performance of optimum capacity achieving source distribution and that of a uniform distribution is illustrated for the class of binary channels. The worst possible loss of achievable rate is less than 5.8% percentage.

Even though not explicit from the formulation, it turns out that both the above criteria yield the same optimum source distribution. For our discussion we will stick to the former rule namely, the *max-min* approach as the rule for finding optimum source pair.

Shulman and Feder [24] have done a detailed investigation of this problem. They have found that, when the whole set of binary channels is considered, the optimum distribution of interest is simply the uniform prior. They have identified the worst channel in the binary compound setup as a limiting Z-channel. For this limiting channel one can achieve a rate more than 0.94208 of the corresponding channel capacity value, when the source is drawn from uniform distribution. The result is also a ramification of the fact that with uniform source distribution the maximum loss for any channel is less than 5.8% reported originally by Majani [25]. We state the two main results of interest from [24]

Theorem 4.2. *The uniform prior, over the class of all binary memoryless channels achieve compound capacity. Moreover, among all binary input chan-*

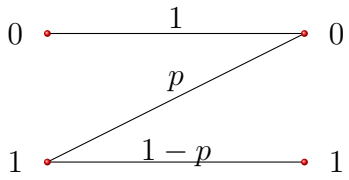


Figure 4.4: Z Channel:The transition probabilities are $\mathbb{P}(Y = 0|X = 0) = 1, \mathbb{P}(Y = 0|X = 1) = p$

nels with the same capacity, the Z-channel has the smallest mutual information, given uniform input distribution.

The computation of the optimum distribution for the limiting Z channel is presented below. We skip the detailed proof of the theorem [24].

4.1.2 Optimum priors on Z channel

The Z channel capacity is

$$\begin{aligned} C &= \max_u I(X;Y) \\ &= \max_u h(up) - uh(p) \end{aligned}$$

Solving $\frac{\partial}{\partial u}\left[h(up) - uh(p)\right] = 0$, we get the capacity achieving input prior $u^\star$ and it is,

$$u^\star = \frac{p^{\frac{p}{1-p}}}{1 + (1-p)p^{\frac{p}{1-p}}}$$

The worst possible Z channel is in the limit $p \to 1$. We find the limit:

$$\begin{aligned} \lim_{p\to 1} u(p) &= \lim_{p\to 1} \frac{p^{\frac{p}{1-p}}}{1 + (1-p)p^{\frac{p}{1-p}}} \\ &= \frac{1}{e} \end{aligned}$$

The achievable rate with uniform prior for this limiting channel can be computed as,

$$\alpha_2 = \lim_{p\to 1} \frac{I\left(Z_p; \frac{1}{2}\right)}{I\left(Z_p; \frac{1}{e}\right)} = \frac{e}{2\log_2 e} = 0.9420847.$$

In Figure 4.3 a comparison between the achievable rates with uniform distribution and that with the optimum individual capacity achieving priors is illustrated. The mutual information with uniform source distribution is computed for all the binary channels and is sketched along with the channel capacities. The gap to capacity and the percentage loss of achievable rates are also illustrated.

Even though we quoted the results only for binary channels, the results presented in [24] applies to other alphabets as well. Whereas the maximum loss is less than 5.8% of the capacity when operating on binary alphabet, the corresponding loss with uniform prior for a larger alphabet channel is significant. In fact, it is conjectured that the achievable rates α_A scale down inversely with the alphabet size A.

$$\alpha_A = \frac{e}{A \log_2 e}.$$

This behaviour is depicted in Figure. 4.5

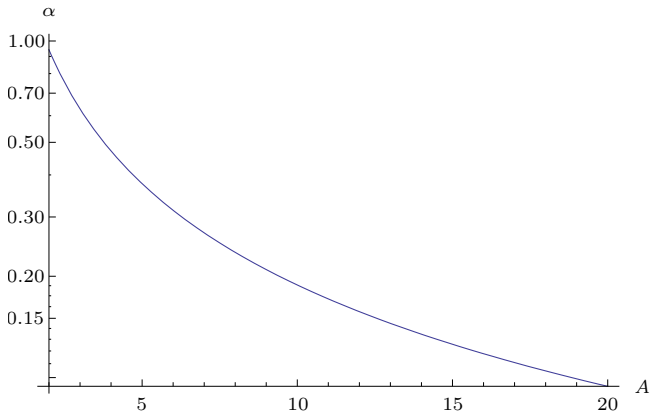


Figure 4.5: Worst case achievable rate with uniform input distribution; Plot shows the achievable rate with the input alphabet $\mathcal{A}$. As the alphabet size increases, the achievable rate decreases inversely. That is $\alpha_A = \frac{e}{A \log_2 e}$

When the problem formulation is restricted to an arbitrary compound set, as opposed to the set whole binary channels, the universal prior may be different than the uniform. Nevertheless, the fact that the worst case loss percentage is utmost 5.792%, it is well justified to fix the source distribution as uniform for the universal coding problem in the binary setup. Just to substantiate this claim in an empirical setup, our search over several random binary channels resulted a much lesser loss than the theoretical minimum.

4.2 Symmetric capacity $I(W)$ of a BMC W

By fixing the input distribution to be uniform we are bound to have a new benchmark on the highest achievable rate. This is the mutual information between the input X and output Y of the channel W when the source prior $\mathbb{P}_X(x)$ is uniform. This is denoted as $I(W)$ and is often referred to as the *symmetric capacity* of the channel W. Recall that we are considering binary channels in this chapter. As discussed earlier, the gap between $I(W)$ and the capacity $C(W)$ is at most $0.058I(W)$. Next, we derive the expression for the symmetric capacity of arbitrary binary channel.

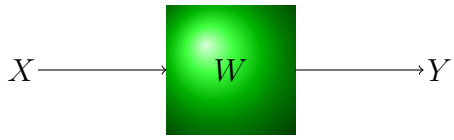


Figure 4.6: Mutual information and channel capacity

Consider a general communication channel model as shown in Figure. 4.6. The input X is drawn from a discrete alphabet $\mathcal{X}$ subject to a distribution P_X. Similarly the output Y represents a random variable with probability distribution P_Y. For a given realization x of the random variable X, the channel produce an output y as realization of Y, according to a probabilistic law $W = \mathbb{P}(y|x)$.

The mutual information $I(X;Y)$ between the input X and output Y quantifies the amount of information exchanged between the two sides of the channel. Mutual information can be computed from the entropy $H(Y)$ and the conditional entropy $H(Y|X)$ using the relationship $I(X;Y) = H(Y) - H(Y|X)$.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\ &= D\left(P_{XY}(x,y)\|P_X(x)P_Y(y)\right) \end{aligned}$$

where $D()$ is the Kullback Leibler divergence, which is a measure of the closeness between two stochastic distributions defined below.

Definition 4.3. *The Kullback Leibler distance (denoted by $D(P\|Q)$) between two distributions P and Q where both P and Q are defined over the same probability support space $supp(P)$ is defined as*

$$D(P\|Q) = \sum_{u\in supp(P)} P(u)\log\frac{P(u)}{Q(u)}. \tag{4.2}$$

Kullbeck-Leibler distance is sometimes referred to as Kullbeck-Leibler divergence or KL distance in short. Even though it is not strictly a metric in a normed vector space, it possess some interesting properties of a metric when viewed in a probability space. KL distance is defined when the underlying support space is the same for the two distributions. If the support space is different for P and Q, the KL divergence between them is undefined.

As a standard notation used throughout this report (much in consistent with the original usage of it by the authors in [21] and [22]) μ denotes the joint

distribution of the input and output of the channel W whereas μ^p stands for the product distribution between them.

$$\begin{aligned}\mu &\doteq P_{X,Y}(x,y) = P_X(x)W(y|x)\\ \mu^p &\doteq P_X(x)P_Y(y) = P_X(x) \circ W(y|x).\end{aligned}$$

Using this simplified notation, the mutual information $I(X;Y)$ can be expressed as

$$I(W) \doteq D(\mu \| \mu^p).$$

Since uniform prior as a universal input distribution is considered, it is significant to compute the mutual information when input is drawn from a uniform probability distribution. This brings in the notion of symmetric capacity of a channel.

Definition 4.4. *The symmetric capacity denoted as $I(W)$ of a channel W is the maximum mutual information between the input and output of the channel when the input is set drawn from uniform distribution.*

From the mutual information expression, by simply substituting the input distribution to be uniform, we can get the symmetric capacity $I(W)$ of a channel W.

Consider the BMC channel W as illustrated in Figure. 4.1. The symmetric capacity is given by,

$$I(W) \doteq D(\mu_u \| \mu_u^p)$$

where μ_u and μ_u^p are the joint and product distributions of input and output when input is set to uniform distribution. For BMC, we can express μ_u and μ_u^p as.

$$\begin{aligned}\mu_u &\doteq \frac{1}{2}W(y|x)\\ \mu_u^p &\doteq \frac{1}{2}(\delta(x-0)+\delta(x-1)) \circ W(y|x).\end{aligned}$$

To reduce the notational explosion, we use μ and μ^p to denote μ_u and μ_u^p in the rest of the report.

For a $BMC(a,b)$ we can now compute the symmetric capacity as below.

$$\begin{aligned}D(\mu\|\mu^p) &= D\left(P_X(x)W(y|x) \,\|P_X(x)\sum_{x'} P_X(x')W(y|x')\right)\\ &= D\left(P_X(x)W(y|x) \,\|P_X(x)\sum_{x'} P_X(x')W(y|x')\right)\\ &= \sum_{x,y} P(x)W(y|x)\log\left(\frac{P(x)W(y|x)}{P(x)\sum_{x'}P(x')W(y|x')}\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{x,y} W(y|x)\log\left(\frac{2W(y|x)}{W(y|0)+W(y|1)}\right)\\
&= \frac{1}{2}W(0|0)\log\left(\frac{2W(0|0)}{W(0|0)+W(0|1)}\right)+\frac{1}{2}W(0|1)\log\left(\frac{2W(0|1)}{W(0|0)+W(0|1)}\right)\\
&\quad+\frac{1}{2}W(1|0)\log\left(\frac{2W(1|0)}{W(1|0)+W(1|1)}\right)+\frac{1}{2}W(1|1)\log\left(\frac{2W(1|1)}{W(1|0)+W(1|1)}\right)\\
&= \frac{1}{2}a\log\left(\frac{2a}{1+a-b}\right)+\frac{1}{2}(1-b)\log\left(\frac{2(1-b)}{1+a-b}\right)\\
&\quad+\frac{1}{2}(1-a)\log\left(\frac{2(1-a)}{1-a+b}\right)+\frac{1}{2}b\log\left(\frac{2b}{1-a+b}\right).
\end{aligned}
$$

The symmetric capacity $I(W)$ of a $BMC(a,b)$ is then,

$$
\begin{aligned}
I(W) &= \frac{1}{2}a\log\left(\frac{2a}{1+a-b}\right)+\frac{1}{2}(1-b)\log\left(\frac{2(1-b)}{1+a-b}\right)\\
&\quad+\frac{1}{2}(1-a)\log\left(\frac{2(1-a)}{1-a+b}\right)+\frac{1}{2}b\log\left(\frac{2b}{1-a+b}\right)
\end{aligned}
$$

It is interesting to look into the geometry of the joint μ_u and product distributions μ_u^p in the probability space of a binary channel. It is shown in Figure. 4.7 and Figure. 4.8.

