# MULTI-USER INFORMATION THEORY: STATE INFORMATION AND IMPERFECT CHANNEL KNOWLEDGE

Pablo Piantanida

UNIVERSITY OF PARIS-SUD XI
SCIENTIFIC UFR OF ORSAY

*THESIS Presented*
*to obtain the degree of*

DOCTOR OF SCIENCES OF THE
UNIVERSITY OF PARIS-SUD XI

# MULTI-USER INFORMATION THEORY: STATE INFORMATION AND IMPERFECT CHANNEL KNOWLEDGE

A dissertation presented

by

**Juan-Pablo Piantanida**

May 14th 2007

The thesis jury is composed of:

Reviewers:

| | | |
|---|---|---|
| Prof. | Muriel Médard | Massachusetts Institute of Technology, |
| Prof. | Ezio Biglieri | Universitat Pompeu Fabra, |

Examinators:

| | | |
|---|---|---|
| Prof. | Amos Lapidoth | Swiss Federal Institute of Technology, |
| Prof. | Philippe Loubaton | Université de Marne la Vallée, |
| Prof. | Jean-Claude Belfiore | École Nationale Supérieure des Télécom., |
| M. | Pierre Duhamel | Directeur de recherche au CNRS. |

Thesis advisor
**Pierre Duhamel**

Author
**Juan-Pablo Piantanida**

# Abstract

The capacity of single and multi-user state-dependent channels under imperfect channel knowledge at the receiver(s) and/or transmitter are investigated. We address these channel mismatch scenarios by introducing two novel notions of reliable communication under channel estimation errors, for which we provide an associated coding theorem and its corresponding converse, assuming discrete memoryless channels. Basically, we exploit for our purpose an interesting feature of channel estimation through use of pilot symbols. This feature is the availability of the statistic characterizing the quality of channel estimates.

In this thesis we first introduce the notion of *estimation-induced outage capacity* for single-user channels, where the transmitter and the receiver strive to construct codes for ensuring reliable communication with a quality of service (QoS), no matter which degree of accuracy estimation arises during a transmission. In our setting, the quality of service constraint stands for achieving target rates with small error probability (the desired communication service), even for very poor channel estimates. Our results provide intuitive insights on the impact of the channel estimates and the channel characteristics (e.g. SNR, number of pilots, feedback rate) on the maximal mean outage rate.

Then the optimal decoder achieving this capacity is investigated. We focus on the family of decoders that can be implemented on most practical coded modulation systems. Based on the theoretical decoder that achieves the capacity, we derive a practical decoding metric for arbitrary memoryless channels that minimizes the average of the transmission error probability over all channel estimation errors. Next, we specialize this metric for the case of fading MIMO channels. According to our notion of outage rates, we characterize maximal achievable information rates of the proposed decoder using Gaussian codebooks. Numerical results show that the derived metric provides significant gains, in terms of achievable information rates and bit error rate (BER), in a bit interleaved coded modulation (BICM) framework, without introducing any additional decoding complexity.

We next consider the effects of imperfect channel estimation at the receivers with imperfect (or without) channel knowledge at the transmitter on the capacity of state-dependent channels with non-causal channel state information at the transmitter. We address this through the notion of reliable communication based on the average of the transmission error probability over all channel estimation errors. This notion allows us to consider the capacity of a composite (more noisy) Gelfand and Pinsker's channel. We derive the optimal Dirty-paper coding (DPC) scheme that achieves the capacity (assuming Gaussian inputs) of the fading Costa channel under the mentioned conditions. The results illustrate a practical trade-off between the amount of training and its impact to the interference cancellation performances of DPC scheme. This approach enable us to study the capacity region of the multiuser Fading MIMO Broadcast Channel (MIMO-BC), where the mobiles (the receivers) only dispose of a noisy estimate of the channel parameters, and these estimates may be (or not) available at the base station (the transmitter). In particular, we observe the surprising result that a BC with a single transmitter and receiver antenna, and imperfect channel estimation at each receiver, does not need the knowledge of estimates at the transmitter to achieve large rates.

Finally, we consider several implementable DPC schemes for multi-user information embedding, through emphasizing their tight relationship with conventional multi-user information theory. We first show that depending on the targeted application and on whether the different messages are asked to have different robustness and transparency requirements, multi-user information embedding parallels the Gaussian BC and the Gaussian Multiple Access Channel (MAC) with non-causal channel state information at the transmitter(s). Based on the theoretical DPC, we propose practical coding schemes for these scenarios. Our results extend the practical implementations of QIM, DC-QIM and SCS from the single user case to the multi-user one. Then, we show that the gap to full performance can be bridged up using finite dimensional lattice codebooks.

# Acknowledgments

I wish to thank several number of people for making my experience during my PhD. a memorable one. First of all, I owe my deepest gratitude to my advisor Mr. Pierre Duhamel for his continual support and guidance over the years. His continual encouragement to formulate novel and relevant research problems, and enthusiasm for all that he does has been truly inspirational. Mr. Duhamel gave me an initial push and always showed great faith in my abilities, allowing me to work independently, but at the same time provided invaluable guidance at the necessary times. He has learned me the importance of the choice of research topics, the teamwork and a lot of things very useful for my research career, for which I will be forever grateful.

I am grateful to Prof. Muriel Médard and Prof. Ezio Biglieri for serving as my thesis reviewers. They provided me a critical reading, valuable suggestions and insightful comments which have been very important for the improvement of my work. Prof. Médard has been a major reference and inspiration for my work. I would also like to thank Professors Philippe Loubaton, Amos Lapidoth and Jean-Claude Belfiore for serving on my orals committee and attending my defense. Prof. Lapidoth has also been a wonderful reference from an information theoretic view-point, that greatly broadened my depth of knowledge of the field, for which he has my admiration.

I would also like to thank Prof. Gerald Matz of Vienna University of Technology, Austria, for his enthusiasm, his contribution and dedication during our collaboration, without him much of this work would not have been possible. I would like to thank all those I interacted with, while interning at the Vienna University, specially Prof. Franz Hlawatsch for receiving me and making of my stay a wonderful experience.

I would also like to thank Mr. Walid Hachem for his interest in my work and his very useful comments, and Prof. Te Sun Han at the Electro-Communication University, Japan, for his helpful discussions via email. I would also like to thank Mr. Samson Lasaulce and Mr. Olivier Rioul for their helpful discussions and encouragement at the begining of my PhD. Mr. Rioul has also been a wonderful teacher that will serve as continual inspiration in my future teaching. I am also thankful to my co-authors Abdellatif Zaidi and Sajad Sadough, whose contributions enriched the work

of this thesis.

I have to thank my friends at the Laboratoire des Signaux et Systèmes and at Supélec, for making the years so enjoyable. I would like to thank Florence, my officemate, for her kindness that contributed to a good working atmosphere. I would also like to thank my parents for encouraging me to be persistent and never give up on something that I want to achieve and also for their love and dedication. Of course, I have to thank all my friends at the University of Buenos Aires, Argentina, for encouraging me to love the research during my graduate studies. Finally, I am particularly indebted to my future wife Marie. We met at the LSS during my first year, and my experience here would not have been the same without her in my life. She has brought so much love to my life and has been a constant source of support and motivation throughout my studies.

*Dedicated to my parents,*
*and to Marie.*

# Table of Contents

# List of Figures

# List of Tables

# Published and Upcoming Works

The material contained in Chapter 2 have been done in collaboration with Prof. G. Matz and have appeared in the following papers:

[1] Piantanida, P., Matz, G. and Duhamel, P., "Outage Behavior of Discrete Memoryless Channels Under Channel Estimation Errors", 2006, Oct. 29 - Nov. 1, *Proc. of IEEE International Symposium on Information Theory and its Applications, ISITA*, Seoul, Korea.
[2] Piantanida, P., Matz, G. and Duhamel, P., "Estimation-Induced Outage Capacity of Ricean Channels", 2006, July 2-5, *Proc. of IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cannes, France.
[3] Piantanida, P., Matz, G. and Duhamel, P., "Outage Behavior of Discrete Memoryless Channels Under Channel Estimation Errors", *Submitted to IEEE Transactions on Information Theory*, 2006, December.

The material contained in Chapter 3 have been done in collaboration with S. Sadough and have appeared in the following papers:

[4] Piantanida, P., Sadough, S. and Duhamel, P., "On the Outage Capacity of a Practical Decoder Using Channel Estimation Accuracy", 2007, *To appear in Proc. of IEEE International Symposium on Information Theory (ISIT)*, Nice, France
[5] Sadough, S. and Piantanida, P. and Duhamel, P., "MIMO-OFDM Optimal Decoding and Achievable Information Rates under Imperfect Channel Estimation", 2007, *Submitted to IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*
[6] Sadough, S., Piantanida, P. and Duhamel, P.,"Achievable Outage Rates with Improved Decoding of Multiband OFDM Under Channel Estimation Errors", 2006, Oct. 29 - Nov. 1, *Proc. of the 40th Asilomar Conference on Signals, Systems and Computers*, California, USA
[7] Piantanida, P, Sadough, S. and Duhamel, P., "On the Outage Capacity of a Practical Decoder Using Channel Estimation Accuracy", *To be submitted to IEEE Trans. on Communications*, 2007.

The material contained in Chapter 4 have appeared in the following papers:

[8] Piantanida, P. and Duhamel, P., "Dirty-paper Coding without Channel Information at the Transmitter and Imperfect Estimation at the Receiver", 2007, *To appear in IEEE International Conference on Communications (ICC)*, Scotland, UK
[9] Piantanida, P. and Duhamel, P., "On the Capacity of the Fading MIMO Broadcast Channel without Channel Information at the Transmitter and Imperfect Estimation at the Receivers ", 2007, *To appear in IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP)*, Hawaii, USA

[10] Piantanida, P. and Duhamel, P., "Achievable Rates for the Fading MIMO Broadcast Channel with Imperfect Channel Estimation", 2006, Sep. 27-29, *Proc. of the Forty-Fourth Annual Allerton Conference on Communication, Control, and Computing*, Illinois, USA

[11] Piantanida, P. and Duhamel, P., "Dirty-paper Coding with Imperfect Channel Estimation Knowledge: Applications to the Fading MIMO Broadcast Channel", *To be submitted to IEEE Transactions on Information Theory*, 2007.

The material contained in Chapter 5 have been done in collaboration with A. Zaidi and have appeared in the following papers:

[12] Piantanida, P., Lasaulce, S. and Duhamel, P., "Broadcast Channels with Noncausal Side Information: Coding theoremf and Application Example", 2005, Feb. 20-25, *Proc. Winterschool on Coding and Information Theory*, Bratislava, Slovakia

[13] Zaidi, A. and Piantanida, P., "MAC Aware Coding Strategy for Multiple User Information Embedding", 2006, May 15-19, *Proc. of IEEE Int. Conf on Audio and Speech Signal Processing, ICASSP*, Toulouse, France

[14] Zaidi, A. and Piantanida, P. and Duhamel, P., "Scalar Scheme for Multiple User Information Embedding", 2005, March 18-23, *Proc. of IEEE Int. Conf. on Audio and Speech Signal Processing, ICASSP*, Philadelphia, USA

[15] Zaidi, A., Piantanida, P. and Duhamel, P., "Broadcast-Aware and MAC-Aware Coding Strategies for Multiple User Information Embedding", *To appear in IEEE Transactions on Signal Processing*, 2007.

Electronic preprints are available on the Internet at the following URL:

`http://www.lss.supelec.fr`

# Chapter 1

# Introduction

In the early 1940s, it was thought (the belief was) that increasing the transmission rate of information over a communication channel increased the probability of error. A communication channel consists of a transmitter (source of information), a transmission medium (with noise and distortion), and a receiver (whose goal is to reconstruct the sender's messages). Claude E. Shannon in his classic papers [1], [2] surprised the communication theory community by proving that this was not true as long as the communication rate was below channel capacity, i.e., the maximum amount of information that can be sent over a noise channel. He showed the basic results for memoryless sources and channels and introduced more general communication models including state-dependent channels.

Shannon's original work focused on memoryless channels whose probability distribution (the noise characteristics of the channel), which is assumed to not change with time, is perfectly known to both the transmitter and the receiver. In this scenario, he proved the existence of good coding and decoding schemes to derive a coding theorem and its converse that allows one to calculate the channel capacity from the noisy characteristics of the channel. While mathematical notions of information had existed before, it was Shannon who made the connection between the construction of optimal codes and an ingenious idea known as "random coding" in order to develop coding theorems and thereby give operational significance to the information measures[1]. The mathematical tools used for these proofs is the concept of *typical sequences* and the

[1]The name "random coding" is a bit misleading since it refers to the random selection of a deterministic code and not a coding systems that operates in a random or stochastic manner.

*concentration of measure phenomenon* as a device to redefine the class of typical sequences and to estimate the residual mass probability of the non-typical sequences (see Csiszàr's tutorial paper [3]).

Information theory or the mathematical theory of communications has two primary goals: The first is the development of the fundamental theoretical limits on the achievable performance when communicating a given information source over given communication channels using optimal (but theoretical) coding schemes from within a prescribed class. The second goal is the development of practical coding schemes, e.g. optimal encoder(s) and decoder(s), that provide performance reasonably good in comparison with the optimal performance given by the theory.

Current research in information theory today is motivated by the increasing interest of its potential applications on the design of single and multi-user communication systems, computer networks, cooperative communications, multi-terminal source coding, multimedia signal processing, etc. There are several similarities in concepts and methodologies between information theory and these current research areas so that the results can be easily extrapolated. A good application example of these ideas is the potential applications of *Dirty-paper Coding* (DPC) for interference cancellation in multi-user communications such as Broadcast channels or applications such as multiple user information embedding (watermarking), in multimedia signal processing.

The developments so far in the engineering community had as significant an impact on the foundations of information theory as they had on applications. In this thesis, by using the relationships between information theory and its applications, we focus on both aspects: (i) The development of capacity expressions providing the ultimate limits of communications under imperfect channel knowledge and (ii) the optimal means of achieving these limits by practical communication systems. The remainder of this chapter provides necessary background material and outlines the contributions of this thesis.

## 1.1 Background

In this section, we review some of fundamental results in information theory and other topics related to the framework of this thesis.

### 1.1.1 Basic Results

Mathematicians and engineers extended Shannon's basic approach to ever more general models of information sources, coding structures, and performance measures. The fundamental ergodic theorem for entropy was extended to the same generality as the ordinary ergodic theorems by McMillan [4] and Breiman [5] and the result is now known as the *Shannon-McMillan-Breiman* theorem (the *asymptotic equipartition* theorem or AEP, the ergodic theorem of information theory, and the entropy theorem). A variety of detailed proofs of the basic coding theorems and stronger versions of the theorems for memoryless, Markov, and other special cases of random processes were developed, notable examples being the work of Feinstein [6] and Wolfowitz [7].

The ideas of measures of information, channels, codes, and communications systems were rigorously extended to more general random processes with abstract alphabets and discrete and continuous time by Khinchine [8] and by Kolmogorov, Gelfand, Yaglom, Dobrushin, and Pinsker [9], [10], [9] and [11]. In addition, the classic notion of entropy was not useful when dealing with processes with continuous alphabets since it is virtually always infinite in such cases. A generalization of the idea of entropy called discrimination was developed by Kullback (cf. [12]). This form of information measure is now more commonly referred to as relative entropy (or Kullback-Leibler number) and it is better interpreted as a measure of similarity between probability distributions than as a measure of information between random variables. Many results for mutual information and entropy can be viewed as special cases of results for relative entropy and the formula for relative entropy arises naturally in some proofs.

Traditional noiseless coding theorems with simpler proofs of the basic results can be found in the literature in a variety of important cases. See, e.g., the texts by Gallager [13], Cover [14], Berger [15], Gray [16], and Csiszàr and Körner [17]. In addition to this bibliography, good surveys of the multi-user information theory may be found in El Gamal and Cover [18], van der Meulen [19], and Berger [20].

### 1.1.2 Related and Subsequent Works

We begin with the model originally addressed by Shannon [1] of a known memoryless channel with (finite) input $\mathscr{X}$ and output $\mathscr{Y}$ alphabets, respectively. The

channel law is defined by the probabilities $W(y|x)$ of receiving $y \in \mathscr{Y}$ when $x \in \mathscr{X}$ is sent. This channel is fixed and assumed to be known at both the transmitter and the receiver. For this model, the capacity is given by [1]

$$C(W) = \max_{P \in \mathscr{P}(\mathscr{X})} I\big(P, W\big),$$

where $\mathscr{P}(\mathscr{X})$ denotes the set of all (input) probability distributions on $\mathscr{X}$ and

$$I\big(P, W\big) = \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} P(x) W(y|x) \log \frac{W(y|x)}{Q(y)},$$

with $Q(y) = \sum_{x \in \mathcal{X}} P(x) W(y|x)$ is the mutual information between the channel input and output.

Within the class of Gaussian channels $W$, we consider constant or additive white Gaussian noise (AWGN) channels, fading channels, and multiple-antenna channels. We refer the reader to the above mentioned texts and for a complete survey of fading channels see Biglieri, Proakis and Shamai [21].

In addition to the Shannon's capacity, the concept of *outage capacity* was first proposed in [22] for fading channels. It is defined as the maximum rate that can be supported with probability $1 - \gamma$, where $\gamma$ is a prescribed outage probability. Furthermore, it has been shown that the outage probability matches well the error probability of actual codes (cf. [23, 24]). This outage probability depends on the codeword error probability, averaged over a random coding ensemble and over all channel realizations. In contrast, ergodic capacity is the maximum information rate for which error probability decays exponentially with the code length.

**State-dependent channels**

In subsequent work, Shannon [25] and others have proposed several different channel models for a variety of situations in which either the encoder or the decoder must be selected without a complete knowledge of the statistic governing the channel over which transmission occurs. Our emphasis in this thesis shall be on single-user and multi-user channels controlled by random states. In such situations where the channel statistic is fully unknown, the most relevant models can be summarized to: (i) compound channels and (ii) arbitrarily varying channels.

(i) Compound DMCs, which models communication over a memoryless channel whose law is unknown but remains fixed throughout a transmission. Both transmitter and receiver are assumed ignorant of the channel law governing the transmission; they only know the family $\mathcal{W}$ to which the law belongs $W \in \mathcal{W}$. We emphasize that in this model no prior distribution is assumed, and codes for these channels must therefore exhibit a small probability of error for every channel in the family. The capacity of a compound DMC is given by the following expression

$$C(\mathcal{W}) = \max_{P \in \mathscr{P}(\mathscr{X})} \inf_{W \in \mathcal{W}} I\big(P, W\big).$$

Obviously, the highest achievable rate cannot exceed the capacity of any channel in the family, but this bound is not tight, as different channels in the family may have different capacity achieving input (cf. [26], [27], [28], [7]). However, if the encoder knows the channel, even if the decoder does not, the capacity is equal to the infimum of the capacities of the channels in the family.

(ii) Arbitrarily varying channels (AVC's) were introduced by Blackwell, Breiman, and Thomasian [29] to model communication situations where the channel statistics ("state") may vary in an unknown and arbitrary manner during the transmission of a codeword, perhaps caused by jamming. Formally, an AVC with input alphabet $\mathscr{X}$, output alphabet $\mathscr{Y}$, and set of possible states $\mathscr{S}$ is defined by the probabilities $W(y|x,s)$ of receiving $y \in \mathscr{Y}$ when $x \in \mathscr{X}$ is sent and $s \in \mathscr{S}$ is the state with probability distribution $P_S(s)$. The capacity problem for AVC's has many variants according to sender's and receivers' knowledge about the states, the state selector's knowledge about the codeword, degree of randomization in encoding and decoding, the error probability criteria adopted, etc. (for further discussions we refer the reader to [30]). Assuming the situation when no information is available to the sender and receiver about the states, nor to the state selector about the codeword sent, and random encoders are permissible. Already the authors in [29] showed that

$$C(W, \mathscr{Q}) = \max_{P \in \mathscr{P}(\mathscr{X})} \min_{P_S \in \mathscr{Q}(\mathscr{S})} I\big(P, W_S\big),$$

where $W_S$ is computed by using $P_S$ and $W$.

In the context of fading channels, it is useful to note that the notions of reliable

communication yielding to the compound channel and the arbitrary varying channel, provide very small values of transmission rates (in most of the cases these are equal to zero). In fact these notions require that the resulting values of capacity can be attained when the channel uncertainty is at its severest during the course of a transmission, and hence error probabilities are evaluated as being the largest with respect to the unknown channels states. In other words, the corresponding notions of reliable transmission are not adapted to wireless communication models.

A variation of these channels has been considered by Kusnetsov and Tsybakov in [31], Heegar and El Gamal in [32] and Gelfand and Pinsker in [33], where the channel states are assumed to be available at the transmitter in a non-causal way. Consider the problem of communicating over a DMC where the transmitter knows the channel states before beginning the transmission (i.e. non-causal state information) but the receiver does not know these. This channel is commonly known as *channel with non-causal state information at the transmitter*. The capacity expression of this channel is given by [33],

$$C\big(W, P_S\big) = \sup_{P(u,x|s)\in\mathscr{P}(\mathscr{U}\times\mathscr{X})} \big\{ I\big(P_U, W\big) - I\big(P_S, P_{U|S}\big) \big\}, \tag{1.1}$$

where $U \in \mathcal{U}$ is an auxiliary random variable chosen so that $U \ominus (X, S) \ominus Y$ form a Markov chain, $I(\cdot)$ is the classical mutual information and $\mathscr{P}$ is the set of all joint probability distributions $P(u, x|s) = \delta\big(x - f(u, s)\big)P(u|s)$ with $f : \mathscr{U} \times \mathscr{S} \mapsto \mathscr{X}$ an arbitrary mapping function and $\delta(\cdot)$ is the *dirac* function. The non-causal side information at the transmitter can substantially increase the capacity.

### Mismatched decoders

The class of decoders called *mismatched decoders* has been of interest since 1970's (cf. [34], [35] and [36]). They are decoders defined by minimizing a "distance" given function $d(\mathbf{x}, \mathbf{y}) \geq 0$, which is defined on channel input and output alphabets. Given an output sequence $\mathbf{y}$ this decoder that uses the metric $d$ declares that the codeword $i$ was sent iff $d(\mathbf{x}_i, \mathbf{y}) < d(\mathbf{x}_j, \mathbf{y})$, for all $j \neq i$, and it declares an error if no such exists. Here the term "distance" is used in the widest sense, no restriction on this is implied. This scenario arises naturally when, due to imperfect channel measurement or for simplicity reasons, the receiver is designed using a suboptimal decoding rule.

Theoretically, one can employ universal decoders (cf. [37], [38] and [39]), however in most practical coded modulation systems it is ruled out by complexity considerations. Thus, due to the simplicity of their implementation mismatched decoders are preferred to all others.

The mismatch capacity [34], which is defined as the supremum of all achievable rates, is unknown. More precisely, the $d$-capacity of a DMC is the supremum of information rates of codes with a given $d$-decoder that yields arbitrarily small error probability. In the special case when $d$ is the hamming distance, $d$-capacity provides the *zero-error* capacity or *erasures-only* capacity. Shannon's *zero-error* capacity can also be regarded as a special case of $d$-capacity, cf. [40]. A lower bound to $d$-capacity follows as a special case of a result in [41]; this bound was obtained also by Hui [42]. Csiszár and Narayan [40] showed that this bound is not tight in general but its positivity is necessary for positive $d$-capacity. Lapidoth [43] showed that $d$-capacity can equal the channel capacity even if the above lower bound is strictly smaller. Other works addressing the problem of $d$-capacity or its special case of zero-error capacity include Merhav, Kaplan, Lapidoth, and Shamai [44], as well as its generalization to the case with arbitrary alphabets [45].

This problem has been studied extensively, and we emphasize that different choices of the code distribution lead to different bounds on the mismatch capacity. In [46], the Gallager upper bound on the average message error probability for DMCs under the random-coding regime was used to derive a bound that is referred to the *Generalized Mutual Information* (GMI). This bound is loosest of the above bounds, but it has the benefit of being applicable to channels with continuous alphabets. As was done in [47], the rate function in this bound is computed by using the Gärtner-Ellis theorem (large deviations principle: LDP).

A special class of mismatched decoders are nearest-neighbor decoders (minimum Euclidean distance decoders) that are often used on additive noise channels, even if the noise is not a white Gaussian process. Incurred performance loss of such decoders, in terms of the achievable rates over single-antenna fading channels, has been studied in [47] and [48]. While in [49] a modified nearest-neighbor decoder, using a weighting factor, for the fading multiple-antenna channel is introduced, and an expression of the GMI of its achievable rates is obtained. A similar investigation was carried out

in [50].

### Broadcast channels

The concept of broadcast channels (BCs) was introduced and first studied by Cover in [51]. It simply consists of a transmitter communicating information simultaneously to several receivers. We remark that this differs from a TV or radio broadcast, in which the transmitter sends the same message to each receiver. Here the transmitter sends different messages to each receiver.

In contrast with point-to-point systems, where the channel capacity is the maximum amount of information that the transmitter can send to the receiver, with arbitrary small error probability. In multi-user communications (with continuous or discrete alphabets), the transmitter can simultaneously transmit to more than one user, and consequently multi-user interference cancellation between different messages is needed. As a consequence, the channel capacity is the set of all simultaneously achievable rate vectors, which become an achievable rate region.

Consider a BC with only two receivers, which consists of an input $X \in \mathscr{X}$ and two outputs $(Y_1, Y_2) \in \mathscr{Y}_1 \times \mathscr{Y}_2$ with a transition probability function $W(y_1, y_2|x)$. The capacity region of this BC only depends on the marginal channels $W(y_1|x)$ and $W(y_2|x)$ (cf. [14], Theorem 14.6). So far conclusive results have been established for special cases only. An achievable rate region for degraded BCs has been proposed by Bergmans in [52]. The physically degraded BC is defined by assuming that $X \ominus Y_1 \ominus Y_2$ form a Markov chain (the output $Y_2$ is a noisy version of $Y_1$). By proving the converse of the corresponding coding theorem, Gallager [53] and Ahlswede [54] obtained the capacity region of BCs with degraded components. However the capacity region for a general non-degraded broadcast channel is still unknown. The largest achievable region for the general case is given by the Marton's region [55] by exploiting the idea of random binning coding (see also [56] for a short proof).

Assume that $(U_1, U_2) \in \mathscr{U}_1 \times \mathscr{U}_2$ are two auxiliary random variables with finite alphabets such that $(U_1, U_2) \ominus X \ominus (Y_1, Y_2)$ form a Markov chain. The Marton's region

(an inner bound of the capacity region) is the set of all rates $(R_1, R_2) \in \mathscr{R}(W)$

$$\begin{aligned}\mathscr{R}(W) = co\Big\{(R_1 \geq 0, R_2 \geq 0) : R_1 &\leq I\big(P_{U_1}, W\big), \\ R_2 &\leq I\big(P_{U_2}, W\big), \\ R_1 + R_2 &\leq I\big(P_{U_1}, W\big) + I\big(P_{U_2}, W\big) \\ &- I\big(P_{U_2}, P_{U_1|U_2}\big), \quad \text{for all } P(u_1, u_2, x) \in \mathscr{P}\Big\},\end{aligned} \tag{1.2}$$

where $co\{\cdot\}$ stands for the convex hull and $\mathscr{P}(\mathscr{U}_1 \times \mathscr{U}_2 \times \mathscr{X})$ denotes the set of all input probability distributions. A complete survey of these channels can be found in [57].

## 1.2 Research Context and Motivation

After a stellar growth over the 90's driven by voice as the killer app, wireless communications is now rapidly moving into a new era propelled by data networking, which has transformed from a niche technology into a vital component of most people's lives. The resultant requirement to combine mobile phone service and rapid growth of the Internet has created an environment where consumers desire seamless, high quality connectivity at all times and from all virtual locations. This brings many technical challenges. This spectacular growth is still occurring in cellular telephony and wireless networking, with no apparent end in sight. In order to satisfy user demand, resulting in constantly increasing of high-information rate transmission (without bandwidth increase), the desired quality of service (QoS) must be guaranteed for each user, even with very poor connection sessions. This means that the system designer must share the available resources (e.g. transmission and training power, number of training symbols, etc.) required to ensure the desired communication service (to achieve target information rates with small error probability).

Supporting the QoS in presence of imperfect channel knowledge is one of the critical requirements of single and multi-user wireless systems. In such communication systems channel estimation is usually performed at the receiver through use of pilot symbols transmitted at the beginning of each frame, and this knowledge is generally sent to the transmitter by some feedback. These channel estimates may strongly differ from the unknown channel, which is a real concern for the design of communication

systems guaranteeing the desired communication service. Specially for radio communications with mobile receivers, where the coherence time of the channel may be too short to permit reliable estimation to the receiver side of the time-varying parameters (the channel states) controlling the communication.

In the described scenario, most classic results concerning the theoretical communication limits and their optimal achieving schemes may turn out to be somewhat limited in practical applications, because these either directly or indirectly assume that the transmitter and receiver perfectly know the channel parameters. For instance, these limits do not incorporate any information about the imperfect channel knowledge. Thus, optimal coding schemes may not be as efficient as intended because its design does not take into account the characterization of the estimation performances. Furthermore, the practical importance of developing new theoretical limits assuming imperfect channel knowledge and QoS requirements, is that this can allow the system designer to decide how allocate the resources needed to achieve the desired communication service.

Therefore, studying the limits of reliable information rates in the case of imperfect channel estimation is an important problem from practical and theoretical viewpoint. This problem was previously tackled by Médard in [58], who derives an inner and outer bound of the capacity for AWGN channels with MMSE channel estimation at the receiver and no information at the transmitter. In [59] Yoo and Goldsmith extend these results to the multiple-antenna fading channel, assuming perfect feedback. This problem was also tackled by Hassibi and Hochwald in [60] for a block-fading channel with training sequences. These bounds are only depending on the variance of the channel estimation error regardless of the channel estimation method. Whereas, its extension to the case of general memoryless channels with an arbitrary estimator function follows from the general framework considered in this work.

This thesis first investigates the fundamental limits of reliable communication over wireless channels with QoS requirements, when the receiver and the transmitter only know noisy estimates (probably very poor estimates) of the channel parameters. As an attempt to deal with this problem of reliable communication over rapidly time-varying channels, an alternative approach consists in relying on the statistic characterizing the quality of channel estimates. This statistic can be used to define the notion of reliable

communication and its associated capacity. Furthermore, through this statistic it is possible to incorporate QoS requirements into the capacity expression.

In addition to studying theoretical limits, using this research outcome for our purpose, optimal decoding for practical communication systems allowing to achieve this capacity under imperfect channel estimation is also investigated. The results obtained in this investigation contain as a special case the improved decoding metric for space-time decoding of fading MIMO (Multiple-Input-Multiple-Output) channels proposed by Tarokh *et al.* [61] and Taricco and Biglieri [62].

Our main questions motivating this research are: (i) How to design communication systems to carry the maximum amount of information by using a minimum of resources, and (ii) how to correct them for imperfect channels knowledge.

Let us now move to a similar discussion concerning a downlink wireless communication channel, the multi-user broadcast channel. Consider, for example, a base station transmitting information over a downlink channel, where the base station (the transmitter) sends at the same time different informations to the mobiles (the receivers). In the case of wireless networks, as Fig. 1.1 shows, the base station may be transmitting a different voice call to a number of mobiles and simultaneously transferring data files to those and other users.

In the recent years, the multiple antenna Gaussian broadcast channel (MIMO-BC) has been extensively studied. Most of the literature focuses on the information-theoretic performances under the assumption on the instantaneous availability at both transmitter and all receivers of the channel matrices controlling the communication. Caire and Shamai in [63], have established an achievable rate region, referred to as the DPC region. They conjectured that this achievable region is the capacity. Recently in [64], Weingarten, Steinberg and Shamai prove this conjecture by showing that the DPC region is equal to the capacity region.

The great attraction of these channels is that under the assumption of perfect channel knowledge, as the signal-to-noise ratio (SNR) tends to infinity, the limiting ratio between the sum-rate capacity and the capacity of a single-user channel that results when the receiver allowed to cooperate is one. Thus, for broadcast channels where the receivers cannot cooperate, the interference cancellation implemented by DPC results in no asymptotic loss. However, as well as for single user wireless

Figure 1.1: Base station transmitting information over a downlink channel.

channels, the assumption of perfect channel knowledge is not applicable to practical BCs. The issue of the effect of the imperfect channel knowledge becomes more severe in this scenario, since the error on the channel estimation of some user affects the performances of many other users if e.g. multi-user interference cancellation is implemented. In particular, the problem may even be more complicated in the situations where no channel information is available at the transmitter, i.e., there is no feedback information from the receiver to the transmitter covering the channel estimates.

For instance, when the channel parameters are not perfectly known at both transmitter and all receivers, there are several questions that must be answered. For example:

(i) First, it is not immediately clear whether it is more efficient to send information to only a single user at a time rather than to use multiuser interference cancellation. Obviously, this answer will depend on the amount and quality of the information available at the transmitter and all receivers. Recently, Lapidoth, Shamai and Wigger [65] have shown that when the transmitter only has an estimate of the channel and the receivers perfectly know the channels, the limiting ratio between the sum-rate capacity and the capacity of a single-user channel with cooperating receivers is upper bounded by 2/3.

(ii) While it is well-known that for systems with perfect channel information significant gains can be achieved by adding antennas at the transmitter and/or receivers (cf. [66], [63]). It is natural to ask if also significant gains can be still achieved with imperfect channel estimation, without excessive increases in the amount of training.

(iii) As we mentioned before DPC scheme was proved to be the optimal way of achieving the boundary points of the capacity region of the MIMO-BC. Nevertheless, is DPC robust to channel estimation errors? if it is not, how to correct this?

The origins of DPC have started in the 1980s with the Gelfand and Pinsker's work [33], where the authors consider the capacity of discrete memoryless state-dependent channels with non-causal channel state information at the transmitter and without information at the receiver (called Gelfand and Pinsker's channel). In "Writing on Dirty Paper" [67], Costa applied this result to an additive white Gaussian noise (AWGN) channel corrupted by an additive Gaussian interfering signal (the channel states) that is non-causally known[2] at the transmitter. He showed the surprising result that choosing an adequate distribution for the codebooks, this channel achieves the same capacity as if the interfering signal was not present. Furthermore, the "interference cancellation" holds for arbitrary power values of the interfering signal compared to the transmission power. Several extensions of this result have been established for non-Gaussian interfering signals and non-stationary/non-ergodic Gaussian interference (cf. [68], [69]).

This result has gained considerable attention during the last years, mainly because of its potential use in communication scenarios where interference cancellation at the transmitter is needed. In particular, many new applications to information embedding (robust watermarking) in multimedia signal processing have emerged over the years [70]. Most notably is the idea of interference cancellation implemented by DPC scheme as well as the optimal way to embed information carrying-signals called *watermarks* into another signal (generally stronger) called *host* signal. The host signal is any multimedia signal, which can be either text, image, audio or video. The embedding must not introduce perceptible distortions to the host, and the watermark should survive common channel degradations. Applications of watermarking include copyright protection, transaction tracking, broadcast monitoring and tamper detection [71], e.g. the transmission of just one bit of information expected to be detectable with very low probability of false alarm, is sufficient to serve as an evidence of copyright.

This thesis investigates in an unified framework both scenarios: the capacity region

[2]The transmitter knows the channel states before beginning the transmission.

of multi-user MIMO broadcast channels and the capacity of channels with channel states non-causally known at the transmitter, under imperfect channel estimation. In addition to these theoretical limits, the role of multi-user state-dependent channels with non-causal channel state information at the transmitter in multiple information embedding is also studied. As well as for multi-user channels, multiple information embedding refers to the situation of embedding several messages into the same host signal, with or without different robustness and transparency requirements. Exploring these connections adds to the general understanding of multiple information embedding, and secondly, also allows us to establish new practical coding schemes.

## 1.3 Overview of Contributions

Through this thesis we address the following specific questions:

**1.** What are the theoretical limits of reliable transmission rates with imperfect channel estimation and quality of service requirements? (see chapter II)

**2.** How those limits can be achieved by using practical decoders in coded modulation systems? (see chapter III)

**3.** What are the fundamental capacity limits of state-dependent channels with non-causal channel state information at the transmitter in presence of imperfect channel knowledge: the fading Costa's channel and the multiple antenna BC? (see chapter IV)

**4.** Can multi-user information theory provide coding strategies for multiple information embedding applications? (see chapter V)

In Chapter 2 we address the above-mentioned channel mismatch scenario by introducing the notion of *estimation-induced outage capacity*, for which we provide an associated coding theorem and its strong converse, assuming a discrete memoryless channel. Basically, the transmitter and the receiver strive to construct codes for ensuring reliable communication with a given quality of service, no matter which degree of accuracy estimation arises during a transmission. In our setting, the quality of

service constraint stands for achieving target rates with small error probability (the desired communication service), even for very poor channel estimates.

We illustrate our ideas via numerical simulations for transmissions over single-user Ricean fading channels, with and without channel estimates available at the transmitter assuming maximum-likelihood (ML) channel estimation at the receiver. We also consider the effects of imperfect channel information at the transmitter, i.e., there is a rate-limited feedback link from the receiver back to the transmitter conveying the channel estimates. These results provide intuitive insights on the impact of the channel estimates and the channel characteristics (SNR, Ricean K-factor, training sequence length, feedback rate, etc.) on the mean outage capacity. For both perfect and rate-limited feedback channel, we derive optimal transmitter power allocation strategies that achieve the mean outage capacity.

In Chapter 3 we investigate the optimal decoder achieving this capacity with imperfect channel estimation. First, by searching into the family of nearest neighbor decoders, which can be easily implemented on most practical coded modulation systems, we derive a decoding metric that minimizes the average of the transmission error probability over all channel estimation errors. This metric, for arbitrary memoryless channels, achieves the capacity of a composite (more noisy) channel. Next, we specialize the general expression to obtain its corresponding decoding metric for fading MIMO channels.

According to the notion of estimation-induced outage rates introduced in Chapter 2, we characterize maximal achievable information rates associated to the proposed decoder. These achievable rates, for uncorrelated Rayleigh fading, are compared to both those of the classical mismatched ML decoder and the ultimate limits given by the estimation-induced outage capacity, which uses a theoretical decoder (i.e. the best possible decoder in presence of channel estimation errors). Numerical results show that the derived metric provides significant gains for the considered scenario, in terms of achievable information rates and bit error rate (BER), in a bit interleaved coded modulation (BICM) framework, without introducing any additional decoding complexity.

In Chapter 4 we examine the effect of imperfect channel estimation at the receiver with imperfect (or without) channel knowledge at the transmitter on the capacity of

state-dependent channels with non-causal channel state information at the transmitter. We address this problem through the notion of reliable communication based on the average of the error probability over all channel estimation errors, assuming a DMC. This notion allows us to consider the capacity of a composite (more noisy) Gelfand and Pinsker's channel. We first derive the optimal DPC scheme (assuming Gaussian codebooks) that achieves the capacity of the single-user fading Costa's channel with ML channel estimation. These results illustrate a practical trade-off between the amount of training and its impact to the interference cancellation performances of DPC scheme. These are useful in realistic scenarios of multiuser wireless communications and information embedding applications (e.g. robust watermarking). We also studied optimal training design adapted to each of these applications.

Next, we exploit the tight relation between the largest achievable rate region (Marton's region) for arbitrary BCs and channels with non-causal channel state information at the transmitter to extend this region to the case of imperfect channel knowledge. We then derive achievable rate regions and optimal DPC schemes, for a base station transmitting information over a multiuser Fading MIMO-BC, where the receivers only dispose of a noisy estimate of the channel parameters, and these estimates may be (or not) available to the transmitter. We provide numerical results for a two-users MIMO-BC with ML or minimum mean square error (MMSE) channel estimation. The results illustrate an interesting practical trade-off between the benefit of a high number of transmit antennas and the amount of training needed. In particular, we observe the surprising result that a BC with a single transmitter and receiver antenna, and imperfect channel estimation at the receivers, does not need the knowledge of estimates at the transmitter to achieve large rates.

In Chapter 5 we presents several implementable DPC based schemes for multiple user information embedding, through emphasizing their tight relationship with conventional multiple user information theory. We first show that depending on the targeted application and on whether the different messages are asked to have different robustness and transparency requirements, multiple user information embedding parallels one of the well-known multi-user channels with non-causal channel state information at the transmitter. The focus is on the Gaussian BC and the Gaussian Multiple Access Channel (MAC). For each of these channels, two practically feasible

transmission schemes are compared. The first approach consists in a straightforward -rather intuitive- superimposition of DPC schemes and the second consists in a joint design of these DPC schemes.

The joint approach is based on the ideal DPC for the corresponding channel. Our results extend on one side the practical implementations QIM, DC-QIM and SCS from the single user case to the multiple user one, and on another side provide a clear evaluation of the improvements brought by joint designs in practical situations. Then, we broaden our view to discuss the framework of more general lattice-based (vector) codebooks and show that the gap to full performance can be bridged up using finite dimensional lattice codebooks. Performance evaluations, including Bit Error Rates and achievable rate region curves are provided for both methods, illustrating the improvements brought by a joint design.

Finally, we discuss conclusions and possible extensions of this thesis in Chapter VI. The following table lists some abbreviations used throughout the thesis.

| | |
|---|---|
| QoS | Quality of Service |
| AWGN | Additive White Gaussian Noise |
| BC | Broadcast Channel |
| MAC | Multiple-Access Channel |
| DMC | Discrete Memoryless Channel |
| MIMO | Multiple Input Multiple Output (Multiple Antenna) |
| MIMO-BC | MIMO Broadcast Channel |
| DPC | Dirty Paper Coding |
| TDMA | Time-Division Multiple Access |
| CSI | Channel State Information |
| CSIR | Channel State Information at the Receiver |
| CSIT | Channel State Information at the Transmitter |
| CEE | Channels Estimation Errors |
| BICM | Bit Interleaved Coded Modulation |
| BER | Bit Error Rate |
| Tx | Transmitter |
| Rx | Receiver |
| PM | Probability Mass |
| PDF | Probability Density Function |
| QIM | Quantization Index Modulation |
| SCS | Scalar Costa Scheme |
| ML | Maximum-Likehood |
| MMSE | Minimum Mean Square Error |

Table 1.1: Table of abbreviations.

# Chapter 2

# Outage Behavior of Discrete Memoryless Channels Under Channel Estimation Errors

Classically, communication systems are designed assuming perfect channel state information at the receiver and/or transmitter. However, in many practical situations, only a noisy estimate of the channel is available that may strongly differs from the true channel. We address this channel mismatch scenario by introducing the notion of *estimation-induced outage capacity*, for which we provide an associated coding theorem and its strong converse, assuming a discrete memoryless channel.

Basically, the transmitter and the receiver strive to construct codes for ensuring reliable communication with a quality of service (QoS), no matter which degree of accuracy estimation arises during a transmission. In our setting, the quality of service constraint stands for achieving target rates with small error probability (the desired communication service), even for very bad channel estimates.

We illustrate our ideas via numerical simulations for transmissions over Ricean fading channels with different quality of services, without channel information at the transmitter and with maximum-likelihood (ML) channel estimation at the receiver. We also consider the effects of imperfect channel information at the transmitter, i.e., there is a rate-limited feedback link from the receiver back to the transmitter conveying the channel estimates. Our results provide intuitive insights on the impact of

the channel estimates and the channel characteristics (SNR, Ricean K-factor, training sequence length, feedback rate, etc.) on the mean outage capacity. For both perfect and rate-limited feedback channel, we derive optimal transmitter power allocation strategies that achieve the mean outage capacity. We furthermore compare our results with the achievable rates of a communication system where the receiver uses a mismatched ML decoder based on the channel estimate.

## 2.1 Introduction

Channel uncertainty, caused e.g. by time variations/fading, interference, or channel estimation errors, can severely impair the performance of wireless systems. Even if the channel is quasi-static and interference is small, uncertainty induced by imperfect channel state information (CSI) remains. As a consequence, studying the limits of reliable information rates in the case of imperfect channel estimation is an important problem. The various amount of information available to the transmitter and/or receiver and the error probability criteria of interest, capturing the channel uncertainty, lead to different capacity measures. Indeed, depending on the target communication and the available resources, each scenario has to identify the adequate notion of reliable transmission, so that in practice the resulting capacity matches well the observed rates.

In selecting a model for a communication scenario, several factors must be considered. These include the physical and statistical nature of the channel disturbances (e.g. fading distribution, channel estimation errors, practical design constraints, etc.), the information available to the transmitter and/or to the receiver and the presence of any feedback link from the receiver to the transmitter (for further discussions we refer the reader to [30]). Let us first review the model for communication under channel uncertainty over a memoryless channel with input alphabet $\mathscr{X}$ and output alphabet $\mathscr{Y}$ [30]. A specific instance of the unknown channel is characterized by a transition probability mass (PM) $W(\cdot|x,\theta) \in \mathcal{W}_\Theta$ with a fixed but unknown channel state $\theta \in \Theta \subseteq \mathbb{C}^d$. Here, $\mathcal{W}_\Theta = \big\{W(\cdot|x,\theta)\colon\, x \in \mathscr{X},\, \theta \in \Theta\big\}$ is a family of conditional transition PMs on $\mathscr{Y}$, parameterized by a random vector $\theta \in \Theta$ with probability density function (pdf) $\psi(\theta)$. In practical wireless systems we may distinguish two

different scenarios.

A first situation is described by two facts: (i) the transmitter and the receiver are designed without full knowledge of the characteristics of the law governing the channel variations $(\psi(\theta), \mathcal{W}_\Theta)$, (ii) the receiver may dispose only of a noisy estimate $\hat{\theta}$ of the CSI. A reasonable approach for this case consists in using mismatched decoders (cf. [34], [42], [40] and [44]). The decoding rule is restricted to be a metric of the interest, which perhaps is not necessarily matched to the channel. Recent additional results obtained by Lapidoth et. al. [48,72] show that in absence of CSI the asymptotic MIMO capacity grows double-logarithmically as a function of SNR. This line of work was initiated by Marzetta and Hochwald [73], and then explored by Zheng and Tse [74], to study the non-coherent capacity of MIMO channels under a block-fading assumption. The authors show that the capacity increases logarithmically in the SNR but with a reduced slope.

Another scenario concerns the case where the law governing the channel variations is known at the transmitter and at the receiver. Caire and Shamai [75] have examined the case of imperfect CSI at the transmitter (CSIT) and perfect CSI at the receiver (CSIR), so that power allocation strategies can be employed.

### 2.1.1 Motivation

The results recalled above are derived assuming that either no CSI or perfect CSI is available at the receiver. However, in many practical situations, the receiver disposes only of a noisy channel estimate (which may in some circumstances be a poor estimate). In that scenario, the resulting capacity will crucially relies on the error probability criteria adopted. On the other hand, most practical constraints of a communication system are concerned with the quality of service (QoS). These constraints require to guarantee a given target rate $R$ with small error probability for each user, no matter which degree of accuracy estimation arises during the communication. To this end, depending on the channel characteristics, the system designer must share the available resources (e.g. power for transmission and training, the amount of training used, etc.), so that the requirements can be satisfied.

Throughout the chapter we assume that the channel state, which neither the transmitter nor the receiver know exactly, remains constant within blocks of duration

$T$ symbol periods (coherence time), and these states for different blocks are i.i.d. $\theta \sim \psi(\theta)$. Note that the value of $T$ is related to the product of the coherence time and the coherence bandwidth of a wireless channel. The receiver only knows an estimate $\hat{\theta}_{\mathrm{R}}$ of the channel state and a *characterization of the estimator performance* in terms of the conditional pdf $\psi(\theta|\hat{\theta}_{\mathrm{R}})$ (this can be obtained using $\mathcal{W}_{\Theta}$, the estimation function and the *a priori* distribution of $\theta$). Moreover, a noisy feedback channel provides the transmitter with $\hat{\theta}_{\mathrm{T}}$, a noisy version of $\hat{\theta}_{\mathrm{R}}$ (e.g. due to quantization or feedback errors). In what follows we assume that $\theta \multimap \hat{\theta}_{\mathrm{R}} \multimap \hat{\theta}_{\mathrm{T}}$ form a Markov chain, with the joint distribution of $(\hat{\theta}_{\mathrm{T}}, \hat{\theta}_{\mathrm{R}}, \theta)$ given by $\psi(\hat{\theta}_{\mathrm{T}}, \hat{\theta}_{\mathrm{R}}, \theta)$. The scenario underlying these assumptions is motivated by current wireless systems, where e.g. $T$ for mobile receivers may be too short to permit reliable estimation of the fading coefficients. However, in spite of this difficulty, the system designer must guarantee the desired quality of service.

The concept of outage capacity was first proposed in [22] for fading channels. It is defined as the maximum rate that can be supported with probability $1 - \gamma_{_{QoS}}$, where $\gamma_{_{QoS}}$ is a prescribed outage probability. Furthermore, it has been shown that the outage probability matches well the error probability of actual codes (cf. [23, 24]). In contrast, ergodic capacity is the maximum information rate for which error probability decays exponentially with the code length. In our setting, a transceiver using $\hat{\theta} = (\hat{\theta}_{\mathrm{R}}, \hat{\theta}_{\mathrm{T}})$ instead of $\theta$ obviously might not support an information rate $R$, even if $R$ is less than the channel capacity under perfect CSI (even arbitrarily small rates might not be supported if $\hat{\theta}$ and $\theta$ happen to be strongly different). Consequently, *outages induced by channel estimation errors* will occur with a certain probability $\gamma_{_{QoS}}$. This outage probability depends on the codeword error probability, averaged over a random coding ensemble and over all channel realizations given the estimated state.

In this chapter we provide an explicit expression to evaluate the trade-off between the maximal outage rate versus the outage probability $\gamma_{_{QoS}}$, that we denote by *estimation-induced outage capacity* $\bar{C}(\gamma_{_{QoS}})$. Due to the independence of different blocks (coherence intervals), it is sufficient to study the estimation-induced outage rate $C(\gamma_{_{QoS}}, \hat{\theta})$ for a single block (cf. related discussions in [76]), for which the unknown channel state is fixed with estimate $\hat{\theta} = (\hat{\theta}_{\mathrm{T}}, \hat{\theta}_{\mathrm{R}})$. Then, we consider the

performance measure

$$\bar{C}(\gamma_{_{QoS}}) = E_{\hat{\theta}}\big\{C(\gamma_{_{QoS}}, \hat{\theta})\big\}, \tag{2.1}$$

which describes the average of information rates over all channel estimates $(\hat{\theta}_{\mathrm{T}}, \hat{\theta}_{\mathrm{R}})$, with prescribed outage probability $\gamma_{_{QoS}}$. The expectation in (2.1) is taken with respect to the joint distribution $\psi(\hat{\theta}) = \psi(\hat{\theta}_{\mathrm{T}}, \hat{\theta}_{\mathrm{R}})$ and reflects an average over a large number of coherence intervals. Our time-varying channel model is relevant for communication systems with small training overhead, where a quality of service in terms of achieving target rates with small error probability must be ensured, although significant channel variations occur, e.g. due to user mobility.

### 2.1.2 Related works

Assume a wireless channel where the coherence time is sufficiently long (this is often a reasonable assumption for a fixed wireless environment), then the transmitter can send a training sequence that allows the receiver to estimate the channel state. In this case, the average of the error probability over all channel estimation errors $\mathcal{E} = \theta - \hat{\theta}_{\mathrm{R}}$ seems to be a reasonable criterion to define the notion of reliable communication, together with the associated definition of achievable rates. By considering this notion of reliable communication, Medard [58] derives capacity bounds for additive white Gaussian noise (AWGN) channels with MMSE channel estimation at the receiver and no CSIT. These bounds are only depending on the variance of the estimation error $\sigma_{\mathcal{E}}^2$ regardless of the channel estimation method. These results have been extended to flat-fading channels in [77, 78]. Recent work by Yoo and Goldsmith [59] derives a capacity lower bound for MIMO fading channels by assuming a perfect feedback link. Unfortunately, Gaussian input distribution are not optimal inputs for maximizing the capacity. Because of the difficulty of computing this maximization only lower and upper bounds are known, these are tight for accurate estimations.

In our setting, this notion of reliable communication relied to the pdf of $\theta$ given $\hat{\theta}_{\mathrm{R}}$, corresponds to consider the capacity of the following *composite channel* model

$$\widetilde{W}(y|x, \hat{\theta}_{\mathrm{R}}) = \int_{\Theta} W(y|x, \theta) d\psi(\theta|\hat{\theta}_{\mathrm{R}}), \tag{2.2}$$

resulting from the average of the unknown channel $W(y|x,\theta)$ over all channel estimation errors, given the estimate $\hat{\theta}_{\mathrm{R}}$. The maximal achievable rate "the capacity", defined for the average of the error probability over all channel estimation errors, is given by

$$\widetilde{C}(\hat{\theta}) = \max_{P(\cdot|\hat{\theta}_{\mathrm{T}})\in\mathscr{P}(\mathscr{X})} I\big(P, \widetilde{W}(\cdot|\cdot,\hat{\theta}_{\mathrm{R}})\big), \tag{2.3}$$

where $I\big(P, \widetilde{W}(\cdot|\cdot,\hat{\theta}_{\mathrm{R}})\big)$ is the mutual information computed with the composite channel (2.2) and the input distribution $P \in \mathscr{P}(\mathscr{X})$. This expression is the capacity of general DMCs for the corresponding bounds found in [58] and [59]. Its proof follows from Shannon's coding theorem, since the resulting error probability of the composite channel is defined in terms of the conditional transition PM $\widetilde{W}(\cdot|x,\hat{\theta}_{\mathrm{R}})$ (cf. [7]). This capacity can be attained by using the maximum-likelihood (ML) decoding metric based on the transition PM (2.2).

The exposed notion of reliable communication, which leads to the capacity (2.3), reproduces well the observed rates in realistic communications when accurate channel estimates are available. However, if it is not the case, the average of the error probability over all estimation errors cannot ensure (in practice) reliable decoding in the case of significant channel variations and coarse estimations. Thus, the capacity measure (2.3) might be not adequate for communication systems with very small training overhead.

This chapter is organized as follows. In section 2.2, we first formalize the notion of *estimation-induced outage capacity* for general DMCs. Then, we present a coding theorem providing the explicit expression for the corresponding capacity. In section 2.3 the proof of the theorem and its converse are presented. An application example for the considered scenario involving a fading Ricean channel with AWGN, without feedback CSI and maximum likelihood (ML) channel estimation, is considered in section 2.4. The mean outage capacity is also compared to the achievable outage rates of a system using the mismatched ML decoder, based on the channel estimate. Then, assuming an instantaneous and error-free feedback, we derive optimal power allocation strategies that maximize the mean outage capacity over all channel estimates. We also consider the effect of rate-limited feedback CSI, deriving the corresponding power allocation strategies. Finally, section 2.5 provides simulations to illustrate

mean outage rates.

## 2.2 Estimation-induced Outage Capacity and Coding Theorem

In this section, we first develop a proper formalization of the notion of *estimation-induced outage capacity* and state a coding theorem.

*Note about notation:* Throughout this section, we use the following notation: $\mathscr{P}(\mathscr{X})$ denotes the set of all atomic (or discrete) probability masses (PMs) on $\mathscr{X}$ with finite number of atoms. Then the $n$th Cartesian power is defined as the sample space of $\mathbf{X} = (X_1, \ldots, X_n)$, with $P^n$-probability mass determined in terms of the $n$th Cartesian power of $P$. The joint PM corresponding to the input $P \in \mathscr{P}(\mathscr{X})$ and the transition PM $W(\cdot|x) \in \mathscr{P}(\mathscr{Y})$ is denoted as $W \circ P \in \mathscr{P}(\mathscr{X} \times \mathscr{Y})$, its marginal on $\mathscr{Y}$ denoted as $WP \in \mathscr{P}(\mathscr{Y})$. The alphabets $\mathscr{X}$ and $\mathscr{Y}$ are assumed finite, and their cardinality is denoted by $\|\cdot\|$, and the complement of any set $\mathscr{A}$ is denoted by $\mathscr{A}^c$. The functional $\mathcal{D}(\cdot\|\cdot)$ and $H(\cdot)$ respectively denote the *Kullback-Leibler* divergence and the entropy. The conditional versions are $\mathcal{D}(\cdot\|\cdot|\cdot)$ and $H(\cdot|\cdot)$, respectively. We use the notion of (conditional) information-typical (I-typical) sets defined in terms of (Kullback-Leibler) divergence, i.e., $\mathcal{T}_P^n(\delta) = \big\{\mathbf{x} \in \mathscr{X} : \mathcal{D}(\hat{P}_n\|P) \leq \delta\big\}$ and $\mathcal{T}_W^n(\mathbf{x}, \delta) = \big\{\mathbf{y} \in \mathscr{Y} : \mathcal{D}(\hat{W}_n\|W|\hat{P}_n) \leq \delta\big\}$ (for further details see Appendix A.1).

### 2.2.1 Problem definition

A message $m$ from the set $\mathcal{M} = \{1, \ldots, \lfloor \exp(nR) \rfloor\}$ is transmitted using a length-$n$ block code defined as a pair $(\varphi, \phi)$ of mappings, where $\varphi : \mathcal{M} \times \Theta \mapsto \mathscr{X}^n$ is the encoder (that makes only use of $\hat{\theta}_{\mathrm{T}}$), and $\phi : \mathscr{Y}^n \times \Theta \mapsto \mathcal{M} \cup \{0\}$ is the decoder (that makes only use of $\hat{\theta}_{\mathrm{R}}$). The random rate, which depends on the unknown channel realization $\theta$ and the estimate $\hat{\theta} = (\hat{\theta}_{\mathrm{T}}, \hat{\theta}_{\mathrm{R}})$ through the probability of error, is given by $\dfrac{1}{n} \log M_{\theta,\hat{\theta}}$. The maximum error probability over all messages is defined as

$$e_{\max}^{(n)}(\varphi, \phi, \hat{\theta}; \theta) = \max_{m \in \mathcal{M}} \sum_{\mathbf{y} \in \mathscr{Y}^n : \phi(\mathbf{y}, \hat{\theta}_{\mathrm{R}}) \neq m} W^n\big(\mathbf{y}|\varphi(m, \hat{\theta}_{\mathrm{T}}), \theta\big). \tag{2.4}$$

**Definition 2.2.1** *For a given channel estimate $\hat{\theta} = \hat{\theta}_0$, and $0 < \epsilon, \gamma_{_{QoS}} < 1$, an outage rate $R \geq 0$ is $(\epsilon, \gamma_{_{QoS}})$-achievable on an unknown channel $W(\cdot|x,\theta) \in \mathcal{W}_\Theta$, if for every $\delta > 0$ and every sufficiently large $n$ there exists a sequence of length-$n$ block codes such that the rate satisfies*

$$\Pr\left(\left\{\theta \in \Lambda_\epsilon^{(n)} : \, n^{-1}\log M_{\theta,\hat{\theta}} \, \geq \, R - \delta\right\}\big|\hat{\theta}\right) \geq 1 - \gamma_{_{QoS}}, \tag{2.5}$$

*where $\Lambda_\epsilon^{(n)} = \left\{\theta \in \Theta : \, e_{\max}^{(n)}(\varphi, \phi, \hat{\theta}; \theta) \leq \epsilon\right\}$ is the set of all channel states allowing for reliable decoding. This definition requires that maximum error probabilities larger than $\epsilon$ occur with probability less than $\gamma_{_{QoS}}$, i.e., $P_{\theta|\hat{\theta}}(\Lambda_\epsilon^{(n)}|\hat{\theta}) \geq 1 - \gamma_{_{QoS}}$.*

*A rate $R \geq 0$ is $\gamma_{_{QoS}}$-achievable if it is $(\epsilon, \gamma_{_{QoS}})$-achievable for every $0 < \epsilon < 1$. Let $C_\epsilon(\gamma_{_{QoS}}, \hat{\theta})$ be the largest $(\epsilon, \gamma_{_{QoS}})$-achievable rate for an outage probability $\gamma_{_{QoS}}$ and a given estimated $\hat{\theta}$. The* estimation-induced outage capacity *of this channel is then defined as the largest $\gamma_{_{QoS}}$-achievable rate, i.e., $C(\gamma_{_{QoS}}, \psi_{\theta|\hat{\theta}}, \hat{\theta}) = \lim\limits_{\epsilon \downarrow 0} C_\epsilon(\gamma_{_{QoS}}, \psi_{\theta|\hat{\theta}}, \hat{\theta})$.*

*Remark:* We would like to point out the main differences between the proposed notion of reliable communication and other notions such as: the average of the transmission error probability over all channel estimation errors and the classical definition of outage capacity.

(i) The practical advantage of the definition 2.2.1 is that for any degree of accuracy estimation, the transmitter and receiver are designed for ensuring reliable communication with probability $1 - \gamma_{_{QoS}}$, no matter which unknown state $\theta$ arises during a transmission. This definition provides a more precise measure of the reliability function compared to the classical definition that ensures reliable communication for the average of the transmission error probability over all channel estimation errors (i.e. the expectation of (2.4) over the pdf $\psi(\theta|\hat{\theta})$).

(ii) We emphasize the fundamental difference between definition 2.2.1 and the classical definition of information outage capacity, in which the instantaneous mutual information specifies the maximum rate with error-free communication[1] depending on each channel state. In the classical definition, when the transmission code rate is greater than the instantaneous mutual information an outage event occurs. In contrast, with channel estimation errors no error-free communications can be ensured,

[1]Here, error-free communications are understood in the sense of asymptotic arbitrary smaller error probabilities $\epsilon$.

the channel realization (even for the "best" ones). Thus, the decoding may fail due to the imperfect channel knowledge. As a consequence, this decoding error is captured by the outage probability that follows the statistic of the channel estimation errors.

In other words, the *estimation-induced outage capacity* is defined as the maximal rate, given an arbitrary channel estimate, ensuring error-free communication with probability $1-\gamma_{_{QoS}}$, i.e., for $(1-\gamma_{_{QoS}})\%$ of channel estimations.

### 2.2.2 Coding Theorem

We next state a theorem quantifying the *estimation-induced outage capacity* $C(\gamma_{_{QoS}},\hat{\theta})$ for our scenario $\hat{\theta}=(\hat{\theta}_{\mathrm{T}},\hat{\theta}_{\mathrm{R}})$ where $\theta \multimap \hat{\theta}_{\mathrm{R}} \multimap \hat{\theta}_{\mathrm{T}}$ form a Markov chain. This means that an estimate $\hat{\theta}_{\mathrm{R}}$ of the channel state is known at the decoder and only its noisy version $\hat{\theta}_{\mathrm{T}}$ is available at the encoder. Classically, we impose an input constraint that depends on the transmitter CSI, and require that $\Gamma(P)=\sum_{x\in\mathscr{X}}\Gamma(x)P(x|\hat{\theta}_{\mathrm{T}})$ is less than $\mathcal{P}(\hat{\theta}_{\mathrm{T}})$. Here, $\Gamma(\cdot)$ is an arbitrary non-negative function, and $P(\cdot|\hat{\theta}_{\mathrm{T}})\in\mathscr{P}_{\Gamma}$ denotes the input distribution depending on $\hat{\theta}_{\mathrm{T}}$ and $\mathscr{P}_{\Gamma}(\hat{\theta}_{\mathrm{T}})=\{P\in\mathscr{P}(\mathscr{X}):\ \Gamma(P)\leq\mathcal{P}(\hat{\theta}_{\mathrm{T}})\}$. Let $\mathcal{W}_{\Theta}=\big\{W(\cdot|x,\theta):\ x\in\mathscr{X},\,\theta\in\Theta\big\}$ be the family of DMCs, parameterized by a random vector $\theta\in\Theta$.

**Theorem 2.2.1** *Given* $0\leq\gamma_{_{QoS}}<1$ *the* estimation-induced outage capacity *of an unknown DMC* $W\in\mathcal{W}_{\Theta}$ *is given by*

$$C(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta})=\max_{P(\cdot|\hat{\theta}_{T})\in\mathscr{P}_{\Gamma}(\hat{\theta}_{T})}\mathscr{C}(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta},P),\tag{2.6}$$

*where*

$$\mathscr{C}(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta},P)=\sup_{\Lambda\subset\Theta:\ \Pr(\Lambda|\hat{\theta})\geq 1-\gamma_{_{QoS}}}\ \inf_{\theta\in\Lambda} I\big(P,W(\cdot|\cdot,\theta)\big).\tag{2.7}$$

*In addition,* $C_{\epsilon}(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta})=C(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta})$ *for all* $0<\epsilon<1$.

