EDITE
n° ordre: xxxx

Thèse

présentée pour obtenir

Le grade de docteur de

l'ÉCOLE NATIONALE SUPÉRIEURE DE TÉLÉCOMMUNICATIONS
ENST

spécialité

ÉLECTRONIQUE ET COMMUNICATIONS

par

ABDELLATIF ZAIDI

Codage avec information adjacente.
Application à la transmission sécurisée de signaux multimédia dans un environnement cellulaire.

Soutenue le 13 décembre 2005 devant la commission d'examen:

Mr. Jean-Claude BELFIORE	*ENST Paris, France*
Mr. Ahmed H. TEWFIK	*University of Minnesota, USA*
Mr. Fernando PÉREZ GONZÀLES	*University of Vigo, Spain*
Mr. Benoit MACQ	*Université Catholique de Louvain, Belgium*
Mr. J. Joseph BOUTROS	*ENST Paris, France*
Mr. Pierre DUHAMEL	*LSS/CNRS, France*

To my parents,

my source of seed funding in human capital.

A man can only be happy following his own inmost need.

D. H. Lawrence

L'attrait du danger est au fond de toutes les passions.

Il n'ya pas de volupté sans vertige.

Le plaisir mêlé de peur enivre.

Anatole France

Acknowledgment

I feel very fortunate to have learned from and have worked with so many incredible individuals from both Signals and Systems Lab., National Center of Scientific Research and National School of Telecommunications, ENST Paris, during my Ph.D. study. Foremost, i would like to thank my advisor Professor Pierre Duhamel for his guidance and support throughout the past three years. I have benefited tremendously from his unique blend of energy, vision, technical insights and practical sensibility. Professor Duhamel's profound thinking, generosity and integrity will be an inspiring role model for my future career. Pierre: i am deeply indebted to your ideas and insights. I also thank my associate advisor Professor J. Joseph Boutros for being a great researcher and also a friend. Joseph introduced me to this exciting, yet energy demanding, research career five years ago. I also gratefully acknowledge the two RNRT projects SDMO and ECRYPT for funding.

I would like to thank all the members of my committee, for having kindly accepted to serve on my defense: Professor Jean Claude Belfiore from ENST, Professor Ahmed H. Tewfik from the university of Minnesota, USA, Professor Fernando Pérez Gonzàles from the university of Vigo Spain and Professor Benoit MACQ from UCL Belgium. Special thanks to the two reviewers of my thesis, Professor Tewfik and Professor Pérez Gonzàles, for their efforts, devotion and availability. Their comments and feedback have been key elements to make the final manuscript have its actual form. It has been a privilege to interact with such bright, talented and engaging people. I am deeply grateful to you, Professor Tewfik, for the many perspectives and insights you provided to me. I really appreciated your call and the informal conversation we had during the preparation of my Ph.D. defense. All my gratitude goes also to Professor Pérez Gonzàles for his careful reading and technical comments.

I am also grateful to my instructors at ENST. Much of the background for the work in this thesis was taught to me by the professors in many courses there. I thank all of them for their dedication to teaching. I also would like to acknowledge the professors at my engineering school, ENSTA. They helped me get started in electrical engineering and taught me the fundamentals.

I thank many of my colleagues with whom i have had the good fortune to collaborate. Joint work with Pablo Piantanida led to some of the results in this thesis. I also have benifited from many hours of stimulating discussions with Resa Khani, Samson Lasaulce, Michel Kieffer, Gerald Matz, Claude Delpha, and earlier with Thomas Helie. I also owe them a great deal for their friendship. During my stay at LSS, i have had opportunities to collaborate on various subjects. Some results not included in this manuscript and just mentioned in the perspective chapter have been obtained within joint work with some trainees whom i supervised. I thank all of them for their work. Special thanks to Mohammed Berry, Ismehene

Chahbi and Francesca Bassi for their work on Slepian-Wolf coding and Wyner-Ziv Coding. The joint work with some of you has been ”special for me”. Once more, i thank Pierre for letting me broaden my research interests. I am also fortunate to be in a dynamic group whose members are always fun to be with. The group members and visitors include Jean Marcel Mamfoumbi (needless to say that without Marcello and his twisted spirit, many coffee breaks would have been less funny), Mahieddine Ichir (Pasteur: i will never forget those two weeks at New York, Philadelphia and Washington with Belka), Olivier Feron, Thomas Rodet, Andrew Klein, Brice Djeumou, Belkacemi Hocine, Sophie Fourcade, Natalia Bahamonde, Marie Mathian, Anissa Zergainoh and Salma Ben Jemaa. Next, i want to thank other people who happened to share my office during those three years and who had to cope with my mood of the day: Maxime Ossonce, Jean Philippe Boyer, Vianney Munoz Jimenez and Ramin Khalili. Ramin: Many of your jokes enlightened my bad days where nothing where running fine. Good luck for your future research. I want also to thank all the researchers that i have met during conferences: their comments and feedback have been valuable to keep this thesis going forward. I specially enjoyed Slava’s enthusiasm during *after-session hours* and fruitful discussions with Professor Pierre Moulin.

Finally, and most importantly, i would like to deeply thank my family for all of their help and support. My brother Moncef has always set a good example for me, leading the way and showing me the ropes. My mom has given me all the loving care that only a mother can provide. You kept me going all these years long. My dad had always represented for me the role model, confidant, and most trusted advisor. I will never fail in believing in you, dad. I owe to you an ever-lasting debt: you will be in my heart of hearts, for ever.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Much of communication theory has been developed for point-to-point communications. In these scenarios, there is only a single source of information at each processing stage: the message at the encoder, and the received signal at the decoder. With this as the default scenario, other sources of useful information available at either the encoder and/or the decoder are called State Information or more commonly "Side" Information (SI). The central characteristic differentiating communication with side information from more conventional communication is the causal or non-causal presence of this interference-like information. The modern theory of communication over channels with causal or non-causal side information started in 1980 with Gel'fand and Pinsker [GP80]. Slightly later, in 1983, Heegard and El Gamal published their relevant work [HG83]. This result, though non-constructive, revealed the fundamental performance limit in a communication channel subject to a state information-like interference. Further, it laid down a foundation for the modern science of this new branch of digital communications. For the next twenty years, a prime goal of both communication researchers and engineers has been to devise practical methods to approach these ultimate limits, in a variety of applications fields.

Side information related systems cover a wide range of data transmission and data compression applications. Data transmission models have found practical usefulness, for example, in wireless communication where the fading coefficient is the state information at the transmitter, in capacity calculation for defective memory where the defective memory cell is the state information at the encoder, in Digital Subscriber lines (DSL) where the cross-talk created by different telephone lines bundled together on the way to the central office is the state information at the transmitter. Data compression models have found practical usefulness, for example, in Distributed Source Coding (DSC) where the noisy version of the source is the state information at the decoder, in sensor networking where a common information shared by different nodes is the state information at the decoder and in high-definition television (HDTV) systems where the noisy analog version of the TV signal is the state information at the decoder. The use of side information in data transmission and data compression is highlighted below.

The focus in this thesis is on another problem, somehow related to data transmission and data compression: Information Embedding. The beauty of information embedding lies in several aspects. First, it elegantly connects information theory to the two rich areas of coding and communication theories. Second, it readily embodies the problem of communication with state information at the encoder. Third, implementable solutions for the problem of information embedding provide useful insights into the dual problem of source coding with state information at the decoder.

1.1 Information Embedding and Related Problems

Information Embedding deals with the problem of transmitting one signal, generally weak, within another, generally strong. The code or signal designed for the transmission is called ”embedded code” or ”embedded signal”. The use of embedded codes has a number of important multimedia applications. The most important is digital watermarking. Transmission of signals by embedding them into other signals is a ”non-conventional” transmission problem. However, it uses guidelines from, and also provides insights on, more conventional data transmission and data compression.

1.1.1 Digital Watermarking

Digital watermarking is a major branch of information embedding. It can be defined as the *imperceptible, robust, secure communication of information* by embedding it in, and retrieving it from, the original cover signal. The imperceptibility requirement refers to the fact that embedding should not (and must not) cause perceptible damage or distortion to the host signal. The robustness requirement refers to the ability of the embedded signal to survive intentional or non-intentional channel degradations. The basic idea is that the embedded information, i.e., the watermark message, travels with the multimedia data wherever the watermarked data goes. The security requirement refers to the ability for this embedded information to convey to the receiver personal or private information about the embedder, with no possibility of being replaced, falsified or altered by a non-authorized third party. A straightforward application strongly linked to digital watermarking is that of ownership proof and copyright protection. In this application, the ”weak signal” (”embedded signal” or ”watermark”) notifies and enforces the ”strong signal” (”host or cover” signal) against unauthorized copying or duplication. This problem arises due to the relative ease with which multimedia contents can be created and distributed. Typically, the digital watermark is embedded into the multimedia content- an audio signal, a video signal, or an image, for example- and (i) identifies the content owner or producer, (ii) identifies the recipient or purchaser, (iii) enables a standards-compliant device to either play or duplicate the content, or (iv) prevents a standards-compliant device from playing or duplicating the content. The problem of ensuring copyright of multimedia at the client side lies in the fact that traditional data protection technologies such as encryption or scrambling cannot be applied exclusively as they are prone to digital copying or analog-re-recording (this is the commonly referred to as "analog hole"!!). In addition to being easily duplicated, digital multimedia signals are also easily altered. Thus, authentication of, or detection of tampering with, multimedia signals is another application of digital watermarking methods. So-called "fragile" watermarks change whenever the composite signal is "sufficiently" altered, thus providing a means of detecting tampering. Other applications, also based on information embedding principles ans sometimes stated as alternative resorts for security purposes and data integrity issues, include covert communication (called also "steganography") and, more generally, low probability of detection communications [CW01, Ram98].

1.1.2 Conventional Data Transmission

The use of information embedding techniques is not limited to digital watermarking and security applications. Other applications from more conventional communication borrow the same principles, though applied in a slightly different manner. For instance, it has recently been recognized [CDW01, BBCS05, KSS04] that codes based on information embedding can be used as scalable alternatives to superposition codes for the Broadcast Channel (BC). Also, information embedding codes have potential applications in Multiple Access Channel (MAC) problems [KSS04]. Yet, multiple-antenna communication systems in general, and especially in a multi-user network environment, promise an intensive use of Dirty-Paper-Coding techniques (which are inherently linked to information embedding) [CS03, VT03, VJG03]. Moreover, though not always recognized as such, bandwidth-conserving hybrid transmission relies on information embedding techniques so as to make possible the re-use and sharing of existing spectrum and bandwidth. The aim is to either backwards-compatibly increase the capacity of an existing communication network, i.e. a "legacy" network, or allow a new network to be backwards-compatibly overlayed on top of the legacy network [CDW01]. In this case, the host signal, being the signal corresponding to the legacy network, and the embedded signal are two different signals that are multiplexed, i.e., transmitted simultaneously over the same channel in the same bandwidth.
In fact, embedded codes can have significant advantages over other codes from other classes, already recognized as being "good enough", in many other scenarios of interest.

1.1.3 Conventional Data Compression

As an important information-theoretic duality between data transmission and data compression do exist, information embedding techniques have also promising potential use in data compression and source coding, with side information at the decoder. Embedded codes shed light on a great variety of challenging information-theoretic source-coding problems. For instance, the problem of optimally encoding (maximally compressing) one source flow so as to be reliably reconstructed (i.e, with sufficient quality) at a certain distant decoder where a "noisy version" of the source is made available has found in information embedding codes "potential" good candidates for coding. Also, the problem of separately encoding different flows from different sources and transmitting them so as to be gathered and jointly decoded at a single remote source has potential solutions in embedded codes [DW04].

In fact, data compression and data transmission being ”two extreme points of communication theory”, or equivalently somewhat dual, as Cover asserted in [CT91], embedded codes play, in data compression, the role they play in data transmission.

1.2 Motivations

This thesis deals with the design of information embedding techniques, tailored so as to be used in a digital watermarking and data hiding context. Apart from some of its additional requirements (imperceptibility, for example), information embedding shares the same principles, goals, strengths and weaknesses with conventional communication. It also, thereby, faces the same trade-offs between, for example, the rate at which information can be transmitted and the probability of error in recovering it, or also, between the optimality and feasibility of codes designed for. The solutions for the problems encountered, however, are not usually the same, or more precisely, not exactly the same.

As a recent research topic, information embedding has lack of solutions to a great number of new problems. The essential of this work is devoted to finding solutions to some of these. We concentrate on the following:

1. Design efficient constructions and provide new insights into the conception of high-rate/low-error data embedding techniques. Maximizing the amount of information that one could transmit within a given signal, with sufficiently low probability of error, is currently hard to achieve by straightforward application of the already existing techniques.

2. Conceive efficient coding strategies so as to reliably transmit different embedded signals within the same cover signal. Transmitting different signals within the same host naturally raises in, for example, situations where different watermarks are either directed to different usages (tampering, identification, authentication) and/or encoded separately, by different entities.

3. Show how these techniques, designed so as to be efficient in the situations where the channel is perfectly known, should adapt in case of a certain channel uncertainty. In a context of information embedding, uncertainties on the channel may be caused, for example, by some imperfect knowledge of the host signal itself and/or some (intentional or non-intentional) channel variations.

Due to its connections to a variety of research areas, among them information theory, communication theory, multimedia signal processing, statistical estimation and mathematics, different approaches are possible to address the above mentioned problems. In this work, we principally follow information theoretic and communication points-of-view. Some aspects from other approaches are sometimes considered, but only partially. Such line of work is motivated by the following reasons.

(i) Though recognized as such, the analogy with conventional transmission has, until now, been exploited only little. Information embedding, which is a new research topic, should however rely on the strong background of standard communication. This is true for code design, power allocation and interference subtraction, for example. Furthermore, multi-user information embedding which is still a new research topic should get benefit from recent advances in network information theory.

(ii) Conversely, information embedding, by means of its simple communication model (Gaussian channel, non-causal state, no precoder, single channel instead of parallel channels, ...) should provide basic coding principles to more conventional, but more complex, communication scenarios.

(iii) As the array of applications, richness of the theory and interconnections to other problems continue to grow, the results in this thesis may find practical usefulness in a number of related problems.

1.3 Thesis Summary

In this work, we concentrate on the design of coding strategies for reliable transmission of large amount of information over an information embedding channel. The channel can be (i) single-user, i.e., one "watermark" or "embedded signal" transmitted from one point to another (ii) multi-user, i.e, different watermarks or embedded signals directed to different usages, (iii) known perfectly, i.e., the transmitter has full knowledge of the channel or (iv) known imperfectly, i.e., the transmitter knows the channel, only with a certain uncertainty. The coding strategies developed for these situations are built in part upon Costa's famous Dirty-Paper Coding (DPC) [Cos83] and tie in with a growing body of work focusing on side-information coding fundamentals, constructions and dualities.

Chapter 2: This chapter states the general problem of coding with state information together with its Gaussian version, equivalently known as Costa problem, DPC or "Writing on Dirty Paper" [Cos83]. We also give a short review of the application of interest, with a particular emphasis on information embedding. Also, we provide in this chapter a parallel between binning-based codes, usually used for information-theoretic analysis, and low-complexity algebraic codes, suboptimal but more feasible in practice. In particular, two feasible schemes referred to as Scalar Costa Scheme (SCS) and Quantization Index Modulation (QIM), respectively, are reviewed in details. These schemes will be used as baseline for performance comparison throughout this thesis.

Chapter 3: This chapter is composed of two parts. In the first part, we heavily rely on the work in [ESZ00] to extend scalar-codebook based techniques (SCS and QIM) to the case of lattice codebooks. Lattice-based codebooks should be regarded as "Multidimensional constellation" with respect to scalar codebooks, which can be viewed as Pulse Amplitude Modulation (PAM) constellations. However, by opposition to infinite dimensional lattice coding considered in [ESZ00], we are interested in finite dimensional implementable solutions. In particular, we address the problem of finding good trade-offs between the amounts of information one signal can carry (payload or transmission rate) and the reliability by which this information can be recovered at the receiver. This naturally leads to the problem of codebook selection which we address, first through some examples, based on the algebraic structure of the lattice and then, through a more general approach. The total gain provided by the resulting codebook, over scalar codebooks (cubic lattice), is measured by both coding and shaping gains. We show in this part of the chapter that these quantities are not decoupled in finite dimensional embedding, but rather interacting. In the second part of this chapter, we first argue that the problem of information embedding is indeed a joint source-channel coding problem. We then provide means (through an example) of constructing good practical nested codes.

Chapter 4: While emphasizing the tight relationship with conventional multiple user information theory, we present in this chapter several implementable DPC-based schemes for multiple user information embedding. We first show that depending on the targeted application and on whether the different messages are required to have different robustness and transparency requirements or not, multiple user information embedding parallels one of the multi-user channels with state information available at the transmitter, for which recent theory is well developed. The focus is on the physically degraded 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. The joint approach is based on the ideal DPC for the corresponding channel. These results extend the practical implementations QIM, DC-QIM and SCS that have been originally conceived for one user to the multiple user case. After presenting the key features of 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 performances can be bridged up using finite dimensional lattice codebooks. Performance evaluations, including Bit Error Rates (BER) and capacity region curves are provided for both methods, illustrating the improvements brought by a joint design.

Chapter 5: This chapter is concerned with evaluating channel capacity sensitivity to the imperfect knowledge of the state information. In coding with state information applications, this may occur in the situations where there is a certain mismatch between the true state information taken into account at the encoder and that seen by the decoder. In information embedding applications, noisy host signals connect Data Hiding coding strategies to the problem at hand. In general, this leads to a performance degradation. In this chapter, we consider the general case of channel sensitivity to two-sided noisy state information: S1 known at the encoder and S2 known at the decoder. The problem of information embedding specializes to the case where S2 is null. We first consider the Gaussian case and show that closed form expressions for channel capacity degradation, due to some unknown perturbing noise, do exist. We then address the case of arbitrarily distributed signals and we show that (under appropriate assumptions), both lower and upper bounds on channel capacity decrease can be found. This is made possible by using De Bruijn Identity which connects Entropy to Fisher Information. The tightness of these bounds is discussed. Coding with a nominal state information slightly perturbed by a weak noise finds applications in many practical situations. Examples include watermark channels subject to time delay desynchronization, where the receiver may not be fully synchronized with the transmitter. Another example concerns the situations where the encoder has access to only a short description (a quantized version, for example) of the state information.