4.2.1 Mismatched decoding

In order to achieve the highest rate of information transfer (the Shannon capacity) of a Discrete Memoryless Channel, it is necessary to employ an optimum decoding rule. Optimum decoding rule such as Maximum Likelihood (ML) requires the knowledge of the channel law. When the decoder is ignorant of the exact channel law, a sub optimum decoder needs to be used. Such a sub optimum decoder will be using a decoding metric not necessarily matched to the channel and for this reason, it is known as mismatched decoding. This problem has been studied extensively in [35], [34] and recently in [37].

It is clear that, by using a mismatched decoder the achievable limit of information transfer is compromised. For instance, an ML decoder tuned to a channel W_1 when the actual channel is W_0 would result in a reduction of the achievable rate. So far, we have only stated the achievability aspects of mismatched decoders in qualitative terms, but the following lemma [6] will assert this claim quantitatively.

Let X, Y be finite sets, P_X, P_Y be the probability distributions on X and Y respectively. We consider a discrete memoryless channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $\mathbb{P}(y|x)$. We generate a code book with N codewords of length n, $\mathcal{C}(n) = \{x_1, \ldots, x_N\}$, drawn i.i.d. according to P_X^n. We denote by P_Y, the induced marginal distribution on Y, i.e. $P_Y(y) = \sum_{x\mathcal{X}} P_{Y|X}(y|x)P_X(x)$. Therefore, if a codeword, say x_1, is

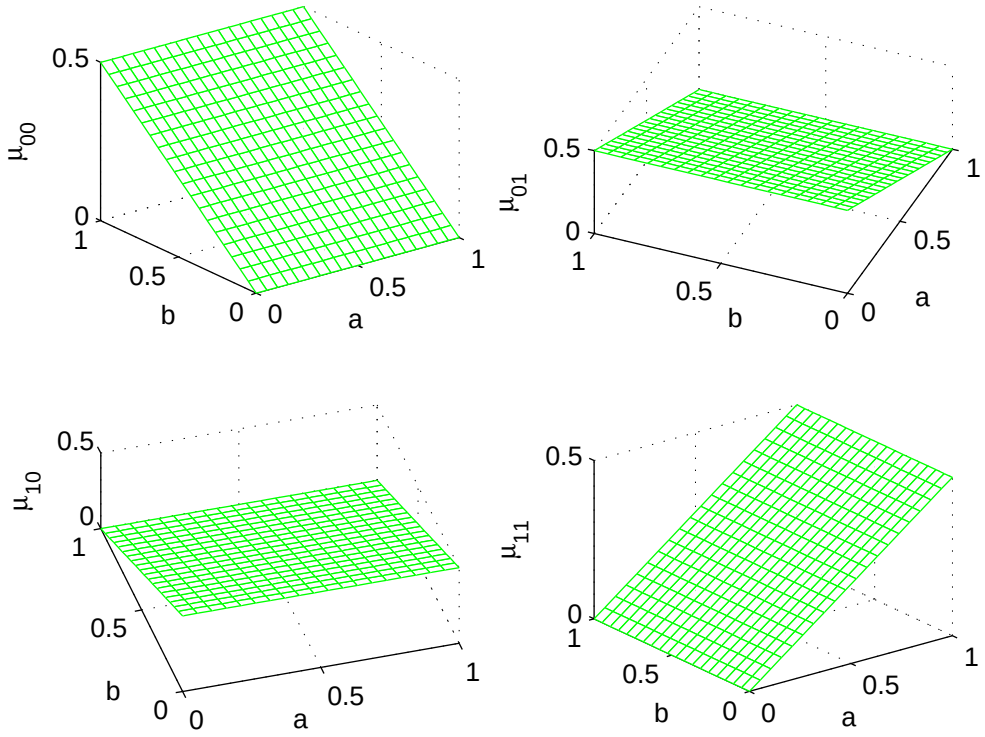


Figure 4.7: Illustration of μ and μ^p for binary channels: Individual components are shown.

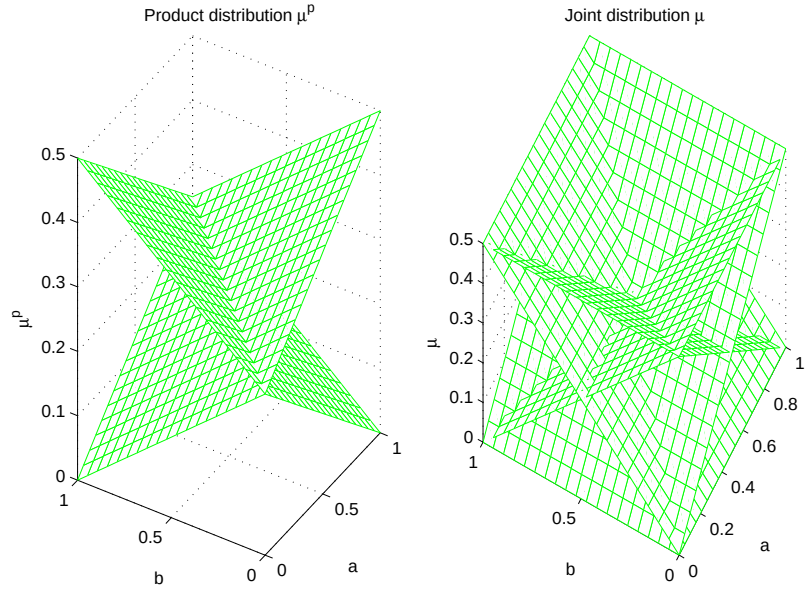


Figure 4.8: Illustration of μ and μ^p for binary channels.

transmitted and if y is the received message, the joint distribution of (x_1, y) is given by $P_{Y|X} \circ P_X$, which we will also denote by μ , and the joint distribution of (x_i, y) for $i \neq 1$ is given by $P_X \times P_Y$, which we will also denote by μ^p.

Upon receiving y, the decoder find a codeword x_i that maximize a given score function $f(x_i, y)$,

$$\hat{x} = \arg\max_i f(x_i, y)$$

Theorem 4.5. *Let X, Y denote the input and output of a discrete memoryless channel W_0. If the decoding rule uses a mismatched metric d, then using a random codebook it is possible to achieve a rate R given by,*

$$R(P_X, W_0, d) = \inf_{\mu \in \mathcal{A}} \quad D(\mu \| \mu_0^p) \tag{4.3}$$

where

$$\mathcal{A} = \{\mu : \mu_X = P_X, \mu_Y = (\mu_0)_Y, \mathbb{E}_\mu[d] \geq \mathbb{E}_{\mu_0}[d]\}$$

and $\mu = P_X \circ W(y|x)$ represents the joint distribution of (X, Y) when the channel is W and $\mu_0 = P_X \circ W_0 Y|X$ is the corresponding distribution when the channel is W_0. The product distribution of the input and outputs of the channel W is $\mu^p = P_X \circ P_Y$. Similarly μ_0^p is the product distribution when the channel is W_0.

4.3 Compound Capacity of BMCs

The compound capacity for a class of BMCs is defined analogous to the case of compound DMCs. We fix the source distribution to be the universal input distribution, which we have seen that is indeed the uniform prior, for the set of all BMCs. The optimum distribution for an arbitrary subset of BMCs may be different from uniform, but we know that, the achievable rate is fairly close to the optimum. In fact, the percentage loss suffered by choosing uniform input distribution is less than 5.8%.

We define the term *compound symmetric capacity* to refer the largest rate possible over a compound set when the input distribution is uniform.

Definition 4.6. *Given a set S of $BMC(a, b), a, b \in [0, 1]$ channels, the compound symmetric capacity is defined as the infimum of the individual symmetric capacities of the component channels when the input distribution is fixed to uniform. Clearly compound symmetric capacity is upper bounded by the compound capacity of any set.*

$$C_{\mathcal{S}}(W) = \inf_{W \in \mathcal{DMC}} I(W) = \inf_{W \in \mathcal{DMC}, P_X(0) = \frac{1}{2}} D(\mu \| \mu^p)$$

4.4 Generalized Linear Decoders

In chapter 3 (see section ?? 3.2) we discussed generalized linear decoders and their achievable limits. Generalized decoders generalizes the concept of a mismatched decoder, by introducing multiple (mismatched) metrics. The largest rate possible over a binary memoryless channel, with a generalized linear decoder is summarized in the following lemma.

Lemma 4.7. *For a BMC W_0, using a random codebook drawn from an input distribution P_X and a linear decoder induced by a finite number K of single letter metric $\{d_k\}_{k=1}^K$, the following data rate can be achieved.*

$$R\left(P_X, W_0, \{d_k\}_{k=1}^K\right) = \inf_{\mu \in \mathcal{A}} \quad D\left(\mu \| \mu_0^p\right) \tag{4.4}$$

where

$$\mathcal{A} = \left\{ \mu : \mu_X = P_X, \mu_Y = (\mu_0)_Y \, , \bigvee_{k=1}^K \mathbb{E}_\mu[d_k] \geq \bigvee_{k=1}^K \mathbb{E}_{\mu_0}[d_k] \right\}$$

We will investigate the generalized linear decoder performance for the binary memoryless channels in more details. Let us look at each of the constraints in the rate formula stated in Eq. 4.5.