In this theorem, we used the mutual information

$$I\big(P,W(\cdot|\cdot,\theta)\big)=\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}}P(x)W(y|x,\theta)\log\frac{W(y|x,\theta)}{Q(y|\theta)},$$

with $Q(y|\theta)=\sum_{x\in\mathscr{X}}P(x)W(y|x,\theta)$. We emphasize that the supremum in (2.7) is taken over all subsets $\Lambda$ of $\Theta$ that have (conditional) probability at least $1-\gamma_{_{QoS}}$.

Theorem 2.2.1 provides an explicit way to evaluate the maximal outage rate versus outage probability $\gamma_{Q_{oS}}$ for an unknown channel that has been estimated with a given accuracy, characterized by $\psi(\theta|\hat{\theta})$.

*Remark:* (i) A proof of the Theorem 2.2.1 is needed because the classical definition of outage capacity in terms of instantaneous mutual information cannot be used since it requires perfect CSI which here is available neither at the transmitter nor at the receiver. A sketch of the proof of Theorem 2.2.1 is relegated to section 2.3. For further details and technical discussions the reader is referred to Appendix A.2. Observe that if perfect CSIR is available then $\Lambda_\epsilon = \Theta$, and the instantaneous mutual information is attainable. Thus, every rate $R$ can be associated to the set $\Lambda_R = \{\theta \in \Theta : I(P, W(\cdot|\cdot, \theta)) \geq R - \delta\}$ whose probability is $1 - \gamma_{Q_{oS}}$. Therefore, in that case with perfect CSI, the channel can be modeled as a compound channel (cf. [28]), whose transition probability depends on a random parameter $\theta \in \Theta$. However, in our setting this is different, since the instantaneous mutual information is not achievable and $\Lambda_\epsilon \subset \Theta$.

(ii) Theorem 2.2.1 is proved for DMCs by using well-known techniques based on typical sequences (cf. Appendix A.1). Extension of the concept of *types* to continuous alphabets are not known [3]. Consequently, for continuous-alphabet channels, the capacity analysis may need to be conducted over the weak topology (requiring completely different analytical tools from measure theory). Instead there are several continuous-alphabet problems whose simplest (or the only) available solution relies upon the method of types, via discrete approximations. For example, the proof of a general version of Sanov's theorem in [79], or the capacity subject to a state constraint of an AVC with general alphabets and states have been determined in this way (cf. [80]). Theorem 2.2.1 can be extended in the same way to continuous alphabets, subject to some constraints, in locally compact Hausdorff (LCH) spaces, e.g. alphabets are like $\mathbb{R}^k$ (or $\mathbb{C}^k$) which are separable spaces. For simplicity, this extension is not included in this chapter.

### 2.2.3 Impact of the channel estimation errors on the estimation-induced outage capacity

To evaluate the rate loss due to imperfect channel estimation we first provide general bounds on the mean outage capacity (2.1). Note that with high-accuracy estimations, the conditional pdf $\psi(\theta|\hat{\theta})$ is close to a dirac distribution, and the resulting averaged outage rate is equal to the ergodic capacity $C_E$ with perfect CSI. We first compare the mean (over all channel estimates) outage rate $\bar{C}(\gamma_{_{QoS}})$ to the Ergodic capacity. Then, this maximal mean outage rate is compared to the average of the capacity (2.3), which is defined in terms of the average error probability.

Assume that the optimal set of probability distributions $\mathcal{W}_{\Lambda^*}$, which is obtained by maximizing expression (2.7) over all sets $\Lambda \subset \Theta$ having probability at least $1-\gamma_{_{QoS}}$, is a convex set. We also assume that the composite channel $\widetilde{W}_{\hat{\theta}} \in \mathcal{W}_{\Lambda^*}$[2], where $\widetilde{W}_{\hat{\theta}}$ is given by expression (2.2). Let $\bar{\theta}(\hat{\theta})$ be the channel state (depending on $\hat{\theta}$) that provides the infimum in (2.7). Under these conditions and assuming any PM $P \in \mathscr{P}(\mathscr{X})$ the following inequalities hold,

$$\bar{C}(\gamma_{_{QoS}}) \leq C_E - E_{\theta,\hat{\theta}}\big[\mathcal{D}(W_\theta \| W_{\bar{\theta}(\hat{\theta})}|P) - \mathcal{D}(W_\theta P \| W_{\bar{\theta}(\hat{\theta})}P)\big], \tag{2.8}$$

$$\bar{C}(\gamma_{_{QoS}}) \leq E_{\hat{\theta}}\big[\widetilde{C}(\hat{\theta})\big] - E_{\hat{\theta}}\big[\mathcal{D}(\widetilde{W}_{\hat{\theta}} \| W_{\bar{\theta}(\hat{\theta})}|P) - \mathcal{D}(\widetilde{W}_{\hat{\theta}} P \| W_{\bar{\theta}(\hat{\theta})}P)\big]. \tag{2.9}$$

The second term on the right side of both inequalities is a positive quantity; and the equality only holds for linear families of probability distributions. The proof of both inequalities follows as consequence of Theorem A.3.1 in Appendix A.3. We emphasize that our setting requires reliable transition for $(1-\gamma_{_{QoS}})\%$ of channels (or estimates), which differers than the average of channel estimation errors. Consequently, smaller values of $\bar{C}(\gamma_{_{QoS}})$ are expected, comparing to those obtained through the average of the error probability $E_{\hat{\theta}}\big[\widetilde{C}(\hat{\theta})\big]$.

## 2.3 Proof of the Coding Theorem and Its Converse

In this section we approach the problem of determining the capacity by using the tools of information theory, according to the definition in section 2.2.1. The proof of

[2]Often this is a reasonable assumption with small outage probabilities $0 \leq \gamma_{_{QoS}} < 1$.

Theorem 2.2.1 is based on an extension of the maximal code lemma [17] to bound the minimum size of the images for the considered channels, according to the notion of *estimation-induced outage capacity*. This extension is based on robust I-typical sets (further details are provided in Appendix A.2).

### 2.3.1 Generalized Maximal Code Lemma

Let $\mathscr{I}_\Lambda$ denote the set of all common $\eta$-images $\mathscr{B}^n \subseteq \mathscr{Y}^n$ associated to a set $\mathscr{A}^n \subset \mathscr{X}^n$ via the collection of simultaneous DMCs $\mathcal{W}_\Lambda$,

$$\mathscr{I}_\Lambda(\mathscr{A}^n,\eta)=\Big\{\mathscr{B}^n\colon \inf_{\theta\in\Lambda} W^n(\mathscr{B}^n|\mathbf{x},\theta)\geq\eta \text{ for all } \mathbf{x}\in\mathscr{A}^n\Big\}.$$

In the following, we denote as

$$\mathrm{g}_\Lambda(\mathscr{A}^n,\eta)=\min_{\mathscr{B}^n\in\mathscr{I}_\Lambda(\mathscr{A}^n,\eta)}\|\mathscr{B}^n\|, \tag{2.10}$$

the minimum of the cardinalities of all common $\eta$-images $\mathscr{B}^n$. For a given channel estimate $\hat{\theta}=(\hat{\theta}_{\mathrm{T}},\hat{\theta}_{\mathrm{R}})$ with degraded CSIT $\theta \ominus \hat{\theta}_{\mathrm{R}} \ominus \hat{\theta}_{T}$, a code $\big(\mathbf{x}_1(\hat{\theta}_{\mathrm{T}}),\ldots,\mathbf{x}_M(\hat{\theta}_{\mathrm{T}});\ \mathscr{D}_1^n(\hat{\theta}),\ldots,\mathscr{D}_M^n(\hat{\theta})\big)$ according to the definition provided in section 2.2.1 consists of a set of codewords $\mathbf{x}_m(\hat{\theta}_{\mathrm{T}})$ and associated decoding sets $\mathscr{D}_m^n(\hat{\theta})$ (i.e., the decoder reads $\phi(\mathbf{y},\hat{\theta})=m$ iff $\mathbf{y}\in\mathscr{D}_m^n(\hat{\theta})$). For any set $\mathscr{A}^n$, we call a code admissible if: (i) $\mathbf{x}_m(\hat{\theta}_{\mathrm{T}})\in\mathscr{A}^n$, (ii) all decoding sets $\mathscr{D}_m^n(\hat{\theta})\subseteq\mathscr{Y}^n$ are mutually disjoint, and (iii) the set

$$\Lambda_\epsilon=\Big\{\theta\in\Theta:\max_{m\in\mathcal{M}} W^n\big((\mathscr{D}_m^n(\hat{\theta}))^c|\mathbf{x}_m(\hat{\theta}_{\mathrm{T}}),\theta\big)\leq\epsilon\Big\}, \tag{2.11}$$

satisfies $\Pr(\Lambda_\epsilon|\hat{\theta})\geq 1-\gamma_{QoS}$. Any input distribution satisfying the input constraint $\mathscr{P}(\hat{\theta}_{\mathrm{T}})$ is denoted as $P(\cdot|\hat{\theta}_{\mathrm{T}})$.

**Theorem 2.3.1** *Let two arbitrary numbers $0<\epsilon,\delta<1$ be given. There exists a positive integer $n_0$ such that for all $n\geq n_0$ the following two statements hold.*

*1) Direct Part: For any $\mathscr{A}^n\subset\mathcal{T}^n_{P|\hat{\theta}_T}(\delta,\hat{\theta}_T)$ and any random set $\Lambda\subset\Theta$ with $\Pr(\Lambda|\hat{\theta})\geq 1-\gamma_{QoS}$, there exists an admissible sequence of length-$n$ block codes of size*

$$M_{\theta,\hat{\theta}}\geq\exp\big[-n\big(H(\mathcal{W}_\Lambda|P)-\delta\big)\big]\,g_\Lambda(\mathscr{A}^n,\epsilon-\delta), \tag{2.12}$$

*for all $\theta\in\Lambda$, where $\Lambda_\epsilon=\Lambda$.*

*2) Converse Part: For* $\mathscr{A}^n = \mathcal{T}^n_{P|\hat{\theta}_T}(\delta, \hat{\theta}_T)$*, the size of any admissible sequence of length-n block codes is bounded as*

$$M_{\theta,\hat{\theta}} \leq \exp\big[-n\big(H(\mathcal{W}_{\Lambda_\epsilon}|P) + \delta\big)\big]\, g_{\Lambda_\epsilon}(\mathscr{A}^n, \epsilon + \delta), \tag{2.13}$$

*for all* $\theta \in \Lambda_\epsilon$*.*

The proof of this theorem easily follows from basic properties of I-typical sequences and the concept of robust I-typical sets, recalled in Appendix A.2. Whereas, Theorem 2.2.1 is obtained based on the following corollary.

**Corollary 2.3.1** *For a given channel estimate* $\hat{\theta}$*, a given outage probability* $\gamma_{_{QoS}}$*, any* $0 < \epsilon, \delta < 1$ *and any PM* $P(\cdot|\hat{\theta}_T) \in \mathscr{P}(\mathscr{X})$*, let* $\mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P)$ *be defined by expression* (2.7)*. Then the following statements holds:*

*(i) There exists an optimal sequence of block codes of length n and size* $M_{\theta,\hat{\theta}}$*, whose maximum error probabilities larger than* $\epsilon$ *occur with probability less than* $\gamma_{_{QoS}}$*, such that*

$$\Pr\Big(n^{-1}\log M_{\theta,\hat{\theta}} \geq R - 2\delta\big|\hat{\theta}\Big) \geq 1 - \gamma_{_{QoS}} \tag{2.14}$$

*for all rate* $R \leq \mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P)$*, provided that* $n \geq n_0(|\mathscr{X}|, |\mathscr{Y}|, \epsilon, \delta)$*.*

*(ii) For any block codes of length n, size* $M_{\theta,\hat{\theta}}$ *and codewords in* $\mathcal{T}^n_{P|\hat{\theta}_T}(\delta, \hat{\theta})$*, whose maximum error probabilities larger than* $\epsilon$ *occur with probability less than* $\gamma_{_{QoS}}$*, the largest code size satisfies*

$$\Pr\Big(n^{-1}\log M_{\theta,\hat{\theta}} > R + 2\delta\big|\hat{\theta}\Big) < \gamma_{_{QoS}} \tag{2.15}$$

*for all rate* $R \geq \mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P)$*, whenever* $n \geq n_0(|\mathscr{X}|, |\mathscr{Y}|, \epsilon, \delta)$*.*

***Proof:*** *From the direct part of Theorem 2.3.1 and Lemma A.2.2, it is easy to see that there exists admissible codes such that*

$$n^{-1}\log M_{\theta,\hat{\theta}} \geq n^{-1}\log g_\Lambda\big(\mathscr{A}^n, \epsilon - \delta\big) - H(\mathcal{W}_\Lambda|P) - \delta, \tag{2.16}$$

*for all* $\theta \in \Lambda$ *and sets* $\Lambda \subset \Theta$ *(having probability at least* $1 - \gamma_{_{QoS}}$*). Let* $\hat{\mathscr{D}}^n$ *be the common* $(\epsilon - \delta)$*-image of minimal size* $\|\hat{\mathscr{D}}^n\| = g_\Lambda\big(\mathscr{A}^n, \epsilon - \delta\big)$*. Then it is easy to show*

*that* $\inf_{\theta\in\Lambda} W_\theta P^n(\hat{\mathscr{D}}^n) \geq (\epsilon-\delta)^2$. *By applying Lemma A.1.4 (see Appendix A.1) to this relation and substituting it in (2.16), we obtain for all* $n \geq n_0'(|\mathscr{X}|, |\mathscr{Y}|, \epsilon, \delta)$,

$$\begin{aligned} n^{-1}\log M_{\theta,\hat{\theta}} &\geq \sup_{\theta\in\Lambda} H(W_\theta P) - H(\mathcal{W}_\Lambda|P) - 2\delta \\ &\geq \inf_{\theta\in\Lambda} I(P, W(\cdot|\cdot,\theta)) - 2\delta, \end{aligned} \tag{2.17}$$

*for all* $\theta \in \Lambda$, *where the last inequality follows from the concavity of the entropy function with respect to* $W_\theta$. *Finally, taking the supremum in* (2.17) *with respect to all sets* $\Lambda \subset \Theta$ *having probability at least* $1-\gamma_{_{QoS}}$ *yields the lower bound (2.14)*

$$\begin{aligned} n^{-1}\log M_{\theta,\hat{\theta}} &\geq \mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P) - 2\delta \\ &\geq R - 2\delta, \end{aligned} \tag{2.18}$$

*for all rate* $R \leq \mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P)$ *and* $\theta \in \Lambda^*$, *which is attained by some code with* $\Lambda_\epsilon = \Lambda^*$.

*Next we prove the upper bound (2.15). From the converse part of Theorem 2.3.1 and Proposition A.2.1, we have*

$$n^{-1}\log M_{\theta,\hat{\theta}} \leq n^{-1}\log g_{\Lambda_\epsilon}\big(\mathscr{A}^n, \epsilon+\delta\big) - H(\mathcal{W}_{\Lambda_\epsilon}|P) + \delta, \tag{2.19}$$

*for all* $\theta \in \Lambda_\epsilon$. *Since* $\mathscr{A}^n = \mathcal{T}^n_{P|\hat{\theta}_T}(\delta, \hat{\theta})$ *implies that any common* $(\epsilon+\delta)$*-image of* $\mathscr{A}^n$ *will be included in* $\bigcap_{\theta\in\Lambda_\epsilon} \mathcal{T}^n_{W_\theta P}(\delta_n')$, *Proposition A.1.1-(iv) (see Appendix A.1) ensures that there exists* $n \geq n_0''(|\mathscr{X}|, |\mathscr{Y}|, \epsilon, \delta)$ *such that,*

$$n^{-1}\log g_{\Lambda_\epsilon}\big(\mathscr{A}^n, \epsilon+\delta\big) \leq \inf_{\theta\in\Lambda_\epsilon} H(W_\theta P) + \delta. \tag{2.20}$$

*Then by applying equation (2.20) to equation (2.19), and then by taking its supremum with respect to all sets* $\Lambda \subset \Theta$ *having probability at least* $1-\gamma_{_{QoS}}$, *we obtain*

$$\begin{aligned} n^{-1}\log M_{\theta,\hat{\theta}} &\leq \mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P) + 2\delta, \\ &\leq R + 2\delta. \end{aligned} \tag{2.21}$$

*for all* $R \geq \mathscr{C}(\gamma_{_{QoS}}, \hat{\theta}, P)$ *and* $\theta \in \Lambda_\epsilon$ *with* $\Pr(\theta \notin \Lambda_\epsilon|\hat{\theta}) < \gamma_{_{QoS}}$, *and this concludes the proof.* ∎

We note that, codes achieving capacity (2.7) can be viewed as codes for a simultaneous channel $\mathcal{W}_{\Lambda^*}$, which has been determined by the decoder. Hence, this outage capacity $C(\gamma_{_{QoS}}, \hat{\theta})$ is seen to equal the maximum capacity of all compound channels that are contained in $\mathcal{W}_\Theta$ and, conditioned on $\hat{\theta}$, have sufficiently high probability.

## 2.4 Estimation-induced Outage Capacity of Ricean Channels

In this section, we illustrate our results via a realistic single user mobile wireless system involving a Ricean block flat-fading channel, where the channel state is described by a single fading coefficient. The channel states of each block are assumed i.i.d. and unknown at both transmitter and receiver. Each of these blocks are preceded by a length-$N$ training sequence $\mathbf{x}_T = [x_0, \ldots, x_{N-1}]$ known by the receiver. This enables maximum-likelihood (ML) estimation of the fading coefficient $\theta$ at the receiver yielding the estimate $\hat{\theta}_{\mathrm{R}}$.

In many wireless systems, CSI at the transmitter is provided by the receiver via a feedback channel. This allows the transmitter to perform power control. Below, we consider the following three feedback schemes: (i) no feedback channel is available, i.e., absence of CSIT. We compare our results with the capacity of a system where the receiver uses a mismatched ML decoder based on $\hat{\theta}_{\mathrm{R}}$; (ii) an instantaneous and error-free feedback channel is available ($\hat{\theta}_{\mathrm{T}} = \hat{\theta}_{\mathrm{R}}$); (iii) an instantaneous and rate-limited feedback channel is available. Here the CSI is quantized using a quantization codebook which is known at both transmitter and receiver (we construct this codebook using the well-known Lloyd-Max algorithm [81]).

### 2.4.1 System Model

We consider a single user, narrowband and block flat-fading communication model for wireless environments given by (all quantities are complex-valued)

$$Y[i] = H[i]X[i] + Z[i]. \tag{2.22}$$

Here, $Y[i]$ is the discrete-time received signal, $X[i]$ denotes the transmit signal, $H[i]$ is the fading coefficient, and $Z[i]$ is the additive noise. The transmit signal is subject to the average power constraint $\Gamma(P) = E_P\{|X[i]\}^2 \leq \mathcal{P}(\hat{\theta}_{\mathrm{T}})$ with $E_{\hat{\theta}_{\mathrm{T}}}\{\mathcal{P}(\hat{\theta}_{\mathrm{T}})\} \leq \bar{P}$, and the noise $Z[i]$ is i.i.d. zero-mean, circularly complex Gaussian, i.e., $Z(i) \sim C\mathcal{N}(0, \sigma_Z^2)$. To model Ricean fading, the channel state $\theta = H[i]$ is assumed to be circularly complex Gaussian with mean $\mu_h$ and variance $\sigma_h^2$, $\theta \sim \psi(\theta) = C\mathcal{N}(\mu_h, \sigma_h^2)$. The Rice

factor is defined as $K_h = \frac{|\mu_h|^2}{\sigma_h^2}$. Furthermore, noise and fading coefficient are statistically independent and their statistics are known at the encoder and decoder. Note that (2.22) models a memoryless channel with channel law $W(\cdot|x,\theta) = C\mathcal{N}(\theta x, \sigma_Z^2)$. The mutual information $I(X;Y|H=h)$ of this channel is maximized with an input distribution for $X[i]$ that is circularly complex Gaussian with zero mean and variance $\mathcal{P}(\hat{\theta}_{\mathrm{T}})$.

Assume that the specific realization of the complex fading coefficient $H[i]$ is unknown at the transmitter and at the receiver side but fixed during a coherence interval. Furthermore, a maximum-likehood (ML) estimate $\hat{\theta}_{\mathrm{R}} = \hat{H}[i]$ of $H[i]$ is assumed to be known at the receiver; this can be achieved by dedicating in each block a short time period to training. In particular, before sending a codeword, at the beginning of each block a training sequence $\mathbf{x}_T$ of length $N$ and total power $\|\mathbf{x}_T\|^2 = NP_T$ that is known by the receiver is transmitted. Within the training period, this results in an instantaneous signal-to-noise ratio (SNR)

$$\mathrm{SNR}_T = \frac{NP_T}{\sigma_Z^2}. \tag{2.23}$$

Note that in this model we have not considered the expense of the power used in training. The ML estimate of $\theta = H[i]$ using the receive sequence $\mathbf{y}_T = (y_0, \ldots, y_{N-1})$ corresponding to the training sequence $\mathbf{x}_T$ is given by

$$\hat{\theta}_{\mathrm{R}} = \frac{\mathbf{x}_T^H \mathbf{y}_T}{NP_T} = H + \mathcal{E}\,, \tag{2.24}$$

where $\mathcal{E} \sim C\mathcal{N}(0, \sigma_{\mathcal{E}}^2)$ with an estimation error given by $\sigma_{\mathcal{E}}^2 = E_{\theta|\hat{\theta}_{\mathrm{R}}}\left[(\theta - \hat{\theta}_{\mathrm{R}})^2 | \hat{\theta}_{\mathrm{R}}\right] = \mathrm{SNR}_T^{-1}$. The performance of this ML estimator can be characterized via the pdf of the channel state estimate,

$$\psi(\hat{\theta}_{\mathrm{R}}|\theta) = W^N\big(\mathcal{A}(\mathbf{x}_T, \hat{\theta}_{\mathrm{R}})|\mathbf{x}_T, \theta\big), \tag{2.25}$$

where $\mathcal{A}(\mathbf{x}_T, \hat{\theta}_{\mathrm{R}}) = \left\{\mathbf{y} \in \mathbb{C}^N : \frac{\mathbf{x}_T^H \mathbf{y}}{NP_T} = \hat{\theta}_{\mathrm{R}}\right\}$. With (2.25), this conditional pdf of the estimated state $\hat{\theta}$ can be shown to equal $\psi\big(\hat{\theta}_{\mathrm{R}}|\theta\big) = C\mathcal{N}\big(\theta, \sigma_{\mathcal{E}}^2\big)$. Using this pdf and the channel's *a priori* distribution $\psi(\theta)$, the *a posteriori* distribution of $\theta$ given $\hat{\theta}_{\mathrm{R}}$ can be expressed as

$$\psi(\theta|\hat{\theta}_{\mathrm{R}}) = \frac{\psi(\hat{\theta}_{\mathrm{R}}|\theta)\psi(\theta)}{\displaystyle\int_{\mathbb{C}} \psi(\hat{\theta}_{\mathrm{R}}|\theta) d\psi(\theta)} = C\mathcal{N}\big(\tilde{\mu}(\hat{\theta}_{\mathrm{R}}), \tilde{\sigma}^2\big), \tag{2.26}$$

where

$$\tilde{\mu}(\hat{\theta}_{\mathrm{R}}) = \rho\mu_h + (1-\rho)\hat{\theta}_{\mathrm{R}}, \quad \text{with} \quad \rho = \frac{\sigma_{\mathcal{E}}^2}{\sigma_{\mathcal{E}}^2 + \sigma_h^2} \tag{2.27a}$$

$$\tilde{\sigma}^2 = \rho\sigma_h^2. \tag{2.27b}$$

### 2.4.2 Global Performance of Fading Ricean Channels

Evaluating (2.7) requires to solve an optimization problem where we have to determine the optimum set $\Lambda^*$, and the associated channel state $\theta^* \in \Lambda^*$ minimizing mutual information. However, in our case it can be observed that the mutual information depends only on $|\theta|$. Thus, for the optimization we can replace the sets $\Lambda$ of complex fading coefficients with sets $\tilde{\Lambda}$ of positive real values $r = |\theta|$. For a given channel estimate $\hat{\theta}_0 = (\hat{\theta}_{\mathrm{T},0}, \hat{\theta}_{\mathrm{R},0})$ that corresponds to the ML estimate of $\theta$ and its corresponding feedback channel, the conditional pdf $\psi(\theta|\hat{\theta} = \hat{\theta}_0)$ can be easily obtained from (2.26). Using these results, the pdf of $r = |\theta|$ given the estimated channel $\hat{\theta}_0$ can be shown to be Ricean:

$$\psi\big(r|\hat{\theta} = \hat{\theta}_0\big) = \frac{r}{\tilde{\sigma}^2/2} \exp\left(-\frac{r^2 + |\tilde{\mu}(\hat{\theta}_{\mathrm{R},0})|^2}{\tilde{\sigma}^2}\right) I_0\left(\frac{|\tilde{\mu}(\hat{\theta}_{\mathrm{R},0})|r}{\tilde{\sigma}^2/2}\right). \tag{2.28}$$

Here, $I_0$ is the zero'th order modified Bessel function of the first kind, and $\tilde{\mu}(\hat{\theta})$ and $\tilde{\sigma}^2$ are specified in (2.27). Consequently, the optimization problem now reduces to finding the optimum positive real interval $\tilde{\Lambda}^* = [r^*, \infty[$ having probability $1-\gamma_{_{QoS}}$ (computed with the pdf in (2.28)). This follows from the fact that the mutual information is a monotone and increasing function in $r$. Moreover, the optimal set $\tilde{\Lambda}^*$ is convex and compact, thus the infimum in the capacity expression actually equals the minimum capacity value over all $r$ in the set $\tilde{\Lambda}^*$. It follows that $r^*$ is the $\gamma_{_{QoS}}$-percentile[3] of $\psi\big(r|\hat{\theta} = \hat{\theta}_0\big)$:

$$\Pr\left(\theta \in \tilde{\Lambda}^* | \hat{\theta} = \hat{\theta}_0\right) = \int_{r^*}^{\infty} d\psi\big(r|\hat{\theta} = \hat{\theta}_0\big) = 1 - \gamma_{_{QoS}}. \tag{2.29}$$

[3]Equation (2.29) can be computed by using the cumulative distribution of a non-central chi-square of two degrees of freedom.

Then, the *estimation-induced outage capacity*, with transmit power constrained to $\mathcal{P}(\hat{\theta}_{\mathrm{T},0})$, can be shown to be given by

$$C(\gamma_{Q_{oS}}, \hat{\theta}_0) = \log_2\left(1 + \frac{\left(r^*(\gamma_{Q_{oS}}, \hat{\theta}_0)\right)^2 \mathcal{P}(\hat{\theta}_{\mathrm{T},0})}{\sigma_Z^2}\right). \tag{2.30}$$

We use this expression to evaluate $\bar{C}(\gamma_{Q_{oS}})$ via the expectation with respect to $\hat{\theta}$ according to (2.1).

We finally note that $\lim_{N\downarrow\infty} \Pr\left(|\theta - \hat{\theta}_{\mathrm{R}}| > \varepsilon | \hat{\theta}_{\mathrm{R}}\right) \to 0$ for any $\varepsilon > 0$. Thus, $\Lambda^* = \{\theta \in \Theta : |\theta - \hat{\theta}_{\mathrm{R}}| \leq \epsilon\}$ contains a smaller and smaller neighborhood of the true parameter $\theta$ and hence by continuity $C(\gamma_{Q_{oS}}, \hat{\theta}) \to \log_2\left(1 + \frac{|\theta|^2 \mathcal{P}(\hat{\theta}_{\mathrm{T}})}{\sigma_Z^2}\right)$ as the training sequence length $N$ tends to infinity. Therefore, the mean outage capacity $\bar{C}(\gamma_{Q_{oS}})$ converges to the ergodic capacity with perfect CSI $C_{\mathrm{E}}$, i.e., $\bar{C}(\gamma_{Q_{oS}}) = E_{\hat{\theta}}\{C(\gamma_{Q_{oS}}, \hat{\theta})\} \to C_{\mathrm{E}}$ for any $0 < \gamma_{Q_{oS}} < 1$.

### 2.4.3 Decoding with the Mismatched ML decoder

Mismatched decoding arises when the decoder is restricted to use a prescribed "metric" $d(\cdot,\cdot)$, which does not necessarily match the channel [44]. Given an output sequence $\mathbf{y}$ and an estimated state $\hat{\theta}_{\mathrm{R}} = \hat{\theta}_0$, a mismatched ML decoder that uses the metric $d_{\hat{\theta}_0}(\mathbf{x}_i, \mathbf{y}) = \left\|\mathbf{y} - \hat{\theta}_0 \cdot \mathbf{x}_i\right\|^2$ declares that the codeword $i$ was sent iff $d_{\hat{\theta}_0}(\mathbf{x}_i, \mathbf{y}) < d_{\hat{\theta}_0}(\mathbf{x}_j, \mathbf{y})$, for all $j \neq i$. Of course, suboptimal performances are expected for this classical decoder, since it does not depends on the law $\psi(\theta|\hat{\theta})$ governing the channel estimation errors. However, we aim at comparing the maximum achievable outage rate (2.1) (obtained from expression (2.30)) with the achievable outage rates $\bar{C}_{\mathrm{ML}}(\gamma_{Q_{oS}})$ of a receiver using this mismatched ML decoding, which does not need to know the law governing the channel variations. For the channel model considered here, the capacity expression provided in [44] specializes to

$$C_{\mathrm{ML}}(\hat{\theta}_0, \theta) = \min_{\mu \in \mathbb{C}:\, \mathbb{R}e\{\mu\hat{\theta}_0\} \geq \mathbb{R}e\{\theta\hat{\theta}_0\}} \log_2\left(1 + \frac{|\mu|^2 \bar{P}}{(|\theta|^2 - |\mu|^2)\bar{P} + \sigma_Z^2}\right), \tag{2.31}$$

which solution is easily obtained as

$$C_{\mathrm{ML}}(\hat{\theta}_0, \theta) = \log_2\left(1 + \frac{|\eta^*|^2 |\hat{\theta}_0|^2 \bar{P}}{(|\theta|^2 - |\eta^*|^2 |\hat{\theta}_0|^2)\bar{P} + \sigma_Z^2}\right), \tag{2.32}$$

with $\eta^* = \dfrac{\mathbb{R}e\{\theta^\dagger \hat{\theta}_0\}}{|\hat{\theta}_0|^2}$. Then, the associated outage probability for a rate $R \geq 0$ is defined as

$$P_{\mathrm{ML}}^{\mathrm{out}}\big(R, \hat{\theta}_0\big) = \Pr\big(\Lambda_{\mathrm{ML}}(R, \hat{\theta}_0)\big|\hat{\theta} = \hat{\theta}_0\big),$$

with $\Lambda_{\mathrm{ML}}(R, \hat{\theta}_0) = \big\{\theta \in \Theta\colon\ C_{\mathrm{ML}}(\hat{\theta}_0, \theta) < R\big\}$, and the maximal outage rate for an outage probability $\gamma_{_{QoS}}$, $C_{\mathrm{ML}}(\gamma_{_{QoS}}, \hat{\theta}_0) = \sup\big\{R \geq 0\colon\ P_{\mathrm{ML}}^{\mathrm{out}}\big(R, \hat{\theta}_0\big) \leq \gamma_{_{QoS}}\big\}$. The average outage rate is then given by

$$\bar{C}_{\mathrm{ML}}(\gamma_{_{QoS}}) = E_{\hat{\theta}}\big\{C_{\mathrm{ML}}(\gamma_{_{QoS}}, \hat{\theta})\big\}. \tag{2.33}$$

Note that for real-valued channels, mismatched ML decoding becomes optimal and (2.32) equals the capacity of the true channel. Hence, a comparison would not make sense in that context.

### 2.4.4 Temporal power allocation for estimation-induced outage capacity

We have proved from (2.6) that the maximal achievable rate for a single user Ricean fading channel is given by (2.30). In this subsection we concentrate on deriving the optimal power allocation strategy to achieve the mean outage capacity (2.1). Since each codeword experiences an additive white Gaussian channel noise, random Gaussian codes with multiple codebooks are employed. Based on the channel estimate known at the transmitter $\hat{\theta}_{\mathrm{T}}$, a codeword is transmitted at a power level given by the optimal power allocation, as demonstrated in [76].

First consider a perfect feedback link from the receiver to the transmitter ($\hat{\theta} = \hat{\theta}_{\mathrm{T}} = \hat{\theta}_{\mathrm{R}}$). For simplicity, we assume an instantaneous and error-free feedback, but the generalization to introduce the effects of feedback delay is rather straightforward. Under these assumptions, from (2.1) and (2.30) the mean outage capacity is given by

$$\bar{C}(\gamma_{_{QoS}}) = \sup_{\mathcal{P}(\hat{\theta})\colon E_{\hat{\theta}}\{\mathcal{P}(\hat{\theta})\} \leq \bar{P}} \int_{\Theta} \log_2\left(1 + \frac{\big(r^*(\gamma_{_{QoS}}, \hat{\theta})\big)^2 \mathcal{P}(\hat{\theta})}{\sigma_Z^2}\right) d\psi(\hat{\theta}), \tag{2.34}$$

where the supremum is over all power allocation non-negative functions $\mathcal{P}(\hat{\theta})$ such that $E_{\hat{\theta}}\{\mathcal{P}(\hat{\theta})\} \leq \bar{P}$. Given a state measurement $\hat{\theta}$, the transmitter selects a code with a power level $\mathcal{P}(\hat{\theta})$ and uses $\hat{\theta}$ and the conditional pdf $\psi\big(r|\hat{\theta}\big)$ to compute $r^*(\gamma_{_{QoS}}, \hat{\theta})$.

Thus, the optimal power allocation maximizing (2.34) is easily derived as the well-known water-filling solution,

$$\mathcal{P}(\hat{\theta})/\sigma_Z^2 = \begin{cases} \dfrac{1}{r_0} - \dfrac{1}{r^*(\gamma_{_{QoS}}, \hat{\theta})}, & r^*(\gamma_{_{QoS}}, \hat{\theta}) \geq r_0 \\ 0, & r^*(\gamma_{_{QoS}}, \hat{\theta}) < r_0 \end{cases} \tag{2.35}$$

where $r_0$ is a positive constant ensuring the power constraint $E_{\hat{\theta}}\{\mathcal{P}(\hat{\theta})\} = \bar{P}$.

The developments so far have assumed an instantaneous and error-free feedback with non-rate-limited. Consider now the situation in which the decoder quantizes and sends to the transmitter the optimal solution $r^*(\gamma_{_{QoS}}, \hat{\theta}_{\mathrm{R}})$, by using an instantaneous and error-free but rate-limited feedback channel. Clearly, the performance is now a function of $R_{FB}$, the amount of feedback bits. In this case, the decoder must select a quantized value among $M_{FB} = \lfloor 2^{R_{FB}} \rfloor$ possibilities in the quantization codebook, which is assumed to be also known at the transmitter. This quantization codebook is usually designed to minimize the average squared error between the input value and the quantized value. For analytical simplicity, we construct the quantization codebook using the optimal non-uniform quantizer $Q\big[\,\cdot\,\big]$ given by the well-known Lloyd-Max algorithm [81]. Then to make benefit of the rate-limited feedback the power allocation (2.35) should be modified accordingly. Note that the considered quantization codebook is not necessarily optimal in the sense of maximizing mean outage rates. Optimal design of quantization codebooks, however, is a much difficult problem. The reason is that the cost function (not necessary the average squared error) can exploit any channel invariance, which may be present in the communication system. For example, in [82] phase-invariance of closed-loop beamforming were used to reduce the number of feedback parameters required (also see [83]).

Let $\hat{\theta}_{\mathrm{T}} \in \big\{\hat{\theta}_{\mathrm{T},1}, \ldots, \hat{\theta}_{\mathrm{T},M_{FB}}\big\}$ be the quantized value $\hat{\theta}_{\mathrm{T}} = Q\big[r^*(\gamma_{_{QoS}}, \hat{\theta}_{\mathrm{R}})\big]$ corresponding to the optimal solution for $r^*(\gamma_{_{QoS}}, \hat{\theta}_{\mathrm{R}})$, which is obtained at the decoder. In this case, by (2.1) and (2.30), the mean outage capacity with rate-limited feedback is given by

$$\bar{C}(\gamma_{_{QoS}}) = \sup_{\mathcal{P}(\hat{\theta}_{\mathrm{T}})} \sum_{i=1}^{M_{FB}} \Pr(\hat{\theta}_{\mathrm{T},i}) \int_{\Lambda_i} C\big(\gamma_{_{QoS}}, \hat{\theta}_{\mathrm{T},i}, \hat{\theta}_{\mathrm{R}}\big) d\psi(\hat{\theta}_{\mathrm{R}}|\hat{\theta}_{\mathrm{T},i}), \tag{2.36}$$

where the supremum is over all non-negative power allocation functions $\mathcal{P}(\hat{\theta}_{\mathrm{T}})$ such that $\sum_{i=1}^{M_{FB}} \mathcal{P}(\hat{\theta}_{\mathrm{T},i}) \Pr(\hat{\theta}_{\mathrm{T},i}) \leq \bar{P}$, and $\Pr(\hat{\theta}_{\mathrm{T},i}) = \Pr(\hat{\theta}_{\mathrm{T}} = \hat{\theta}_{\mathrm{T},i})$ denote the probability

for the state known at the transmitter $\hat{\theta}_{T,i}$ and $\Lambda_i = \left\{\hat{\theta}_{\mathrm{R}} \in \Theta : \hat{\theta}_{\mathrm{T},i} = Q\left[r^*(\gamma_{QoS}, \hat{\theta}_{\mathrm{R}})\right]\right\}$ is the set of states $\hat{\theta}_{\mathrm{R}}$ corresponding to the quantized state $\hat{\theta}_{\mathrm{T},i}$. It is immediate to see that the optimal power allocation function $\mathcal{P}(\hat{\theta}_{\mathrm{T}})$ must satisfy the power constraint with equality. Then, from the Lagrange multipliers and the Kuhn-Tucker conditions [84] we get that $\mathcal{P}(\hat{\theta}_{\mathrm{T}})$ is the solution maximizing (2.36) if it satisfies the following inequality

$$\int_{\Lambda_i} \frac{r^*(\gamma_{QoS}, \hat{\theta}_{\mathrm{R}})}{1 + \left(\dfrac{\mathcal{P}(\hat{\theta}_{\mathrm{T},i})}{\sigma_Z^2}\right) r^*(\gamma_{QoS}, \hat{\theta}_{\mathrm{R}})} d\psi(\hat{\theta}_{\mathrm{R}}|\hat{\theta}_{\mathrm{T},i}) \leq r_0, \tag{2.37}$$

for all $\hat{\theta}_{\mathrm{T},i} \in \left\{\hat{\theta}_{\mathrm{T},1}, \ldots, \hat{\theta}_{\mathrm{T},M_{FB}}\right\}$, with equality for all $\hat{\theta}_{\mathrm{T},i}$ such that $\mathcal{P}(\hat{\theta}_{\mathrm{T},i}) > 0$, where $r_0$ is a given positive constant whose value is fixed in order to satisfy the power constraint with equality. However, expression (2.37) shows that a closer solution to $\mathcal{P}(\hat{\theta}_{\mathrm{T},i})$ cannot be found.

Define a function $L_{\hat{\theta}_{\mathrm{T},i}}(r_0)$ denoting the left-hand side of (2.37) as a function of $r_0 \geq 0$, which is parameterized by $\hat{\theta}_{\mathrm{T},i}$. Then, for a given $\hat{\theta}_{T,i}$, $L_{\hat{\theta}_{\mathrm{T},i}}(r_0)$ is a positive decreasing function whose maximum value is $\bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i}) = E_{\hat{\theta}_{\mathrm{R}}|\hat{\theta}_{\mathrm{T}}}\left\{r^*(\gamma_{QoS}, \hat{\theta}_{\mathrm{R}})|\hat{\theta}_{\mathrm{T}} = \hat{\theta}_{\mathrm{T},i}\right\}$ and it is attained for $\mathcal{P} = 0$. Thus, the solution for (2.37) is parametrized as

$$\mathcal{P}(\hat{\theta}_{\mathrm{T},i}) = \begin{cases} L^{-1}_{\hat{\theta}_{\mathrm{T},i}}(r_0), & \text{if } 0 < r_0 < \bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i}) \\ 0, & \text{otherwise} \end{cases} \tag{2.38}$$

where the value of $r_0$ is determined by solving

$$\sum_{i=1}^{M_{FB}} \mathcal{P}(\hat{\theta}_{\mathrm{T},i}) \Pr(\hat{\theta}_{\mathrm{T},i}) = \bar{P}. \tag{2.39}$$

For practical computation we can parameterize both the average power $\bar{P}$ and the solution $\mathcal{P}(\hat{\theta}_{T,i})$ in terms of $r_0 \in [0, \max_{\hat{\theta}_{\mathrm{R},i}} \bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i})]$. Since $L^{-1}_{\hat{\theta}_{\mathrm{T},i}}(r_0)$ is decreasing in $r_0$, then $\bar{P}$ is also a decreasing function of $r_0$. For a given $r_0$ (i.e. given $\bar{P}$), positive power is allocated only for values $\hat{\theta}_{\mathrm{T},i} \in \left\{\hat{\theta}_{\mathrm{T},1}, \ldots, \hat{\theta}_{\mathrm{T},M_{FB}}\right\}$ such that $\bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i}) > r_0$. Consequently, this optimal power allocation $\mathcal{P}(\hat{\theta}_{\mathrm{T},i})$ has a water-filling nature, similar to the optimal power allocation in the case of non-rate-limited feedback, found in (2.35). However, obtaining the optimal solution of $\mathcal{P}(\hat{\theta}_{\mathrm{T}})$ may be computationally intensive. We have observed that in most applications, rates close to the optimal can

be achieved using the following suboptimal power allocation function:

$$\mathcal{P}(\hat{\theta}_{\mathrm{R},i})/\sigma_Z^2 = \begin{cases} \dfrac{1}{r_0} - \dfrac{1}{\bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i})}, & \bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i}) \geq r_0 \\ 0, & \bar{r}(\gamma_{QoS}, \hat{\theta}_{\mathrm{R},i}) < r_0 \end{cases} \tag{2.40}$$

where $r_0$ is determined by the power constraint (2.39).

## 2.5 Simulation results

In this section, numerical results are presented based on Monte Carlo simulations. We consider the three scenarios described in section 2.4 that are motivated by real environments of mobile wireless systems.

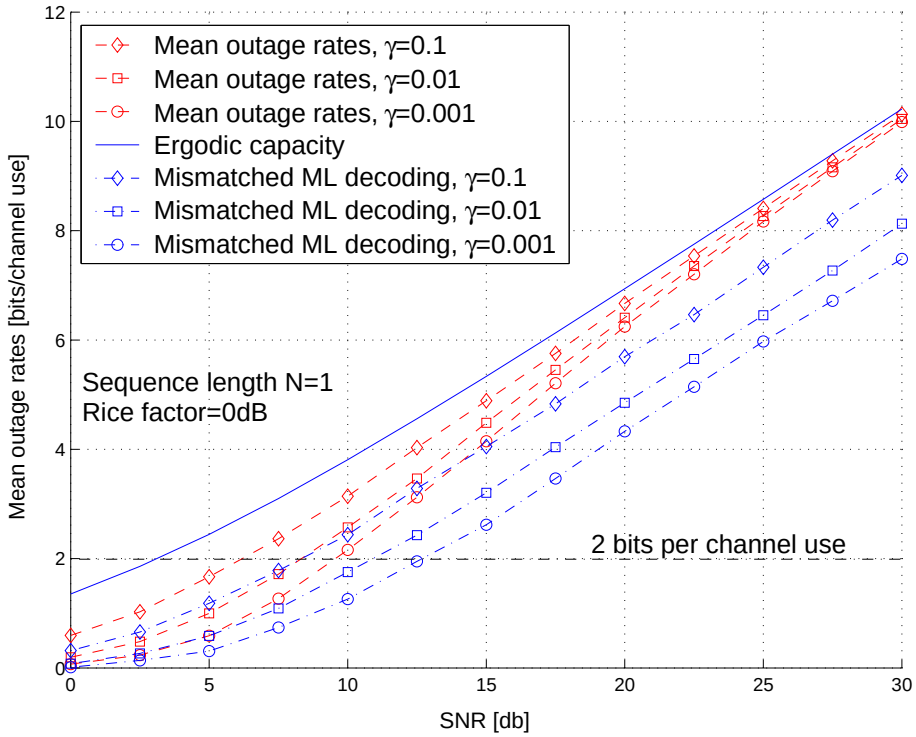



Figure 2.1: Average of estimation-induced outage capacity without feedback (no CSIT) and achievable rates with mismatched ML decoding vs SNR, for various outage probabilities.

(i) We suppose a communication system where no CSIT is available. Fig. 2.1 shows the average of *estimation-induced outage capacity* $\bar{C}(\gamma_{QoS})$ from (2.1) (in bits per channel use) versus the signal-to-noise ratio $\mathrm{SNR} = |\mu_h|^2\bar{P}/\sigma_Z^2$ for different outage probabilities $\gamma_{QoS} = \{10^{-1}, 10^{-2}, 10^{-3}\}$. Here, the transmitter does not know the channel estimate, and consequently no power control is possible. The channel's Rice

factor was $K_h = 0\,\text{dB}$, the power and the length of the training sequence are $P_T = P$ and $N = 1$, respectively. Note that with this length, e.g. at $\text{SNR} = 0\,\text{dB}$ ($= \text{SNR}_T$), the estimation error is still large ($\sigma_\mathcal{E}^2 = 1$) to use the notion of reliable communication based on the average of the error probability over all channel estimation errors. This scenario has been outlined in the introduction section, exposing that the *estimation-induced outage capacity* provides a more realistic measure of the limits of reliable rates effectively supported. For comparison, we also show the mean outage rate $\bar{C}_{\text{ML}}(\gamma_{QoS})$ of mismatched ML decoding (2.33). We observe that the mean outage rate $\bar{C}(\gamma_{QoS})$ is still quite large, in spite of the small training sequence. However, achieving 2 bits ($\gamma_{QoS} = 0.01$) with imperfect channel information requires 5.5 dB more than in the case with perfect CSI. In comparison, the mean outage rate $\bar{C}_{\text{ML}}(\gamma_{QoS})$ with mismatched ML decoding is significantly smaller. Indeed, in order to achieve the target rate of 2 bits, a communication system using this mismatched decoder would requires 2.5 additional dB. This means that the accuracy of the channel estimate in this case is too small to allow for ML decoding.

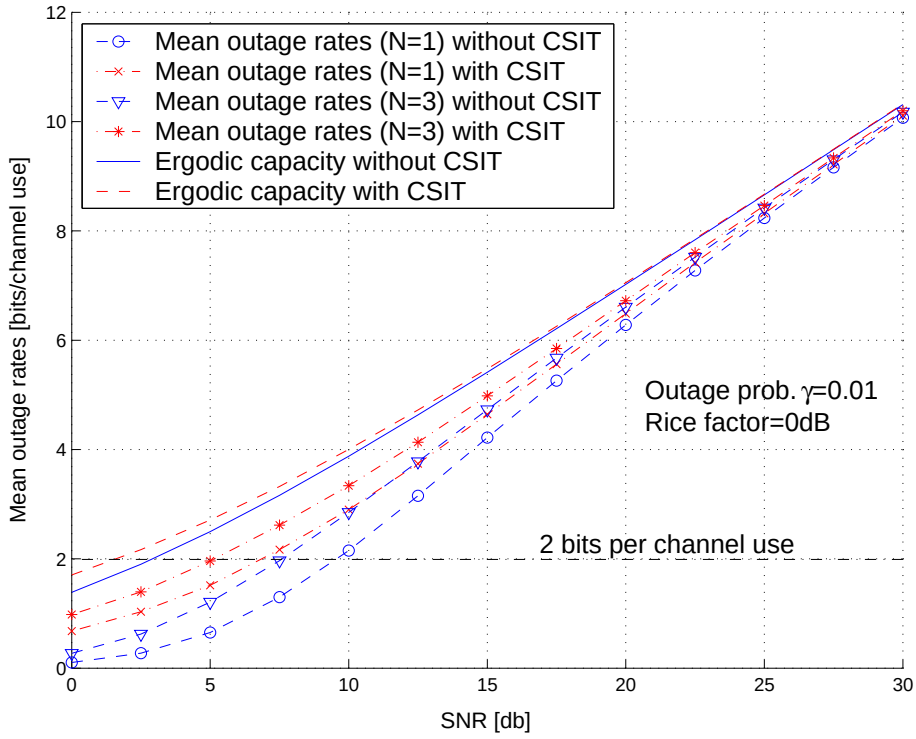



Figure 2.2: Average of estimation-induced outage capacity for different amounts of training, without feedback (no CSIT) and with perfect feedback (CSIT=CSIR) vs. SNR.

(ii) Fig. 2.2 shows the average *estimation-induced outage capacity* in bits per channel use for different amounts of training, with both perfect and no feedback/CSIT versus the signal-to-noise ratio, for an outage probability $\gamma_{_{QoS}} = 10^{-2}$. For comparison, we show ergodic capacity under perfect CSI. In this case, the power allocation function is given by the optimal solution (2.35). It is seen that the average rate increases with the amount of CSIR and CSIT. To achieve 2 bits without feedback/CSIT, it is seen that a scheme with estimated CSIR and $N = 3$ ($\nabla$ markers) requires 7.5 dB, i.e., 4.5 dB more than in the case with perfect CSIR (solid line). Whereas if the training length is further reduced to $N = 1$ ($\circ$ markers), this gap increases to 6.5 dB. In the case of perfect feedback (CSIT=CSIR), the SNR requirements for 2 bits are 2 dB (perfect CSIR, dashed line), 5 dB (estimated CSIR with $N=3$, $*$ markers), and 7 dB (estimated CSIR with $N=1$, $\times$ markers), respectively. Thus, with feedback the gap between estimated and perfect CSI is slightly smaller than without feedback (3 dB and 5 dB with $N=3$ and $N=1$, respectively). Observe that for values of SNR larger than 10 dB similar performance are achieved without feedback channel and $N = 3$ comparing to a system with a feedback link and $N = 1$. Therefore, using this information a system designer may decide to use training sequences of length $N = 3$ instead of implementing a feedback channel.

(iii) Fig. 2.3 shows the average of *estimation-induced outage capacity* for an outage probability $\gamma_{_{QoS}} = 0.01$ and rate-limited feedback/CSIT versus the signal-to-noise ratio. We suppose error-free feedback link of two bits ($R_{FB} = 2$) with training sequences of length $N = 1$. Here, we used the power allocation function given by the suboptimal solution (2.40). For comparison, we show the average of *estimation-induced outage capacity* without CSIT and with perfect feedback, and we also show the ergodic capacity under perfect CSI and feedback. Observe that at 2 bits the gap between the average outage capacity without feedback and rate-limited feedback is 0.75 dB/2 bits. Whereas the gap between the average of outage capacity with 2 bits of feedback and with non-limited rate is still 2.5 dB.

Finally, we study the impact of the imperfect channel estimation on the mean outage rate for different fading statistics (different Rice factors) and perfect feedback (CSIT=CSIR). Fig. 2.4 shows the average of *estimation-induced outage capacity* for Rice factors $K_h = \{-15, 0, 25\}$ dB and different amounts of training $N = \{1, 3\}$.

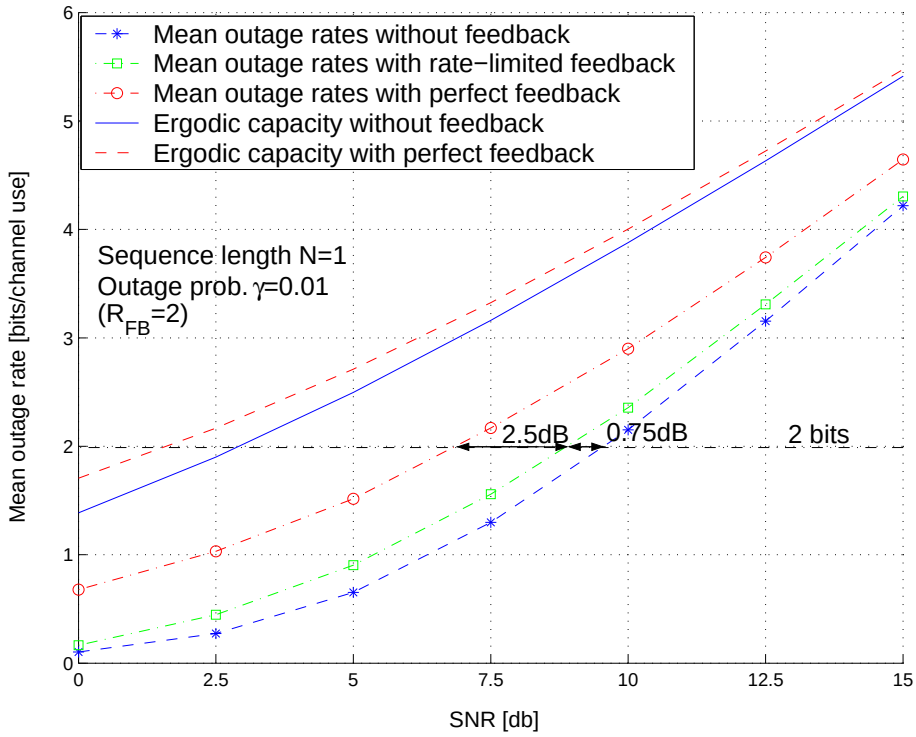



Figure 2.3: Average of estimation-induced outage capacity for different amounts of training with rate-limited feedback CSI ($R_{FB} = 2$) vs. SNR.

For comparison, the ergodic capacity under perfect CSI is also plotted. We observe that increasing the Rice factor from (A) to (B) and (C) increases the impact of the estimation errors on the mean outage rates. On the other hand, for high value of $K_h = 25\,\text{dB}$ (i.e. smaller variance values $\sigma_h^2$) the mean outage rates are not sensitive to the amount of training. While for smaller values of Rice factor $K_h = -15\,\text{dB}$ it is more important to achieve accuracy channel estimations. This impact on the mean outage rates, due to accuracy measurements of $\hat{\theta}$, depends on the trade-off between the estimation error $\sigma_\varepsilon^2$ and the variance of the fading process $\sigma_h^2$ (see expression (2.27)). Therefore, this analysis could serve as a basis to decide in practical situations whether or not robust channel estimation is necessary depending on the nature of the fading process. Of course, the worst case is observed for the range of middle values of Rice factors (i.e. $K_h = 0\,\text{dB}$), since for these values the uncertainty about the quality of channel estimates is maximal.

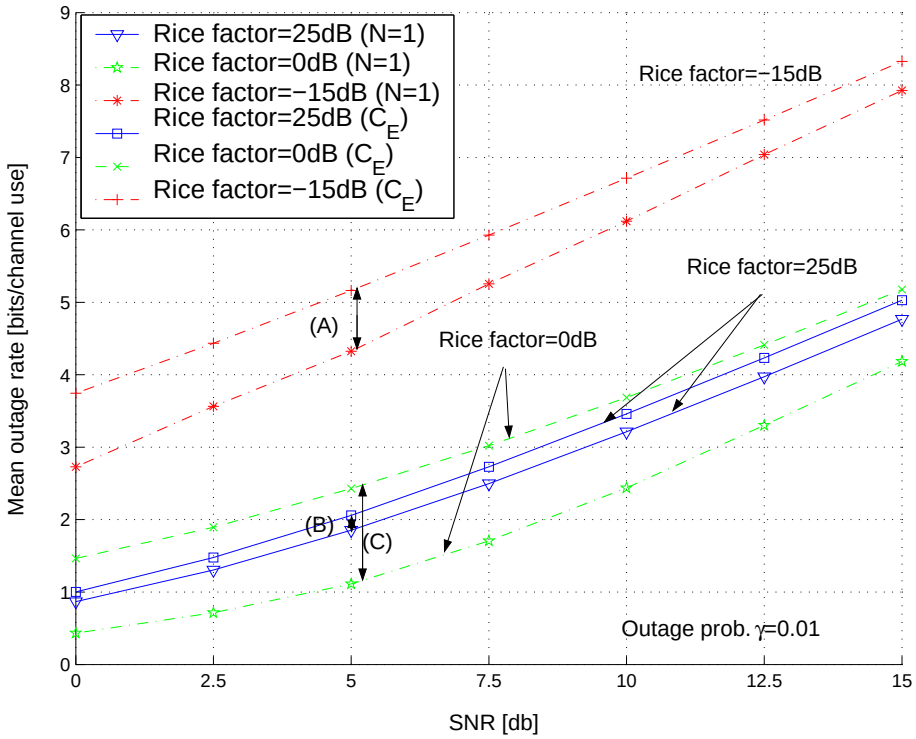



Figure 2.4: Average of estimation-induced outage capacity for different rice factors and amounts of training with perfect feedback (CSIT=CSIR) vs. SNR.

## 2.6 Summary

In this chapter we have studied the problem of reliable communications over unknown DMCs when the receiver and the transmitter only know a noisy estimate of the channel state. We proposed to characterize the information theoretic limits of such scenarios in terms of the novel notion of *estimation-induced outage capacity*. The transmitter and receiver strive to construct codes for ensuring the desired communication service, i.e. for achieving target rates with small error probability, no matter which degree of accuracy estimation arises during a transmission. We provided an explicit expression characterizing the trade-off between the maximum achievable outage rate (i.e. maximizing over all possible transmitter-receiver pairs) satisfying the QoS constraint. We proved the corresponding associated coding theorem and its strong converse. A Ricean fading model is used to illustrate our approach by computing its mean outage capacity. Our results are useful for a system designer to assess the amount of training and feedback required to achieve target rates over a given channel.

Finally, we studied the maximum achievable outage rate of a native system whose

receiver uses the mismatched maximum-likelihood decoder based on the channel estimate. Results indicate that this type of decoding can be largely suboptimal for the considered class of channels, at least if the training phase is short and the channel state information inaccurate. An improved decoder should use a metric based on maximizing *a posteriori* probability, e.g. ML metrics conditioned on the channel estimate as MAP detectors. It will be attractive to study practical coding schemes satisfying the QoS constraints and achieving rates close to the average of *estimation-induced outage capacity*.

Possibly straightforward applications of these results are practical time-varying systems with small training overhead and quality of service constraints, such as OFDM systems. Another application scenario arises in the context of cellular coverage, where the average of *estimation-induced outage capacity* would characterize performance over multiple communication sessions of different users in a large number of geographic locations (cf. [85]). In that scenario, the system designer must ensure a quality of service during the connection session, i.e., reliable communication for $(1 - \gamma_{QoS})$-percent of users, for any degree of accuracy estimation.

# Chapter 3

# On the Outage Capacity of a Practical Decoder Using Channel Estimation Accuracy

The optimal decoder achieving the outage capacity under imperfect channel estimation is investigated. First, by searching into the family of nearest neighbor decoders, which can be easily implemented on most practical coded modulation systems, we derive a decoding metric that minimizes the average of the transmission error probability over all channel estimation errors. This metric, for arbitrary memoryless channels (DMCs), achieves the capacity of a composite (more noisy) channel. Next, we specialize our general expression to obtain its corresponding decoding metric for fading MIMO channels. According to the notion of estimation-induced outage capacity (EIO capacity) introduced in our previous work (see chapter 2), we characterize maximal achievable information rates associated to the proposed decoder. In the case of uncorrelated Rayleigh fading, these achievable rates are compared to the rates achieved by the classical mismatched maximum-likelihood (ML) decoder and the ultimate limits given by the EIO capacity. The latter uses the best theoretical decoder in presence of channel estimation errors. Our results are useful for designing a communication system (transmission power, training sequence length, training power, etc.) where a prescribed quality of service (QoS) in terms of achieving target rates with small error probability, must be satisfied even in presence of very poor

channel estimates. Numerical results show that the derived metric provides significant gains for the considered scenario, in terms of achievable information rates and bit error rate (BER), in a bit interleaved coded modulation (BICM) framework, without introducing any additional decoding complexity.

## 3.1 Introduction

Consider a practical wireless communication system, where the receiver disposes only of noisy channel estimates that may in some circumstances be poor estimates, and these estimates are not available at the transmitter. This constraint constitutes a practical concern for the design of such communication systems that, in spite of their knowledge limitations, have to ensure communications with a prescribed quality of service (QoS). This QoS requires to guarantee transmissions with a given target information rate and small error probability, no matter which degree of accuracy estimation arises during the transmission. The described scenario addresses two important questions: (i) What are the theoretical limits of reliable transmission rates, using the best possible decoder in presence of imperfect channel state information at the receiver (CSIR) and (ii) how those limits can be achieved by using practical decoders in coded modulation systems ? Of course, these questions are strongly related to the notion of capacity that must take into account the above mentioned constraints.

We have addressed in chapter 2 the first question (i), for arbitrary memoryless channels (DMCs), by introducing the notion of *Estimation-induced outage capacity* (EIO capacity). This novel notion characterizes the information-theoretic limits of such scenarios, where the transmitter and receiver strive to construct codes for ensuring the desired communication service, no matter which degree of accuracy estimation arises during the transmission. The explicit expression of this capacity allows one to evaluate the trade-off between the maximal achievable outage rate (i.e. maximizing over all possible transmitter-receiver pairs) versus the outage probability $\gamma_{QoS}$ (the QoS constraint). This can be used by a system designer to optimally share the available resources (e.g. power for transmission and training, the amount of training used, etc.), so that the communication requirements be satisfied. Nevertheless, the theoret-

ical decoder used to achieve the latter capacity cannot be implemented on practical communication systems.

The second question (ii) concerning the derivation of a practical decoder, which can achieve information rates closed to the EIO capacity, is addressed in this chapter. Classically, to deal with imperfect channel state information (CSI) one sub-optimal technique, known as mismatched maximum-likehood (ML) decoding (cf. [35]), consists in replacing the exact channel by its estimate in the decoding metric. However, this scheme is not appropriate in presence of channel estimation errors (CEE), at least for small number of training symbols [62]. Indeed, intensive recent research has been conducted. In [86] and [87] the authors analyze bit error rate (BER) performances of this decoder in the case of an orthogonal frequency division multiplexing (OFDM) system. References [88] considered a training-based MIMO system and showed that for compensating the performance degradation due to CEE, the number of receive antennas should be increased, which may become a limiting item for mobile applications. On the other hand, the performance of Bit Interleaved Coded Modulation (BICM) over fading MIMO channels with perfect CSI was studied for instance, in [89], [90] and [91]. Cavers in [92], derived a tight upper bound on the symbol error rate of PSAM for 16-QAM modulations. A similar investigation was carried out in [93] showing that for iterative decoding of BICM at low SNR, the quality of channel estimates is too poor for being used in the mismatched ML decoder.

As an alternative to the aforementioned decoder, Tarokh *et al.* in [61] and Taricco and Biglieri in [62], proposed an improved ML detection metric and applied it to a space-time coded MIMO system, where they showed the superiority of this metric in terms of BER. Interestly enough, this decoding metric can be formally derived as a special case of the general framework presented in this chapter. So far, most of the research in the field were focused on evaluating the performances of mismatched decoders in terms of BER (cf. [35]), but still not providing an answer to the question (ii). In [49], the authors investigate achievable rates of a weighting nearest-neighbor decoder for multiple-antenna channel. Moreover, in section 2.4.3 we have showed that the achievable rates using the mismatched ML decoding are largely sub-optimal (at least for limited number of training symbols) compared to the ultimate limits given by the EIO capacity (see also [94]). In this chapter, according to the notion of

EIO capacity, we investigate the maximal achievable information rate with Gaussian codebooks of the improved decoder in [62]. Furthermore, we show that this decoder achieves the capacity of a composite (more noisy) channel.

This chapter is organized as follows. In section 3.2, we briefly review our notion of capacity. Then, by using the tools of information theory, we search into the family of decoders that can be easily implemented on most practical coded modulation systems to derive the general expression of the decoder. This decoder minimizes the average of the transmission error probability over all CEE and consequently, achieves the capacity of the composite channel. We accomplish this by exploiting an interesting feature of the theoretical decoder that achieves the EIO capacity. This feature is the availability of the statistic characterizing the quality of channel estimates, i.e., the *a posteriori* probability density function (pdf) of the unknown channel conditioned on its estimate. In section 3.3 we describe the fading MIMO model. In section 3.4, we specialize our expression of the decoding metric for the case of MIMO channels and use this for iterative decoding of MIMO-BICM. In section 3.5, we compute achievable information rates of a receiver using the proposed decoder and compare these to the EIO capacity and the rates of the classical mismatched approach. Section 3.6 illustrates via simulations, conducted over uncorrelated Rayleigh fading, the performance of the improved decoder in terms of achievable outage rates and BER, comparing to those provided by the mismatched ML decoding.