Chapter 6: In this chapter, we consider an alternative definition of channel capacity, using a game theory approach. This involves a min-max optimization problem between the embedder (encoder/decoder) and an eventual attacker (channel). The set of parameters over which the payoff function is optimized are the induced distortions. These distortions have then to be properly measured so as to lead to accurate solutions of the optimization problem. In this chapter, we first provide means of evaluating these channel distortions. We then evaluate the capacity loss of common information embedding systems when facing an important class of channel attacks, amplitude scaling plus additive noise. Analysis is specialized to the situation when communication can be modeled by transmission over an Additive White Gaussian Noise and Jitter (AWGN&J) channel. The second part of this chapter concentrates on finding optimal embedder (encoder) and attacker (channel) strategies. The payoff function is the detection probability and embedding is based on Spread Spectrum. The embedder wants to reliably transmit information, under any distortion constrained channel attack strategy. Conversely, the attacker wants to impair this transmission for any power constrained information embedding strategy.

Chapter 7: This chapter presents a practical information embedding application, treated within the context of Secured Diffusion of audio contents (Music) in a mObile cellular network, (SDMO). Of special interest are the following problems.

1. Transmission Rate evaluation. The results in this part heavily rely on the materials in Chapter 3.
2. Embedding two different watermarks within the same host signal. The two watermarks are intended to two different usages. The robust watermark aims at ownership identification whereas the fragile watermark aims at identifying tempering. The results in this part heavily rely on the materials in Chapter 4.
3. Channel capacity sensitivity to jitter-like attacks. The results in this part heavily rely on the materials in Chapter 5.

Chapter 8: We conclude this thesis in Chapter 8, where we also discuss some possible directions for future work.

Chapter 2

Information Embedding and Coding With State Information at the Encoder

In this chapter, we first present a communication model for information embedding. We also discuss its tight relationship with conventional communication. It is shown that information embedding can be viewed as an instance of communication over channels with state information (SI) non-causally available at the transmitter, a situation which is commonly known as "Gel'fand-Pinsker problem". More precisely, coding for information embedding amounts to the Gaussian version of Gel'fand-Pinsker setting, also known as "Costa problem". Both theoretical techniques, optimal but computationally non feasible, and practical implementations, sub-optimal but computationally feasible, are discussed.

2.1 Information Embedding

Consider a host signal vector $\mathbf{s} \in \mathbb{R}^n$ into which we want to embed some information m. In typical information embedding applications (digital watermarking, for example), this host signal could be a vector of pixel values, text, audio or speech samples. Alternatively, it could be the representation of the host signal in some transform domain (such as discrete cosine transform coefficients and wavelet coefficients). Typically, the message m can be a watermark or an authentication signal, as in classical ownership-proof applications. In most recent applications, the host signal $\mathbf{s}$ can be any block of data from a given *host data set* and the message m can be any information one would want to transmit within this data. In our attempt to emphasize the very general framework, we only ask the cover signal $\mathbf{s}$ to be strong "enough" so as to be able to "carry" the message m, with no particular assumption regarding the nature of signals, their statistical distributions and/or their use. We assume that the samples of the host signal $\mathbf{s}$ take values into a finite cardinality set $\mathcal{S}$, i.e., $|\mathcal{S}| < \infty$, and that the message m is an integer, taking values into a certain alphabet

$$\mathcal{M} = \{1, 2, \ldots, M\},$$

of cardinality $M = |\mathcal{M}|$. We wish to embed at rate R bits per dimension (i.e, bits per host sample) so that, if each index (or message) is embedded into a n-length vector of the host, the embedding rate is given by

$$R = \frac{1}{n} \log_2(M). \tag{2.1}$$

The transmitter wants to embed m into the cover signal $\mathbf{s}$, with, hopefully, no "serious" degradation caused to the host itself. The receiver wants to reliably recover the transmitted message, hopefully, even if the signal was altered in the channel. The received signal may be altered either by non-intentional channel degradations (e.g., ambient noise) and/or by deliberate manipulations due to some malicious *attacker* in the channel. The aim of this eventual attacker may depend on the application. In ownership applications for example, a possible attack may consist in falsifying the watermarked content, seriously corrupting it or even completely destroying it. Alternatively, degradations to the watermarked content may be due to *legal* signal manipulations such as compression, conversion A/D, etc... *Intentional* attacks concern digital watermarking applications, mainly. In almost all other information embedding applications, degradations that the embedded signal encounters are most of the times non-intentional, i.e, simply due to the ambient noise. The impairment caused to the composite signal, in the channel, may be measured, qualitatively, by the difference in quality between the received and the transmitted signals.

2.1.1 Mathematical Model for Information Embedding

An information embedding system may be represented by the block diagram shown in Fig.2.1. The message m to be embedded is chosen from the alphabet $\mathcal{M}$, i.e., $m \in \{1, 2, \ldots, M\}$. The composite signal $\mathbf{c} = \mathbf{x} + \mathbf{s}$ transmitted over the channel is subject to a variety of channel degradations. Given some received signal $\mathbf{y}$, possibly corrupted, the decoder outputs an estimate $\hat{m}$ of the transmitted message m. Basically, the information embedding system consists in an encoding (i.e., embedding) function and a decoding (i.e., recovering) function. Embedding an index $m \in \mathcal{M}$ into (a vector of) the cover signal $\mathbf{s} \in \mathcal{S}^n$ amounts to mapping m and $\mathbf{s}$ to a certain, properly chosen, composite signal $\mathbf{c} \in \mathcal{C}^n$. A cryptographic key $\mathbf{k} \in \mathcal{K}^n$, not shown in Fig.2.1, can be used as a source of common randomness that is known to the decoder, in order to secure the transmission. We denote by $X^n(\cdot, \cdot)$ the encoding function,

$$\begin{aligned} X^n \quad : \quad & \mathcal{S}^n \times \mathcal{M} \times \mathcal{K}^n \longrightarrow \mathcal{X}^n \\ & (\mathbf{s}, m, \mathbf{k}) \longmapsto \mathbf{x}. \end{aligned} \tag{2.2}$$

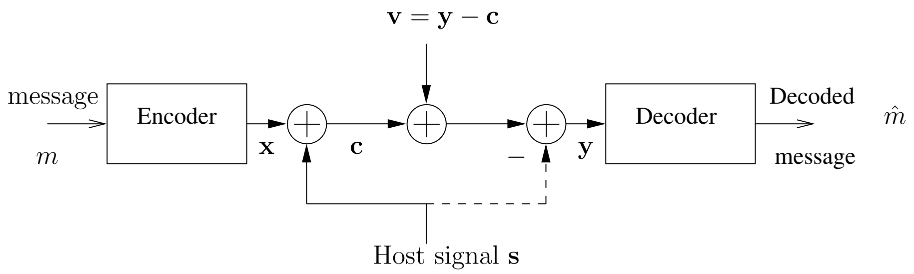


Figure 2.1: An abstract communication model for blind (solid line) and non-blind (dashed line) information embedding.

The receiver receives the signal $\mathbf{y}$, sum of the composite signal $\mathbf{c}$ and some perturbation vector $\mathbf{v}$ due to channel degradations. Two different situations arise. If the receiver has access to the original host signal $\mathbf{s}$, acting as channel interference, it first subtracts this interference and then estimates m, from the remaining signal. Decoding with knowledge of the original host at the receiver is referred to as "non-blind" information embedding. Non-blind information embedding is of practical usefulness in only some few specific applications where only some "authorized" or "privileged" receivers should successfully decode the transmitted message. A most common (but alas more complex!!) situation is when the receiver has no access to the original host signal $\mathbf{s}$ and is referred to as "blind" information embedding. In an unifying approach, we denote by $\mathbf{y} \in \mathcal{Y}^n$ the sequence from which the message m is decoded and, loosely, continue to refer to it as "received signal". Thus, this received signal $\mathbf{y}$ is either the output of the channel (blind case), or the output of the channel from which the host is already removed (non-blind case). The role of the decoder is to reliably recover the transmitted message m , even in presence of channel perturbations $\mathbf{v}$. The decoding (or recovering) function

$$\begin{aligned} \hat{W} \quad : \quad & \mathcal{Y}^n \times \mathcal{K}^n \longrightarrow \mathcal{M} \\ & (\mathbf{y}, \mathbf{k}) \longmapsto \hat{m} \end{aligned} \tag{2.3}$$

is such that $\hat{m} = \hat{W}(\mathbf{y}; \cdot)$ is the best estimate of m.

2.1.2 A Non-Conventional Power-limited Channel

In point-to-point communication, a conventional communication channel has one input and one output. The degradation represents the distortion due to this channel. The channel model shown in Fig.2.1 is non conventional, for at least four reasons.

(i) First, the channel has two inputs: the message m to be transmitted and the host signal $\mathbf{s}$, as it is noticeable from (2.3). The second input (i.e., the host $\mathbf{s}$) plays also the role of interference in the channel.

(ii) Second, the model of channel degradations is sufficiently general to include both random and deterministic perturbation vectors and both signal-dependent and signal-independent perturbation vectors. This makes the set of admissible channel perturbations larger than that in classical communication. The ability of the receiver to recover the transmitted message, even under severe channel conditions, characterizes the system *robustness* to channel degradations. Channel degradations are measured by, for example, the well known squared distortion-measure

$$D_a \triangleq \mathbb{E}_{\mathbf{V}}\{\|\mathbf{y} - \mathbf{c}\|^2\}. \tag{2.4}$$

(iii) Third, embedding the message m into the host $\mathbf{s}$ should not cause serious degradations to the cover signal. This means that, while carrying m, this cover signal should remain "useful". The distortion introduced by the encoding process is measured by, for example,

$$D_E \triangleq \mathbb{E}_{\mathbf{X}}\{\|\mathbf{x}\|^2\}. \tag{2.5}$$

That the embedding should cause no perceptible distortion to the host is sometimes called the *transparency requirement*. The imperceptibility of the embedded signal $\mathbf{x}$ should be guaranteed by means of some perceptual analysis previous to the embedding operation, something which is intrinsically dependent on the type of host signal in question.

(iv) Fourth, by opposition to conventional communication where recovering the signal $\mathbf{s}$ at the receiver is not required, one may be interested, in information embedding, in both reliably transmitting the message m and also recovering the cover signal $\mathbf{s}$. This branch of information embedding is sometimes called "reversible information embedding".

The transparency requirement means that the composite signal should look like the original. One way to express this concept of "resemblance" is to bound the energy (or equivalently the variance) of the embedded signal $\mathbf{x}$. Denoting by P the maximal tolerable embedding distortion, this implies that

$$\mathbb{E}_{\mathbf{X}}\{\|\mathbf{x}\|^2\} \leq P. \tag{2.6}$$

Of course the bound P on tolerable distortions is host-signal dependent. Intuitively, one can expect that "strong" host signals can carry much information, thus allowing larger P. The aim of an information embedder (the encoder) is to reliably transmit the maximum amount of information, for a given distortion budget. Equivalently, it can be viewed as minimizing the incurred distortion, for a given amount of information to transmit. Hence, one can view information embedding problems as "non-conventional" power-limited communication. Obviously, this non-conformity makes information embedding somewhat non-common and imposes additional constraints on system design. However, information embedding can also be viewed as being conventional.

2.1.3 A Conventional Communication Channel

We mentioned above that an information embedding system can be "blind" or "non-blind". In the "non-blind" case, the overall system amounts to transmitting an index (message) or an ensemble of indexes through a noisy channel. For that, the encoder devises an appropriate embedded signal $\mathbf{x}$ and transmits it through the channel. The receiver tries to recover the transmitted message from the noisy (possibly deliberately corrupted) signal, as in classical communication. In the blind case, the situation is different, in that the host signal (which is a part of the channel) carries the transmitted message but, at the same time, may inhibit its retrieval. Of course, simply ignoring the presence of the host signal and designing the embedded signal independently (i.e., based only on the bounded-energy constraint) would make the system seem more "conventional". However, this is not preferred as it will be shown later in this chapter.

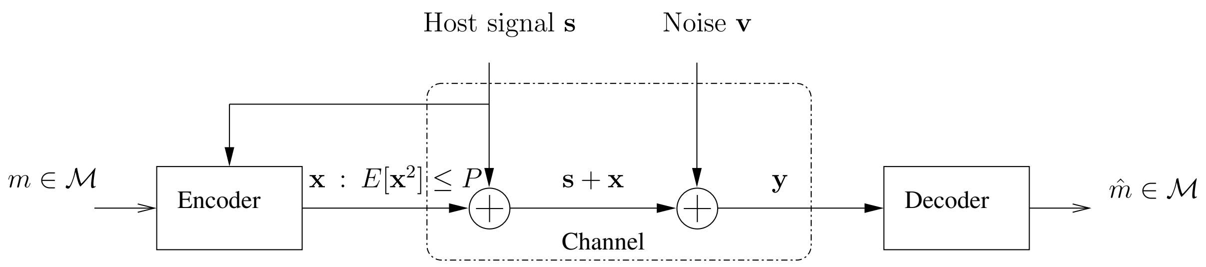


Figure 2.2: Blind information embedding viewed as communication over a channel with side information at the encoder.

Recently, Cox et Al. [CMM99] have recognized that one may view blind information embedding as communication with state information known at the encoder. In such a description, it is the host signal itself which is considered as side information. Communication with state information, either at the encoder and/or at the decoder, had already strong background and had already captured a vast amount of attention when Cox made his statement. However, in information embedding, one can consider Cox work as the starting point. Considering the cover signal as state information at the encoder makes the information embedding problem equivalent to communication over a channel, a part of which is known at the transmitter. With this view, the model in Fig.2.1 simplifies to that in Fig.2.2. In this model, the channel comprises both the noise $\mathbf{v}$ due to intentional and/or non-intentional degradations and the host itself, viewed as interference. Of course, even when considered as communication with state information available at the transmitter, information embedding remains somehow specific, by its additional requirements stated above. However, some of these simply translate to constraints (generally upper bounds) on the communication parameters. For instance, the transparency requirement is just a channel-input power constraint. The robustness requirement, as for it, states a bound on the strength of admissible channel degradations.

Considered as this, information embedding turns to be a conventional communication problem. Hence, the performance of a system designed for embedding information may be measured by that of the equivalent (conventional) system, viewed as a system for transmitting information. Common performance criteria are channel capacity and probability of error. These are classically defined as follows.

1. Following Shannon original work [Sha49], and more precisely his "channel coding theorem", the "information" channel capacity is defined as the supremum of all achievable rates. A rate R subject to distortion D_E is said to be achievable if there exists a sequence of codes $(\mathcal{M}, X^n, \hat{W})$, $n \geq 1$, subject to distortion D_E, with rate R such that the maximal probability of error $P_e^{(n)}$ tends to 0 as $n \longrightarrow +\infty$.

2. The probability of error $P_e^{(n)}$ is defined as the *average* (over the index m) probability of decoding some index $m' \neq m$ given that $m \in \mathcal{M}$ was transmitted. Assuming uniform distribution of the

messages over the set $\mathcal{M}$, this is given by

$$P_e^{(n)} \triangleq \frac{1}{M} \sum_{m \in \mathcal{M}} \Pr\left\{\hat{W}(\mathbf{y}, \mathbf{k}) \neq m | m\right\}, \tag{2.7}$$

where $\hat{W}(\cdot;\cdot)$ is the decoding function defined above.

Now that connection with communication with side information is established, performance of information embedding systems can be studied within this framework. We will first consider the basic concept of coding for channels with state information at the encoder, with a particular emphasis on its Gaussian version, commonly known as ”Costa problem”. Also, throughout the rest of this thesis, we will interchangeably use the terms ”transmission”, ”embedding” and ”communication”. When dealing with transmission, the term ”reliable” or ”reliably”, extensively used in this work, can mean either that one can guarantee that $\hat{m} = m$ or that the probability of error $P_e^{(n)}$ is small enough.

2.2 Channel Coding with State Information at the Encoder

In the context of coding and transmission, the terms ”channel state”, ”state information” and ”side information” equivalently refer to the situation when, apart from the message at the encoder and the received signal at the decoder, there is an additional source of information, available either at the transmitter or at the receiver. The use of this state information depends on whether this information is made available at the encoder or at the decoder. For instance, whereas the receiver generally observes the channel state in a non-causal manner (for it can always wait until the end of the transmission before decoding), the transmitter can observe the Channel State Information (CSI) causally or non-causally. In the causal case, the transmitter at time n knows the CSI sequence from time 1 to n only. In the non-causal case, the transmitter observes the entire CSI sequence before the transmission of any symbol begins. The prime works on coding for channels with non-causal CSI began with Gel'fand and Pinsker [GP80], in the case where the state information is known to the transmitter and with Heegard and El Gamal [HG83], in the case where the state information is known to the receiver. Earlier, in 1958, Shannon [Sha58] suggested optimal coding for channels with causal state information at the encoder. As mentioned before, information embedding is an instance of channel coding with CSI non-causally known to the transmitter, a situation which is commonly known as ”Gel'fand-Pinsker problem”.

2.2.1 Gel'fand-Pinsker Problem

Consider the channel model shown in Fig.2.3. We wish to send an index $W \in \{1, 2, \ldots, M\}$ to the receiver, in n uses of the channel. When neither the sender nor the receiver knows the state information $S^n = (S_1, S_2, \ldots, S_n)$, channel capacity is provided by the Shannon mutual information [CT91] that the channel output Y^n conveys about its input X^n, i.e.,

$$C_{00} = \max_{p(x)} I(X;Y), \tag{2.8}$$

where, as in the rest of this chapter, the first subscript under capacity C denotes the availability of state information to the sender, and the second subscript that to the receiver. $I(\cdot;\cdot)$ denotes Shannon mutual

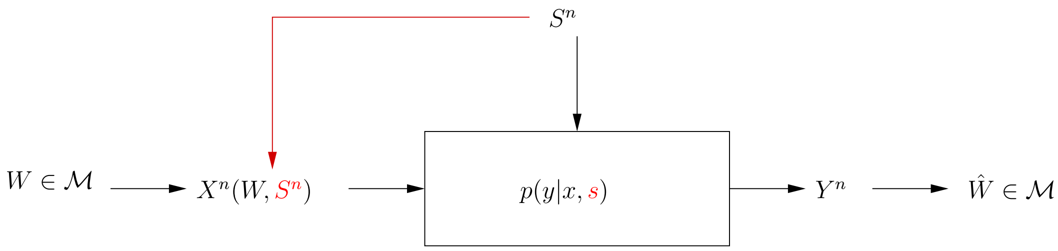


Figure 2.3: Channel with non-causal state information at the transmitter (Gel'fand-Pinsker Problem).

information. When only the sender non-causally knows the state information $S^n = (S_1, S_2, \ldots, S_n)$, channel capacity has been established by Gel'fand and Pinsker in [GP80] and can be formalized as in the following theorem.