4.4.1 Linear Decoder Constraints

One of the constraint in the optimization term in Eq. 4.5 is

$$\bigvee_{k=1}^K \mathbb{E}_\mu[d_k] \geq \bigvee_{k=1}^K \mathbb{E}_{\mu_0}[d_k]$$

First, let us look at the constituent terms, namely $\mathbb{E}_\mu[d_k]$ and $\mathbb{E}_{\mu_0}[d_k]$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mu_0}[d_k] &= \sum_y \frac{1}{2}\left(W_0(y|0)d_k(0,y) + W_0(y|1)d_k(1,y)\right) \\
&= \sum_y \frac{1}{2}\left(W_0(y|0)\log\left(\frac{W_k(y|0)}{\sum_{x'} P(x')W_k(y|x')}\right) + W_0(y|1)\log\left(\frac{W_k(y|1)}{\sum_{x'} P(x')W_k(y|x')}\right)\right) \\
&= \frac{1}{2}\left(W_0(0|0)\log\left(\frac{W_k(0|0)}{\sum_{x'} P(x')W_k(0|x')}\right) + W_0(0|1)\log\left(\frac{W_k(0|1)}{\sum_{x'} P(x')W_k(0|x')}\right)\right) \\
&\quad + \frac{1}{2}\left(W_0(1|0)\log\left(\frac{W_k(1|0)}{\sum_{x'} P(x')W_k(1|x')}\right) + W_0(1|1)\log\left(\frac{W_k(1|1)}{\sum_{x'} P(x')W_k(1|x')}\right)\right) \\
&= \frac{1}{2}W_0(0|0)\log\left(\frac{W_k(0|0)}{P(0)W_k(0|0) + P(1)W_k(0|1)}\right) \\
&\quad + \frac{1}{2}W_0(0|1)\log\left(\frac{W_k(0|1)}{P(0)W_k(0|0) + P(1)W_k(0|1)}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}W_0(1|0)\log\left(\frac{W_k(1|0)}{P(0)W_k(1|0)+P(1)W_k(1|1)}\right)\\
&+\frac{1}{2}W_0(1|1)\log\left(\frac{W_k(1|1)}{P(0)W_k(1|0)+P(1)W_k(1|1)}\right)\\
=\;&\frac{1}{2}\left(W_0(0|0)\log\left(\frac{2W_k(0|0)}{W_k(0|0)+W_k(0|1)}\right)+W_0(0|1)\log\left(\frac{2W_k(0|1)}{W_k(0|0)+W_k(0|1)}\right)\right)\\
&+\frac{1}{2}\left(W_0(1|0)\log\left(\frac{2W_k(1|0)}{W_k(0|1)+W_k(1|0)}\right)+W_0(1|1)\log\left(\frac{2W_k(1|1)}{W_k(1|0)+W_k(1|1)}\right)\right)\\
=\;&\frac{1}{2}\left[a_0\log\left(\frac{2a_k}{1+a_k-b_k}\right)+(1-b_0)\log\left(\frac{2\,(1-b_k)}{1+a_k-b_k}\right)\right]+\\
&\frac{1}{2}\left[(1-a_0)\log\left(\frac{2\,(1-a_k)}{1+b_k-a_k}\right)+b_0\log\left(\frac{2b_k}{1+b_k-a_k}\right)\right].
\end{aligned}
$$

Similarly, we can derive $\mathbb{E}_{\mu_0}[d_k]$. The two formulae are summarized below along with the corresponding constraint.

$$\max_{k=1\ldots K}\mathop{\mathbb{E}}_{\mu}[d_k]\geq\max_{k=1\ldots K}\mathop{\mathbb{E}}_{\mu_0}[d_k]$$

where,

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mu_0}[d_k]\;=\;&\frac{1}{2}\left[a_0\log\left(\frac{2a_k}{1+a_k-b_k}\right)+(1-b_0)\log\left(\frac{2\,(1-b_k)}{1+a_k-b_k}\right)\right]\\
&+\frac{1}{2}\left[(1-a_0)\log\left(\frac{2\,(1-a_k)}{1+b_k-a_k}\right)+b_0\log\left(\frac{2b_k}{1+b_k-a_k}\right)\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mu}[d_k]\;=\;&\frac{1}{2}\left[a\log\left(\frac{2a_k}{1+a_k-b_k}\right)+(1-b)\log\left(\frac{2\,(1-b_k)}{1+a_k-b_k}\right)\right]\\
&+\frac{1}{2}\left[(1-a)\log\left(\frac{2\,(1-a_k)}{1+b_k-a_k}\right)+b\log\left(\frac{2b_k}{1+b_k-a_k}\right)\right]
\end{aligned}
$$

The other pair of constraints in the rate equation are $\mu_X = P_X$ and $\mu_Y = (\mu_0)_Y$. They can be expressed in terms of the channel parameters a and b of a BMC. Recall that a and b are the transition probabilities of a BMC, where $\mathbb{P}(Y=0|X=0)=a$ and $\mathbb{P}(Y=1|X=1)=b$.

$$
\begin{aligned}
(\mu)_Y\;&=\;\mathbb{E}_X[\mu]\\
&=\;\mathbb{E}_X[P_X(x)W(y|x)]\\
&=\;\sum_{x\in\mathcal{X}}P_X(x)W(y|x)\\
&=\;\frac{1}{2}\sum_{x\in\mathcal{X}}W(y|x)
\end{aligned}
$$

$$\begin{aligned}(\mu_0)_Y &= \mathbb{E}_X[\mu_0] \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}} W_0(y|x)\end{aligned}$$

The marginal constraint $\mu_Y = (\mu_0)_Y$ is then given by,

$$\sum_{x\in\mathcal{X}} W(y|x) = \sum_{x\in\mathcal{X}} W_0(y|x)$$

This, for a binary channel becomes,

$$\begin{aligned}W(y|0) + W(y|1) &= W_0(y|0) + W_0(y|1) \\ a + 1 - b &= a_0 + 1 - b_0 \\ a - b &= a_0 - b_0\end{aligned}$$

for $y \in \{0, 1\}$

Finally, we can express $D(\mu\|\mu_0^p)$ in terms of the channel transition probabilities.

$$D(\mu\|\mu_0^p) = D\left(P_X(x)W(y|x)\,\|P_X(x)\sum_{x'} P_X(x')W_0(y|x')\right)$$

$$\begin{aligned}&= D\left(P_X(x)W(y|x)\,\|P_X(x)\sum_{x'} P_X(x')W_0(y|x')\right) \\ &= \sum_{x,y} P(x)W(y|x)\log\left(\frac{P(x)W(y|x)}{P(x)\sum_{x'} P(x')W(y|x')}\right) \\ &= \sum_{x,y} W(y|x)\log\left(\frac{2W(y|x)}{W(y|0)+W(y|1)}\right) \\ &= W(0|0)\log\left(\frac{2W(0|0)}{W_0(0|0)+W_0(0|1)}\right) + W(0|1)\log\left(\frac{2W(0|1)}{W_0(0|0)+W_0(0|1)}\right) \\ &\quad + W(1|0)\log\left(\frac{2W(1|0)}{W_0(1|0)+W_0(1|1)}\right) + W(1|1)\log\left(\frac{2W(1|1)}{W_0(1|0)+W_0(1|1)}\right) \\ &= a\log\left(\frac{2a}{1+a_0-b_0}\right) + (1-b)\log\left(\frac{2(1-b)}{1+a_0-b_0}\right) \\ &\quad + (1-a)\log\left(\frac{2(1-a)}{1-a_0+b_0}\right) + b\log\left(\frac{2b}{1-a_0+b_0}\right)\end{aligned}$$

Using the derived expressions, we can express the achievable rate R under mismatched setting over a binary memoryless channel, in terms of the channel parameters.

$$R = \inf_{a,b\in\Lambda} f(a, b, a_0, b_0)$$

where,

$$\begin{aligned} f\left(a,b,a_0,b_0\right) &= a\log\left(\frac{2a}{1+a_0-b_0}\right)+(1-b)\log\left(\frac{2(1-b)}{1+a_0-b_0}\right)+ \\ &\quad (1-a)\log\left(\frac{2(1-a)}{1-a_0+b_0}\right)+b\log\left(\frac{2b}{1-a_0+b_0}\right) \end{aligned}$$

and

$$\Lambda=\left\{a,b|a-b=a_0-b_0,\max_{k=1\ldots K} g\left(a_j,b_j,a,b\right)\geq \max_{j=1\ldots K} g_0\left(a_j,b_j,a_0,b_0\right)\right\}.$$

The functional g and g_0 are the expressions for $\underset{\mu_0}{\mathbb{E}}\left[d_k\right]$ and $\underset{\mu}{\mathbb{E}}\left[d_k\right]$ computed earlier. They are defined as,

$$\begin{aligned} g\left(a,b,a_j,b_j\right) &= \frac{1}{2}\left[a_0\log\left(\frac{2a_k}{1+a_k-b_k}\right)+(1-b_0)\log\left(\frac{2\left(1-b_k\right)}{1+a_k-b_k}\right)\right] \\ &\quad +\frac{1}{2}\left[(1-a_0)\log\left(\frac{2\left(1-a_k\right)}{1+b_k-a_k}\right)+b_0\log\left(\frac{2b_k}{1+b_k-a_k}\right)\right] \\ g_0\left(a_j,b_j,a_0,b_0\right) &= \frac{1}{2}\left[a\log\left(\frac{2a_k}{1+a_k-b_k}\right)+(1-b)\log\left(\frac{2\left(1-b_k\right)}{1+a_k-b_k}\right)\right] \\ &\quad +\frac{1}{2}\left[(1-a)\log\left(\frac{2\left(1-a_k\right)}{1+b_k-a_k}\right)+b\log\left(\frac{2b_k}{1+b_k-a_k}\right)\right] \end{aligned}$$

We can summarise it into the following lemma.

Lemma 4.8. *Let $W(a,b)$ denote a binary memoryless channel with transition probabilities $1-a$ and $1-b$ where $\mathbb{P}(Y=0|X=0)=a$ and $\mathbb{P}(Y=1|X=1)=b$. Using a random codebook drawn from an input distribution P_X and a linear decoder induced by a finite number K of single letter metric $\{d_k\}_{k=1}^K$, the following data rate can be achieved.*

$$\begin{aligned} R &= \inf_{\substack{a-b=a_0-b_0 \\ 0\leq a,b\leq 1 \\ \vee_{k=1}^K g(a,b,a_k,b_k)\geq \vee_{k=1}^K g(a_0,b_0,a_k,b_k)}} f\left(a,b,a_0,b_0\right) \\ &= \min_{k=1,\ldots,K} \inf_{\substack{a-b=a_0-b_0 \\ 0\leq a,b\leq 1 \\ g(a,b,a_k,b_k)\geq \vee_{k=1}^K g(a_0,b_0,a_k,b_k)}} f\left(a,b,a_0,b_0\right) \end{aligned}$$

where

$$\begin{aligned} f\left(a,b,a_0,b_0\right) &= a\log\left(\frac{2a}{1+a_0-b_0}\right)+(1-b)\log\left(\frac{2(1-b)}{1+a_0-b_0}\right)+ \\ &\quad (1-a)\log\left(\frac{2(1-a)}{1-a_0+b_0}\right)+b\log\left(\frac{2b}{1-a_0+b_0}\right) \end{aligned}$$

and

$$\begin{aligned} g\left(a,b,a_j,b_j\right) &= \frac{1}{2}\left[a_0\log\left(\frac{2a_k}{1+a_k-b_k}\right)+(1-b_0)\log\left(\frac{2\left(1-b_k\right)}{1+a_k-b_k}\right)\right] \\ &\quad +\frac{1}{2}\left[(1-a_0)\log\left(\frac{2\left(1-a_k\right)}{1+b_k-a_k}\right)+b_0\log\left(\frac{2b_k}{1+b_k-a_k}\right)\right] \end{aligned}$$

The performance of a mismatched decoder has direct value in the compound channel setup, where a single encoding and decoding strategy need to be devised for a set of channels, without assuming any explicit knowledge of the channel of communication.