Notational conventions are as follows. Upper and lower case bold symbols are used to denote matrices and vectors; $\mathbb{I}_M$ represents an $(M \times M)$ identity matrix; $\mathbb{E}_{\mathbf{X}}\{\cdot\}$ refers to expectation with respect to the random vector $\mathbf{X}$; $|\cdot|$ and $\|\cdot\|_F$ denote matrix determinant and Frobenius norm, respectively; $(\cdot)^T$ and $(\cdot)^\dagger$ denote vector transpose and Hermitian transpose, respectively.

## 3.2 Decoding under Imperfect Channel Estimation

Throughout this section we focus on deriving a practical decoder for general memoryless channels that achieves information rates close to the EIO capacity (the ultimate bound).

### 3.2.1 Communication Model Under Channel Uncertainty

A specific instance of the memoryless channel is characterized by a transition probability $W(y|x,\theta) \in \mathcal{W}_\Theta$ with an unknown channel state $\theta$, over the general input and output alphabets $\mathscr{X}, \mathscr{Y}$. Here, $\mathcal{W}_\Theta = \{W(\cdot|x,\theta) \colon x \in \mathscr{X}, \theta \in \Theta\}$ is a family of conditional pdf parameterized by the vector of parameters $\theta \in \Theta \subseteq \mathbb{C}^d$, where $d$ denotes the number of parameters. Throughout the chapter we assume that the channel state, which neither the transmitter nor the receiver know exactly, remains constant within blocks of symbols, related to the product of the coherence time and the coherence bandwidth of a wireless channel, and these states for different blocks are i.i.d. $\theta \sim \psi(\theta)$. The transmitter does not know the channel state and the receiver only knows an estimate $\hat{\theta}$ and a *characterization of the estimator performance* in terms of the conditional pdf $\psi(\theta|\hat{\theta})$ (this can be obtained using $\mathcal{W}_\Theta$, the estimation function and $\psi(\theta)$). A decoder using $\hat{\theta}$, instead of $\theta$, obviously might not support an information rate $R$ (even small rates might not be supported if $\hat{\theta}$ and $\theta$ are strongly different). Consequently, outage events induced by CEE will occur with a certain probability $\gamma_{_{QoS}}$. The scenario underlying these assumptions is motivated by current wireless systems, where the coherence time for mobile receivers may be too short to permit reliable estimation of the fading coefficients and in spite of this fact, the desired communication service must be guaranteed. This leads to the following notion of capacity.

### 3.2.2 A Brief Review of Estimation-induced Outage Capacity

A message $m \in \mathcal{M} = \{1, \ldots, \lfloor \exp(nR) \rfloor\}$ is transmitted using a pair $(\varphi, \phi)$ of mappings, where $\varphi : \mathcal{M} \mapsto \mathscr{X}^n$ is the encoder, and $\phi : \mathscr{Y}^n \times \Theta \mapsto \mathcal{M}$ is the decoder (that utilizes $\hat{\theta}$). The random rate, which depends on the unknown channel realization $\theta$ through its probability of error, is given by $n^{-1} \log M_{\theta,\hat{\theta}}$. The maximum error probability (over all messages)

$$e_{\max}^{(n)}(\varphi, \phi, \hat{\theta}; \theta) = \max_{m \in \mathcal{M}} \int_{\{\mathbf{y} \in \mathscr{Y}^n : \phi(\mathbf{y}, \hat{\theta}) \neq m\}} dW^n\big(\mathbf{y}|\varphi(m), \theta\big), \tag{3.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$. For a given channel estimate $\hat{\theta}$, and $0 < \epsilon, \gamma_{_{QoS}} < 1$, an outage rate $R \geq 0$ is $(\epsilon, \gamma_{_{QoS}})$-achievable if for every $\delta > 0$ and every sufficiently large $n$ there

exists a sequence of length-$n$ block codes such that the rate satisfies the quality of service

$$\Pr\left(\Lambda_\epsilon(R,\hat{\theta})\big|\hat{\theta}\right) = \int_{\Lambda_\epsilon(R,\hat{\theta})} d\psi(\theta|\hat{\theta}) \geq 1-\gamma_{_{QoS}}, \tag{3.2}$$

where $\Lambda_\epsilon(R,\hat{\theta}) = \left\{\theta \in \Delta_\epsilon^{(n)} : n^{-1}\log M_{\theta,\hat{\theta}} \geq R-\delta\right\}$ stands for the set of all channel states allowing for the desired transmission rate $R$, and $\Delta_\epsilon^{(n)} = \left\{\theta \in \Theta : e_{\max}^{(n)}(\varphi,\phi,\hat{\theta};\theta) \leq \epsilon\right\}$ is the set of all channel states allowing for reliable decoding (arbitrary small error probability). This definition requires that maximum error probabilities larger than $\epsilon$ occur with probability less than $\gamma_{_{QoS}}$. The practical advantage of such definition is that for $(1-\gamma_{_{QoS}})\%$ of channel estimates, the transmitter and receiver strive to construct codes for ensuring the desired communication service. The EIO capacity is then defined as the largest $(\epsilon,\gamma_{_{QoS}})$-achievable rate, for an outage probability $\gamma_{_{QoS}}$ and a given channel estimate $\hat{\theta}$, as

$$C(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta}) = \lim_{\epsilon\downarrow 0}\sup_{\varphi,\phi}\left\{R\geq 0 :\ \Pr\left(\Lambda_\epsilon(R,\hat{\theta})|\hat{\theta}\right) \geq 1-\gamma_{_{QoS}}\right\}, \tag{3.3}$$

where the maximization is taken over all encoder and decoder pairs. In section 2.3, we proved the following coding Theorem that provides an explicit way to evaluate the maximal outage rate (3.3) versus outage probability $\gamma_{_{QoS}}$ for an estimate $\hat{\theta}$, characterized by $\psi(\theta|\hat{\theta})$.

**Theorem 3.2.1** *Given an outage probability $0 \leq \gamma_{_{QoS}} < 1$, the EIO capacity is given by*

$$C(\gamma_{_{QoS}},\psi_{\theta|\hat{\theta}},\hat{\theta}) = \max_{P\in\mathscr{P}_\Gamma(\mathscr{X})}\ \sup_{\Lambda\subset\Theta:\ \Pr(\Lambda|\hat{\theta})\geq 1-\gamma_{_{QoS}}}\ \inf_{\theta\in\Lambda} I\big(P,W(\cdot|\cdot,\theta)\big), \tag{3.4}$$

*where $I(\cdot)$ denotes the mutual information of the channel $W(y|x,\theta)$ and $\mathscr{P}_\Gamma(\mathscr{X})$ is the set of input distributions that does not depend on $\hat{\theta}$, satisfying the input constraint $\int g(x)dP(x) \leq \Gamma$ for a nonnegative cost function $g : \mathscr{X} \to [0,\infty)$.*

The existence of a decoder $\phi$ in (3.3) achieving the capacity (3.4) is proved using a random-coding argument, based on the well-known method of typical sequences [17]. Nevertheless, this decoder cannot be implemented on practical communication systems.

### 3.2.3 Derivation of a Practical Decoder Using Channel Estimation Accuracy

We now consider the problem of deriving a practical decoder that achieves the capacity (3.4). Assume that we restrict the searching of decoding functions $\phi$, maximizing (3.3), to the class of additive decoding metrics, which can be implemented on realistic systems. This means that for a given channel output $\mathbf{y} = (y_1, \ldots, y_n)$, we set the decoding function

$$\phi_{\mathcal{D}}(\mathbf{y}, \hat{\theta}) = \arg \min_{m \in \mathcal{M}} \mathcal{D}^n\big(\varphi(m), \mathbf{y}|\hat{\theta}\big), \tag{3.5}$$

where $\mathcal{D}^n\big(\mathbf{x}, \mathbf{y}|\hat{\theta}\big) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}\big(x_i, y_i|\hat{\theta}\big)$ and $\mathcal{D} : \mathscr{X} \times \mathscr{Y} \times \Theta \mapsto \mathbb{R}_{\geq 0}$ is an arbitrary per-letter additive metric. Consequently, the maximization in (3.3) is actually equivalent to maximizing over all decoding metrics $\mathcal{D}$. However, we note that this restriction does not necessarily lead to an optimal decoder achieving the capacity.

*Problem statement:* In order to find the optimal decoding metric $\mathcal{D}$ maximizing the outage rates in (3.3), for a given outage probability $\gamma_{_{QoS}}$ and channel estimate $\hat{\theta}$, it is necessary to look at the intrinsic properties of the capacity definition. Observe that the size of the set of all channel states allowing for reliable decoding $\Delta_\epsilon^{(n)}$ is determined by the decoding function $\phi$ chosen and the maximal achievable rate $R$, constrained to the outage probability (3.2), is then limited by this size. Thus, for a given decoder $\phi$, there exists an optimal set $\Lambda_\epsilon^* \subseteq \Delta_\epsilon^{(n)}$ of channel states with conditional probability larger than $1 - \gamma_{_{QoS}}$, providing the largest achievable rate, which follows as the minimal instantaneous rate for the worst $\theta \in \Lambda_\epsilon^*$. The optimal set $\Lambda_\epsilon^*$ is equal to the set $\Lambda^*$ maximizing the expression (3.4). Hence, an optimal decoding metric must guarantee minimum error probability (3.1) for every $\theta \in \Lambda^*$.

The computation of such a metric becomes very difficult (not necessary feasible by using the class of decoders in (3.5)), since the maximization in (3.3) by using $\phi_{\mathcal{D}}$ is not an explicit function of $\mathcal{D}$. However, it is interesting to note [40], that if the set $\Lambda^*$ defines a compact and convex set of channels $\mathcal{W}_{\Lambda^*}$, then the optimal decoding metric can be chosen as the ML decoder $\mathcal{D}^*(x, y|\hat{\theta}) = -\log W(y|x, \theta^*)$, where $\theta^*$ is the channel state minimizing the mutual information in (3.4). The receiver can thus be a ML receiver with respect to the worst channel in the family. However, in most practical cases, the channel states are represented by vectors of complex coefficients

that do not lead to convex sets of channels.

*Optimal decoder for composite channels:* Instead of trying to find an optimal decoding metric minimizing the error probability (3.1) for every $\theta \in \Lambda^*$, we propose to look at the decoding metric minimizing the average of the transmission error probability over all CEE. This means,

$$\mathcal{D}_{\mathcal{M}} = \arg\min_{\mathcal{D}} \int_{\Theta} e_{\max}^{(n)}(\varphi, \phi_{\mathcal{D}}, \hat{\theta}; \theta) d\psi(\theta|\hat{\theta}), \tag{3.6}$$

where $e_{\max}^{(n)}$ is obtained by replacing (3.5) in (3.1). Actually, for $n$ sufficiently large, this optimization problem can be resolved by setting

$$\mathcal{D}_{\mathcal{M}}(x, y|\hat{\theta}) = -\log \widetilde{W}(y|x, \hat{\theta}) \quad \text{with} \quad \widetilde{W}(y|x, \hat{\theta}) = \int_{\Theta} W(y|x, \theta) d\psi(\theta|\hat{\theta}), \tag{3.7}$$

$\widetilde{W}$ is the channel resulting from the average of the unknown channel over all CEE, given the estimate $\hat{\theta}$. Here, we do not go into the details of how the optimal metric (3.7) minimizes (3.6), since it can be obtained by following an analogy with the proof based on the method of types in [40]. Basically, the average of the transmission error probability in (3.6) leads to the composite channel $\widetilde{W}(y|x, \hat{\theta})$. We then take the logarithm of this composite channel to obtain its ML decoder (3.7), which minimizes (with $n$ sufficiently large) the error probability (3.6).

*Remark:* We emphasize that this decoder cannot guarantee small error probabilities for every channel state $\theta \in \Lambda^*$, and consequently it only achieves a lower bound of the EIO capacity (3.4). Nevertheless, this decoder archives the capacity of the composite channel. Therefore, the remaining question to answer is how much lower are the achievable outage rates using the metric (3.7), comparing to the theoretical decoder achieving the EIO capacity. In section 3.5, we evaluate the metric (3.7) and its achievable information rates for fading MIMO channels.

## 3.3 System Model

### 3.3.1 Fading MIMO Channel

We consider a single-user MIMO system with $M_T$ transmit and $M_R$ receiver antennas transmitting over a frequency non-selective channel and refer to it as a MIMO channel. Fig. 3.1 depicts the BICM coding scheme used at the transmitter. The

binary data sequence $\mathbf{b}$ is encoded by a non-recursive and non-systematic convolutional (NRNSC) code, before being interleaved by a quasi-random interleaver. The output bits $\mathbf{d}$ are gathered in subsequences of $B$ bits and mapped to complex M-QAM ($M = 2^B$) vector symbols $\mathbf{x}$ with average power $\frac{tr(\mathbf{x}\mathbf{x}^\dagger)}{M_T} = \bar{P}$. We also send some pilot symbols at the beginning of each data frame for channel estimation. The symbols of a frame are then multiplexed for being transmitted through $M_T$ antennas. Assuming a frame of $L$ transmitted symbols associated to each channel matrix $\mathbf{H}_k$, the received signal vector $\mathbf{y}_k$ of dimension $(M_R \times 1)$ is given by

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{z}_k, \quad k = 1, \ldots, L, \tag{3.8}$$

where $\mathbf{x}_k$ is the $(M_T \times 1)$ vector of transmitted symbols, referred to as a compound symbol. Here, the entries of the random matrix $\mathbf{H}_k$ are independent identically distributed (i.i.d.) zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variables. Thus, the channel state $\theta = \mathbf{H}_k$ is distributed as $\mathbf{H}_k \sim \psi_H(\mathbf{H}) = C\mathcal{N}\big(\mathbf{0}, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma_H}\big)$

$$C\mathcal{N}\big(\mathbf{0}, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma_H}\big) = \frac{1}{\pi^{M_R M_T} |\Sigma_H|^{M_T}} \exp\Big[ - tr\big(\mathbf{H}\mathbf{\Sigma_H}^{-1}\mathbf{H}^\dagger\big)\Big], \tag{3.9}$$

where $\mathbf{\Sigma_H}$ is the Hermitian covariance matrix of the columns of $\mathbf{H}$ (assumed to be the same for all columns), i.e., $\mathbf{\Sigma_H} = \sigma_h^2 \mathbb{I}_{M_R}$. The noise vector $\mathbf{z}_k \in \mathbb{C}^{M_R \times 1}$ consists of ZMCSCG random vector with covariance matrix $\mathbf{\Sigma_0} = \sigma_Z^2 \mathbb{I}_{M_R}$. Both $\mathbf{H}_k$ and $\mathbf{z}_k$ are assumed ergodic and stationary random processes, and the channel matrix $\mathbf{H}_k$ is independent of $\mathbf{x}_k$ and $\mathbf{z}_k$.

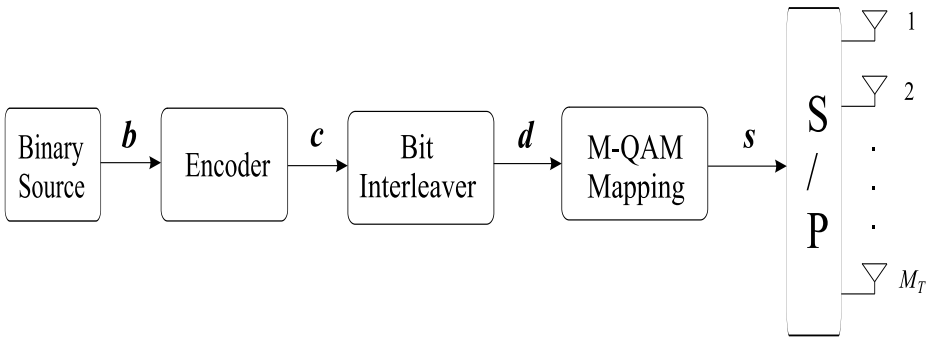



Figure 3.1: Block diagram of MIMO-BICM transmission scheme.

### 3.3.2 Pilot Based Channel Estimation

Assuming that the channel matrix is time-invariant over an entire frame, channel estimation is usually performed on the basis of known training (pilot) symbols transmitted at the beginning of each frame. The transmitter, before sending the data $\mathbf{x}_k$, sends a training sequence of $N$ vectors $\mathbf{X}_T = (\mathbf{x}_{T,1}, \ldots, \mathbf{x}_{T,N})$. According to the observation of the channel model (3.8), this sequence is affected by the channel matrix $\mathbf{H}_k$, allowing the receiver to observe separately $\mathbf{Y}_{T,k} = \mathbf{H}_k \mathbf{X}_{T,k} + \mathbf{Z}_{T,k}$, where $\mathbf{Z}_{T,k}$ is the noise matrix affecting the transmission of training symbols. We assume that the coherence time is much longer than the training time and the average energy of the training symbols is $P_T = \frac{1}{NM_T} tr\big(\mathbf{X}_T \mathbf{X}_T^\dagger\big)$.

We focus on the estimation of $\mathbf{H}_k$, from the observed signals $\mathbf{Y}_{T,k}$ and $\mathbf{X}_{T,k}$. In the ML sense this estimate is obtained by minimizing $\|\mathbf{Y}_{T,k} - \mathbf{H}_k \mathbf{X}_T\|^2$ with respect to $\mathbf{H}_k$. This yields $\widehat{\mathbf{H}}_{\mathrm{ML},k} = \mathbf{Y}_{T,k}\mathbf{X}_T^\dagger\big(\mathbf{X}_T\mathbf{X}_T^\dagger\big)^{-1} = \mathbf{H}_k + \mathcal{E}_k$, where $\mathcal{E}_k = \mathbf{Z}_{T,k}\mathbf{X}_T^\dagger\big(\mathbf{X}_T\mathbf{X}_T^\dagger\big)^{-1}$ denotes the estimation error matrix [62]. Since to estimate the $M_R \times M_T$ channel matrix, we need at least $M_R M_T$ independent measurements, and each symbol time yields $M_R$ samples at the receiver, we must have $N \geq M_T$. Moreover, matrix $\mathbf{X}_T$ must have full rank $M_T$ and consequently the matrix $\mathbf{X}_T\mathbf{X}_T^\dagger$ must be nonsingular. We suppose orthogonal training sequences, i.e., we refer to a matrix $\mathbf{X}_T$ with orthogonal rows, such that $\mathbf{X}_T\mathbf{X}_T^\dagger = NP_T\mathbb{I}_{M_T}$. Next, denoting $\underline{\mathcal{E}}_j$ the $j$th column of the error matrix $\mathcal{E}$, we can write $\mathbf{\Sigma}_{\mathcal{E}} = \mathbb{E}_{\mathcal{E}}\big\{\underline{\mathcal{E}}_j\underline{\mathcal{E}}_j^\dagger\big\} = \mathrm{SNR}_T^{-1}\mathbb{I}_{M_R}$ with $\mathrm{SNR}_T = \dfrac{NP_T}{\sigma_Z^2}$, yielding a white error matrix, i.e. the entries of $\mathcal{E}$ are i.i.d. ZMCSCG random variables with variance $\sigma_{\mathcal{E}}^2 = \mathrm{SNR}_T^{-1}$. Thus, for each frame, the conditional pdf of $\hat{\theta} = \widehat{\mathbf{H}}_{\mathrm{ML}}$ given $\theta = \mathbf{H}$ is the complex normal matrix pdf

$$\psi_{\widehat{H}_{\mathrm{ML}}|H}(\widehat{\mathbf{H}}_{\mathrm{ML}}|\mathbf{H}) = C\mathcal{N}\big(\mathbf{H}, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma}_{\mathcal{E}}\big). \tag{3.10}$$

## 3.4 Metric Computation and Iterative Decoding of BICM

In this section, we specialize the expression (3.7) to derive the decoding metric for MIMO channels (3.8) and then we consider MIMO-BICM decoding with the derived metric.

### 3.4.1 Mismatched ML Decoder

The classical mismatched ML decoder consists of the likelihood function of the channel pdf using the channel estimate $\widehat{\mathbf{H}}_{\mathrm{ML}}$. This leads to the following Euclidean distance

$$\mathcal{D}_{\mathrm{ML}}\big(\mathbf{x},\mathbf{y}|\widehat{\mathbf{H}}_{\mathrm{ML}}\big) = -\log W(\mathbf{y}|\mathbf{x},\widehat{\mathbf{H}}_{\mathrm{ML}}) = \|\mathbf{y}-\widehat{\mathbf{H}}_{\mathrm{ML}}\mathbf{x}\|^2 + \text{const.} \tag{3.11}$$

### 3.4.2 Metric Computation

We now specialize the expression (3.7) in the case of a MIMO channel (3.8). To this end, we need to derive the pdf $\psi_{H|\widehat{H}_{\mathrm{ML}}}(\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}})$, which can be obtained by using the pdf (3.10) and (3.9) (see Appendix B.1). Thus,

$$\psi_{H|\widehat{H}_{\mathrm{ML}}}(\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}}) = C\mathcal{N}\big(\mathbf{\Sigma}_{\Delta}\widehat{\mathbf{H}}_{\mathrm{ML}}, \mathbb{I}_{M_T}\otimes\mathbf{\Sigma}_{\Delta}\mathbf{\Sigma}_{\mathcal{E}}\big), \tag{3.12}$$

where $\mathbf{\Sigma}_{\Delta} = \mathbf{\Sigma}_{\mathbf{H}}(\mathbf{\Sigma}_{\mathcal{E}}+\mathbf{\Sigma}_{\mathbf{H}})^{-1} = \mathbb{I}_{M_R}\delta$ and $\delta = \dfrac{\mathrm{SNR}_T\sigma_h^2}{\mathrm{SNR}_T\sigma_h^2+1}$. The availability of the distribution (3.12) characterizing the CEE is the key feature of pilot assisted channel estimation. Then, by averaging the channel $W(\mathbf{y}|\mathbf{x},\mathbf{H})$ over all CEE, i.e. using the pdf (3.12), and after some algebra we obtain the composite channel (cf. Appendix B.1)

$$\widetilde{W}(\mathbf{y}|\mathbf{x},\widehat{\mathbf{H}}_{\mathrm{ML}}) = C\mathcal{N}\big(\delta\widehat{\mathbf{H}}_{\mathrm{ML}}\mathbf{x}, \mathbf{\Sigma_0} + \delta\mathbf{\Sigma}_{\mathcal{E}}\|\mathbf{x}\|^2\big). \tag{3.13}$$

Finally, from (3.13) the optimal decoding metric for the MIMO channel (3.8) is reduced to

$$\mathcal{D}_{\mathcal{M}}^{\mathrm{MIMO}}\big(\mathbf{x},\mathbf{y}|\widehat{\mathbf{H}}_{\mathrm{ML}}\big) = M_R\log(\sigma_Z^2+\delta\sigma_{\mathcal{E}}^2\|\mathbf{x}\|^2) + \frac{\|\mathbf{y}-\delta\widehat{\mathbf{H}}_{\mathrm{ML}}\mathbf{x}\|^2}{\sigma_Z^2+\delta\sigma_{\mathcal{E}}^2\|\mathbf{x}\|^2}. \tag{3.14}$$

This metric coincides with that proposed for space-time decoding, from independent results in [62]. We note that under near perfect CSI, obtained when $N\to\infty$,

$$\lim_{N\to\infty}\frac{\mathcal{D}_{\mathcal{M}}^{\mathrm{MIMO}}\big(\mathbf{x},\mathbf{y}|\widehat{\mathbf{H}}_{\mathrm{ML}}\big)}{\mathcal{D}_{\mathrm{ML}}\big(\mathbf{x},\mathbf{y}|\widehat{\mathbf{H}}_{\mathrm{ML}}\big)} = 1, \quad \textit{almost surely.} \tag{3.15}$$

Consequently, we have the expected result that the metric (3.14) tends to the classical mismatched ML decoding metric (3.11), when the estimation error $\sigma_{\mathcal{E}}^2\to 0$.

### 3.4.3 Receiver Structure

The problem of decoding MIMO-BICM has been addressed in [95] under the assumption of perfect CSIR. Here we consider the same problem with CEE, for which we use the metric (3.14) in the iterative decoding process of BICM. Basically, the receiver consists of the combination of two sub-blocks operating successively. The block diagram of the transmitter and the receiver are shown in Fig. 3.1 and Fig. 3.2, respectively. The first sub-block, referred to as soft symbol to bit MIMO demapper, produces bit metrics (probabilities) from the input symbols and the second one is a soft-input soft-output (SISO) trellis decoder. Each sub-block can take advantage of the a posteriori (APP) provided by the other sub-block as an additive information. Here, SISO decoding is performed using the well known forward-backward algorithm [96]. We recall the formulation of the soft MIMO detector.

Suppose first the case where the channel matrix $\mathbf{H}$ is perfectly known at the receiver. The MIMO demapper provides at its output the extrinsic probabilities on coded and interleaved bits $\mathbf{d}$. Let $d_{k,i}$, $i = 1, ..., BM_T$, be the interleaved bits corresponding to the $k$-th compound symbol $\mathbf{x}_k \in Q$ where the cardinality of $Q$ is equal to $2^{BM_T}$. The extrinsic probability $P_{\text{dem}}(d_{k,j})$ of the bit $d_{k,j}$ (bit metrics) at the MIMO demapper output is calculated as

$$P_{\text{dem}}(d_{k,j} = 1) = K \sum_{\substack{\mathbf{x}_k \in Q \\ d_j = 1}} \prod_{\substack{i=1 \\ i \neq j}}^{BM_T} P_{\text{dec}}(d_i) \exp\big[ - \mathcal{D}(\mathbf{x}_k, \mathbf{y}_k | \mathbf{H}_k) \big], \tag{3.16}$$

where $\mathcal{D}(\mathbf{x}_k, \mathbf{y}_k | \mathbf{H}_k) = -\log W(\mathbf{y}_k | \mathbf{x}_k, \mathbf{H}_k)$ and $K$ is the normalization factor satisfying $P_{\text{dem}}(d_{k,j} = 1) + P_{\text{dem}}(d_{k,j} = 0) = 1$ and $P_{\text{dec}}(d_{k,j})$ is the *prior* information on bit $d_{k,j}$, coming from the SISO decoder. The summation in (3.16) is taken over the product of the channel likelihood given a compound symbol $\mathbf{x}_k$, and the *a priori* probability on this symbol (the term $\prod P_{\text{dec}}$) fed back from the SISO decoder at the previous iteration. Concerning this latter term, the *a priori* probability of the bit $d_{k,j}$ itself has been excluded, so as to let the exchange of extrinsic information between the channel decoder and the MIMO demapper. Also, note that this term assumes independent coded bits $d_{k,i}$, which is true for random interleaving of large size. At the first iteration, where there is no *a priori* information available, we set $P_{\text{dec}}(d_{k,i}) = 1/2$.

Notice that by replacing the unknown channel involved in (3.16) by its channel

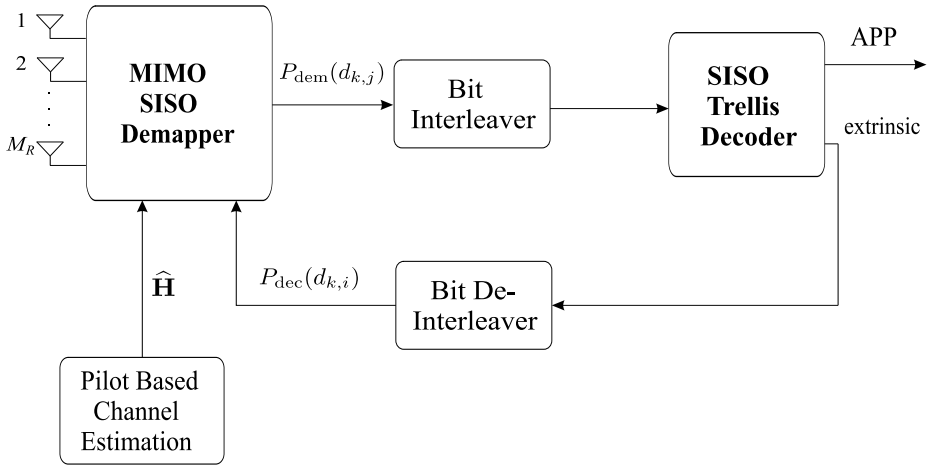



Figure 3.2: Block digram of MIMO-BICM receiver.

estimate $\widehat{\mathbf{H}}_k$, we obtain the mismatched ML decoder of MIMO-BICM. Instead of this (mismatch approach (3.11)), we propose to introduce the demaping rule given by $\mathcal{D}_{\mathcal{M}}^{\mathrm{MIMO}}(\mathbf{x}_k, \mathbf{y}_k|\widehat{\mathbf{H}}_k)$ (3.14) in (3.16), which is adapted to the CEE. This yields to the same equation that (3.16) with its appropriate constant $K$.

## 3.5 Achievable Information Rates over MIMO Channels

In this section we derive the achievable information rates in the sense of outage rates, associated to a receiver using the decoding rule (3.5) based on the metric (3.14) and on the mismatched ML metric (3.11).

### 3.5.1 Achievable Information Rates Associated to the Improved Decoder

Assume a given pair of matrices $(\mathbf{H}, \widehat{\mathbf{H}})$, characterizing a specific instance of the channel realization and its estimate. We first derive the instantaneous achievable rates $C_{\mathcal{M}}^{\mathrm{MIMO}}(\mathbf{H}, \widehat{\mathbf{H}})$ for MIMO channels $W(\mathbf{y}|\mathbf{x}, \mathbf{H}) = C\mathcal{N}\big(\mathbf{Hx}, \mathbf{\Sigma_0}\big)$, associated to a receiver using the derived metric (3.14). This is done by using the following Theorem [44], which provides the general expression for the maximal achievable rate with a given decoding metric.

**Theorem 3.5.1** *For any pair of matrices* $(\mathbf{H}, \widehat{\mathbf{H}})$*, the maximal achievable rate associated to a receiver using a metric* $\mathcal{D}(\mathbf{x}, \mathbf{y}|\widehat{\mathbf{H}})$ *is given by*

$$C_{\mathcal{D}}(\mathbf{H}, \widehat{\mathbf{H}}) = \sup_{P_X \in \mathscr{P}_\Gamma(\mathscr{X})} \inf_{V_{Y|X} \in \mathcal{V}(\mathbf{H}, \widehat{\mathbf{H}})} I(P_X, V_{Y|X}), \tag{3.17}$$

*where the mutual information functional*

$$I(P_X, V_{Y|X}) = \iint \log_2 \frac{V_{Y|X}(\mathbf{y}|\mathbf{x}, \Upsilon)}{\int V_{Y|X}(\mathbf{y}|\mathbf{x}', \Upsilon) dP_X(\mathbf{x}')} dP_X(\mathbf{x}) dV_{Y|X}(\mathbf{y}|\mathbf{x}, \Upsilon), \tag{3.18}$$

*and* $\mathcal{V}(\mathbf{H}, \widehat{\mathbf{H}})$ *denotes the set of test channels, i.e., all possibles uncorrelated MIMO channels* $V_{Y|X}(\mathbf{y}|\mathbf{x}, \Upsilon) = C\mathcal{N}(\Upsilon\mathbf{x}, \mathbf{\Sigma})$*, verifying that*[1]

$$(c_1) : tr\big(\mathbb{E}_P\left\{\mathbb{E}_V\{\mathbf{y}\mathbf{y}^\dagger\}\right\}\big) = tr\big(\mathbb{E}_P\left\{\mathbb{E}_W\{\mathbf{y}\mathbf{y}^\dagger\}\right\}\big),$$
$$(c_2) : \mathbb{E}_P\left\{\mathbb{E}_V\big\{\mathcal{D}(\mathbf{x}, \mathbf{y}|\widehat{\mathbf{H}})\big\}\right\} \leq \mathbb{E}_P\left\{\mathbb{E}_W\big\{\mathcal{D}(\mathbf{x}, \mathbf{y}|\widehat{\mathbf{H}})\big\}\right\}.$$

In order to solve the constrained minimization problem in Theorem (3.5.1) for our metric $\mathcal{D} = \mathcal{D}_{\mathcal{M}}$ (expression (3.14)), we must find the channel $\Upsilon \in \mathbb{C}^{M_R \times M_T}$ and the covariance matrix $\mathbf{\Sigma} = \mathbb{I}_{M_R}\sigma^2$ defining the test channel $V_{Y|X}(\mathbf{y}|\mathbf{x}, \Upsilon)$ that minimizes the relative entropy (3.18). On the other hand, through this chapter we assume that the transmitter does not dispose of the channel estimates, and consequently no power control is possible. Thus, we choose the sub-optimal input distribution $P_X = C\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_P})$ with $\mathbf{\Sigma_P} = \mathbb{I}_{M_T}\bar{P}$. We first compute the constraint set $\mathcal{V}(\mathbf{H}, \widehat{\mathbf{H}})$, given by $(c_1)$ and $(c_2)$, and then we factorize the matrix $\mathbf{H}$ to solve the minimization problem. Before this, to compute the constraint $(c_2)$, we need the following result (Appendix B.2).

**Lemma 3.5.1** *Let* $\mathbf{A} \in \mathbb{C}^{M_R \times M_T}$ *be an arbitrary matrix and* $\mathbf{X}$ *be a random vector with pdf* $C\mathcal{N}(0, \mathbf{\Sigma_P})$*. For every real positive constants* $K_1, K_2 > 0$*, the following equality holds*

$$\mathbb{E}_{\mathbf{X}}\left[\frac{\|\mathbf{A}\mathbf{X}\|^2 + K_1}{\|\mathbf{X}\|^2 + K_2}\right] = \frac{\|\mathbf{A}\|_F^2}{n+1} + \left(\frac{K_1}{K_2} - \frac{\|\mathbf{A}\|_F^2}{n+1}\right)\left(\frac{K_2}{\bar{P}}\right)^{n+1} \exp\left(\frac{K_2}{\bar{P}}\right)\Gamma\left(-n, K_2/\bar{P}\right), \tag{3.19}$$

*where* $n = M_T - 1$ *with* $n \in \mathbb{N}_+$ *and* $\Gamma(-n, t) = \frac{(-1)^n}{n!}\Big[\Gamma(0, t) - \exp(-t)\sum_{i=0}^{n-1}(-1)^i \frac{i!}{t^{i+1}}\Big]$*,* $\mathbf{\Sigma_P} = \mathbb{I}_{M_T}\bar{P}$ *and* $\Gamma(0, t) = \int_t^{+\infty} u^{-1}\exp(-u)du$ *denotes the exponential integral function.*

[1] Our constraint $(c_1)$ is different of that provided in [44], since here the channel noise is i.i.d. and consequently we can only satisfy the equality of the matrix traces and not of the covariance matrices.

From Lemma 3.5.1 and some algebra, it is not difficult to show that the constraints require that

$$(\mathrm{c}_1): \quad tr\big(\Upsilon\mathbf{\Sigma_P}\Upsilon^\dagger + \mathbf{\Sigma}\big) = tr\big(\mathbf{H}\mathbf{\Sigma_P}\mathbf{H}^\dagger + \mathbf{\Sigma_0}\big), \tag{3.20}$$

$$(\mathrm{c}_2): \quad \|\Upsilon + a_{\mathcal{M}}\widehat{\mathbf{H}}\|_F^2 \leq \|\mathbf{H} + a_{\mathcal{M}}\widehat{\mathbf{H}}\|_F^2 + \mathrm{C}, \tag{3.21}$$

$$\begin{aligned} a_{\mathcal{M}} &= \delta(\delta\sigma_{\mathcal{E}}^2\bar{P} - \lambda_n\sigma_Z^2)\big[M_T\delta\sigma_{\mathcal{E}}^2\lambda_n\bar{P} + \lambda_n\sigma_Z^2 - \delta\sigma_{\mathcal{E}}^2\bar{P}\big]^{-1}, \\ \mathrm{C} &= M_T\lambda_n\big[\|\mathbf{H}\|_F^2 - \|\Upsilon\|_F^2 + \bar{P}^{-1}\big(tr(\mathbf{\Sigma_0}) - tr(\mathbf{\Sigma})\big)\big]\big[1 - \frac{\sigma_Z^2}{\delta\bar{P}\sigma_{\mathcal{E}}^2}\lambda_n - M_T\lambda_n\big]^{-1}, \\ \lambda_n &= \left(\frac{\sigma_Z^2}{\delta\bar{P}\sigma_{\mathcal{E}}^2}\right)^n \exp\left(\frac{\sigma_Z^2}{\delta\bar{P}\sigma_{\mathcal{E}}^2}\right)\Gamma\left(-n, \frac{\sigma_Z^2}{\delta\bar{P}\sigma_{\mathcal{E}}^2}\right), \quad \text{with } n = M_T - 1. \end{aligned}$$

From expression (3.21) and computing the relative entropy, the minimization in (3.17) writes

$$C_{\mathcal{M}}^{\mathrm{MIMO}}(\mathbf{H}, \widehat{\mathbf{H}}) = \begin{cases} \min\limits_{\Upsilon} \quad \log_2 \det\left(\mathbb{I}_{M_R} + \Upsilon\mathbf{\Sigma_P}\Upsilon^\dagger\mathbf{\Sigma}^{-1}\right), \\ \text{subject to} \quad \|\Upsilon + a_{\mathcal{M}}\widehat{\mathbf{H}}\|_F^2 \leq \|\mathbf{H} + a_{\mathcal{M}}\widehat{\mathbf{H}}\|_F^2 + \mathrm{C}, \end{cases} \tag{3.22}$$

where $\mathbf{\Sigma}$ must be chosen such that $tr\big(\Upsilon\mathbf{\Sigma_P}\Upsilon^\dagger + \mathbf{\Sigma}\big) = tr\big(\mathbf{H}\mathbf{\Sigma_P}\mathbf{H}^\dagger + \mathbf{\Sigma_0}\big)$. In order to obtain a simpler and more tractable expression of (3.22), we consider the following decomposition of the matrix $\mathbf{H} = \mathbf{U}\,\mathrm{diag}(\underline{\lambda})\mathbf{V}^\dagger$ with $\underline{\lambda} = (\lambda_1, \ldots, \lambda_{M_R})^T$. Let $\mathrm{diag}(\underline{\mu})$ be a diagonal matrix such that $\mathrm{diag}(\underline{\mu}) = \mathbf{U}^\dagger\Upsilon\mathbf{V}$, whose diagonal values are given by the vector $\underline{\mu} = (\mu_1, \ldots, \mu_{M_R})^T$. We define $\widetilde{\mathbf{H}}^\dagger = \mathbf{V}^\dagger\widehat{\mathbf{H}}^\dagger\mathbf{U}$, the vector $\tilde{\mathbf{h}}^\dagger = \mathrm{diag}(\widetilde{\mathbf{H}}^\dagger)^T$ resulting of its diagonal and let $b_{\mathcal{M}} = \|\mathbf{H} + a_{\mathcal{M}}\widehat{\mathbf{H}}\|_F^2 - a_{\mathcal{M}}^2(\|\widetilde{\mathbf{H}}\|_F^2 - \|\tilde{\mathbf{h}}\|^2)$. Using the above definitions and some algebra, the optimization (3.22) becomes equivalent to

$$C_{\mathcal{M}}^{\mathrm{MIMO}}(\mathbf{H}, \widehat{\mathbf{H}}) = \begin{cases} \min\limits_{\underline{\mu}} \quad \sum\limits_{i=1}^{M_R} \log_2\left(1 + \dfrac{\bar{P}|\mu_i|^2}{\sigma^2(\underline{\mu})}\right), \\ \text{subject to} \quad \|\underline{\mu} + a_{\mathcal{M}}\tilde{\mathbf{h}}\|^2 \leq b_{\mathcal{M}}, \end{cases} \tag{3.23}$$

with $\sigma^2(\underline{\mu}) = \frac{\bar{P}}{M_R}(\|\underline{\lambda}\|^2 - \|\underline{\mu}\|^2) + \sigma_Z^2$. The constraint set in the minimization (3.23), which corresponds to the set of vectors $\{\underline{\mu} \in \mathbb{C}^{M_T \times 1} : \|\underline{\mu} + a_{\mathcal{M}}\tilde{\mathbf{h}}\|^2 \leq b_{\mathcal{M}}\}$, is a closed convex polyhedral set. Thus, the infimun in (3.23) is attainable at the extremal of the set given by the equality (cf. [84]). Furthermore, for every vector $\underline{\mu}$ such that $\|\underline{\mu}\|^2 \leq \|\underline{\lambda}\|^2$, we observe that the expression (3.23) is a monotone increasing function of the square norm of $\underline{\mu}$. As a consequence, it is sufficient to find the optimal vector $\underline{\mu}_{\mathcal{M}}^{\mathrm{opt}}$ by minimizing the square norm over the constraint set. This becomes a classical

minimization problem that can be easily solved by using Lagrange multipliers. The corresponding achievable rates are then presented in the following corollary.

**Corollary 3.5.1** *Given a pair of matrices* $(\mathbf{H}, \widehat{\mathbf{H}})$ *the following information rates can be achieved by a receiver using the decoding rule* (3.5) *based on the metric* (3.14)*, for uncorrelated MIMO channels,*

$$C_{\mathcal{M}}^{MIMO}(\mathbf{H}, \widehat{\mathbf{H}}) = \log_2 det\left(\mathbb{I}_{M_R} + \Upsilon_{opt}\mathbf{\Sigma_P}\Upsilon_{opt}^{\dagger}\sigma^{-2}(\underline{\mu}_{\mathcal{M}}^{opt})\right), \tag{3.24}$$

*where the optimal solution* $\Upsilon_{opt} = \mathbf{U}\, diag(\underline{\mu}_{\mathcal{M}}^{opt})\mathbf{V}^{\dagger}$ *with*

$$\underline{\mu}_{\mathcal{M}}^{opt} = \begin{cases} \left(\dfrac{\sqrt{b_{\mathcal{M}}}}{\|\tilde{\mathbf{h}}\|} - |a_{\mathcal{M}}|\right)\tilde{\mathbf{h}} & \text{if } b_{\mathcal{M}} \geq 0, \\ \underline{0} & \text{otherwise,} \end{cases} \tag{3.25}$$

and $\sigma^2(\underline{\mu}_{\mathcal{M}}^{\text{opt}}) = \frac{\bar{P}}{M_R}(\|\underline{\lambda}\|^2 - \|\underline{\mu}_{\mathcal{M}}^{\text{opt}}\|^2) + \sigma_Z^2$.

### 3.5.2 Achievable Information Rates Associated to the Mismatched ML decoder

Next, we aim at comparing the achievable rates obtained in (3.24) to those provided by the classical mismatched ML decoder (3.11). Following the same steps as above, we can compute the achievable rates associated to the mismatched ML decoder. In this case, the minimization problem writes

$$C_{\text{ML}}^{\text{MIMO}}(\mathbf{H}, \widehat{\mathbf{H}}) = \begin{cases} \min\limits_{\Upsilon} \quad \log_2 \det\left(\mathbb{I}_{M_R} + \Upsilon\mathbf{\Sigma_P}\Upsilon^{\dagger}\mathbf{\Sigma}^{-1}\right), \\ \text{subject to} \quad \mathbb{R}e\{tr(\mathbf{H}\mathbf{\Sigma_P}\widehat{\mathbf{H}}^{\dagger})\} \leq \mathbb{R}e\{tr(\Upsilon\mathbf{\Sigma_P}\widehat{\mathbf{H}}^{\dagger})\}, \end{cases} \tag{3.26}$$

where $\mathbf{\Sigma}$ must be chosen such that $tr\left(\Upsilon\mathbf{\Sigma_P}\Upsilon^{\dagger}+\mathbf{\Sigma}\right) = tr\left(\mathbf{H}\mathbf{\Sigma_P}\mathbf{H}^{\dagger}+\mathbf{\Sigma_0}\right)$. The resulting achievable rates are given by

$$C_{\text{ML}}^{\text{MIMO}}(\mathbf{H}, \widehat{\mathbf{H}}) = \log_2 \det\left(\mathbb{I}_{M_R} + \Upsilon_{\text{opt}}\mathbf{\Sigma_P}\Upsilon_{\text{opt}}^{\dagger}\sigma^{-2}(\underline{\mu}_{\text{ML}}^{\text{opt}})\right), \tag{3.27}$$

where $\Upsilon_{\text{opt}} = \mathbf{U}\,\text{diag}(\underline{\mu}_{\text{ML}}^{\text{opt}})\mathbf{V}^{\dagger}$ and

$$\begin{aligned} \sigma^2(\underline{\mu}_{\text{ML}}^{\text{opt}}) &= \frac{\bar{P}}{M_T}(\|\underline{\lambda}\|^2 - \|\underline{\mu}_{\text{ML}}^{\text{opt}}\|^2) + \sigma_Z^2, \\ \underline{\mu}_{\text{ML}}^{\text{opt}} &= \frac{\mathbb{R}e\{tr(\Lambda^{\dagger}\tilde{\mathbf{h}})\}}{\|\tilde{\mathbf{h}}\|^2}\tilde{\mathbf{h}}. \end{aligned} \tag{3.28}$$

### 3.5.3 Estimation-Induced Outage Rates

Through this section, we have so far considered instantaneous achievable rates over MIMO (3.24) channels. We now provided its associated outage rates, according to the notion of EIO capacity defined in section 3.2.2. In order to compute these outage rates, it is necessary to calculate the outage probability as a function of the outage rate. Given outage rate $R \geq 0$ and channel estimate $\widehat{\mathbf{H}}$, the outage probability is defined as

$$P_{\mathcal{M}}^{\text{out}}(R, \widehat{\mathbf{H}}) = \int_{\left\{\mathbf{H} \in \mathbb{C}^{M_R \times M_T} : C_{\mathcal{M}}(\mathbf{H}, \widehat{\mathbf{H}}) < R\right\}} d\psi_{\mathbf{H}|\widehat{\mathbf{H}}}(\mathbf{H}|\widehat{\mathbf{H}}),$$

then the maximal outage rate for an outage probability $\gamma_{_{QoS}}$ is given by

$$C_{\mathcal{M}}^{\text{out}}(\gamma_{_{QoS}}, \widehat{\mathbf{H}}) = \sup_{R} \left\{ R \geq 0 : P_{\mathcal{M}}^{\text{out}}(R, \widehat{\mathbf{H}}) \leq \gamma_{_{QoS}} \right\}. \tag{3.30}$$

Since this outage rate still depends on the channel estimate, we consider the average over all channel estimates as $\overline{C}_{\mathcal{M}}^{\text{out}}(\gamma_{_{QoS}}) = \mathbb{E}_{\widehat{\mathbf{H}}}\left\{C_{\mathcal{M}}^{\text{out}}(\gamma_{_{QoS}}, \widehat{\mathbf{H}})\right\}$. These achievable rates are upper bounded by the mean outage rates given by the EIO capacity, which provides the maximal outage rate (i.e. maximizing over all possible receiver using the channel estimates), achieved by a theoretical decoder. In our case, this capacity is given by $\overline{C}(\gamma_{_{QoS}}) = \mathbb{E}_{\widehat{\mathbf{H}}}\left\{C(\gamma_{_{QoS}}, \widehat{\mathbf{H}})\right\}$, where $C(\gamma_{_{QoS}}, \widehat{\mathbf{H}})$ can be computed from (3.4) by setting $\theta = \mathbf{H}$ and $\hat{\theta} = \widehat{\mathbf{H}}$.

## 3.6 Simulation Results

In this section we provide numerical results to analyze the performance of a receiver using the decoder (3.5) based on the metric (3.14). We consider uncorrelated Rayleigh fading MIMO channels, assuming that the channel changes for each compound symbol inside the frame of $N_c = 50$ symbols. This assumption was made because of BICM, in oder to let the interleaver to work. The performances are measured in terms of BER and achievable outage rates. The binary information data is encoded by a rate 1/2 non-recursive non-systematic convolutional (NRNSC) channel code with constraint length 3 defined in octal form by $(5, 7)$. The interleaver is a random one operating over the entire frame with size $N_c\, M_T \log_2(B)$ bits and the symbols belonging to a

16-QAM constellation with Gray and set-partition labeling. Besides, it is assumed that the average pilot symbol energy is equal to the average data symbol energy.

### 3.6.1 Bit Error Rate Analysis of BICM Decoding Under Imperfect Channel Estimation

Here, we compare BER performances between the proposed decoder (3.14) and the mismatched decoder (3.11) for BICM decoding (section IV). Fig. 3.3 and 3.4 show, for a $2 \times 2$ MIMO channel ($M_T = M_R = 2$), the increase in the required $E_b/N_0$ caused by decoding with the mismatched ML decoder in presence of CEE. For comparison, BER obtained with perfect CSIR are also presented. In this case, we need at least 2 pilot symbols to estimate the channel matrix $\mathbf{H}$, since $N \geq M_T$. Thus, we insert $N = 2, 4$ or 8 pilots per frame for channel training. At BER $= 10^{-4}$ and $N = 2$, we observe about 1.4 dB of SNR gain by using the proposed decoder. We also note that the performance loss of the mismatched receiver with respect to our receiver becomes insignificant for $N \geq 8$. This can be explained from (3.15), since by increasing the number of pilot symbols both decoders coincide. Results show that the decoder under investigation outperforms the mismatched decoder, especially when few numbers of pilots are dedicated for training.

### 3.6.2 Achievable Outage Rates Using the Derived Metric

Numerical results concerning achievable information rates decoding with the investigated metric over fading MIMO channels are based on Monte Carlo simulations.

Fig. 3.5 compares average outage rates (in bits per channel use) over all channel estimates, of both mismatched ML decoding (given by expression (3.27)) and the proposed metric (given by (3.24)) versus the SNR. The $2 \times 2$ MIMO channel is estimated by sending $N = 2$ pilot symbols per frame, and the outage probability has been fixed to $\gamma_{QoS} = 0.01$. For comparison, we also display the upper bound of these rates given by the EIO capacity (obtained by evaluating the expression (3.4)), and the capacity with perfect channel knowledge. It can be observed that the achievable rate using the mismatched ML decoding is about 5 dB (at a mean outage rate of 6 bits) of SNR far from the EIO capacity. Whereas, we note that the proposed decoder achieves higher

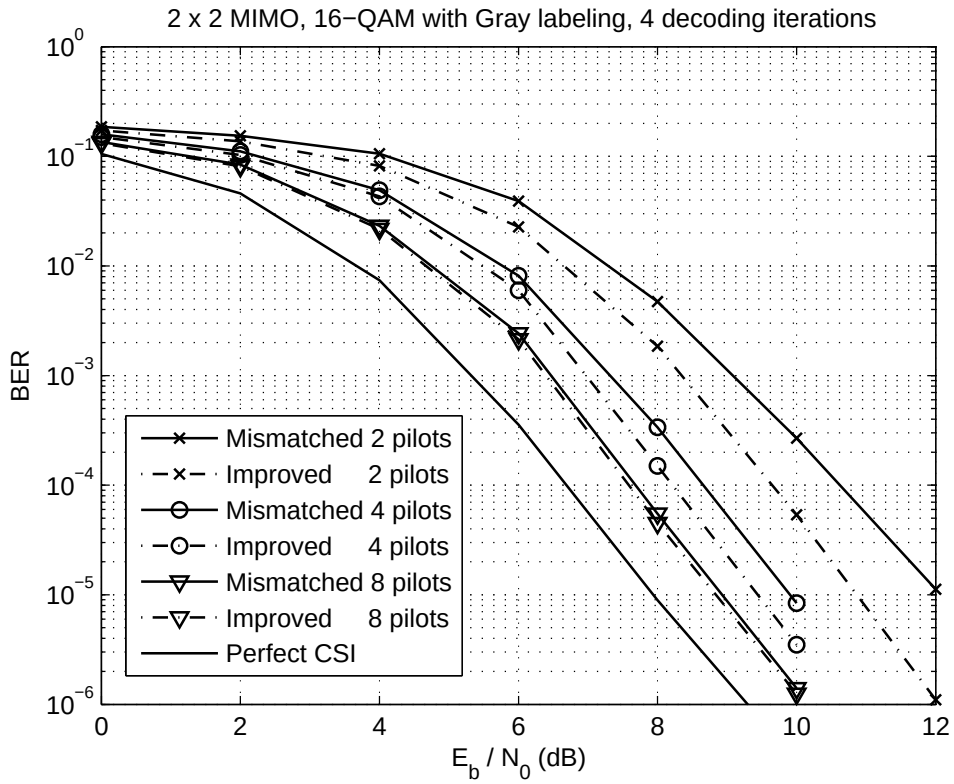



Figure 3.3: BER performances over $2 \times 2$ MIMO with Rayleigh fading for various training sequence lengths and Gray labeling.

rates for any SNR values and decreases by about 1.5 dB the aforementioned SNR gap.

Similar plots are shown in Fig. 3.6 in the case of a $4 \times 4$ MIMO channel estimated by sending training sequences of length $N = 4$. Again, it can be observed that the modified decoder achieves higher rates than the mismatched decoder. However, we note that the performance degradation using the mismatched decoder has decreased to less than 1 dB (at a mean outage rate of 10 bits). This observation is a consequence of using orthogonal training sequences that requires $N \geq M_T$, since the CEE can be reduced by increasing the number of antennas [97].

Note that, the achievable rates of the proposed decoder are still about 3 dB far from the ultimate performance given by the EIO capacity. However, it provides significative gains in terms of information rates compared to the classical mismatch approach.

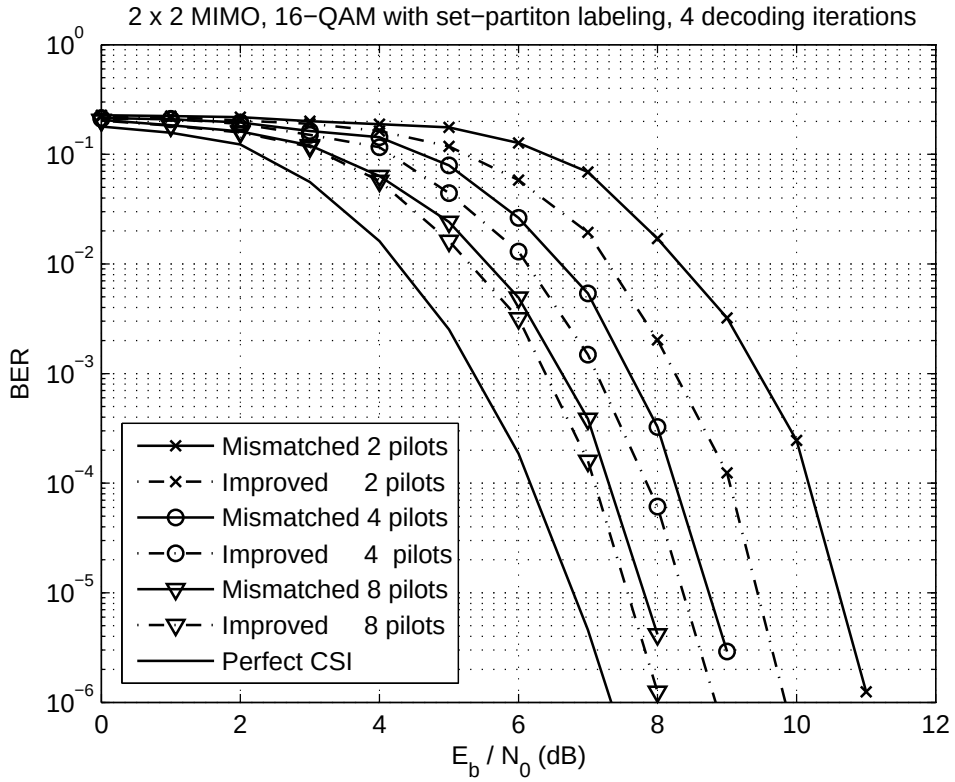



Figure 3.4: BER performances over $2 \times 2$ MIMO with Rayleigh fading for various training sequence lengths and set-partition labeling.

## 3.7 Summary

This chapter studied the problem of reception in practical communication systems, when the receiver has only access to noisy estimates of the channel and these estimates are not available at the transmitter. Specifically, we focused on determining the optimal decoder that achieves the EIO capacity of arbitrary memoryless channels under imperfect channel estimation. By using the tools of information theory, we derived a practical decoding metric that minimizes the average of the transmission error probability over all CEE. This decoder is not optimal in the sense that it cannot achieve the EIO capacity. In contrast, this decoder achieves the capacity of a composite (more noisy) channel.

By using the general decoding metric, we analyzed the case of uncorrelated fading MIMO channels. Then, we used this metric for iterative BICM decoding of MIMO systems with ML channel estimation. Moreover, we obtained the maximal achievable rates, using Gaussian codebooks, associated to the proposed decoder and compared

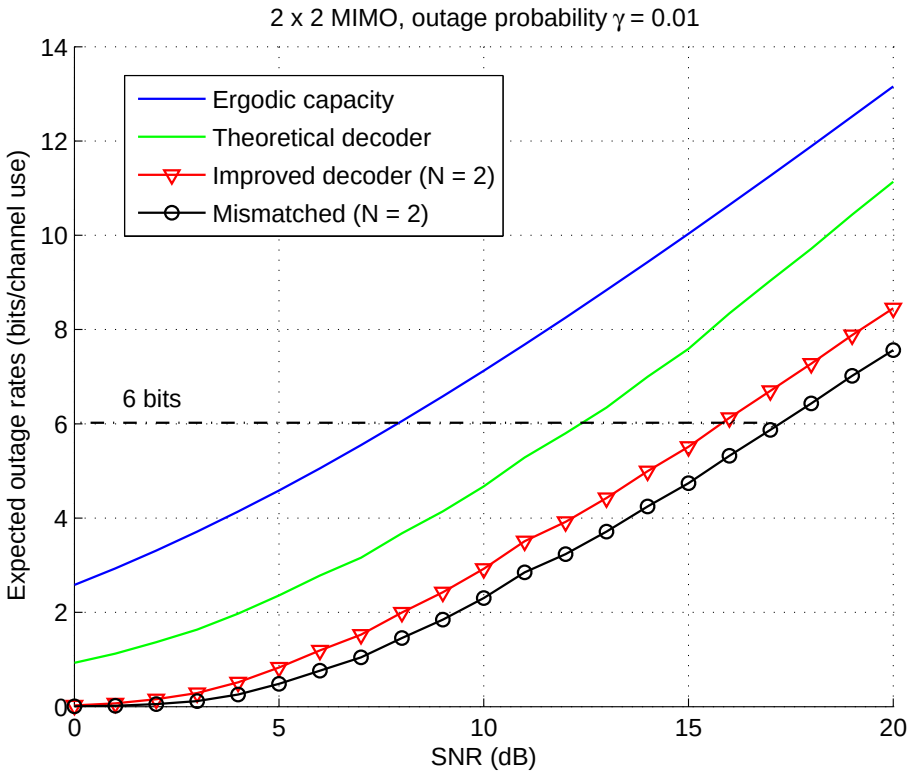



Figure 3.5: Expected outage rates over $2 \times 2$ MIMO with Rayleigh fading versus SNR ($N = 2$).

these rates to those of the classical mismatched ML decoder. Simulation results indicate that mismatched ML decoding is sub-optimal under short training sequences, in terms of both BER and achievable outage rates, and confirmed the adequacy of the proposed decoder.

Although we showed that the proposed decoder outperforms classical mismatched approaches, the derivation of a practical decoder that maximizes the EIO capacity (over all possible theoretical decoders) under imperfect channel estimation, is still an open problem in its full generality. Nevertheless, other types of decoding metrics incorporating also the outage probability value, have yet to be fully explored.

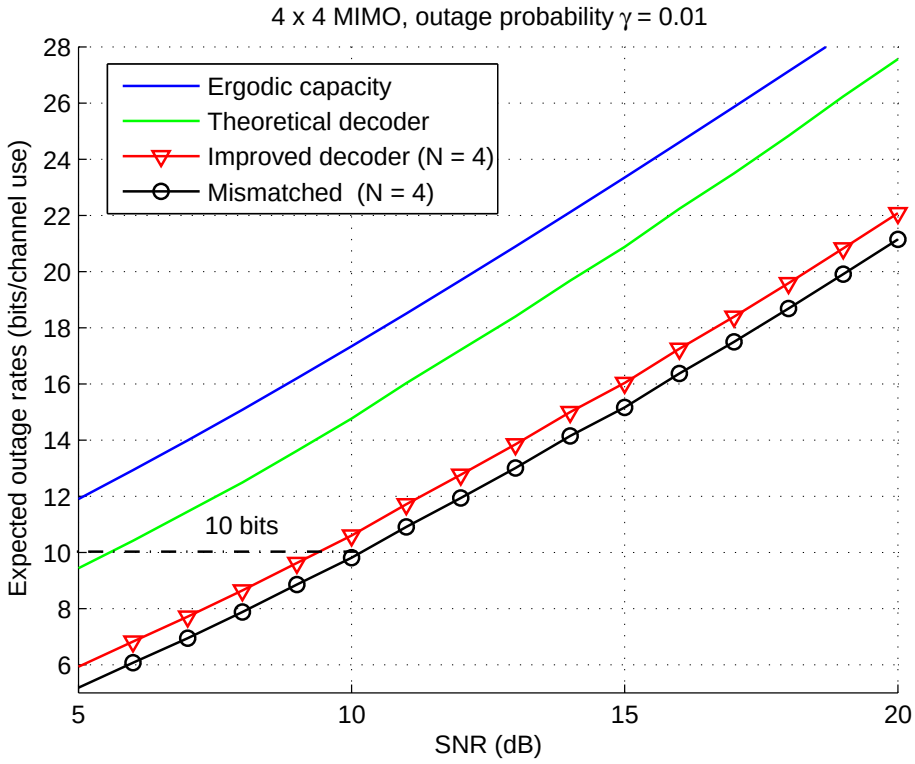



Figure 3.6: Expected outage rates over $4 \times 4$ MIMO with Rayleigh fading versus SNR ($N = 4$).

# Chapter 4

# Dirty-Paper Coding with Imperfect Channel Knowledge: Applications to the Fading MIMO Broadcast Channel

The effect of imperfect channel estimation at the receiver with imperfect (or without) channel knowledge at the transmitter on the capacity of state-dependent channels with non-causal channel state information at the transmitter is examined. We address this problem through the notion of reliable communication based on the average of the transmission error probability over all channel estimation errors, assuming a discrete memoryless channel. This notion allows us to consider the capacity of a composite (more noisy) Gelfand and Pinsker's channel. We first derive the optimal Dirty-paper coding (DPC) scheme, by assuming Gaussian inputs, achieving the capacity of the single-user fading Costa channel with maximum-likehood (ML) channel estimation. Our results, for uncorrelated Rayleigh fading, illustrate a practical trade-off between the amount of training and its impact to the interference cancellation performances of DPC scheme. These are useful in realistic scenarios of multiuser wireless communications and information embedding applications (e.g. robust watermarking). We also studied optimal training design adapted to each of these applications.

Next, we exploit the tight relation between the largest achievable rate region (Mar-

ton's region) for arbitrary broadcast channels and channels with non-causal channel state information at the transmitter to extend this region to the case of imperfect channel knowledge. We derive achievable rate regions and optimal DPC schemes assuming Gaussian codebooks, for a base station transmitting information over a multiuser Fading MIMO Broadcast Channel (MIMO-BC), where the mobiles (the receivers) only dispose of a noisy estimate of the channel parameters, and these estimates may be (or not) available at the base station (the transmitter).

These results are particularly useful for a system designer to assess the amount of training data and the channel characteristics (e.g. SNR, fading process, power for training, number of antennas) to achieve target rates. We provide numerical results for a two-users MIMO-BC with ML or minimum mean square error (MMSE) channel estimation. The results illustrate an interesting practical trade-off between the benefit of an elevated number of transmit antennas and the amount of training needed. In particular, we observe the surprising result that a BC with a single transmitter and receiver antenna, and imperfect channel estimation at the receivers, does not need the knowledge of estimates at the transmitter to achieve large rates compared to time-division multiple access (TDMA).