Theorem 1 (Gel'fand and Pinsker [GP80]) *The capacity of a Discrete Memoryless Channel (DMC) with input X and output $Y = X + S + V$, where S is non-causally known to the transmitter is given by*

$$C_{10} = \max_{p(u,x|s)} \{I(U;Y) - I(U;S)\}, \tag{2.9}$$

where the maximum is over all joint distributions of the form $p(s)p(u,x|s)p(y|x,s)$ and U is a random variable taking values in a bounded cardinality set $\mathcal{U}$ ($|\mathcal{U}| \leq |\mathcal{X}|+|\mathcal{S}|$) and chosen such that $U \to (X,S) \to Y$ form a Markov Chain.

The proof of this result can be found in [GP80]. Only the direct coding theorem (achievability) is briefly reviewed below for it uses a random binning argument that outlines the key ideas behind the code construction undertaken in this thesis. The ingredients needed for the proof such as typicality and joint typicality notions are briefly defined in Appendix A. The general form of (2.9) inspires the following comments.

1. Channel capacity (2.9) can be understood in light of the following rough argument. The main idea is to transfer the information conveying role of the channel input X to some fictitious input U so that the channel behaves like a discrete memoryless channel $U \longrightarrow Y$. The capacity of this fictitious channel is obtained by maximizing the Shannon mutual information $I(U;Y)$ which represents the total number of bits, per channel use, that can be transmitted through the channel. However, there is a cost for such a transfer: $I(U;S)$ bits, per channel use, have to be allocated to the state itself. The difference $I(U;Y) - I(U;S)$ is the number of bits, per channel use, that can be allocated to the index W.

2. The joint distribution $p(u,x,s,y)$ can be restricted to the form $p(s)p(u|s)p(x|u,s)$. The marginal distribution $p(x|u,s)$ can be taken as a deterministic function $x = f(u,s)$ without loss of capacity [GP80], i.e.,

$$p(u,x,s,y) = \begin{cases} p(s)p(u|s)p(y|x,s) & \text{if } x = f(u,s) \\ \\ 0 & \text{otherwise} \end{cases} . \tag{2.10}$$

3. The conditional distribution $p(u,x|s)$ describes the coding strategy that achieves all rates less than $I(U;Y) - I(U;S)$. The marginal distribution $p(u)$ is used to create the codewords and the

conditional distribution $p(x|u,s)$ is used to form the input sequence $\mathbf{x}$ from the codeword $\mathbf{u}$ and the known noise sequence $\mathbf{s}$.

4. Though initially established for discrete memoryless channels (DMC), Gel'fand-Pinsker capacity expression (2.9) can be extended to memoryless channels with discrete time and continuous alphabets [Gal68] by considering the supremum of $I(U_d;Y_p) - I(U_d;S_q)$ over all finite alphabet variables U_d and all partitions Y_p and S_q of the channel output and state alphabets.

2.2.1.1 Direct Coding Theorem and Random Binning

The idea of binning is inherent to coding theory, especially for theoretical analysis and represents a key element in the solutions of information network problems. This section provides a brief overview. The basic principle consists in partitioning a given set of sequences or codewords, drawn according to a certain probability mass function, into different sub-sets. Each sub-set is then used to identify either a message (index) to be transmitted (in data transmission applications) or a source vector to be quantized (in data compression applications). However, dividing a set of codewords into smaller sub-sets (called also bins) must obey some combinatorial requirements, so as to be efficient in binning for coding. For instance, the overall codewords as well as the codewords collapsed inside each bin must not only have the appropriate probability mass function, but also the appropriate cardinality (number of codewords). A brief description can be found in Appendix A and a thorough focus is available in [CT91].

Proof 1 (Gel'fand and Pinsker [GP80]) *We wish to show that, for any $\epsilon > 0$ and sufficiently large n, there exists an (n,M)-code with probability of error $P_e \leq \epsilon$ and $M \geq 2^{n[I(U;Y)-I(U;S)-\epsilon]}$. The rigorous proof is rather lengthy. The main idea can be shortly exposed as follows. For any message $m \in \mathcal{M}$, choose $J = 2^{n[I(U;S)+\delta]}$ ($\delta > 0$ is small) words $\mathbf{U}_{j,m} \in \mathcal{U}^n$, indexed by j, $j = 1, 2, \ldots, J$, with distribution $p(\mathbf{U})$. Then with probability close to 1 for any typical word $\mathbf{S} \in \mathcal{S}^n$ and for any $m \in \mathcal{M}$, one can find at least one word $\mathbf{X}_{j,m}$ such that $\mathbf{S}$ and $\mathbf{U}_{j,m}$ are jointly typical, for the joint distribution $p_{\mathbf{US}}$.*

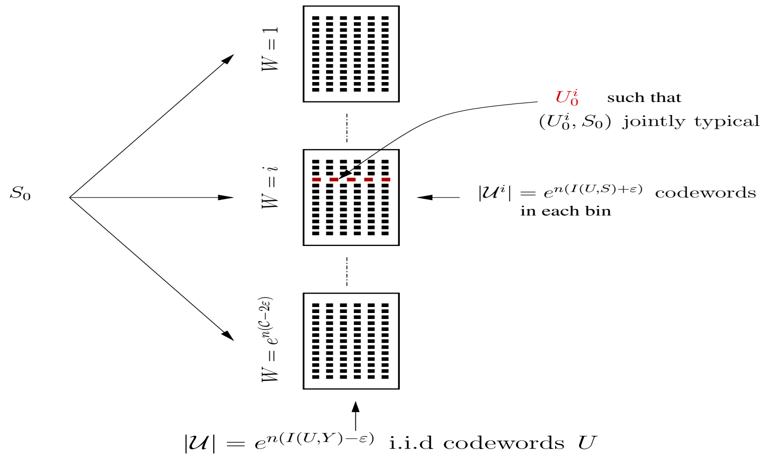


Figure 2.4: Illustration of the generation of probabilistic codes U for the solution of Gel'fand-Pinsker problem.

An illustration of the binning-based generation of the codes that achieve channel capacity (2.9) is shown in Fig.2.4. Note that in general, there is some loss in performance in not knowing the state information

S at the decoder as well, meaning that the one side transmitter state information capacity C_{10} is inferior to the two-sided state information capacity C_{11}, i.e.,

$$C_{10} \leq C_{11}. \tag{2.11}$$

The term "two-sided" state information refers to the case where the state information is available at both the transmitter and the receiver. The situations in which equality holds in (2.11) are referred to as having a "public-private equivalence" (PPE) property. PPE refers to the non-decrease of channel capacity if the decoder has or not access to the state S in the model shown in Fig.2.2. Situations with PPE are relevant for practical usefulness. An example of such situations, "Costa problem", has been addressed by Costa in [Cos83].

2.2.2 Gaussian Channel: Costa Problem

Costa problem is an instance of Gel'fand-Pinsker problem. It corresponds to the situation when (i) both the state information S and the channel noise V are independent identically distributed (i.i.d.) Gaussian and, (ii) the input X is power-constrained, i.e.,

$$\mathbb{E}[X^2] \leq P. \tag{2.12}$$

Costa problem focuses on both the theoretical capacity limit and coding strategies to achieve it. Several obvious encoding schemes to communicate over Costa channel can be considered. These schemes may be justified by the temptation to reduce the problem to a classic one. For instance, the following coding strategies can be envisaged.

1. The transmitter could attempt to pre-subtract the interference S at the transmitter and transmit $X' = X - S$. The received signal would then be $Y' = X' + S + V = X - S + S + V = X + V$, thus eliminating the interference. However, the problem with this naive approach stems from the power constraint: assuming X and S to be independent, the average transmit power would be $\mathbb{E}[X'^2] = \mathbb{E}[X^2] + \mathbb{E}[S^2]$. As the interference S may be arbitrarily strong, this would entail a severe power penalty, and hence a reduced transmission rate.

2. Similarly, the transmitter could attempt to use a fraction $0 \leq \alpha \leq \min\{1, Q/P\}$ of the transmission power P to partially cancel S, i.e., transmit $X' = X - \sqrt{\frac{\alpha P}{Q}}S$. The received signal would be $Y' = X + (1 - \sqrt{\dfrac{\alpha P}{Q}})S + V$. This approach would yield a transmission rate of only $\dfrac{1}{2}\log\left(1 + \dfrac{(1-\alpha)P}{N + (\sqrt{Q} - \sqrt{\alpha P})^2}\right)$.

In fact, these two schemes are two forms of one single intuitive procedure which consists in erasing (a part of) the interference prior to transmitting. Surprisingly, Costa showed in [Cos83] that the optimal encoder should not fight against the side interference S by pre-subtracting it. Instead, it should use it constructively. Further, Costa showed that, by doing this, the power constrained encoder can reliably transmit all rates less than $\frac{1}{2}\log(1 + \frac{P}{N})$ bits/symbol, independently on the strength of the interfering signal S. Costa result can be formalized as in the following theorem.

Theorem 2 (Max H. M. Costa [Cos83]) *The capacity of the channel $Y = X + S + V$, where $S \sim \mathcal{N}(0, QI)$, non-causally known to the transmitter, and $V \sim \mathcal{N}(0, NI)$ are multivariate Gaussian random variables (I is the identity matrix) and the input $X \in \mathbb{R}^n$ satisfies the power constraint $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P$, is given by*

$$C_{10} = \frac{1}{2}\log\left(1 + \frac{P}{N}\right). \tag{2.13}$$

Prior to dealing with the proof, note that the capacity (2.13) is that of an Additive White Gaussian Noise (AWGN) channel with Signal-to-Noise Ratio (SNR) of P/N [dB]. Also, note the surprising fact that this capacity depends on the variance of the known noise S. Hence, the achievable rates would not change if the state information, acting as noise, were not present or were also known at the decoder and could be subtracted off. This means that Costa model has PPE. The proof for Costa capacity (2.13) involves two parts, the achievability proof and the converse. These basic steps can be summarized as follows.

Proof 2 *As special case of Gelfand-Pinsker setting, the achievability part can be proved in a similar way. First generate $2^{n(I(U;Y)-\epsilon)}$ i.i.d sequences U, according to the uniform distribution over the set of typical U. Next, distribute these sequences uniformly over 2^{nR} bins. For each sequence U, let $\iota(U)$ be the index of the bin containing U. For encoding, given the state vector $\mathbf{s}$ and the message W, look in the bin W for a sequence U such that (U, S) is jointly typical. Declare an error if no such U can be found. If the number of sequences in bin W is larger than $2^{n(I(U;Y)+\delta)}$, the probability of finding no such U decreases to zero exponentially as n increases. Next, choose X such that (X, U, S) is jointly typical and send it through the channel. At the decoder, look for the unique sequence U such that (U, Y) is jointly typical. Declare an error if more than one or no such sequence exist. Then set the estimate $\hat{W}$ of W equal to the index of the bin containing the obtained sequence U. If the transmission rate satisfies $R < I(U, Y) - I(U, S) - \epsilon - \delta$, the probability of error averaged over all codes decreases exponentially to zero as $n \longrightarrow +\infty$. This shows the existence of a code that achieves rate R with arbitrarily small probability of error.*
The converse is shown using a simple argument. From (2.10), we see that the optimal codebook for maximizing the mutual interference difference $I(U, Y) - I(U; S)$ is a function of both the transmitted codeword X and the state S. In [Cos83], Costa considered a codebook U of the form

$$U = X + \alpha S, \tag{2.14}$$

where α is a parameter to be determined in the sequel. Using straightforward calculation, the information that the received signal $Y = X + S + V$ conveys about the codebook U, at the receiver, can be written as

$$\begin{aligned} I(U;Y) &= H(Y) - H(Y|U) \\ &= H(X+S+V) + H(X+\alpha S) - H(X+S+V; X+\alpha S) \\ &= \frac{1}{2}\log\left(\frac{(P+Q+N)(P+\alpha^2 Q)}{PQ(1-\alpha)^2 + N(P+\alpha^2 Q)}\right). \end{aligned} \tag{2.15}$$

Similarly, the information that the state information S conveys about the codebook U, at the transmitter, writes

$$I(U;S) = \frac{1}{2}\log\left(\frac{P+\alpha^2 Q}{P}\right). \tag{2.16}$$

Combining (2.15) and (2.16), we get the transmission rate $R(\alpha) = I(U;Y) - I(U;S)$ *as*

$$R(\alpha) = \frac{1}{2}\log\left(\frac{(P+Q+N)}{PQ(1-\alpha)^2 + N(P+\alpha^2 Q)}\right). \tag{2.17}$$

Maximizing over $p(u,x|s)$ *in (2.10) reduces, in this case, to a maximization over the parameter* α *and gives*

$$\begin{aligned} C_{10} &\triangleq \max_{\alpha} R(\alpha) \\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right), \end{aligned} \tag{2.18}$$

attained with the parameter α *set to its optimal value (Costa parameter)*

$$\alpha = \frac{P}{P+N}. \tag{2.19}$$

Now, since the capacity of the channel cannot exceed $C_{11} \triangleq \max_{p(x|s)} I(X;Y|S) = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$*, for this is the capacity (in the Gaussian case) when both the encoder and the decoder know the sequence* S*, the optimality of Costa scheme is established.*

Costa result is more commonly known as "Dirty Paper Coding" (DPC) or, equivalently, as "Writing on Dirty Paper" (WDP). The design of the codebook together with the input distribution $p(u,x|s)$ according to Costa's DPC can be summarized in the form

$$X \sim \mathcal{N}(0,P) \;\text{ independent of }\; S, \tag{2.20a}$$

$$U = X + \alpha S \;\text{ with }\; \alpha = P/(P+N). \tag{2.20b}$$

The denomination WDP refers to a famous analogy between transmitting information over a channel with part of channel interference non-causally known to the encoder, and writing on a sheet of paper, with dirty spots on it.

2.2.2.1 Writing on Dirty Paper

Consider a sheet of paper covered with independent dirt spots of normally distributed intensity. We wish to write a message, directed to some reader, on this sheet of paper. This seemingly insignificant problem has tight relationship with the problem of coding with SI available at the encoder. Costa noticed this: "in some sense, the probability of writing a message on this sheet of paper is analogous to that of sending information through the channel of Fig.2.3". Similarly to the transmitter who has full knowledge of channel SI, the writer knows the location and intensity of the dirt spots. Also, similarly to the receiver who has to recover the transmitted message without having access to the channel SI, the reader can not distinguish the ink marks applied by the writer from the dirty spots. Paralleling the discussion stated above and denoting by "b" and "c" the message to be written and the dirty sheet of paper, respectively, two different scenarios are possible.

1. A non-informed writer writes the message "b" on top of the dirt spots "c", thus causing the corresponding reader to not distinguish the message from the dirty spots. This is a blind or non-informed approach. The writer (wearing bad glasses, probably) does not see non-dirty parts of the sheet of paper where the message should be written, as one could expect intuitively.

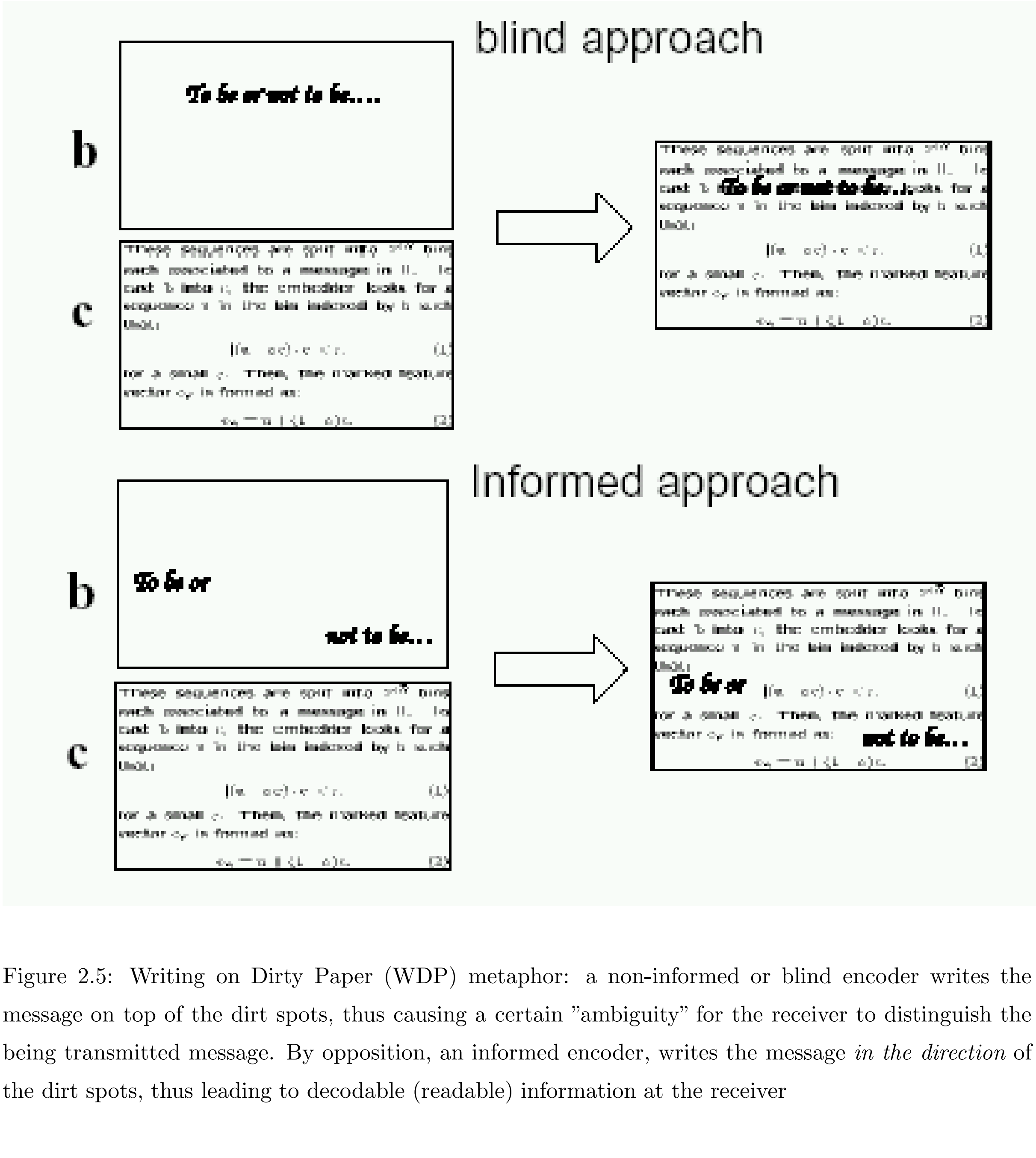


Figure 2.5: Writing on Dirty Paper (WDP) metaphor: a non-informed or blind encoder writes the message on top of the dirt spots, thus causing a certain ”ambiguity” for the receiver to distinguish the being transmitted message. By opposition, an informed encoder, writes the message *in the direction* of the dirt spots, thus leading to decodable (readable) information at the receiver

2. An informed writer (wearing good glasses, this time) first observes the non-dirty parts of the dirty paper (i.e., blanks in "c") and then writes down on these "clean" parts the message "b". The reader, knowing the exact location of the dirt spots in the sheet of paper, looks for the message "b" in those "clean" parts. This is the informed approach.