4.4.2 Compound channels and linear decoders

Instead of a single channel W_0, if we consider a family of channels $\mathcal{S}$, this constitute the compound channel setup. As introduced earlier, the notion of compound channels assume that, the transmitter and receivers do not know the underlying channel law. Naturally, in such a case, the receiver will be using a mismatched decoder. From the mismatched rate established in the previous section, it is fairly straightforward to compute the compound channel capacity using a linear decoder. This will be the infimum of the mismatched rate over all possible channels W_0 in the set $\mathcal{S}$. The following lemma present the achievable rate over a compound set of DMCs using a set of mismatched metrics $\{d_k\}_{k=1}^K$.

Lemma 4.9. *For a family $\mathcal{S}$ of discrete memoryless channels W_0, using a random codebook drawn from an input distribution P_X and a linear decoder induced by a finite number K mismatched single letter metric $\{d_k\}_{k=1}^K$, the following data rate can be achieved.*

$$R\left(P_X,\mathcal{S},\{d_k\}_{k=1}^K\right)=\inf_{W_0\in\mathcal{S}}\inf_{\mu\in\mathcal{A}} \quad D\left(\mu\|\mu_0^p\right) \tag{4.5}$$

where

$$\mathcal{A}=\left\{\mu:\mu_X=P_X,\mu_Y=(\mu_0)_Y\,,\bigvee_{k=1}^K\mathbb{E}_\mu[d_k]\geq\bigvee_{k=1}^K\mathbb{E}_{\mu_0}[d_k]\right\}$$

In the last few sections, we discussed the largest possible rate over a binary memoryless channel, using a mismatched decoder. No explicit information on the functional relationship of the decoding metric to the channel law is assumed. We know that, the decoding metric for a maximum likelihood (ML) rule is the likelihood of the channel. As briefed earlier, such a likelihood metric is linear as well. It is an interesting question to consider the likelihood of a channel as a decoding metric and seek the achievable rate under mismatched decoding.

4.5 Decoders using likelihood:One sided channels

Abbe and Zheng [22] looked at the possibility of using a decoding metric similar to the maximum likelihood decoder. The idea is to use a single letter decoding metric as the likelihood (or log likelihood) of the channel rule[2]. As the channel rule is not assumed in the receiver, it rely on the log likelihood of a mismatched channel as the decoding metric. Question is, whether we can achieve the compound channel capacity using such a decoder based on the log likelihood of mismatched channel.

In general, it is not yet known whether a linear decoder based on the log likelihood can achieve compound capacity. Abbe and Zheng formulated a sufficient condition on the set in order to admit a linear log likelihood decoder which achieve compound capacity. They called such sets as *one sided*. They proved that when a set $\mathcal{S}$ is one sided, using the log-likelihood of the worst possible channel[3] compound capacity can be achieved.

They have extended the sufficiency condition beyond one sided sets by introducing a log likelihood type decoder using multiple metrics. Such a decoding scheme is called Generalized Maximum a Posteriori (GMAP) rule. It is proved that, a set which is a finite union of one sided sets is capacity achieving under a GMAP decoding rule. While the decoder for a (single) one sided set made use of log likelihood, the GMAP rule for union of one sided sets is not quite the same as Generalized Log Likelihood Ratio Test (GLRT) []. GMAP uses the log of the aposteriori distribution of the worst channels in each of the one sided subset in the set $\mathcal{S}$. GLRT on the other hand uses the log of the likelihood distribution (in place of the posteriori distribution) of the worst channels in each of the subsets.

Since our focus is mainly on the class of binary memoryless channels, we will do a detailed investigation on the characteristics of one sidedeness in the binary context. Moreover the binary setup offers us the luxury of geometrical and visualization in a two dimensional space spanned by the two crossover probabilities a and b of BMC.

Interestingly, as we will soon find out, the one sided notion in a binary setup brings in nicer characteristic regions.

4.5.1 One sided binary channels

Recall that a one sided set (see Def. 3.6) is,

A set $\mathcal{S}$ is one sided if

$$D\left(\mu_0 \| \mu_{\mathcal{S}}^p\right) \geq D\left(\mu_0 \| \mu_{\mathcal{S}}\right) + D\left(\mu_{\mathcal{S}} \| \mu_{\mathcal{S}}^p\right)$$

[2]Recall that, the log likelihood over a DMC becomes addition of single letter log likelihoods.

[3]Worst possible channel refers to the channel which has the least mutual information subject to the fixed input distribution.

where,

$$W_{\mathcal{S}} = \arg \min_{W \in cl(S)} I\left(P_X, W\right) \tag{4.6}$$

and $\mu_0 = P_X \circ W_0, \mu_S = P_X \circ W_S$, are the joint distributions over the channel W_0 and W_S respectively.

Closure of a set $cl(S)$ is the smallest closed set containing S.

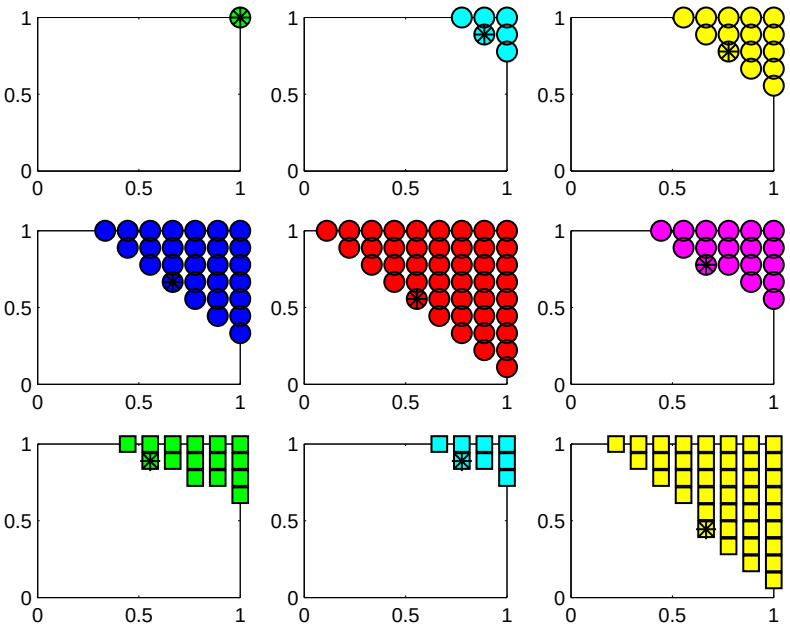


Figure 4.9: One sided region and the corresponding dominant channels.Here a uniform grid of binary channels is simulated and the identified one sided regions dominated by the respective channels picked are shown.

Since the region characterizing onesidedness is represented by inequality involving three divergence metrics, we compute those under the binary setup. First, recall that $D\left(\mu \| \mu^p\right)$ is just the mutual information $I\left(P_X, W\right)$ and it can be computed for the BMCs as follows:

$$\begin{aligned} D\left(\mu \| \mu^p\right) &= D\left(P_X(x)W\left(y|x\right) \| P_X(x) \sum_{x'} P_X(x')W\left(y|x'\right)\right) \\ &= D\left(P_X(x)W\left(y|x\right) \| P_X(x) \sum_{x'} P_X(x')W\left(y|x'\right)\right) \\ &= \sum_{x,y} P(x)W(y|x) \log\left(\frac{P(x)W(y|x)}{P(x)\sum_{x'} P(x')W(y|x')}\right) \end{aligned}$$

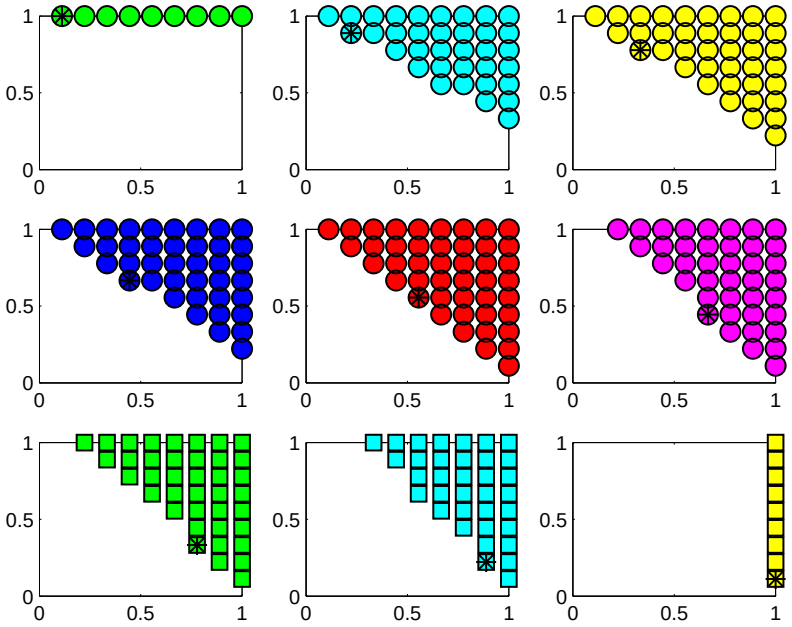


Figure 4.10: One sided region shown for various dominant channels. Here a uniform grid of binary channels is simulated and the identified one sided regions dominated by the respective channels picked are shown.