## 4.1 Introduction

Consider the problem of communicating over a discrete memoryless channel (DMC) defined by a conditional distribution $W(y|x,s)$ where $X \in \mathscr{X}$ is the channel input, $S \in \mathscr{S}$ is the random channel state with distribution $P_S$ and $Y \in \mathscr{Y}$ is the channel output. The transmitter knows the channel states before beginning the transmission (i.e. non-causal state information) but the receiver does not know these. This channel is commonly known as *channel with non-causal state information at the transmitter*. The capacity expression of this channel has been derived by Gelfand and Pinsker in [33],

$$C\big(W, P_S\big) = \sup_{P(u,x|s)\in\mathscr{P}} \big\{ I\big(P_U, W\big) - I\big(P_S, P_{U|S}\big) \big\}, \tag{4.1}$$

where $U \in \mathcal{U}$ is an auxiliary random variable chosen so that $U \ominus (X,S) \ominus Y$ form a Markov Chain, $I(\cdot)$ is the classical mutual information and $\mathscr{P}$ is the set of all joint probability distributions $P(u,x|s) = \delta\big(x - f(u,s)\big)P(u|s)$ with $f : \mathscr{U} \times \mathscr{S} \mapsto \mathscr{X}$

an arbitrary mapping function and $\delta(\cdot)$ is the *dirac* function. In "Writing on Dirty Paper" [67], Costa applied this result to an additive white Gaussian noise (AWGN) channel corrupted by an additive Gaussian interfering signal $S$ that is non-causally known at the transmitter. The channel state $S$ is a Gaussian variable with power $Q$ independent of the Gaussian noise $Z$; the channel output $Y = X + S + Z$ and its input $X$ of limited-power $\bar{P}$ (often $\ll Q$). He showed the simple but surprising result that choosing the auxiliary variable $U = X + \alpha S$ with an appropriate value $\alpha^* = \bar{P}(\bar{P}+\sigma_Z^2)^{-1}$, where $\sigma_Z^2$ being the AWGN variance, this coding scheme referred as *Dirty-paper coding (DPC)*, allows one to achieve the same capacity as if the interfering signal $S$ was not present.

This result has gained considerable attention during the last years, mainly because of its potential use in communication scenarios where interference cancellation at the transmitter is needed. In particular, information embedding (robust watermarking for multimedia security applications) [98] and multiuser interference cancellation for Broadcast Channels (BC) [63] are instances of such scenarios. Indeed, this result has been the focus of intense study and some remarkable progress has already been made in several of its applications. However, there is still an important question regarding the assumptions under which interference cancellation through the use of DPC holds. This assumes that both the transmitter and receiver perfectly know the channel statistic $W$ controlling the communication. Therefore, it is not clear if the surprising performances of DPC still hold in practical situations where imperfect (or no) channel knowledge is available. Throughout this chapter, we investigate this question in the context of the fading Costa channel and the Fading Multiple-Input-Multiple-Output Broadcast Channel (MIMO-BC).

### 4.1.1 Related and Subsequent Work

The capacity region of a general BC is still unknown. Whereas Marton in [55] found an achievable rate region for the general discrete memoryless broadcast channel, which is the largest known inner bound to the capacity region. In the recent years, the Fading MIMO-BC has been extensively studied. Most of the literature focuses on the information-theoretic performances under the assumption on the availability of the time-varying channel matrices at both transmitter and all receivers. Caire and

Shamai in [63], have established an achievable rate region, referred to as the DPC region. They conjectured that this achievable region is the capacity. Recently in [64], Weingarten, Steinberg and Shamai prove this conjecture by showing that the DPC region is equal to the capacity region. Furthermore, this region is shown to be tight to the inner bound given by the Marton's region.

The great attraction of the fading MIMO-BC is that under the assumption of perfect channel knowledge, as the signal-to-noise ratio (SNR) tends to infinity, the limiting ratio between the sum-rate capacity and the capacity of a single-user channel that results when the receiver allowed to cooperate is one. Thus, for a BC where the receivers cannot cooperate, the interference cancellation implemented by DPC results in no asymptotic loss.

Nevertheless, it is well-known that the performances of wireless systems are severely affected if only noisy channel estimates are available (cf. [58], [59] and chapter 2). Of particular interest is the issue of the effect of this imperfect knowledge on the multiuser interference cancellation implemented by DPC scheme. In such scenario, the error on the channel estimation of some user affects the achievable rates of many other users. Furthermore, the problem may even be more serious in practical situations where no channel information is available at the transmitter, i.e., there is no feedback information from the receiver to the transmitter covering the channel estimates.

Consequently, when the channel is imperfectly known (or unknown), it is not immediately clear whether it is more efficient to send information to only a single user at a time (i.e. time-division multiple-access TDMA) rather than to use multiuser interference cancellation (cf. [99] and [100]). In addition to this, from a practical point of view, the system designer must decide the amount of training and power required to achieve a target pair of rates.

For these reasons, the limits of reliable information rates of Fading MIMO-BCs with imperfect channel information is an important problem. Indeed, intensive recent research has been conducted, e.g. Sharif and Hassibi in [101] proposed an opportunistic coding scheme that employs only partial information. They show that the optimal scaling factor of the sum-rate capacity is the same one as obtained with perfect channel knowledge using DPC. References in [102] already derive a lower bound of the capacity of MIMO-BC with MMSE channel estimation and perfect feedback. This

approach parallels that by Yoo and Goldsmith [59], which was initially introduced by Medard in [58], where the authors have been derived similar bounds on the capacity of single-user MIMO channels. Whereas in [65], Lapidoth, Shamai and Wigger show that when the transmitter only has an estimate of the channel and the receivers have perfect channel knowledge, the limiting ratio between the sum-rate capacity and the capacity of a single-user channel with cooperating receivers is upper bounded by 2/3. Recently, Jindal in [103] investigates a system where each receiver has perfect channel knowledge, but the transmitter only receives quantized information regarding the channel instantiation. A similar work has been carried out in [104], considering downlink systems with more users than transmitter antennas and finite rate feedback at the transmitter.

### 4.1.2 Outline of This Work

In the first part of this chapter (section 4.2), we consider the natural extension of DMCs $W(y|x,s,\theta)$ with channel states $S$ non-causally known at the transmitter, to the more realistic case where neither the transmitter nor the receiver know the random parameters $\theta$ controlling the communication. We assume that the receiver obtains an estimate $\hat{\theta}$ during a phase of independent training and its estimate may be (or not) available at the transmitter. We address this problem through the notion of reliable communication based on the average of the error probability over all channel estimation errors (CEE). This is done by incorporating in the capacity definition the statistic characterizing the quality of channel estimates, i.e., the *a posteriori* pdf of the unknown channel conditioned on its estimate (it is available from the family of channel pdfs controlling the communication and the estimator chosen). This novel notion allows us to make a connection between the capacity of the Gelfand and Pinsker's channel (4.1) and the capacity of a composite (more noisy) channel. Based on this setting, we formulate the analogue of the Marton's region for arbitrary discrete memoryless BCs with imperfect channel estimation.

In the second part of this chapter (section 4.3), based on our previous approach, we first consider the special case of a single-user fading Costa channel modeled as $Y = H(X+S)+Z$, where $\theta = H$ is the random channel estimated at the receiver by using maximum-likelihood (ML) channel estimation. We study the cases where these

channel estimates may be (or not) available at the transmitter. Here, we determine the optimal trade-off between the amount of training required for channel estimation and the corresponding achievable rates using an optimal DPC scheme under CEE. We observe that depending on the targeted application, multiuser interference cancellation or robust watermarking, two different training scenarios are relevant, for which adequate training design is proposed. Then, in section 4.4 we focus on the capacity region of the multiuser Fading MIMO-BC with imperfect channel estimation. We assume that the channel is estimated at each receiver using ML or minimum mean square error (MMSE) channel estimation. Two scenarios are considered: (i) We first assume that an instantaneous error-free feedback provides the transmitter with the channel estimates of each receiver and (ii) we suppose that there is no feedback from the receivers back to the transmitter conveying these channel estimates. For each of these scenarios, we derive the corresponding optimal DPC scheme and its achievable rate region, assuming Gaussian codebooks.

The proposed framework in this work is sufficiently general to involve the most important application scenarios in information embedding and multiuser communications. In particular, this can be easily extended by using recent results (e.g. [103] and [104]) to the more general scenarios considering both noisy feedback and imperfect channel estimation. Section 4.5 illustrates average rates over all channel estimates of the fading Costa channel, for different amount of training. Moreover, we use a two-users uncorrelated Rayleigh-fading MIMO-BC to show average rates for different amount of training and antenna configurations. Finally, section 4.4 concludes the chapter.

Notational conventions are as follows: upper and lower case bold symbols are used to denote matrices and vectors; $\mathbb{I}_M$ represents an $(M \times M)$ identity matrix; $\mathbb{E}_{\mathbf{X}}\{\cdot\}$ refers to expectation with respect to the random vector $\mathbf{X}$; $|\cdot|$ denotes matrix determinant; $(\cdot)^T$ and $(\cdot)^\dagger$ denote vector transpose and Hermitian transpose, respectively.

## 4.2 Channels with non-Causal CSI and Imperfect Channel Estimation

In this section, we first introduce the single-user DMC with non-causal channel state information at the transmitter and the notion of reliable communication based on the average of the error probability over all CEE. This notion allows us to consider the capacity of a composite (more noisy) channel. Subsequently we use a similar approach to find the equivalent Marton's region for the case of BCs with imperfect channel estimation.

### 4.2.1 Single-User State-Dependent Channels

Consider a general model for communication under channel uncertainty over DMCs with input alphabet $\mathscr{X}$, output alphabet $\mathscr{Y}$ and states $\mathscr{S}$ (cf. [33] and [30]). A specific instance of the unknown channel is characterized by a transition probability mass (PM) $W(\cdot|x,s,\theta) \in \mathcal{W}_\Theta$ with a random state $s \in \mathscr{S}$ perfect known by the transmitter and a fixed but unknown channel $\theta \in \Theta \subseteq \mathbb{C}^d$. Here, $\mathcal{W}_\Theta = \big\{W(\cdot|x,s,\theta) \colon x \in \mathscr{X}, s \in \mathscr{S}, \theta \in \Theta\big\}$ is a family of conditional transition PMs on $\mathscr{Y}$, parameterized by a vector $\theta \in \Theta$, which each realization follows i.i.d. $\theta_i \sim f_\theta(\theta)$.

Assume that the coherence time is sufficiently long and thus the transmitter can send a training sequence that allows the receiver to estimate the channel $\theta_i$. Thus, the receiver only knows a channel estimate $\hat{\theta}_i$ and a characterization of the estimator performance in terms of the conditional probability density function (pdf) $f_{\theta|\hat{\theta}}(\theta|\hat{\theta})$. This can be easily obtained using $\mathcal{W}_\Theta$, the estimator function and $f_\theta(\theta)$. In this context we identify two different scenarios: (i) The transmitter knows the channel estimates $\hat{\theta}_i$ and (ii) the transmitter does not know the channel estimates, only its statistic $f_{\hat{\theta}}(\hat{\theta})$ is available. The memoryless extension of $W(\cdot|x,s,\theta)$ within a block of length $n$ is given by $W^n(\mathbf{y}|\mathbf{x},\mathbf{s},\underline{\theta}) = \prod_{i=1}^n W(y_i|x_i,s_i,\theta_i)$ where $\mathbf{x} = (x_1,\ldots,x_n)$, $\mathbf{s} = (s_1,\ldots,s_n)$ and each realization follows independent and identically distributed (i.i.d.) $s_i \sim P_S(s)$ and $\mathbf{y} = (y_1,\ldots,y_n)$. The sequence of channel state $\mathbf{s}$ is perfectly known at the transmitter before sending $\mathbf{x}$ and unknown at the receiver.

### 4.2.2 Notion of Reliable Communication and Coding Theorem

A message $m$ from the set $\mathcal{M} = \{1, \ldots, \lfloor 2^{n\bar{R}} \rfloor\}$ is transmitted using a length-$n$ block code defined as a pair $(\varphi, \phi)$ of mappings, where $\varphi : \mathcal{M} \times \mathscr{S}^n \times \Theta^n \mapsto \mathscr{X}^n$ is the encoder (that utilize $\underline{\hat{\theta}}$ if available), and $\phi : \mathscr{Y}^n \times \Theta^n \mapsto \mathcal{M}$ is the decoder (that utilizes $\underline{\hat{\theta}}$). Note that the encoder uses the realization of the state sequence $\mathbf{s}$, which is exploited for encoding the information messages $m \in \mathcal{M}$. The average rate over all channel estimates $\hat{\theta}$, is given by $\mathbb{E}_{\hat{\theta}}\{n^{-1}\log_2 M_{\hat{\theta}}\}$ and the maximum (over all messages) of the average of the error probability over all CEE

$$\bar{e}^{(n)}_{\max}(\varphi, \phi, \underline{\hat{\theta}}) = \max_{m \in \mathcal{M}} \mathbb{E}_{\underline{\theta}\mathbf{S}|\underline{\hat{\theta}}}\Big\{ \sum_{\mathbf{y} \in \mathscr{Y}^n : \phi(\mathbf{y}, \underline{\hat{\theta}}) \neq m} W^n\big(\mathbf{y}|\varphi(m, \mathbf{s}, \underline{\hat{\theta}}), \mathbf{s}, \underline{\theta}\big)\Big\}. \tag{4.2}$$

where the joint pdf $P(\underline{\theta}, \mathbf{s}|\underline{\hat{\theta}}) = \prod_{i=1}^{n} f_{\theta|\hat{\theta}}(\theta_i|\hat{\theta}_i) P_S(s_i)$.

For a given $0 < \epsilon < 1$, a mean rate $\bar{R} \geq 0$ is $\epsilon$-achievable on an estimated channel, if for every $\delta > 0$ and every sufficiently large $n$ there exists a sequence of length-$n$ block codes such that the rate satisfies $\mathbb{E}_{\hat{\theta}}\{n^{-1}\log_2 M_{\hat{\theta}}\} \geq \bar{R} - \delta$ and $\bar{e}^{(n)}_{\max}(\varphi, \phi, \underline{\hat{\theta}}) \leq \epsilon$. This definition requires that maximum of the averaged error probability occurs with probability less than $\epsilon$. For a more robust notion of reliability over single-user channels we refer the reader to chapter 2. Then, a mean rate $\bar{R} \geq 0$ is achievable if it is $\epsilon$-achievable for every $0 < \epsilon < 1$, and let $\bar{C}_\epsilon$ be the largest $\epsilon$-achievable rate. The capacity is then defined as the largest achievable mean rate, $\bar{C} = \lim_{\epsilon \downarrow 0} \bar{C}_\epsilon$. We next state a theorem quantifying this capacity.

**Theorem 4.2.1** *The capacity of a DMC $W(\cdot|x, s, \theta)$ with non-causal channel state information at the transmitter and imperfect channel estimation, is given by $\bar{C}_{01}$ when the channel estimates are not available at the transmitter and othercase $\bar{C}_{11}$,*

$$\bar{C}_{01}(W) = \sup_{P(u,x|s) \in \mathscr{P}_{01}} \mathbb{E}_{\hat{\theta}}\big\{\mathscr{C}\big(P(u, x|s), \hat{\theta}\big)\big\}, \tag{4.3}$$

$$\bar{C}_{11}(W) = \mathbb{E}_{\hat{\theta}}\Big\{ \sup_{P_{\hat{\theta}}(u,x|s) \in \mathscr{P}_{11}} \mathscr{C}\big(P_{\hat{\theta}}(u, x|s), \hat{\theta}\big)\Big\}, \tag{4.4}$$

*where*

$$\mathscr{C}\big(P(u, x|s), \hat{\theta}\big) = I\big(P_U, \widetilde{W}_{\hat{\theta}}\big) - I\big(P_S, P_{U|S}\big). \tag{4.5}$$

In this theorem $\mathscr{P}_{11}$ denotes the set of probability distributions so that $(U,\hat{\theta}) \multimap (X,S,\theta) \multimap Y$ form a Markov chain, while we emphasize that the supremum in (4.4) is taken over the set $\mathscr{P}_{01}$ of input distributions not depending on the channel estimates $\hat{\theta}$. The test channel is given by

$$\widetilde{W}(y|u,\hat{\theta}) = \sum_{(x,s)\in\mathscr{X}\times\mathscr{S}} \delta\big(x - f(u,s)\big) P_S(s) \widetilde{W}(y|x,s,\hat{\theta}), \tag{4.6}$$

and the composite (more noisy) channel $\widetilde{W}(y|x,s,\hat{\theta}) = \mathbb{E}_{\theta|\hat{\theta}}\big\{W(y|x,s,\theta)\big\}$, where $\mathbb{E}_{\theta|\hat{\theta}}\{\cdot\}$ denotes the expectation with the conditional pdf $f_{\theta|\hat{\theta}}$ characterizing the channel estimation errors. We also used the mutual information

$$I\big(P_U, \widetilde{W}_{\hat{\theta}}\big) = \sum_{u\in\mathscr{U}} \sum_{y\in\mathscr{Y}} P(u)\widetilde{W}(y|u,\hat{\theta}) \log_2 \frac{\widetilde{W}(y|u,\hat{\theta})}{Q(y|\hat{\theta})},$$

with $Q(y|\hat{\theta}) = \sum_{u\in\mathscr{U}} P(u)\widetilde{W}(y|u,\hat{\theta})$. The exposed situation can be reduced to that of Gelfand and Pinsker's channel [33], and hence does not lead to a new mathematical problem. The main differences are presented in appendix C.1.

### 4.2.3 Achievable Rate Region of Broadcast Channels with Imperfect Channel Estimation

We now explore the strong connection between the Marton's region and our previous formulation for channels with non-causal state information, to obtain a natural extension of this region for the case of imperfect channel estimation.

A broadcast channel is composed of one sender and many receivers. The objective is to broadcast information from a sender to the many receivers. Here, we consider broadcast channels with only two receivers since multiple receivers cases can be similarly treated. The discrete memoryless BC with one sender and two receivers consists of an input $X \in \mathscr{X}$ and two outputs $(Y_1, Y_2) \in \mathscr{Y}_1 \times \mathscr{Y}_2$ with a transition probability function $W(y_1,y_2|x,\underline{\theta}) \in \mathcal{W}_\Theta$, which is parameterized by the vectors of parameters $\underline{\theta} = (\theta_1,\theta_2) \in \Theta$, such that $Y_i \multimap (X,\theta_i) \multimap \theta_j$ with $j \neq i$ form a Markov chain, for which the joint realization follows i.i.d. $\underline{\theta} \sim f_{\underline{\theta}}(\underline{\theta})$. The capacity region of this BC only depends on the marginal PMs $W(y_1|x,\theta_1)$ and $W(y_2|x,\theta_2)$ (cf. [14], Theorem 14.6). We assume that each receiver $i$ only knows its channel estimate $\hat{\theta}_i$ and a characterization

of the estimator performance in terms of the conditional pdf

$$f_{\theta|\hat{\theta}}(\theta_i|\hat{\theta}_i) = \int_{\Theta}\int_{\Theta} f_{\underline{\theta}\hat{\theta}_j|\hat{\theta}_i}(\underline{\theta},\hat{\theta}_j|\hat{\theta}_i)d\theta_j d\hat{\theta}_j, \quad \text{with } j \neq i. \tag{4.7}$$

We emphasize that in this model the joint vector $\underline{\theta}$ of channel parameters may have correlated components $\theta_i$ and in such case each marginal pdf in (4.7) contains the estimation error of the other channel, which will be present in the capacity expression. Following the same steps as before, we can obtain the memoryless $n$-th extension of this channel and then define the average of the error probability (over all CEE) corresponding to each user. Next, we state the following achievable rate region.

**Theorem 4.2.2** *Let $(U_1, U_2) \in \mathscr{U}_1 \times \mathscr{U}_2$ be two arbitrary auxiliary random variables with finite alphabets such that $(U_1, U_2, \hat{\underline{\theta}}) \multimap (X, \underline{\theta}) \multimap (Y_1, Y_2)$ form a Markov chain. The following rate region is an inner bound of the capacity region of the discrete memoryless BC $W(y_1, y_2|x, \underline{\theta})$ with imperfect channel estimation*

$$\begin{aligned}\mathscr{R}(W) = co\Big\{(\bar{R}_1 \geq 0, \bar{R}_2 \geq 0) : \bar{R}_1 &\leq \mathbb{E}_{\hat{\underline{\theta}}}\big\{I\big(P_{U_1}, \widetilde{W}_{\hat{\theta}_1}\big)\big\},\\ \bar{R}_2 &\leq \mathbb{E}_{\hat{\underline{\theta}}}\big\{I\big(P_{U_2}, \widetilde{W}_{\hat{\theta}_2}\big)\big\},\\ \bar{R}_1 + \bar{R}_2 &\leq \mathbb{E}_{\hat{\underline{\theta}}}\big\{I\big(P_{U_1}, \widetilde{W}_{\hat{\theta}_1}\big) + I\big(P_{U_2}, \widetilde{W}_{\hat{\theta}_2}\big)\\ &\quad - I\big(P_{U_2}, P_{U_1|U_2}\big)\big\}, \quad \textit{for all } P_{\hat{\underline{\theta}}}(u_1, u_2, x) \in \mathscr{P}\Big\},\end{aligned} \tag{4.8}$$

*where $\mathscr{P}$ is the set of all distribution $P_{\hat{\underline{\theta}}}(u_1, u_2, x)$ such that $(U_1, U_2, \hat{\underline{\theta}}) \multimap (X, \underline{\theta}) \multimap (Y_1, Y_2)$ form a Markov chain and $co\{\cdot\}$ stands for convex hull. We emphazise that for the case where the channel estimates $\hat{\underline{\theta}}$ are not available at the transmitter the achievable region still holds, but the distributions in $\mathscr{P}$ must not depend on the channel estimates.*

The marginal distributions of the composite BC channel

$$\widetilde{W}(y_i|u_i, \hat{\theta}_i) = \sum_{(x,u_j)\in\mathscr{X}\times\mathscr{U}_j} \delta\big(x - f(u_1, u_2)\big) P_{U_1U_2}(u_1, u_2)\widetilde{W}(y_i|x, \hat{\theta}_i), \tag{4.9}$$

$j \neq i$ and $\widetilde{W}(y_i|x, \hat{\theta}_i) = \mathbb{E}_{\theta_i|\hat{\theta}_i}\big\{W(y_i|x, \theta_i)\big\}$, where $\mathbb{E}_{\theta_i|\hat{\theta}_i}\{\cdot\}$ denotes the expectation with the conditional pdf $f_{\theta_i|\hat{\theta}_i}(\theta_i|\hat{\theta}_i)$ characterizing the CEE. The achievability proof of this theorem relies on the fact that the composite BC with imperfect channel estimation can be seen as a more noisy BC. Then, by applying Marton's coding

scheme with the statistic of codewords adapted to the composite BC, the averaged error probability of each user grows to zero as the size of these codewords $n \to \infty$.

We remark that for any joint distribution $P_{\underline{\hat{\theta}}}(u_1, u_2, x) \in \mathscr{P}$ the rate pair

$$\begin{aligned} R_1 &= \mathbb{E}_{\underline{\hat{\theta}}}\big\{I\big(P_{U_1}, \widetilde{W}_{\hat{\theta}_1}\big) - I\big(P_{U_2}, P_{U_1|U_2}\big)\big\}, \\ R_2 &= \mathbb{E}_{\underline{\hat{\theta}}}\big\{I\big(P_{U_2}, \widetilde{W}_{\hat{\theta}_2}\big)\big\}, \end{aligned} \tag{4.10}$$

can be achieved by using interference cancellation. This means that user 1 with codewords $U_1$ is considering $U_2$ as the state sequence which is non-causally known at the transmitter. Thus, the channel seen by user 1 is a single-user channel with interference $U_2$ as considered in theorem (4.2.1). In general, the set of achievable rates can be increased by reversing the roles of user 1 and 2, and then the region (4.8) follows [56]. This approach of ordering the users and encoding each user by considering the effect of previous users as non-causally known interference is refereed as *successive encoding* strategy, which was recently showed to achieve the capacity region of the Gaussian MIMO-BC with perfect channel information [64].

Based on the results derived through this section, in the following two sections we consider the capacity of the fading Costa channel and then the capacity region of the Fading MIMO-BC, both with imperfect channel estimation at the receiver(s) and channel estimates available (or not) at the transmitter.

## 4.3 On the Capacity of the Fading Costa Channel with Imperfect Estimation

Throughout this section we consider a memoryless fading Costa channel with Gaussian codebooks. We first derive adequate channel training adapted to each application scenario, assuming ML channel estimation. Then, from Theorem (4.5) we find the optimal DPC scheme and its maximal achievable rates.

### 4.3.1 Fading Costa Channel and Optimal Channel Training

The discrete-time channel at time $t$ is $Y(t) = H(t)\big(X(t) + S(t)\big) + Z(t)$, where $X(t) \in \mathbb{C}$ is the transmitter symbol and $Y(t) \in \mathbb{C}$ is the received symbol. Here,

$H(t) \in \mathbb{C}$ is the complex random channel ($\theta = H$) whose entries are i.i.d. zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variables $f_\theta(\theta) = C\mathcal{N}(0, \sigma_H^2)$. The noise $Z(t) \in \mathbb{C}$ consists of i.i.d. ZMCSCG random variables with variance $\sigma_Z^2$. The channel state $S(t) \in \mathbb{C}$ consists of i.i.d. ZMCSCG random variables with variance $Q$. The quantities $H(t)$, $Z(t)$, $S(t)$ are assumed ergodic and stationary random processes, and the channel matrix $H(t)$ is independent of $S(t)$, $X(t)$ and $Z(t)$. This leads to a stationary and discrete-time memoryless channel $W\big(y|x,s,H\big)$ with pdf

$$W(y|x,s,H) = C\mathcal{N}\big(H(x+s), \sigma_Z^2\big). \tag{4.11}$$

The average symbol energy at the transmitter is constrained to satisfy $\mathbb{E}_X\{X(t)X(t)^\dagger\} \leq \bar{P}$. We next focus on training sequence design for channel estimation.

A standard technique to allow the receiver to estimate the channel matrix consists of transmitting training sequences, i.e., a set of symbols whose location and values are known to the receiver. From a practical point of view, we assume that the channel is constant during the transmission of an entire codeword so that the transmitter, before sending the data $\mathbf{x}$, sends a short training sequence of $N$ symbols $\mathbf{x}_T = (x_{T,1}, \ldots, x_{T,N})$. The average energy per training symbol is $P_T = \frac{1}{N} tr\big(\mathbf{x}_T \mathbf{x}_T^\dagger\big)$. Thus, in practical applications two different scenarios are relevant:

(i) *The channel affects the training sequence only*, i.e. the decoder observes $\mathbf{y}_T = H\mathbf{x}_T + \mathbf{z}_T$, where $\mathbf{z}_T$ is the noise affecting the transmission of training symbols. This scenario arises, e.g., in BCs where the transmitter does not send the sequence $\mathbf{x}_T$ during the training phase. In that case, an optimal training is obtained by sending an arbitrary constant symbol, $x_{T,i} = x_0$ for all $i = 1, \ldots, N$. So that a ML estimate $\hat{\theta} = \hat{H}_{\mathrm{ML}}$ is obtained at the receiver from the observed output. The ML estimate of $H$ is given by (see chapter 2)

$$\widehat{H}_{\mathrm{ML}} = \big(\mathbf{x}_T^\dagger \mathbf{x}_T\big)^{-1} \mathbf{x}_T^\dagger \mathbf{y}_T = H + \mathcal{E}, \tag{4.12}$$

where $\mathcal{E} = \big(\mathbf{x}_T^\dagger \mathbf{x}_T\big)^{-1} \mathbf{x}_T^\dagger \mathbf{z}_T$ is the estimation error with a *noise reduction factor* $\eta = N^{-1}$

$$\sigma_{\mathcal{E}}^2 = \mathrm{SNR}_T^{-1} \quad \text{and} \quad \mathrm{SNR}_T = \frac{P_T}{\eta \sigma_Z^2}. \tag{4.13}$$

(ii) *The channel affects both the training sequence and the state sequence*, which is unknown at the receiver, i.e. the decoder observes $\mathbf{y}_T = H(\mathbf{x}_T + \mathbf{s}_T) + \mathbf{z}_T$, where $\mathbf{s}_T$ is

the state sequence affecting the channel as multiplicative noise. This scenario arises in robust digital watermarking where the channel means an unknown multiplicative attack on the host signal $\mathbf{s}_T$ that is used for training. Here, because the presence of $\mathbf{s}_T$ with average energy per symbol $Q \gg P_T$, the scenario is much more complicated than (i). In other words, as a consequence of this a different method for channel estimation is needed.

We note that the transmitter, before sending the training sequence, perfectly knows the state sequence $\mathbf{s}_T$. Therefore, it can be used for adapting the training sequence to reduce the multiplicative noise at the transmitter. Consider the mean estimator $\widehat{H}_\Delta = \langle \mathbf{y}_T \rangle = H\bar{\nu} + \langle \mathbf{z}_T \rangle$, where $\bar{\nu} = \langle \mathbf{x}_T \rangle + \langle \mathbf{s}_T \rangle$ and $\langle \cdot \rangle$ denotes the mean operator. Obviously, if for some length $N$ the transmitter disposes of enough power $P_T$ to get $\bar{\nu} = 1$ the interference could completely be removed from $\mathbf{y}_T$. Of course, in most of practical cases this is not possible for all realizations of the random sequences $\mathbf{s}_T$, and only part of these sequences can be removed. We can state this more formally as the following optimization problem. Given some arbitrary pair $(\Delta, \gamma)$ with $0 \leq (\Delta, \gamma) < 1$, we find the optimal training sequence $\mathbf{x}_T^*$ and its required length $N^*$ such that

$$\mathbf{x}_T^* = \begin{cases} \text{Minimize} & \|\mathbf{x}_T\|^2/N, \\ \text{Subjet to} & \displaystyle\int_{\{\mathbf{s}_T:\, \bar{\nu}^2 < (1-\Delta)P_T\}} df(\mathbf{s}_T) \leq \gamma, \end{cases} \tag{4.14}$$

where $(1-\Delta)P_T$ represents the power remaining for channel training after removing $\mathbf{s}_T$. This means that for $100 \times (1-\gamma)\%$ of channel estimates the multiplicative interference introduced by $\mathbf{s}_T$ can be removed at the transmitter, elsewhere the training fails. We call $\gamma$ the *failure tolerance level*. Then, the solution of (4.14) is easily found to be $\mathbf{x}_T^*(\mathbf{s}_T) = (x_0^*, \ldots, x_0^*)$ with

$$x_0^*(\mathbf{s}_T) = \begin{cases} \sqrt{(1-\Delta)P_T} - \langle \mathbf{s}_T \rangle & \text{if } \|\mathbf{x}_T^*(\mathbf{s}_T)\|^2 \leq NP_T, \\ 0 & \text{elsewise}, \end{cases} \tag{4.15}$$

and $N^*$ is chosen such that the probability that the training power $P_T$ is not enough to remove the interference be smaller than the *failure tolerance level*, i.e.

$$\int_{\{\mathbf{s}_T:\, \|\mathbf{x}_T^*(\mathbf{s}_T)\|^2 > N^* P_T\}} df(\mathbf{s}_T) \leq \gamma.$$

It follows that $N^*$ can be computed by using the cumulative function of a non-central chi-square of two degrees of freedom $\mathrm{cdf}\big(r; 2, 2N^*P_T(1-\Delta)Q^{-1}\big) = 1-\gamma$ with $r = \frac{2N^*}{Q}P_T$. Actually, the channel estimate can be written as $\widehat{H}_\Delta = H + \mathcal{E}_\Delta$, where $\mathcal{E}_\Delta = \sqrt{\eta_\Delta}\,\langle \mathbf{z}_T \rangle$ is the estimation error with

$$\sigma^2_{\mathcal{E}_\Delta} = \mathrm{SNR}^{-1}_{T,\Delta} \quad \text{and} \quad \mathrm{SNR}_{T,\Delta} = \frac{P_T}{\eta_\Delta \sigma^2_Z}, \tag{4.16}$$

and $\eta_\Delta = \big(N(1-\Delta)\big)^{-1}$ is the *noise reduction factor*. We note that $\eta_\Delta > \eta$, where $\eta$ is the noise reduction factor without the interference sequence present during the phase of training.

From the expression (4.12) and some algebra, we compute the *a posteriori* pdf of $H$ given $\widehat{H}_{\mathrm{ML}}$

$$f_{H|\widehat{H}_{\mathrm{ML}}}(H|\widehat{H}_{\mathrm{ML}}) = C\mathcal{N}(\delta\widehat{H}_{\mathrm{ML}}, \delta\sigma^2_{\mathcal{E}}), \tag{4.17}$$

where $\delta = (\sigma^2_H + \mathrm{SNR}^{-1}_T)^{-1}\sigma^2_H$ and the analogue pdf $f_{H|\widehat{H}_\Delta}(H|\widehat{H}_\Delta)$ follows by substituting $\widehat{H}_\Delta$, $\delta_\Delta = (\sigma^2_H + \mathrm{SNR}^{-1}_{T,\Delta})^{-1}\sigma^2_H$ and $\sigma^2_{\mathcal{E}_\Delta}$ (instead of $\widehat{H}_{\mathrm{ML}}$, $\delta$ and $\sigma^2_{\mathcal{E}}$) in (4.17).

### 4.3.2 Achievable Rates and Optimal DPC Scheme

We now evaluate the test channel (4.11) in the capacity expression (4.4) to derive maximal achievable rates with imperfect channel estimation. This requires to determine the optimum distribution $P_{\underline{\hat{\theta}}}(u, x|s)$ maximizing the capacity. We begin by computing the composite channels $\widetilde{W}(y|x, s, \widehat{H}_{\mathrm{ML}})$ and $\widetilde{W}(y|x, s, \widehat{H}_\Delta)$ associated to each estimation scenario (i) and (ii), respectively. From (4.11) and (4.17) we obtain

$$\widetilde{W}\big(y|x, s, \widehat{H}_{\mathrm{ML}}\big) = C\mathcal{N}\big(\delta\widehat{H}_{\mathrm{ML}}(x+s), \sigma^2_Z + \delta\sigma^2_{\mathcal{E}}(|x|^2 + |s|^2)\big), \tag{4.18}$$

where $\widetilde{W}\big(y|x, s, \widehat{H}_\Delta\big)$ follows by substituting $\widehat{H}_\Delta$, $\delta_\Delta$ and $\sigma^2_{\mathcal{E}_\Delta}$ in (4.18). Actually, we only need to consider the capacity of the composite channel (4.18) associated to the scenario (i), since that corresponding to the scenario (ii) differs only by constant quantities.

A careful examination of the composite channel (4.18) shows that Gaussian codebooks may not necessary achieve the capacity (4.4) (see [105] and [94] for a similar discussions in the context of non-coherent capacity and performance of nearest-neighbor decoding, respectively). The reason is that actually part of the channel noise, due to

the estimation errors, is correlated to the channel input. Since we aim to compute optimal DPC schemes, through this chapter we assume Gaussian inputs, which only leads to a lower bound of the capacity. However, in section 4.5 numerical result show that this assumption does not decrease significatively the capacity (at least for middle and high SNR).

*1) Channel estimates known at the transmitter:* Obviously, if the channel estimates $\widehat{H}_{\mathrm{ML}}$ are known at the transmitter, the optimal Gaussian input distribution is shown to be given by

$$P_{\widehat{H}_{\mathrm{ML}}}(u,x|s) = \begin{cases} P(x) & \text{if} \quad u = x + \alpha^*(\widehat{H}_{\mathrm{ML}})s, \\ 0 & \text{elsewhere,} \end{cases} \tag{4.19}$$

where $P(x) = C\mathcal{N}\big(0,\bar{P}\big)$, and $\bar{P}$ is the power constraint and

$$\alpha^*(\widehat{H}_{\mathrm{ML}}) = \frac{\delta^2|\widehat{H}_{\mathrm{ML}}|^2\bar{P}}{\delta^2|\widehat{H}_{\mathrm{ML}}|^2\bar{P} + \sigma_Z^2 + \delta\sigma_{\mathcal{E}}^2(\bar{P}+Q)}. \tag{4.20}$$

By evaluating the capacity expression (4.4) in the composite channel (4.11) and using the optimal input (4.19), the maximal achievable rate (respect to Gaussian codebooks) denoted $\bar{C}_{11}$ is then

$$\bar{C}_{11} = \mathbb{E}_{\widehat{H}_{\mathrm{ML}}}\Big\{ \log_2\left(1 + \frac{\delta^2|\widehat{H}_{\mathrm{ML}}|^2\bar{P}}{\sigma_Z^2 + \delta\sigma_{\mathcal{E}}^2(\bar{P}+Q)}\right)\Big\}. \tag{4.21}$$

*2) Channel estimates unknown at the transmitter:* The problem in this case is more complicated since the transmitter is not aware to the knowledge of the channel estimate $\widehat{H}_{\mathrm{ML}}$, and consequently the optimal parameter (4.20) cannot be computed. However, assuming Gaussian inputs, which means that $P\big(u,x|s\big)$ is a conditional joint Gaussian pdf. The optimal DPC scheme can be shown to be given by

$$P\big(u,x|s\big) = \begin{cases} P(x) & \text{if} \quad u = x + \alpha s, \\ 0 & \text{elsewhere,} \end{cases} \tag{4.22}$$

where $\alpha \in [0,1]$ is the parameter maximizing the capacity expression in (4.4). Hence, given $\alpha$ the achievable rates can be computed by replacing (4.18) and (4.22) in (4.5). Thus, using some algebra we obtain

$$I_\alpha\big(P_U;\widetilde{W}_{\widehat{H}}\big) = \log_2\left(\frac{(\mathbb{P}+\mathbb{Q}+\mathbb{N})(\mathbb{P}+\alpha^2\mathbb{Q})}{\mathbb{P}\mathbb{Q}(1-\alpha)^2 + \mathbb{N}(\mathbb{P}+\alpha^2\mathbb{Q})}\right), \tag{4.23}$$

$$I_\alpha\big(P_S; P_{U|S}\big) = \log_2\left(\frac{\mathbb{P}+\alpha^2\mathbb{Q}}{\mathbb{P}}\right), \tag{4.24}$$

where $\mathbb{P} = \delta^2|\widehat{H}_{\rm ML}|^2\bar{P}$, $\mathbb{Q} = \delta^2|\widehat{H}_{\rm ML}|^2 Q$ and $\mathbb{N} = \sigma_Z^2 + \delta\sigma_{\tilde{\mathcal{E}}}^2(\bar{P}+Q)$. Given $0 \le \alpha \le 1$, by using (4.23) and (4.24), the capacity expression in (4.4) denoted $\bar{C}_{01}(\alpha)$ that is function of $\alpha$, writes as

$$\bar{C}_{01}(\alpha) = \mathbb{E}_{\widehat{H}_{\rm ML}}\Big\{\log_2\left(\frac{\mathbb{P}(\mathbb{P}+\mathbb{Q}+\mathbb{N})}{\mathbb{P}\mathbb{Q}(1-\alpha)^2+\mathbb{N}(\mathbb{P}+\alpha^2\mathbb{Q})}\right)\Big\}. \tag{4.25}$$

Actually, it remains to find the optimal parameter $\alpha$ maximizing (4.25).

Let us first consider the more intuitive suboptimal choice given by the average over all channel estimates of the optimal parameter $\alpha^*(\widehat{H}_{\rm ML})$ in (4.20), i.e. $\bar{\alpha} = \mathbb{E}_{\widehat{H}_{\rm ML}}\big\{\alpha^*(\widehat{H}_{\rm ML})\big\}$ with $f_{\widehat{H}_{\rm ML}}(\widehat{H}_{\rm ML}) = C\mathcal{N}\big(0, \sigma_H^2+\sigma_{\tilde{\mathcal{E}}}^2\big)$. Thus, it is not difficult to show that

$$\bar{\alpha} = 1 - \frac{1}{\rho}\exp\left(\frac{1}{\rho}\right)E_1\left(\frac{1}{\rho}\right), \quad \text{with } \rho = \frac{\delta\bar{P}\sigma_H^2}{\mathbb{N}}, \tag{4.26}$$

where $E_1(z) = \displaystyle\int_z^\infty t^{-1}\exp(-t)dt$ denotes the exponential integral function. Therefore, the rates in (4.25) can be achieved using the DPC scheme (4.22) with parameter $\bar{\alpha}$ (4.26).

Another possibility is to find directly by maximizing (4.25) the optimal parameter $\alpha^*$. To this end, we observe that

$$\alpha^* = \arg\min_{0\le\alpha\le1}\mathbb{E}_{\widehat{H}_{\rm ML}}\Big\{\log_2\big(\mathbb{P}\mathbb{Q}(1-\alpha)^2+\mathbb{N}(\mathbb{P}+\alpha^2\mathbb{Q})\big)\Big\}. \tag{4.27}$$

Using some algebra the expression (4.27) writes as

$$\alpha^* = \arg\min_{0\le\alpha\le1}\Big\{\log_2(\bar{P}/Q+\alpha^2) + \frac{1}{\log(2)}\exp\left(\frac{\rho(\bar{P}/Q+\alpha^2)}{(1-\alpha)^2}\right)E_1\left(\frac{\rho(\bar{P}/Q+\alpha^2)}{(1-\alpha)^2}\right)\Big\}. \tag{4.28}$$

Unfortunately, there is no explicit solution of (4.28). However, this maximization can be numerically solved to then compute $\bar{C}_{01}(\alpha^*)$. The derived results through this section are also valid for the composite channel corresponding to the channel training of scenario (ii).

## 4.4 On the Capacity of the Fading MIMO-BC with Imperfect Estimation

We first introduce the channel estimation model and review the characterization of the DPC region for the multiuser Fading MIMO-BC with perfect channel information, since this will serve as a basis to derive the corresponding achievable rate region with imperfect channel estimation. Then, from Theorem 4.2.2 we obtain two achievable regions assuming ML or MMSE channel estimation at each receiver and Gaussian codebooks. Here, as well as in previous section, we assume two scenarios: (i) The channel estimates of each receiver are available at the transmitter and (ii) these estimates are unknown at the transmitter.

### 4.4.1 MIMO-BC and Channel Estimation Model

We consider a memoryless Fading MIMO-BC with $m$-users. Assume that the transmitter has $M_T$ antennas and each receiver has $M_R$ ($M_T \geq M_R$) antennas. The channel output at time $t$ is $\mathbf{y}_k(t) = \mathbf{H}_k(t)\mathbf{x}(t) + \mathbf{z}_k(t),\ k = 1, \ldots, K$ where $\mathbf{x}(t) \in \mathbb{C}^{M_T \times 1}$ is the vector of transmitter symbols and $\mathbf{y}_k(t) \in \mathbb{C}^{M_R \times 1}$ is the vector of received symbols at $k$-terminal. Here, $\theta_k = \mathbf{H}_k(t) \in \mathbb{C}^{M_R \times M_T}$ is the complex random matrix of the terminal $k$ whose entries $\big(\mathbf{H}_k(t)\big)_{i,j}$ are independent identically distributed (i.i.d.) zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variables $C\mathcal{N}(0, \sigma^2_{H,k})$. Thus, these matrices are distributed i.i.d. $\mathbf{H}_k(t) \sim f_H(\mathbf{H}_k)$ with pdf

$$C\mathcal{N}\big(0, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma_{H,k}}\big) = \frac{1}{\pi^{M_R M_T} |\mathbf{\Sigma_{H,k}}|^{M_T}} \exp\Big[ - tr\big(\mathbf{H}_k \mathbf{\Sigma_{H,k}}^{-1} \mathbf{H}_k^{\dagger}\big)\Big], \tag{4.29}$$

where $\mathbf{\Sigma_{H,k}}$ is the Hermitian covariance matrix of the columns of $\mathbf{H}_k$ (assumed to be the same for all columns), i.e., $\mathbf{\Sigma_{H,k}} = \sigma^2_{H,k}\mathbb{I}_{M_R}$. The noise vector $\mathbf{z}_k(t) \in \mathbb{C}^{M_R \times 1}$ at $k$-terminal consists of ZMCSCG random vector with covariance matrix $\mathbf{\Sigma_{0,k}} = \sigma^2_{Z,k}\mathbb{I}_{M_R}$. Both $\mathbf{H}_k(t)$ and $\mathbf{z}_k(t)$ are assumed ergodic and stationary random processes, and the channel matrix $\mathbf{H}_k(t)$ is independent of $\mathbf{x}(t)$ and $\mathbf{z}_k(t)$. This leads to a stationary and

discrete-time memoryless BC

$$\mathbf{W}\big(\mathbf{y}_1,\ldots,\mathbf{y}_m|\mathbf{x},\underline{\mathbf{H}}\big)=\prod_{k=1}^{K}\mathbf{W}_k(\mathbf{y}_k|\mathbf{x},\mathbf{H}_k),\quad \text{with } \mathbf{W}_k(\mathbf{y}_k|\mathbf{x},\mathbf{H}_k)=C\mathcal{N}\big(\mathbf{H}_k\mathbf{x},\mathbf{\Sigma_{0,k}}\big), \tag{4.30}$$

where $\underline{\theta}=\underline{\mathbf{H}}=(\mathbf{H}_1,\ldots,\mathbf{H}_K)$. The average symbol energy at the transmitter is constrained to satisfy $tr\big(\mathbb{E}_{\mathbf{X}}(\mathbf{x}(t)\mathbf{x}(t)^\dagger)\big)\leq\bar{P}$.

We assume the standard technique to allow the receivers to estimate the channel matrix based on the use of training sequences (this estimation scenario corresponds to that of (i) explained in section 4.3). This supposes that the channel matrices are quasi-constant during the transmission of an entire codeword so that the channel is *information stable* [106] and the transmitter, before sending the data $\mathbf{X}$, sends a training sequence of $N$ vectors $\mathbf{X}_T=(\mathbf{X}_{T,1},\ldots,\mathbf{X}_{T,N})$. This sequence is affected by the channel matrix $\mathbf{H}_k$, allowing each $k$-receiver to observe separately $\mathbf{Y}_{T,k}=\mathbf{H}_k\mathbf{X}_T+\mathbf{Z}_{T,k}$, where $\mathbf{Z}_{T,k}$ is the noise matrix affecting the transmission of training symbols. The average energy of the training symbols is $\bar{P}_T=\frac{1}{NM_T}tr\big(\mathbf{X}_T\mathbf{X}_T^\dagger\big)$. We focus on ML and MMSE estimation of the channel matrix $\mathbf{H}_k$, for each user $k=1,\ldots,K$, from the observed signals $\mathbf{Y}_{T,k}$ and $\mathbf{X}_T$. Consider the following estimators:

(i) The ML estimator is obtained by minimizing $\|\mathbf{Y}_{T,k}-\mathbf{H}_k\mathbf{X}_T\|^2$ with respect to $\mathbf{H}_k$, yielding

$$\widehat{\mathbf{H}}_{\mathrm{ML},\mathbf{k}}=\mathbf{Y}_{T,k}\mathbf{X}_T^\dagger\big(\mathbf{X}_T\mathbf{X}_T^\dagger\big)^{-1}=\mathbf{H}_k+\mathcal{E}_k, \tag{4.31}$$

where $\mathcal{E}_k=\mathbf{Z}_{T,k}\mathbf{X}_T^\dagger\big(\mathbf{X}_T\mathbf{X}_T^\dagger\big)^{-1}$ denotes the estimation error matrix. Since to estimate the $M_R\times M_T$ channel matrix, we need at least $M_RM_T$ independent measurements so that each symbol time yields $M_R$ samples at the receiver. Therefore, the matrix $\mathbf{X}_T$ must be full rank $M_T$ and thus the matrix $\mathbf{X}_T\mathbf{X}_T^\dagger$ must be nonsingular. This can be satisfied using orthogonal training sequences with $N\geq M_T$, which means that the matrix $\mathbf{X}_T$ has orthogonal rows, such that $\mathbf{X}_T\mathbf{X}_T^\dagger=NP_T\mathbb{I}_{M_T}$. Next, denoting $\big(\mathcal{E}_k\big)_j$ the $j$th column of $\mathcal{E}_k$, we can write $\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}=\mathbb{E}_{\mathcal{E}}\big\{\big(\mathcal{E}_k\big)_j\big(\mathcal{E}_k\big)_j^\dagger\big\}=\mathrm{SNR}_{T,k}^{-1}\mathbb{I}_{M_R}$ with $\mathrm{SNR}_{T,k}=\frac{N\bar{P}_T}{\sigma_{Z,k}^2}$, yielding a white error matrix where the entries of $\mathcal{E}_k$ are i.i.d. ZMCSCG random variables with variance $\mathrm{SNR}_{T,k}^{-1}$. Thus, the conditional pdf of $\widehat{\mathbf{H}}_{\mathrm{ML},k}$ given $\mathbf{H}_k$ is $f_{\widehat{H}_{\mathrm{ML}}|H}(\widehat{\mathbf{H}}_{\mathrm{ML},k}|\mathbf{H}_k)=C\mathcal{N}\big(\mathbf{H}_k,\mathbb{I}_{M_T}\otimes\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\big)$.

(ii) An MMSE estimate of $\mathbf{H}_k$ can be obtained by the linear transformation $\mathbf{Y}_{T,k}\mathbf{T}_{F,k}$, with $\mathbf{T}_{F,k}$ the $N\times M_T$ matrix that minimizes the mean square error

$\mathbb{E}\|\mathbf{Y}_{T,k}\mathbf{T}_{F,k} - \mathbf{H}_k\|^2$. This, together with the definition of the error matrix yields

$$\widehat{\mathbf{H}}_{\mathrm{MMSE},\mathbf{k}} = \widehat{\mathbf{H}}_{\mathrm{ML},\mathbf{k}}\mathbf{A}_{\mathrm{MMSE},\mathbf{k}}, \tag{4.32}$$

$$\mathbf{A}_{\mathrm{MMSE},\mathbf{k}} = \delta_k\,\mathbb{I}_{M_T} \quad \text{with } \delta_k = \frac{\mathrm{SNR}_{T,k}\sigma^2_{H,k}}{\mathrm{SNR}_{T,k}\sigma^2_{H,k}+1}, \tag{4.33}$$

where $\mathbf{A}_{\mathrm{MMSE},\mathbf{k}}$ is an invertible biasing matrix (cf. [62]). In particular, from (4.33), it is easy to show that the conditional pdf $f_{\widehat{H}_{\mathrm{MMSE}}|H}(\widehat{\mathbf{H}}_{\mathrm{MMSE},k}|\mathbf{H}_k) = C\mathcal{N}\big(\delta_k\,\mathbf{H}_k, \mathbb{I}_{M_T} \otimes \delta_k^2\,\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\big)$.

### 4.4.2 Achievable Rates and Optimal DPC scheme

Consider now the problem of finding the capacity region of the multiuser Fading MIMO-BC $\mathbf{W}$ given by (4.30) under CEE. Let us first review, by assuming perfect channel information at both transmitter and each receiver, the optimal design of *successive interference cancellation*, obtained with DPC scheme.

*DPC scheme for BCs*: A *successive encoding* strategy corresponds to the following approach: (i) the users are ordered and (ii) each user is encoded by considering the previous users as non-causally known interference. In the DPC scheme, users codeword $\{\mathbf{x}_k\}_{k=1}^K$ are independent Gaussian vectors $\mathbf{x}_k \sim C\mathcal{N}(0,\mathbf{P}_k)$ with their corresponding covariance matrices $\{\mathbf{P}_k \succeq 0\}_{k=1}^K$ and added up to form the transmitted codeword $\mathbf{x} = \sum_{i=1}^{k-1}\mathbf{x}_i + \mathbf{x}_k + \mathbf{s}_{\Sigma,k+1}^K$ with $\mathbf{s}_{\Sigma,k+1}^K = \sum_{i=k+1}^{K}\mathbf{x}_i$ and $k \in \{1,\ldots,K\}$. The encoder considers the interference $\mathbf{s}_{\Sigma,k+1}^K$, due to users $i > k$, to encode the user codeword $\mathbf{x}_k$. The remaining codewords $(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1})$ are considered by the $k$-th decoder as additional channel noise $\widetilde{\mathbf{z}}_{\Sigma,1}^{k-1} = \sum_{i=1}^{k-1}\mathbf{x}_i$. Then, the $k$-th codeword $\mathbf{x}_k$ is obtained by letting $\mathbf{x}_k = \mathbf{u}_k - \mathbf{F}_k(\mathbf{H}_k)\,\mathbf{s}_{\Sigma,k+1}^K$, where $\mathbf{u}_k$ is the auxiliary random vector chosen according to the message for the $k$-th user and $\{\mathbf{F}_k \succeq 0\}_{k=1}^K$ with $\mathbf{F}_k \in \mathbb{C}^{M_R \times M_R}$ are the optimal precoding matrices. These matrices together with the covariance matrices determine the joint pdf of the auxiliary random vectors $P_{\underline{\mathbf{H}}}(\mathbf{x},\mathbf{u}_1,\ldots,\mathbf{u}_K)$. The optimal matrices are shown to be [69]

$$\mathbf{F}_k^*(\mathbf{H}_k) = \mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\dagger\big(\mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^\dagger + \mathbb{N}_k(\mathbf{H}_k)\big)^{-1}, \tag{4.34}$$

where $\mathbb{N}_k = \mathbf{\Sigma}_{\mathbf{0},\mathbf{k}} + \mathbf{H}_k\mathbf{P}_{\Sigma,1}^{k-1}\mathbf{H}_k^\dagger$ and $\mathbf{P}_{\Sigma,1}^{k-1} = \sum_{i=1}^{k-1}\mathbf{P}_i$.

Let $\pi$ be a permutation defined on the set of index $\{1,\ldots,K\}$, such that $\pi$ determines the encoding order for the DPC scheme, i.e., the message of user $\pi(k)$ is

encoded first while the message of user $\pi(k-1)$ is encoded second and so on. Then, by searching the best choice between all permutations of the encoding order, this coding scheme has been shown in [64] to be optimal (this achieves the capacity) for the Fading MIMO-BC with perfect channel information.

**Theorem 4.4.1 (Capacity region)** *The capacity region $\bar{\mathscr{R}}_{BC}^{(DPC)}$ of the Fading MIMO-BC* $\mathbf{W}$ *with $K$-users and perfect channel information at both transmitter and all receivers is given by*

$$\bar{\mathscr{R}}_{BC}^{(DPC)}(\bar{P}) = co\Big\{ \bigcup_{\substack{\pi,\{\mathbf{P}_k\succeq 0\}\,\forall\, k:\\ tr(\sum_k \mathbf{P}_k)\leq \bar{P}}} \mathcal{A}\big(\pi,\{\mathbf{P}_k\}_{k=1}^K,\mathbf{W}\big)\Big\}, \tag{4.35}$$

*where* $\mathcal{A}\big(\pi,\{\mathbf{P}_k\}_{k=1}^K,\mathbf{W}\big) = \big\{\mathbf{R}\in\mathbb{R}_+^K:\ R_k \leq R_{\pi(k)}^{DPC},\ k=1,\ldots,K\big\}$, *and*

$$R_{\pi(k)}^{DPC} = \mathbb{E}_{\underline{\mathbf{H}}}\Big\{ \log_2 \frac{\Big|\mathbf{H}_{\pi(k)}\Big(\sum_{i=1}^{k}\mathbf{P}_{\pi(i)}\Big)\mathbf{H}_{\pi(k)}^{\dagger} + \mathbf{\Sigma}_{\mathbf{0},\pi(\mathbf{k})}\Big|}{\Big|\mathbf{H}_{\pi(k)}\Big(\sum_{j=1}^{k-1}\mathbf{P}_{\pi(j)}\Big)\mathbf{H}_{\pi(k)}^{\dagger} + \mathbf{\Sigma}_{\mathbf{0},\pi(\mathbf{k})}\Big|}\Big\}. \tag{4.36}$$

This region $\bar{\mathscr{R}}_{\mathrm{BC}}^{(\mathrm{DPC})}$ is the convex hull of the union of all sets $\mathcal{A}\big(\pi,\{\mathbf{P}_k\}_{k=1}^K,\mathbf{W}\big)$ of achievable rates over all permutations $\pi$ and admissible covariance matrices $\{\mathbf{P}_k\succeq 0\}_{k=1}^K$.

We now consider the already described scenarios of channel estimation, for which we study two cases: (i) We assume that all channel estimates are perfectly known at the transmitter side and (ii) all these channel estimates are not available at the transmitter.

*1) Channel estimates known at the transmitter:* We now focus on the capacity of this BC with imperfect channel estimation at the receivers and assuming that the channel estimates are perfectly known at the transmitter. This can be done by evaluating in the achievable region 4.2.2 the marginal channel pdfs of the (more noisy) composite MIMO-BC given by (4.9). Here, we use the simple extension of that region formulated for two users, to the general case of $K$-users. Thus, we obtain the following achievable rate region.

**Theorem 4.4.2 (Achievable rate region)** *An achievable region $\widetilde{\mathscr{R}}_{11}^{(DPC)}$ for the Fading MIMO-BC with ML or MMSE channel estimation and all these estimates $(\widehat{\mathbf{H}}_1,\ldots,\widehat{\mathbf{H}}_K)$ perfectly known at the transmitter, is given by*

$$\widetilde{\mathscr{R}}_{11}^{(DPC)}(\bar{P}) = co\Big\{ \bigcup_{\substack{\pi,\{\mathbf{P}_k\succeq 0\}\ \forall\, k:\\ tr(\sum_k \mathbf{P}_k)\leq \bar{P}}} \mathcal{A}\big(\pi, \{\mathbf{P}_k\}_{k=1}^K, \widetilde{\mathbf{W}}\big)\Big\}, \tag{4.37}$$

*where* $\mathcal{A}\big(\pi, \{\mathbf{P}_k\}_{k=1}^K, \widetilde{\mathbf{W}}\big) = \big\{\mathbf{R}\in\mathbb{R}_+^K:\ R_k \leq \widetilde{R}_{\pi(k)}^{DPC},\ k=1,\ldots,K\big\}$, *and*

$$\widetilde{R}_{\pi(k)}^{DPC} = \mathbb{E}_{\underline{\widehat{\mathbf{H}}}}\Big\{ \log_2 \frac{\Big|\delta_{\pi(k)}^2 \widehat{\mathbf{H}}_{\pi(k)}\Big(\sum_{i=1}^{k}\mathbf{P}_{\pi(i)}\Big)\widehat{\mathbf{H}}_{\pi(k)}^{\dagger} + \widetilde{\mathbf{\Sigma}}_{\mathbf{0},\pi(\mathbf{k})}\Big|}{\Big|\delta_{\pi(k)}^2 \widehat{\mathbf{H}}_{\pi(k)}\Big(\sum_{j=1}^{k-1}\mathbf{P}_{\pi(j)}\Big)\widehat{\mathbf{H}}_{\pi(k)}^{\dagger} + \widetilde{\mathbf{\Sigma}}_{\mathbf{0},\pi(\mathbf{k})}\Big|}\Big\}, \tag{4.38}$$

*with* $\widetilde{\mathbf{\Sigma}}_{\mathbf{0},\pi(\mathbf{k})} = \mathbf{\Sigma}_{\mathbf{0},\pi(\mathbf{k})} + \delta_{\pi(\mathbf{k})}\bar{P}\mathbf{\Sigma}_{\mathcal{E},\pi(\mathbf{k})}$ *and* $\delta_{\pi(k)}$ *defined by* $\delta_{\pi(k)} = \dfrac{SNR_{T,\pi(k)}\sigma_{H,\pi(k)}^2}{SNR_{T,\pi(k)}\sigma_{H,\pi(k)}^2 + 1}$.

***Proof:*** *In order to prove the achievability of this region, we show in Appendix C.2 that the marginal pdf* $\{\widetilde{\mathbf{W}}_k\}_{k=1}^K$ *corresponding to the composite MIMO-BC are*

$$\widetilde{\mathbf{W}}_k(\mathbf{y}_k|\mathbf{x},\widehat{\mathbf{H}}_k) = C\mathcal{N}\big(\delta_k\widehat{\mathbf{H}}_k\mathbf{x}, \mathbf{\Sigma}_{\mathbf{0},\mathbf{k}} + \delta_k\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\|\mathbf{x}\|^2\big), \tag{4.39}$$

*where* $\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}} = SNR_{T,k}^{-1}\mathbb{I}_{M_R}$ *and* $\delta_k$ *is given by* (4.33). *In particular, we show that the expression of the achievable region is independent of the considered type of estimation ML or MMSE, since both estimations lead to the same composite channel* (4.39). *Actually, it remains to evaluate these marginal pdfs in Theorem 4.2.2 to determine the joint distribution* $P_{\underline{\widehat{\mathbf{H}}}}(\mathbf{x},\mathbf{u}_1,\ldots,\mathbf{u}_K)$ *that achieves the boundary points of* (4.8). *We already observe that part of the channel noise in* (4.39) *due to the estimation errors is correlated to the channel input, as well as for the channel considered in section 4.3. This implies that in contrast to the classical case, where perfect channel information is available, here a joint Gaussian density* $P_{\underline{\widehat{\mathbf{H}}}}$ *is not expected to be optimal to characterize the boundary points of this region. However, we focus on the optimal DPC scheme based on Gaussian codebooks, since numerical result show that this assumption does not decrease significatively the capacity. By using DPC coding scheme and some algebra, it is not difficult to show that the optimal precoding matrices are*

$$\left\{\begin{array}{rcl} \widehat{\mathbf{F}}_k^*(\widehat{\mathbf{H}}_k) &=& \delta_k^2\widehat{\mathbf{H}}_k\mathbf{P}_k\widehat{\mathbf{H}}_k^{\dagger}\big(\delta_k^2\widehat{\mathbf{H}}_k\mathbf{P}_k\widehat{\mathbf{H}}_k^{\dagger} + \mathbb{N}_k(\widehat{\mathbf{H}}_k)\big)^{-1}, \\ \mathbf{x}_k &=& \mathbf{u}_k - \widehat{\mathbf{F}}_k^*(\widehat{\mathbf{H}}_k)\,\mathbf{s}_{\Sigma,k+1}^K, \end{array}\right. \tag{4.40}$$

*where* $\mathbb{N}_k = \mathbf{\Sigma}_{\mathbf{0},\mathbf{k}} + \delta_k\bar{P}\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}} + \delta_k^2\widehat{\mathbf{H}}_k\mathbf{P}_{\Sigma,1}^{k-1}\widehat{\mathbf{H}}_k^{\dagger}$ *and* $\widehat{\mathbf{H}}_k$ *is the estimated channel matrix for the k terminal. The definitions of the remaining quantities are equal to those*

*of the DPC scheme with perfect channel information, i.e. users codeword* $\{\mathbf{x}_k\}_{k=1}^K$ *are independent Gaussian vectors* $\mathbf{x}_k \sim C\mathcal{N}(0, \mathbf{P}_k)$ *with corresponding covariance matrices* $\{\mathbf{P}_k \succeq 0\}_{k=1}^K$*, etc.* ■

The sum-rate capacity of the considered MIMO-BC is equal to the maximum sum-rate achievable on the dual uplink with power constraint $\bar{P}$ and is given by

$$C_{\text{BC}}^{\text{sum}}(\bar{P}) = \mathbb{E}_{\underline{\widehat{\mathbf{H}}}}\Big\{ \max_{\substack{\{\mathbf{P}_k \succeq 0\}\,\forall\, k: \\ tr(\sum_k \mathbf{P}_k) \leq \bar{P}}} \Big| \mathbb{I}_{M_R} + \sum_{k=1}^{K} \gamma_k^2 \widehat{\mathbf{H}}_k \mathbf{P}_k \widehat{\mathbf{H}}_k^{\dagger} \Big| \Big\}, \tag{4.41}$$

where $\gamma_k^2 = \dfrac{\text{SNR}_{T,k}\delta_k^2}{\text{SNR}_{T,k}\sigma_{Z,k}^2 + \delta_k \bar{P}}$. Note that (4.41) is a concave maximization, for which efficient numerical algorithms exist (cf. [107]).

*2) Channel estimates unknown at the transmitter:* We now focus on the capacity of the MIMO-BC with imperfect channel estimation at the receivers and assuming that these channel estimates are unknown at the transmitter. The situation here is significantly different of that with perfect channel knowledge (cf. [63]) or when the channel estimates are also availables at the transmitter in Theorem (4.4.2). The reason is that the transmitter cannot use the instantaneous channel estimates to find the optimal precoding matrices needed for the DPC scheme. By using the successive encoding strategy of DPC and Theorem 4.2.2, we first determine an achievable rate region for the composite MIMO-BC, which results of imperfect channel estimation at the receivers. Then, we investigate optimal precoding matrices $\underline{\mathbf{F}} = (\mathbf{F}_1, \ldots, \mathbf{F}_K)$, inspired by the optimal solution (4.40) when the estimates are availables at the transmitter.