2.2.2.2 Geometrical interpretation

There is an interesting interpretation of random binning as *sphere covering* at the encoder and *sphere packing* at the decoder. In this section, we reexamine the code generation and the encoding-decoding procedures given above to provide a brief geometrical interpretation of Costa's DPC. A detailed interpretation is given in [SEG00a]. This geometrical description is made possible by representing random vectors by points in $\mathbb{R}^n$, where orthogonality between Gaussian random vectors stand for their independence. A random vector of length n is represented by a point in the n-dimensional space $\mathbb{R}^n$. Hence, the host signal $\mathbf{S}$ of power Q and the embedded signal $\mathbf{X}$ of power P lie on the surface of the spheres $\mathcal{S}_S$ and $\mathcal{S}_X$ centered at the origin and of radii $\sqrt{nQ}$ and $\sqrt{nP}$, respectively. The codebook $\mathcal{U}$, formed by codewords $\mathbf{U} = \mathbf{X} + \alpha\mathbf{S}$ of power $P + \alpha^2 Q$, where $\alpha = P/(P+N)$, can be represented by points on the surface of the n-dimensional sphere $\mathcal{S}_U$ centered at the origin and of radius $\sqrt{n(P+\alpha^2 Q)}$. In addition, the codebook $\mathcal{U}$ contains $2^{(n(I(U;Y)-\epsilon))}$ code-vectors $\mathbf{U}$, each drawn according to $\mathbf{U} \sim \mathcal{N}(0, (P+\alpha^2 Q)I)$. These code-vectors are randomly and equiprobably assigned to $2^{n(C-2\epsilon)}$ distinct bins, denoted by $\mathcal{U}_W$, where W is the bin index. Each bin $\mathcal{U}_W$ contains $2^{n(I(U;S)+\epsilon)}$ code-vectors. By means of the geometrical *sphere representation* introduced above, the encoder and the decoder can be viewed as performing sphere covering and sphere packing, respectively.

***(i)* Sphere covering at the encoder** : Given an index $W = W_0$ and a state vector $\mathbf{S}_0$, the encoder looks in the bin identified by W_0 (i.e., $\mathcal{U}_{W_0}$) for a code-vector $\mathbf{U}_0$ such that the pair $(\mathbf{U}_0, \mathbf{S}_0)$ is jointly typical. This is equivalent to searching for the code-vector $\mathbf{U}_0$ that satisfies

$$\mathbf{U}_0 = \operatorname*{argmin}_{\mathbf{U} \in \mathcal{U}_{W_0}} \|\mathbf{U} - \alpha\mathbf{S}_0\|. \tag{2.21}$$

Next, the encoder transmits $\mathbf{X}_0 = \mathbf{U}_0 - \alpha\mathbf{S}_0$ over the channel. Since $\alpha\mathbf{S}_0$ lies within distance $\sqrt{nP}$ of $\mathbf{U}_0$, the encoder chooses the correct code-vector $\mathbf{U}_0$ if the state $\alpha\mathbf{S}_0$ lies in the *bin-encoding sphere* centered at $\mathbf{U}_0$ and of radius $\sqrt{nP}$. Moreover, the composite signal $\mathbf{X}_0 + \mathbf{S}_0 = \mathbf{U}_0 + (1-\alpha)\mathbf{S}_0$ can be put in the form $\mathbf{X}_0 + \mathbf{S}_0 = \beta\mathbf{U}_0 + \mathbf{Z}$, where $\mathbf{Z}$ is orthogonal to $\mathbf{U}$ and $\mathbb{E}[\mathbf{Z}^2] = (1-\alpha)^2 \frac{PQ}{P+\alpha^2 Q}$. Hence, by transmitting $\mathbf{X}_0 = \mathbf{U}_0 - \alpha\mathbf{S}_0$, the encoder steers the state $\mathbf{S}_0$ towards $\beta\mathbf{U}_0$. This explains why we mentioned above that the optimal encoder should not erase the state, but designs codewords *in the direction* of the state. If we think of the encoder as being quantization-based (this will be justified in Section 2.4), this is equivalent to steer the state $\mathbf{S}_0$ toward its quantizer representative $\mathcal{Q}(\mathbf{S}_0)$ by transmitting the quantization error $\mathbf{X}_0 = \mathcal{Q}(\mathbf{S}_0) - \mathbf{S}_0$. Now, since $\alpha\mathbf{S}_0$ lies near the surface of a sphere of radius $\sqrt{n\alpha^2 Q}$ and since $\mathbf{U}_0$ lies near the surface of a sphere of radius $\sqrt{n(P+\alpha^2 Q)}$, fulfilling the power constraint can be viewed as covering the hull between spheres of radii $\sqrt{n(P+\alpha^2 Q)}$ and $\sqrt{n\alpha^2 Q}$ with bin-encoding spheres of radius $\sqrt{nP}$. The required number of bin-encoding spheres is lower bounded by the ratio of the hull volume to the volume of the bin-encoding sphere, i.e.,

$$\frac{A_n(n(P+\alpha^2 Q))^{n/2}}{A_n(nP)^{n/2}} = 2^{n(I(U;S)+\epsilon)}, \tag{2.22}$$

where A_n is a constant that depends on n [CT91].

(ii) **Sphere packing at the decoder:** Given a received sequence $\mathbf{Y} = \mathbf{Y}_0$, search in the entire codebook $\mathcal{U} = \bigcup_W \mathcal{U}_W$ for the (unique) code-vector $\hat{\mathbf{U}}$ such that the pair $(\hat{\mathbf{U}}, \mathbf{Y}_0)$ is jointly typical. This is equivalent to search for the (unique) code-vector $\hat{\mathbf{U}}$ that satisfies

$$\hat{\mathbf{U}} = \underset{\mathbf{U}\in\mathcal{U}}{\operatorname{argmin}} \quad \|\mathbf{Y}_0 - \beta\mathbf{U}\|, \tag{2.23}$$

where

$$\beta = \frac{\mathbb{E}[\mathbf{U}^T\mathbf{Y}_0]}{\mathbb{E}[\mathbf{U}^T\mathbf{U}]} = \frac{P+\alpha Q}{P+\alpha^2 Q}.$$

Next, return the estimated index $\hat{W}$ as the index of the bin $\mathcal{U}_{\hat{W}}$ containing $\hat{\mathbf{U}}$. Since $\mathbf{Y}_0 = \mathbf{X}_0 + \mathbf{S}_0 + \mathbf{V}$ can be put in the form $\mathbf{Y}_0 = \beta\mathbf{U}_0 + \mathbf{Z} + \mathbf{V}$, the probability of error in decoding the appropriate index depends mainly on the radius of the *decoding sphere* centered at $\beta\mathbf{U}_0$ and of radius $\sqrt{n(\mathbb{E}[\mathbf{Z}^2]+N)}$. Since the received sequence $\mathbf{Y}_0$ lies near the surface of a sphere of radius $\sqrt{n(P+Q+N)}$, the number of *distinguishable* (reliably decodable) code-vectors is upper bounded by the number of decoding spheres that can be packed into a sphere of radius $\sqrt{n(P+Q+N)}$, i.e.,

$$\frac{A_n(n(P+Q+N))^{n/2}}{A_n(n(\mathbb{E}[\mathbf{Z}^2]+N))^{n/2}} = 2^{n(I(U;Y)-\epsilon)}. \tag{2.24}$$

Combining (2.22) and (2.24), we get the number of different messages (indexes) that can be reliably communicated as

$$\frac{2^{n(I(U;Y)-\epsilon)}}{2^{n(I(U;S)+\epsilon)}} = 2^{n(C-2\epsilon)}. \tag{2.25}$$

This is because all $2^{n(I(U;S)+\epsilon)}$ code-vectors gathered in the same bin $\mathcal{U}_W$convey the same index W.

2.2.3 Extensions

As mentioned before, the initial Costa's DPC scheme is derived under the assumption that both the non-causally known noise S and the unknown noise V are independent and white Gaussian. Also, the encoder is assumed to be power-constrained. This result has since been extended to different distributions on the two noise sources. For instance, it has been shown in [YSJ+01] that the known noise S does not affect the capacity as long as both noise sources are Gaussian, but not necessarily identically distributed. The resulting coding scheme is named "Writing on Colored Paper", by reference to the possibility for the two noise sources to be non white. Also, a *sufficient* condition for the so called Public Private Equivalence (PPE) property to hold is provided in [CL02b]. The resulting coding scheme is equivalently referred to as either "Generalized Writing on Dirty Paper" (GWDP) or "Writing on non Gaussian paper", for it allows the state information to have any distribution, not necessarily deterministic as stated in [ESZ00]. These two extensions are briefly discussed, below.

2.2.3.1 Colored Gaussian Channel with Side Information

In this section, we consider the channel model shown in Fig.2.2. We look at a single block of n transmissions

$$Y^n = X^n + S^n + V^n, \tag{2.26}$$

where S^n and V^n are independent Gaussian sequences (not necessarily identically distributed) with arbitrary finite-dimensional covariance matrices K_{ss} and K_{vv} respectively. The sequence S^n is entirely known non-causally to the transmitter, but not to the receiver. The encoder maps a codeword index $W \in \{1, \ldots, 2^{nR}\}$ and a side information S^n to a block of n transmissions. The decoder maps the channel output to a codeword index. The capacity of such a channel has been provided in [YSJ+01].

Theorem 3 (Wei Yu et al. [YSJ+01]) *Consider a block of n transmissions in a Gaussian channel $Y^n = X^n + S^n + V^n$, where S^n and V^n are independent Gaussian sequences, with S^n known non-causally to the transmitter. Suppose $|K_{ss}| > 0$. The capacity of the channel under a power constraint P is*

$$C_n = \max_{K_{xx}} \frac{1}{2n} \log \frac{|K_{xx} + K_{vv}|}{|K_{vv}|}, \tag{2.27}$$

provided that the maximization is over covariance matrices K_{xx} such that $\frac{1}{n}Tr(K_{xx}) \leq P$, and the maximizing K_{xx} is such that $|K_{xx}| > 0$ ($|A|$ denotes the determinant of matrix A).

The proof of the achievability can be found in [YSJ+01, Yu02] where it is shown that all rates of the form

$$R_n = \frac{1}{n} \left(I(U^n; Y^n) - I(U^n; S^n) \right) - \epsilon, \tag{2.28}$$

are achievable for all joint Gaussian distributions $p(u^n|x^n, s^n)p(x^n)p(s^n)$. The converse closely follows that of Costa's original DPC. Consider a codebook U^n in the form $U^n = X^n + FS^n$, where F is an $n \times n$ matrix. The mutual information $I(U^n; Y^n)$ that the received sequence Y^n conveys about the codebook U^n is given by a straightforward generalization of (2.15), as

$$\begin{aligned} I(U^n; Y^n) &= H(U^n) + H(Y^n) - H(U^n; Y^n) \\ &= \frac{|K_{xx} + FK_{ss}F^T|.|K_{xx} + K_{ss} + K_{vv}|}{|\mathrm{Cov}(Y, U)|}. \end{aligned} \tag{2.29}$$

$\mathrm{Cov}(Y, U)$ is the covariance matrix of $(Y; U)$, given by

$$\mathrm{Cov}(Y, U) = \begin{pmatrix} K_{xx} + FK_{ss}F^T & K_{xx} + FK_{ss} \\ K_{xx} + K_{ss}F^T & K_{xx} + K_{ss} + K_{vv} \end{pmatrix}.$$

Similarly, the mutual information $I(U^n; S^n)$ that the known sequence S^n conveys about the codebook U^n at the transmitter generalizes that given by (2.16), as

$$I(U^n; S^n) = \frac{1}{2} \log \frac{|K_{xx} + FK_{ss}F^T|}{|K_{xx}|}. \tag{2.30}$$

The transmission rate can be obtained by combining (2.29) and (2.30) to get an explicit expression for

$$R_n(F) = \frac{1}{n} \left(I(U^n; Y^n) - I(U^n; S^n) \right). \tag{2.31}$$

The optimal matrix F that maximizes $R_n(F)$ has a similar expression to that of Costa's parameter $\alpha = P/(P+N)$ (which is optimal in the scalar case) and writes

$$F = K_{xx} \left(K_{xx} + K_{vv}\right)^{-1}, \tag{2.32}$$

independently on the covariance matrix K_{ss} of the known noise sequence S^n. Thus, the maximal transmission rate is given by

$$\begin{aligned}
\max_F R_n(F) &= \frac{1}{2n} \log \frac{|K_{xx}|}{|K_{xx} - K_{xx}(K_{xx}+K_{vv})^{-1}K_{xx}|}, && (2.33)\\
&= \frac{1}{2n} \log \frac{|K_{xx}+K_{vv}|}{|K_{vv}|}, && (2.34)\\
&= \frac{1}{n} I(X^n; Y^n|S^n). && (2.35)
\end{aligned}$$

Equation (2.34) results from the use of Shur's complement formula for matrices determinant, i.e.,

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D|.|A - BD^{-1}C| = |A|.|D - CA^{-1}B|.$$

Equation (2.35) holds because this is the mutual information formula for a vector Gaussian channel without interfering signal S^n [CT91]. If the channel $Y^n = X^n + S^n + V^n$ is a memoryless channel, i.e.,

$$p(y^n|x^n, s^n) = \prod_{k=1}^{n} p(y_k|x_k, s_k), \tag{2.36}$$

where each use of the channel involves a vector input and a vector output and coding is done over many uses of the channel, channel capacity is given by (2.35). The reason is that a memoryless vector channel can be transformed into n parallel sub-channels through a diagonalization of the noise covariance, as pointed out in [YSJ$^+$01]. If the channel works on a single block of n transmissions, an additional maximization of (2.35) over $p(x^n)$ is needed. The reason is that $I(X^n; Y^n|S^n)$ depends on the $p(x^n)$ (and not on $p(u^n|x^n, s^n)$). Maximization over $p(x^n)$ amounts to that over K_{xx} and gives

$$\max_{p(x^n)} \frac{1}{n} R_n(F) = \max_{K_{xx}} \frac{1}{2n} \log \frac{|K_{xx}+K_{vv}|}{|K_{vv}|}. \tag{2.37}$$

Finally, since the capacity of the Gaussian vector channel without interference is given by (2.37) and since the capacity of the channel with interference cannot exceed that of the channel without interference, (2.37) is indeed the required capacity of the vector channel. Hence, the assumption that (U^n, S^n) takes the form $U^n = X^n + FS^n$ is without loss of generality. Further, just as in the i.i.d case, neither optimal F nor capacity depend on the distribution of the non-causal state information S^n. Curiously enough, the optimal F takes the form of the optimal non-causal Wiener filter for estimating X^n from the noisy observation $X^n + V^n$.

2.2.3.2 Writing on non-Gaussian paper

We now consider a more generalized extension of Costa initial WDP to the situation where only the unknown noise is Gaussian (the known noise S can have *any* distribution and the unknown noise is not necessarily i.i.d). Coding for such channel, sometimes referred to as "Generalized Writing on Dirty

Paper" (GWDP) or equivalently as "Writing on non Gaussian paper", has been reported in [CL02b]. It is shown that there is no loss in capacity in having non-Gaussian state S and non i.i.d. channel noise V. The proof relies on the following two assertions.

$$X - \alpha(X+V) \quad \text{and} \quad X+V \quad \text{are independent.} \tag{2.38a}$$

$$X - \alpha(X+V) \quad \text{and} \quad S \quad \text{are independent.} \tag{2.38b}$$

Assertion (2.38a) follows since $X - \alpha(X+V)$ and $X+V$ are jointly Gaussian and uncorrelated. The uncorrelation between $X - \alpha(X+V)$ and $X+V$ will always hold as long as the parameter α is set to its optimal value, i.e., $\alpha = P/(P+N)$, and can be simply seen from

$$\mathbb{E}[(X - \alpha(X+V))(X+V)] = P - \alpha(P+N).$$

Assertion (2.38b) follows since S is independent of both X and V. Now, using (2.38), the following equalities follow.

$$\begin{aligned}
H(U|S) &= H(X+\alpha S|S) = H(X|S) = H(X). \\
H(U|Y) &= H(X+\alpha S|Y), \\
&= H(X+\alpha S - \alpha Y|Y), \\
&= H(X - \alpha(X+V)|Y), \\
&= H(X - \alpha(X+V)|Y), \\
&= H(X - \alpha(X+V)), && (2.39) \\
&= H(X - \alpha(X+V)|(X+V)), \\
&= H(X|X+V). && (2.40)
\end{aligned}$$

Eqs. (2.39) and (2.40) hold because of (2.38). Maximizing the transmission rate $R = I(U;Y) - (U;S) = H(U|S) - H(U|Y)$ over $p(x)$, capacity writes

$$\begin{aligned}
C &= \max_{p(x)} H(U|S) - H(U|Y), \\
&= \max_{p(x)} [H(X) - H(X|X+V)], \\
&= \max_{p(x)} I(X;X+V), \\
&= \frac{1}{2}\log(1+\frac{P}{N}).
\end{aligned}$$

Hence, the AWGN capacity can be (theoretically) attained even with non Gaussian state, a situation which is current in practice.

2.3 Binning Coding v.s. Algebraic Coding

In section 2.2, we provided a quick view of the general setup of coding with SI non-causally known to the transmitter, which represents the theoretical foundation for the problem of information embedding. However, performance limits therein are shown to be achievable by use of random codes. These are probabilistic in nature, and thus computationally prohibitive to be implemented in practice. Hence, simplifications are required to make this approach feasible in real life. Recently, adhering to Costa setting, Chen and Wornell [CW01] and Eggers and al. [EBTG03] designed practical quantization-based schemes to achieve the side-information capacity for watermarking applications. A similar work on quantized projections appeared earlier in 19978 in [SZTB98]. The codebook entries are chosen to be quantizers representatives. These two sample-wise schemes are referred to as "Quantization Index Modulation" (QIM) and "Scalar Costa Scheme" (SCS), respectively. By opposition to random codes, these quantization-based codes can be viewed as being algebraic. By algebraic, it is meant that there is some structure in the codebook entries (broadly, a group structure). Such a structure not only simplifies the sharing (between the encoder and the decoder) of a huge number of codewords, but also simplifies the search and storing procedures of the codebook entries at both the encoder and the decoder. For instance, the encoder (and/or the decoder) does not need store all the codebook entries. Only a subset of these, together with some codebook parameters, are sufficient to generate the remaining codewords. Also, the search can be made easier. However, there is a cost to pay for such a simplification: algebraic codes are not optimal, in that they do not achieve the ultimate performance provided by random codes. Ignoring the relative computational complexity, the efficiency of an algebraic code can be measured by the extent by which it approaches these ultimate performance or, equivalently, by the gap between the two.