$$
\begin{aligned}
&= \sum_{x,y} W(y|x) \log\left(\frac{2W(y|x)}{W(y|0)+W(y|1)}\right) \\
&= W(0|0)\log\left(\frac{2W(0|0)}{W(0|0)+W(0|1)}\right) + W(0|1)\log\left(\frac{2W(0|1)}{W(0|0)+W(0|1)}\right) \\
&\quad + W(1|0)\log\left(\frac{2W(1|0)}{W(1|0)+W(1|1)}\right) + W(1|1)\log\left(\frac{2W(1|1)}{W(1|0)+W(1|1)}\right) \\
&= a\log\left(\frac{2a}{1+a-b}\right) + (1-b)\log\left(\frac{2(1-b)}{1+a-b}\right) \\
&\quad + (1-a)\log\left(\frac{2(1-a)}{1-a+b}\right) + b\log\left(\frac{2b}{1-a+b}\right)
\end{aligned}
$$

The channel $W_{\mathcal{S}}$ satisfying Eq. 4.6 is referred to as the dominant channel.

$$
\begin{aligned}
W_{\mathcal{S}} &= \arg\min_{a,b\in[0,1]} a\log\left(\frac{2a}{1+a-b}\right) + (1-b)\log\left(\frac{2(1-b)}{1+a-b}\right) \\
&\quad + (1-a)\log\left(\frac{2(1-a)}{1-a+b}\right) + b\log\left(\frac{2b}{1-a+b}\right)
\end{aligned}
$$

The other term $D(\mu_0\|\mu_{\mathcal{S}})$ has the following form and can be simplified further into a form involving the BMC parameters.

$$
D(\mu\|\mu_0) = D(P_X W\|P_X W_0)
$$

$$
\begin{aligned}
&= \sum_{x,y} P_X(x)W(y|x)\log\left(\frac{P_X(x)W(y|x)}{P_X W_0(y|x)}\right)\\
&= \frac{1}{2}\sum_{x,y} W(y|x)\log\left(\frac{W(y|x)}{W_0(y|x)}\right)\\
&= a\log\left(\frac{a}{a_0}\right)+b\log\left(\frac{b}{b_0}\right)+(1-a)\log\left(\frac{1-a}{1-a_0}\right)+(1-b)\log\left(\frac{1-b}{1-b_0}\right)
\end{aligned}
$$

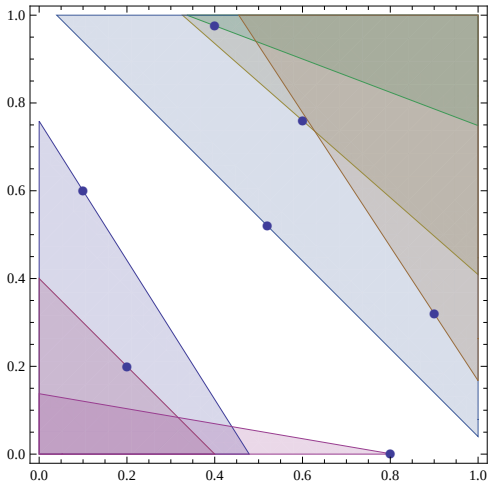


Figure 4.11: One sided regions of dominant channels

Using the expressions just derived for the different divergences involved in Eq. **??** we can write the one sidedness condition for binary memoryless channels as,

$$
\begin{aligned}
0 \;\leq\; & -\log\left[\frac{1}{1-a_s}\right]-\log\left[\frac{1}{1-b_s}\right]+\log\left[\frac{2}{1+a_s-b_s}\right]\\
& -a_s\log\left[\frac{2a_s}{1+a_s-b_s}\right]-\log\left[\frac{2-2b_s}{1+a_s-b_s}\right]\\
& +b_s\log\left[\frac{2-2b_s}{1+a_s-b_s}\right]+\log\left[\frac{2}{1-a_s+b_s}\right]-\log\left[\frac{2-2a_s}{1-a_s+b_s}\right]\\
& +a_s\log\left[\frac{2-2a_s}{1-a_s+b_s}\right]-b_s\log\left[\frac{2b_s}{1-a_s+b_s}\right]\\
& +b_0\left(\log\left[\frac{1}{1-b_s}\right]-\log\left[\frac{2}{1+a_s-b_s}\right]-\log\left[\frac{1}{b_s}\right]+\log\left[\frac{2}{1-a_s+b_s}\right]\right)\\
& +a_0\left(\log\left[\frac{2}{1-a_s}\right]-\log\left[\frac{1}{a_s}\right]+\log\left[\frac{2}{1+a_s-b_s}\right]-\log\left[\frac{2}{1-a_s+b_s}\right]\right).
\end{aligned}
$$

Upon simplification, this becomes,

$$
(a-a_s)\log\left(\frac{a_s}{1-a_s}\frac{1-a_s+b_s}{1+a_s-b_s}\right)+(b-b_s)\log\left(\frac{b_s}{1-b_s}\frac{1+a_s-b_s}{1-a_s+b_s}\right) \;\geq\; 0
$$

Slope $\eta(W_s)$ of the one sided region dominated by a chosen channel $W_s = (a_s, b_s)$ is then given by,

$$\eta(W_s) = \frac{\log\left(\frac{a_s}{1-a_s}\frac{1-a_s+b_s}{1+a_s-b_s}\right)}{\log\left(\frac{b_s}{1-b_s}\frac{1+a_s-b_s}{1-a_s+b_s}\right)} \tag{4.7}$$

Dominant region of a chosen channel $W_s = (a_s, b_s)$ is simply the region given by,

$$\Gamma_{\mathcal{S}}^{\text{onesided}}(W_s) = \left\{a, b | (a-a_s)\log\left(\frac{a_s}{1-a_s}\frac{1-a_s+b_s}{1+a_s-b_s}\right) + (b-b_s)\log\left(\frac{b_s}{1-b_s}\frac{1+a_s-b_s}{1-a_s+b_s}\right) \geq \right. \tag{4.8}$$

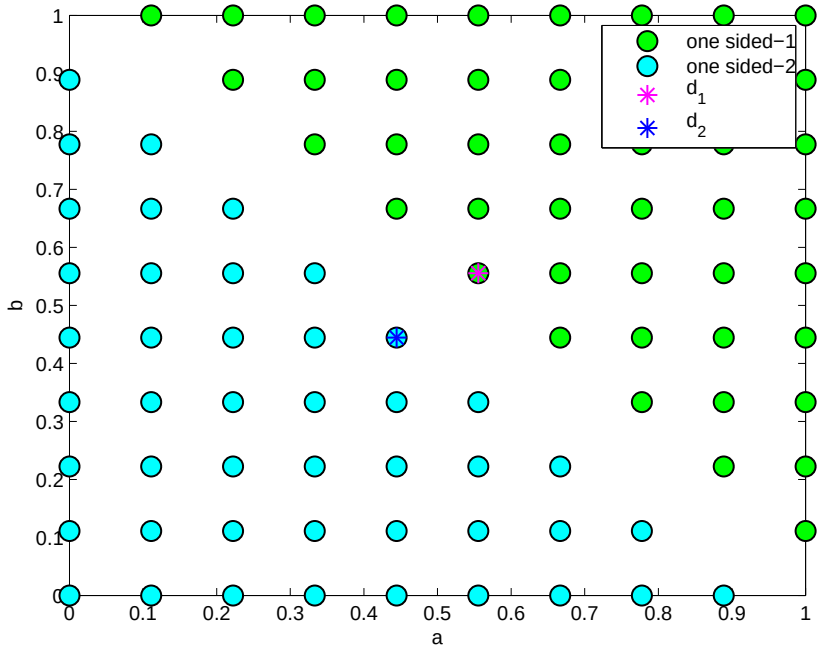


Figure 4.12: DMC: union of two one sided sets. The dominant channels in each of the one sided subsets is marked with $*$. The dominant channels in this case are binary symmetric channels.

4.6 Universality beyond union of one sets

The story of our journey with a quest to find linear decoders for compound DMC reach at at a point with the following milestones.

1. If the set is one sided, a linear decoder can be designed which in conjunction with a code can achieve the symmetric compound capacity

2. If a given compound set can be circumscribed by a union of several one sided sets (call it the super one sided set), we can design a generalized linear decoder which can achieve rate up to the worst possible channel (in terms of the symmetric capacity) in the super set. The rate thus achieved will (in general) be less than the symmetric compound capacity of the original set.

Clearly for arbitrary compound set, it is not known whether a linear decoder exist which can achieve the compound capacity. We shall now claim that compound BMC admit a linear decoder which is compound capacity achieving. In fact, for BMC, we can achieve all the way up to the symmetric capacity of respective channels. In this respect, we show that a linear decoder exist which meets the MMI performance in the case of binary memoryless channels.

We state the main result in the following theorem.

Theorem 4.10. *There exists a linear codebook $\mathcal{C}$ and a generalized linear decoder $\mathcal{D}$ such that by using this codebook $\mathcal{C}$ on any channel W, the symmetric capacity $I(W)$ is achieved.*

Over a compound set of binary memoryless channels, reliable communication can be achieved when the rate R is less than the compound capacity $C_{\mathcal{S}}$, using a linear universal decoder.

This theorem states that for the class of binary memoryless channels, universal linear decoders can indeed achieve the compound symmetric capacity. When the input is fixed to be uniform distribution, using a random codebook and a linear decoder, without any knowledge on the underlying channel in which transmission take place, one can achieve the maximum possible rate, the capacity.The additive nature of the decoder is helpful in simpler implementation of the receive circuitry.

Lemma 4.11. *With Generalized Maximum Aposteriory (GMAP) rule, induced by two single letter metrics d_1, d_2,*

$$\begin{aligned} d_1(y,x) &= \log \frac{W_1(y|x)}{W_1(y|0)+W_1(y|1)}, W_1 = BMC(a_1,a_1) \\ d_2(y,x) &= \log \frac{W_2(y|x)}{W_2(y|0)+W_2(y|1)}, W_2 = BMC(a_2,a_2), \end{aligned}$$

the compound symmetric capacity can be achieved for any binary memoryless channel W. Moreover, the two metrics d_1 and d_2 satisfy complementary symmetric property $a_1 + a_2 = 1$.

The induced single letter metrics actually correspond to complementary BSC channels. Except the pure noise BSC channel (i.e., $BSC(1/2)$) any other pair of such pair can be chosen as a decoding metric.