**Theorem 4.4.3 (Achievable rate region)** *An achievable region* $\widetilde{\mathscr{R}}_{01}^{(DPC)}$ *for the Fading MIMO-BC with ML or MMSE channel estimation, and assuming that the channel estimates are not available at the transmitter, is given by*

$$\widetilde{\mathscr{R}}_{01}^{(DPC)}(\bar{P}, \underline{\mathbf{F}}) = co\Big\{ \bigcup_{\substack{\pi, \{\mathbf{P}_k \succeq 0\}\,\forall\, k: \\ tr(\sum_k \mathbf{P}_k) \leq \bar{P}}} \mathcal{B}\big(\pi, \{\mathbf{P}_k\}_{k=1}^K, \widetilde{\mathbf{W}}, \underline{\mathbf{F}}\big) \Big\}, \tag{4.42}$$

$$\mathcal{B}\big(\pi, \{\mathbf{P}_k\}_{k=1}^K, \widetilde{\mathbf{W}}, \underline{\mathbf{F}}\big) = \big\{ \mathbf{R} \in \mathbb{R}_+^K : \; R_k \leq \widetilde{R}_{\pi(k)}^{\widetilde{DPC}}(F_{\pi(k)}),\; k = 1, \ldots, K \big\},\ \textit{and}$$

$$\widetilde{R}^{\widetilde{DPC}}_{\pi(k)}(\mathbf{F}_{\pi(k)}) = \mathbb{E}_{\underline{\widehat{\mathbf{H}}}}\Big\{ \log_2 \frac{|\mathbb{P}_{\pi(k)}||\mathbb{P}_{\pi(k)} + \mathbb{Q}_{\pi(k)} + \mathbb{N}_{\pi(k)}|}{\begin{vmatrix} \mathbb{P}_{\pi(k)} + \mathbf{F}_{\pi(k)}\mathbb{Q}_{\pi(k)}\mathbf{F}^{\dagger}_{\pi(k)} & \mathbb{P}_{\pi(k)} + \mathbf{F}_{\pi(k)}\mathbb{Q}_{\pi(k)} \\ \mathbb{P}_{\pi(k)} + \mathbb{Q}_{\pi(k)}\mathbf{F}^{\dagger}_{\pi(k)} & \mathbb{P}_{\pi(k)} + \mathbb{Q}_{\pi(k)} + \mathbb{N}_{\pi(k)} \end{vmatrix}} \Big\}, \quad (4.43)$$

$$\begin{aligned} \mathbb{P}_{\pi(k)} &= \delta^2_{\pi(k)}\widehat{\mathbf{H}}_{\pi(k)}\mathbf{P}_{\pi(k)}\widehat{\mathbf{H}}^{\dagger}_{\pi(k)}, \\ \mathbb{Q}_{\pi(k)} &= \delta^2_{\pi(k)}\widehat{\mathbf{H}}_{\pi(k)}\mathbf{P}^{m}_{\Sigma,\pi(k)+1}\widehat{\mathbf{H}}^{\dagger}_{\pi(k)}, \qquad \mathbf{P}^{k}_{\Sigma,j} = \sum_{j=i}^{k}\mathbf{P}_j, \\ \mathbb{N}_{\pi(k)} &= \mathbf{\Sigma}_{\mathbf{0},\pi(\mathbf{k})} + \delta_{\pi(k)}\bar{P}\mathbf{\Sigma}_{\mathcal{E},\pi(\mathbf{k})} + \delta^2_{\pi(k)}\widehat{\mathbf{H}}_{\pi(k)}\mathbf{P}^{\pi(k)-1}_{\Sigma,1}\widehat{\mathbf{H}}^{\dagger}_{\pi(k)}. \end{aligned}$$

The derivation of this achievable region follows from Theorem 4.2.2 by evaluating (4.8) in the composite MIMO-BC (4.39), the details are presented in appendix C.3. Actually, it remains to find the optimal precoding matrices $\underline{\mathbf{F}} = (\mathbf{F}_1, \ldots, \mathbf{F}_K)$ maximizing the rates in (4.43). We emphasize that this maximization must be taken over all matrices not depending on the channel estimates $\underline{\widehat{\mathbf{H}}}$ (these are assumed to be unknown at the transmitter).

Consider first the more intuitive suboptimal choice for $\mathbf{F}_k$, $k = 1, \ldots, K$, that consists in taking the average over all channel estimates of the optimal matrices (4.40) with channel estimates availables at the transmitter. This amounts to the following computation

$$\bar{\mathbf{F}}_k = \mathbb{E}_{\underline{\widehat{\mathbf{H}}}}\big\{\mathbb{P}_k(\widehat{\mathbf{H}}_k)\big(\mathbb{P}_k(\widehat{\mathbf{H}}_k) + \mathbb{N}_k(\widehat{\mathbf{H}}_k)\big)^{-1}\big\}, \quad (4.44)$$

where the channel estimates follows as $\widehat{\mathbf{H}}_k \sim f_{\widehat{\mathbf{H}}}(\widehat{\mathbf{H}}_k) = \mathcal{CN}\big(0, \mathbb{I}_{M_T} \otimes \sigma^2_{\hat{H},k}\mathbb{I}_{M_R}\big)$ with $\sigma^2_{\hat{H},k} = \sigma^2_{\mathcal{E},k} + \sigma^2_{H,k}$. By using some algebra, in appendix C.4 we prove the following statement.

**Lemma 4.4.1** *The average over all channel estimates of the optimal precoding matrices in* (4.44) *is given by*

$$\bar{\mathbf{F}}_k = \mathbb{I}_{M_R}\frac{1}{M_R}\left[1 - \rho_k^{n+1}\exp(\rho_k)\Gamma(-n, \rho_k)\right], \quad (4.45)$$

*where* $\rho_k = \dfrac{M_T}{M_R}\dfrac{tr(\mathbf{\Sigma}_{\mathbf{0},\mathbf{k}} + \delta_k\bar{P}\mathbf{\Sigma}_{\mathcal{E},k})}{\delta^2_k\sigma^2_{\hat{H},k}tr(\mathbf{P}^k_{\Sigma,1})}$ *and* $n = M_T M_R - 1$ *with* $n \in \mathbb{N}_+$,

$$\Gamma(-n, t) = \frac{(-1)^n}{n!}\Big[\Gamma(0, t) - \exp(-t)\sum_{i=0}^{n-1}(-1)^i\frac{i!}{t^{i+1}}\Big],$$

*and* $\Gamma(0,t) = \int_t^{+\infty} u^{-1}\exp(-u)du$ *denotes the exponential integral function.*

The other (obviously optimal, but solvable numerically only) possibility is to find directly the optimal matrix $\mathbf{F}_k^*$ maximizing the rates in (4.43). We observe that these matrices can be found as follows

$$\mathbf{F}_k^* = \arg\min_{\mathbf{F}\succeq 0} \mathbb{E}_{\underline{\widehat{\mathbf{H}}}}\Big\{ \log_2 \begin{vmatrix} \mathbb{P}_k + \mathbf{F}\mathbb{Q}_k\mathbf{F}^\dagger & \mathbb{P}_k + \mathbf{F}\mathbb{Q}_k \\ \mathbb{P}_k + \mathbb{Q}_k\mathbf{F}^\dagger & \mathbb{P}_k + \mathbb{Q}_k + \mathbb{N}_k \end{vmatrix} \Big\}. \tag{4.46}$$

To solve expression (4.46), we note that the transmitter does not have access to the channel estimates and consequently no spatial power optimization can be implemented. Therefore, the solution is shown to be given by $\mathbf{F}_k^* = \alpha_k^* \mathbb{I}_{M_R}$ and the covariance matrices $\{\mathbf{P}_k = \mathbb{I}_{M_T} P_k\}_{k=1}^K$ such that $\sum_{k=1}^K P_k = M_T^{-1}\bar{P}$ (cf. [66]), where by using elementary algebra it is not difficult to show that

$$\alpha_k^* = \arg\min_{0\le\alpha\le 1} \Big\{ \lambda(\alpha)\Big[ \exp\left(\frac{\beta_{-,k}(\alpha)}{4\alpha}\right)\Gamma\left(0,\frac{\beta_{-,k}(\alpha)}{4\alpha}\right) - \exp\left(\frac{\beta_{+,k}(\alpha)}{4\alpha}\right)\Gamma\left(0,\frac{\beta_{+,k}(\alpha)}{4\alpha}\right)\Big]\Big\}, \tag{4.47}$$

with constants

$$\begin{aligned} &\lambda(\alpha) = \frac{A_{0,k}A_{1,k}^{-1}}{A_{3,k}\sqrt{B_k^2 - 4\alpha}}, \\ &\beta_{\pm,k}(\alpha) = B_k \pm \sqrt{B_k^2 - 4\alpha} \quad \text{and} \quad B_k = \frac{A_{0,k}}{A_{1,k}A_{3,k}}\left(\frac{2A_{1,k}A_{2,k}}{A_{0,k}} - 1\right), \\ &A_{0,k} = \delta_k^4(P_k + P_{\Sigma,k+1}^m\alpha)^2 \quad \text{and} \quad A_{1,k} = \delta_k^2(P_k + P_{\Sigma,k+1}^m\alpha^2), \\ &A_{2,k} = \delta_k^2\bar{P} \quad \text{and} \quad A_{3,k} = \sigma_{Z,k}^2 + \delta_k\sigma_{\mathcal{E},k}^2\bar{P}. \end{aligned} \tag{4.48}$$

Unfortunately, (4.47) does not lead to an explicit solution for $\alpha_k^*$. However, this maximization can be numerically solved for each $k = 1,\ldots,K$, to compute (4.43) and then $\bar{\mathscr{R}}_{01}(\bar{P}, \underline{\mathbf{F}}^*)$. Both solutions were tested, and we observed that the achievable rates with $\underline{\bar{\mathbf{F}}}$ are very close to those provided by the optimal solution $\underline{\mathbf{F}}^*$. As a result, we have chosen in the simulations below to use the mean parameter for designing the "close to optimal" DPC scheme.

## 4.5 Simulation Results and Discussions

In this section, numerical results are presented based on Monte Carlo simulations. We first illustrate achievable rates for the Fading Costa channel according to the derived results in section 4.3. Then, using results in section 4.4, we illustrate achievable

rates of a realistic downlink wireless communication scenario involving a two-users ($m = 2$) Fading MIMO Broadcast Channel.

### 4.5.1 Achievable rates of the Fading Costa Channel

(i) *Channel training and optimal DPC design:* We start by considering the channel training scenario described in 4.3 that arises in robust watermarking applications when the channel coefficient during the training phase affects both the training sequence and the state sequence. Fig. 4.1 shows the noise reduction factor $\eta_\Delta$ versus the training sequence length $N$, for various failure tolerance levels $\gamma \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. The power of the state sequence $Q$ is 20 dB larger than that corresponding to the training sequence $P_T$. Let us suppose that, e.g., we want to get an estimation error 10 times less than the channel noise (i.e. $\eta_\Delta = 10^{-1}$), with a failure tolerance level $\gamma = 10^{-2}$. From Fig. 4.1 we can observe that the required training length is $N = 500$. Whereas to get equal performances, when the state sequence is not present during the training phase, would only require $N = 10$.

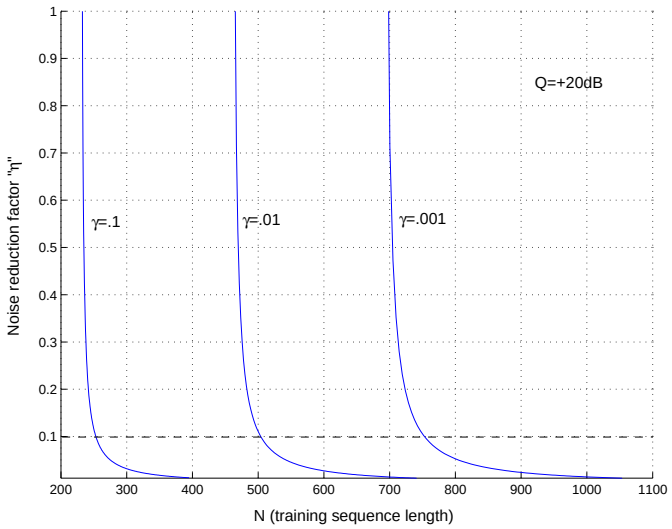



Figure 4.1: Noise reduction factor $\eta_\Delta$ versus the training sequence lengths $N$, for various probabilities $\gamma$.

Fig. 4.2 shows both the mean parameter $\bar{\alpha}$ (4.26) and the optimal parameter $\alpha^*$ (4.47) versus the signal-to-noise ratio, for various training sequence lengths $N$. The state sequence power $Q$ is +20 dB larger than that of the channel input $\bar{P}$, and the training power is $P_T = \bar{P}$. We can observe that both parameters are relatively close

for many SNR values. Furthermore, even in the SNR ranges where the values seem to be quite different, we have observed that the achievable rates with $\bar{\alpha}$ are very close to those provided by the optimal solution $\alpha^*$. Therefore, we can conclude that the mean parameter can be used to design the optimal DPC scheme.

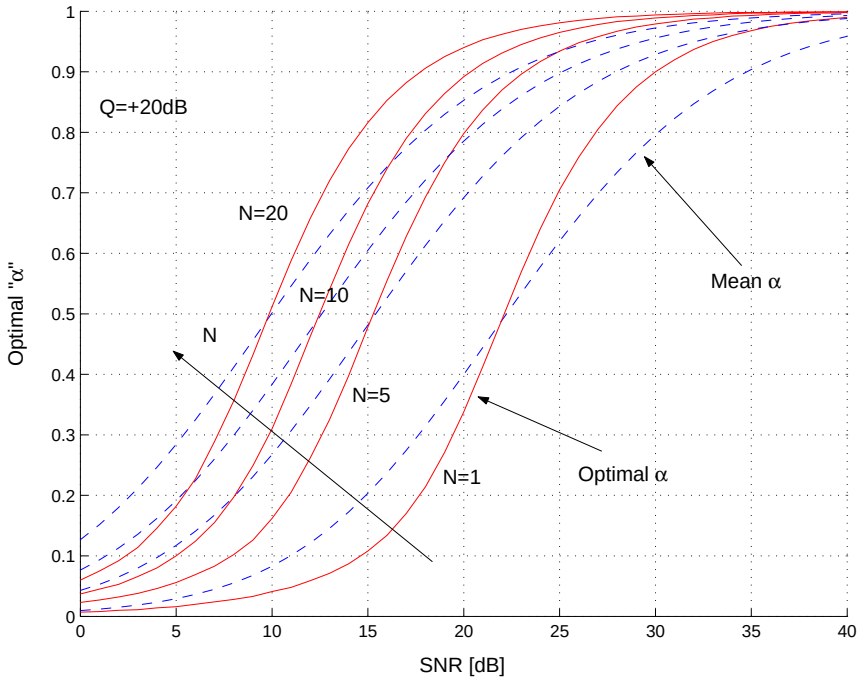



Figure 4.2: Optimal parameter $\alpha^*$ (solid lines) versus the SNR, for various training sequence lengths $N$. Dashed lines show mean alpha $\bar{\alpha}$.

(ii) *Achievable rates:* Fig. 4.3 shows achievable rates (4.25) (in bits per channel use) with channel estimates unknown at the transmitter versus the SNR, for various training sequence lengths $N \in \{1, 10, 20\}$ (dashed line). For comparison we also show achievable rates (4.21) with channel estimates known at the transmitter (danshed-dot line) and with perfect channel knowledge at both transmitter and receiver (solid line). It is seen that the average rates tend to increase rather fast with the amount of training. For example, to achieve 2 bits with channel estimates unknown at the transmitter. Observe that a scheme with estimated channel and $N = 10$ requires 18 dB, i.e., 11 dB more than with perfect channel information. Whereas, if the training length is further reduced to $N = 1$, this gap increases to 27 dB. On the other hand, when the channel estimates are known at the transmitter, the SNR required for 2 bits is only 1 dB less than the case with channel estimates unknown. This rate gain is slightly smaller, and consequently we can conclude that for the fading Costa channel

with a single transmitter and receiver antenna, the knowledge of the channel estimates at the transmitter is not really necessary with the proposed DPC scheme.

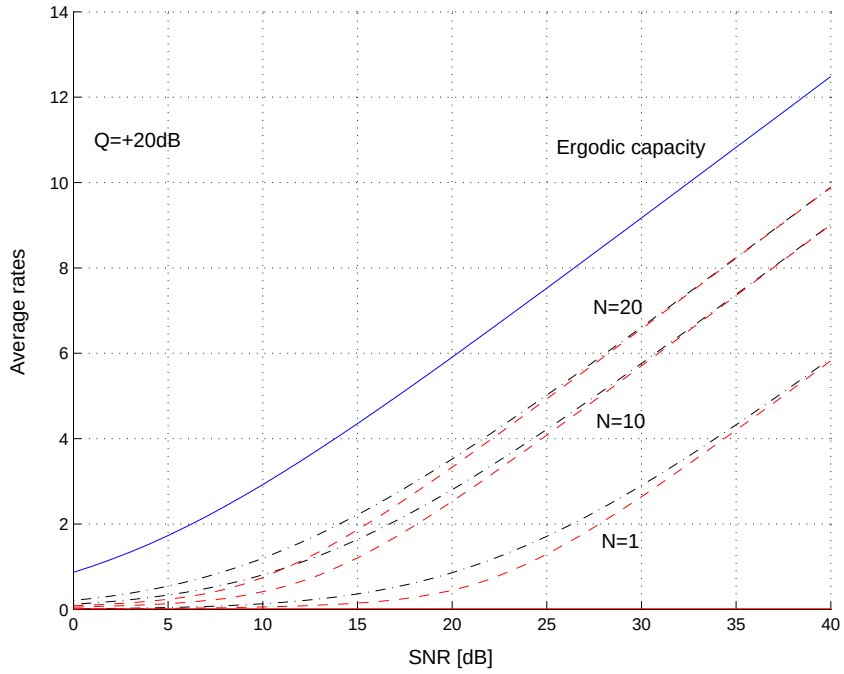



Figure 4.3: Achievable rates with channel estimates known at the transmitter (dashed-dot lines) versus the SNR, for various training sequence lengths $N$. Dashed lines suppose channel estimates unknown at the transmitter. Solid line shows the capacity with the channel known at both the transmitter and the receiver.

Finally, we study the impact of the power state sequence on the achievable rates. Fig. 4.4 shows similar plots for different values of $+Q \in \{+20, +30, +40\}$, i.e., $Q$ is times larger (in dB) than the channel input power $\bar{P}$, and training sequence length is $N = 10$. We can observe that the performance are very sensitive to the power $Q$. This is because with imperfect channel estimation the capacity still depends on $Q$ (cf. (4.25)), while with perfect channel information the state sequence is canceled at the transmitter independent of the power $Q$.

### 4.5.2 Achievable Rates of the Fading MIMO-BC

We first consider a base station (the transmitter) with three antennas ($M_T = 3$) and mobiles (the receivers) with two antennas ($M_R = 2$). We show the average of achievable rates over all channel estimates, for different amount of training $N$. For comparison, we also show the time-division rate region where the transmitter sends information to only a single user at a time and the ergodic capacity (4.35) with perfect

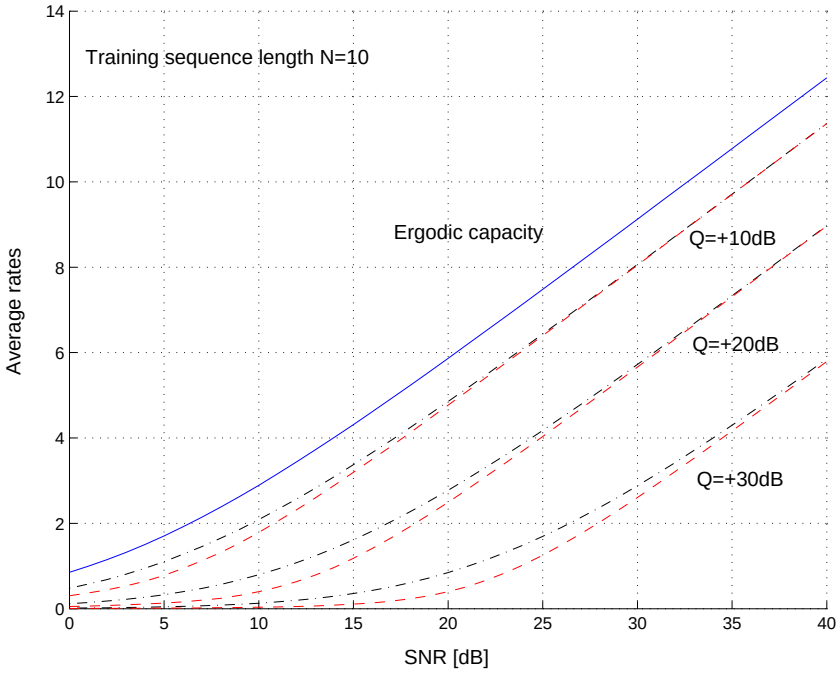



Figure 4.4: Similar plots for different power values of the state sequence $Q$.

channel knowledge. For numerical results, we assume that the transmitter is subject to a short-term power constraint, so that the transmitter must satisfy power constraint $\bar{P}$ for every fading state. This implies that there can be no adaptive power allocation over time, only spatial power allocation if channel estimates available at the transmitter is used. Suppose very different signal-to-noise ratios $\text{SNR}_1 = 0\text{dB}$ and $\text{SNR}_2 = 10\text{dB}$, and equal fading distributions $\sigma^2_{H,1} = \sigma^2_{H,2} = 1$. Here, the training assumes same channel SNR than transmission, i.e., $\bar{P}_T = \bar{P}$. This is specially important to avoid noise saturation over the achievable rates. We assume the two scenarios studied in section 4.4: (i) The channel estimates of each receiver are available at the transmitter and (ii) these estimates are unknown at the transmitter.

(i) In this case the channel estimates are available at the transmitter and consequently spatial power allocation is possible. However, the expressions (4.36) and (4.38) are not concave functions of the covariance matrices, and thus filing these region borders is numerically difficult. Instead, we consider a simplified power allocation scheme that maximizes the sum-rate capacity and achieves average rates close to optimal performances. By assuming power-sharing between the two users and a given encoding order, i.e. each user has power $\bar{P}_k$ with $tr(\mathbf{P_k}) \leq \bar{P}_k$ such that

$\bar{P} = \alpha\bar{P}_1 + (1-\alpha)\bar{P}_2$, we can obtain optimal covariance matrices $\{\mathbf{P_1}, \mathbf{P_2}\}$ maximizing the sum-rate capacity. Then, we swap the encoder order, which allows us to explore both possibilities, and choice the best one. This yields to the especializated algorithm with individual power constraints developed in [107]. We then investigate the performance in terms of the average of achievable sum-rate versus the amount of training, for different number of transmit antennas.

Fig. 4.5 shows the average of the achievable region (in bits per channel use) with perfect CSI (Ergodic capacity) and with estimated CSI (i.e. ML or MMSE channel estimation), for different amount of training $N = \{4, 10\}$. Observe that the achievable rates using imperfect channel estimation are still quite large irrespective of the small training sequence length $N = 4$ (dashed line), i.e. 1.4 bits less comparing to the capacity with perfect CSI (solid line). In comparison, only 0.6 bits less are expected with $N = 10$. Suppose now that user-2 is sending information at a rate $R_2 = 4$ bits, a relevant question to ask is the following: In presence of imperfect channel estimation with a given amount of training, how large performance gains can be achieved for user-1 by using the DPC scheme adapted to the channel estimation errors instead of the classical DPC substituting the unknown channel matrices by its corresponding estimates (dashed-dot lines) ? We note that this gain is about $+0.2$ bits with $N = 10$ and $+0.3$ bits with $N = 4$.

Fig. 4.6 shows the average performance in terms of achievable sum-rate for different training sequence lengths $N \in \{2, 100\}$ and different number of transmit antennas $M_T \in \{2, 4, 8, 16, 32\}$ with two receiver antennas $M_R = 2$. This allows to evaluate the amount of training necessary to achieve a certain mean sum-rate for a given number of transmit antennas. It is seen that a small increase in the training sequence length can cause significant improvement in the mean sum-rate. We observe that for large training sequence lengths and smaller number of transmit antennas, in this case $M_T \leq 8$, the mean sum-rate has close performance to the sum-rate capacity. However, increasing the number of transmit antennas requires very large amount of training, with a very slow convergence to its performance limits.

(ii) Consider now that the base station and the mobiles have a single antenna ($M_T = M_R = 1$). We show the average (over all channel estimates) of achievable rates (4.43) with channel estimates unknown at the transmitter and using the mean

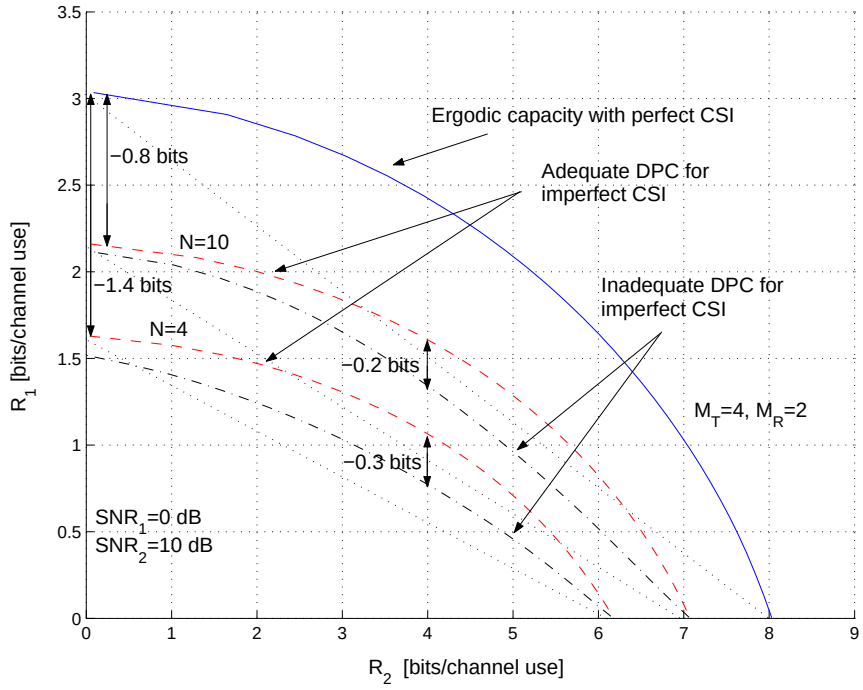



Figure 4.5: Average (over all channel estimates) of achievable rate region with ML or MMSE channel estimation at both transmitter and all receivers (dashed curves), for $N = \{4, 10\}$. Dashed-dot curves show similar plots using the classical DPC substituting unknown channel matrices by its corresponding estimates.

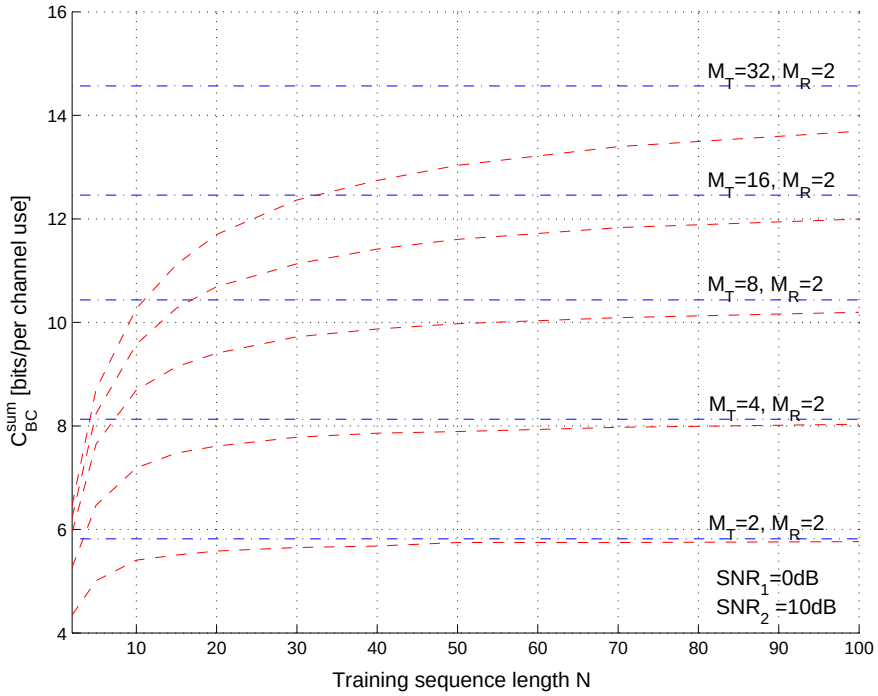



Figure 4.6: Average of sum-rate capacity with ML or MMSE channel estimation (dashed lines) versus the amount of training, for different number of transmit antennas. Dashed-dot lines show average of sum-rate capacity with perfect CSI.

parameter (4.45) in the precoding matrices, for different amount of training $N$. For comparison, we also show similar plots with channel estimates known at the transmitter, the time-division rate region and the ergodic capacity under perfect channel information. Then, we investigate these achievable rates by increasing the number of transmitter and receiver antennas. For which we assume a transmitter with four antennas ($M_T = 4$) and receivers with two antennas ($M_R = 2$).

Fig. 4.7 shows the average of the achievable rates with both: channel estimates available at both transmitter and all receivers (Theorem (4.4.2)) and with channel estimates only available at the receivers (Theorem (4.4.3)), for different amount of training $N = \{5, 20\}$. Observe that the achievable rates with channel estimation are still quite large irrespective of the small training sequence length $N = 5$ (dashed and danshed-dot lines), i.e. 0.2 bits less comparing to the capacity with perfect channel information (solid line). Suppose now that user-2 needs to send information at a rate $R_2 = 1.5$ bits. We want to determine, how large performance gains can be achieved for user-1, when the channel estimates are not availables at the transmitter. We investigate this by observing the gain for the first user when the second user is transmitting at 1.5 bits. Note that this gain is $-0.1$ bits (with $N = 20$) and $-0.22$ bits (with $N = 5$) less compared to the case of perfect channel information. On the other hand, only 0.04 bits more are expected when the transmitter knows the channel estimates. This rate gain is slightly smaller, and consequently we can conclude that the knowledge of the channel estimates at the transmitter is not really necessary with the proposed DPC scheme.

Fig. 4.8 shows similar plots with $M_T = 4$ and $M_R = 2$ and $N = \{5, 40\}$. In this multiple antenna scenario, without channel information at the transmitter, there can be no adaptive spatial power allocation. However, at equal power, it is seen that a small increase in the number of transmitter antennas can cause significant improvement, comparing with the single antenna case. We recall that the short-term power constraint is averaged over all transmitter antennas, so that this power constraint is independent of the number of transmitter antennas. Consider now that user-2 needs to send information at a rate $R_2 = 3$ bits. We observe that, with channel estimates available at the transmitter, significant gains can be achieved compared to the case where the estimates are unknown at the transmitter (approximately 1.4 bits

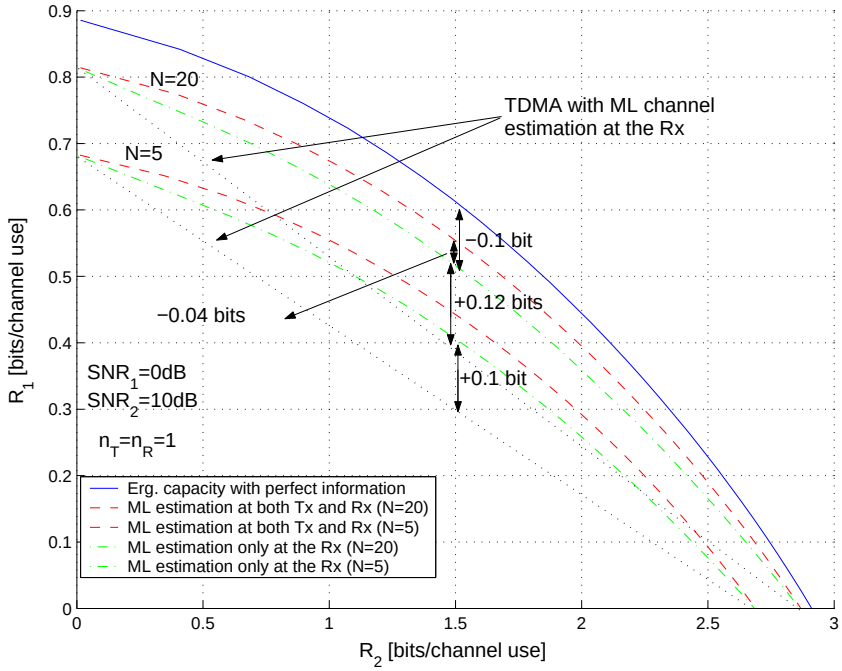



Figure 4.7: Average of achievable rate region with a single antenna BC ($M_T = M_R = 1$) and channel estimates known at the transmitter (dashed lines) versus the SNR, for training sequence lengths $N = \{5, 20\}$. Dashed-dot lines assume channel estimates unknown at the transmitter. Solid line shows the capacity with perfect channel knowledge.

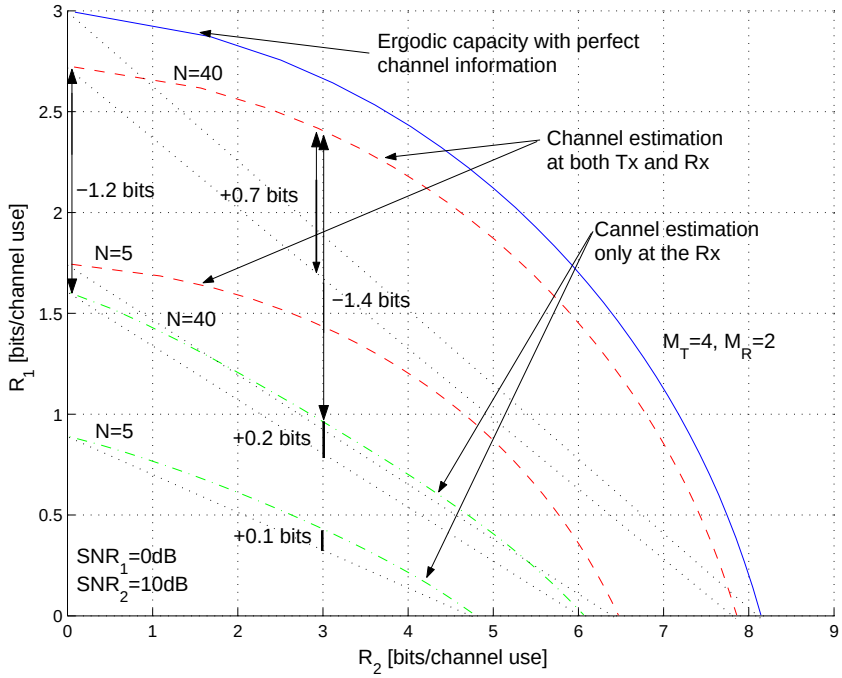



Figure 4.8: Similar plots of the achievable rate region with $N = \{5, 40\}$, four transmitter antennas ($M_T = 4$) and two receiver antennas ($M_R = 2$).

with $N = 40$). Whereas, a multiple antenna BC achieves rates close to those of the time-division multiple access (dot line). The gain, by using DPC instead of TDMA, is reduced to only 0.2 bits with $N = 40$, while not signicative gain is observed for $N = 5$ (only 0.1 bits). Note that this gain is equal to that obtained with a single antenna. Thus, for a MIMO-BC, taking a real benefit from a large number of transmit antennas would require an instantaneous knowledge of channel estimates at the transmitter. If it is not the case, TDMA provides similar performances to MIMO Broadcast channels.

## 4.6 Summary

In this chapter we studied the problem of communicating reliably over imperfectly known channels with channel states non-causally known at the transmitter. The general framework considered through a novel notion of reliable communication under imperfect channel knowledge, enables us to easily extend existing capacity expressions that assume perfect channel knowledge to the more realistic case with imperfect channel estimation. The key feature for this purpose is our notion of reliable communication that transforms the mismatched scenario given by the CEE, into a composite (more noisy) state-dependent channel. We assumed two scenarios: (i) The receiver only has access to noisy estimates of the channel and these estimates are perfectly known at the transmitter and (ii) there is no channel information available at the transmitter and imperfect information is available at the receiver. In this scenario, we proposed to characterize the information-theoretic limits based on the average of the transmission error probability over all CEE. This basically means that the transceiver does not require small instantaneous transmission error probabilities, but rather the average over all channel estimation errors must be arbitrary small. Inspired by a similar approach, we consider a natural extension of the Marton's region for arbitrary broadcast channels, obtaining explicit expressions for general DMCs of the corresponding maximal achievable rates.

We next used the capacity expression to obtain achievable rates for the fading Costa channel with ML channel estimation and Gaussian inputs. We also studied optimal training design adapted to each application scenario, e.g., BCs or robust watermarking. The somewhat unexpected result is that, while it is well-known that

DPC for such class of channel requires perfect channel knowledge at both transmitter and receiver, without channel information at the transmitter, significant gains compared to TDMA can be still achieved by using the proposed (adapted to the channel estimation errors) DPC scheme. Further numerical results in the context of uncorrelated fading show that, under the assumption of imperfect channel information at the receiver, the benefit of channel estimates known at the transmitter does not lead to large rate increases.

In a similar manner, using the achievable region for general BCs, we characterized an achievable rate region for the Fading MIMO-BC, assuming ML or MMSE channel estimation. We considered both scenarios: (i) The transmitter and all receivers only know a noisy estimate of the channel matrices and (ii) the more complicate case where there is no channel information available at the transmitter. We derive the optimal DPC scheme under the assumption of Gaussian inputs, for which we observed the expected result that both estimators lead to the same capacity region. The "close to optimal" DPC scheme in scenario (ii), without knowledge of channel estimates, follows as the average over all channel estimates of the optimal DPC scheme implemented for the case where the transmitter knows the estimates.

Our results are useful to assess the amount of training data to achieve target rates. Interesting is that a BC with a single transmitter and receiver antenna and no channel information at the transmitter can still achieve significant gains compared to TDMA using the proposed DPC scheme. Furthermore, in this case the benefit of channel estimates known at the transmitter does not lead to large rate increases. However, we also showed that, for multiple antenna BCs, in order to achieve large gain rates the transmitter requires the knowledge of all channel estimates, i.e., some feedback channel (perhaps rate-limited) must go from the receivers to the transmitter, conveying these channel estimates. Clearly, while it is well-known that for systems with many users significant gains can be achieved by adding base station antennas, under imperfect channel estimation, benefiting of a large number antennas requires very large amount of training. Consequently, in practice depending on the degree of accuracy channel estimation, this benefit may not hold.

# Chapter 5

# Broadcast-Aware and MAC-Aware Coding Strategies for Multiple User Information Embedding

Multiple user information embedding is concerned with embedding several messages into the same host signal. This chapter presents several implementable "Dirty-paper coding" (DPC) based schemes for multiple user information embedding, through emphasizing their tight relationship with conventional multiple user information theory.

We first show that depending on the targeted application and on whether the different messages are asked to have different robustness and transparency requirements or not, multiple user information embedding parallels one of the well-known multi-user channels with state information available at the transmitter. The focus is on the Gaussian Broadcast Channel (BC) and the Gaussian Multiple Access Channel (MAC). For each of these channels, two practically feasible transmission schemes are compared. The first approach consists in a straightforward- rather intuitive- superimposition of DPC schemes. The second consists in a joint design of these DPC schemes, which is based on the ideal DPC for the corresponding channel.

These results extend on one side the practical implementations QIM, DC-QIM and SCS from the single user case to the multiple user one, and on another side provide a clear evaluation of the improvements brought by joint designs in practical

situations. After presenting the key features of the joint design within the context of structured scalar codebooks, we broaden our view to discuss the framework of more general lattice-based (vector) codebooks and show that the gap to full performance can be bridged up using finite dimensional lattice codebooks. Performance evaluations, including Bit Error Rates and achievable rate region curves are provided for both methods, illustrating the improvements brought by a joint design.

## 5.1 Introduction

Research on information embedding has gained considerable attention during the last years, mainly due to its potential application in multimedia security. Digital watermarking and data hiding techniques, which are a major branch of information embedding, refer to the situation of embedding information carrying-signals called *watermarks* into another signal, generally stronger, called *cover* or *host* signal. The cover signal is any multimedia signal. It can be either image, audio or video. The embedding must not introduce perceptible distortions to the host, and the watermark should survive common channel degradations. These two requirements are often called *transparency requirement* and *robustness requirement*, respectively. Being conflicting, these two requirements, together with the interference stemming from the host signal itself, have for long time limited the use of digital watermarking to applications where little information (payload) has to be embedded. These include copyright protection [71], for example, where the transmission of just one bit of information, expected to be detectable with very low probability of false alarm, is sufficient to serve as an evidence of copyright. In these applications, the watermark is in general a pseudo-noise sequence obtained by means of conventional Spread-Spectrum Modulations (SSM) techniques. SSM techniques do not allow the encoder to exploit knowledge of the host signal in the design of the transmitted codewords and are consequently interference limited by construction.

Information embedding can also be viewed as power-limited communication over a "super"-channel with state (or side) information non-causally known to the transmitter [108, 109]. The channel input is the watermark and the available state information is the cover or host signal itself. An achievable rate, for a watermarking system,

consists in any rate of payload that can be successfully decodable. The capacity, or more precisely the data hiding capacity, is the supremum of all achievable rates. Based on this equivalence many host-interference rejecting schemes have been proposed [108, 110] in this still emerging field. It has then become possible to embed large amount of information while at the same time satisfying the two requirements above.

The most relevant work in this area is the initial Costa's "Writing on Dirty Paper" [111], commonly known as "Costa's problem". Costa was the first to examine the Gaussian dirty chapter problem. He obtained the remarkable result that an additive Gaussian interference which is non-causally known only at the encoder incurs no loss of capacity, relative to the Gaussian interference-free channel. The theoretical proof of "Costa's problem" is based on an optimal random binning argument for i.i.d. Gaussian codebook. This technique had been proved to be optimal for more general problems in "coding for channels with random parameters" studied in [112] and [113]. Binning consists in a probabilistic construction of codewords. However, this probabilistic construction is convenient only for theoretical analysis, not for practical coding applications. The schemes proposed by Chen and Wornell [108] and Eggers et al. [110], in the context of information embedding, adhere to Costa's setting in that the interference due to the host signal is nearly removed, thus achieving close to the side-information capacity. In addition, these schemes are feasible in practice, for that randomize codewords are replaced by low-complexity quantization-based algebraic codewords. These two sample-wise schemes are referred to as "Quantization Index Modulation" (QIM) and "Scalar Costa Scheme" (SCS), respectively.

During the last years, both QIM and SCS have been thoroughly studied and extended into different directions such as non-ergodic and correlated Gaussian channel noise [69], non uniform quantizers [114] and recently to lattice codebooks [115–117]. This chapter extends these schemes to another direction: multiple information embedding. Multiple information embedding refers to the situation of embedding several messages into the same host signal, with or without different robustness and transparency requirements. Of course finding a single unifying mathematical analysis to general multiple information embedding situations under broad assumptions seems to be a hard task. Instead, this chapter addresses the very common situations of multiple user information embedding, from an information theoretic point-of-view. The basic

problem is that of finding the set of rates at which the different watermarks can be simultaneously embedded. This problem has tight relationship, as well as in the case of single embedding, to conventional multiple user information theory. Consider for example watermark applications such as copy control, transaction tracking, broadcast monitoring and tamper detection. Obviously, each application has its own robustness requirement and its own targeted data hiding rate. Thus, embedding different watermarks intended to different usages into the same host signal naturally has strong links with transmitting different messages to different users in a conventional multi-user transmission environment. The design and the optimization of algorithms for multiple information embedding applications should then benefit from recent advances and new findings in multi-user information theory [118].

In this chapter, we first argue that many multiple information embedding situations can be nicely modeled as communication over either a Broadcast Channel (BC) or a Multiple Access Channel (MAC), both with state information available at the transmitter(s). Next, we rely heavily on the general theoretical solutions for these channels (cf. [118]) to devise efficient practical encoding schemes. The resulting schemes consist, in essence, of applying the initial QIM or SCS as many times as the number of different watermarks to be embedded. This choice conforms the near-to-optimum performance of both QIM and SCS in the single user case. However, we show that these schemes should be appropriately designed when it comes to the multi-user case. A joint design is required so as to closely approach the theoretical performance limits. For instance, for both the resulting BC-based and MAC-based schemes, the improvement brought by this joint design is pointed out through comparison with the straightforward -rather intuitive- corresponding scheme which is obtained by simply super-imposing (i.e with no joint design) scalar schemes (or DPCs for the ideal coding).

We introduce the notion of "awareness" to refer to this joint design. An interesting contribution at this stage is then that awareness helps in improving system performance. Awareness in the BC case basically implies that the encoder responsible for embedding the robust watermark is aware that a fragile signal is also embedded (with a known power) and thus, it modifies the coding scheme accordingly. This allows increasing the rate for the robust watermark. Similarly, awareness in the MAC case

takes advantage at the embedder from the knowledge that a peeling-off decoder is used, i.e., that the better watermark is subtracted, an operation that changes the channel seen by the embedder. Again, the way to account for this MAC-awareness is to change the coding parameters. This increases the rate at which the worse watermark can be reliably communicated. The improvement brought up by awareness is demonstrated through both achievable rate region and Bit Error Rate (BER) analysis. We finally show that performance can further be made closer to the theoretical limits by considering lattice-based codebooks. Some finite-dimensional lattices with good packing and quantization properties are considered for illustration.

The rest of the chapter is organized as follows. After introducing the notation we recall in section 5.2 some fundamental principles of the DPC technique. Also we give a brief review of the formal statement of the information embedding problem as communication with side information available only at the transmitter, together with the state of the art of the sub-optimal practical coding schemes. These schemes will serve as baseline for the construction of the proposed approaches throughout the chapter. Then we turn in section 5.3 to a detailed discussion on multiple information embedding applications. Two mathematical models corresponding to the multiple information embedding problem viewed either as communication over a degraded Broadcast Channel (BC) with state information at the transmitter or as communication over a Multiple Access Channel (MAC) with state information at the transmitters are provided. Corresponding performance analyses are undertaken in sections 5.4.1 and 5.4.2, respectively. For each of theses two mathematical models, analysis is carried out within the context of two watermarks using scalar-valued codebooks. Section 5.5 extends these results to the more general case of an arbitrary number of watermarks using high dimensional lattice-based codebooks. Finally, we close with a discussion followed by some concluding remarks in section 5.6.

### 5.1.1 Notation

Throughout the chapter, boldface fonts denote vectors. We use uppercase letters to denote random variables, lowercase letters for their individual values, e.g. $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and calligraphic fonts for sets , e.g. $\mathcal{X}$. Unless otherwise specified, vectors are assumed to be in the $n$-dimensional Euclidean space $(\mathbb{R}^n, \|.\|)$ where $\|.\|$

denotes the Euclidean norm of vectors. For a generic random vector $\mathbf{X}$, we use $\mathbb{E}_{\mathbf{X}}[.]$ to denote the expectation taken with respect to $\mathbf{X}$ and $f_{\mathbf{X}}(.)$ to denote its probability density function (PDF). The Gaussian distribution with mean $\mu$ and square deviation $\sigma^2$ is denoted by $\mathcal{N}(\mu, \sigma^2)$. A random variable $\mathbf{X}$ with conditional PDF given $\mathbf{S}$ is denoted by $\mathbf{X}|\mathbf{S}$.

## 5.2 Information Embedding and DPC

In this section, we first give a brief review of the information embedding problem as DPC. The resulting framework uses DPC principles to provide the ultimate theoretical performance which is used as baseline for comparison in the rest of this chapter. Next, both the well-known Scalar Costa Scheme (SCS) [110] and Quantization Index Modulation [108] are briefly reviewed together with their achievable performance.

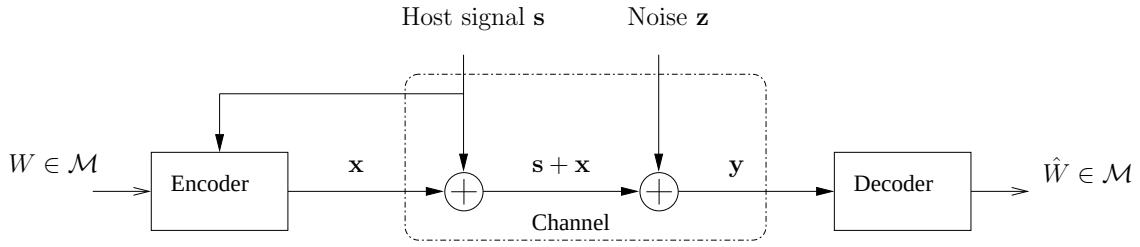



Figure 5.1: Blind information embedding viewed as DPC over a Gaussian channel.

### 5.2.1 Information Embedding as Communication with Side Information

Fig. 5.1 depicts a block diagram of the blind information embedding problem considered as a communication problem. A message $m$ has to be sent to a receiver through some channel called the *watermark channel*. This channel is assumed to be i.i.d. Gaussian. We denote the Gaussian channel noise by $\mathbf{Z}$, with $Z_i \sim \mathcal{N}(0, N)$. The message $m$ may be represented by a sequence $\{W\}$ of $\mathcal{M}$-ary symbols, with $\mathcal{M} = \{1, \ldots, M\}$, so as the transmission of the message $m$ amounts to that of the corresponding symbols $\{W\}$. Thus, from now on, we will concentrate on the reliable transmission of $W$. Also, we will loosely use the term "message" to refer to the symbol

$W$ itself, instead of $m$. Prior to transmission, the message $W$ is encoded into a signal $\mathbf{X}$ called the watermark which is then embedded into the cover signal $\mathbf{S} \in \mathbb{R}^n$, thus forming the watermarked or composite signal $\mathbf{S} + \mathbf{X}$.

We assume that the cover signal $S_i \sim \mathcal{N}(0, Q)$ is Gaussian i.i.d. distributed and the watermarker $\mathbf{X}$ must satisfy the input power constraint $\mathbb{E}[\mathbf{X}^2] \leq P$. $M$ is the greatest integer smaller than or equal to $2^{nR}$ and $R$ is the transmission rate, expressed in number of bits per host sample that the encoder can reliably transmit. The watermark must be embedded without introducing any perceptible distortion to the host signal. This corresponds to the input power constraint in conventional power-limited communication and is commonly called the *transparency* requirement. *The robustness* requirement -as for it- refers to the ability of the watermark to survive channel degradations. Rather than considering watermarking as communication over a very noisy channel where the cover signal $\mathbf{S}$ acts as self-interference as in Spread-Spectrum Modulations (SSM), it has been realized [109,119] that blind watermarking can be viewed as communication with state information non-causally known at the transmitter. The state information being the cover signal $\mathbf{S}$ (entirely known at the transmitter). The relevant work is the initial Costa's "Writing on Dirty Paper" [111], also commonly known as "Dirty-paper coding" (DPC). Costa was the first to show the remarkable result that the interference $\mathbf{S}$, non-causally known only to the encoder, incurs no loss in capacity relative to the standard interference-free AWGN channel, i.e.

$$C = \frac{1}{2} \log \left( 1 + \frac{P}{N} \right). \tag{5.1}$$

The achievability of this capacity is based on random binning arguments for general channels with state information [112]. This consists in a random construction of Gaussian codebook $\{\mathcal{U}_1, \ldots, \mathcal{U}_M\}$ and random partition of its codewords into "bins". In the Gaussian case (side information $\mathbf{S}$ and noise $\mathbf{Z}$ i.i.d. Gaussian), Costa showed that with the choice of the input distribution $p(u, x|s)$ such that $\mathbf{X} \sim \mathcal{N}(0, P)$ independent of $\mathbf{S}$, and

$$\mathbf{U} = \mathbf{X} + \alpha \mathbf{S} \text{ with } \alpha = P/(P + N), \tag{5.2}$$

this capacity is attained. The ideal DPC is however not feasible in practice due to the huge random codewords size needed for efficient binning. Therefore some sub-optimal

lower-complexity practical schemes have been proposed in [108] and in [110]. A brief review is given in the following section.

### 5.2.2 Sub-optimal Coding

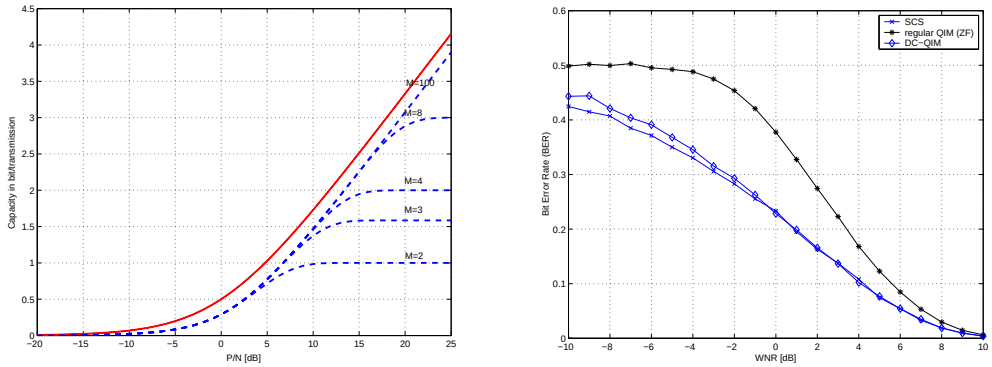
Following Costa's ideal DPC, Chen *et al.* proposed the use of structured quantization-based codebooks in [108]. The resulting embedding scheme is referred to as Quantization Index Modulation (QIM). Whereas in [110], Eggers *et al.* designed a practical "Scalar Costa Scheme" (SCS) where the random codebook $\mathbf{U}$ is chosen to be a concatenation of dithered scalar uniform quantizers. The watermark signal is a scaled version of the quantization error, i.e,

$$x_k = \widetilde{\alpha}\left(\mathcal{Q}_\Delta\Big(s_k - \frac{W}{M}\Delta\Big) - \Big(s_k - \frac{W}{M}\Delta\Big)\right), \tag{5.3}$$

with $\Delta = \sqrt{12P}/\widetilde{\alpha}$, $\widetilde{\alpha} = \sqrt{P/(P+2.71N)}$ and $\mathcal{Q}_\Delta$ is the uniform scalar quantizer with constant step size $\Delta$. Decoding is also based on scalar quantization of the received signal $\mathbf{y} = \mathbf{x} + \mathbf{s} + \mathbf{z}$ followed by a thresholding procedure. That is, the estimate $\widehat{W}$ of the transmitted message $W$ is the closest integer to $r_k M/\Delta$, with $r_k = \mathcal{Q}_\Delta(y_k) - y_k$. The optimum parameter $\widetilde{\alpha} = \sqrt{P/(P+2.71N)}$ is obtained by numerically maximizing the Shannon mutual information $I(W;r)$[1]. With this setting, SCS performs close to the optimal DPC. The above mentioned QIM which corresponds to the inflation parameter $\alpha = 1$ is less efficient, especially at relatively high noise levels. This QIM embedding function is referred to as *regular* QIM. Regular QIM can be slightly modified so as to increase its immunity to noise. The resulting scheme, called Distortion-Compensated QIM (DC-QIM), corresponds to $\alpha = P/(P+N)$ and performs very close to SCS as shown in Fig. 5.2.

We observe that SCS and DC-QIM schemes, though clearly sub-optimal, perform close to the ideal DPC. This constitutes the main motivation focus adapting them to the multiple watermarking situation.

[1]Caution should be exercised here as $\mathbf{r}$ is the error quantization of the received signal, not the received signal itself.




(a) Capacity (b) Bit Error Rate

Figure 5.2: Performance of Scalar Costa Scheme (SCS), regular and Distortion-Compensated QIM in terms of both (a) Capacity in bit per transmission and (b) Bit Error Rate, BER. (a) $M$-ary SCS capacity (dashed) and full AWGN capacity (solid). (b) SCS outperforms -by far- regular QIM in terms of BER. A slight improvement over DC-QIM is observed at very low Watermark-to-Noise Ratio WNR $= 10\log_{10}(P/N)$.

## 5.3 Multiple User Information Embedding: Broadcast and MAC Set-ups

In an information embedding context, "multiple user" refers to the situation where several messages $W_i$ have to be embedded into a common cover signal $\mathbf{S}$. The embedding may or may not require different robustness and transparency requirements. This means that each of these messages can be *robust*, *semi-fragile* or *fragile*. Also, depending on the targeted application, the watermarking system may require either joint or separate decoding. For joint decoding, think of one single *trusted* authority checking for several (say $K$) watermarks at once. For separate (or distributed) decoding, think of several (say $L$) authorities each checking for its own watermark.

In order to emphasize the very general case, one may even imagine these decoders having access to different noisy versions of the same watermarked content. This is due to the possibly different channel degradations the watermarked content may experience depending on the receiver location (think of a watermarked image being transmitted over a mobile network, with watermarking verification performed at different nodes of this network). As in decoding process, we may wish that the encoding of these messages be performed either jointly or separately. Some of the situations

of concern are given by the illustrative examples described above, with the receivers playing the role of the transmitters and vice-versa. Of course, though intentionally kept in its very general form, this model may not include some specific multiple information embedding situations. This is due to the difficulty of finding a single unifying approach. Nevertheless, the framework that we proposed is sufficiently general to involve the most important multiple information embedding scenarios. For instance two classes of such scenarios, that we will recognize as being equivalent to communication over a degraded Broadcast Channel (BC) and a Multiple Access Channel (MAC) in subsections 5.3.1 and 5.3.2 respectively, are worthy of deep investigations. To simplify the exposition, we first restrict our attention to the two-watermarks embedding scenario. Extension to the general case then follows.

### 5.3.1 A Mathematical Model for BC-like Multiuser Information Embedding

Consider an information embedding system aiming at embedding two messages $W_1$ and $W_2$, assumed to be $M_1$-ary and $M_2$-ary respectively, into the same cover signal $\mathbf{S} \sim \mathcal{N}(0, Q)$. We suppose that one single *trusted authority* (the same encoder) has to embed these two messages and that embedding should be performed in such a way that the corresponding two watermarks correspond to two different usages (separate decoders). For example, the watermark $\mathbf{X}_2$ (carrying $W_2$) should be very robust whereas the watermark $\mathbf{X}_1$ (carrying $W_1$) may be of lesser robustness. This means that the watermark $\mathbf{X}_2$ must survive channel degradations up to some noise level $N_2$ larger than $N_1$, i.e. $N_2 \gg N_1$. Furthermore, the previously mentioned transparency requirement implies that the two watermarks put together must satisfy the input power constraint $P$, i.e. $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ is constrained to have $\mathbb{E}_{\mathbf{X}}[\mathbf{X}^2] = P$. Assuming in dependent watermarks[2] $\mathbf{X}_1$ and $\mathbf{X}_2$, we suppose with no loss of generality that $\mathbb{E}_{\mathbf{X}_1}[\mathbf{X}_1^2] = \gamma P$ and $\mathbb{E}_{\mathbf{X}_2}[\mathbf{X}_2^2] = (1-\gamma)P$, where $\gamma \in [0, 1]$ may be arbitrarily chosen to share power between both watermarks.

In practice, this multiple watermarking scenario can be used to serve multiple purposes. In the scope of watermarking of medical images for example, we may wish

[2]A justification of this assumption will be provided in section 5.4.

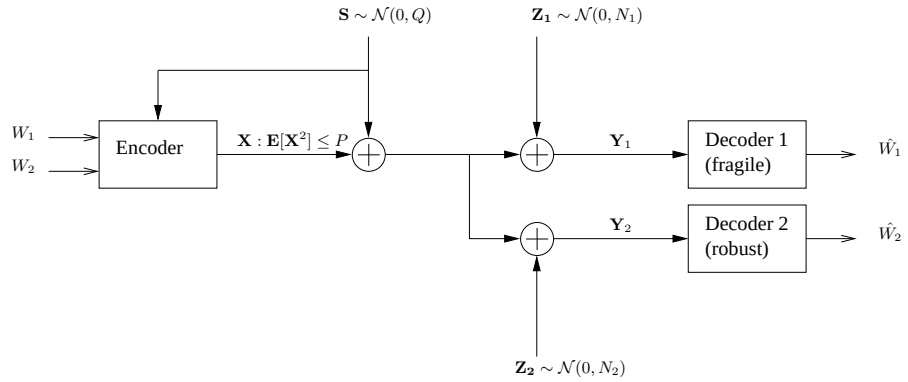



Figure 5.3: Two users information embedding viewed as communication over a two-users Gaussian Broadcast Channel (GBC).

to store the patient information into the corresponding image, in a secure and private way. This information is sometimes called the "annotation part" of the watermark and is hence required to be sufficiently robust. Further, we may wish to use an additional possibly fragile "tamper detection part" to detect tampering. Another example stems from proof-of-ownership applications: we may wish to use one watermark to convey ownership information (should be robust) and a second watermark to check for content integrity (should be semi-fragile or fragile). A third example concerns watermarking for distributed storage. Suppose that a multimedia content (e.g. video or audio) has to be stored in different storage devices. Furthermore, we want to protect this multimedia content against piracy, by the use of a watermark. As the alteration level induced by the storage and extraction processes may differ from one device to another, the encoding technique must enable the reliably decoded rate to adapt to the actual alteration level. Of course many other examples and applications can be listed. We just mention here that the model at hand can be applied every time one watermarking authority (i.e, one transmitter) has to simultaneously embed several watermarks in such a way that these watermarks satisfy different robustness requirements.

Assuming Gaussian channel noises $\mathbf{Z}_i \sim \mathcal{N}(0, N_i)$, with $i = 1, 2$, a simplified block diagram of the transmission scheme of interest is shown in Fig. 5.3. Decoder $i$ decodes $\widehat{W_i}$ from the received signal $\mathbf{Y}_i = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{S} + \mathbf{Z}_i$ at rate $R_i$. An error occurs if $\widehat{W_i} \neq W_i$. Functionally, this is the very transmission diagram of a two users Gaussian Broadcast Channel (GBC) with state information available at the transmitter but not at the receivers. In addition, the watermark $\mathbf{X}_2$ having to be robust plays the role of

the message directed to the "degraded user" in a broadcast context. Conversely, the watermark $\mathbf{X}_1$ plays the role of the message directed to the "better user". Also, here we have considered only two watermarks. The similarity with a $L$-users BC will be retained if, instead of just two watermarks, $L$ watermarks are to be simultaneously embedded by the same so-called *trusted* authority.

### 5.3.2 A Mathematical Model for MAC-like Multiuser Information Embedding

We now consider another situation. Again, the watermarking system aims at embedding two independent messages $W_1$ and $W_2$ into the same cover signal $\mathbf{S}$. However, the present situation is different in that, this time, (i) embedding is performed by two different authorities, each having to embed its own message satisfying a given power requirement and (ii) at the receiver, a single *trusted* authority having to check for both watermarks. We assume no particular cooperation between the two embedding authorities, meaning that the watermarks $\mathbf{X}_1$ (carrying $W_1$) and $\mathbf{X}_2$ (carrying $W_2$) should be designed independently of each other. In addition, watermarks $\mathbf{X}_1$ and $\mathbf{X}_2$ must satisfy independent power constraints $P_1$ and $P_2$, respectively. Thus, two individual power constraints must be satisfied, which differs from the above scenario (BC-like) in which the power constraint $P$ is taking over both watermarks $X_1 + X_2$.

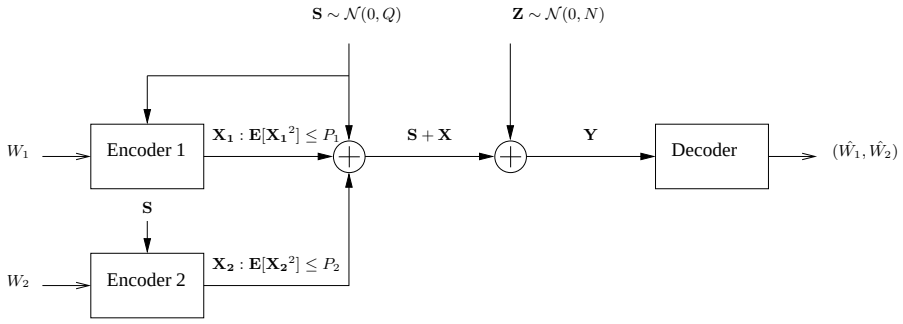



Figure 5.4: Two users information embedding viewed as communication over a (two users) Multiple Access Channel (MAC).