2.4 Sub-optimal Algebraic-based Coding Techniques.

Broadly, information embedding schemes can be divided into two main classes: (i) host-interference non-rejecting methods and (ii) host-interference rejecting methods. Host interference non-rejecting methods do not allow the encoder to exploit the knowledge of the host signal in the design of the transmitted codewords and are consequently interference limited by construction. The simplest methods consist in adding a pseudo-noise sequence to the host signal and are often referred to as Spread-Spectrum Modulations (SSM). When the knowledge of the host signal at the encoder is adequately exploited in system design, the resulting information embedding system can be made host interference free.

2.4.1 Spread-Spectrum Modulations (SSM)

So far, we have argued that information embedding can be viewed as communication over a very noisy channel. Motivated by the observation that digital communication systems for transmission over very noisy channels, possibly subject to intentional disturbs (such as jamming or interferences) are almost usually build upon Spread Spectrum (SS) technology, early approaches for information embedding were based on spread spectrum. However, while the term "spread spectrum" (SS) refers to expanding the bandwidth of the transmitted signal with comparison to that of the source [Pro01] in conventional communication, it has a slightly different translation in information embedding. It refers to spreading the message to be transmitted over many samples of the original cover signal, using some pseudo-random spreading sequence. A simplified diagram of the simplest (additive) SS-based information embedding is depicted in Fig.2.6.

Here, we consider bipolar transmission of binary messages, i.e., $\mathcal{M} = \{-1, 1\}$. Extension to non-binary

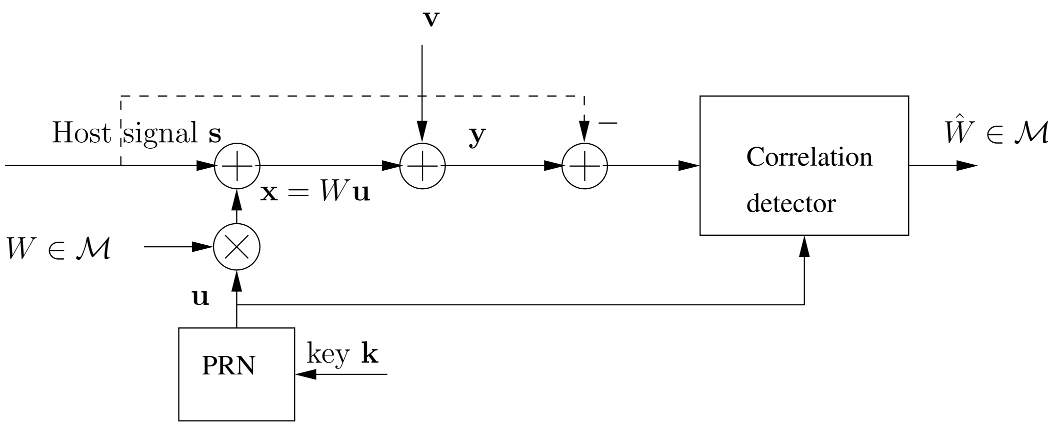


Figure 2.6: Blind (solid line) and non-blind (dashed line) additive spread-spectrum-based information embedding.

alphabets is straightforward. Taking the index W as input, the encoder forms the codeword $\mathbf{x}$ of length n as $\mathbf{x} = W\mathbf{u}$ and transmits it over the channel. The sequence $\mathbf{u}$ is produced by a Pseudo Random Number generator (PRN) using a secret key $\mathbf{k} \in \mathbf{K}$. In principle, the sequence $\mathbf{u}$ can be drawn according to any given probability mass function. Upon reception, the decoder computes the cross-correlation between the received signal $\mathbf{y} = \mathbf{x} + \mathbf{s} + \mathbf{v}$ and the sequence $\mathbf{u}$. In this case, this is given by the ratio of their normalized inner product

$$< \mathbf{y}, \mathbf{u} > \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i u_i$$

to the normalized Euclidean norm $\|\mathbf{u}\|$ of the sequence $\mathbf{u}$, i.e.,

$$\begin{aligned} y &\triangleq \frac{< \mathbf{y}, \mathbf{u} >}{\|\mathbf{u}\|}, \\ &= W + \frac{< \mathbf{s}, \mathbf{u} >}{\|\mathbf{u}\|} + \frac{< \mathbf{v}, \mathbf{u} >}{\|\mathbf{u}\|}, \\ &= W + s + v. \end{aligned} \tag{2.41}$$

The two quantities s and v denote the (normalized) projection of the host $\mathbf{s}$ and the noise $\mathbf{v}$ upon the sequence $\mathbf{u}$. Additive spread spectrum relies on the assumption that the pseudo-random sequence $\mathbf{u}$ is uncorrelated with both the host signal $\mathbf{s}$ and the unknown channel noise $\mathbf{v}$, i.e., $s = v = 0$. Hence, the receiver can recover the transmitted message W by simply computing the correlation (2.41). However, the accuracy of the measure (2.41) depends on the length n of the involved signals. Very large n lead to more accurate correlation evaluation. For finite-length signals, a hypothesis test is needed. The larger n, the more precise this hypothesis test. Very long sequences are not preferred in practice however, because the interfering signals $\mathbf{s}$ and $\mathbf{v}$ are strengthened also, thereby.

The maximum rate of spread-spectrum information embedding can be easily determined in the case of an i.i.d Gaussian host signal $\mathbf{s} \sim \mathcal{N}(0, Q)$ and an i.i.d Gaussian channel noise $\mathbf{v} \sim N(0, N)$. In this case, the channel in Fig.2.6 is equivalent to an AWGN channel having the same SNR, i.e., P/N [dB] in non-blind communication and $P/(N + Q)$ [dB] in blind communication. In non-blind reception, the decoder subtracts the cover signal $\mathbf{s}$ from the received signal $\mathbf{y}$ prior to decoding, thus making the scheme interference-free. Blind SS, as for it, has poor performance because of strong host interference. Blind and non-blind SS-based channel capacities are given by

$$C_{\text{Blind SS}} = \frac{1}{2}\log\left(1+\frac{P}{N+Q}\right), \tag{2.42a}$$

$$C_{\text{Non-blind SS}} = \frac{1}{2}\log\left(1+\frac{P}{N}\right). \tag{2.42b}$$

Note that (2.42a) and (2.42b) can be only achieved with ideal coding and signal shaping. Also, due to information embedding requirements and especially the transparency requirement, we have $Q \gg P$ and $Q \gg N$ in common information embedding scenarios. Thus blind SS suffers significantly from original signal interference and its efficiency is mainly determined by the Document-to-Watermark Ratio $\text{DWR} = 10\log(Q/P)$ [dB]. Capacities curves of blind and non-blind SS, shown in Fig.2.7, are depicted for DWR = 20 dB. It can be seen that the non-knowledge of the host signal at the receiver in blind

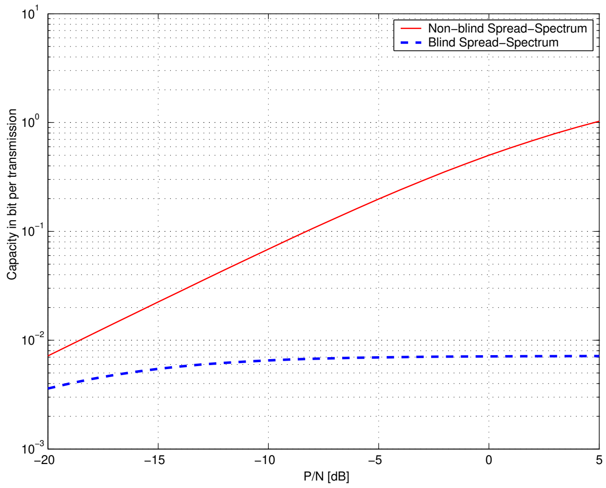


Figure 2.7: Channel capacity, in bit per transmission, for both non-blind and blind Spread Spectrum v.s $\text{SNR} = P/N$ [dB]. The depicted curves are for DWR = 20 dB. Blind SS suffers great performance loss due the host interference.

SS significantly reduces its achievable rates. A more general approach to spread the information to be transmitted over many elements of the cover signal is called Spread-Transform (ST).

2.4.1.1 Spread-Transform Information Embedding

ST-information embedding has been proposed by Swanson *et al.* in [SZT96], and later by Chen and Wornell in [CW99]. A detailed description of this approach can be found in [CW99, Egg01]. Here, only the principle is examined. In ST information embedding, the message is not embedded into the original signal $\mathbf{s}$, but onto the projection $\bar{\mathbf{s}}$ of $\mathbf{s}$ onto a properly chosen random sequence $\mathbf{t}$. Denoting the spreading factor by τ, the number of original data elements belonging to one element in the transform domain ST, each τ consecutive elements of the host signal $\mathbf{s}$ are transformed into one element of $\bar{\mathbf{s}}$ according to

$$\bar{s}_k = \sum_{i=\tau k}^{\tau k+\tau-1} s_i t_i, \quad k = 1, 2, \ldots \tag{2.43}$$

Similarly, the same transformation applied to the embedded signal $\mathbf{x}$ yields a shorter signal $\bar{\mathbf{x}}$. Due to information embedding requirements stated above, especially that related to the transparency requirement, the inverse spread transform operation must be applied to the composite signal in the transform domain $\bar{\mathbf{s}} + \bar{\mathbf{x}}$. This gives a composite signal $\mathbf{c} = \mathbf{s} + \mathbf{x}$ such that

$$x_n = \bar{x}_k t_n, \quad k = \lceil n/\tau \rceil, \tag{2.44}$$

where "$\lceil . \rceil$" denotes rounding to the next larger integer value. At the receiver, decoding must be performed in the same transform domain. Hence, the received sequence $\mathbf{y}$ has to be projected onto $\mathbf{t}$, too. This provides

$$\bar{y}_k = \sum_{i=\tau k}^{\tau k+\tau-1} y_i t_i, \quad k = 1, 2, ... \tag{2.45}$$

Note that the denomination "spread transform" stems from the fact that the information to be embedded into $\bar{\mathbf{s}}$ is spread into τ original elements by the inverse transform. A part from security issues, the major advantage of ST, is that of canceling any component of the unknown channel noise that is orthogonal to the spreading direction $\mathbf{t}$. The latter observation is the key idea behind the enhancement in SNR observed in the transform domain,

$$\mathrm{SNR}_\tau = \mathrm{SNR} + 10 \log_{10} \tau \quad [\mathrm{dB}], \tag{2.46}$$

where SNR_τ denotes the SNR in the transform domain.

2.4.2 Side Information Quantization and QIM

We begin this section by showing, with qualitative arguments, why quantization is a basic operation, specifically in information embedding and more generally, in all coding with state information systems. We then give a brief description of the basic principle of Quantization Index Modulation [CW01] followed by two typical applications: first as a class of powerful embedding functions in the context of information embedding and then, as an alternative understanding of the famous Tomlinson Harashima Precoding (THP) [Tom71, MH69, HM72], in the context of inter-symbol interference (ISI) mitigation in classical communication. Other forms of indexed-quantization-based schemes are briefly reviewed, with particular emphasis on the famous Scalar Costa Scheme (SCS) [EBTG03] described in Section 2.4.4.

2.4.2.1 Quantization-based coding for SI systems

Quantization satisfies the requirements needed by the optimal encoder and the optimal decoder stated in Section 2.2.2, as it will be argued below. Of course, this is not a rigorous proof of why it is precisely quantization that should be considered for practical implementation of the optimal coding. However, this (at least) justifies the high interest quantization is gaining each time coding for SI systems is of concern. The suitability of quantization for such systems can be explained as follows.

1. We mentioned in Section 2.2.2 that, for encoding, the optimal encoder *steers* the host signal $\mathbf{s}$ toward the nearest code-vector $\mathbf{u}$ in the bin $\mathcal{U}_W$ identified by the index W to be transmitted. This is precisely what a quantizer does. Given some quantizer $\mathcal{Q}(.)$, each vector $\mathbf{r} \in \mathbb{R}^n$ to be quantized is *steered* toward its reconstruction point (vector) $\hat{\mathbf{r}} = \mathcal{Q}(\mathbf{r})$ by adding to it its quantization error $\mathcal{Q}(\mathbf{r}) - \mathbf{r}$. Hence, quantization has already, in it, the fundamental concept of *steering* codewords

in a given direction. To have the arrangement in bins stated required by the optimal coding, one need simply choose a set of quantizers, and index them by the set of indexes to be transmitted (this fixes the number of quantizers to be exactly $|\mathcal{M}|$).

2. Paralleling the optimal encoding stated in Section 2.2.2, each reconstruction point of each quantizer (bin) is the center of a *bin-quantization cell* $\mathcal{V}(\mathcal{Q})$ which can be viewed as a *bin-encoding sphere*. Similarly, paralleling the optimal decoding stated in Section 2.2.2, each received sequence $\mathbf{y}$ is quantized to the center of the nearest *decoding quantization cell*, which can be viewed as a *decoding sphere*.

3. In addition, fulfilling the power constraint in the optimal encoding has a *sphere covering* interpretation, as mentioned before. Quantizing at the encoder has a similar *sphere covering* interpretation. When quantizing a signal $\mathbf{r}$ with power Q under the (distortion) constraint that the quantization error $\mathcal{Q}(\mathbf{r}) - \mathbf{r}$ has variance P, the (volume of the) quantization cell has to be properly designed. Satisfying the distortion constraint P can be viewed as covering the hull between the positions of $\mathbf{r}$ and those of its representative $\hat{\mathbf{r}} = \mathcal{Q}(\mathbf{r})$ with bin-quantization cells $\mathcal{V}(\mathcal{Q})$. A similar interpretation applies for the decoder. Further, quantization cells for the same quantizer do not intersect and those for different quantizers may intersect, exactly as for the optimal probabilistic coding stated above.

The beauty of quantization in coding for information embedding systems lies on the above features. An additional argument that suggests quantization as an ideal suitable tool in coding for embedding information stems from the embedding-specific requirements stated above. For instance, the transparency requirement, which expresses some kind of closeness of the composite signal to the original, may be ideally fulfilled by properly designing the quantizers. Under certain assumptions[1], the quantizer representative $\hat{\mathbf{r}} = \mathcal{Q}(\mathbf{r})$ of a signal $\mathbf{r}$ is close, in the sense of the Euclidean distance, to this signal itself. Note that Chen and Wornell differently (but qualitatively, too) argue in [CW01] why quantization is well suited for information embedding systems. The resulting scheme is named as "Quantization Index Modulation" (QIM).

2.4.2.2 Indexed quantization for IE: principle and optimality

Quantization Index Modulation (QIM) refers to embedding information by first modulating an index or sequence of indexes with the embedded information and then quantizing the host signal with the associated quantizer or sequence of quantizers. The optimality of QIM for embedding information is assessed in two steps, first through an example (so as to illustrate the basic principle) and then, through a more rigorous development.

***(i)* QIM through an example** Consider the case where we wish to embed one bit of information per host sample. So, the index W is in $\{1, 2\}$, meaning that we need two quantizers. Their corresponding sets of reconstruction points in $\mathbb{R}^n$ are indicated in Fig.2.8 by **O** for the first quantizer and by ✗ for the second quantizer. The denomination QIM stems from *modulating* the quantization by the index to be

[1]This is the case when, for example, the high resolution quantization assumption is satisfied.

transmitted. Namely, if $W = 1$, the host signal $\mathbf{s}$ is quantized with the **O**-quantizer to the nearest **O** point if $W = 1$ and with the ✗-quantizer to the nearest ✗ point if $W = 2$. Denoting by $\mathcal{Q}(\mathbf{s})$ the reconstruction point of $\mathbf{s}$, the embedded codeword $\mathbf{x}$ is set to the quantization error $\mathcal{Q}(\mathbf{s}) - \mathbf{s}$. Hence, the composite signal $\mathbf{c} = \mathbf{x} + \mathbf{s}$ is represented by an **O** point if $W = 1$ and by an ✗ point if $W = 2$. The basic principle

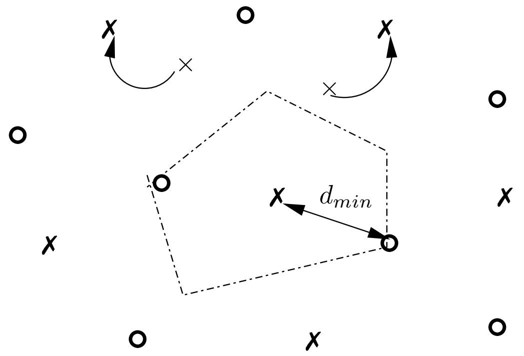


Figure 2.8: QIM information embedding. The points marked with **O**'s and ✗'s are the reconstruction points of two different quantizers: the **O**-quantizer, associated with index $W = 1$, and the ✗-quantizer, associated with the index $W = 2$. The minimum distance d_{min} determines the system immunity to channel noise.

of QIM relies on the following property: as the host signal $\mathbf{s}$ varies, the composite signal $\mathbf{c}$ varies from one **O** point (if the message to be embedded is $W = 1$) to another or from one **O** point (if the message to be embedded is $W = 2$) to another, but it never varies between an **O** point and an ✗ point. If we assume no perturbation in the channel (i.e., $\mathbf{v} = 0$), the latter property means that, for the same transmitted message, the receiver sees a received signal $\mathbf{y}$ which is always in the same set of reconstruction points, independently on the host signal. For instance, the receiver would see always an **O** point if $W = 1$ is transmitted, independently on the host signal $\mathbf{s}$. Similarly, the receiver would see an ✗ point if $W = 2$ is transmitted, independently on the host signal $\mathbf{s}$. That the index of the quantizer (1 for the **O**-quantizer and 2 for the ✗-quantizer) does not vary with the host signal allows to identify the transmitted message with no error, even in presence of infinite energy host signal. Hence, if the unknown channel noise $\mathbf{v}$ is zero, QIM would allow to completely reject the interference due the host, exactly as the more complex optimal coding does. If channel perturbations are not zero but are not too severe so as to move an **O** point to an ✗ point or vice versa, the decoder would still decode the correct message. This would be performed by rounding the received signal to the nearest reconstruction point in the set of reconstruction points formed by the union of **O** points and ✗ points. In fact, it is the minimum distance between the set of the reconstruction points of different quantizers that effectively determines the performance of the QIM embedding system. Considering that $M = |\mathcal{M}|$ indexes are to be transmitted, and thus M different quantizers denoted by $\{\mathcal{Q}_i(.)\}, i = 1, 2, \ldots, M$, this minimum distance is defined as

$$d_{min} \triangleq \min_{(i,j):i \neq j} \min_{(\mathbf{c}^{(i)}, \mathbf{c}^{(j)})} \|\mathbf{c}^{(i)} - \mathbf{c}^{(j)}\|, \tag{2.47}$$

where $\mathbf{s}^{(i)}$ is the composite signal obtained by use of the quantizer $\mathcal{Q}_i(.)$. For equiprobable indexes W in the set $\mathcal{M}$, the minimum-distance decoder makes decision according to the rule

$$\hat{W} = \underset{W \in \mathcal{M}}{\operatorname{argmin}} \quad \|\mathbf{y} - \mathcal{Q}_W(\mathbf{y})\|. \tag{2.48}$$

Intuitively, the minimum distance measures the strength of the perturbations that are tolerated by the system. For example, if the unknown channel noise $\mathbf{v}$ is Gaussian and has power (per-sample) N, then the minimum distance decoder will make the correct decision as long as $(d_{min}/2)^2 > nN$. Thus, at high SNR, the probability of error $P_e \triangleq \Pr(\hat{W} \neq W)$ can be approximated by

$$P_e \approx \Phi\left(\sqrt{\frac{d_{min}^2}{4N}}\right), \tag{2.49}$$

where $\Phi(u) = \displaystyle\int_u^{+\infty} \frac{1}{\sqrt{2\pi}} \exp -\frac{u^2}{2} du$ is the tail probability of the Gaussian PDF. In addition to the reliability of transmission, measured by the probability of error (2.49), QIM is characterized by the encoding distortion D_E induced to the host by the embedding process. Chen and Wornell [CW01] remarkably noticed that the initial QIM scheme, also sometimes referred to as *regular* QIM, can be improved so as to have better rate-distortion-robustness rates by appropriately scaling the quantizers. The resulting scheme is named "Distortion Compensation QIM" (DC-QIM).