Corollary 4.12. *For the class of binary memoryless channels, when the input is drawn from uniform distribution, linear decoder can achieve MMI performance.*

Proposition 4.13. *For BMC W_0, a mismatched decoder tuned to one of the channels $W_1 = (a_1, a_1)$ or $W_2 = (1 - a_1, 1 - a_1)$ can achieve the capacity.*

Let us consider two regions of the BMC grid. Let is call $\mathcal{B}^+(a, b)$ be the region with $a+b \leq 1$ and $\mathcal{B}^-(a, b)$ corresponding to channels satisfying $a+b \geq 1$. The above proposition says that, the mismatched rate with a decoder tuned to a BSC from either $\mathcal{B}^+$ or $\mathcal{B}^-$ is equal to the maximum achievable rate with a decoder tuned to the true channel w_0. In fact, the true channel rate is achieved when the channel (BSC) corresponding to the decoder and the true channel both belong to the same channel regions. If $W_1 \in \mathcal{B}^+$ and $W_0 \in \mathcal{B}^+$ or $W_1 \in \mathcal{B}^-$ and $W_0 \in \mathcal{B}^{R-}$ then, the true rate is achieved, otherwise, the mismatched rate can be zero.

Lemma 4.14. *For any BMC $W_0 \in \mathcal{B}^+$, with a mismatched decoder tuned to a BSC W_1 from the same region, i.e., $W_1 \in \mathcal{B}^+$ achievable rate is given by*

$$I_{mis}(W_0, W_1) = D(\mu \| \mu_0^p)$$

Proof.

$$\begin{aligned}
I_{\text{mis}}(W_0, W_1) &= \inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu[d_1] \geq \mathbb{E}_{\mu_0}[d_1]}} D(\mu \| \mu_0^p) \\
&= \inf_{\substack{\mu:\frac{a-a_0}{b-b_0}=\frac{1+a_0-b_0}{1+b_0-a_0} \\ a+b \leq a_0+b_0, a_1 \leq \frac{1}{2}}} D(\mu \| \mu_0^p) \\
&= \begin{cases} \inf\limits_{\substack{\mu:\frac{a-a_0}{b-b_0}=\frac{1+a_0-b_0}{1+b_0-a_0} \\ a+b \leq a_0+b_0, a_1 \leq \frac{1}{2}}} D(\mu \| \mu_0^p), & W_1 \in \mathcal{B}^+ \\ \inf\limits_{\substack{\mu:\frac{a-a_0}{b-b_0}=\frac{1+a_0-b_0}{1+b_0-a_0} \\ a+b \geq a_0+b_0, a_1 \geq \frac{1}{2}}} D(\mu \| \mu_0^p), & W_1 \in \mathcal{B}^- \end{cases} \\
&= \begin{cases} D(\mu_0 \| \mu_0^p) & a_1 \leq \frac{1}{2} \\ 0 & a_1 \geq \frac{1}{2} \end{cases} \\
&= D(\mu_0 \| \mu_0^p)
\end{aligned}$$

$\mu^p = \mu_0^p$ translate to the following condition form BMCs.

$$\frac{a - a_0}{b - b_0} = \frac{1 + a_0 - b_0}{1 + b_0 - a_0}$$

and the inequality $\mathbb{E}_\mu[d_1] \geq \mathbb{E}_{\mu_0}[d_1]$ to $a + b \leq a_0 + b_0$.

□

We now prove the main theorem:

Proof. We need to prove,

$$\inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu d_1 \vee \mathbb{E}_\mu d_2 \geq \mathbb{E}_{\mu_0} d_1 \vee \mathbb{E}_{\mu_0} d_2}} D(\mu \| \mu_0^p) = D(\mu_0 \| \mu_0^p)$$

Note that,

$$\mu_0^p = \tilde{\mu}_0^p.$$

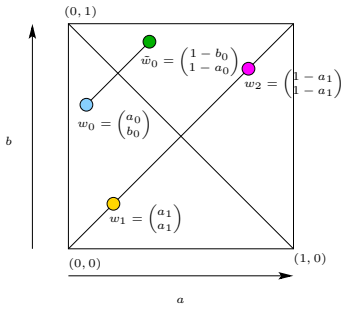


Figure 4.13: Complementary channels in the DMC set

w.l.o.g assume $a < \frac{1}{2}$ and $a_0 + b_0 < 1$, then

$$\mathbb{E}_{\mu_0}[d_1] \vee \mathbb{E}_{\mu_0}[d_2] = \mathbb{E}_{\mu_0}[d_1]$$

$$\begin{aligned}
\inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu[d_1]\vee\mathbb{E}_\mu[d_2]\geq\mathbb{E}_{\mu_0}[d_1]\vee\mathbb{E}_{\mu_0}[d_2]}} D(\mu\|\mu_0^p) &= \inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu[d_1]\vee\mathbb{E}_\mu[d_2]\geq\mathbb{E}_{\mu_0}[d_1]}} D(\mu\|\mu_0^p) \\
&= \min\left(\inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu[d_1]\geq\mathbb{E}_{\mu_0}[d_1]}} D(\mu\|\mu_0^p), \inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu[d_2]\geq\mathbb{E}_{\mu_0}[d_1]}} D(\mu\|\mu_0^p) \right) \\
&= \min\left(\inf_{\substack{\mu:\mu^p=\mu_0^p \\ \mathbb{E}_\mu[d_1]\geq\mathbb{E}_{\mu_0}[d_1]}} D(\mu\|\mu_0^p), \inf_{\substack{\mu:\mu^p=\tilde{\mu}_0^p \\ \mathbb{E}_\mu[d_2]\geq\mathbb{E}_{\tilde{\mu}_0}[d_2]}} D(\mu\|\tilde{\mu}_0^p) \right) \\
&= \min\left(D(\mu_0\|\mu_0^p), D(\tilde{\mu}_0\|\tilde{\mu}_0^p) \right) \\
&= D(\mu_0\|\mu_0^p)
\end{aligned}$$

where we defined $\tilde{W}_0$ as

$$\begin{aligned}
W_0 &= \begin{bmatrix} a_0 & 1-b_0 \\ 1-a_0 & b_0 \end{bmatrix} \\
\tilde{W}_0 &= \begin{bmatrix} 1-b_0 & a_0 \\ b_0 & 1-a_0 \end{bmatrix}.
\end{aligned}$$

$$\mathop{\mathbb{E}}_{\mu_0}[d_1] = \frac{1}{2}\left[(a_0+b_0)\log(2a_1) + (2-a_0-b_0)\log(2(1-a_1))\right]$$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\tilde{\mu}_0}[d_2] &= \frac{1}{2}\left[(a_0+b_0)\log(2a_2)+(2-a_0-b_0)\log(2(1-a_2))\right] \\
&= \frac{1}{2}\left[(1-b_0+1-a_0)\log(2[1-a_1])+(2-1+b_0-1+a_0)\log(2a_1)\right] \\
&= \frac{1}{2}\left[(2-b_0-a_0)\log(2[1-a_1])+(b_0+a_0)\log(2a_1)\right] \\
&= \mathop{\mathbb{E}}_{\mu_0}[d_1]
\end{aligned}$$

□

4.6.1 Alternate Proof of Theorem 4.10

We consider two metrics, namely $d_1 = (a_1, a_1)$ and $d_2 = (1-a_1, 1-a_1)$. With these two metrics, we need to prove that, the achievable rate on any $W_0 = (a_0, b_0)$ is independent of d_1, d_2 and are equal to $D(\mu_0\|\mu_0^p)$.

We need to prove,

$$\inf_{\mu:\max_{k=1,2}\mathbb{E}[d_k]\geq\max_{k=1,2}\mathbb{E}[d_k]} D(\mu\|\mu_0^p) = D(\mu_0\|\mu_0^p)$$

First consider the constraints:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mu}[d_1]-\mathop{\mathbb{E}}_{\mu}[d_2] &= \frac{1}{2}(a\log(2a_1)+(1-b)\log(2(1-a_1))) \\
&\quad +\frac{1}{2}(b\log(2a_1)+(1-a)\log(2(1-a_1))) \\
&\quad -\frac{1}{2}(a\log(2(1-a_1))+(1-b)\log(2a_1)) \\
&\quad +\frac{1}{2}(b\log(2(1-a_1))+(1-a)\log(2a_1)) \\
&= (1-a+b)\log\left(\frac{1-a_1}{a_1}\right) \\
&= \begin{cases} \geq 0, & if \quad a+b\leq 1, a_1\leq\frac{1}{2} \\ \leq 0, & if \quad a+b\geq 1, a_1\leq\frac{1}{2} \end{cases}
\end{aligned}$$

Note that, here we fix $a_1 \leq \frac{1}{2}$ w.l.g, since the two metrics are chosen from the two regimes (rephrase later).

Hence,

$$\max\left(\mathop{\mathbb{E}}_{\mu}[d_1], \mathop{\mathbb{E}}_{\mu}[d_2]\right) = \begin{cases} \mathop{\mathbb{E}}_{\mu}[d_1], & if \quad a+b\leq 1 \\ \mathop{\mathbb{E}}_{\mu}[d_2], & if \quad a+b\geq 1 \end{cases}$$

Similarly,

$$\max\left(\mathop{\mathbb{E}}_{\mu}[d_1], \mathop{\mathbb{E}}_{\mu_0}[d_2]\right) = \begin{cases} \mathop{\mathbb{E}}_{\mu_0}[d_1], & if \quad a_0+b_0\leq 1 \\ \mathop{\mathbb{E}}_{\mu}[d_2], & if \quad a_0+b_0\geq 1 \end{cases}$$

So, we have four regimes and if we can prove that, the optimum in all these regimes is the same, then we are done. The four regimes $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$ are:

$$\Gamma_1 \quad = \quad a_0 + b_0 \leq 1, \quad a + b \leq 1$$

Set the Lagrangian $L(a, \lambda)$,

$$L(a, \lambda) = f(a) + \lambda g(a)$$

where $f(a) = D\left(\mu_0 \| \mu_0^p\right)$ with the marginal constraint $a - b = a_0 - b_0$. The constraint $g(a) \leq 0$ where,

$$\begin{aligned} g(a) &= \mathop{\mathbb{E}}_{\mu}[d_1] - \mathop{\mathbb{E}}_{\mu_0}[d_1] \\ &= \frac{1}{2}(a + b - a_0 - b_0) \log\left(\frac{a_1}{1 - a_1}\right) \end{aligned}$$

Optimum conditions are,

$$\begin{aligned} \frac{\partial}{\partial a} L(a, \lambda) &= 0 \\ g(a) &= 0 \\ \lambda &\geq 0 \end{aligned}$$

at the optimum point $a = a^\star$.

$$g(a) \quad = \quad (a_0 + b_0 - a - b) \log\left(\frac{a_1}{1 - a_1}\right) = 0$$

when $a = a_0$ and $b = b_0$ for all a_1. This is because,

$$\begin{aligned} a_0 + b_0 - a - b &= 0 \\ a - a_0 &= b - b_0 \end{aligned}$$

where the second equation is the marginality constraint $(\mu)_Y = (\mu_0)_Y$.