In practice, this multiple watermarking scenario can be used to serve multiple purposes. Loosely speaking, every watermarking system addressing the same application multiple times is concerned. An example stemming from proof-of-ownership applications is as follows. Consider two different creators independently watermark-

ing the same original content $\mathbf{S}$, as it is common for large artistic works such as feature films and music recordings. Each of the two watermarks may contain private information. A common *trusted* authority may have to check for both watermarks. This is the case when an authenticator agent needs to track down the initial owner of an illegally distributed image, for example. A second example is the so-called hybrid in-band on-channel digital audio broadcasting [108]. In this application, we would like to simultaneously transmit two digital signals within the same existing analog (AM and/or FM) commercial broadcast radio without interfering with conventional analog reception. Thus, the analog signal is the cover signal and the two digital signals are the two watermarks. These two digital signals may be designed independently. One digital signal may be used as an enhancement to refine the analog signal and the other as supplemental information such as station or program identification. A third application concerns distributed (i.e., at different places) watermarking: some fingerprinting can be embedded right at the camera, while possible annotations can be added next to the storage device.

Assuming a Gaussian channel noise $\mathbf{Z} \sim \mathcal{N}(0, N)$ corrupting the watermarked signal $\mathbf{S} + \mathbf{X}$, a simplified diagram is shown in Fig. 5.4. The encoder $i$, $i = 1, 2$, encodes $W_i$ into $\mathbf{X}_i$ at rate $R_i$. The decoder outputs $(\widehat{W_1}, \widehat{W_2})$. An error occurs if $(\widehat{W_1}, \widehat{W_2}) \neq (W_1, W_2)$. Functionally, this is the very transmission diagram of a two users Gaussian Multiple Access Channel (MAC) with state information available at the transmitters but not to the receiver. Note that, here, we have considered only two watermarks. The similarity with a $K$-users MAC will be retained if, instead of just two authorities, $K$ different embedding authorities, each encoding its own message are considered.

The above discussion indicates that there are strong similarities between multiple information embedding and conventional multiple user communication. In sections 5.4 and 5.5, we rely on recent findings in multi-user information theory [118] to devise efficient implementable multiple watermarking schemes and address their practical achievable performance. Also, in our attempt to further highlight the analogy with conventional multi-user communication, we will sometimes use the terms "multiple users", "degraded user" and "better user" to loosely refer to "multiple watermarks", "the receiver decoding the more noisy watermarked content" and "the receiver decoding

the less noisy watermarked content", respectively.

## 5.4 Information Embedding over Gaussian Broadcast and Multiple Access Channels

In this section, we are interested in designing efficient low-complexity multiuser information embedding schemes for each of the two situations considered in section 5.3. We first present a straightforward rather intuitive method based on super-imposing two SCSs. This simple method can be thought as being "coding-unaware". Next, we use the similarity between multi-user information embedding problem and transmission over Gaussian BC and MAC to design more efficient multiple watermarking schemes. We reefer to these latter strategies as being "broadcast-aware" and "MAC-aware", respectively. The improvement brought by "awareness" is illustrated through both achievable rate regions and BER enhancements. Note that we will assume, throughout this section, that the flat-host assumption is satisfied as long as quantization is concerned.

### 5.4.1 Broadcast-Aware Coding for Two-Users Information Embedding

A simple approach for designing a coding system for the two users information embedding problem considered in subsection 5.3.1 consists in using two independent single-user DPCs (or SCSs for the corresponding suboptimal practical implementation).[3]

#### Broadcast-unaware coding (double DPC)

In essence, the ideal coding is based on successive encoding at the transmitter as follows:

(i) Use a first DPC (denoted by DPC2) taking into account the known state $\mathbf{S}$ and the power of unknown noise $\mathbf{Z}_2$ to form the most robust watermark $\mathbf{X}_2$ intended

[3]Note that this is not the most naive way of working, each DPC being tuned based on all information available.

to the degraded user. By using (5.2), DPC1 is given by $\mathbf{X}_2 = \mathbf{U}_2 - \alpha_2 \mathbf{S}$ with

$$\mathbf{U}_2|\mathbf{S} \sim \mathcal{N}\left(\alpha_2 \mathbf{S}, (1-\gamma)P\right), \text{ with } \alpha_2 = \frac{(1-\gamma)P}{(1-\gamma)P + N_2}. \tag{5.4}$$

(ii) Use a second DPC (denoted by DPC1) taking into account the known state $\mathbf{S} + \mathbf{X}_2$, sum of the cover signal $\mathbf{S}$ and the already formed watermark $\mathbf{X}_2$, and the power of unknown noise $\mathbf{Z}_1$ to form the less robust watermark $\mathbf{X}_1$ intended to the better user. By using (5.2), DPC1 is given by $\mathbf{X}_1 = \mathbf{U}_1 - \alpha_1(\mathbf{S} + \mathbf{X}_2)$ with

$$\mathbf{U}_1|\mathbf{U}_2, \mathbf{S} \sim \mathcal{N}\left(\alpha_1(\mathbf{S} + \mathbf{X}_2), \gamma P\right), \text{ with } \alpha_1 = \frac{\gamma P}{\gamma P + N_1}. \tag{5.5}$$

(iii) Finally, transmit the composite signal $\mathbf{S} + \mathbf{X}$ over the watermark channel, with $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ being the composite watermark. The received signals are $\mathbf{Y}_1 = \mathbf{X} + \mathbf{S} + \mathbf{Z}_1$ and $\mathbf{Y}_2 = \mathbf{X} + \mathbf{S} + \mathbf{Z}_2$.

Note that the watermark $\mathbf{X}_2$ should be embedded first because of the following intuitive reason. When considering the extreme case where the watermark $\mathbf{X}_1$ is fragile, this watermark should be by design, damaged by any operation that alters the cover signal $\mathbf{S}$. Since robust embedding is such an operation, the fragile watermark should be embedded last. The theoretical achievable region $\mathscr{R}_{\text{BC}}$ with DPC1 and DPC2 is given by

$$\begin{aligned} \mathscr{R}_{\text{BC}}(P) = \bigcup_{0 \le \gamma \le 1} \Big\{ (R_1, R_2) : R_1 &\le \frac{1}{2}\log_2\left(1 + \frac{\gamma P}{N_1}\right), \\ R_2 &\le R(\alpha_2, (1-\gamma)P, Q, \gamma P + N_2) \Big\}, \end{aligned} \tag{5.6}$$

where $R(\alpha, P, Q, N) = \frac{1}{2}\log_2\left(P(P+Q+N)/(PQ(1-\alpha)^2 + N(P+\alpha^2 Q))\right)$ and $Q$ is the power of the host signal $\mathbf{S}$ . Using straightforward algebra, which is omitted for brevity, it can be shown that the rates in (5.6) can be obtained by evaluating the achievable region [118]

$$\begin{aligned} \mathscr{R}_{\text{BC}}(P_{U_1U_2|S}) = \Big\{ (R_1, R_2) : R_1 &\le I(U_1; Y_1|U_2) - I(U_1; S|U_2), \\ R_2 &\le I(U_2; Y_2) - I(U_2; S) \Big\}, \end{aligned} \tag{5.7}$$

with the choice of $U_1$ and $U_2$ given by (5.5) and (5.4), respectively.

Using (5.3) and following the way a single user SCS is derived from the corresponding single-user DPC, a suboptimal practical two-users scalar information embedding

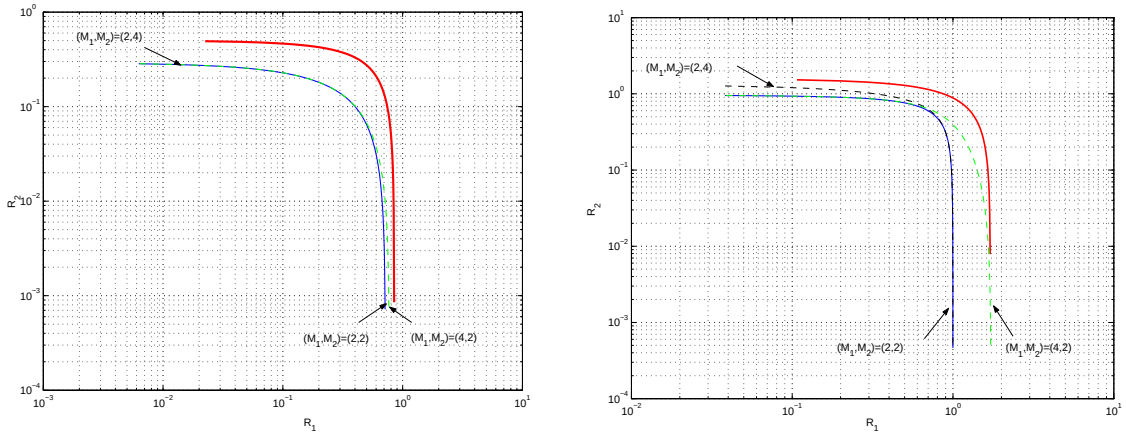
scheme can be derived by independently super-imposing two SCSs (denoted by SCS1 and SCS2 and taken as scalar versions of DPC1 and DPC2, respectively). SCS1 and SCS2 are applied sequentially, starting with SCS2 for the design of the watermark $\mathbf{x}_2$ as an appropriate scaled version of the quantization error of the cover signal $\mathbf{s}$. Then, SCS1 designs the watermark $\mathbf{x}_1$ as an appropriate scaled version of the quantization error of the sum signal $\mathbf{s}+\mathbf{x}_2$. The corresponding uniform scalar quantizers $\mathcal{Q}_{\Delta_1}$ and $\mathcal{Q}_{\Delta_2}$ have step sizes $\Delta_1=\sqrt{12\gamma P}/\widetilde{\alpha_1}$ and $\Delta_2=\sqrt{12(1-\gamma)P}/\widetilde{\alpha_2}$, where

$$(\widetilde{\alpha_1},\widetilde{\alpha_2})=\left(\sqrt{\frac{\gamma P}{\gamma P+2.71N_1}},\sqrt{\frac{(1-\gamma)P}{(1-\gamma)P+2.71N_2}}\right). \tag{5.8}$$

Note that the flat-host assumption on signals $\mathbf{s}$ and $\mathbf{s}+\mathbf{x}_2$ is assumed to hold as supposed above. We denote by $(\widetilde{R_1},\widetilde{R_2})$ the transmission throughput achieved by this set-up. This rate pair is computed numerically. Results are depicted in Fig. 5.5 and are compared to the theoretical rate pair $(R_1,R_2)\in\mathscr{R}_{\text{BC}}$ given by (5.6), for two examples of channel parameters. The noise in first example, (i.e., the one such that $P/N_2=0$ dB) may model a channel attack which has the same power as the composite watermark $\mathbf{X}=\mathbf{X}_1+\mathbf{X}_2$. The performance of this first approach is worthy of some brief discussion.

(i) From (5.6), we see that DPC1- as given by (5.5)- is optimal. The achievable rate $R_1$ corresponds to that of a channel with not only no interfering cover signal $\mathbf{S}$, but also no interference signal $\mathbf{X}_2$. Thus, the message $W_1$ can be sent at its maximal rate, as if it were embedded alone. From "Decoder 1" point of view, the channel from $W_1$ to $\mathbf{Y}_1$ is functionally equivalent to a single-user channel from $W_1$ to $\mathbf{Y}_1'=\mathbf{Y}_1-\mathbf{U}_2=\mathbf{X}_1+(1-\alpha_2)\mathbf{S}+\mathbf{Z}_1$, having just $(1-\alpha_2)\mathbf{S}$ as state information, not $\mathbf{S}+\mathbf{X}_2$. Yet, it is not that $\mathbf{Y}_1$ is a single-user channel, but rather that the amount of reliably decodable information $W_1$ is exactly the same as if $W_1$ were transmitted alone over $\mathbf{Y}_1'$. Moreover DPC2- as given by (5.4) is not optimal. The reason is that the achievable rate $R_2$ given by (5.6) is inferior to $\frac{1}{2}\log_2(1+(1-\gamma)P/(\gamma P+N_2))$. The latter rate is that of a watermark signal subject to the full interference penalty from both the cover signal $\mathbf{S}$ and the watermark $\mathbf{X}_1$.

(ii) SCS1 performs close to optimality. The scalar channel having a message $W_1$




(a) Rates for $P/N_1 = 5$ dB and $P/N_2 = 0$ dB.

(b) Rates for $P/N_1 = 12$ dB and $P/N_2 = 9$ dB.

Figure 5.5: Theoretical and feasible transmission rates for broadcast-like multiple user information embedding for two examples of SNR. For each SNR, the upper curve corresponds to the theoretical rate region $\mathscr{R}_{\mathrm{BC}}$ (5.6) of the double DPC and the lower curve corresponds to the achievable rate region $(\widetilde{R_1}, \widetilde{R_2})$ of the two superimposed SCSs with quantization parameters given by (5.8). Dashed line correspond to (2-ary,4-ary) and (4-ary,2-ary) transmissions.

as input and the quantization error as output is functionally equivalent to that from $W_1$ to $\mathbf{r}'_1 = Q_{\Delta_1}(\mathbf{y}'_1) - \mathbf{y}'_1$, where $\mathbf{y}'_1$ is the single-user channel suffering only partly from the interference $\mathbf{X}_2$[4]. The practical transmission rate over this channel is given by the mutual information $I(W_1; r'_1)$, the maximum of which (i.e $\widetilde{R_1}$) is obtained with the choice (5.8) of $\widetilde{\alpha_1}$. However, being derived from DPC2 -which is itself non optimal- SCS2 is obviously suboptimal. Consequently the parameter $\widetilde{\alpha_2}$ chosen does not maximize the mutual information $I(W_2; r_2)$, with $\mathbf{r}_2 = Q_{\Delta_2}(\mathbf{y}_2) - \mathbf{y}_2$.

In the following section, we show that the encoding of $W_2$ can be improved so as to bring the rate $\widetilde{R_2}$ close to $R_2^{(\max)} = \frac{1}{2}\log_2(1 + (1-\gamma)P/(\gamma P + N_2))$. The corresponding scheme, which we call "Joint scalar DPC" in the sequel, improves system performance by making multiple information embedding broadcast-aware.

[4]Note that in the equivalent channel $\mathbf{y}'_1 = \mathbf{x}_1 + (1-\alpha_2)\mathbf{s} + \mathbf{z}_1$, the watermark $\mathbf{x}_1$ is formed as a scaled version of the quantization error of the channel state $(1-\alpha_2)\mathbf{s}$ and not $\mathbf{s} + \mathbf{x}_2$ as before.

**Broadcast-aware coding (joint DPC)**

In section 5.3.1, we have shown that the communication scenario depicted in Fig. 5.3 is basically that of a degraded GBC with state information non-causally known to the transmitter but not to the receivers. In [118], it has been shown that the capacity region $\mathcal{C}_{\text{BC}}$ of this channel is given by

$$\begin{aligned}\mathcal{C}_{\text{BC}}(P) = \bigcup_{0\leq\gamma\leq 1}\Big\{(R_1,R_2): R_1 &\leq \frac{1}{2}\log_2\left(1+\frac{\gamma P}{N_1}\right),\\ R_2 &\leq \frac{1}{2}\log_2\left(1+\frac{(1-\gamma)P}{\gamma P+N_2}\right)\Big\},\end{aligned} \tag{5.9}$$

which is that of a GBC with no interfering signal $\mathbf{S}$. This region can be attained by an appropriate successive encoding scheme that uses two well designed DPCs. The encoding of $W_1$ (DPC1) is still given by (5.5). For the encoding of $W_2$ however, the key point is to consider the unknown watermark $\mathbf{X}_1$ as noise. We refer to this by saying that the encoder is "aware" of the existence of the watermark $\mathbf{X}_1$ and takes it into account. The resulting DPC (again denoted by DPC2) uses the cover signal $\mathbf{S}$ as channel state and $\mathbf{Z}_2+\mathbf{X}_1$ as total channel noise:

$$\mathbf{U}_2|\mathbf{S}\sim\mathcal{N}(\alpha_2\mathbf{S},(1-\gamma)P) \text{ with } \alpha_2=\frac{(1-\gamma)P}{(1-\gamma)P+(N_2+\gamma P)}, \tag{5.10}$$

and $\mathbf{X}_2=\mathbf{U}_2-\alpha_2\mathbf{S}$. Obviously, this encoding does not remove the interference due to $\mathbf{X}_1$. Nevertheless, DPC1 is optimal in that it attains the maximal possible rate $R_2^{(\max)}$ at which $W_2$ can be sent together with $W_1$.

**Feasible rate region**

Consider now a scalar implementation of this Joint DPC scheme consisting in two successive SCSs. DPC2 can be implemented by a scalar scheme SCS2, quantizing the cover signal $\mathbf{s}$ and outputting the watermark $\mathbf{x}_2$ as an appropriate scaled version of the quantization error. We denote by $\widetilde{\alpha_1}$ and $\Delta_1$ the corresponding scale factor and quantization step size, respectively. DPC1 can be implemented by a scalar scheme SCS1, quantizing the newly available signal $\mathbf{s}+\mathbf{x}_2$ and outputting the watermark $\mathbf{x}_1$ as an appropriately scaled version of the quantization error. We denote by $\widetilde{\alpha_2}$ and $\Delta_2$ the corresponding scale factor and quantization step size, respectively. Let $\mathbf{Y'}_1=\mathbf{Y}_1-\mathbf{U}_2$ be the channel functionally equivalent to $\mathbf{Y}_1$ introduced above. The resulting achievable rate region $\widetilde{\mathscr{R}}_{\text{BC}}$, practically feasible with this coding, is given by

$$\begin{aligned}\widetilde{\mathscr{R}}_{\text{BC}}(P) = \bigcup_{0\leq\gamma\leq 1}\Big\{(\widetilde{R_1},\widetilde{R_2}) : \widetilde{R_1} &\leq \max_{\alpha_1\in[0,1]} I\big(W_1; \underbrace{\mathcal{Q}_{\Delta_1(\alpha_1,\gamma)}(\mathbf{y}_1') - \mathbf{y}_1'}_{\mathbf{r}_1'}\big),\\ \widetilde{R_2} &\leq \max_{\alpha_2\in[0,1]} I\big(W_2; \underbrace{\mathcal{Q}_{\Delta_2(\alpha_2,\gamma)}(\mathbf{y}_2) - \mathbf{y}_2}_{\mathbf{r}_2}\big)\Big\}.\end{aligned} \tag{5.11}$$

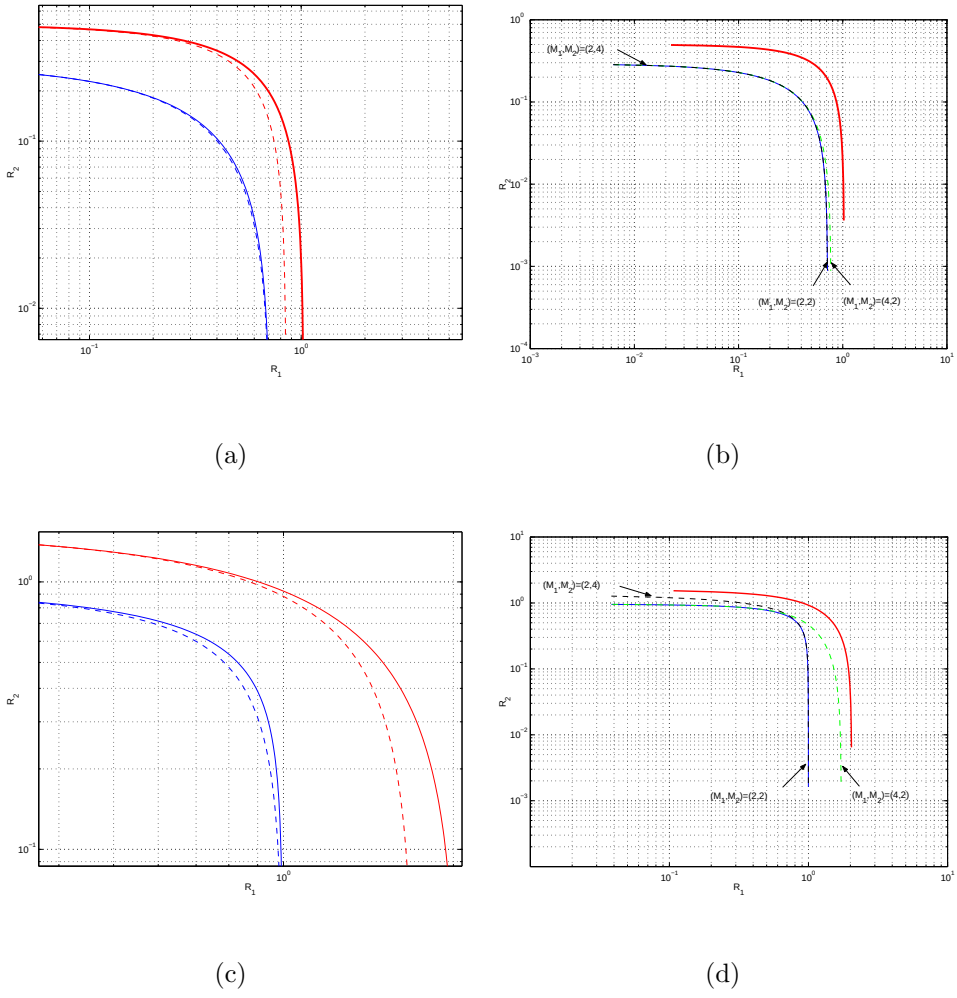
The proof simply follows from the discussion above regarding the equivalent channels from $W_1$ to $\mathbf{r}_1'$ for the message $W_1$ and from $W_2$ to $\mathbf{r}_2$ for the message $W_2$. Each of these two channels conforms the single user channel considered in the initial work [110] and has hence a similar expression of the transmission rate. The inflation parameters pair $(\widetilde{\alpha_1},\widetilde{\alpha_2})$ maximizing the right hand side terms of (5.11) is given by

$$(\widetilde{\alpha_1},\widetilde{\alpha_2}) = \left(\sqrt{\frac{\gamma P}{\gamma P + 2.71 N_1}}, \sqrt{\frac{(1-\gamma)P}{(1-\gamma)P + 2.71(\gamma P + N_2)}}\right). \tag{5.12}$$

The region (5.11), obtained through a Monte-Carlo based integration, is depicted in Fig. 5.6 and is compared to the ideal DPC region $\mathcal{C}_{\text{BC}}$ given by (5.9), for two choices of channel parameters: weak channel noise (Fig. 5.6(c) and Fig. 5.6(d)) and strong channel noise (Fig. 5.6(a) and Fig. 5.6(b)). The latter may model, for example, a channel attack with power equal to that of the composite watermark $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$, as mentioned above. Note that we need to compute the conditional probabilities $p_{\mathbf{r}_1'}(\mathbf{r}_1'|W_1)$ and $p_{\mathbf{r}_2}(\mathbf{r}_2|W_2)$. These are computed using the high resolution quantization assumption $Q \gg P$, which is relevant in most watermarking applications.

Improvement over the "Double DPC" is made possible by increasing the rate $R_2$ at which the robust watermark can be sent. It is precisely "awareness" that allows such improvement. However, note that this improvement is more significantly for high SNR as shown in Fig. 5.6(c). Whereas for low SNR, this improvement (thought still theoretically possible) is almost not visible for scalar codebooks, as shown in Fig. 5.6(a). This can be interpreted as follows: The above mentioned "awareness", which can be viewed as a power saving technique for the "degraded user", does not sensibly improve the overall communication when the channel is very bad.[5] Both theoretical and feasible rate regions of the BC-aware scheme are also depicted for non-binary inputs in Fig. 5.6(d) and Fig. 5.6(b). It can be seen that, depending on the SNR,

[5]Note however that, this should not be considered as a drawback since when the channel is very bad capacity is not needed, but reliability transmission.




(a) (b)

(c) (d)

Figure 5.6: The improvement brought by "BC-awareness" (with binary inputs) is depicted for (a) $P/N_1 = 5$ dB, $P/N_2 = 0$ dB and (c) $P/N_1 = 12$ dB, $P/N_2 = 9$ dB. Solid line corresponds to the rate region of the BC-aware scheme achievable theoretically (upper) and practically (lower). Dashed line corresponds to the rate region of the BC-unaware scheme achievable theoretically (upper) and practically (lower). (b) and (d): achievable rate region of the BC-aware scheme for $M_1$-ary and $M_2$-ary alphabets depicted for (b) $P/N_1 = 5$ dB, $P/N_2 = 0$ dB and (d) $P/N_1 = 12$ dB, $P/N_2 = 9$ dB.

the practically feasible rate region (5.11) can more-or-less approach the theoretical capacity region $\mathcal{C}_{\text{BC}}$, by increasing the sizes $M_1$ and $M_2$ of the input alphabets $\mathcal{M}_1$ and $\mathcal{M}_2$.[6]

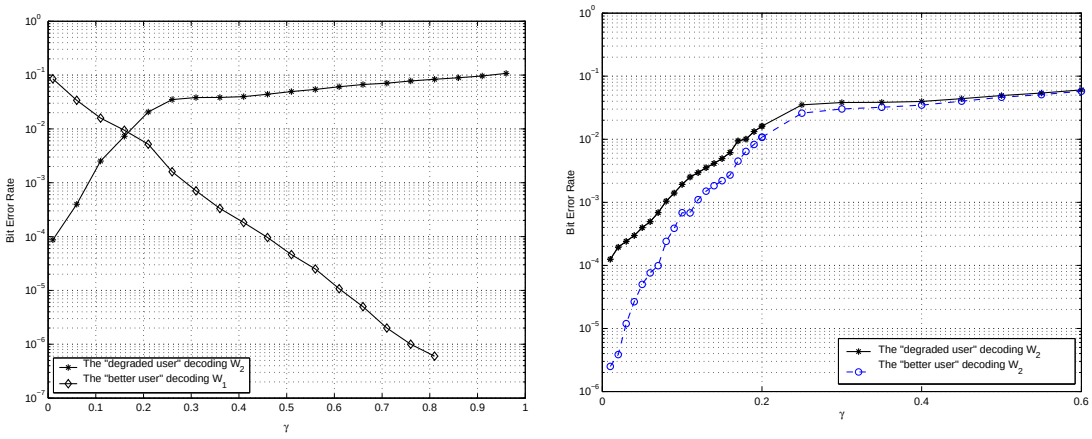
**Bit Error Rate analysis and discussion**

Another performance analysis is based on measured BERs for hard decision based decoding of binary scalar DPC. Results are obtained with Monte Carlo based simulation and are depicted in Fig. 5.7. Note that the set of channel parameters chosen in Fig. 5.7 may model a wide range of admissible channel attacks on the individual watermarks, since the individual SNRs, $\text{SNR}_1 = 10\log_{10}(\gamma P/N_1)$ and $\text{SNR}_2 = 10\log_{10}((1-\gamma)P/(\gamma P + N_2))$, vary from $-8$ dB to 12 dB and from $-15$ dB to 9 dB respectively as the power-sharing parameter $\gamma$ varies from 0 to unity. However, this may be not a good choice to model a strong attack on the composite watermark $\mathbf{X}_1+\mathbf{X}_2$ (for example, one such that $P/N_2 = 0$ dB). For such an attack, the individual rates are very low and the BERs are very bad. In principle, it would be possible to use any provably efficient error correction code for each of the channels $\mathbf{Y}_1$ and $\mathbf{Y}_2$ taken separately. However, at low SNR ranges, it is well known that repetition coding is almost optimal. The curves in Fig. 5.7(a) are obtained with $(\rho_1, \rho_2) = (4, 4)$, meaning that $W_1$ and $W_2$ are repeated 4 times each.

We observe that as $\gamma \in [0, 1]$ increases, the power part of the signal $\mathbf{X}$ allocated to the watermark carrying $W_1$ becomes larger and that allocated to the watermark carrying $W_2$ becomes smaller. This causes the corresponding BER curves to monotonously decrease and increase, respectively. Also, it can be checked that, when plotted separately, these curves are identical to those of a SCS with a signal-to-noise power ratio equal to $\text{SNR}_1$ and $\text{SNR}_2$, respectively. This conforms the assumption made above regarding the functionally equivalent channels $\mathbf{y}_1'$ and $\mathbf{y}_2$. The curves depicted in Fig. 5.7 also motivate the following discussion.

(i) In practical situations, the repetition factors $\rho_1$ and $\rho_2$ should be chosen in light of the desired transmission rates and robustness requirements. The choice $(\rho_1, \rho_2) = (4, 4)$ made above should be taken just as a baseline example. Channel

[6]However, a gap of about 1.53 dB should remain visible, i.e., $R_1 - \widetilde{R_1} > 1.53$ dB and $R_2 - \widetilde{R_2} > 1.53$ dB.




(a) (b)

Figure 5.7: Broadcast-aware multiple user information embedding. (a): Bit Error Rates for binary transmission using repetition coding. (b): Each decoder can only decode "his" own watermark. Thought much less noisy, the "best user" performs only slightly better than the "degraded user" in decoding message $W_2$. The messages $W_1$ and $W_2$ are repeated 4 times each, i.e. $(\rho_1, \rho_2) = (4, 4)$ and channel parameters are such that $P/N_1 = 12$ dB and $P/N_2 = 9$ dB.

coding as a means of providing additional redundancy obviously strengthens the watermark immunity to channel degradations. However, such a redundancy inevitably limits the transmission rate. This means that for equal targeted transmissions rates $R_1$ and $R_2$, the repetition factors $\rho_1$ and $\rho_2$ should satisfy $\rho_2 \geq \rho_1$.

(ii) The scalar DPC considered here for multiple watermarking is constructed using insights from coding for broadcast channels [120, 121], as mentioned above. Interestingly, in such channels the user who experiences the better channel (less noisy) has to reliably decode the message assigned to the (degraded) user who experiences the worst channel (more noisy). In an information embedding context, this means that the robust watermark, which is supposed to survive channel degradation levels up to $N_2$, should be reliably decodable if, actually, the channel noise is less-powerful. However, this strategy, which is inherently related to the principle of superposition coding at the transmitter combined with successive decoding (peeling off technique) at the "better user" (Decoder 1) [122], makes more sense in the situations where the "better user" is unable to reliably

decode its own message if it does not primarily subtract off the interference due to the message assigned to the "degraded user". The DPC-based scheme is fundamentally different in that the interference is already subtracted off at the encoder. As a consequence, the "better user" does not need to decode the message of the degraded user.[7]

(iii) There could however have advantages and disadvantages for the DPC-based scheme described above to follow such a strategy. An obvious disadvantage concerns security issues. In a transmission scheme where security is a major issue, the "better user" should not be able to reliably decode the message assigned to the "degraded user". By opposition, an obvious advantage stems from the following observation. If channel quality is improved, resulting in better SNR in the transmission of $W_2$, the "degraded user", being at present a "better user", should be able to reliably decode much more information $W_2$ than it does with the old channel quality. For the above described DPC-based scheme, to fulfill this additional requirement, one should focus on maximizing (over $\alpha_1$) the conditional mutual information $I(W_1; r_1|W_2)$. This would however lead to a suboptimal choice $\widetilde{\alpha_1'}$ of the inflation parameter $\alpha_1$ for the transmission of $W_1$, an d consequently to a smaller transmission rate $\widetilde{R_1} = I(W_1; r_1')|_{\alpha_1=\widetilde{\alpha_1'}}$.

(iv) The present DPC-scheme, as is, does not fully satisfy the above mentioned broadcast property. From Fig. 5.7(b), we observe that the "better user" does not fully exploit the fact of being much less noisy (than the degraded user) to more reliably decode $W_2$: The improvement in BER upon the "degraded user is very small and is even negligible, as shown in Fig. 5.7(b). And even though this improvement seems to behave like the improvement in SNR (which is maximal at $\gamma = 0$), it is actually smaller than the one, $10\log_{10}((\gamma P + N_2)/(\gamma P + N_1))$ dB, which should be visible if the "better user" were able to reliably decode $W_2$ as in superposition coding.

[7]Note that by opposition to superposition coding, there is an important embedding ordering at the encoder. The benefit of such ordering is a decoupling of the receivers and hence a more scalable system. Each receiver needs only know its own codebook to extract its message.

### 5.4.2 MAC-Aware Coding for Two Users Information Embedding

In this section we are interested in designing implementable multiple watermarking schemes for the situation described in subsection 5.3.2. Paralleling the development made in section 5.4, we provide a performance analysis for two MAC-aware and unaware multiple watermarking strategies.

#### MAC-unaware coding (double DPC)

The situation described in subsection 5.3.2 corresponds in essence to two Costa's channels. A simple approach for designing a watermark system for this situation consists in two single-user DPCs (or SCSs for the corresponding practical implementation). Let $\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{S} + \mathbf{Z}$ denote the received signal. Upon reception, the receiver should reliably decode the messages $W_1$ and $W_2$ having been embedded into the watermarks $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. However, since decoding is performed jointly, the successful decoding of one of the two messages should benefit of the other message. This is illustrated through the following possible coding.

(i) Encoder 2 uses a DPC (DPC2) taking into account the known state $\mathbf{S}$ and the power of unknown noise $\mathbf{Z}$ to form the watermark $\mathbf{X}_2$ of power $P_2$ and carrying $W_2$ as $\mathbf{X}_2 = \mathbf{U}_2 - \alpha_2\mathbf{S}$, where

$$\mathbf{U}_2 \sim \mathcal{N}\left(\alpha_2\mathbf{S}, P_2\right), \text{ with } \alpha_2 = \frac{P_2}{P_2 + N}. \tag{5.13}$$

At reception, the decoder first decodes $W_2$ and then cleans up the channel by subtracting the interference penalty $\mathbf{U}_2$ that the transmission of $W_2$ causes to that of $W_1$.[8] Thus the channel for $W_1$ is made equivalent to $\mathbf{Y}_1 = \mathbf{Y} - \mathbf{U}_2 = \mathbf{X}_1 + (1 - \alpha_2)\mathbf{S} + \mathbf{Z}$. This "cleaning up" step is inherently associated with *successive* decoding and is sometimes referred to as the *peeling-off* technique. Hence, encoder 1 can reliably transmit $W_1$ over the channel $\mathbf{Y}_1$ by using a second DPC (DPC1).

[8]Note that, theoretically, the decoder looks for the (unique) codeword $\mathbf{U}_2$ such that $(\mathbf{U}_2, \mathbf{Y})$ is jointly typical. In practice however, the decoder only knows an estimate $\hat{\mathbf{U}}_2$ of the codeword $\mathbf{U}_2$ even if $W_2$ is decoded perfectly, since the host $\mathbf{S}$ is unknown at the receiver (see discussion in Section 5.4.2).

(ii) Encoder 1 forms $\mathbf{X}_1$ as $\mathbf{X}_1 = \mathbf{U}_1 - \alpha_1 \mathbf{S}$, where

$$\mathbf{U}_1 | \mathbf{S} \sim \mathcal{N}(\alpha_1 \mathbf{S}, P_1), \text{ with } \alpha_1 = (1-\alpha_2)\frac{P_1}{P_1+N} = \frac{NP_1}{(P_1+N)(P_2+N)}. \quad (5.14)$$

The rate pair $(R_1, R_2) \in \mathscr{R}_{\text{MAC}}$ achieved by the considered two DPCs are those corresponding to the corner point (B1) of the achievable region $\mathscr{R}_{\text{MAC}}$ depicted in Fig. 5.8, and are given by

$$R_1(B1) = \frac{1}{2}\log_2\left(1 + \frac{P_1}{N}\right), \quad (5.15a)$$

$$R_2(B1) = \frac{1}{2}\log_2\left(\frac{P_2(P_2+Q+N+P_1)}{P_2Q(1-\alpha_2)^2 + (N+P_1)(P_2+\alpha_2^2 Q)}\right). \quad (5.15b)$$

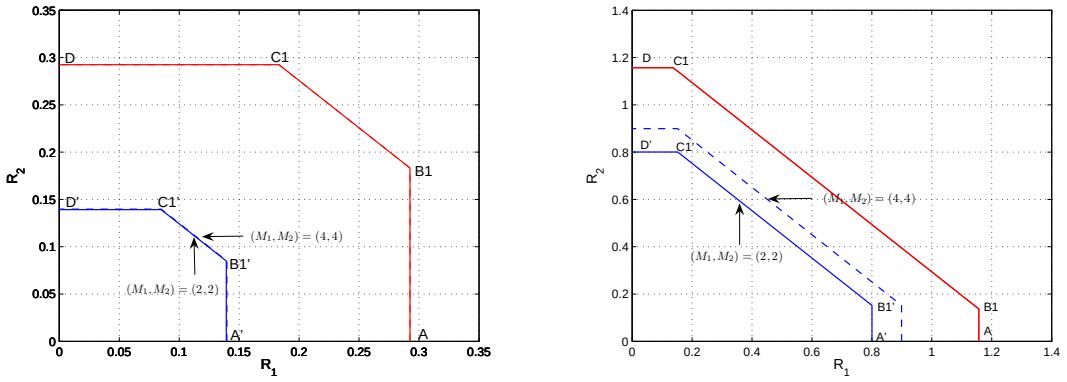
Using straightforward algebra which is omitted for brevity, it can be shown that the rates in (5.15) correspond to a corner point in the rate region obtained by evaluating the achievable region [118]

$$\begin{aligned} \mathscr{R}_{\text{MAC}}(P_1, P_2) = \Big\{(R_1, R_2) : R_1 &\leq I(U_1; Y|U_2) - I(U_1; S|U_2), \\ R_2 &\leq I(U_2; Y|U_1) - I(U_2; S|U_1), \\ R_1 + R_2 &\leq I(U_1, U_2; Y) - I(U_1, U_2; S), \Big\}, \end{aligned} \quad (5.16)$$

with the choice of codebooks $U_1$ and $U_2$ given by (5.13) and (5.14), respectively. Following the same principle, similar DPC schemes allowing to attain the corner points (A), (C1) and (D) can be designed. The corner point (A) corresponds to the watermark $\mathbf{X}_1$ (i.e, the information $W_1$) being sent at its maximum achievable rate whereas the watermark $\mathbf{X}_2$ (i.e, the information $W_2$) not transmitted at all. The two corner points (C1) and (D) correspond to the points (B1) and (A), respectively, with the roles of the watermarks $\mathbf{X}_1$ and $\mathbf{X}_2$ reversed. Any rate pair lying on the lines connecting these corner points can be attained by time sharing. We concentrate on the corner point (B1) and consider a practical implementation of this theoretical set-up. This can be performed by using two SCSs, SCS1 and SCS2, consisting of scalar versions of DPC1 and DPC2. The uniform scalar quantizers $\mathcal{Q}_{\Delta_1}$ and $\mathcal{Q}_{\Delta_2}$ have step sizes $\Delta_1 = \sqrt{12P_1}/\widetilde{\alpha_1}$ and $\Delta_2 = \sqrt{12P_2}/\widetilde{\alpha_2}$, where

$$(\widetilde{\alpha_1}, \widetilde{\alpha_2}) = \left((1-\alpha_2)\sqrt{\frac{P_1}{P_1+2.71N}}, \sqrt{\frac{P_2}{P_2+2.71N}}\right), \quad (5.17)$$

conform the codebooks choice in (5.13) and (5.14).[9] Note that the signal $\mathbf{S}$ is assumed to be flat-host as mentioned above. The feasible transmission rate pair achieved by this practical coding corresponds to the corner point (B1') in the diagrams shown in Fig. 5.8. Note that results ate depicted for two choices of channel parameters: strong channel noise (shown in Fig. 5.8(a)) and weak channel noise (shown in Fig. 5.8(b)). The strong noise may model a channel attack which has the same power as the composite watermark $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$. The performance of this first approach can




(a) Rates for $P_1 = P_2; (P_1 + P_2)/N = 0$ dB. (b) Rates for $P_1 = P_2; (P_1 + P_2)/N = 9$ dB.

Figure 5.8: Theoretical and feasible transmission rates for MAC-like multiple user information embedding. The frontier with corner points (A), (B1), (C1), and (D) corresponds to the theoretical rate pair $(R_1, R_2) \in \mathscr{R}_{\text{MAC}}$ of the double ideal DPC. The frontier with corner points (A'), (B1'), (C1'), and (D') corresponds to the feasible rate pair $(\widetilde{R_1}, \widetilde{R_2})$ of the two superimposed SCSs. Dashed line corresponds to practical rates obtained with the use of quaternary alphabets.

be summarized as follows.

(i) From (5.15b), we see that DPC1- as given by (5.14)- is optimal. The interference due to the cover signal $\mathbf{S}$ and the second watermark $\mathbf{X}_2$ is completely canceled. Hence, the watermark $\mathbf{X}_1$ can be sent at its maximal rate $R_1$, as if it were alone over the watermark channel. The channel from $W_1$ to $\mathbf{Y}$ is functionally equivalent to that from $W_1$ to $\mathbf{Y}_1 = \mathbf{Y} - \mathbf{U}_2$. However, DPC2- as given by (5.13)- is non optimal, because the rate $R_2$ given by (5.15b) is inferior to $\frac{1}{2}\log_2(1 + P_2/(P_1 + N))$, which is that of a watermark subject to the full

[9] Note that the choice $(\widetilde{\alpha_1}, \widetilde{\alpha_2})$ in (5.17) does not maximize the input-output mutual information. Rather, it directly traces the way in which the codebooks are generated in (5.13) and (5.14).

interference penalty from both the cover signal $\mathbf{S}$ and the watermark $\mathbf{X}_1$.

(ii) SCS1 performs close to optimality. The scalar channel is equivalent to that from $W_1$ to $\mathbf{r}_1 = \mathcal{Q}_{\Delta_1}(\mathbf{y}_1) - \mathbf{y}_1$. The practical transmission rate over this channel is given by the mutual information $I(W_1; r_1)$, the maximum of which (i.e $\widetilde{R_1}$) is obtained with the choice (5.17) of $\widetilde{\alpha_1}$. However, SCS2 is non optimal, simply because DPC2 is not. The inflation parameter $\widetilde{\alpha_2}$ does not maximize the mutual information $I(W_2; r)$, with $\mathbf{r} = \mathcal{Q}_{\Delta_2}(\mathbf{y}) - \mathbf{y}$. Thus, the achievable rate $\widetilde{R_2}$ is not maximal and corresponds to $\widetilde{R_2} = I(W_2; r)|_{\alpha_2 = \widetilde{\alpha_2}}$.

The encoding of $W_2$ can be improved so as to bring the achievable rate $\widetilde{R_2}(B1')$ close to $R_2^{(\max)} = \frac{1}{2}\log_2\left(1 + \frac{P_2}{P_1+N}\right)$. The corresponding scheme, called "joint DPC", enhances the performance by making multiuser information embedding MAC-aware.

**MAC-aware coding (joint DPC)**

In subsection 5.3.2, we argued that the communication scenario depicted in Fig. 5.4 is basically that of a Gaussian Multiple Access Channel (GMAC) with state information non-causally known to the transmitters but not to the receiver. In [118], it is reported that the capacity region $\mathcal{C}_{\text{MAC}}$ of this channel is given by

$$\begin{aligned} \mathcal{C}_{\text{MAC}}(P_1, P_2) = \Big\{(R_1, R_2) : R_1 &\leq \frac{1}{2}\log_2\left(1 + \frac{P_1}{N}\right), \\ R_2 &\leq \frac{1}{2}\log_2\left(1 + \frac{P_2}{N}\right), \\ R_1 + R_2 &\leq \frac{1}{2}\log_2\left(1 + \frac{P_1 + P_2}{N}\right)\Big\}, \end{aligned} \tag{5.18}$$

which is that of a GMAC with no interfering signal $\mathbf{S}$. This region, with corner points (A), (B), (C) and (D), is shown in Fig. 5.9 and can be attained by an appropriate successive encoding scheme that uses well designed DPCs. Consider for example the corner point (B). The encoding of $W_1$ is again given by (5.14), recognized above to be optimal[10]. The encoding DPC2 of $W_2$ however should be changed so as to consider the watermark $\mathbf{X}_1$ as noise. We refer to this situation by saying that the encoder should be "aware" of the existence of $\mathbf{X}_1$ and acts accordingly. The resulting DPC

[10]Note however that as $\alpha_1$ depends on $\alpha_2$, the optimal inflation parameter for DPC1 becomes $\alpha_1 = P_1/(P_1 + P_2 + N)$.

(again denoted by DPC2) uses the cover signal $\mathbf{S}$ as channel state and the signal $\mathbf{Z} + \mathbf{X}_1$ as total channel noise:

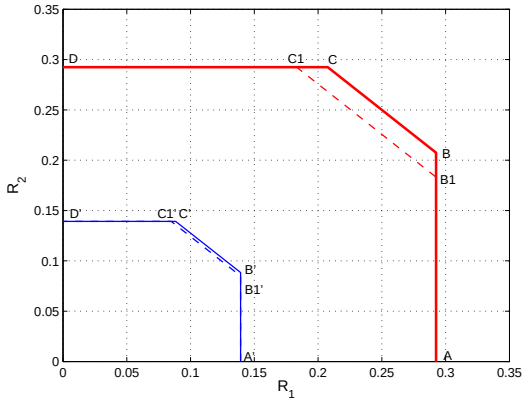
$$\mathbf{U}_2|\mathbf{S} \sim \mathcal{N}(\alpha_2\mathbf{S}, P_2), \text{ with } \alpha_2 = \frac{P_2}{P_2 + (P_1 + N)}. \tag{5.19}$$

Obviously the interference due to $\mathbf{X}_1$ is not removed. However, this scheme is optimal in that it achieves the maximum rate $R_2^{(\max)}$ at which the message $W_2$ can be sent as long as the message $W_1$ is sent at its maximum rate.

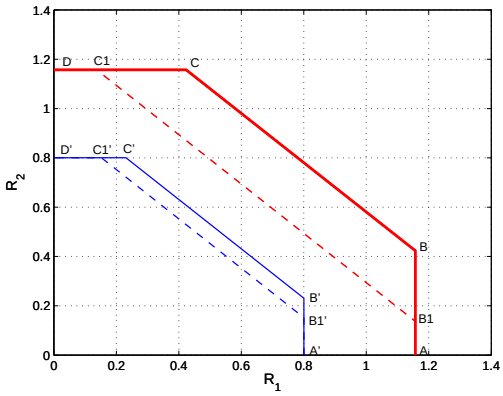
**Feasible rate region**

We consider now a practical implementation for this joint scheme through two jointly designed SCSs with parameters $(\widetilde{\alpha_1}, \Delta_1)$ and $(\widetilde{\alpha_2}, \Delta_2)$, respectively. This results in a maximal feasible transmission rate $\widetilde{R_2}$ given, as before, by $\widetilde{R_2} = \max_{\alpha_2 \in [0,1]} I(W_2; r)$. However, the corresponding scale parameter $\alpha_2$ is set this time to its optimal choice, i.e, $\widetilde{\alpha_2} = \sqrt{P_2/(P_2 + 2.71(N + P_1))}$.[11] The resulting transmission rate pair $(\widetilde{R_1}, \widetilde{R_2})$ is represented by the corner point (B') in Fig. 5.9 for two examples of channel conditions: weak noise (shown in Fig. 5.9(b)) and strong noise modelling a strong channel attack on the composite watermark $\mathbf{X} = \mathbf{X}_1 + \gamma X_2$ (shown in Fig. 5.9(a)). Reversing the roles of the watermarks $\mathbf{X}_1$ and $\mathbf{X}_2$, the joint design also pushes out the corner point (C1') to (C'). More generally any rate pair on the region frontier delimited by the corner points (A'), (B'), (C') and (D') is made practically feasible by subsequent time-sharing. When the message $W_i$ *travels alone* over the watermark channel, the equivalent channel is $\mathbf{Y}_i = \mathbf{Y} - \mathbf{U}_j$, $(i, j) \in \{1, 2\} \times \{1, 2\}, i \neq j$. Hence, $W_i$ can be sent at its maximum feasible rate, which is given by $\max_{\alpha_i \in [0,1]} I(W_i; r_i)$, with $\mathbf{r}_i = \mathcal{Q}_{\Delta_i}(\mathbf{y}_i) - \mathbf{y}_i$. When the two messages *travel together*, the maximal sum of the two feasible rates corresponds to one of the two (say $W_1$) set to its maximal feasible rate and the other ($W_2$) facing a total channel noise of $\mathbf{z} + \mathbf{x}_1$. Of course, we can reverse the roles of $W_1$ and $W_2$, and the maximal feasible sum rate remains unchanged. Consequently, the

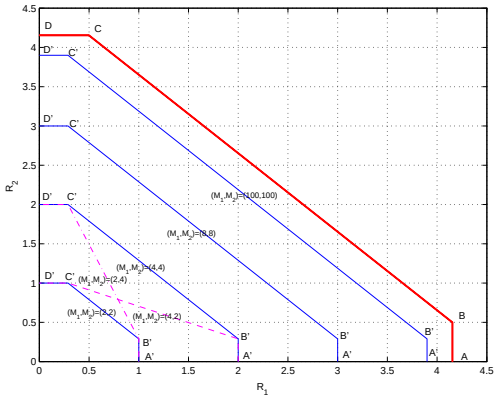
[11] Note that the optimal inflation parameter for SCS1 is $\widetilde{\alpha_1} = (P_1 + N)\sqrt{P_1/P_1 + 2.71N}/(P_1 + P_2 + N)$.




(a)

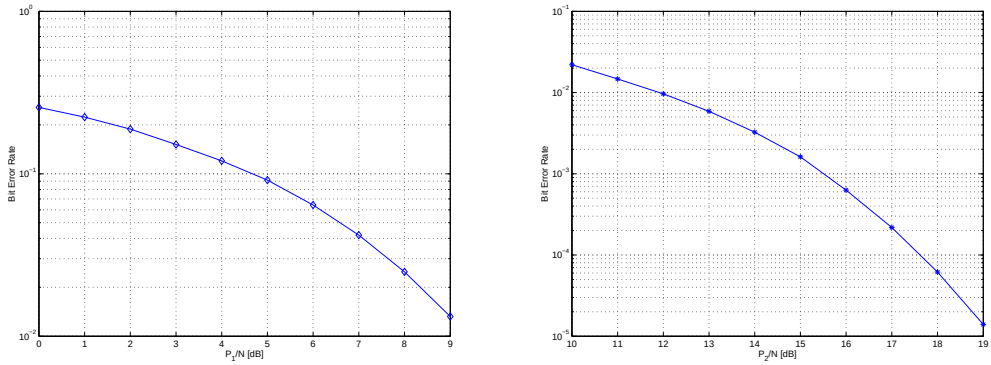



(b)




(c)

Figure 5.9: MAC-like multiple user information embedding. The improvement brought by "awareness" is depicted for (a) strong channel noise, $P_1 = P_2$, $(P_1 + P_2)/N = 0$ dB and (b) weak channel noise, $P_1 = P_2$, $(P_1 + P_2)/N = 9$ dB. Solid line delineates the capacity region of the MAC-aware scheme achievable theoretically (upper) and practically (lower). Dashed line delineates the rate region of the MAC-unaware scheme achievable theoretically (upper) and practically (lower). (c) Capacity region of the MAC-aware scheme with ($M_1-$ary,$M_2-$ary) input alphabets for very high SNR.




(a) Decoding of $W_1$. (b) Decoding of $W_2$.

Figure 5.10: MAC-like multiple user information embedding bit error rates. The two messages $W_1$ and $W_2$ are sent at rates $(\widetilde{R_1}, \widetilde{R_2})$ corresponding to the corner point (B') in the capacity region diagram shown in Fig. 5.9.

achievable rate region $\widetilde{\mathscr{R}}_{\text{MAC}}$ is given by

$$\begin{aligned}\widetilde{\mathscr{R}}_{\text{MAC}}(P_1,P_2) = \Big\{(\widetilde{R_1},\widetilde{R_2}) : \widetilde{R_1} &\leq \max_{\alpha_1\in[0,1]} I\big(W_1; \mathcal{Q}_{\Delta_1(\alpha_1,P_1)}(\mathbf{y}_1) - \mathbf{y}_1\big),\\ \widetilde{R_2} &\leq \max_{\alpha_2\in[0,1]} I\big(W_2; \mathcal{Q}_{\Delta_2(\alpha_2,P_2)}(\mathbf{y}_2) - \mathbf{y}_2\big),\\ \widetilde{R_1}+\widetilde{R_2} &\leq \max_{\alpha_1\in[0,1]} I\big(W_1; \mathcal{Q}_{\Delta_1(\alpha_1,P_1)}(\mathbf{y}_1) - \mathbf{y}_1\big)\\ &\quad + \max_{\alpha_2\in[0,1]} I\big(W_2; \mathcal{Q}_{\Delta_2(\alpha_2,P_2)}(\mathbf{y}) - \mathbf{y}\big)\Big\}.\end{aligned} \tag{5.20}$$

Fig. 5.9 shows the achievable rate region $\widetilde{\mathscr{R}}_{\text{MAC}}$ gain brought by the joint design of the DPCs in approaching the theoretical limit $\mathcal{C}_{\text{MAC}}$ (5.18). This improvement, which is more visible at large SNR (i.e., weak channel noise), is more significant in the situations where $W_1$ and $W_2$ are both transmitted with non-zero rates. In this case, for a given transmission rate $\widetilde{R_2}$ of $W_2$, the maximal transmission rate at which $W_1$ can be sent is larger and equivalently for any rate $\widetilde{R_1}$. Moreover the gap to the theoretical limit $\mathcal{C}_{\text{MAC}}$ can be reduced by use of sufficiently large size alphabets $\mathcal{M}_1$ and $\mathcal{M}_2$ as shown in Fig. 5.9(c). Of course, this is achieved at the cost of a slight increase in encoding and decoding complexities.

### Bit Error Rate analysis and discussion

Consider the coding scheme given by (5.14) and (5.19). The *peeling off* technique aims to clean up the channel before decoding $W_1$, by subtracting the codeword $\mathbf{U}_2$.

This is good for performance evaluation and for theoretically proving the achievability of the corner point (B) of the capacity region. However, in practice, the decoder does not know the exact codeword $\mathbf{U}_2$ that "Encoder 2" had used. Instead, it has access to an estimation $\widehat{\mathbf{U}}_2$ of $\mathbf{U}_2$, which is determined as the (unique) codeword being typically joint with the received signal $\mathbf{Y}$. Of course, the accuracy of this estimation, and hence that of decoding message $W_1$, depends on the value of SNR2. For instance, a bad SNR2 will likely cause decoding of $W_2$ to fail. Thus, the estimate $\widehat{\mathbf{U}}_2$ does not resemble the exact $\mathbf{U}_2$ and it is rather seen as an additional noise source. However, at good (high) SNR2, the estimate $\widehat{\mathbf{U}}_2$ of codeword $\mathbf{U}_2$ is accurate and the *peeling off* technique is efficient as shown in Fig. 5.10. For instance, at the same SNR, decoding message $W_1$ is more accurate than that of $W_2$, though $P_2 = 10P_1$.

## 5.5 Multi-User Information Embedding and Structured Lattice-Based Codebooks

In this section, we extend the results obtained in section 5.4 in the context of two watermarks to the general multiple watermarking case. We also broaden our view to consider the high dimensional lattice-based codebooks case.

### 5.5.1 Broadcast-Aware Information Embedding: the Case of $L$ - Watermarks

The results in subsection 5.4.1 can be straightforwardly extended to the situation where, instead of just two messages, $L$ messages $W_i$, $i = 1, 2, \ldots, L$, have to be embedded into the same cover signal $\mathbf{S}$. The composite watermark is $\mathbf{X} = \sum_{i=1}^{L} \mathbf{X}_i$. The watermark $\mathbf{X}_i$ has power $P_i$ and carries the message $W_i$, where $\sum_{i=1}^{L} P_i = P$. We consider a Gaussian Broadcast Channel $\mathbf{Z}_i \sim \mathcal{N}(0, N_i)$ and assume without loss of generality that $N_1 \leq N_2 \leq \ldots \leq N_L$. This means that the watermarks should be designed in such a way that $\mathbf{X}_i$ is less robust than $\mathbf{X}_j$ for $i \leq j$. Following the joint DPC scheme above, the watermarks should be ordered according to their relative strengths and put on top of each other. This means that the most robust (that

is $\mathbf{X}_L$) should be embedded first whereas the most fragile (that is $\mathbf{X}_1$) should be embedded last. For $i$ ranging from $L$ to 1, the watermark signal $\mathbf{X}_i$ is obtained by applying an i-th DPC (denoted here by DPCi). The available state information to be used is $\mathbf{S}_i = \mathbf{S} + \sum_{j=i+1}^{L} \mathbf{X}_j$, the sum of the cover signal $\mathbf{S}$ and the already embedded watermarks $\mathbf{X}_j$, $j > i$. The channel noise is $\mathbf{Z}_i + \sum_{j=1}^{i-1} \mathbf{X}_j$, the sum of the ambient noise $\mathbf{Z}_i$ and the not-yet embedded watermarks $\mathbf{X}_j$, $j < i$, accumulated and taken as an additional noise component. Note that the Gaussiannity of this noise term and its statistic independence from both $\mathbf{X}_i$ and $\mathbf{S}_i$ as well as the statistic independence of $\mathbf{X}_i$ on $\mathbf{S}_i$ conform to the statistical independence between the state information, the watermark and the noise in the original Costa set-up [111]. Thus, the optimal inflation parameter for DPCi is $\alpha_i = P_i/(N_i + \sum_{j=1}^{i} P_j)$ and the corresponding maximal achievable rate $R_i$ is given by

$$R_i = \frac{1}{2}\log_2\left(1 + \frac{P_i}{N_i + \sum_{j=1}^{i-1} P_j}\right). \tag{5.21}$$

A scalar implementation of this broadcast-based joint DPC for embedding $L$ watermarks, consists in $L$ SCSs jointly designed. Similarly to the 2-watermark case and using the equivalent channel $\mathbf{y}'_i = \mathbf{y}_i - \sum_{j=i+1}^{L} \mathbf{u}_j$ for SCSi, $i = 1, 2, \ldots, L$, the corresponding achievable rate region is given by the union of all rate $L$-tuples $(\widetilde{R_1}, \ldots, \widetilde{R_L})$ simultaneously satisfying

$$\widetilde{R_i} \leq \max_{\alpha_i \in [0,1]} I\big(W_i; \mathcal{Q}_{\Delta_i(\alpha_i, P_i)}(\mathbf{y}'_i) - \mathbf{y}'_i\big). \tag{5.22}$$

The union is taken over all power assignments $\{P_i\}$, $i = 1, 2, \ldots, L$, satisfying the average power constraint $\sum_{j=1}^{L} P_i = P$. The inflation parameter maximizing the right hand side term of (5.22) is

$$\widetilde{\alpha_i} = \sqrt{\frac{P_i}{P_i + 2.71\left(N_i + \sum_{j=1}^{i-1} P_j\right)}}. \tag{5.23}$$

### 5.5.2 MAC-Aware Information Embedding: The Case of $K$-Watermarks

The results in subsection 5.4.2 can be straightforwardly extended to the situation where, instead of just two messages, $K$ messages $W_i$, $i = 1, \ldots, K$, have to be independently encoded into the same cover signal $\mathbf{S}$ and jointly decoded, by the same watermarking authority. We suppose that the watermark $\mathbf{X}_i$, carrying $W_i$, $i = 1, \ldots, K$, has power $P_i$. Also we denote by $\mathbf{Z} \sim \mathcal{N}(0, N)$ the channel noise, assumed to be i.i.d. Gaussian. Functionally, this is a $K$-user GMAC with state information available at the transmitters but not to the receiver, as argued in subsection 5.3.2. The capacity region of such a channel follows a straightforward generalization of (5.18). This region is given by the union of all rate $K$-tuples simultaneously satisfying

$$\begin{aligned} R_i &\leq \frac{1}{2}\log_2\left(1+\frac{P_i}{N}\right), \ i = 1, 2, \ldots, K, \\ \sum_{j=1}^{K} R_j &\leq \frac{1}{2}\log_2\left(1+N^{-1}\sum_{i=1}^{K} P_i\right), \end{aligned} \tag{5.24}$$

where the union is taken over all power assignments $\{P_i\}$, $i = 1, \ldots, K$. Following the two-message case considered above, any corner point of this region can be attained by applying $K$ well designed DPCs. Consider for example the corner point (B) corresponding to the message $W_1$ transmitted at its maximum rate. Upon reception of $\mathbf{Y} = \sum_{i=1}^{K} \mathbf{X}_i + \mathbf{S} + \mathbf{Z}$, the receiver should perform successive decoding so as to reliably decode the $K$-tuple $(W_1, W_2, \ldots, W_K)$.

In order to attain the corner point (B), decoding should be performed in such a way that $W_K$ is decoded first, $W_1$ is decoded last and $W_j$ is decoded before $W_i$ for $j > i$. Consequently, coding consists in a set of $K$ DPCs, denoted by $\{\text{DPCi}\}$, with $i$ ranging from $K$ to 1. At the receiver, the decoder sees the equivalent channel $\mathbf{Y} - \sum_{j>i} \mathbf{U}_j$ in the decoding of the message $W_i$. Thus, an optimal DPCi for this equivalent channel is given by: $\mathbf{X}_i = \mathbf{U}_i - \alpha_i \mathbf{S}$ where $\mathbf{U}_i | \mathbf{S} \sim \mathcal{N}(\alpha_i \mathbf{S}, P_i)$ and $\alpha_i = P_i/(\sum_{j=1}^{K} P_j + N)$. With this theoretical set-up, it is possible to reliably transmit all the messages together, with $W_i$ sent at rate $R_i = \frac{1}{2}\log_2\big(1 + P_i/(\sum_{j=1}^{i-1} P_j + N)\big)$. This rate is the maximal rate at which $W_i$ can be transmitted as long as the other messages $W_j$, $j \neq i$, are simultaneously transmitted at non zero rates. A scalar implementation of this ($K$ users) GMAC-

based joint DPC scheme consists in successively applying $K$ well designed SCSs. Equivalent channel for SCSi is $\mathbf{y}_{i,b} = \mathbf{y} - \sum_{j=i+1}^{K} \mathbf{u}_j$, which is the received signal assuming interference from only the (i-1) *before-hand* watermarks $\mathbf{x}_j$, $j < i$ and no *post-hand* interference from the remaining $(K-i)$ watermarks $\mathbf{x}_j$, $j > i$. We also denote by $\mathbf{y}_i \triangleq \mathbf{y}_{i,0} = \mathbf{x}_i + \mathbf{s} + \mathbf{z}$ the received signal assuming neither beforehand nor post-hand interferences. The set of feasible rates achieved by this practical coding can be obtained as a straightforward generalization of (5.20). The corresponding achievable rate region is given by the convex hull of all rate $K$-tuples $(\widetilde{R_1}, \ldots, \widetilde{R_K})$ simultaneously satisfying

$$\begin{aligned} \widetilde{R_i} &\leq \max_{\alpha_1 \in [0,1]} I\big(W_i; \mathcal{Q}_{\Delta_i}(\mathbf{y}_i) - \mathbf{y}_i\big), \;\; i = 1, 2, \ldots, K, \\ \sum_{j=1}^{K} \widetilde{R_j} &\leq \sum_{j=1}^{K} \max_{\alpha_j \in [0,1]} I\big(W_j; \mathcal{Q}_{\Delta_j}(\mathbf{y}_{j,b}) - \mathbf{y}_{j,b}\big). \end{aligned} \tag{5.25}$$

The maximum of the mutual information $I(W_i; \mathcal{Q}_{\Delta_i}(\mathbf{y}_i) - \mathbf{y}_i)$ is attained with the optimal choice of $\alpha_i \in [0,1]$ given by

$$\widetilde{\alpha_i} = \Big(1 - \sum_{j=i+1}^{K} \alpha_j\Big) \sqrt{\frac{P_i}{P_i + 2.71N}}, \;\; \text{with } \widetilde{\alpha_K} = \sqrt{\frac{P_K}{P_K + 2.71N}}.$$

### 5.5.3 Lattice-Based Codebooks for BC-Aware Multi-User Information Embedding

The gap to the ideal capacity region of the sample-wise joint scalar DPC practical capacity region shown in Fig. 5.6 can be partially bridged using structured finite-dimensional lattice-based codebooks. Lattices have been studied in [123] and considered for first time in the context of single-user watermarking in [115]. Consequent works [116, 117] extended these results to different scenarios. In what follows, only the required ingredients are briefly reviewed. The reader may refer to [124] for a full discussion.

Consider the transmission scheme depicted in Fig. 5.11 where $\Lambda$ is some $n$-dimensional lattice. This scheme is a generalization to the lattice codebook case of a slight variation of the scalar case considered in subsection 5.4.1 [12]. The function

[12] More precisely, this is a generalization to the lattice case of a DC-QIM based two users watermarking scheme. DC-QIM is considered because it is more convenient and also it has very close performance to SCS as has been reported in 5.2.2.