(ii) **Optimality of DC-QIM** Consider a quantizer $\mathcal{Q}_i$, $i \in \mathcal{M}$. Scaling this quantizer by a factor $\alpha \in [0, 1]$ means that all its reconstruction points have to be scaled by $1/\alpha$. Likewise, two points separated by a distance d before scaling are separated by d/α after scaling. Thus, scaling increases the minimum distance (2.47) and thus, reduces the probability of error (2.49) by a factor of $1/\alpha$. However, scaling also introduces an additional distortion by increasing the distortion D_E by a factor $1/\alpha^2$. DC-QIM relies on the idea that compensating this additional distortion is possible by adding back a fraction $(1-\alpha)$ of the quantization error. This results in an embedding function in the form

$$\mathbf{x}(\mathbf{s}; W) = \mathcal{Q}_W(\alpha\mathbf{s}) - \alpha\mathbf{s}. \tag{2.50}$$

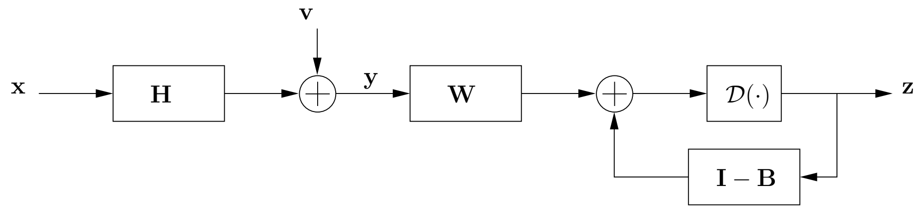
However, while removing the additional distortion, the quantization error added back represents a source of interference at the receiver. If the quantization error $\mathbf{x}$ should satisfy a power constraint of P, the power of the distortion-compensation term is $(1-\alpha)^2\frac{P}{\alpha}$. Hence, the signal-to-noise ratio at the receiver writes

$$\mathrm{SNR}(\alpha) = \frac{d_{min}^2/\alpha}{(1-\alpha)^2 P/\alpha^2 + N}. \tag{2.51}$$

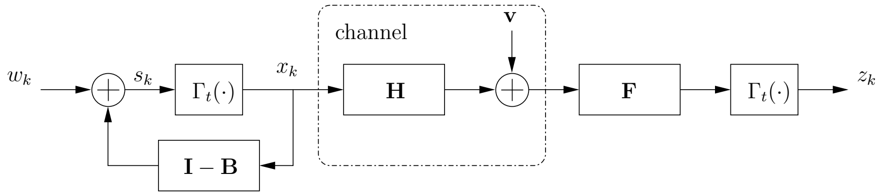
Since decreasing the parameter α, increases the minimum distance d_{min} but also strengthens this interference term, one optimality criterion for choosing α is to maximize (2.51). The solution for this simple optimization problem is given by

$$\alpha = \frac{P}{P+N}, \tag{2.52}$$

which is nothing but the initial optimal Costa's parameter.



(a) Channel model and MMSE-DFE



(b) THP precoding

Figure 2.9: Communication system using Tomlinson-Harashima Precoding (THP).

2.4.3 Indexed quantization for precoding for ISI channels

In this section, we go one step further in illustrating the use of (indexed) quantization in side information systems, by viewing the famous Tomlinson-Harashima Precoding (THP) [Tom71, HM72] scheme for inter-symbol interference (ISI) channels as a form of indexed quantization. Of course, there are other precoding schemes [EF92, LTF93, Lor93], but this section is concerned primarily with THP. The reason is that many of these are simply advanced forms of THP (Zero Forcing THP (ZF-THP), Minimum Mean-Square Error THP (MMSE-THP)). While this understanding of THP shows that the idea of indexed quantization is not new, this establishes a strong link between the very "mature" coding for ISI channels and the very new, but yet well developed, coding for information embedding. Alternatively, this also shows that the results provided in the rest of this work have potential use in classical communication.

2.4.3.1 Precoding for ISI channels

Inter-symbol interference (ISI) is a significant obstacle against reliable digital communication through band-limited channels. A classical situation where ISI occurs is that of communication over channels where different symbols directed to the same user interfere. Another more involved situation is that in which different symbols directed to different users interfere. In both situations, the role of *precoding* is to make the channel ISI-free. The channel model is shown in Fig.2.9(a) where $\mathbf{x}$ is the channel input and $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{v}$ is its output. The degradations that the transmitted signal may encounter in the channel are represented by the channel matrix $\mathbf{H}$ and the Gaussian noise $\mathbf{v}$. We want to output an optimal estimation $\mathbf{z}$ of $\mathbf{x}$. Obviously, this can be given by the MMSE estimation of $\mathbf{x}$ given the observation $\mathbf{y}$. The MMSE estimation is the one that minimizes the squared power of the error $\mathbf{e} = \mathbf{x} - \mathbf{z}$. The optimal MMSE filter $\mathbf{W}$ can be divided into two parts: (i) a *feed-forward* filter $\mathbf{K}_f = \mathbf{H}^T\mathbf{H}$ corresponding to the fordward channel $\mathbf{x} \longmapsto \mathbf{w} = \mathbf{H}^T\mathbf{y} = \mathbf{K}_f\mathbf{x} + \mathbf{v}'$ where $\mathbf{v}' = \mathbf{H}^T\mathbf{v}$ and, (ii) a *backward* filter $K_b = (\mathbf{H}^T\mathbf{H} + \mathbf{K}_{xx})^{-1}$ corresponding to the backward channel $\mathbf{w} \longmapsto \mathbf{x} = \mathbf{K}_b\mathbf{w} + \mathbf{e}$, where $\mathbf{K}_{xx}$ is the covariance matrix of the signal $\mathbf{x}$. In practice the backward channel involves a decision-based feedback equalizer. The block labeled $\mathcal{D}(\cdot)$ in Fig.2.9(a) represents a symbol decision device.

2.4.3.2 Decision-Feedback Equalizer (DFE)

Using a Cholesky factorization, the channel matrix $\mathbf{K}_b$ (which is also the covariance matrix of the error $\mathbf{e}$) can be diagonalized as $\mathbf{K}_b^{-1} = \mathbf{B}^T\mathbf{\Delta}\mathbf{B}$. This diagonalization forms a Decision-Feedback Equalizer (DFE). The feed-forward filter is is formed by the concatenation of the two filters $\mathbf{\Delta}^{-1}\mathbf{B}^{-T}$ and $\mathbf{H}^T$. The feedback filter is the filter $\mathbf{I}-\mathbf{B}$. The DFE process is also shown in Fig.2.9(a). The finite-length minimum mean-square error decision-feedback equalizer (MMSE-DFE) has proven to be an effective structure for combating ISI. The design of the MMSE-DFE filters requires the (full) knowledge of the ISI term, treated as channel state information (CSI) at the receiver (this CSI is most of the times obtained through training). In practice, the CSI at the receiver is noisy. Potential noise sources include estimation and/or channel time variations. Also, a phenomenon that might significantly degrade the MMSE-DFE performance is catastrophic error propagation. A means of circumventing this problem is to assign this task to the transmitter. If the CSI is available at the transmitter, then the feedback portion of MMSE-DFE can be designed and implemented at the transmitter. This structure is also known as the MMSE-Tomlinson-Harashima precoder.

2.4.3.3 Tomlinson-Harashima Precoding (THP)

Tomlinson [Tom71] and Harashima [MH69, HM72] independently introduced precoding as a technique for inter-symbol interference mitigation. The structure that they presented is referred to as the Tomlinson-Harashima Precoder (THP). The general THP system is shown in Fig.2.9. $\mathbf{F}$ and $\mathbf{B}$ are the two feed-forward and feed-backward filters, respectively. The THP precoder output x_k is filtered by $\mathbf{H}$ and Gaussian noise n_k is added producing the sequence seen at the receiver. As shown in Fig.2.9, the signal x_k transmitted over the channel is formed by modulo-reducing the s_k. The modulo-t function Γ_t is used for mapping the real numbers $\mathbb{R}$ to $(-t/2, t/2]$ where t is any positive real number. This modulo-operation can be viewed as the signal-dependent addition $\Gamma_t(s_k) = s_k + a_k$, where a_k is the integer multiple of t for which $x_k \in (-t/2, t/2]$. Alternatively, $\Gamma_t(\cdot)$ can be viewed as a real-valued quantizer. This quantization is message-indexed (through w_k). In fact, by moving the channel knowledge from the receiver to the encoder (which reduces computational complexity in a variety of situations as in broadcast situations) the channel in Fig.2.9 is made equivalent to that in Fig.2.3 for which we argued, in the previous section, that QIM is well suited. Hence, THP relies basically on a form of indexed quantization. Of course, this does not mean that THP has been somehow inspired by the way in which QIM is designed, simply because it has been proposed in 1972, i.e., earlier than QIM. However, this means that, naturally, the two solutions of the two related problems (information embedding and communication over ISI channels) bear resemblance to each other. While not surprising, this understanding of THP as QIM has, nevertheless, not been recognized as such until very recently. More precisely, this analogy between THP and QIM can be further extended. For instance, like QIM, THP has two forms: its simplest form ZF-THP and its more involved form MMSE-THP. Broadly speaking, MMSE-THP has the same advantages, over ZF-THP, that DC-QIM has over regular (or ZF-)QIM. Also, consistent with this analogy, the MMSE choice of the optimal filter in MMSE-THP resembles the optimal choice of parameter α in DC-QIM. In fact, the influence of the inflation parameter α itself can be understood as filtering as mentioned above.

2.4.4 The Scalar Costa Scheme (SCS)

In section 2.2.2, we mentioned that the optimal dirty paper coding, DPC, is largely impractical for it relies on random codes and requires an exhaustive search strategy for selecting the appropriate codeword. In Costa's DPC, the codebook $\mathcal{U}^n$ is constructed as

$$\begin{aligned} \mathcal{U}^n &= \{\mathbf{u}_k = \mathbf{x}_k + \alpha\mathbf{s}_k \mid k \in \{1, 2, \ldots, L_u\}, \\ & \quad \mathsf{x} \sim \mathcal{N}(0, PI_n), \ \mathsf{s} \sim \mathcal{N}(0, QI_n)\}, \end{aligned} \tag{2.53}$$

where $\mathbf{x}$ and $\mathbf{s}$ are realizations of two n-dimensional independent random processes x and s with Gaussian PDF, I_n denotes the n-dimensional identity matrix and $\alpha = P/(P+N)$ is the optimal Costa parameter. The codebook $\mathcal{U}^n$ has cardinality $L_u = \lceil 2^{n(I(U;Y)-\epsilon)} \rceil$, ($\epsilon$ is an arbitrary small positive number), and is partitioned into L_M disjoint sub-codebooks $\{\mathcal{U}_i^n\}$, $i = 1, 2, \ldots, L_M$, in such a way that the total codebook $\mathcal{U}^n$ writes

$$\mathcal{U}^n = \mathcal{U}_1^n \cup \mathcal{U}_2^n \cup \ldots \cup \mathcal{U}_W^n \cup \ldots \cup \mathcal{U}_{L_M}^n. \tag{2.54}$$

The size L_u of the codebook $\mathcal{U}^n$ can become very large, even for small values of the length n and the size M of the alphabet $\mathcal{M}$, thus making the problem of storing and searching the codebook difficult. While leaving the main concept of Costa's DPC unchanged, Eggers and al. [EBTG03] proposed the use of a structured codebook. This is chosen to be a product of dithered uniform scalar quantizers, thus the denomination "Scalar Costa Scheme" (SCS). With respect to the optimal DPC, the codebook $\mathcal{U}^n$ can be written in the form

$$\mathcal{U}^n = \overbrace{\mathcal{U}^1 \circ \mathcal{U}^1 \circ \cdots \circ \mathcal{U}^1}^{n \text{ times}}, \tag{2.55}$$

where $\mathcal{U}^1$ is a one-dimensional component codebook. $\mathcal{U}^1$ is separated into M disjoint sub-codebooks so that

$$\mathcal{U}^1 = \mathcal{U}_0^1 \cup \mathcal{U}_1^1 \cup \cdots \cup \mathcal{U}_W^1 \cup \cdots \cup \mathcal{U}_{M-1}^1. \tag{2.56}$$

The entries of codebook $\mathcal{U}$ are formed with the reconstruction points (vectors) of the uniform scalar quantizer $\mathcal{Q}_\Delta(.)$ of constant step size Δ, i.e.,

$$\mathcal{U}^1 = \left\{ u = k\alpha\Delta + W\frac{\alpha\Delta}{M} \mid W \in \mathcal{M}, k \in \mathbb{Z} \right\} \tag{2.57}$$

and the W-th sub-codebook of $\mathcal{U}^1$ is given by

$$\mathcal{U}_W^1 = \left\{ u = k\alpha\Delta + W\frac{\alpha\Delta}{M} \mid k \in \mathbb{Z} \right\}. \tag{2.58}$$

Note that, as in the previous sections, a secure pseudo-random sequence $\{k_n\}$ can be introduced as an additional shift in the codebook $\mathcal{U}^1$, for security purposes. This encryption procedure does not modify the codebook properties and for instance the minimum distance d_{min} between the M different sub-codebooks remains unchanged.

2.4.4.1 SCS encoder

The encoder designs the codeword $\mathbf{x}$ to be a scaled version of the quantization error of the host signal $\mathbf{s}$, i.e., $\mathbf{x} = \mathbf{u} - \alpha\mathbf{s} = \alpha\mathbf{q}$. The quantization is performed in a sample-wise operation, i.e.,

$$x_n = \alpha \left\{ \mathcal{Q}\left(s_n - \Delta\left(\frac{W}{M} + k_n\right)\right) - \left(s_n - \Delta\left(\frac{W}{M} + k_n\right)\right)\right\}. \tag{2.59}$$

Under the well known high resolution quantization assumption $Q \gg P$, this encoding process incurs no noticeable distortion to the host.

2.4.4.2 SCS decoder

Decoding is also based on uniform scalar quantization of the received signal $\mathbf{y} = \mathbf{x} + \mathbf{s} + \mathbf{v}$ followed by a thresholding procedure. For instance, the decoder first computes the quantization error

$$r_n = \mathcal{Q}_\Delta\{y_n - k_n\Delta\} - (y_n - k_n\Delta) \tag{2.60}$$

and then sets the estimate $\hat{W}$ of the transmitted index W as the closest integer to $\dfrac{r_k}{\Delta/M}$. They are precisely the same principles of QIM that make the SCS independent on the characteristics of the host signal $\mathbf{s}$, at least when the channel noise $\mathbf{v}$ is not too strong.

2.4.5 Channel capacity

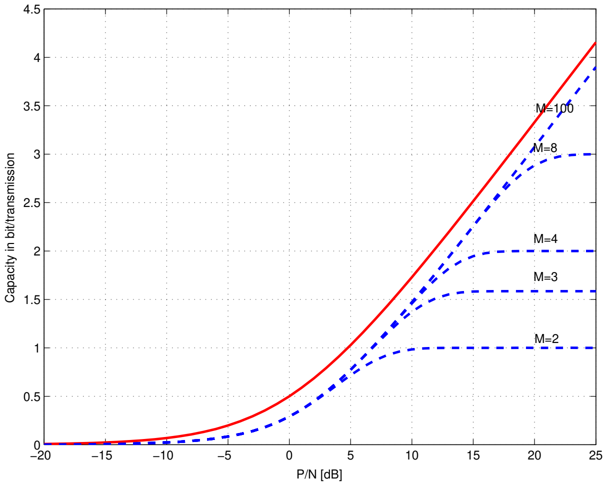
We begin this section by the following note. In this section, as well as in the rest of this work, we will loosely use the term ”channel capacity” to refer to the maximum transmission rate achievable with a given scheme. Of course, strictly speaking this is not the *information capacity* originally considered by Shannon. While Shannon's information capacity is independent on the choice of the encoder and/or the decoder, both of these are fixed here by fixing the scheme. Instead, one should rather speak of the *feasible capacity*, for this is the best transmission rate that the use of the scheme would allow. Shannon's information capacity can then be viewed as the maximum *feasible capacity.* The maximization is over all possible encoders and all possible decoders. However, we will ignore the discrepancy between the two in the following, as it is usual in classical communication.