We also need to check $\lambda \geq 0$ condition, which is true (to be written). Other regimes too have the same optimum point since, the condition $g(a) = 0$ is achieved by the same optimum point.

4.7 Summary and discussions on the main results

In this chapter we proved that for any binary memoryless channel (BMC) one can design a linear decoder and achieve rates up to capacity. When the input

distribution is fixed to uniform the rate thus achieved is the symmetric capacity. With this result the quest to find a linear mismatched decoder whose performance is as good as the optimum decoder (which uses channel knowledge) is established. Hence this serve as a favourable alternative to the MMI decoder because of the linear structure present.

Because of this optimal performance in mismatched setting, the same decoder serve as a universal decoder for any compound BMC. Thus, we established the fact that, for compound BMC, linear decoders indeed exist beyond one sided sets. Since the decoder proposed uses metrics independent of the set, we can get away from the burden to compute the dominant channels in the set, as required in the case of compound sets which are union of one sided channels.

Compound Capacity Achieving Codes

5

So far, the discussions we had on the compound channels were on the existence of a code using which one can in principle achieve reliable communication over a family of channels. We stressed the importance of a decoder to operate without the knowledge of the channel of communication. Our subsequent (main) focus turned to the linear decoders because of its appealing additive property which can be exploited when the code used is suitably structured. One of the main results established in this thesis has been on the existence of a linear decoder for any compound binary memoryless channel. The proof of existence of a good code in the compound setting was based on the random code ensemble argument.

However, we have not identified any structured codes to validate the claim of exploiting the linear decoder structure. While this problem (finding structured codes) is a relatively hard problem in coding theory, in this chapter we make a gentle attempt to do a preliminary investigation on the suitability of one class of codes, namely *polar codes* as a universal coding strategy for the compound BMC set. In the interest of time, we limit the discussion mainly to polar codes under successive decoding rule (a low complexity decoding scheme proposed in the original work introducing polar codes [51]), skipping technical details of the encoding and decoding schemes.

5.1 Polar codes

Recently, Arikan introduced a family of error correcting codes known as *polar codes* [51]. These are the first cases of provably symmetric capacity achieving codes for binary channels. It has been since proved that, Polar codes are useful for a wide variety of problems in information theory including source coding[50]. A comprehensive treatment on Polar codes and its application to

source and channel coding is presented in a recent work of Korada [50].

Our interest in Polar codes is motivated by its capacity achieving property on binary memoryless channels. It is natural to investigate the suitability of such codes to the compound channel coding problem. In a recent development authors in [49] proved that under successive cancellation decoding (successive decoding has an appealing complexity order of $\mathcal{O}(n \log n)$ where n is the block-length of the code), polar codes cannot achieve compound capacity. However for the class of degraded channels, by using the codes for the dominant channel, it is possible to achieve the symmetric compound capacity. We have tried to characterize the class of such degraded channels for the binary symmetric channels.

Our main motivation on Polar codes in the compound channel setting is to seek whether Polar codes can serve as compound capacity achieving under GMAP decoding.

5.2 Universal Polar codes

In general polar codes do not achieve compound capacity. A formal proof is presented in [49]. A sketch of the proof is given below.

Let us consider two channels V and W. Since

$$\begin{aligned} \min(x, y) &\leq x \\ \min(x, y) &\leq y \end{aligned}$$

we can write,

$$\begin{aligned} I\left(V^+\right) \wedge I\left(W^+\right) &\leq I\left(V^+\right) \\ I\left(V^+\right) \wedge I\left(W^+\right) &\leq I\left(W^+\right) \\ I\left(V^-\right) \wedge I\left(W^-\right) &\leq I\left(V^-\right) \\ I\left(V^-\right) \wedge I\left(W^-\right) &\leq I\left(W^+\right) \end{aligned}$$

and hence,

$$\begin{aligned} \frac{1}{2}\left[I\left(V^+\right) \wedge I\left(W^+\right) + I\left(V^-\right) \wedge I\left(W^-\right)\right] &\leq \frac{1}{2}\left[I\left(V^+\right) + I\left(V^-\right)\right] \\ &= I(V) \\ \frac{1}{2}\left[I\left(V^+\right) \wedge I\left(W^+\right) + I\left(V^-\right) \wedge I\left(W^-\right)\right] &\leq \frac{1}{2}\left[I\left(W^+\right) + I\left(W^-\right)\right] \\ &= I(W) \\ &\leq I(V) \wedge I(W). \end{aligned}$$

It is easy to show that this form a degrading sequence. In general, we can write,

$$\begin{aligned}\frac{1}{2^n}\sum_{i=1}^{n} I\left(V^{\sigma_n}\right) \wedge I\left(W^{\sigma_n}\right) &\geq \frac{1}{2^{n-1}}\sum_{i=1}^{n-1} I\left(V^{\sigma_{n-1}}\right) \wedge I\left(W^{\sigma_{n-1}}\right) \\ &\cdots \\ &\geq I(V) \wedge I(W).\end{aligned}$$

where $\sigma_n = \{\sigma_n(i)\}, i = 1, 2, \ldots, n : \sigma_n(i) \in \{+, i\}$

5.2.1 Compound BSC

Because of the cascading nature of the polar code construction, it is possible to construct a universal polar code for compound BSC set. This is because of the fact that good polar code indices for a degraded BSC is a subset of polar code indices for a better BSC. For example $BSC\left(\frac{1}{4}\right)$ is a degraded version of $BSC\left(\frac{1}{8}\right)$. In other words, a polar code constructed for $BSC\left(\frac{1}{4}\right)$ stands as a good code for $BSC\left(\frac{1}{8}\right)$ as well. In general $BSC\left(\frac{1}{4}\right)$ can be written as a cascade of $BSC\left(\frac{1}{8}\right)$ and another symmetric channel. In this case the compound rate is the minimum and this rate can be achieved by polar codes constructed for the dominant channel in the set.

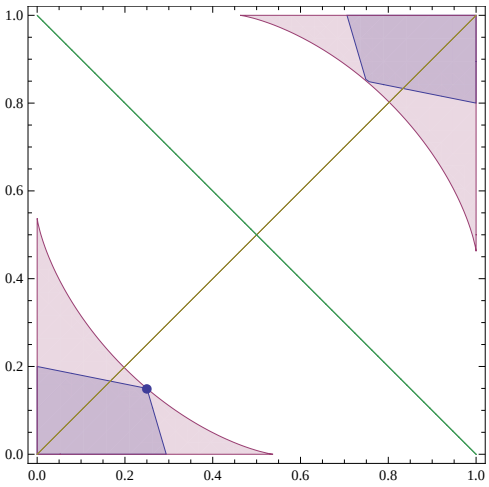


Figure 5.1: Polar code performance

5.2.2 Polar codes for compound binary set

Figure to be added if this section is found meaningful: Cascade of two channels: First channel is BSC with transition probability $1 - z$. Second channel is $(a, b) = (x, y)$. The resulting channel is (α, β). See notebook (black cover).

$$zx + (1 - z)(1 - y) = \alpha$$

$$\begin{aligned} zy + (1-z)(1-x) &= \beta \\ z(1-x) + (1-z)y &= 1-\alpha \\ (1-z)x + z(1-y) &= 1-\beta \end{aligned}$$

This gives us the condition $x-y = \alpha-\beta$. For arbitrary $\alpha \in [0,1], \beta \in [0,1]$, we need to find x, y. It turns out that there is a unique solution:

$$\begin{aligned} x &= \frac{(1+a-b)z + b - 1}{2z-1} \\ y &= \frac{(1-a+b)z + a - 1}{2z-1} \end{aligned}$$

Example 5.1. $\alpha = 0.2, z = 1/8, \beta = 0.4. x = 0.66667, y = 0.866667.$

5.3 Degraded Compound set

Definition 5.2. *A channel W_2 is said to be a degraded version of channel W_1 if there exists a channel W such that $W_2 = W_1 W$. That is, W_2 can be expressed as a cascade of the channels W_1 followed by W.*

For a given compound rate $C_{\mathcal{S}}$, we can characterize the set of DMCs $W \in \mathcal{S}$ such that the worst channel $W_0 \in \mathcal{S}$ is degraded a version of all channels in the set. Since the good channel indices for a degraded channel W_0 is a subset of the good channel indices of any of $W \in \mathcal{S}$, polar codes can be used as universal for such a set.

Figure. 5.1 illustrate the set of such a degraded compound set for a chosen rate. Clearly, all channels with rates R are not guaranteed to have universal polar codes. However a subset of such a set (the smaller shaded region) is a sufficient set which admit universal polar codes.

5.3.1 Gap to capacity

For arbitrary compound BMCs, if the polar codes constructed for the least rate (the compound capacity C_S), it is interesting to compute the worst possible loss. This is shown in Figure. 5.2. The worst possible gap happens when we consider arbitrary compound BMC with $R = \frac{1}{2}$.

The absolute gap to compound capacity is sketched against all rates in Figure.5.3

5.4 Polar universal codes under GMAP

The preliminary investigation on the universal property of Polar codes is based on the code construction property with degraded channels and that with successive decoding rule. The results thus serve as a sufficient condition to use

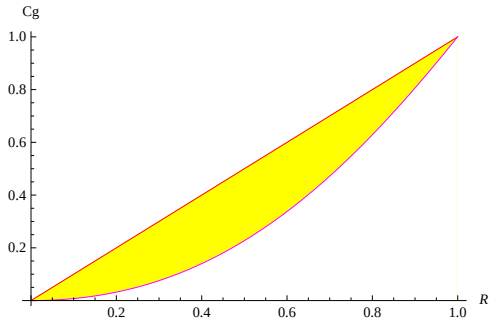


Figure 5.2: Universal Polar codes: Shaded region shows the gap region when polar codes are used as universal codes for arbitrary compound BMCs. The performance shown here is under successive decoding. The Gap is expected to minimize or disappear when GMAP decoding is used.

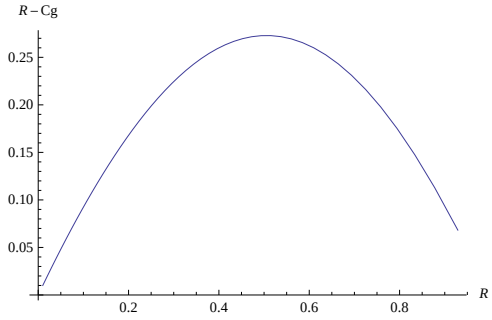


Figure 5.3: Universal Polar codes:The gap to compound capacity is shown for different rates.

under the GMAP decoding rule which we are interested in. More investigation and analysis is required to study the universal properties of Polar codes under GMAP rule. It may still be possible to achieve compound capacity with GMAP beyond degraded class of channels.