$\iota_1(.)$ is used for arbitrary mapping the set of indexes $W_1 \in \mathcal{M}_1 = \{1, \ldots, M_1\}$ to a certain set of vectors $\mathcal{C}_{w_1} = \{\mathbf{c}_{w_1} \ : \ w_1 = 1, \ldots, M_1\}$ to be specified in the sequel. The function $\iota_2(.)$ does similarly for the set of indexes $W_2 \in \mathcal{M}_2 = \{1, \ldots, M_2\}$. With respect to the scalar codebook case, $\mathcal{C}_{w_i}$, $i = 1, 2$, is a lattice codebook whose entries must be appropriately chosen so as to maximize the encoding performance. For each

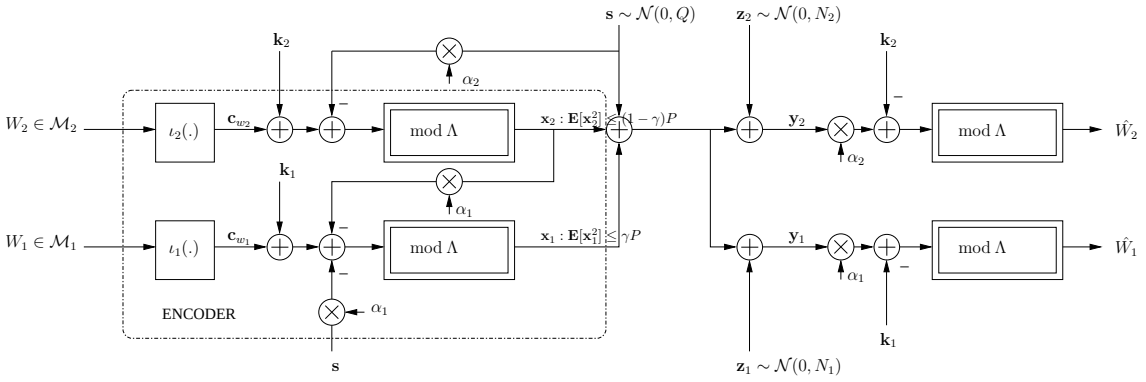



Figure 5.11: Lattice-based scheme for multiple information embedding over a Gaussian Broadcast Channel (GBC).

$W_i \in \mathcal{M}_i$, with $i = 1, 2$, the codeword $\iota_i(W_i) = \mathbf{c}_{w_i}$ is the *coset leader* of the coset $\Lambda_{w_i} = \mathbf{c}_{w_i} + \Lambda$ relative to the lattice $\Lambda$. The codebook $\mathcal{C}_{w_i}$ is shared between the encoder and the decoder $i$ and is assumed to be uniformly distributed over the fundamental cell $\mathcal{V}(\Lambda)$ of the lattice $\Lambda$. Also, we assume *common randomness*, meaning that the key $\mathbf{k}_i$, $i = 1, 2$, is known to both the encoder and the decoder $i$. Apart from obvious security purposes, these keys will turn out to be useful in attaining the capacity region.

In the following, we consider cover signal vectors (frames) of length $n$. Following (5.3), the encoding and decoding functions for the lattice-based joint DPC given by (5.5) and (5.10) write

$$\begin{aligned} \mathbf{x}_2(\mathbf{s}; W_2, \Lambda) &= (\mathbf{c}_{w_2} + \mathbf{k}_2 - \alpha_2 \mathbf{s}) \bmod \Lambda, \\ \mathbf{x}_1(\mathbf{s}; W_1, \Lambda) &= (\mathbf{c}_{w_1} + \mathbf{k}_1 - \alpha_1(\mathbf{s} + \mathbf{x}_2)) \bmod \Lambda, \\ \widehat{W_i} &= \text{argmin}_{W_i \in \mathcal{M}_i} \|(\alpha_i \mathbf{y}_i - \mathbf{k}_i - \mathbf{c}_{w_i}) \bmod \Lambda\|, \ \ i = 1, 2. \end{aligned} \tag{5.26}$$

The modulo reduction operation is defined as $\mathbf{x} \bmod \Lambda \triangleq \mathbf{x} - \mathcal{Q}_\Lambda(\mathbf{x}) \in \mathcal{V}(\Lambda)$ where the $n$-dimensional quantization operator $\mathcal{Q}_\Lambda(.)$ is such that quantization of $\mathbf{x} \in \mathbb{R}^n$ results in the closest lattice point $\boldsymbol{\lambda} \in \Lambda$ to $\mathbf{x}$.

We focus on the practically feasible rate region achieved by (5.26). To this end, we rely on a previous works relative to practical achievable rates with lattice codebooks in the context of a single-user watermark [115]. Here, the situation is different since two watermarks are concerned, but the key ideas remain the same. Thus, details are skipped and we only mention the key steps, in processing the received signals $\mathbf{y}_1$ and $\mathbf{y}_2$. Each of the channels $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is similar to the one in [115, 117], with however a different state information and channel noise. The establishment of the results below relies principally on the properties of a Modulo Lattice Additive Noise (MLAN) channel [125] and on the following two important properties of the mod-$\Lambda$ operation:

$$\text{(P1)} \quad \forall(\boldsymbol{\lambda}, \mathbf{a}) \in \Lambda \times \mathbb{R}^n, (\mathbf{a} + \mathbf{v} + \boldsymbol{\lambda}) \bmod \Lambda \ = (\mathbf{a} + \mathbf{v}) \bmod \Lambda. \tag{5.27a}$$

$$\text{(P2)} \quad \forall\, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2n}, ((\mathbf{x} \bmod \Lambda) + \mathbf{y}) \bmod \Lambda = (\mathbf{x} + \mathbf{y}) \bmod \Lambda. \tag{5.27b}$$

Upon reception of $\mathbf{y}_i$, $i = 1, 2$, "receiver i" computes the signal $\mathbf{r}_i = (\alpha_i \mathbf{y}_i - \mathbf{k}_i) \bmod \Lambda$. Using $(P_1)$ and $(P_2)$ and straightforward algebra calculations, it can be shown that

$$\mathbf{r}_1 = (\mathbf{c}_{w_1} + \alpha_1 \mathbf{z}_1 - (1 - \alpha_1)\mathbf{x}_1) \bmod \Lambda, \tag{5.28a}$$

$$\mathbf{r}_2 = (\mathbf{c}_{w_2} + \alpha_2(\mathbf{z}_2 + \mathbf{x}_1) - (1 - \alpha_2)\mathbf{x}_2) \bmod \Lambda. \tag{5.28b}$$

Hence, the "degraded user" (more noisy watermarked content) sees the equivalent channel noise $\widetilde{\mathbf{V}_2} = (\alpha_2(\mathbf{Z}_2 + \mathbf{X}_1) - (1 - \alpha_2)\mathbf{X}_2) \bmod \Lambda$ and the "better user" (less noisy watermarked content) sees the equivalent channel noise $\widetilde{\mathbf{V}_1} = (\alpha_1 \mathbf{Z}_1 - (1 - \alpha_1)\mathbf{X}_1) \bmod \Lambda$. Now, using the important *Inflated Lattice Lemma* reported in [126], $\mathbf{Y}_1$ and $\mathbf{Y}_2$ turn to be two MLAN channels with channel noises $\widetilde{\mathbf{V}_1}$ and $\widetilde{\mathbf{V}_2}$, respectively. The MLAN channel has been first considered in [127, 128]. It is shown that when modulo reduction is with respect to some lattice $\Lambda$ and when the channel noise $\mathbf{V}$ is i.i.d. Gaussian, capacity in bits per dimension can be written as

$$C(\Lambda) = \frac{1}{n} \left(\log_2(V(\Lambda)) - h(\mathbf{V})\right), \tag{5.29}$$

where $h(\cdot)$ denotes differential entropy. Hence, the practically achievable rates $R_1(\Lambda)$ and $R_2(\Lambda)$ are given by (5.29), with the channel noise $\mathbf{V}$ being replaced by $\widetilde{\mathbf{V}_1}$ and $\widetilde{\mathbf{V}_2}$, respectively. The maximally achievable rates are obtained by maximizing these expressions over $\alpha_1$ and $\alpha_2$, respectively. The corresponding achievable rate region $\bar{\mathscr{R}}_{\text{BC}}$ is given by

$$\bar{\mathscr{R}}_{\text{BC}}(P) = \bigcup_{0\le\gamma\le 1} \Big\{(\widetilde{R}_1, \widetilde{R}_2): \quad \widetilde{R}_1 \le \max_{\alpha_1\in[0,1]} \frac{1}{n}\Big(\log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_1}(\alpha_1,\gamma)\big)\Big), \\ \widetilde{R}_2 \le \max_{\alpha_2\in[0,1]} \frac{1}{n}\Big(\log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_2}(\alpha_2,\gamma)\big)\Big)\Big\}. \tag{5.30}$$

Note that from the right hand side term of (5.30), we have $\bar{\mathscr{R}}_{\text{BC}} \subset \mathcal{C}_{\text{BC}}$, where $\mathcal{C}_{\text{BC}}$ is the full capacity region of a Gaussian BC with state information at the encoder (5.9). In general no closed form of (5.30) can be derived and the optimal pair $(\alpha_1, \alpha_2)$ has to be computed numerically to evaluate the differential entropy $h(\widetilde{\mathbf{V}_i})$, $i = 1, 2$. However, closed form approximations can be found in some special situations as shown hereafter.

(i) As the dimensionality $n$ of the lattice goes to infinity, the PDFs of the noises $\widetilde{\mathbf{V}_1}$ and $\widetilde{\mathbf{V}_2}$ tend to Gaussian distributions as quantization errors with respect to this lattice. Consequently, the optimal inflation parameters $\alpha_1$ and $\alpha_2$ minimizing $h(\widetilde{\mathbf{V}_1})$ and $h(\widetilde{\mathbf{V}_2})$ are those which minimize the variances of $\widetilde{\mathbf{V}_1}$ and $\widetilde{\mathbf{V}_2}$, respectively. These are $\alpha_1 = \gamma P/(\gamma P + N_1)$ and $\alpha_2 = (1-\gamma)P/(P + N_2)$. The ideal capacity region is attained with such a choice.

(ii) For finite-dimension lattice reduction however, the PDFs of $\widetilde{\mathbf{V}_1}$ and $\widetilde{\mathbf{V}_2}$ are not strictly Gaussian, but rather the convolution of a Gaussian with a uniform distribution. The equality $(\alpha_1, \alpha_2) = \left(\frac{\gamma P}{\gamma P + N_1}, \frac{(1-\gamma)P}{N_2 + P}\right)$ does not hold strictly but remains a quite accurate approximation. Considering this approximation leads to $\mathbb{E}_{\widetilde{\mathbf{V}}_1}[\widetilde{\mathbf{V}}_1^2] = \alpha_1 N_1$ and $\mathbb{E}_{\widetilde{\mathbf{V}}_2}[\widetilde{\mathbf{V}}_2^2] = \alpha_2(N_2 + \gamma P)$. Now, given that[13] $h(\widetilde{\mathbf{V}_1}) \le \log(2\pi e \alpha_1 N_1)$ and $h(\widetilde{\mathbf{V}_2}) \le \log 2\pi e \alpha_2(N_2 + \gamma P)$, we get

$$R_1(\Lambda) \ge \frac{1}{n}\left(\frac{1}{2}\log\left(1 + \frac{\gamma P}{N_1}\right) - \frac{1}{2}\log 2\pi e G(\Lambda)\right), \tag{5.31a}$$

$$R_2(\Lambda) \ge \frac{1}{n}\left(\frac{1}{2}\log\left(1 + \frac{(1-\gamma)P}{N_2 + \gamma P}\right) - \frac{1}{2}\log 2\pi e G(\Lambda)\right). \tag{5.31b}$$

This means that by using appropriate lattices for modulo-reduction, we are able to make the gap to the full theoretical capacity region smaller then $\log 2\pi e G(\Lambda)$. This can be achieved by selecting lattices that have good quantization proper-

[13] This is because the normal distribution is the one that maximizes entropy for a given second moment.

ties. These are those for which the normalized second moment $G(\Lambda)$ approaches $1/2\pi e$.

The $n$-dimensional lattices considered for Monte-Carlo achievable rate region integration are summarized in table 5.1, together with their most important parameters. Achievable rate region curves in bits per dimension are plotted in Fig. 5.12(a) where

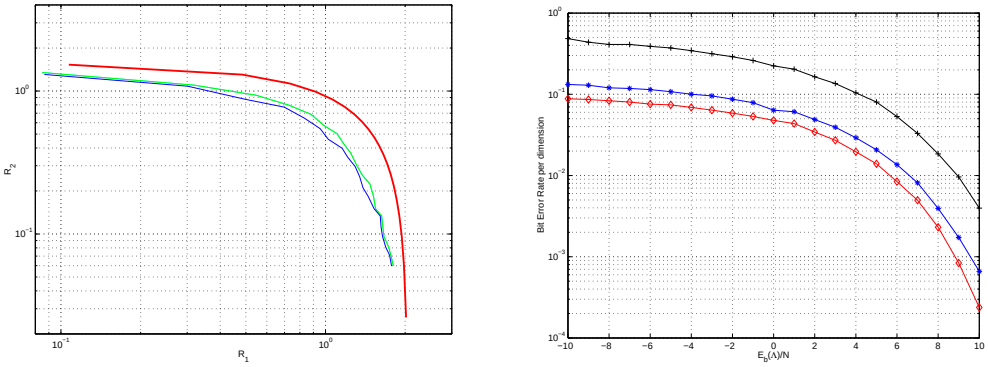
| Lattice | Name | $n$ | $G(\Lambda)$ | $\gamma_s(\Lambda)$ [dB] | $\gamma_s(\Lambda)$ [bit per dimension] |
|---|---|---|---|---|---|
| $\mathbb{Z}$ | Integer Lattice | 1 | $\frac{1}{12}$ | 0.00 | 0.000 |
| $A_2$ | Hexagonal Lattice | 2 | $\frac{5}{36\sqrt{3}}$ | 0.17 | 0.028 |
| $D_4$ | $4D$ Checkerboard L. | 4 | 0.0766 | 0.37 | 0.061 |

Table 5.1: Lattices with their important parameters

we observe that the use of the hexagonal lattice $A_2$, for example, enlarges the set of the rate pairs practically feasible, with respect to the scalar lattice $\mathbb{Z}$. Of course, this improvement goes along with a slight increase in computational cost. The same improvement can be observed through BER enhancement visible in Fig. 5.12(b). Note that Fig. 5.12(b) only shows the BER (against the per-bit per-dimension SNR $E_b(\Lambda)/N_1$) relative to the transmission of message $W_1$ with normalized rates. The BER curves corresponding to the transmission of message $W_2$ can be obtained by shifting to the right those of $W_1$ by the factor $\beta_{\text{BC}}(R_1, R_2) = \frac{R_1}{R_2} \times \frac{(1-\gamma)P}{\gamma P} \times \frac{N_1}{\gamma P + N_2}$ [dB].

### 5.5.4 Lattice-based codebooks for MAC-aware multi-user information embedding

The gap to the capacity region $\mathcal{C}_{\text{MAC}}$ (5.18) of the achievable rate region $\widetilde{R}_{\text{MAC}}$ (5.20) shown in Fig. 5.9 and corresponding to the sample-wise joint scalar DPC can be partially bridged using finite-dimensional lattice-based codebooks. The resulting transmission scheme is depicted in Fig. 5.13 where $\Lambda$ is some $n$-dimensional lattice. The functions $\iota_i(.)$, $i = 1, 2$ and the lattice codebooks $\mathcal{C}_{w_i}$, $i = 1, 2$ are defined in a similar way to that in the broadcast case addressed above. We focus on the improvement of the feasible rate pair $(R_1(\Lambda), R_2(\Lambda))$ brought by the use of the lattice codebooks $\mathcal{C}_{w_i}$, $i = 1, 2$, with comparison to the baseline scalar codebooks considered




(a) Achievable rate region with lattices $\mathbb{Z}$ and $A_2$.

(b) Bit Error Rates with lattices $\mathbb{Z}$ and $A_2$ and $D_4$.

Figure 5.12: Performance improvement in multiple user information embedding rates and BER due to the use of lattice codebooks. (a): achievable rate region for BC-like multiple user information embedding and (b): Corresponding BERs corresponding to the transmission of message $W_1$. From bottom to top: lattices Checkerboard $D_4$, Hexagonal $A_2$ and Cubic $\mathbb{Z}$.

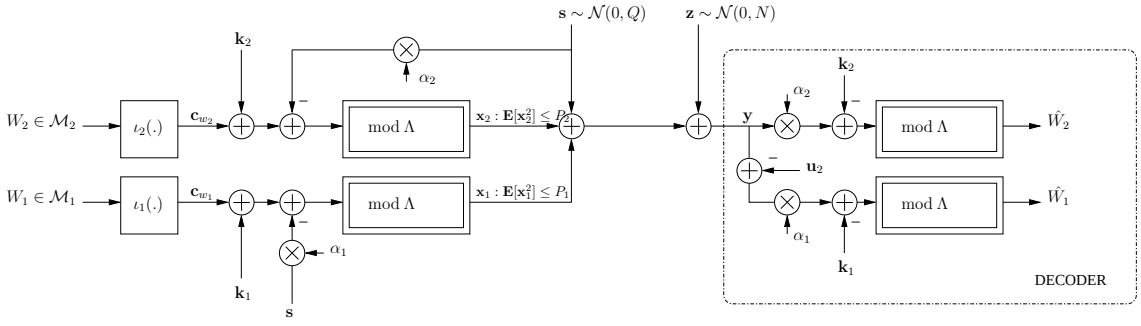



Figure 5.13: Lattice-based scheme for multiple information embedding over a Gaussian Multiple Access Channel (GMAC).

in subsection 5.4.2. Consider, for example, the corner point (B') of the capacity region shown in Fig. 5.9. The encoding and decoding of $W_1$ and $W_2$ are performed according to

$$\begin{aligned}
\mathbf{x}_1(\mathbf{s}; W_1, \Lambda) &= (\mathbf{c}_{w_1} + \mathbf{k}_1 - \alpha_1(1-\alpha_2)\mathbf{s}) \bmod \Lambda, \\
\mathbf{x}_2(\mathbf{s}; W_2, \Lambda) &= (\mathbf{c}_{w_2} + \mathbf{k}_2 - \alpha_2\mathbf{s}) \bmod \Lambda, \\
\widehat{W_1} &= \operatorname{argmin}_{W_1 \in \mathcal{M}_1} \|(\alpha_1\mathbf{y}_1 - \mathbf{k}_1 - \mathbf{c}_{w_1}) \bmod \Lambda\|, \\
\widehat{W_2} &= \operatorname{argmin}_{W_2 \in \mathcal{M}_2} \|(\alpha_2\mathbf{y} - \mathbf{k}_2 - \mathbf{c}_{w_2}) \bmod \Lambda\|.
\end{aligned} \tag{5.32}$$

where $\mathbf{y}_1 = \mathbf{y} - (\mathbf{x}_2 + \alpha_2\mathbf{s})$. Upon reception, the receiver first computes the error signal $\mathbf{r} = (\alpha\mathbf{y} - \mathbf{k}_2) \bmod \Lambda$. In a similar way to that for the broadcast case, it can be shown that $\mathbf{r} = (\mathbf{c}_{w_2} + \alpha_2(\mathbf{z} + \mathbf{x}_1) - (1-\alpha_2)\mathbf{x}_2) \bmod \Lambda$. Hence the equivalent channel for the transmission of $W_2$ is an MLAN channel with (Gaussian) channel noise $\widetilde{\mathbf{v}_2} = (\alpha_2(\mathbf{z} + \mathbf{x}_1) - (1-\alpha_2)\mathbf{x}_2) \bmod \Lambda$. Next, the receiver computes $\mathbf{r}_1 = (\alpha\mathbf{y}_1 - \mathbf{k}_1) \bmod \Lambda$, which can be shown to equal $(\mathbf{c}_{w_1} + \alpha_1\mathbf{z} - (1-\alpha_1)\mathbf{x}_1) \bmod \Lambda$, completely independent of $\mathbf{x}_2$. Hence the equivalent channel for the transmission of $W_1$ is another MLAN channel with (Gaussian) channel noise $\widetilde{\mathbf{v}_1} = (\alpha_1\mathbf{z} - (1-\alpha_1)\mathbf{x}_1) \mod \Lambda$. Consequently, by using (5.32) the achievable rate pair $(R_1(B'), R_2(B'))$ corresponding to the corner point (B') of the capacity region $\mathcal{C}_{\text{MAC}}$ is given by

$$R_1(B') = \max_{\alpha_1 \in [0,1]} \frac{1}{n} \Big( \log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_1}(\alpha_1, P_1)\big) \Big), \tag{5.33a}$$

$$R_2(B') = \max_{\alpha_2 \in [0,1]} \frac{1}{n} \Big( \log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_2}(\alpha_2, P_2)\big) \Big). \tag{5.33b}$$

Note that $(R_1, R_2) \in \mathcal{C}_{\text{MAC}}$. Similarly to the development made in the broadcast case, the achievable rate region by using the modulo reduction with respect to the lattice $\Lambda$ straightforwardly generalizes (5.20) and it is given by

$$\begin{aligned}
\bar{\mathscr{R}}_{\text{MAC}}(P_1, P_2) = \Big\{ (\widetilde{R}_1, \widetilde{R}_2) : \widetilde{R}_1 &\le \max_{\alpha_1 \in [0,1]} \frac{1}{n} \Big( \log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_1}(\alpha_1, P_1)\big) \Big), \\
\widetilde{R}_2 &\le \max_{\alpha_2 \in [0,1]} \frac{1}{n} \Big( \log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_2}(\alpha_2, P_2)\big) \Big), \\
\widetilde{R}_1 + \widetilde{R}_2 &\le \max_{\alpha_1 \in [0,1]} \frac{1}{n} \Big( \log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}_1}(\alpha_1, P_1)\big) \Big) \\
&\quad + \max_{\alpha_2 \in [0,1]} \frac{1}{n} \Big( \log_2(V(\Lambda)) - h\big(\widetilde{\mathbf{V}}(\alpha_2, P_2)\big) \Big) \Big\},
\end{aligned} \tag{5.34}$$

where $\widetilde{\mathbf{V}_i} = (\alpha_i\mathbf{Z} - (1-\alpha_i)\mathbf{X}_i) \bmod \Lambda$, $i = 1, 2$ and $\widetilde{\mathbf{V}} = (\alpha_2(\mathbf{Z} + \mathbf{X}_1) - (1-\alpha_2)\mathbf{X}_2) \bmod \Lambda$.

The improvement brought by lattice coding is illustrated in Fig. 5.12(b). The curves correspond to the transmission of message $W_1$. As in the broadcast case, the BER curves corresponding to the transmission of message $W_2$ can be obtained by translating to the right those of $W_1$, by $\beta_{\text{MAC}}(R_1, R_2) = \frac{R_1 P_2 N}{R_2 P_1 (N+P_1)}$[dB].

## 5.6 Summary

In this chapter, we investigated practical joint scalar schemes for multiple user information embedding. For instance, two different situations of embedding several messages into one common cover signal are considered. The first situation is recognized as being equivalent to communication over a Gaussian BC with state information non-causally known at the transmitter but not at the receivers. The second is argued as to be analog to communication over a Gaussian MAC with state information known non-causally at the transmitters but not at the receiver. Next, based on this equivalence with multi-user information theory, two practically feasible scalar schemes for simultaneously embedding two messages into the same host signal are proposed. These schemes carefully extend the initial QIM and SCS schemes, that were originally conceived for embedding one watermark, to the two-watermark case. The careful design concerns the joint encoding as well as the appropriate order needed so as to reliably embed the different watermarks. A central idea for the joint design is "awareness".

The improvement brought by this awareness is shown through comparison to the corresponding rather intuitive schemes, obtained through superimposition, as many times as needed, of the single user schemes QIM and SCS. Performance is analyzed in terms of both achievable rate region and BER. Finally, the proposed schemes are straightforwardly extended to the arbitrary number of watermarks case and also to the vector case through lattice-based codebooks. Results are supported by illustrative achievable rate region and BER curves obtained through Monte-Carlo integration and Monte-Carlo-simulation, respectively.

# Chapter 6

# Conclusions and Future Work

In this thesis we have studied the problem of reliable communication over single and multi-user wireless channels when the receiver(s) and the transmitter only know noisy estimates of the time-varying channel parameters. In particular, we established a fundamental connection between the more common technique to obtain receiver channel knowledge through use of pilot symbols and the notion of reliable communication under channel estimation errors. This connection for arbitrary channel estimators follows from the statistic of the channel estimation errors (CEE), i.e. the probability distribution function of the unknown channel given its estimate. Furthermore, it appears to be an effective way to introduce the imperfect channel knowledge in the capacity definition. We proposed to characterize the information theoretic limits of such scenarios in terms of two novel notions: the (i) *estimation-induced outage capacity* and (ii) the average (over all channel estimation errors) of the transmission error probability, which leads to the capacity of a composite (more noisy) channel.

With regards to the practical consequences of this research, many of these outcomes have been applied to develop practical coding schemes for applications like watermarking and the optimal design of decoders adapted to the CEE. All this leads to a number of results and still open questions in this thesis.

The transceiver in the *estimation-induced outage capacity* strives to construct codes for ensuring the desired communication service, i.e. for achieving target rates with small error probability, no matter which degree of accuracy estimation arises during a transmission. We proved a coding theorem and its strong converse that provides an explicit expression of the outage capacity within this constraint. This

capacity expression allows us to evaluate the trade-off between the maximal achievable outage rate (i.e. maximizing over all possible transmitter-receiver pairs) versus the outage probability (the QoS constraint). This trade-off can be used by a system designer to optimally share the available resources (e.g. power for transmission and training, number of feedback bits, the amount of training used, etc.), so that the communication requirements be satisfied.

Possibly straightforward applications of these results are practical time-varying systems with small training overhead and quality of service constraints. Particularly in mobile wireless environments where channels change rapidly, and as consequence it may not be feasible to obtain reliable estimation of the channel parameters. Another application scenario arises in the context of cellular coverage, where this capacity would characterize performance over multiple communication sessions of different users in a large number of geographic locations (cf. [85]). In that scenario based on our results, the system designer can ensure reliable communication for $(1 - \gamma_{QoS})$-percent of users during the connection session.

In addition to studying the capacity under the above mentioned constraints, we also considered the problem of reception in practical communication systems. Specifically, we focused on determining the optimal decoder that achieves the *estimation-induced outage capacity* for arbitrary DMCs. Inspired by the theoretical decoder that achieves the capacity we derived a practical decoding metric adapted to the channel estimation errors. Performances of this decoder in terms of achievable information rates and BER of iterative MIMO-BICM decoding were studied for the case of uncorrelated fading MIMO channels and compared to those of the classical mismatched ML decoding, which replaces the unknown channel by its estimate. Simulation results indicate that the mismatched ML decoding is sub-optimal compared to the proposed decoder under short training sequences, in terms of both BER and achievable information rates.

Although we showed that the proposed decoding metric outperforms classical mismatched approaches, this only achieves a lower bound of the *estimation-induced outage capacity*. This decoder ensures reliable communication for the average (over all CEE) of the transmission error probability, but it does not guarantee small error probabilities for every channel state in the optimal set of states maximizing the outage

capacity. In contrast, this decoder achieves the capacity of a composite (more noisy) channel. Nevertheless, different variations of the decoding metric incorporating not only the statistic of the channel estimates, but also the optimal set of states, have yet to be fully explored.

We also extensively investigated the problem of communicating reliably over imperfectly known channels with channel states non-causally known at the transmitter, which is of particular importance to increase data rates in next generation wireless systems. We addressed this, through the second notion of reliable communication based on the average of the transmission error probability over all CEE. This basically means that the transceiver does not require small instantaneous transmission error probabilities, but rather its average over all CEE must be arbitrary small. This notion enable us to easily extend existing capacity expressions that assume perfect channel knowledge to the more realistic case with imperfect channel estimation, transforming the mismatched scenario into composite (more noisy) state dependent channels. We also considered the natural extension of the Marton's region for arbitrary broadcast channels to the case with imperfect channel knowledge.

Two scenarios are studied: (i) the receiver(s) only has access to noisy estimates of the channel and these estimates are perfectly known at the transmitter and (ii) no channel information is available at the transmitter and imperfect information is available at the receiver(s). Then, we used the capacity expressions to derive achievable rates and optimal DPC schemes with Gaussian codebooks for the fading Costa's channel and the Fading MIMO-BC, assuming ML or MMSE channel estimation. Our results for downlink communications, are useful to assess the amount of training data to achieve target rates.

The somewhat unexpected result is that, while it is well-known that DPC for such class of channels requires perfect channel knowledge at both the transmitter and the receiver, without channel information at the transmitter, significant gains can be still achieved by using the proposed (adapted to the CEE) DPC scheme. Further numerical results in the context of uncorrelated fading show that, under the assumption of imperfect channel information at the receiver, the benefit of channel estimates known at the transmitter does not lead to large rate increases. The "close to optimal" DPC scheme used in this scenario (without knowledge of channel estimates)

follows as the average over all channel estimates of the optimal DPC scheme when the transmitter knows the estimates.

Obtaining receiver channel knowledge in practical communication systems is feasible through the use of a few number of pilot symbols, but transmitter channel knowledge generally requires feedback from the receivers. One surprising conclusion to be drawn from this research is that a BC with a single transmitter and receiver antenna and no channel information at the transmitter can still achieve significant gains compared to TDMA using the proposed DPC scheme. Furthermore, in this case the benefit of channel estimates known at the transmitter does not lead to large rate increases. However, we also showed that, for multiple antenna BCs, in order to achieve large gain rates compared with TDMA the transmitter requires the knowledge of all channel estimates, i.e., some feedback channel (perhaps rate-limited) must go from the receivers to the transmitter, conveying these channel estimates.

Interestedly, while it is well-known that for systems with many users significant gains can be achieved by adding base station antennas, under imperfect channel estimation, benefiting of a large number antennas requires very large amount of training and feedback channel. For practical multiple-antenna systems, this feedback may require substantial bandwidth and may in fact be difficult to obtain within a fast enough time scale, and consequently depending on the degree of accuracy channel estimation, this benefit may not hold.

This work establishes the bases for further research considering also the effects of rate-limited feedback channel that may provide the transmitter with degraded versions of the channel estimates at the receiver(s). Thus, it is of great interest to study the large gray area between the two extreme cases (i)-(ii), where the receivers dispose of imperfect channel estimation while the transmitter may (or not) know all these channel estimates. Future research directions may include, in addition to instantaneous information, information regarding the quality of channel estimates at the transmitter. For example, the pdf of the channel estimate (unknown at the transmitter) given its degraded (more noisy) estimate resulting of rate limited feedback, can be used to derive the optimal DPC in a similar manner as well as we did for the case (ii). Answering this and related questions will allow to better understand the benefit of adding multiple base station antennas in practical downlink systems.

In the final chapter of this thesis we studied the role of multi-user state dependent channels with non-causal channel state information at the transmitter in multi-user information embedding. We investigated practical joint scalar schemes for multiple user information embedding. For instance, two different situations of embedding several messages into one common cover signal are considered: (i) The first situation is recognized as being equivalent to communication over a Gaussian BC with state information non-causally known at the transmitter but not at the receivers and (ii) the second over a Gaussian MAC with state information known non-causally at the transmitters but not at the receiver.

Next, based on this equivalence with multi-user information theory, two practically feasible scalar schemes for simultaneously embedding two messages into the same host signal are proposed. These schemes extend the initial QIM and SCS schemes, that were originally conceived for embedding one watermark, to the two-watermark case. The careful design concerns the joint encoding as well as the appropriate order needed so as to reliably embed the different watermarks. The central idea for this joint design is "awareness". Performance is analyzed in terms of both achievable rate region and Bit Error Rate. Finally, the proposed schemes are straightforwardly extended to the arbitrary number of watermarks case and also to the vector case through lattice-based codebooks.

The notions of reliable communication studied in this thesis require complete knowledge of the statistics characterizing the channel variations (e.g. the pdf of the fading process). However, for certain scenarios this assumption may not hold, and consequently the statistic of the CEE (the pdf of the unknown channel given its estimate) cannot be computed. This leads to a different mathematical problem, which is connected with AVCs. Thus, it would be interesting as future work, to investigate this capacity with partial knowledge of the statistics characterizing the channel variations.

# Appendix A

# Information-typical Sets

Information divergence of probability distributions can be interpreted as a (non-symmetric) analogue of Euclidean distance [129]. With this interpretation, several results of these sequences are intuitive "information-typical sets" counterparts of standard "strong-typical sets" [3]. The definition of *I-typical* sets using the information divergence was first suggested by Csiszár and Narayan [130].

Throughout this appendix, we use the following notation: The empirical PM $\hat{P}_n$ associated a sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n$ is $\hat{P}_n(\mathbf{x}, \mathscr{A}) = N(\mathscr{A}|\mathbf{x})/n$ with $N(\mathscr{A}|\mathbf{x}) = \sum_{i=1}^{n} 1_{\mathscr{A}}(x_i)$, and $\hat{W}_n$ is the empirical transition PM associated with $\mathbf{x}$ and $\mathbf{y} = (y_1, \ldots, y_n) \in \mathscr{Y}^n$. The set $\mathscr{P}_n(\mathscr{X}) \subset \mathscr{P}(\mathscr{X})$ denotes the set of all rational point probability masses on $\mathscr{X}$, and its cardinality is bounded by $\|\mathscr{P}_n(\mathscr{X})\| \leq (1+n)^{|\mathscr{X}|}$ (cf. [17]). A function mapping $\theta \in \Theta \mapsto W(\cdot|\cdot, \theta) \in \mathscr{P}(\mathscr{Y})$ is a stochastic transition PM, i.e., for each $\theta \in \Theta$ this mapping defines a transition PM, and for every subset $\mathscr{B} \subset \mathscr{Y}$ the function mapping $\theta \mapsto W(\mathscr{B}|\cdot, \theta)$ is $\Theta$-measurable. We shall use the total variation or variational distance defined by $\mathcal{V}(P, Q) = 2 \sup_{\mathscr{A} \subseteq \mathscr{X}} |P(\mathscr{A}) - Q(\mathscr{A})|$, and its conditional version of Pinsker's inequality $\mathcal{V}(W \circ P, V \circ P) \leq \sqrt{\mathcal{D}(W\|V|P)/2}$ (cf. [17]). The *support* of a transition PM $W$ is the set $Supp(W) = \{b \in \mathscr{Y} : W(b|a) > 0 \text{ for all } P(a) > 0\}$. Given any set $\mathcal{W} \subset \mathscr{P}(\mathscr{Y})$, there is one PM that contains all the others supports and this will be called the *support* of $\mathcal{W}$, denoted $Supp(\mathcal{W})$. It follows that $\mathcal{D}(W\|V|P) < \infty$ iff $Supp(W) \subset Supp(V)$. Let $Q, P \in \mathscr{P}(\mathscr{X})$ be two PMs, then $Q$ is said to be absolutely continuous with respect to $P$, writes $Q \ll P$, if $Q(\mathscr{A}) = 0$ for every set $\mathscr{A} \subset \mathscr{X}$ for which $P(\mathscr{A}) = 0$.

## A.1 Definitions and Basic Properties

**Definition A.1.1** *For any PM $P \in \mathscr{P}_n(\mathscr{X})$, the set of all sequences $\mathbf{x} \in \mathscr{X}^n$ with* type *$P$ is defined by $\mathcal{T}_P^n = \big\{\mathbf{x} \in \mathscr{X}^n : \mathcal{D}(\hat{P}_n \| P) = 0\big\}$, where $\hat{P}_n(\mathbf{x}, \cdot)$ is the empirical probability.*

**Definition A.1.2** *For any PM $P \in \mathscr{P}(\mathscr{X})$, the set of all sequences $\mathbf{x} \in \mathscr{X}^n$ called* I-typical *with constant $\delta > 0$ is defined by $\mathcal{T}_P^n(\delta) = \big\{\mathbf{x} \in \mathscr{X}^n : \mathcal{D}(\hat{P}_n \| P) \leq \delta\big\}$, where $\hat{P}_n(\mathbf{x}, \cdot)$ is the empirical probability, such that $\hat{P}_n(\mathbf{x}, \cdot) \ll P$.*

**Definition A.1.3** *For any transition PM $W(\cdot|x) \in \mathscr{P}(\mathscr{Y})$, the set of all sequences $\mathbf{y} \in \mathscr{Y}^n$ under the condition $\mathbf{x} \in \mathscr{X}$ called* conditional I-typical *with constant $\delta > 0$ is defined by $\mathcal{T}_W^n(\mathbf{x}, \delta) = \big\{\mathbf{y} \in \mathscr{Y}^n : \mathcal{D}(\hat{W}_n \| W|\hat{P}_n) \leq \delta\big\}$, where $\hat{W}_n(b|a)N(a|\mathbf{x}) = N(a, b|\mathbf{x}, \mathbf{y})$ is the transition empirical probability, such that $\hat{W}_n(\cdot|a) \ll W(\cdot|a)$ for each $a \in \mathscr{X}$.*

**Lemma A.1.1** *(Uniform continuity of the entropy function) Let $P, Q \in \mathscr{P}(\mathscr{X})$ be PMs and $V(\cdot|x), W(\cdot|x) \in \mathscr{P}(\mathscr{Y})$ be two transition PMs. Then*

$$\begin{aligned}
&\textit{(i) If } \mathcal{V}(P, Q) \leq \Theta \leq 1/2, && \Rightarrow \ \big|H(P) - H(Q)\big| \leq -\Theta \log \frac{\Theta}{|\mathscr{X}|}. \\
&\textit{(ii) If } \mathcal{V}(V \circ P, W \circ P) \leq \Theta \leq 1/2, && \Rightarrow \ \big|H(V|P) - H(W|P)\big| \leq -\Theta \log \frac{\Theta}{|\mathscr{X}||\mathscr{Y}|}.
\end{aligned}$$

See Lemma 1.2.7 in [17].

**Proposition A.1.1** (Properties of I-typical sequences)

*(i) Any sequence $\mathbf{x} \in \mathcal{T}_P^n(\delta)$ implies $\mathcal{V}\big(\hat{P}_n(\mathbf{x}, \cdot), P\big) \leq \sqrt{\delta/2}$. Moreover any sequence $\mathbf{y} \in \mathcal{T}_W^n(\mathbf{x}, \delta)$ implies $\mathcal{V}(\hat{W}_n \circ \hat{P}_n, W \circ \hat{P}_n) \leq \sqrt{\delta/2}$ for all $\mathbf{x} \in \mathscr{X}^n$.*

*(ii) There exists sequences $(\delta_n)_{n \in \mathbb{N}_+}$ and $(\delta'_n)_{n \in \mathbb{N}_+}$ in $\mathbb{R}_+$ with $(\delta_n, \delta'_n) \to 0$ and $n \log^{-1}(n+1) \to \infty$ as $n \to \infty$, depending only on $|\mathscr{X}|$ and $|\mathscr{Y}|$ so that for every PM $P \in \mathscr{P}(\mathscr{X})$ and transition PM $W(\cdot|x) \in \mathscr{P}(\mathscr{Y})$, $P^n\big(\mathcal{T}_P^n(\delta_n)\big) > 1 - \epsilon_n$ and $W^n\big(\mathcal{T}_W^n(\delta'_n)|\mathbf{x}\big) > 1 - \epsilon'_n$, with*

$$\begin{aligned}
\epsilon_n &= \exp\big\{ - n\big(\delta_n - n^{-1}|\mathscr{X}| \log(n+1)\big)\big\}, \\
\epsilon'_n &= \exp\big\{ - n\big(\delta'_n - n^{-1}|\mathscr{X}\|\mathscr{Y}| \log(n+1)\big)\big\}.
\end{aligned}$$

*Note that $\log(n+1) < \sqrt{n}$ and consequently these sequences vent to zero with a convergence rate smaller than that obtained for strong typical sets [3].*

*(iii) For any PMs $P,Q \in \mathscr{P}(\mathscr{X})$ and transition PMs $W(\cdot|x), V(\cdot|x) \in \mathscr{P}(\mathscr{Y})$ and $\delta > 0$*

$$\begin{aligned} &\textit{If}\ \ \mathcal{D}(Q\|P) \le \delta \quad &\Rightarrow\ \ &|H(Q) - H(P)| \le -\sqrt{\delta/2}\log\frac{\sqrt{\delta/2}}{|\mathscr{X}|}. \\ &\textit{If}\ \ \mathcal{D}(W\|V|P) \le \delta \ &\Rightarrow\ \ &|H(W|P) - H(V|P)| \le -\sqrt{\delta/2}\log\frac{\sqrt{\delta/2}}{|\mathscr{X}\|\mathscr{Y}|}. \end{aligned}$$

*(iv) There exists sequences $(\epsilon_n)_{n\in\mathbb{N}_+}$ and $(\epsilon'_n)_{n\in\mathbb{N}_+}$ in $\mathbb{R}_+$ with $(\epsilon_n, \epsilon'_n) \to 0$ depending only on $|\mathscr{X}|$ and $|\mathscr{Y}|$ so that for every PM $P \in \mathscr{P}(\mathscr{X})$ and transition PM $W(\cdot|x) \in \mathscr{P}(\mathscr{Y})$*

$$\begin{aligned} &\Big|\frac{1}{n}\log|\mathcal{T}^n_P(\delta_n)| - H(P)\Big| \le \epsilon_n, \\ &\Big|\frac{1}{n}\log|\mathcal{T}^n_W(\mathbf{x},\delta'_n)| - H(W|P)\Big| \le \epsilon'_n, \quad \textit{for every } \mathbf{x} \in \mathcal{T}^n_P(\delta_n). \end{aligned}$$

***Proof:*** *Assertion (i) immediately follows from Pinsker's inequality. Assertion (iii) follows from (i) and the uniform continuity Lemma A.1.1 of the entropy function. Assertion (iv) immediately follows by defining I-typical sets using $(\delta_n, \delta'_n)$ sequences and from the claim (iii), i.e. $\mathcal{D}(\hat{P}_n\|P) \le \delta_n$ and $\mathcal{D}(\hat{W}_n\|W|\hat{P}_n) \le \delta'_n$, where the existence of such sequences was proved in the claim (ii). For the claim (ii) it is sufficient to prove the second assertion*

$$\begin{aligned} W^n\big([\mathcal{T}^n_W(\mathbf{x},\delta'_n)]^c|\mathbf{x}\big) &= \sum_{V_n:\mathcal{D}(V_n\|W|\hat{P}_n)>\delta'_n} W^n\big(\mathcal{T}^n_{V_n}(\mathbf{x})|\mathbf{x}\big) \\ &\le \sum_{V_n:\mathcal{D}(V_n\|W|\hat{P}_n)>\delta'_n} \exp(-n\mathcal{D}(V_n\|W|\hat{P}_n)) \\ &\le (1+n)^{|\mathscr{X}\|\mathscr{Y}|}\exp(-n\delta'_n) \\ &= \exp\big\{-n\big(\delta'_n - n^{-1}|\mathscr{X}\|\mathscr{Y}|\log(n+1)\big)\big\}. \end{aligned}$$ ∎

**Lemma A.1.2** *(*Uniform continuity of I-divergences*)*

*(i) For any transition PMs $W(\cdot|x), V(\cdot|x), Z(\cdot|x) \in \mathscr{P}(\mathscr{Y})$ and a PM $P \in \mathscr{P}(\mathscr{X})$, such that $\mathcal{D}(Z\|W|P) \le \epsilon$ for some $\epsilon > 0$. Then there exists $\delta > 0$ such that $|\mathcal{D}(Z\|V|P) - \mathcal{D}(W\|V|P)| \le \delta$ and $\delta \to 0$ as $\epsilon \to 0$, with $\delta = -\sqrt{\epsilon/2}\log\big(\sqrt{\epsilon/2}/(|\mathscr{X}||\mathscr{Y}|^2)\big)$.*

*(ii) Similarly for $P,Q,Z \in \mathscr{P}(\mathscr{X})$ such that $\mathcal{D}(Z\|Q) \le \epsilon$ for some $\epsilon > 0$. Then there exists $\delta' > 0$ such that $|\mathcal{D}(Z\|P) - \mathcal{D}(Q\|P)| \le \delta'$ and $\delta' \to 0$ as $\epsilon \to 0$, with $\delta' = -\sqrt{\epsilon/2}\log\big(\sqrt{\epsilon/2}/|\mathscr{X}|^2\big)$.*

***Proof:*** *We only prove the first statement, since (ii) follows immediately. Observe that from Proposition A.1.1 (i) and Lemma A.1.1 we have that* $\mathcal{D}(Z\|W|P) \leq \epsilon$ *implies* $|H(V|P) - H(W|P)| \leq -\sqrt{\epsilon/2}\log\frac{\sqrt{\epsilon/2}}{|\mathscr{X}||\mathscr{Y}|}$. *By considering the following inequalities:*

$$\begin{aligned}
|\mathcal{D}(Z\|V|P) - \mathcal{D}(W\|V|P)| &\leq |H(V|P) - H(W|P)| \\
&+ \sum_{a\in\mathscr{X}}\sum_{b\in\mathscr{Y}} P(a)|W(b|a) - V(b|a)|\log|\mathscr{Y}| \\
&\leq -\sqrt{\epsilon/2}\log\big(\sqrt{\epsilon/2}/(|\mathscr{X}||\mathscr{Y}|)\big) + \sqrt{\epsilon/2}\log|\mathscr{Y}| \\
&= \delta.
\end{aligned}$$

■

**Lemma A.1.3 (*Large probability of I-typical sets*)** *Let* $\mathcal{T}_P^n(\delta)$ *and* $\mathcal{T}_W^n(\mathbf{x},\delta)$ *be an I-typical and conditional I-typical sets, respectively. The probability that a sequence does not belong to these sets vent to zero, i.e.*

$$\begin{aligned}
\lim_{n\to\infty} P^n\Big([\mathcal{T}_P^n(\delta)]^c\Big) &= 0, \\
\lim_{n\to\infty} W^n\Big([\mathcal{T}_W^n(\mathbf{x},\delta)]^c|\mathbf{x}\Big) &= 0.
\end{aligned}$$

*Furthermore,* $\mathcal{D}(\hat{P}_n||P) \to 0$ *and* $\mathcal{D}(\hat{W}_n||W|\hat{P}_n) \to 0$ *with probability* 1 *with* $n \to \infty$.

***Proof:*** *We observe from assertion (ii)*

$$W^n\big(\big\{\mathbf{y}\in\mathscr{Y}^n : \mathcal{D}(\hat{W}_n\|W|\hat{P}_n)\big) > \delta\big\}\big|\mathbf{x}\big) \leq \exp\big[-n\big(\delta - n^{-1}|\mathscr{X}\|\mathscr{Y}|\log(n+1)\big)\big],$$

*for every* $\mathbf{x}\in\mathcal{T}_P^n(\delta)$, *and then it expression goes to zero as* $n\to\infty$. *The second assertion follows from the fact that,* $\sum_{n=1}^{\infty}\Pr\big(\big\{\mathcal{D}(\hat{W}_n\|W|\hat{P}_n) > \delta\big\}|\mathbf{x}\big) < \infty$, *and by applying Borel-Cantelli Lemma [131], we obtain* $\Pr\Big(\limsup_{n\to\infty}\big\{\mathcal{D}(\hat{W}_n\|W|\hat{P}_n) > \delta\big\}\big|\mathbf{x}\Big) = 0$. *This concludes the proof, since this holds for every* $\delta > 0$. ■

**Lemma A.1.4** *Given* $0 < \eta < 1$, *and PMs* $W(\cdot|x,\theta) \in \mathscr{P}(\mathscr{Y})$ *with* $\theta\in\Theta$ *and* $P\in\mathscr{P}(\mathscr{X})$. *Let* $\Lambda\subset\Theta$ *be a set of parameters, then there exists sequences* $(\epsilon_n)_{n\in\mathbb{N}_+}$ *and* $(\epsilon_n')_{n\in\mathbb{N}_+}$ *in* $\mathbb{R}_+$ *with* $(\epsilon_n,\epsilon_n')\to 0$ *depending only on* $|\mathscr{X}|$, $|\mathscr{Y}|$ *and* $\eta$, *so that:*
*(i) If* $\mathscr{A}^n\subset\mathscr{X}$, $\inf_{\theta\in\Lambda} W_\theta P^n(\mathscr{A}) \geq \eta$, *then* $\frac{1}{n}\log\|\mathscr{A}^n\| \geq \sup_{\theta\in\Lambda} H\big(W_\theta P\big) - \epsilon_n$.
*(ii) If* $\mathscr{B}^n\subset\mathscr{Y}$, $\inf_{\theta\in\Lambda} W^n(\mathscr{B}|\mathbf{x},\theta) \geq \eta$, *then* $\frac{1}{n}\log\|\mathscr{B}^n\| \geq \sup_{\theta\in\Lambda} H\big(W(\cdot|\cdot,\theta)|P\big) - \epsilon_n'$, *for any* $\mathbf{x}\in\mathcal{T}_P^n(\delta_n)$.

This Lemma simply follows from the proof of Corollary 1.2.14 in [17] and previous lemmas.

## A.2 Auxiliary results

This appendix introduces a few concepts shedding more light on the encoder and decoder required to achieve outage rates and furthermore provides some auxiliary technical results required for the formal proof of Theorem 2.2.1 in Section 2.3.

*Unfeasibility of Mismatched Typical Decoding:* Consider a DMC $W(\cdot|x,\theta) \in \mathcal{W}_\Theta$ and its (noisy) estimate $V(\cdot|x) = W(\cdot|x,\hat{\theta}) \in \mathcal{W}_\Theta$. The following Lemma proves that typical set decoding based on $V$ leads to a block-error probability that approaches one when the channel is not perfectly known ($W \neq V$).

**Lemma A.2.1** *Consider two channels $W(\cdot|x), V(\cdot|x) \in \mathcal{W}_\Theta$ such that $\mathcal{D}(W\|V|P) > \xi > 0$ for any input distribution $P$ and let $\mathcal{T}_W^n(\mathbf{x},\delta_n)$, $\mathcal{T}_V^n(\mathbf{x},\delta_n) \subset \mathscr{Y}^n$ denote two associated conditional I-typical sets for arbitrary $\mathbf{x} \in \mathcal{T}_P^n(\delta_n)$. Then, (i) there exists an index $n_0 \in \mathbb{N}_+$ such that for $n \geq n_0$ the conditional I-typical sets $\mathcal{T}_W^n(\mathbf{x},\delta_n)$ and $\mathcal{T}_V^n(\mathbf{x},\delta_n)$ are disjoint, i.e. $\mathcal{T}_W^n(\mathbf{x},\delta_n) \cap \mathcal{T}_V^n(\mathbf{x},\delta_n) = \emptyset$; (ii) the $W$-probability of $\mathcal{T}_V^n(\mathbf{x},\delta_n)$ converges to zero, $\lim_{n\to\infty} W^n\big(\mathcal{T}_V^n(\mathbf{x},\delta_n)\big|\mathbf{x}\big) = 0$; (iii) furthermore, $\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) \to \mathcal{D}(W\|V|P)$ with probability 1.*

Results (i) and (ii) reveal that the standard concept of typical sequences (respect to $V$) merely specifies some local structure in a small neighborhood of $V(\cdot|x)$ but not in the whole space (as outlined in [132]). In other words, this standard concept should be useful only to decode over perfectly known channels. However, this does not establish that any decoder based on method of types is not useful to decode on estimated channels. This only shows that for any $0 < \epsilon < 1$, there is no exists decoding sets $\{\mathscr{D}_i^n\}$ with $\mathscr{D}_i^n \subseteq \mathcal{T}_V^n(\mathbf{x}_i,\delta_n)$ associated to codewords $\{\mathbf{x}_i\} \subseteq \mathcal{T}_P^n(\delta_n)$, such that $W^n(\mathscr{D}_i^n|\mathbf{x}_i) > 1-\epsilon$ for all $n \geq n_0$.

***Proof:*** *In order to prove (i) we must show that for every $\xi > 0$ with $W(\cdot|x), V(\cdot|x)$ and $P$ verifying $\mathcal{D}(W\|V|P) > \xi$, with the assumption that $\mathcal{D}(\hat{W}_n\|W|\hat{P}_n) \leq \delta_n$ (using $\delta$-sequences). Then, there exists $n_0 = n_0(|\mathscr{X}|,|\mathscr{Y}|,\delta_n,\xi) \in \mathbb{N}_+$ such that $\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) > \delta_n$ for all $n \geq n_0$. To this end, we know from Lemma A.1.2 that $\mathcal{D}(\hat{W}_n\|W|\hat{P}_n) \leq \delta_n$ implies $|\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) - \mathcal{D}(W\|V|P)| \leq \delta_n'$, with $\delta_n' = -\sqrt{\delta_n/2}\log\big(\sqrt{\delta_n/2}/(|\mathscr{X}||\mathscr{Y}|^3)\big)$. We have also used the fact that $|\mathcal{D}(W\|V|\hat{P}_n) - \mathcal{D}(W\|V|P)| \leq \sqrt{2\delta_n}\log|\mathscr{Y}|$ for sufficiently large $n$, with $\mathcal{D}(\hat{P}_n\|P) \leq \delta_n$. As a result*

$\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) \geq \mathcal{D}(W\|V|P) - \delta'_n > \xi - \delta'_n$, *since there exits* $n_0 = n_0(|\mathscr{X}|, |\mathscr{Y}|, \delta_n, \xi) \in \mathbb{N}_+$ *such that* $\xi - \delta'_n > \delta_n$ *for all* $n \geq n_0$, *and* $(\delta_n, \delta'_n) \to 0$ *as* $n \to \infty$. *In particular, this is also possible for any* $\xi > 0$, *concluding the proof of (i). We now prove the assertion (ii),*

$$\begin{aligned} W^n\big(\mathcal{T}_V^n(\mathbf{x},\delta)\big|\mathbf{x}\big) &= \sum_{Z_n:\mathcal{D}(Z_n\|V|\hat{P}_n)\leq\delta} W^n\big(\mathcal{T}_{Z_n}^n(\mathbf{x})\big) \\ &\leq \sum_{Z_n:\mathcal{D}(Z_n\|W|\hat{P}_n)\leq\delta} \exp\big(-n\mathcal{D}(Z_n\|W|\hat{P}_n)\big) \\ &\overset{(a)}{\leq} \sum_{Z_n\in\mathscr{P}_n(\mathscr{Y})} \exp(-n\delta) \\ &\leq \exp\big\{-n(\delta - n^{-1}|\mathscr{X}\|\mathscr{Y}|\log(n+1))\big\}, \end{aligned} \tag{A.1}$$

*where (a) follows from assertion (i) which proves that* $\mathcal{D}(Z_n\|W|\hat{P}_n) \leq \delta$ *and* $\mathcal{D}(W\|V|P) > \delta$ *imply* $\mathcal{D}(Z_n\|V|\hat{P}_n) > \delta$ *for all* $n \geq n_0$. *For this reason if* $\mathcal{D}(Z_n\|V|\hat{P}_n) \leq \delta$ *then* $\mathcal{D}(Z_n\|W|\hat{P}_n) > \delta$ *and* $\mathcal{D}(W\|V|P) \leq \xi$. *Finally, we now prove assertion (iii). From continuity Lemma A.1.2 we can assert that there exists* $n_0 \in \mathbb{N}_+$ *such if* $\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) \leq \delta$ *then* $|\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) - \mathcal{D}(W\|V|P)| \leq \eta$. *Whereas, it also implies that for an arbitrary* $\eta > 0$ *there exits* $n_0 \in \mathbb{N}_+$ *and some* $\delta > 0$ *such if* $|\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) - \mathcal{D}(W\|V|P)| > \eta$ *then* $\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) > \delta$. *Now apply this relation in order to bound the following probability:* $\Pr\big(\big\{|\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) - \mathcal{D}(W\|V|P)| > \eta\big\}\big|\mathbf{x}\big) \leq \exp\big\{-n(\delta - n^{-1}|\mathscr{X}\|\mathscr{Y}|\log(n+1))\big\}$ *for any* $n \geq n_0$. *Thus,* $\sum_{n=n_0}^{\infty} \Pr\big(\big\{|\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) - \mathcal{D}(W\|V|P)| > \eta\big\}\big|\mathbf{x}\big)$ *converges for each* $\eta > 0$, *and the proof is concluded by applying Borel-Cantelli Lemma [131].* ■

*Robust Decoders:* Let $\mathscr{A}^n \subset \mathscr{X}^n$ denote a set of transmit sequences and let $W_\theta(\cdot|x) = W(\cdot|x,\theta)$. A set $\mathscr{B}^n \subset \mathscr{Y}^n$ (depending on $\Lambda \subset \Theta$) is called a *robust* $\epsilon$-*decoding set* for a sequence $\mathbf{x} \in \mathscr{A}^n$ and an unknown DMC $W(\cdot|x,\theta) \in \mathcal{W}_\Theta$, if the conditional (w.r.t. $\hat{\theta}$) probability of all $\theta$, for which the $W^n(\cdot|\mathbf{x},\theta)$-probability of $\mathscr{B}^n$ exceeds $1-\epsilon$, is at least $1-\gamma_{QoS}$, i.e., $\Pr\big(W^n(\mathscr{B}^n|\mathbf{x},\theta) > 1-\epsilon\big|\hat{\theta}\big) \geq 1-\gamma_{QoS}$.

A set $\mathscr{B}^n \subset \mathscr{Y}^n$ of received sequences is called a *common* $\eta$-*image* $(0 < \eta \leq 1)$ of a transmit set $\mathscr{A}^n \subset \mathscr{X}^n$ for the collection of DMCs $\mathcal{W}_\Lambda$, iff $\inf_{\theta\in\Lambda} W^n(\mathscr{B}^n|\mathbf{x},\theta) \geq \eta$ for all $\mathbf{x} \in \mathscr{A}^n$.

Finally, $\Lambda \subset \Theta$ is called a *confidence set* for $\theta$ given $\hat{\theta}$, if $\Pr(\theta \notin \Lambda|\hat{\theta}) < \gamma_{QoS}$ where $\gamma_{QoS}$ represents the confidence level.

**Proposition A.2.1** *If $\Lambda$ is a confidence set with confidence level $\gamma_{QoS}$ and $\mathscr{B}^n$ is a common $\eta$-image for the associated collection of DMCs, then $\mathscr{B}^n$ is also a robust $\epsilon$-decoding set with $\epsilon = 1 - \eta$.*

The statement follows from the fact that any transition PM is $\Theta$-measurable and from basic properties of measurable functions (see [131, p. 185]).

*Robust I-Typical Sets:* We next elaborate the explicit construction of robust $\epsilon$-decoding sets by introducing the concept of *robust I-typical sets*. A robust I-typical set is defined as

$$\mathscr{B}^n_\Lambda(\mathbf{x}, \delta_n) = \bigcup_{\theta \in \Lambda} \mathcal{T}^n_{W_\theta}(\mathbf{x}, \delta_n),$$

with arbitrary $\Lambda \subset \Theta$ and $\delta$-sequence $\{\delta_n\}$.

The next result provides a relation of robust I-typical sets and robust $\epsilon$-decoding sets.

**Lemma A.2.2** *For any $0 < \gamma_{QoS}, \epsilon < 1$, a necessary and sufficient condition for a robust I-typical set $\mathscr{B}^n_\Lambda(\mathbf{x}, \theta)$ to be a robust $\epsilon$-decoding set with probability $1 - \gamma_{QoS}$ is that $\Lambda$ be a confidence set.*

***Proof:*** *We start proving the necessary part of this condition, namely $\Pr\left(\Lambda|\hat{\theta}\right) \geq 1 - \gamma_{QoS}$ implies $\Pr\left(W^n(\mathscr{B}^n_\Lambda|\mathbf{x}, \theta) > 1 - \epsilon \middle| \hat{\theta}\right) \geq 1 - \gamma_{QoS}$. It straightforwardly show that $\mathscr{B}^n_\Lambda(\mathbf{x}, \delta_n)$ is a common $\eta$-image for the collection of DMCs $\mathcal{W}_\Lambda$ with $\eta = 1 - \epsilon$ (see Proposition A.1.1-ii). Hence, the necessity is a direct consequence of Proposition A.2.1. Now prove the sufficiency condition. To this end, we will show that if $\Pr\left(\theta \notin \Lambda|\hat{\theta}\right) \geq 1 - \gamma_{QoS}$ then $\Pr\left(W^n(\mathscr{B}^n_\Lambda|\mathbf{x}, \theta) > 1 - \epsilon \middle| \hat{\theta}\right) < \gamma_{QoS}$. As a consequence of this assumption, we have $\Pr\left(\mathcal{D}(V\|W_\theta|P) \neq 0\right) \geq 1 - \gamma_{QoS}$ for all transition PM $V(\cdot|x) \in \mathcal{W}_\Lambda$ (with $V \neq W_\theta$), where we have used the uniform continuity of information divergences. This implies that for each $V(\cdot|x) \in \mathcal{W}_\Lambda$ there exists $\xi > 0$ such that $\Pr\left(\mathcal{D}(V\|W_\theta|P) > \xi\right) \geq 1 - \gamma_{QoS}$. Therefore from Lemma A.2.1 (i), there exists $n_0 \in \mathbb{N}_+$ such that $\mathcal{T}^n_V(\mathbf{x}, \delta_n) \cap \mathcal{T}^n_{W_\theta}(\mathbf{x}, \delta_n) = \emptyset$ with probability $1 - \gamma_{QoS}$, for $\delta_n > 0$ and all $n \geq n_0$. Consequently, there exists also $n_0' \in \mathbb{N}_+$ such that $W^n\left(\mathscr{B}^n_\Lambda|\mathbf{x}, \theta\right) \leq W^n\left([\mathcal{T}^n_{W_\theta}]^c\middle|\mathbf{x}, \theta\right)$ with probability $1 - \gamma_{QoS}$, for all $n \geq n_0'$. Finally as above, this and Proposition A.1.1-(ii) imply for sufficiently "n" large, $\Pr\left(W^n(\mathscr{B}^n_\Lambda|\mathbf{x}, \theta) \leq \epsilon \middle| \hat{\theta}\right) \geq 1 - \gamma_{QoS}$, concluding the proof.* ∎

**Theorem A.2.1 (Cardinality of robust I-typical sets)** *For any collection of DMCs $\mathcal{W}_\Lambda$ and associated robust I-typical set $\mathscr{B}_\Lambda^n(\mathbf{x},\delta_n)$ with $\mathbf{x}\in\mathcal{T}_P^n(\delta_n)$, there exists an index $n_0$ such that for all $n\geq n_0$ the size $\|\mathscr{B}_\Lambda^n(\mathbf{x},\delta_n)\|$ of the robust I-typical set is bounded as follows:*

$$\left|\frac{1}{n}\log\|\mathscr{B}_\Lambda^n(\mathbf{x},\delta_n)\|-H(\mathcal{W}_\Lambda|P)\right|\leq\eta_n.$$

*Here, $H(\mathcal{W}_\Lambda|P)=\sup\limits_{V\in\mathcal{W}_\Lambda}H(V|P)$ and $\eta_n\to 0$ as $\delta_n\to 0$ and $n\to\infty$.*

The quantity $H(\mathcal{W}_\Lambda|P)$ may be interpreted as the conditional entropy of the set $\mathcal{W}_\Lambda$ and can be shown to equal the I-projection [129] of the uniform distribution on $\mathcal{W}_\Lambda$.

**Corollary A.2.1** *Assume same assumptions made in Theorem A.2.1, then*

$$\lim_{n\to\infty}\|\mathscr{B}_\Lambda^n(\mathbf{x},\delta_n)\|=H(\mathcal{W}_\Lambda|P),$$

*for every sequence $\mathbf{x}\in\mathcal{T}_P^n(\delta_n)$.*

Before proving Theorem A.2.1, we need the following result.