We consider the channel capacity obtained by use of the suboptimal schemes QIM and SCS. We first notice that the above referred to as *regular* QIM is a special case of the coding process in the SCS, obtained with the choice $\alpha = 1$ in (2.59). Capacity of the SCS is obtained by numerically maximizing, over α, the mutual information $I(r; W)$ between the encoder input W and the decoder output r. Capacity of *regular* QIM follows straightforwardly, i.e.,

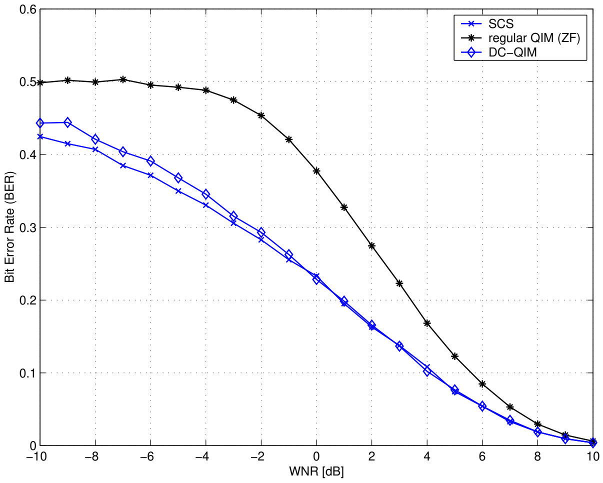
$$C_{\text{SCS}} = \max_\alpha I(r; W), \tag{2.61a}$$

$$C_{\text{QIM}} = I(r; W)|_{\alpha=1}. \tag{2.61b}$$

The mutual information $I(r; W)$ has no closed form expression since the conditional distributions $p(r|k)$ and $p(r|W, k)$ can be obtained only numerically. That the maximization must be performed over α follows the same reason as for the optimal DPC coding. The optimum scale parameter α that achieves



(a) Capacity



(b) Bit Error Rate

Figure 2.10: 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. Left, M-ary SCS capacity (dashed) approaches the full AWGN capacity (solid) as $M \to +\infty$. Right, SCS outperforms -by far- regular QIM in terms of BER. A slight improvement over DC-QIM is observed at very low $\text{SNR} = 10\log_{10}(P/N)$.

the scalar capacity in (2.61a) is

$$\alpha = \sqrt{\frac{P}{P + 2.71N}}, \tag{2.62}$$

which is, as for the optimal DPC scheme, independent of the host signal $\mathbf{s}$. The performance of both scalar QIM and SCS are depicted in Fig.2.10.

2.4.6 Gaps to Capacity

Capacity and BER curves are depicted in Fig.2.10. Observe that, with respect to the performance of blind SS depicted in Fig.2.7, the binary SCS performs close to the ideal DPC at low SNR. Further, the important gap visible at high SNR can be bridged up by use of M-ary transmissions. The same observation is valid for the BER: small BERs are possible at large SNR. However, there is a relatively important gap to full AWGN performance, even with infinite alphabet size and for both low and high SNR ranges. This means that even asymptotically (in both SNR and alphabet size), both SCS and QIM are suboptimal. Observe for instance the poor error rates at low to medium SNRs, with SNR $= 10\log_{10}(P/N)$. For example we have BER $\approx 10^{-1}$ at SNR ≈ 4 dB. Such error rates are of course too high to enable reliable transmission. This sub-optimality is precisely the one mentioned above when comparing algebraic coding to random coding. For instance, this gap is due to the use of a uniform scalar quantizer codebook $\mathbf{U} = \Delta\{\pm\alpha/2, \pm 3\alpha/2, \ldots, \pm(M-1)\alpha/2\}$, from an information theoretic point of view. From a communication point of view, such a codebook amounts to the use of an M-point Pulse Amplitude Modulation (PAM) signal, which is non-optimal as it is well known from classical digital communication theory. In Chapter 3, we will concentrate on the design of appropriate codebooks that partially bridge up the gap mentioned above. We will see that a certain *shaping gain* can be obtained at low SNR. Also, a certain *coding gain* can be obtained at high SNR, when the codebook is properly designed.

2.5 Summary

In this chapter, we provided a brief description of the problem of Channel Coding with Side Information (CCSI) non-causally available at the transmitter, also commonly known as "Gel'fand-Pinsker problem". We also addressed its Gaussian version, or Costa problem, together with the optimal Dirty Paper Coding (DPC). Though theoretically optimal, DPC is unfeasible in practice due to the huge size of the involved random codebook. However, DPC represents the theoretical foundation for the suboptimal low-complexity techniques which are relevant for practical implementation. As quantization is a key element in the design of algebraic implementable codebooks for the solutions of side-information systems, most relevant coding techniques are quantization-based. We also provided in this chapter a brief analysis of the most important information embedding methods, that are Quantization Index Modulation (QIM) and the Scalar Costa Scheme (SCS). QIM and SCS, which are two forms of one single coding strategy based on dithered quantizers, outperform, by-far, Spread Spectrum Modulations (SSM) techniques. SSM methods are among the earliest embedding functions considered in information embedding, but they greatly suffer from the interference due to the host signal (the side information).

Chapter 3

Lattices and Nested Lattices for Source-Channel Coding in Information Embedding

The content of this chapter has been partially published in [ZD05a, ZD05c, ZD06a].

In the previous chapter, we studied the performance of both the famous Scalar Costa Scheme (SCS) and Quantization Index Modulation (QIM), with comparison to the optimal Dirty Paper Coding (DPC). We also mentioned that, due to the sample-wise quantization, a certain gap to the ultimate DPC performance exists. Finite-dimensional lattice quantization, which is an important class of structured Vector Quantization (VQ), should improve reachable rates [EZ04] and hence, partially reduce this gap. In the first part of this chapter, coset-based codes, which are often proposed as an alternative to the theoretical probabilistic random binning in network coding, are used to devise a structured high dimensional Costa scheme for information embedding. Both asymptotic and finite-dimensional performance are considered within the context of communication over a Modulo Lattice Additive Noise (MLAN) channel. Next, we address the problem of optimal codebook selection by exploiting the appealing algebraic structure of the lattice. Three possible choices for channel codewords are compared, raising the question of an unavoidable trade-off between reliable transmission (low error rates) and payload (high transmission rates). Then, guidelines taken from shaping of multidimensional constellations [FW89, GDF89] are used to reformulate the problem of codebook selection as the search for a good lattice support region from which codebook elements should be selected. In the second part of this chapter, we use a binning interpretation to argue that information embedding can also be understood as a source-channel coding problem and that nesting of lattices provides means of constructing efficient low complexity good source-channel codes. By emphasizing the interaction between shaping (provided by the coarse lattice) and coding (provided by the fine lattice) we give insight -through an example- into the construction of *efficient* fine/coarse lattices.

3.1 Preliminaries on Lattices

This section provides a brief introduction to lattices. Only the ingredients required in the rest of this chapter are reviewed. An extensive focus can be found in [CS88].

3.1.1 Lattices

Algebraically, an n-dimensional real lattice Λ is a discrete additive subgroup of $\mathbb{R}^n$ defined as $\Lambda = \{\mathbf{G}.\mathbf{u} : \mathbf{u} \in \mathbb{Z}^n\}$, where $\mathbf{G}$ is an $n \times n$ full-rank generator matrix. Geometrically, a lattice Λ is an infinite regular array that covers n-space uniformly. For example (a) the simplest n-dimensional lattice is the integer lattice $\mathbb{Z}^n$ which consists of all n-vectors with integer coordinates, (b) the lattice family A_n, $n \in \mathbb{N}$, is defined as $A_n = \{(x_0, x_1, \ldots, x_n) \in \mathbb{Z}^{n+1} \,:\, x_0 + \ldots + x_n = 0\}$ and (c) the lattice family D_n is defined as $D_n = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n \,:\, x_1 + \ldots + x_n = \text{even}\}$. The fundamental Voronoi region $\mathcal{V}$ of Λ is the set of points $\mathbf{x} \in \mathbb{R}^n$ that are closer to $\mathbf{0}$ than to any other lattice point $\boldsymbol{\lambda} \in \Lambda$, i.e,

$$\mathcal{V}(\Lambda) \triangleq \left\{ \mathbf{x} \,:\, \|\mathbf{x}\| \leq \|\mathbf{x} - \boldsymbol{\lambda}\|, \, \forall \, \boldsymbol{\lambda} \in \Lambda \right\}.$$

For example, the Voronoi region of $\mathbb{Z}^n$ consists of all n-vectors that lie within a cubic region of unit volume, centered at the origin. The fundamental volume of Λ is the volume of its Voronoi region, i.e,

$$V(\Lambda) \triangleq \text{Vol}\left(\mathcal{V}(\Lambda)\right) = \int_{\mathcal{V}(\Lambda)} d\mathbf{x} = \sqrt{\det(\mathbf{G}^T \mathbf{G})}.$$

The second moment of $\mathcal{V}(\Lambda)$, or simply of Λ, is defined as

$$\sigma^2(\Lambda) \triangleq \frac{1}{nV(\Lambda)} \int_{\mathcal{V}(\Lambda)} \|\mathbf{x}\|^2 d\mathbf{x}$$

and its normalized second moment is the dimensionless quantity $G(\Lambda) \triangleq V(\Lambda)^{-\frac{2}{n}} \sigma^2(\Lambda)$. $G(\Lambda)$ is a dimensionless measure of the covering efficiency of Λ. The normalized second moment of the integer lattice $\mathbb{Z}^n$ is $G(\mathbb{Z}^n) = 1/12$. The covering radius $r_{cov}(\Lambda)$ is the radius of the smallest n-dimensional ball centered at the origin that contains $\mathcal{V}(\Lambda)$. The packing radius $\rho(\Lambda)$ is the radius of the biggest n-dimensional ball centered at the origin and contained in $\mathcal{V}(\Lambda)$. The points of $\mathbb{R}^n$ located at the vertices of $\mathcal{V}(\Lambda)$ are called lattice holes. Kissing points are points in $\mathcal{V}(\Lambda)$ that are at distance $\rho(\Lambda)$ from Λ. The kissing number $K(\Lambda)$ is the number of kissing points, or equivalently, the number of neighbors of any lattice point $\boldsymbol{\lambda} \in \Lambda$. Finally, the minimum distance of a lattice Λ is defined -as in coding theory- as the distance between the two points of this lattice that are closest, i.e,

$$d_{min} \triangleq \min_{(i,j):i \neq j} \; \min_{(\boldsymbol{\lambda_i}, \boldsymbol{\lambda_j}) \in \Lambda \times \Lambda} \|\boldsymbol{\lambda_i} - \boldsymbol{\lambda_j}\|.$$

3.1.2 Lattice Codebooks and Nesting of Lattices

A lattice codebook is a finite subset of a lattice or of a translated lattice, and may be specified as follows. Let $\Lambda_{\mathbf{c}} = \mathbf{a} + \Lambda$ be a translated n-dimensional lattice, and $\mathcal{R}$ be an n-dimensional *support region* of non-zero volume. Then the lattice codebook $\mathcal{C}(\Lambda_{\mathbf{c}}, \mathcal{R})$ is defined as

$$\mathcal{C}(\Lambda_{\mathbf{c}}, \mathcal{R}) = \Lambda_{\mathbf{c}} \cap \mathcal{R}.$$

That is, $\mathcal{C}(\Lambda_{\mathbf{c}}, \mathcal{R})$ consists of M points $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\}$ of $\Lambda_{\mathbf{c}}$ that lie in $\mathcal{R}$. If M is large enough, the size of the high-rate lattice codebook $\mathcal{C}(\Lambda_{\mathbf{c}}, \mathcal{R})$ is well approximated by

$$|\mathcal{C}(\Lambda_{\mathbf{c}}, \mathcal{R})| \cong V(\mathcal{R})/V(\Lambda_{\mathbf{c}}),$$

and its coding rate R is well approximated by

$$R \cong \frac{1}{n}\log_2[V(\mathcal{R})/V(\Lambda_{\mathbf{c}})] \text{ bits per dimension.}$$

Consider now a pair of lattices (Λ_1, Λ_2) with Λ_2 being a subgroup of Λ_1, algebraically. Geometrically, Λ_2 is a sub-lattice of Λ_1. The pair (Λ_1, Λ_2) is called nested (or more precisely Λ_2 is nested in Λ_1) in the sense that each point of Λ_2 is also a point of Λ_1 but not vice versa, i.e., $\Lambda_2 \subset \Lambda_1$. A nested lattice then consists of an n-dimensional *lattice partition* Λ_1/Λ_2 where the lattices Λ_1 and Λ_2 are respectively referred to as *fine* lattice and *coarse* lattice. In this case, there exist an $n \times n$ integer matrix $\mathbf{J}$ such that their corresponding generator matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ satisfy $\mathbf{G}_2 = \mathbf{G}_1.\mathbf{J}$ with $\det(\mathbf{J}) \geq 1$. An important parameter is the nesting ratio $\mu(\Lambda_1, \Lambda_2)$ defined as $\mu(\Lambda_1, \Lambda_2) \triangleq \sqrt[n]{\det(\mathbf{J})} = \sqrt[n]{V(\Lambda_2)/V(\Lambda_1)}$. The set of points of Λ_1 that are inside the Voronoi region of Λ_2 forms a lattice codebook $\mathcal{C}(\Lambda_1, \mathcal{V}(\Lambda_2)) = \{\Lambda_1 \cap \mathcal{V}(\Lambda_2)\}$. The elements of this lattice codebook are *coset leaders* (minimum-norm points) of the coarse lattice Λ_2 relative to the fine lattice Λ_1. For each $\mathbf{c} \in \mathcal{C}$ the translated lattice $\Lambda_{\mathbf{c}} = \mathbf{c} + \Lambda_2$ is called a coset of Λ_2 relative to Λ_1. Algebraically, Λ_1/Λ_2 forms a quotient group [1] whose order is related to the coding rate R of the nested lattice code,

$$R \triangleq \frac{1}{n}\log_2|\mathcal{C}| = \frac{1}{n}\log_2|\Lambda_1/\Lambda_2| = \frac{1}{n}\log_2\left(\mu(\Lambda_1, \Lambda_2)\right). \tag{3.1}$$

3.2 Lattice-based Information Embedding

The performance of the sample-wise transmission schemes considered in the previous chapter can be enhanced using structured low-complexity lattice-based codebooks. In Section 3.2.2 the corresponding capacity is derived using important insights from [ESZ00, ZSE02]. Error probability and the problem of codebook selection are addressed in Sections 3.2.3 and 3.2.4, respectively. Results are supported by some realistic finite-dimensional lattice implementations together with their capacity and probability of error curves.

3.2.1 Lattice coding for QIM Information Embedding

Consider the transmission scheme depicted in Fig.3.1 where Λ is some n-dimensional lattice. Assume that the encoder and the decoder share *common randomness* so that the key $\mathbf{K}$ is available to both of them.

[1] We may view each coset as an *equivalence class* for this quotient group and the leader of the coset as a *representative* of this equivalence class.

Apart from obvious security purposes this key will turn to be particularly useful to achieve capacity. Also, consider a certain mapping or indexing function $\iota(.)$ between the set of indexes $m \in \{1, \dots, M\}$

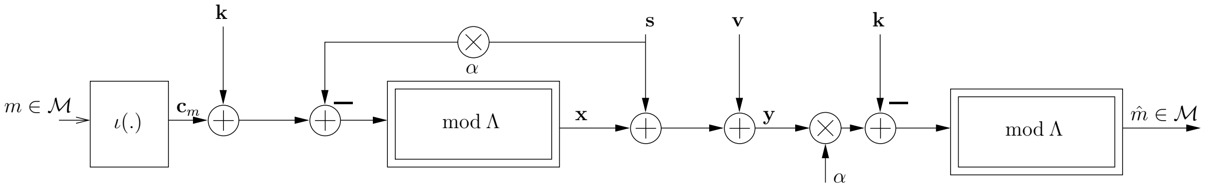


Figure 3.1: Information embedding based on modulo-reduction.

and a set of vectors $\mathcal{C}_m = \{\mathbf{c}_m \,:\, m = 1, \dots, M\}$ to be specified in the sequel. Mapping is performed in an arbitrary manner. Loosely speaking, $\mathcal{C}_m$ may be viewed as a lattice codebook in the sense given in Section 3.1. For each $m \in \mathcal{M}$, the vector $\iota(m) = \mathbf{c}_m$ is the *coset leader* of the coset $\Lambda_m = \mathbf{c}_m + \Lambda$ of the lattice Λ. The set of coset leaders is shared between the encoder and the decoder and is assumed to be uniformly distributed over the fundamental cell $\mathcal{V}(\Lambda)$ of the lattice Λ. The key $\mathbf{k}$ may be used as a dither. Dithering is a well known capacity-maximizing technique [ESZ00] the usefulness of which will be outlined hereafter. Also, we may view a coset Λ_m as a codebook bin in the original random binning coding argument in [Cos83]. In the following, we consider host signal vectors (frames) of length n, i.e. the same as the dimension of the lattice Λ. Also, we use the modulo reduction operation $\bmod \Lambda$ with respect to the Voronoi region $\mathcal{V}$ of the lattice Λ. This modulo operation is defined as $\mathbf{x} \bmod \Lambda \triangleq \mathbf{x} - \mathcal{Q}_\Lambda(\mathbf{x}) \in \mathcal{V}(\Lambda)$ where the n-dimensional quantizer 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 $\mathbf{x}$. The received signal is

$$\mathbf{y} = \mathbf{x} + \mathbf{s} + \mathbf{v}, \tag{3.2}$$

sum of the watermark $\mathbf{x}$, the host signal $\mathbf{s}$ and some extra channel noise $\mathbf{v}$.