5.5 Discussion on Polar universal codes

The results presented in this chapter is a very brief overview of our preliminary investigation on the suitability of Polar codes as a universal coding strategy for compound binary memoryless channels. There is ongoing research being carried out by several researchers in this arena, some of them are being mentioned in the reference section. It is clear that, under the successive decoding rule, polar codes cannot achieve the compound capacity for any arbitrary BMC. We presented some characteristics on the set in order to have a polar universal code with a successive decoder at receiver. The worst case loss incurred when polar codes designed for the degraded channel is computed and is presented.

It shows that, when the compound capacity is higher, the loss is considerably small, while it is significantly high with compound sets with lower capacities. The suitability of polar codes under the generalized linear MAP decoder is still being investigated.

Summary and Open Problems

6

In this thesis, we presented the topic of reliable communication over a family of unknown channels. We looked at the possibility of the existence of linear decoders for arbitrary compound sets of binary memoryless channels. We subsequently proved that, it is indeed possible to have a code which admit linear decoder for any arbitrary compound BMC. The linear decoder thus designed can achieve rates all the way up to the channel capacities of the individual channels in the set. Thus, we can replace the maximum mutual information (MMI) decider with linear decoders when structured universal codes.

While there is no known structured universal codes for DMCs, we looked at Polar codes in a quest to identify universal codes. They are not universal for general BMCs under successive decoding strategy, but their candidature is still being investigated under linear generalized MAP decoding rule.

6.1 Open problems

There are a a lot of open and unsolved problems related to the compound channels. A small list of them, heavily incomplete is presented below.

1. For compound binary memoryless channels, we proved the existence of codes which admit linear decoders. The decoder uses mismatched metrics and using which one can achieve all the way up to the channel capacity of the respective channels (The individual channel capacity can be achieved when the sender knows the rate. In a compound setting with transmitter do not know the rate, one can only hope to send at rate up to the compound rate). The results presented so far holds good only for binary

input binary output channels. It will be interesting to seek the possibility on larger alphabet channels, namely arbitrary DMCs.

2. Universal codes. There are no known codes which are universal. It will be an exciting line of research to identify structured codes which serve as universal codes for arbitrary DMCs. The recent invention of polar codes offered initial promise, but it was soon found out that, they are not universal under the less complex decoder, namely the successive decoder used in polar codes. It is still an open question as to whether the polar codes serve as universal codes under a more general decoding rule, like generalized maximum aposteriori scheme.

3. In order to admit a simpler implementation of the linear decoder, one need good structured codes. One of the exponents of exploiting such simplification has been tree codes like convolutional codes. If one can find codes like convolutional codes, for compound channels, one could hope to take the universal decoders a leap step closer to reality. It will be interesting to see whether such codes, even if not capacity achieving, can be found.

Bibliography

[1] C. E. Shannon, A Mathematical Theory of Communication,Bell Technical 1948

[2] A.J. Viterbi, "Wireless digital communication: a view based on three lessons learned," IEEE Commun. Mag., pp. 33-36, Sept. 1991.

[3] P. Elias, "Coding for noisy channels," IRE Conv. Rec.,March 1955, vol. 3, pp. 37-46

[4] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Trans. Inform. Theory, vol. IT-13, pp. 260269, 1967.

[5] I.Csiszar, J. Korner, *Information Theory:Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[6] D.Blackwell, L.Breiman, A.J. Thomasian, "The Capacity of a class of channels,", *Ann. Mathe.Stat.*, vol. 30, pp. 1229-1241, Dec. 1959.

[7] J.Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. Berlin/Heidelberg: Springer-Verlag, 1978.

[8] W.L. Root, P.P. Varaiya, "Capacity of classes of Gaussian channels,", SIAM J. Appl. Math., vol. 16, pp. 1350-1393, Nov. 1968.

[9] J. Ziv and A.Lempel., "A universal algorithm for sequential data compression," IEEE Transactions on Information Theory, vol. 23, no.3, May 1977, pp. 337–343.

[10] V.Goppa, "Nonprobabilistic mutual information without memory", Problems of Control and Information Theory, vol. 4, pp. 97-102, 1975.

[11] J.Ziv, "Universal decoding for finite state channels", IEEE Transactions on Information Theory, vol. 31, pp. 453-460, July 1985.

[12] I.N.Sanov, On the Probability of large deviations of random variables, Mat.Sb.42 (in Russian). English translation in: Selected Translations in Mathematical Statistics and Probability I (1961) 231-244.

[13] I.Csiszar, P.C Shields, Information Theory and Statistics: A tutorial,Foundations and Trends in Communications and Information theory, Vol 1, Issue 4, 2004, Ed. S. Verdu, Now Publishers

[14] A.Dembo, O.Zeitouni, Large Deviations Techniques and Applications, Springer Series in Applications of Mathematics: Stochastic Modelling and Applied Probability (Number 38), Second Edition 1998.

[15] F.den Hollander, Large deviations,Field Institute Monographs, American Mathematical Society (AMS), (2000)

[16] T.Cover, J.Thomas, Elements of Information theory, 2nd Edition, Wiley International, 2006

[17] R.G.Gallager, Information Theory and Reliable Communication, John Wilye 1968

[18] T.Richardson, R.Urbanke, *Modern Coding theory*, Cambridge University Press, 2007.

[19] David J.C. MacKay,*Information theory, inference and learning algorithms*, Cambridge University Press, 2003.

[20] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding: Turbo Codes,", Proc. 1993 IEEE International Conference on Communications, Geneva, Switzerland, pp. 1064-1070, May 1993.

[21] E.A.Abbe, Local to geometric Methods in Information Theory, PhD thesis, Massachusetts Institute of Technology, 2008.

[22] E.Abbe, L.Zheng, "Linear Universal Decoding for Compound Channels: a Local to Global Geometric Approach," arXiv Preprint 2008.

[23] V.S.Borkar,Method in the Madness:Laws and Order in Probability,Resonance Magazine, Indian Academy of Science, November 2000.

[24] N.Shulman,M.Feder, The uniform Distribution as a Universal Prior, IEEE Trans. on Information Theory, Vol.50.No.6,June 2004

[25] E.E.Majani and H.Rumsey, Two results on binary-input discrete memoryless channels, In Proc. IEEE Int. Symp. Information Theory, Budapest, Hungary, 1991, P 104

[26] R.J.McEliece, Are turbo-like codes effective on nonstandard channels?, IEEE Inform. Theory Soc Newsletter, Vol.51, pp.1-8, Dec 2001

[27] J.Feng, T.G.Kurtz, Large Deviations for Stochastic Processes, Mathematical surveys and monographs, Vol 131, American Mathematical Society (AMS) 2006

[28] S.W.Goulomb,The limit behavior of the Z-channel, IEEE Trans. Information Theory, Vol. IT-26, p. 372, May 1980.

[29] A.Lapidoth,I.E,Telatar, The compound channel capacity of a class of finite state channels, IEEE Trans. Info.Theory, Vol.44,No.3,May 1998

[30] P.Sweeney,Error control coding, John Wiley and Sons,2002

[31] R. G. Gallager, A simple derivation of the coding theorem and some applications, IEEE Transactions on Information Theory, vol. 11, pp. 3-18, Jan. 1965.

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

[33] S. Shamai and I. Sason, Variations on the Gallager bounds, connections, and applications,IEEE Transactions on Information Theory, vol. 48, pp. 3029-3051, Dec. 2002.

[34] N.Merhav, G.Kaplan, A.Lapidoth and S.Shamai:"On information rates for mismatched decoders," IEEE Trans. Info. Theory, vol 40. No.6 pp:1953-1967, November 1994.

[35] I.Csiszer, P.Narayan, "Channel Capacity for a Given Decoding Metric", IEEE Trans.Info. Theory, Vol. 41, No.1, January 1995.

[36] A. Lapidoth and P. Narayan, "Reliable Communication under channel uncertainty," IEEE Trans.Info. Theory, Vol.40, No.10, pp:2148-2177, October 1998.

[37] A.Ganti, A.Lapidoth, I.E.Telatar,"Mismatched Decoding Revisted:General Alphabets, Channels with Memory, and the Wide-band Limit", IEEE. Trans.Info. Theory, Vol.46, No.7, pp:2315-2327, November 2000.

[38] J.P. Bouchaud, M. Mzard, J. Dalibard, Complex systems: cole d't de Physique des Houches, session LXXXV, 3-28

[39] C. E. Shannon, R. G. Gallager, and E. Berlekamp, Lower bounds to error probability for coding on discrete memoryless channels, Information Control, vol. 10, pp. 65103, 1967.

[40] C.Arndt, Information Measures, Information and its description in Science and Engineering, Springer 2004

[41] R.Ahlswede, L.Baumer, N.Cai,H.Aydinian,V.Blinovsky,C.Deppe,H.Mashurian, General Theory of Information Transfer and Combinatorics, Springer 2006

[42] Mark Chu-Carroll,Simplices and Simplicial Complexes: A Science blog http://scienceblogs.com/goodmath/2007/03/simplices_and_simplicial_compl.php

[43] Wikipedia: Pentachoron http://en.wikipedia.org/wiki/Pentachoron

[44] Wikipedia: Simplex http://en.wikipedia.org/wiki/Simplex

[45] R.A.Silverman, On Binary Channels and Their Cascades, IRE Transactions on Information Theory, 1955

[46] S.Muroga, On the capacity of a discrete channel, International Journal of Physics Society, Japan, vol. 8, pp. 484-494: July-August 1953

[47] M.M.Deza,E.Deza,*Dictionary of distances*,Elsevier,2006.

[48] J. Massey, "Variable-length codes and the Fano metric," Information Theory, IEEE Transactions on , vol.18, no.1, pp. 196-198, Jan 1972.

[49] H.Hassani, S.B.Korada, R.Urbanke, "On the compound capacity of polar codes," in preparation 2009

[50] S.B.Korada,Polar Codes for Channel and Source Coding, PhD Thesis, EPFL 2009.

[51] E.Arikan,"Channel Polarization: A method for constructing capacity achieving codes for symmetric binary input memoryless channels," *submitted to IEEE Trans. Inform. Theory*, 2008.

[52] G. D. Forney. The Viterbi algorithm. Proceedings of the IEEE 61(3):268278, March 1973

[53] A.J.Viterbi, J.K.Omura, Principles of digital communication and coding:McGraw-Hill, 1979, New York.