**Theorem A.2.2** *Consider any arbitrary set $\mathcal{W}\subset\mathscr{P}(\mathscr{Y})$ of transition PMs, and a set of sequences $\mathscr{B}_\mathcal{W}^n\subset\mathscr{Y}^n$ defined by $\mathscr{B}_\Sigma^n(\mathbf{x})=\bigcup\limits_{W\in\Sigma}\mathcal{T}_W^n(\mathbf{x})$ for every $\mathbf{x}\in\mathscr{X}^n$, where $\Sigma=\mathcal{W}\cap\mathscr{P}_n(\mathscr{Y})$. Then, the size of $\mathscr{B}_\mathcal{W}^n(\mathbf{x})$ is bounded by*

$$\left|\frac{1}{n}\log\|\mathscr{B}_\mathcal{W}^n(\mathbf{x})\|-\max_{W\in\Sigma}H(W|\hat{P}_n(\cdot|\mathbf{x}))\right|\leq|\mathscr{X}|\mathscr{Y}|n^{-1}\log(1+n).$$

*Furthermore, if the set $\mathcal{W}$ is convex then the upper bound can be replaced by $\|\mathscr{B}_\Sigma^n(\mathbf{x})\|\leq\exp\left\{n\max\limits_{W\in\Sigma}H(W|\hat{P}_n(\cdot|\mathbf{x}))\right\}$.*

The lower bound can be easily proved. The upper bound for any convex set $\mathcal{W}$ easily follows as a generalization from the results found in [133]. For $\mathcal{W}$ non convex, the upper bound is easily obtained in the same way as the lower bound.

***Proof:*** *We first show that the size of $\mathscr{B}_\Lambda^n(\mathbf{x},\delta_n)$ is asymptotically equal to the size of $\mathscr{B}_\Sigma^n(\mathbf{x})=\bigcup\limits_{V\in\Sigma}\mathcal{T}_V^n(\mathbf{x})$ where $\Sigma=\mathcal{W}_\Lambda\cap\mathscr{P}_n(\mathscr{Y})$ is the intersection of $\mathcal{W}_\Lambda$ with the*

*set $\mathscr{P}_n(\mathscr{Y})$ of empirical distributions induced by receive sequences of length $n$. In particular, there exists an index $n_0$ such that for all $n \geq n_0$ and $\mathbf{x} \in \mathcal{T}_P^n(\delta_n)$*

$$\|\mathscr{B}_\Sigma^n(\mathbf{x})\| \leq \|\mathscr{B}_\Lambda^n(\mathbf{x}, \delta_n)\| \leq (1+n)^{|\mathscr{X}\|\mathscr{Y}|}\|\mathscr{B}_\Sigma^n(\mathbf{x})\|. \tag{A.2}$$

*The lower bound in* (A.2) *is trivial. We will next establish that there exists $\epsilon_n > 0$ such that for all $n \geq n_0$*

$$\bigcup_{W \in \mathcal{W}_\Lambda} \mathcal{T}_W^n(\mathbf{x}, \delta_n) \subseteq \bigcup_{V \in \Sigma} \mathcal{T}_V^n(\mathbf{x}, \epsilon_n), \tag{A.3}$$

*from which the upper bound in* (A.2) *follows from*

$$\begin{aligned} \left\| \bigcup_{W \in \mathcal{W}_\Lambda} \mathcal{T}_W^n(\mathbf{x}, \delta_n) \right\| &\overset{(a)}{\leq} \sum_{V \in \Sigma} \|\mathcal{T}_V^n(\mathbf{x}, \epsilon_n)\| \\ &\overset{(b)}{\leq} (1+n)^{|\mathscr{X}\|\mathscr{Y}|}\|\mathscr{B}_\Sigma^n(\mathbf{x})\|, \end{aligned} \tag{A.4}$$

*where (a) follows from equation (A.3) and the union bound, (b) follows from $\|\mathcal{T}_V^n(\mathbf{x}, \delta_n)\| \leq (1+n)^{|\mathscr{X}\|\mathscr{Y}|}\|\mathcal{T}_V^n(\mathbf{x})\|$ and the fact that for every $V, \bar{V} \in \mathscr{P}_n(\mathscr{Y})$ with $V \neq \bar{V}$ and each $\mathbf{x} \in \mathscr{X}^n$ we have $\mathcal{T}_V^n(\mathbf{x}) \cap \mathcal{T}_{\bar{V}}^n(\mathbf{x}) = \emptyset$.*

*Let us now prove expression* (A.3). *Assume that $\mathcal{W}_\Lambda$ is a relatively $\tau_0$-open subset of $\mathcal{W}_\Lambda \cup \mathscr{P}_n(\mathscr{Y})$, i.e., every $W \in \mathcal{W}_\Lambda$ has a $\tau_0$-neighborhood defined in the $\tau_0$-topology [79]. Then there exists $n_0$ such that for any $n \geq n_0$ and $\varepsilon > 0$, the $\varepsilon$-open ball $U_0(W, \varepsilon)$ satisfies $U_0(W, \varepsilon) \cap \mathscr{P}_n(\mathscr{Y}) \subset \mathcal{W}_\Lambda$. Choose $0 < \varepsilon' < \varepsilon$ and pick an empirical transition PM $V \in \mathscr{P}_n(\mathscr{Y})$ such that for all $(a, b) \in \mathscr{X} \times \mathscr{Y}$, $|V(b|a) - W(b|a)| < \varepsilon'_n$ and $V(b|a) = 0$ if $W(b|a) = 0$ for every $a \in \mathscr{X}$ with $P(a) > 0$. The continuity properties of information divergences imply that for any sequence $\mathbf{y} \in \mathcal{T}_W^n(\mathbf{x}, \delta_n)$ (i.e., $\mathcal{D}(\hat{W}_n\|W|\hat{P}_n) \leq \delta_n$), $\left|\hat{W}_n(b|a)\hat{P}_n(a) - W(b|a)\hat{P}_n(a)\right| \leq \sqrt{\delta_n/2}$, hence $\left|\hat{W}_n(b|a)\hat{P}_n(a) - V(b|a)\hat{P}_n(a)\right| \leq \varepsilon' + \sqrt{\delta_n/2}$. Finally, from this equation it is easily show, that there exists an $\epsilon_n > 0$ such that $\mathcal{D}(\hat{W}_n\|V|\hat{P}_n) \leq \epsilon_n$, i.e., $\mathbf{y} \in \mathcal{T}_V^n(\mathbf{x}, \epsilon_n)$. Consequently, we have proved that for any $W \in \mathcal{W}_\Lambda$ and large enough $n$, it is possible to find $V \in \Sigma$ and $\epsilon_n > 0$ such that $\mathcal{T}_W^n(\mathbf{x}, \delta_n) \subseteq \mathcal{T}_V^n(\mathbf{x}, \epsilon_n)$, thus establishing* (A.3). *Using similar arguments as above and the uniform continuity of the entropy function, it can be shown that there exists $n'_0$ and $\xi'_n > 0$ such that for all $n \geq n'_0$ and $\mathbf{x} \in \mathcal{T}_P^n(\delta_n)$*

$$\left| \max_{W \in \Sigma} H(W|\hat{P}_n) - \sup_{V \in \mathcal{W}_\Lambda} H(V|P) \right| \leq \xi'_n, \tag{A.5}$$

*with* $\xi'_n \to 0$ *as* $n \to \infty$. *Theorem A.2.1 then follows by combining the inequalities* (A.2) *with Theorem A.2.2 and inequalities* (A.5), *and setting* $\eta_n = \xi'_n + 2|\mathscr{X}||\mathscr{Y}|n^{-1}\log(n+1)$. *Consequently, there exists* $n''_0 = \max\{n'_0, n_0\}$ *such that for any* $n \geq n''_0$ *this theorem holds.* ∎

*Proof of the Generalized Maximal Code Lemma*: For simplicity we denote $M = M_{\theta,\hat{\theta}}$. Up to now we know that choosing any arbitrary confidence set $\Lambda \subset \Theta$ (defined by $\Pr(\Lambda|\hat{\theta}) \geq 1 - \gamma_{QoS}$). The associated robust I-typical set $\mathscr{B}^n_\Lambda(\mathbf{x}, \delta_n) \subset \mathscr{Y}^n$ constitutes a robust $\epsilon$-decoding set for the simultaneous DMCs $\mathcal{W}_\Lambda$, i.e. $\Lambda_\epsilon = \Lambda$ (see above definitions). To prove the direct part, consider an admissible code that is maximal, i.e., it cannot be extended by arbitrary $(\mathbf{x}_{M+1}; \mathscr{D}^n_{M+1})$ such that the extended code remains admissible.

Define the set $\mathscr{D}^n = \bigcup_{i=1}^{M} \mathscr{D}^n_i$ with $\mathscr{D}^n_i \subseteq \mathscr{B}^n_\Lambda(\mathbf{x}_i, \delta)$, and choose $\delta < \epsilon$ such that $1 - \epsilon > \epsilon - \delta$. Then,

$$\inf_{\theta\in\Lambda} W^n(\mathscr{D}^n|\mathbf{x}_i, \theta) > \epsilon - \delta, \quad \text{for all } \mathbf{x}_i \in \mathscr{A}^n. \tag{A.6}$$

For any $\mathbf{x} \in \mathscr{A}^n \setminus \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, if $W^n(\mathscr{B}^n_\Lambda(\mathbf{x}, \delta) \setminus \mathscr{D}^n|\mathbf{x}, \theta) > 1 - \epsilon$ for all $\theta \in \Lambda$, the code would have an admissible extension, contradicting our initial assumption. Thus, for all $\mathbf{x} \in \mathscr{A}^n \setminus \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, we have

$$\inf_{\theta\in\Lambda} W^n(\mathscr{B}^n_\Lambda \setminus \mathscr{D}^n|\mathbf{x}, \theta) \leq 1 - \epsilon.$$

This equation implies that for all $\theta \in \Lambda$ and large enough $n$

$$W^n(\mathscr{D}^n|\mathbf{x}, \theta) \geq \epsilon - \delta, \quad \text{for all } \mathbf{x} \in \mathscr{A}^n \setminus \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}. \tag{A.7}$$

The inequalities (A.6) and (A.7) together imply that $\mathscr{D}^n$ is a common $(\epsilon-\delta)$-image of the set $\mathscr{A}^n$ via the collection of channels $\mathcal{W}_\Lambda$. By the definition of $\mathrm{g}_\Lambda(\mathscr{A}^n, \epsilon - \delta)$ it follows that

$$\|\mathscr{D}^n\| \geq \mathrm{g}_\Lambda(\mathscr{A}^n, \epsilon - \delta). \tag{A.8}$$

On the other hand, $\mathscr{D}^n_i \subseteq \mathscr{B}^n_\Lambda(\mathbf{x}_i, \delta)$ implies that

$$\begin{aligned} \|\mathscr{D}^n\| &= \sum_{i=1}^{M} \|\mathscr{D}^n_i\| \\ &\leq M_{\theta,\hat{\theta}} \|\mathscr{B}^n_\Lambda(\mathbf{x}, \delta)\| \\ &\leq M_{\theta,\hat{\theta}} \exp\big[n\big(H(\mathcal{W}_\Lambda|P) + \delta\big)\big], \end{aligned} \tag{A.9}$$

for $n$ large enough and all $\theta \in \Lambda$, where the last inequality follows by applying the cardinality upper bound of Theorem A.2.1. The lower bound (2.12) is then immediately obtained by combining (A.8) and (A.9). To prove the second statement (converse part), let $\hat{\mathscr{D}}^n$ be a common $(\epsilon+\delta)$-image via the collection of channels $\mathcal{W}_{\Lambda_\epsilon}$, i.e.,

$$\inf_{\theta\in\Lambda_\epsilon} W^n(\hat{\mathscr{D}}^n|\mathbf{x}_m,\theta) \geq \epsilon+\delta, \quad \text{for } m \in \mathcal{M}, \tag{A.10}$$

that achieves the minimum in (2.10), i.e., $\|\hat{\mathscr{D}}^n\| = \mathrm{g}\Lambda_\epsilon(\mathscr{A}^n, \epsilon+\delta)$. For any admissible code, (2.11) and (A.10) imply

$$\inf_{\theta\in\Lambda_\epsilon} W^n(\mathscr{D}_m^n \cap \hat{\mathscr{D}}^n|\mathbf{x}_m,\theta) \geq \delta \quad \text{for } m \in \mathcal{M}. \tag{A.11}$$

Using Corollary 1.2.14 in [17], we hence obtain

$$\big\|\mathscr{D}_m^n \cap \hat{\mathscr{D}}\big\| \geq \exp\big[n\big(H(\mathcal{W}_{\Lambda_\epsilon}|P)-\delta\big)\big], \tag{A.12}$$

for $n$ large enough. On the other hand, the decoding sets $\mathscr{D}_m^n$ are disjoint and thus

$$\begin{aligned}\mathrm{g}_{\Lambda_\epsilon}(\mathscr{A}^n,\epsilon+\delta) = \|\hat{\mathscr{D}}^n\| &\geq \sum_{i=1}^{M} \|\hat{\mathscr{D}}\cap\mathscr{D}_i^n\| \\ &\geq M_{\theta,\hat{\theta}} \exp\big[n\big(H(\mathcal{W}_{\Lambda_\epsilon}|P)-\delta\big)\big],\end{aligned}$$

where the last inequality follows from (A.12). This inequality is equivalent to (2.13) and concludes the proof of the theorem.

## A.3 Information Inequalities

For any given functions $f_1, f_2, \ldots, f_k$ on $\mathscr{Y}$ and numbers $\lambda_1, \lambda_2, \ldots, \lambda_k$, the set $\mathcal{L} = \big\{W(\cdot|x) : \sum_{b\in\mathscr{Y}} W(b|x) f_i(b) = \lambda_i, \;\; 1 \leq i \leq k\big\}$ if non-empty, is called a *linear family* of probability distributions.

**Theorem A.3.1** *Let $\Lambda \subset \Theta$ be a convex set, with $\mathcal{W}_\Lambda \subset \mathscr{P}(\mathscr{Y})$ and $W(\cdot|x,\theta^*) \in \mathcal{W}_\Lambda$ be a transition PM such that $Supp(W_{\theta^*}) = Supp(\mathcal{W}_\Lambda)$. Then,*

$$I(P, W_{\theta^*}) \leq I(P, W_\theta) + \mathcal{D}(W_\theta P \| W_{\theta^*} P) - \mathcal{D}(W_\theta \| W_{\theta^*}|P) \tag{A.13}$$

*holds for every $\theta \in \Lambda$ and any $P \in \mathscr{P}(\mathscr{X})$. Furthermore, if the asserted inequality holds for some $\theta^* \in \Lambda$ and all $\theta \in \Lambda$ then $\theta^*$ must be the transition PM providing*

*the infimun value of the mutual information, i.e.* $I(P,W_{\theta^*}) = \inf_{\theta\in\Lambda} I(P,W_\theta)$. *Moreover, inequality* (A.13) *is actually an equality if* $\mathcal{W}_\Lambda$ *is a linear family of probability distributions* $\mathcal{L}$.

***Proof:*** *For any arbitrary* $W(\cdot|x) \in \mathcal{W}_\Lambda$, *the convexity of* $\mathcal{W}_\Lambda$ *ensures that* $W_\alpha(\cdot|x) = (1-\alpha)W^*(\cdot|x) + \alpha W(\cdot|x) \in \mathcal{W}_\Lambda$ *for all* $0 \leq \alpha \leq 1$. *Observe that* $W_\alpha(\cdot|x)$ *is linear in* $\alpha$ *and* $I(P,W)$ *is a convex function in* $W$, *then* $I(P,W_\alpha)$ *is also convex function in* $\alpha$. *Hence, the difference quotient of* $I(P,W_\alpha)$ *evaluated in* $\alpha = 0$ *is given by,*

$$\Delta_t(\alpha = 0) = \frac{1}{t}\big[I(P,W_t) - I(P,W^*)\big] \tag{A.14}$$

*with* $\Delta_t(\alpha = 0) \geq 0$ *for each* $t \in (0,1)$. *Thus, there exits some* $0 < \tilde{t} < t$ *such that*

$$0 \leq \Delta_t(\alpha = 0) = \frac{d}{d\alpha} I(P,W_\alpha)\Big|_{\alpha=\tilde{t}}. \tag{A.15}$$

*While,*

$$\frac{d}{d\alpha} I(P,W_\alpha) \;=\; \sum_{a\in\mathscr{X}}\sum_{b\in\mathscr{Y}} P(a)\big(W(b|a) - W^*(b|a)\big)\log\frac{W_\alpha(b|a)}{W_\alpha P(b)} \tag{A.16}$$

*and by taking* $t \to 0$ *in expression* (A.15), *we obtain*

$$\begin{aligned} 0 \;&\leq\; \lim_{\tilde{t}\to 0}\Delta_t(\alpha=0) = \frac{d}{d\alpha} I(P,W_t)\Big|_{\alpha=\tilde{t}} \\ &=\; \sum_{a\in\mathscr{X}}\sum_{b\in\mathscr{Y}} P(a)\big(W(b|a) - W^*(b|a)\big)\log\frac{W^*(b|a)}{W^*P(b)} \\ &=\; I(P,W) + \mathcal{D}(WP\|W^*P) - \mathcal{D}(W\|W^*|P) - I(P,W^*), \end{aligned} \tag{A.17}$$

*where we have used the fact that* $Supp(W) \subseteq Supp(W^*)$. *Thus, this concludes the proof of the inequality, since expression* (A.17) *is always positive. In order to show the equality, observe that under the assumption that* $\mathcal{W}_\Lambda$ *is a linear family. For every* $W(\cdot|x) \in \mathcal{L}$, *there is some* $\alpha < 0$ *such that* $W_\alpha(\cdot|x) = (1-\alpha)W^*(\cdot|x) + \alpha W(\cdot|x) \in \mathcal{L}$. *Therefore, we must have* $(d/dt)I(P,W_\alpha)\big|_{\alpha=0} = 0$, *i.e.* $\sum_{a\in\mathscr{X}}\sum_{b\in\mathscr{Y}} P(a)\big(W(b|a) - W^*(b|a)\big)\log\frac{W^*(b|a)}{W^*P(b)} = 0$, *for all* $W(\cdot|x) \in \mathcal{L}$, *and this proves the equality in* (A.13). ■

# Appendix B

# Auxiliary Proofs

## B.1 Metric evaluation

**Theorem B.1.1** *Let* $\mathbf{H}_i \in \mathbb{C}^{M_R \times M_T}$ *(*$i = 1, 2$*) be circularly symmetric complex Gaussian random matrices with zero means and full-rank Hermitian covariance matrices* $\mathbf{\Sigma}_{ij} = \mathbb{E}\{(\mathbf{H})_i (\mathbf{H})_j^\dagger\}$ *of the columns* $(\mathbf{H})_i$ *of* $\mathbf{H}_i$ *(assumed to be the same for all columns) for* $i = 1, 2$*. Then the random variable* $\mathbf{H}_1 | \mathbf{H}_2 \sim C\mathcal{N}(\mu, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma})$ *is a circularly symmetric complex Gaussian with mean* $\mu = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{H}_2$ *and covariance matrix of its columns* $\mathbf{\Sigma} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$*.*

From (3.9) and (3.10), by choosing $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{\mathbf{H}}$ and $\mathbf{\Sigma}_{22} = \mathbf{\Sigma}_{\mathbf{H}} + \mathbf{\Sigma}_{\mathcal{E}}$ in Theorem B.1.1. We obtain the *a posteriori* pdf $\psi_{\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}}}(\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}}) = C\mathcal{N}\big(\mathbf{\Sigma}_{\mathbf{\Delta}}\widehat{\mathbf{H}}_{\mathrm{ML}}, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma}_{\mathbf{\Delta}}\mathbf{\Sigma}_{\mathcal{E}}\big)$, where $\mathbf{\Sigma}_{\mathbf{\Delta}} = \mathbf{\Sigma}_{\mathbf{H}}(\mathbf{\Sigma}_{\mathcal{E}} + \mathbf{\Sigma}_{\mathbf{H}})^{-1}$. In order to evaluate the general expression of the decoding metric (3.7) for fading MIMO channels, we compute the expectation of $\mathbf{W}(\mathbf{y}|\mathbf{x}, \mathbf{H}) = C\mathcal{N}\big(\mathbf{Hx}, \mathbf{\Sigma_0}\big)$ over $\psi_{\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}}}(\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}})$. To this end, we need the following result (cf. [134]).

**Theorem B.1.2** *For a circularly symmetric complex random vector* $\mathbf{V} \sim C\mathcal{N}(\mu, \Pi)$ *with mean* $\mu = \mathbb{E}_{\mathbf{V}}\{\mathbf{V}\}$ *and covariance matrix* $\mathbf{\Pi} = \mathbb{E}_{\mathbf{V}}\{\mathbf{V}\mathbf{V}^\dagger\} - \mu\mu^\dagger$*, and Hermitian matrix* $\mathbf{A}$ *such that* $\mathbb{I} + \mathbf{\Pi A} \succ 0$*, which means positive definite, we have*

$$\mathbb{E}_{\mathbf{V}}\big[\exp(-\mathbf{V}^\dagger \mathbf{A} \mathbf{V})\big] = |\mathbb{I} + \mathbf{\Pi A}|^{-1} \exp\big[-\mu^\dagger \mathbf{A}(\mathbb{I} + \mathbf{\Pi A})^{-1}\mu\big]. \tag{B.1}$$

From this theorem, we can compute the composite channel $\widetilde{\mathbf{W}}(\mathbf{y}|\mathbf{x}, \widehat{\mathbf{H}})$. Let us define $\mathbf{V} = \mathbf{y} - \mathbf{Hx}$ such that the conditional pdf of $\mathbf{V}$ given $(\widehat{\mathbf{H}}, \mathbf{x})$ is $\mathbf{V}|(\widehat{\mathbf{H}}, \mathbf{x}) \sim C\mathcal{N}(\mu, \mathbf{\Pi})$

with $\mu = \mathbf{y} - \mathbf{\Sigma_\Delta}\widehat{\mathbf{H}}\mathbf{x}$ and $\mathbf{\Pi} = \mathbf{\Sigma_\Delta}\mathbf{\Sigma}_\mathcal{E}\|\mathbf{x}\|^2$. Thus, by defining $\mathbf{A} = \mathbf{\Sigma_0}^{-1}$ from (B.1) and some algebra, we obtain $\widetilde{\mathbf{W}}(\mathbf{y}|\mathbf{x}, \widehat{\mathbf{H}}) = C\mathcal{N}\big(\delta\widehat{\mathbf{H}}\mathbf{x}, \mathbf{\Sigma_0} + \delta\mathbf{\Sigma}_\mathcal{E}\|\mathbf{x}\|^2\big)$.

## B.2 Proof of Lemma 3.5.1

Consider the quadratic expressions $Q_1(\mathbf{X}) = \|\mathbf{AX}\|^2 + K_1$ and $Q_2(\mathbf{X}) = \|\mathbf{X}\|^2 + K_2$, $\mathbf{X}$ is a vector of $M_T$ elements, such that $Q_1, Q_2 > 0$ *almost surely*. The joint generating function of $Q_1$ and $Q_2$, namely, $M_{Q_1,Q_2}(t_1, t_2) = \mathbb{E}_\mathbf{X}\big\{\exp\big(t_1Q_1(\mathbf{X}) + t_2Q_2(\mathbf{X})\big)\big\}$. Evaluating this, we obtain

$$M_{Q_1,Q_2}(t_1, t_2) = \exp\big(t_1K_1 + t_2K_2\big)\big|\mathbb{I}_{M_R} - \big(t_1\mathbf{A}^\dagger\mathbf{A} + t_2\big)\mathbf{\Sigma_P}\big|^{-1/2}. \tag{B.2}$$

Then from the gamma integral and setting $t_2 = -z$ in (C.14)

$$\mathbb{E}_\mathbf{X}\big\{Q_1(\mathbf{X})Q_2^{-1}(\mathbf{X})\big\} = \int_0^\infty \mathbb{E}_\mathbf{X}\big\{Q_1(\mathbf{X})\exp\big[-zQ_2(\mathbf{X})\big]\big\}dz, \tag{B.3}$$

where it is not difficult to show that

$$\begin{aligned}\mathbb{E}_\mathbf{X}\big\{Q_1(\mathbf{X})\exp\big[-zQ_2(\mathbf{X})\big]\big\} &= \frac{\partial M_{Q_1,Q_2}(t_1,-z)}{\partial t_1}\Big|_{t_1=0}, \\ &= \big[K_1 + 2^{-1}tr(\mathbf{A\Sigma_P A}^\dagger)(1+z\bar{P})^{-1}\big] \\ &\quad \times(1+z\bar{P})^{-(M_T/2)}\exp\big(-K_2z\big).\end{aligned} \tag{B.4}$$

Finally, this Lemma follows by solving the integral in (C.15), which leads to expression (3.19).

# Appendix C

# Additional Computations

## C.1 Proof of Theorem 4.2.1

Next we provide an outline of the proof of coding theorem 4.2.1 and its weak converse.

***Proof:*** *The direct part of the theorem easily follows by using the same random coding scheme that is used to achieve the capacity* (4.1) *with perfect channel knowledge. The main deference is that in this case we have to design random codewords (forming the codebook) with the channel statistic corresponding to the composite model* $\widetilde{W}$*. Then, given channel estimates* $\underline{\hat{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$*, it is not difficult to show that the average error probability* $\bar{e}^{(n)}_{\max}(\varphi, \phi, \underline{\hat{\theta}}) \to 0$ *vanishes as* $n \to \infty$*. Whereas, a weak converse follows from the convexity property of the conditional entropy and the Fano's Lemma. As messages* $m \in \{1, \ldots, \lfloor 2^{nR_{\hat{\theta}}} \rfloor\}$ *are assumed to be uniformly distributed, we have:*

$$\begin{aligned} R_{\underline{\hat{\theta}}} &= n^{-1} I\big(m; \widetilde{\mathbf{y}}_{\underline{\hat{\theta}}}\big) + n^{-1} H\big(m|\widetilde{\mathbf{y}}_{\underline{\hat{\theta}}}\big), \\ &\overset{(a)}{\leq} n^{-1} I\big(m; \widetilde{\mathbf{y}}_{\underline{\hat{\theta}}}\big) + n^{-1} \mathbb{E}_{\underline{\theta}|\underline{\hat{\theta}}}\big\{H\big(m|\widetilde{\mathbf{y}}_{\underline{\hat{\theta}},\underline{\theta}}\big)\big\}, \\ &\overset{(b)}{\leq} n^{-1} I\big(m; \widetilde{\mathbf{y}}_{\underline{\hat{\theta}}}\big) + \mathbb{E}_{\underline{\theta}|\underline{\hat{\theta}}}\big\{H_2\big(P^{(n)}_{e,\underline{\hat{\theta}}}(\underline{\theta})\big) + P^{(n)}_{e,\underline{\hat{\theta}}}(\underline{\theta})\big\}, \\ &\overset{(c)}{\leq} n^{-1} I\big(m; \widetilde{\mathbf{y}}_{\underline{\hat{\theta}}}\big) + \big(H_2(\bar{P}^{(n)}_{e,\underline{\hat{\theta}}}) + \bar{P}^{(n)}_{e,\underline{\hat{\theta}}}\big), \end{aligned} \tag{C.1}$$

*where* $\widetilde{\mathbf{y}}_{\underline{\hat{\theta}}} = (\widetilde{Y}_{\hat{\theta}_1,1}, \ldots, \widetilde{Y}_{\hat{\theta}_n,n})$ *is the vector of channel outputs, whose joint probability distribution is computed using the n-extension of the composite channel* $\widetilde{W}^n_{\underline{\hat{\theta}}}$*,* $\mathbf{s} = (S_1, \ldots, S_n)$ *is the sequence of channel states and* $H_2(p) \triangleq -p \log p - (1-p)\log(1-p)$*. (a) Follows from the convexity of the conditional entropy, (b) follows from the Fano's*

*Lemma and (c) follows from the concavity property of the binary entropy $H_2$ respect to the error probability with $\bar{P}_{e,\underline{\hat{\theta}}}^{(n)} \triangleq \mathbb{E}_{\underline{\theta}|\underline{\hat{\theta}}}\{P_{e,\underline{\hat{\theta}}}^{(n)}(\underline{\theta})\}$. Then, from* (C.1) *by bounding the following term as [33]*

$$n^{-1} I\big(m; \widetilde{\mathbf{y}}_{\underline{\hat{\theta}}}\big) \leq \frac{1}{n} \sum_{i=1}^{n} \big[ I\big(U_{\hat{\theta}_i,i}; \widetilde{Y}_{\hat{\theta},i}\big) - I\big(U_{\hat{\theta},i}; S_i\big) \big], \tag{C.2}$$

*the proof follows by taking the average over all channel estimates and noting that the right-hand side in* (C.1) *grows to zero as $\bar{P}_{e,\underline{\hat{\theta}}}^{(n)} \to 0$ when $n \to \infty$.* ∎

## C.2 Composite MIMO-BC Channel

The achievable rate region in Theorem 4.2.2 depends only on the conditional marginal distributions of the composite MIMO-BC, which follows as the average of the unknown marginal channel (4.30) over the *a posterior* pdf. According to the $K$-th extension of the marginal pdfs (4.7), this writes as

$$\widetilde{\mathbf{W}}_k(\mathbf{y}_k|\mathbf{x}, \widehat{\mathbf{H}}_k) = \int \cdots \int_{\mathbb{C}^{M_R \times M_T}} \mathbf{W}_k(\mathbf{y}_k|\mathbf{x}, \mathbf{H}_k)\, \underline{d} f_{\underline{\mathbf{H}}\,\{\underline{\widehat{\mathbf{H}}}\}_k|\widehat{\mathbf{H}}_k}(\underline{\mathbf{H}}, \{\underline{\widehat{\mathbf{H}}}\}_k|\widehat{\mathbf{H}}_k), \tag{C.3}$$

where $\{\underline{\widehat{\mathbf{H}}}\}_k = (\widehat{\mathbf{H}}_1, \cdots, \widehat{\mathbf{H}}_{k-1}, \widehat{\mathbf{H}}_{k+1}, \cdots, \widehat{\mathbf{H}}_K)$ and $\underline{\mathbf{H}} = (\mathbf{H}_1, \ldots, \mathbf{H}_m)$. We note that in this case the matrices $\underline{\mathbf{H}}$ are independents and on the other side $\mathbf{Y}_k \ominus (\mathbf{X}, \mathbf{H}_k) \ominus \{\underline{\mathbf{H}}\}_k$ and $\mathbf{H}_k \ominus \widehat{\mathbf{H}}_k \ominus (\{\underline{\mathbf{H}}\}_k, \{\underline{\widehat{\mathbf{H}}}\}_k)$ form a Markov chain for every $k = \{1, \ldots, K\}$. Thus, we must only compute the pdf $f_{\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}}}(\mathbf{H}_k|\widehat{\mathbf{H}}_{\mathrm{ML},k})$ and $f_{\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{MMSE}}}(\mathbf{H}_k|\widehat{\mathbf{H}}_{\mathrm{MMSE},k})$ for which we need the following theorem.

**Theorem C.2.1** *Let $\mathbf{H}_i \in \mathbb{C}^{M_R \times M_T}$ be circularly symmetric complex Gaussian random matrices with zero means and full-rank Hermitian covariance matrices $\mathbf{\Sigma}_{ij} = \mathbb{E}\{(\mathbf{H})_i (\mathbf{H})_j^{\dagger}\}$ of the columns $(\mathbf{H})_i$ of $\mathbf{H}_i$ (assumed to be the same for all columns) for $i = 1, 2$. Then the random variable $\mathbf{H}_1|\mathbf{H}_2 \sim C\mathcal{N}(\mu, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma})$ is a circularly symmetric complex Gaussian with mean $\mu = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{H}_2$ and covariance matrix of its columns $\mathbf{\Sigma} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.*

From expressions (4.29) and (4.31), by choosing $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{\mathbf{H},\mathbf{k}}$ and $\mathbf{\Sigma}_{22} = \mathbf{\Sigma}_{\mathbf{H},\mathbf{k}} + \mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}$ in Theorem C.2.1, we obtain the *a posteriori* pdf

$$f_{\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{ML}}}(\mathbf{H}_k|\widehat{\mathbf{H}}_{\mathrm{ML},k}) = C\mathcal{N}\big(\mathbf{\Sigma}_{\mathbf{\Delta},\mathbf{k}}\widehat{\mathbf{H}}_{\mathrm{ML},k}, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma}_{\mathbf{\Delta},\mathbf{k}}\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\big), \tag{C.4}$$

where $\mathbf{\Sigma_{\Delta,k}} = \mathbf{\Sigma_{H,k}}(\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}} + \mathbf{\Sigma_{H,k}})^{-1}$. We note from (4.32) that both estimators yield to the same *a posteriori* pdf, since

$$f_{\mathbf{H}|\widehat{\mathbf{H}}_{\mathrm{MMSE}}}(\mathbf{H}_k|\widehat{\mathbf{H}}_{\mathrm{MMSE},k}) = C\mathcal{N}\big(\mathbf{\Sigma_{\Delta,k}}\mathbf{A}_{\mathrm{MMSE},k}^{-1}\widehat{\mathbf{H}}_{\mathrm{MMSE},k}, \mathbb{I}_{M_T} \otimes \mathbf{\Sigma_{\Delta,k}}\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\big). \quad \text{(C.5)}$$

We shall denote this pdf as $f_{\mathbf{H}|\widehat{\mathbf{H}}}(\mathbf{H}_k|\widehat{\mathbf{H}}_k)$ for some arbitrary estimate $\widehat{\mathbf{H}}_k$. Finally, by using (C.4) and the following result (cf. [134]) we can easily evaluate expression (C.3).

**Theorem C.2.2** *For a circularly symmetric complex random vector* $\mathbf{v} \sim C\mathcal{N}(\mu, \Pi)$ *with mean* $\mu = \mathbb{E}_{\mathbf{V}}\{\mathbf{v}\}$ *and covariance matrix* $\mathbf{\Pi} = \mathbb{E}_{\mathbf{V}}\{\mathbf{v}\mathbf{v}^\dagger\} - \mu\mu^\dagger$*, and Hermitian matrix* $\mathbf{A}$ *such that* $\mathbb{I} + \mathbf{\Pi A} \succ 0$*, which means positive definite, we have*

$$\mathbb{E}_{\mathbf{V}}\big[\exp(-\mathbf{v}^\dagger\mathbf{A}\mathbf{v})\big] = |\mathbb{I} + \mathbf{\Pi A}|^{-1}\exp\big[-\mu^\dagger\mathbf{A}(\mathbb{I} + \mathbf{\Pi A})^{-1}\mu\big]. \quad \text{(C.6)}$$

From this theorem, we can compute the marginal distributions of the composite channel $\widetilde{\mathbf{W}}_k(\mathbf{y}_k|\mathbf{x}, \widehat{\mathbf{H}}_k)$. Let us define $\mathbf{v} = \mathbf{y}_k - \mathbf{H}_k\mathbf{x}$ such that the conditional pdf of $\mathbf{v}$ given $(\widehat{\mathbf{H}}_k, \mathbf{x})$ is $\mathbf{v}|(\widehat{\mathbf{H}}_k, \mathbf{x}) \sim C\mathcal{N}(\mu, \mathbf{\Pi})$ with $\mu = \mathbf{y}_k - \mathbf{\Sigma_{\Delta,k}}\widehat{\mathbf{H}}_k\mathbf{x}$ and $\mathbf{\Pi} = \mathbf{\Sigma_{\Delta,k}}\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\|\mathbf{x}\|^2$. Thus, by defining $\mathbf{A} = \mathbf{\Sigma_{0,k}}^{-1}$ from (C.6) and some algebra, we obtain

$$\widetilde{\mathbf{W}}_k(\mathbf{y}_k|\mathbf{x}, \widehat{\mathbf{H}}_k) = C\mathcal{N}\big(\delta_k\widehat{\mathbf{H}}_k\mathbf{x}, \mathbf{\Sigma_{0,k}} + \delta_k\mathbf{\Sigma}_{\mathcal{E},\mathbf{k}}\|\mathbf{x}\|^2\big). \quad \text{(C.7)}$$

## C.3 Evaluation of the Marton's Region for the Composite MIMO-BC

Consider that users codeword $\{\mathbf{x}_k\}_{k=1}^K$ are independent Gaussian vectors $\mathbf{x}_k \sim C\mathcal{N}(0, \mathbf{P}_k)$ with corresponding covariance matrices $\{\mathbf{P}_k \succeq 0\}_{k=1}^K$. Assume arbitrary positive semi-defined matrices $\mathbf{F}_k \in \mathbb{C}^{M_R \times M_R}$ (not depending on the unknown channel estimates), and let $P(\mathbf{x}, \mathbf{u}_1, \ldots, \mathbf{u}_K)$ be the joint pdf of auxiliary random vectors defined as

$$\mathbf{u}_k = \mathbf{x}_k + \mathbf{F}_k\,\mathbf{s}_{\Sigma,k+1}^K, \quad \text{(C.8)}$$

thus this pdf does not depend on the channel estimates $\underline{\widehat{\mathbf{H}}}$. From the extension to $K$-users of Theorem (4.2.2) and by evaluating the composite MIMO-BC and the DPC scheme (C.8), it is not difficult to show that for every realization of channel estimates

$$\widetilde{R}_k(\mathbf{F}_k, \widehat{\mathbf{H}}_k) = I\big(P_{U_k}, \widetilde{W}_{\widehat{\mathbf{H}}_k}\big) - I\big(P_{U_k}, P_{U_1,\ldots,U_{k-1}|U_k}\big), \quad \text{for each } k = \{1, \ldots, K\}. \quad \text{(C.9)}$$

Then, by using standard algebra and taking the average of (C.9) over all channel estimates, we can obtain expression (4.43).

## C.4 Proof of Lemma 4.4.1

Let $\mathbf{A_k} = \widehat{\mathbf{H}}_k \widehat{\mathbf{H}}_k^{\dagger} = \sum_{i=1}^{M_T} \hat{\mathbf{h}}_\mathbf{i} \hat{\mathbf{h}}_\mathbf{i}^{\dagger}$ be an $M_R \times M_R$ random complex matrix whose columns are the vectors $\hat{\mathbf{H}}_\mathbf{1}, \ldots, \hat{\mathbf{h}}_{\mathbf{M_T}}$. Then $\mathbf{A_k}$ follows a nonsingular central Wishart distribution of dimensionality $M_R$ with $M_T$ degree of freedom and associated parameter matrix $\mathbf{\Sigma}_{\widehat{\mathbf{H}},\mathbf{k}} = \sigma^2_{\hat{H},k} \mathbb{I}_{M_R}$, i.e. the pdf of any matrix $\mathbf{A_k} \succeq 0$ is given by

$$f(\mathbf{A_k}) = K^{-1} \big|\mathbf{A_k}\big|^{(M_T - M_R - 1)/2} \exp\big[tr(\mathbf{\Sigma}_{\widehat{\mathbf{H}},\mathbf{k}}{}^{-1} \mathbf{A_k})\big], \tag{C.10}$$

$$K = \big|\mathbf{\Sigma}_{\widehat{\mathbf{H}},\mathbf{k}}\big|^{\frac{M_T}{2}} \Gamma_{M_R}(M_T/2),$$

and

$$\Gamma_{M_R}(M_T/2) = \pi^{M_R(M_R-1)/4} \prod_{j=1}^{M_T} \Gamma\big[(M_T + 1 - j)/2\big].$$

We define the exponential matrix function $f(t) = \exp(t\mathbf{A})$, for all $t \in \mathbb{R}$ and any Hermitian matrix $\mathbf{A} \in \mathbb{C}^{M_R \times M_R}$ with

$$\exp(t\mathbf{A}) = \sum_{j=0}^{\infty} \frac{1}{j!} (t\mathbf{A})^j,$$

and we note that $\dfrac{d}{dt} \exp(t\mathbf{A}) = \exp(t\mathbf{A})\mathbf{A}$. Since $\mathbf{A} = \mathbf{A}^{\dagger}$ it is not difficult to show that the matrix inverse can be written as [135]

$$\mathbf{A}^{-1} = \int_0^{\infty} \exp(-z\mathbf{A}) dz, \tag{C.11}$$

this integral expression is a generalization of the Gamma integral for the matrix case.

Consider now the quadratic expressions $\mathbf{Q_1}(\mathbf{A_k}) = \mathbf{A_k}$ and $\mathbf{Q_2}(\mathbf{A_k}) = \mathbf{A_k} + \mathbf{C_k}$, with $\mathbf{C_k} \succeq 0$ a diagonal matrix and $\mathbf{Q_1}, \mathbf{Q_2} \succeq 0$ *almost surely*. Thus, the derivation of Lemma 4.4.1 follows by calculating the expectation that we denote as $\mathbf{I_k}$, given by

$$\mathbf{I_k} = \mathbb{E}_{\mathbf{A_k}}\{\mathbf{Q_1}(\mathbf{A_k}) \mathbf{Q_2}{}^{-1}(\mathbf{A_k})\}, \tag{C.12}$$

where the integral involved in this expectation must be calculated over all positive semi-definite matrices $\mathbf{A_k} \succeq 0$. We solve (C.12) through the joint generating function

of $\mathbf{Q_1}$ and $\mathbf{Q_2}$, namely,

$$M_{Q_1,Q_2}(\mathbf{T_1},\mathbf{T_2}) = \mathbb{E}_{\mathbf{A_k}}\big\{ \exp\big(\mathbf{T_1}\mathbf{Q_1}(\mathbf{A_k}) + \mathbf{T_2}\mathbf{Q_2}(\mathbf{A_k})\big)\big\}. \tag{C.13}$$

where $\mathbf{T_1}, \mathbf{T_2} \succeq 0$ are arbitrary positive definite matrices.

This expression can be evaluated by using the Wishart distribution (C.10) through the Lebesgue measure in $\mathbb{C}^{M_R\times M_R}$ given by $\mathbf{dA_k} = 2^{M_R}\prod_{j=1}^{M_R} b_{jj}^{M_R+1-j}\mathbf{dB}$, where $\mathbf{A_k} = \mathbf{BB}^\dagger$ with $\mathbf{B} = (b_{ij})$, $b_{ii} > 0\ \forall$, $i$, $b_{ij} = 0, \forall i < j$. Thus, using some algebra from (C.13) we can show that

$$M_{Q_1,Q_2}(\mathbf{T_1},\mathbf{T_2}) = \big|\mathbb{I}_{M_R} - \mathbf{\Sigma}_{\widehat{\mathbf{H}},\mathbf{k}}\mathbf{T_1} - \mathbf{\Sigma}_{\widehat{\mathbf{H}},\mathbf{k}}\mathbf{T_2}\big|^{-M_T/2}\exp\big(\mathbf{T_2}\mathbf{C}\big). \tag{C.14}$$

Then from expression (C.11) the integral $\mathbf{I_k}$ (C.12) writes

$$\mathbb{E}_{\mathbf{A_k}}\big\{\mathbf{Q_1}(\mathbf{A_k})\mathbf{Q_2}(\mathbf{A_k})^{-1}\big\} = \int_0^\infty \mathbb{E}_{\mathbf{A_k}}\big\{\mathbf{Q_1}(\mathbf{A_k})\exp\big[-z\mathbf{Q_2}(\mathbf{A_k})\big]\big\}dz. \tag{C.15}$$

Actually, by setting $\mathbf{T_1} = t\mathbb{I}_{M_R}$ and $\mathbf{T_2} = -z\mathbb{I}_{M_R}$ in (C.14), $\forall\, t, z \in \mathbb{R}_+$, it is not difficult to show that

$$\mathbb{E}_{\mathbf{A_k}}\big\{\mathbf{Q_1}(\mathbf{A_k})\exp\big[-z\mathbf{Q_2}(\mathbf{A_k}))\big]\big\} = \frac{\partial M_{Q_1,Q_2}(t\mathbb{I}_{M_R}, -z\mathbb{I}_{M_R})}{\partial t}\Big|_{t=0}, \tag{C.16}$$

where from (C.14)

$$\frac{\partial M_{Q_1,Q_2}(t\mathbb{I}_{M_R}, -z\mathbb{I}_{M_R})}{\partial t}\Big|_{t=0} = \frac{M_T}{2}\mathbf{\Sigma}_{\widehat{\mathbf{H}},\mathbf{k}}\big(1+z\sigma^2_{\hat{H},k}\big)^{-\left(\frac{M_TM_R}{2}+1\right)}\exp\big(-z\mathbf{C}_k\big). \tag{C.17}$$

Finally, it remains to solve the integral in (C.15) using (C.17) (it can be found in [136]), which leads to the following expression

$$\mathbb{E}_{\mathbf{A_k}}\big\{\mathbf{Q_1}(\mathbf{A_k})\mathbf{Q_2}^{-1}(\mathbf{A_k})\big\} = \frac{1}{M_R}\big[1-\rho_k^{n+1}\exp(\rho_k)\Gamma(-n,\rho_k)\big]\mathbb{I}_{M_R}, \tag{C.18}$$

where $n = M_RM_T - 1$, $\mathbf{C}_k = c_k\mathbb{I}_{M_R}$, $\rho_k = \dfrac{c_k}{\sigma^2_{\hat{H},k}}$ and

$$\Gamma(-n,t) = \frac{(-1)^n}{n!}\Big[\Gamma(0,t) - \exp(-t)\sum_{i=0}^{n-1}(-1)^i\frac{i!}{t^{i+1}}\Big],$$

with $\Gamma(0,t) = \displaystyle\int_t^{+\infty} u^{-1}\exp(-u)du$ denoting the exponential integral function. The Lemma follows from (C.18) and the adequate choice of $c_k$.

# References


[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[2] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record, Part 4*, pp. 142–163, 1959.

[3] I. Csiszár, "The method of types," *IEEE Trans. Information Theory*, vol. IT-44, pp. 2505–2523, October 1998.

[4] B. McMillan, "The basic theorems of information theory," *Ann. of Math. Statist.*, vol. 24, p. 196, 1953.

[5] L. Breiman, "The individual ergodic theorem of information theory," *Ann. of Math. Statist.*, pp. 809–811, 1957.

[6] A. Feinstein, "A new achievable rate region for the interference channel," *IRE Transactions on Information Theory*, pp. 2–20, 1954.

[7] J. Wolfowitz, *Coding Theorems of Information Theory.* Berlin, 1964.

[8] A. J. Khinchine, *On the fundamental theorems of information theory.* Uspekhi Matematicheskikh Nauk., 11:17-75, 1957. Translated in Mathematical Foundations of Information Theory, Dover New York, 1957.

[9] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom, "On the general definitions of the quantity of information," *Dokl. Akad. Nauk*, vol. 111, pp. 745–748, 1956.

[10] A. N. Kolmogorov, A. M. Yaglom, and I. M. Gelfand, "Quantity of information and entropy for continuous distributions," in *3rd All-Union Mat. Conf. Izd. Akad. Nauk. SSSR*, vol. 3, pp. 300–320, 1956.

[11] R. L. Dobrushi, "A general formulation of the fundamental shannon theorem in information theory," in *Translation in Transactions Amer. Math. Soc, series 2*, vol. 33, pp. 323–438, 1956.

[12] S. Kullback, *Information Theory and Statistics*. Dover, New York (reprint of 1959 edition published by Wiley), 1968.

[13] R. G. Gallager, *Information theory and reliable communications*. Wiley, New York, 1968.

[14] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecomunications, Wiley & Sons New York, 1991.

[15] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, N.J., 1971.

[16] R. Gray, *Entropy and Information Theory*. Springer-Verlag, New York, 1990.

[17] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Academic, New York, 1981.

[18] A. A. E. Gamal and T. M. Cover, "Multiple user information theory," *IEEE Transactions on Information Theory*, vol. IT-68, pp. 1466–1483, December 1980.

[19] E. Van der Meulen, "A survey of multi-way channels in information theory," *IEEE Trans. Information Theory*, vol. IT-23, pp. 1–37, 1977.

[20] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications* (G. Longo, ed.), Springer-Verlag, New York, 1977.

[21] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoric and communications aspects," *IEEE Trans. Information Theory*, vol. IT-40, pp. 2619–2692, October 1998.

[22] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Information Theory*, vol. 43, pp. 359–378, May 1994.

[23] R. Knopp and P. Humblet, "On coding for block fading channels," *IEEE Trans. Information Theory*, vol. IT-46, pp. 189–205, Jan 2000.

[24] E. Malkamaki and H. Leib, "Coded diversity on block-fading channels," *IEEE Trans. Information Theory*, vol. IT-45, pp. 771–781, Mar 1999.

[25] C. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp. 289–293, 1958.

[26] D. Blackwell, L. Breiman, and A. Thomasian, "The capacity of a class of channels," *Ann. Math. Stat.*, vol. 30, pp. 1229–1241, 1959.

[27] R. L. Dobrushin, "Optimun information transmission through a channel with unknown parameters," *Radio Eng. Electron.*, vol. 4, no. 12, pp. 1–8, 1959.

[28] J. Wolfowitz, "Simultaneous channels," *Arch. Rat. Mech. Anal.*, vol. 4, pp. 371–386, 1960.

[29] D. Blackwell, L. Breiman, and A. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Stat.*, vol. 31, pp. 558–567, 1960.

[30] A. Lapidoth, "Reliable communication under channel uncertainty," *IEEE Trans. Information Theory*, vol. 44, pp. 2148–2177, October 1998.

[31] A. Kusnetsov and T. B.S., "Coding in memory with defective cells," *Prob. Peredach. Inform.*, vol. 10, no. 2, pp. 52–60, April-June 1974.

[32] C. Heegar and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Information Theory*, vol. IT-29, pp. 731–739, 1983.

[33] S. I. Gelfand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[34] T. R. M. Fischer, "Some remarks on the role of inaccuracy in shannon's theory of information transmission," in *Trans. 8th Prague Conf. on Information Theory*, pp. 211–226, 1971.

[35] D. Divsalar, *Performance of mismatched receivers on bandlimited channels*. PhD thesis, Ph.D. dissertation, Univ. of California, Los Angeles, 1979.

[36] J. Omura and B. Levitt, "Coded error probability evaluation for antijam communication systems," *IEEE Transactions on Communications*, vol. 30, pp. 896–903, May 1982.

[37] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Information Theory*, vol. 23, pp. 337– 343, May 1977.

[38] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Information Theory*, vol. 44, pp. 1726–1745, Sep 1998.

[39] O. Shayevitz and M. Feder, "Universal decoding for frequency-selective fading channels," *IEEE Trans. Information Theory*, vol. 51, pp. 2770– 2790, Aug 2005.

[40] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Information Theory*, vol. IT-41, no. 1, pp. 35–43, 1995.

[41] I. Csiszár, "Graph decomposition: a new key to coding theorems," *IEEE Trans. Information Theory*, vol. IT-27, pp. 5–12, January 1981.

[42] J. Hui, "Fundamental issues of multiple accessing," tech. rep., Ph.D. dissertation, M.I.T., ch. IV, 1983.

[43] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Information Theory*, vol. IT-42, pp. 1439–1452, Sept. 1996.

[44] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Information Theory*, vol. IT-40, pp. 1953–1967, Nov. 1994.

[45] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: general alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Information Theory*, vol. 46, pp. 2315–2328, Nov. 2000.

[46] G. Kaplan and S. Shamai (Shitz), "Information rates and error exponents of compound channels with application to antipodal signaling in a fading," *Environment, AEU (Electronics and Communication)*, vol. 47, no. 4, p. 228■230, 1993.

[47] A. Lapidoth, "Nearest neighbor decoding for additive non-gaussian noise channels," *IEEE Trans. Information Theory*, vol. 42, pp. 1520–1529, Sep 1996.

[48] A. Lapidoth and S. Shamai, "Fading channels: How perfect need "perfect side information" be ?," *IEEE Trans. Information Theory*, vol. 48, pp. 1118–1134, May 2002.

[49] H. Weingarten, Y. Steinberg, and S. Shamai, "Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels weingarten," *IEEE Trans. Information Theory*, vol. 50, pp. 1665– 1686, Aug 2004.

[50] D. Samardzija and N. Mandayam, "Pilot-assisted estimation of mimo fading channel response and achievable data rates," *IEEE Transactions on Signal Processing*, vol. 51, pp. 2882– 2890, Nov 2003.

[51] T. Cover, "Broadcast channels," *IEEE Trans. Information Theory*, vol. IT-18, pp. 2–14, 1972.

[52] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Information Theory*, vol. IT-19, pp. 197–207, 1973.

[53] R. G. Gallager, "Capacity and coding for degraded broadcast channels," *Problemy Peredaci Informaccii*, vol. 10, no. 3, pp. 3–14, 1974.

[54] R. Ahlswede and J. Körner, "Source coding with side information and a converse for the degraded broadcast channel," *IEEE Trans. Information Theory*, vol. IT-21, pp. 629–637, 1975.

[55] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Information Theory*, vol. IT-25, pp. 306–311, 1979.

[56] A. El Gamal and E. Van der Meulen, "A proof of Marton's coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Information Theory*, vol. IT-27, pp. 120–122, 1981.

[57] T. Cover, "Comments on broadcast channels," *IEEE Trans. Information Theory*, vol. IT-44, pp. 2524–2530, 1998.

[58] M. Médard, "The effect upon channel capacity in wireless communication of perfect and imperfect knownledge of the channel," *IEEE Trans. Information Theory*, vol. IT-46, pp. 933–946, May 2000.

[59] T. Yoo and A. Goldsmith, "Capacity of fading MIMO channels with channel estimation error," in *Proceedings of International Conf. on Comunications (ICC)*, June 2004.

[60] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Transactions on Information Theory*, vol. IT-49, pp. 951–961, April 2003.

[61] V. Tarokh, A. Naguib, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communication:performance criteria in the presence of channel estimation errors,mobility, and multiple paths," *IEEE Transactions on Communications*, pp. 199–207, Feb 1999.

[62] G. Taricco and E. Biglieri, "Space-time decoding with imperfect channel estimation," *IEEE Trans. on Wireless Communications*, vol. 4, pp. 2426 – 2467, July 2005.

[63] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna gaussian broadcast channel," *IEEE Trans. Information Theory*, vol. IT-49, pp. 1691–1706, july 2003.

[64] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Information Theory*, pp. 3936–3964, Sep. 2006.

[65] A. Lapidoth, S. Shamai, and M. Wigger, "On the capacity of a MIMO Fading Broadcast Channel with imperfect transmitter side-information," in *Proceedings of Allerton Conf. on Commun., Control, and Comput.*, Sep. 2005.

[66] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. on Telecomm. ETT*, vol. 10, pp. 585–596, Nov. 1999.

[67] M. Costa, "Writing on dirty paper," *IEEE Trans. Information Theory*, vol. IT-29, pp. 439–441, 1983.

[68] A. S. Cohen and A. Lapidoth, "Generalized writing on dirty paper," in *Proc. ISIT 2002*, (Lausanne-Switzerland), July 2002.

[69] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, "Writing on colored paper," in *Proc. IEEE ISIT*, (Washington D.C.), p. 302, June 2001.

[70] P. Moulin and J. O'Sullivan, "Information-theoretic analysis," in *Int. Symp. Information Theory (Sorrento, Italy)*, p. 19, June 2000.

[71] I. Cox, M. Miller, and A. McKellips, "Electronic watermarking: the first 50 years," in *Proc. Int. Workshop on Multimedia Signal Processing*, pp. 225–230, 2001.

[72] A. Lapidoth and S. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Information Theory*, vol. 49, pp. 2426 – 2467, Oct. 2003.

[73] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayleigh flat fading," *IEEE Trans. Information Theory*, vol. IT-45, pp. 139–157, Jan. 1999.

[74] L. Zheng and D. Tse, "Communication on the grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Information Theory*, vol. IT-48, pp. 359 – 383, Feb. 2002.

[75] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Trans. Information Theory*, vol. IT-45, no. 6, pp. 2007–2019, 1999.

[76] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Information Theory*, vol. IT-43, pp. 1986–1992, 1997.

[77] T. E. Klein and R. Gallager, "Power control for additive white gaussian noise channel under channel estimation errors," in *In Proc. IEEE ISIT*, p. 304, June 2001.

[78] J. Diaz, Z. Latinovic, , and Y. Bar-Ness, "Impact of imperfect channel state information upon the outage capacity of rayleigh fading channels," in *Proceeding of GLOBECOM 04*, pp. 887–892, 2004.

[79] I. Csiszár, "Sanov property, generalize I-projection and a conditional limit theorem," *Ann. Probability*, vol. 12, pp. 768–793, 1984.

[80] I. Csiszár, "Arbitrarily varying channels with general alphabets and states," *IEEE Trans. Information Theory*, vol. IT-38, pp. 1725–1742, 1992.

[81] A. Gersho and R. Gray, *Vector quantization and signal compression.* Norwell, Massachusetts: Kluwer Academic Publishers, 1992.

[82] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *Selected Areas in Communications*, vol. 16, pp. 1423–1436, Oct. 1998.

[83] G. Jongren, M. Skoglund, and B. Ottersten, "Combining beamforming and orthogonal space-time block coding," vol. 48, pp. 611–627, Mar 2002.

[84] J. Hirriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, 1993.

[85] J. Luo, L. Lin, R. Yates, and P. Spasojevic, "Service outage based power and rate allocation," *IEEE Trans. Information Theory*, vol. IT-49, pp. 323–330, Jan 2003.

[86] K. Ahmed, C. Tepedelenhoglu, and A. Spanias, "Effect of channel estimation on pair-wise error probability in OFDM," in *Proc. of Int. Conf. of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 745–748, May 2004.

[87] A. Leke and J. M. Cioffi, "Impact of imperfect channel knowledge on the performance of multicarrier systems," in *IEEE Global Telecommun. Conf*, vol. 4, pp. 951–955, Nov. 1998.

[88] P. Garg, R. K. Mallik, and H. M. Gupta, "Performance analysis of space-time coding with imperfect channel estimation," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 257–265, Jan. 2005.

[89] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Information Theory*, vol. IT-44, pp. 927–945, May 1998.

[90] E. Zehavi, "8-PSK trellis codes for a rayleigh channel," *IEEE Trans. Communications*, vol. 40, pp. 873–887, May 1992.

[91] X. Li, A. Chindapol, and J. A. Ritcey, "Bit-interleaved coded modulation with iterative decoding and 8-PSK modulation," *IEEE Trans. Communications*, vol. 50, pp. 1250–1257, Aug. 2002.

[92] J. K. Cavers, "An analysis of pilot symbol assisted modulation for rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 40, pp. 686–693, Nov. 1991.

[93] Y. Huang and J. A. Ritcey, "16-QAM BICM-ID in fading channels with imperfect channel state information," *IEEE Trans. Communications*, vol. 2, pp. 1000–1007, Sept. 2003.

[94] A. Lapidoth and S. Shamai, "Fading channels: how perfect need 'perfect side information' be?," *IEEE Transactions on Information Theory*, vol. 48, pp. 1118–1134, May 2002.

[95] J. J. Boutros, F. Boixadera, and C. Lamy, "Bit-interleaved coded modulations for multiple-input multiple-output channels," in *Int. Symp. on Spread Spectrum Tech. and Applications*, pp. 123–126, Sept. 2000.

[96] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Information Theory*, pp. 284–287, March 1974.

[97] P. Garg, R. K. Mallik, and H. M. Gupta, "Performance analysis of space-time coding with imperfect channel estimation," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 257–265, Jan. 2005.

[98] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Information Theory*, vol. 49, March 2003.

[99] N. Jindal and A. Goldsmith, "Dirty paper coding versus TDMA for MIMO broadcast channels," *IEEE Trans. Information Theory*, vol. 5, pp. 1783–1794, May 2005.

[100] S. Yang and J.-C. Belfiore, "The impact of channel estimation error on the DPC region of the two-user gaussian broadcast channel," in *Proceedings of Allerton Conf. on Commun., Control, and Comput.*, Sep. 2005.

[101] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channel with partial side information," *IEEE Trans. Information Theory*, vol. 51, pp. 506–522, Feb. 2005.

[102] A. F. Dana, M. Sharif, and B. Hassibi, "On the capacity region of MIMO gaussian broadcast channels with estimation error," in *ISIT 2006, Washington, Seattle*, July 2006.

[103] N. Jindal, "Mimo broadcast channels with finite rate feedback," *IEEE Trans. Information Theory*, vol. 52, pp. 5045–5059, Nov. 2006.

[104] T. Yoo, N. Jindal, and A. Goldsmith, "Finite-rate feedback mimo broadcast channels with a large number of users," in *Proc. of IEEE International Symp. on Information Theory*, (Seattle, USA), July 2006.

[105] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete time memoryless rayleigh fading channels," *IEEE Trans. Information Theory*, vol. IT-47, pp. 1290–1301, May 2001.

[106] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Information Theory*, vol. 40, pp. 1147–1157, 1994.

[107] N. Jindal, R. Wonjong, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna gaussian broadcast channels," *IEEE Trans. Information Theory*, vol. 51, pp. 1570– 1580, April 2005.

[108] B. Chen and G. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, pp. 1423–1443, may 2001.

[109] I. Cox, M. Miller, and A. McKellips, "Watermarking as communication with side information," in *Proc. Int. Conference on Multimedia Computing and Systems*, pp. 1127–1141, July 1999.

[110] J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, "Scalar costa scheme for information embedding," *IEEE Transactions on Signal Processing*, pp. 1003–10019, 2003.

[111] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. on IT*, vol. IT-29, pp. 439–441, may 1983.

[112] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and IT.*, vol. 9, pp. 19–31, 1980.

[113] C. D. Heegard and A. A. E. Gamal, "On the capacity of computer memory with defects," *IEEE Transactions on Information Theory*, vol. IT-29, pp. 731–739, September 1983.

[114] N. Liu and K. P. Subbalakshmi, "Non-uniform quantizer design for image data hiding," in *Proc. of IEEE Int. Conf. on Image Processing, ICIP*, vol. 4, (Singapore), pp. 2179– 2182, October 2004.

[115] R. F. H. Fischer, R. Tzschoppe, and R. Bäuml, "Lattice costa schemes using subspace projection for digital watermarking," in *Proc. ITG Conference on Source and Channel Coding*, 2004.

[116] P. Moulin and R. Koetter, "Data-hiding codes," in *IEEE Int. Conference on Image Processing*, (Singapore), October 2004.

[117] A. Zaidi and P. Duhamel, "Modulo lattice additive noise channel for QIM watermarking," in *proc of Int. Conf. Image Processing ICIP*, (Genova, Italy), pp. 993–996, september 2005.

[118] Y.-H. Kim, A. Sutivong, and S. Sigurjonsson, "Multiple user writing on dirty paper," in *Proc. ISIT 2004*, (Chicago-USA), p. 534, June 2004.

[119] B. Chen and G. Wornell, "Achievable performance of digital watermarking systems," in *Proc. Int. Conference on Multimedia Computing and Systems*, vol. 87, (Florence, Italy), pp. 13–18, june 1999.

[120] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. IT-18, pp. 2–14, Junuary 1972.

[121] T. M. Cover, "Comments on broadcast channels," *IEEE Transactions on Information Theory*, vol. IT-44, pp. 2524–2530, October 1988.

[122] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: John Willey & Sons INC., 1991.

[123] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Transactions on Information Theory*, vol. IT-48, pp. 1250–1276, June 2002.

[124] J. H. Conway and N. J. A. Sloane, *Sphere Packing, Lattices and Groups.* New York: third edition, John Willey & Sons INC., 1988.

[125] G. D. Forney, M. D. Trott, and S. Y. Chung, "Sphere-bound-achieving cosets codes and multilevel coset codes," *IEEE Trans. on IT*, vol. IT-46, pp. 820–850, 2000.

[126] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice strategies for cancelling known interference," in *Int. Symps. on IT and Its Applications, ISITA*, (Honolulu, Hawaii), pp. 681–684, 2000.

[127] G. D. Forney and L. F. Wei, "Multidimensional constellations- part I: Introductions figures of merit, and generalized crosss constellations," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 877–892, August 1989.

[128] J. G. D. Forney, "Multidimensional constellations- part II: Voronoi constellations," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 941–958, 1989.

[129] I. Csiszár, "Information projections revisited," *IEEE Trans. Information Theory*, vol. IT-49, pp. 1474–1490, June 2003.

[130] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Positivity, constraints," *IEEE Trans. Information Theory*, vol. IT-34, no. 2, pp. 181–193, 1988.

[131] P. Billingsley, *Probability and Measure*. New York, Wiley, 3rd ed., 1995.

[132] T. S. Han and K. Kobayashi, "Exponential- type error probabilities for multiterminal hypothesis testing," *IEEE Trans. Information Theory*, vol. IT-35, pp. 2–14, January 1989.

[133] J. L. Massey, "On the fractional weight of distinct binary $n$-tuples," *IEEE Trans. Information Theory*, vol. IT-20, p. 131, January 1974.

[134] M. Schwartz, W. Bennett, and S. Stein, *Communication Systems and Techniques*. New York McGraw-Hill, 1996.

[135] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*. Cambridge University Press, 1986.

[136] I. Gradshteyn and I. Ryzhik, *Table of Integrals and Products*. Academic, New York, 1965.