3.2.1.1 Outline of lattice-based structured binning

Prior to dealing with optimal coding for communication over the channel depicted in Fig.3.1, we establish a brief parallel with the original random binning technique used for channels with state information. Such a parallel helps the understanding of lattice quantization as a search for (distortion) joint-typical sequences and makes the encoding/decoding process (3.3) straightforward. In [Cos83], the encoder first partitions the codebook entries uniformly into $M = 2^{nR}$ bins and then, given a sequence $\mathbf{S}$ and an index $m \in \mathcal{M}$, searches in bin m for a sequence $\mathbf{U}$ such that $(\mathbf{U}, \mathbf{S})$ is jointly typical. Next, the encoder chooses a sequence $\mathbf{X}$ such that $(\mathbf{X}, \mathbf{U}, \mathbf{S})$ is jointly typical and sends it through the channel. Upon reception of $\mathbf{Y} = \mathbf{X} + \mathbf{S} + \mathbf{V}$, the decoder searches in the entire codebook (union of bins) for the (unique) sequence $\mathbf{U}$ such that $(\mathbf{U}, \mathbf{Y})$ is jointly typical and then sets the estimate $\hat{m}$ of m as the index of the bin containing the obtained sequence $\mathbf{U}$. In the structured lattice-based algebraic scheme in Fig.3.1, we may view the $M = 2^{nR}$ cosets $\Lambda_m, m = 1, \dots, M$ as bins. Given $m \in \mathcal{M}$, the encoder quantizes $\alpha\mathbf{s} - \mathbf{k}$ to the nearest point in Λ_m, obtaining $\mathbf{u} = \mathcal{Q}_{\Lambda_m}(\alpha\mathbf{s} - \mathbf{k})$. It then transmits $\mathbf{x} = (\alpha\mathbf{s} - \mathbf{k}) - \mathbf{u} = (\alpha\mathbf{s} - \mathbf{k}) \bmod \Lambda_m$. Note that loosely speaking, $\mathbf{U}$ is D_E-distortion typical with $\mathbf{S}$, meaning that $(\mathbf{U}, \mathbf{S})$ is jointly typical and that, in addition, the norm of the quantization error $\mathbf{X}$ is inferior to D_E. A good lattice for quantization would be one that minimizes the norm of this error. Upon reception, the receiver, not knowing the exact bin Λ_m that has been used in the encoding process, quantizes the scaled received signal $\alpha\mathbf{y} - \mathbf{k}$ to the nearest point in the structure formed by the union of all cosets

$$\bigcup_{m\in\mathcal{M}} \Lambda_m,$$

and sets the estimate $\hat{m}$ as the index of the coset (bin) containing the obtained point. Encoding and decoding are given by

$$\mathbf{x}(\mathbf{s}; m, \Lambda) = (\mathbf{c}_m + \mathbf{k} - \alpha\mathbf{s}) \bmod \Lambda, \tag{3.3a}$$

$$\hat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \quad \min_{\boldsymbol{\lambda}\in\Lambda_m} \|\alpha\mathbf{y} - \mathbf{k} - \boldsymbol{\lambda}\|. \tag{3.3b}$$

It is important to note that by the properties of dithered quantization, the input constraint

$$\frac{1}{n}\mathbb{E}_{\mathbf{K}}[\mathbf{X}^2|\mathbf{S} = \mathbf{s}, \mathbf{C}_m = \mathbf{c}_m] = P, \tag{3.4}$$

is fulfilled, independently on the individual values of $\mathbf{c}_m$ and $\mathbf{s}$. Note also that the one-dimensional SCS is obtained with the particular case of an integer lattice $\Lambda = \mathbb{Z}$ where signals are scaled according to $\mathbf{s}' = \mathbf{s}/\alpha$, $\mathbf{x}' = \mathbf{x}/\alpha$ and $\mathbf{v}' = \mathbf{v}/\alpha$. The optimum value $\alpha = \frac{P}{P+N}$ of DC-QIM inflation parameter $\alpha \in (0, 1]$ is chosen such that it increases the inter-quantizers (cosets) minimum distance, keeps the embedding distortion unchanged and minimizes the channel noise interference at the decoder. In this lattice scheme, regular QIM corresponds -as for the scalar case- to $\alpha = 1$ and Dither Modulation (DM) with constant step size Δ proposed in [CW01] is obtained with the cubic lattice $\Lambda = \Delta\mathbb{Z}^N$.

3.2.2 Capacity analysis

First, recall the following two important properties of the mod-Λ operation defined above. These properties will be extensively used throughout this chapter.

$$\text{(P1)} \quad (\mathbf{a} + \mathbf{v} + \boldsymbol{\lambda}) \bmod \Lambda = (\mathbf{a} + \mathbf{v}) \bmod \Lambda, \quad \forall\, (\boldsymbol{\lambda}, \mathbf{a}) \in \Lambda \times \mathbb{R}^n. \tag{3.5a}$$

$$\text{(P2)} \quad ((\mathbf{x} \bmod \Lambda) + \mathbf{y}) \bmod \Lambda = (\mathbf{x} + \mathbf{y}) \bmod \Lambda, \quad \forall\, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2n}. \tag{3.5b}$$

Consider now the mod-Λ channel depicted in Fig.3.2(a). This channel has been first considered in [FTC00] where it has been shown that, assuming a channel noise $\mathbf{V}$ independent of the input $\mathbf{C}_m$, capacity is achieved if $\mathbf{C}_m$ is uniformly distributed over $\mathcal{V}(\Lambda)$. In this case, the capacity $C(\Lambda)$ of the channel satisfies (in bits per-dimension)

$$C(\Lambda) = \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\tilde{\mathbf{V}})\right) < \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right), \tag{3.6}$$

where $h(.)$ denotes differential entropy and the noise term $\tilde{\mathbf{V}} = \mathbf{V} \bmod \Lambda \in \mathcal{V}(\Lambda)$ is the quantization error, with respect to Λ, of the WGN noise $\mathbf{V}$. The right hand side term of (3.6) is the full capacity $C^{\max}$ of an AWGN channel with Signal-to-Noise Ratio P/N. A key idea in deriving the capacity of the

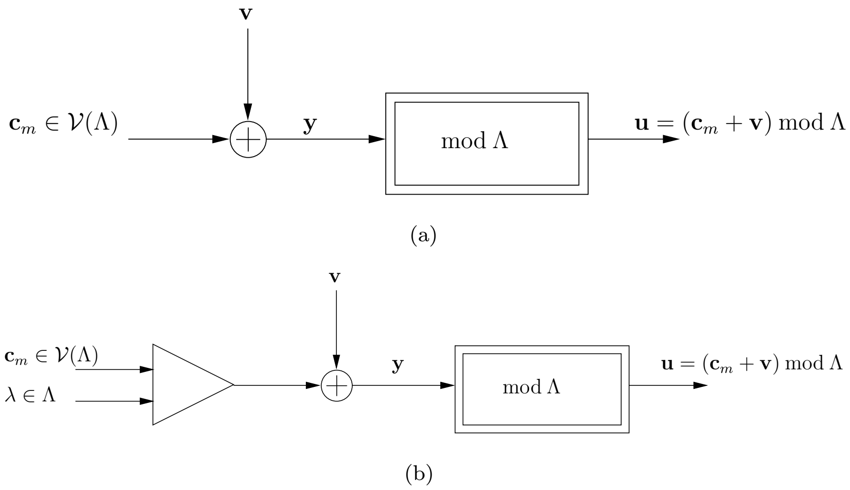


Figure 3.2: (a) The typical mod-Λ channel. (b) Equivalent channel: adding a lattice point $\lambda \in \Lambda$ at the channel input does not change its output.

channel in Fig.3.1 is that the output of a mod-Λ channel does not change if a lattice point $\lambda \in \Lambda$ is added at its input as shown in Fig.3.2(b). This is due to the property (P1) of mod-Λ reduction. Consider now the case $\alpha = 1$ (regular lattice QIM). This case clearly shows an important property of mod-Λ channels. Generalization to $\alpha \neq 1$ is undertaken in what follows. Using the distributive property (P2) of the modulo operation, equation (3.3b) can be re-written as

$$\hat{m} = \underset{m=1,\ldots,M}{\operatorname{argmin}} \|(\mathbf{y} - \mathbf{k} - \mathbf{c}_m) \bmod \Lambda\|,$$

with

$$\begin{aligned}(\mathbf{y} - \mathbf{k} - \mathbf{c}_m) \bmod \Lambda &= ((\mathbf{c}_m + \mathbf{k} - \mathbf{s}) \bmod \Lambda + \mathbf{s} + \mathbf{v} - \mathbf{k} - \mathbf{c}_m) \bmod \Lambda, \\ &= \mathbf{v} \bmod \Lambda. \end{aligned} \tag{3.7}$$

Then the modulo decoder sees the signal $\mathbf{y} - \mathbf{k} = \mathcal{Q}_\Lambda(\mathbf{y} - \mathbf{c}_m - \mathbf{k}) + \mathbf{c}_m + \tilde{\mathbf{v}}$, with $\tilde{\mathbf{v}} = \mathbf{v} \bmod \Lambda$ being the equivalent channel noise. Hence the channel input is $\mathcal{Q}_\Lambda(\mathbf{y} - \mathbf{c}_m - \mathbf{k}) + \mathbf{c}_m$, which is the sum of one lattice point $\mathcal{Q}_\Lambda(\mathbf{y} - \mathbf{c}_m)$ and the vector $\mathbf{c}_m$ uniformly distributed over $\mathcal{V}(\Lambda)^2$. The communication channel depicted in Fig.3.1 is thus equivalent to that in Fig.3.2(b), which is itself equivalent to that in Fig.3.2(a). The channel noise $\tilde{\mathbf{v}}$ has a probability density function (PDF) given by the restriction of a Gaussian PDF over $\mathcal{V}(\Lambda)$,

$$f_{\tilde{\mathbf{V}}}(\tilde{\mathbf{v}}) = \begin{cases} (2\pi N)^{-n/2} \sum_{\boldsymbol{\lambda} \in \Lambda} \exp\left(-\frac{\|\tilde{\mathbf{v}} - \boldsymbol{\lambda}\|^2}{2N}\right), & \text{if } \tilde{\mathbf{v}} \in \mathcal{V}(\Lambda) \\ 0, & \text{if } \tilde{\mathbf{v}} \notin \mathcal{V}(\Lambda). \end{cases} \tag{3.8}$$

Therefore, lattice (regular) QIM capacity is also given by (3.6) and the lattice watermarking scheme is equivalent to communicating over an MLAN channel. However, in general no closed form for (3.6) can

[2]It now should become clear why we made such an assumption in Section 3.2.1

be derived and numerical integration is needed to evaluate the differential entropy.

We now turn to the general case $\alpha \neq 1$. The receiver computes $\mathbf{y}' = (\alpha\mathbf{y} - \mathbf{k}) \bmod \Lambda$. Using (P2) and writing $\alpha\mathbf{y} = \mathbf{y} - (1-\alpha)\mathbf{y}$, $\mathbf{y}'$ can be rewritten [ZSE02] as

$$\mathbf{y}' = (\mathbf{c}_m + \alpha\mathbf{v} - (1-\alpha)\mathbf{x}) \bmod \Lambda. \tag{3.9}$$

The equivalent channel noise $\tilde{\mathbf{v}} = (\alpha\mathbf{v} - (1-\alpha)\mathbf{x}) \bmod \Lambda$ generalizes that corresponding to the Zero-Forcing (ZF) approach. However, in order to satisfy the MLAN channel requirements, the noise $\tilde{\mathbf{V}}$ has to be statistically independent of the input $\mathbf{C}_m$. This is ensured by the following *Inflated Lattice Lemma* reported in [ESZ00].

Lemma 1 (Inflated Lattice Lemma [ESZ00]) *The channel from $\mathbf{C}_m$ to $\mathbf{Y}'$, defined by (3.2), (3.3a) and (3.9) is equivalent in distribution to the channel*

$$\mathbf{Y}' = (\mathbf{C}_m + \mathbf{V}')\, mod\, \Lambda, \tag{3.10}$$

where $\mathbf{V}'$ is independent of $\mathbf{C}_m$ and is given by

$$\mathbf{V}' = (\alpha\mathbf{V} - (1-\alpha)\mathbf{U})\, mod\, \Lambda, \tag{3.11}$$

and $\mathbf{U}$ is a random variable uniformly distributed over $\mathcal{V}(\Lambda)$ and is statistically independent of $\mathbf{V}$.

Note that the independence is achieved even if the high resolution quantization assumption $Q \gg P$ is violated. The key idea for the proof is based on the fact that dithering (by the use of the key $\mathbf{K}$) makes $\mathbf{X}$ (almost) uniform over $\mathcal{V}(\Lambda)$. Thus, transmission over the channel in Fig.3.1 is equivalent to that over an MLAN channel (modulo Λ) with input $\mathbf{c}_m$ and noise $\tilde{\mathbf{v}}$. However, due to the inflation parameter α, the equivalent noise $\tilde{\mathbf{v}}$ is no longer the restriction of a Gaussian noise over $\mathcal{V}(\Lambda)$, but the convolution of a uniform self noise $(1-\alpha)\mathbf{x}$ and the ambient Gaussian noise $\alpha\mathbf{v}$. Consequently, equation (3.6) is slightly modified and capacity is given by the supremum of (3.6) over all values of parameter $\alpha \in (0, 1]$. This capacity is attained with a uniform input and it satisfies (in bits per dimension)

$$C(\Lambda) = \max_{\alpha} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\tilde{\mathbf{V}})\right) < \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right). \tag{3.12}$$

Note that with respect to the ZF approach, the optimal inflation parameter in (3.12) enables the receiver to pull down the noise $\tilde{\mathbf{v}}$ before lattice decoding. By this operation, lattice decoding achieves close to Maximum-Likelihood (ML) decoding which corresponds to the right hand side term of (3.12).

3.2.2.1 Asymptotes and approximations

In general no closed form of (3.12) can be derived and numerical integration is needed to evaluate the differential entropy $h(\tilde{\mathbf{V}})$. However, several special cases are worthy of some discussion.

(i) In the ZF-approach ($\alpha = 1$), as WNR $= 10\log_{10}(P/N)$ becomes large ($N \to 0$), the probability that the noise $\mathbf{V}$ falls outside $\mathcal{V}(\Lambda)$ decreases, the Λ-aliased noise $\tilde{\mathbf{V}}$ becomes approximately equal to $\mathbf{V}$ and the capacity $C(\Lambda)$ tends to $(1/n)\log_2(V(\Lambda)) - \frac{1}{2}\log_2(2\pi e N)$. On the other hand, as the (per-dimension) noise variance N becomes large, the mod-Λ channel becomes very noisy and its capacity tends to zero. Consequently, $C(\Lambda)$ can be asymptotically approximated by the piece-wise linear-curve $C^{\text{approx}} = \max\{0, \frac{1}{2}\log_2\left(\frac{V(\Lambda)^2}{(2\pi e N)^n}\right)\}$ for both low and high WNRs.

(ii) As the dimensionality n of the lattice goes to infinity, the PDF of the noise $\tilde{\mathbf{V}}$ tends to a Gaussian distribution. Thus, the optimal choice for parameter α is the one that minimizes the variance of $\tilde{\mathbf{V}}$, that is $\alpha = P/(P+N)$. With such a choice, the ultimate capacity C^{max} is attained.

(iii) For finite-dimensional lattice reduction, the PDF $f_{\tilde{\mathbf{V}}}(\tilde{\mathbf{V}})$ of the noise $\tilde{\mathbf{V}}$ is not strictly Gaussian. The optimal inflation parameter α has to be computed numerically. An approximation of the solution can be obtained by minimizing the variance of $\tilde{\mathbf{V}}$ and leads to $\alpha = P/(P+N)$, for which $\mathbb{E}_{\tilde{\mathbf{V}}}[\tilde{\mathbf{V}}^2] = \frac{PN}{P+N} = \alpha N$. In this case, we have

$$h(\tilde{\mathbf{V}}) \leq h\left(\alpha \mathbf{V} - (1-\alpha)\mathbf{X}\right) \leq \log(2\pi e \alpha N).$$

The first inequality follows since the modulo operation can only decrease the entropy. The second follows since for a given second moment a Gaussian random variable has the largest entropy. Consequently, a lower bound on $C(\Lambda)$ is

$$C(\Lambda) \geq \frac{1}{n}\left(\frac{1}{2}\log(1+P/N) - \frac{1}{2}\log(2\pi e G(\Lambda))\right), \tag{3.13}$$

meaning that for a given lattice Λ the theoretical gap to the full capacity $C^{\max}$ may be made smaller than $\log(2\pi e G(\Lambda))$. Good lattices for quantization ($G(\Lambda) \rightarrow \frac{1}{2\pi e}$) even nullify this gap.

3.2.2.2 Capacity and Shaping Gain

The volume $V(\Lambda)$ in (3.12) characterizes the average transmit power $\sigma^2(\Lambda) = P$ needed to transmit the set of indexes $m \in \mathcal{M}$. With respect to the baseline cubic lattice $\mathbb{Z}^n$, the reduction in this transmission power is given by the shaping gain $\gamma_s(\Lambda) = 1/12G(\Lambda)$ of the lattice Λ. Substituting $V(\Lambda)$ in (3.12) by its expression as a function of $\gamma_s(\Lambda)$, equation (3.12) becomes

$$C(\Lambda) \quad = \quad \max_{\alpha} \frac{1}{2}\log_2\left(12\sigma^2(\Lambda)\gamma_s(\Lambda)\right) - \frac{1}{n}h(\tilde{\mathbf{V}}). \tag{3.14}$$

Since the ultimate shaping gain (as the dimensionality of the lattice goes to infinity) is $\frac{\pi e}{6}$, the ultimate capacity gain provided by lattice coding over scalar coding approaches is 1.53 dB (i.e., 0.255 bit per dimension). Here also, the parallel with digital communications is strong. This shaping gain is the same as the one provided by multidimensional constellations over PAM constellations. However the limit of 1.53 dB can never be attained with finite dimensional lattice embedding. According to [KP93] the shaping gain of finite constellations is approximately given by

$$\gamma_s(\Lambda) \approx \frac{\pi e}{6}(1 - 2^{-2R}). \tag{3.15}$$

Thus, asymptotically, the capacity curve of (3.14) corresponding to the use of the lattice Λ can be viewed as the translation to the left (lower WNRs) of that corresponding to the cubic lattice $\mathbb{Z}^n$ by a factor equal to the shaping gain (3.15). This is supported by the finite dimensional capacity curves shown in Fig.3.3. These curves are obtained through Monte-Carlo integration.

3.2.2.3 Simulations and discussion

The n-dimensional lattices considered for Monte-Carlo capacity integration are summarized in Table 3.1, together with their most important parameters.

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
E_7	7-dimensional E_7 L.	7	0.0732	0.56	0.093
E_8	Gosset Lattice	8	0.0717	0.65	0.108

Table 3.1: Some finite-dimensional lattices with their important parameters

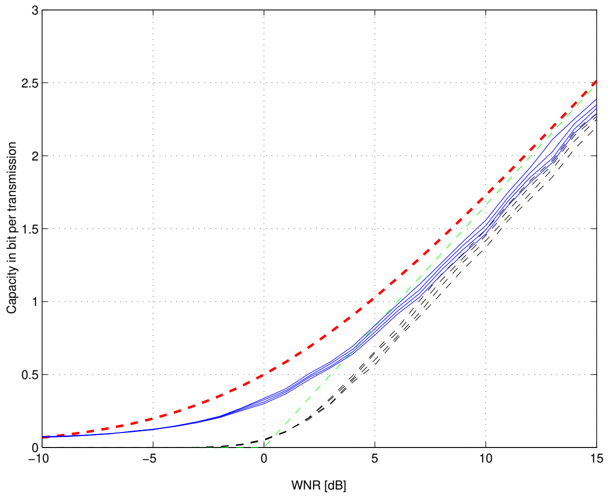


Figure 3.3: Capacity curves of lattice based transmission for some finite-dimensional lattices over the Watermark-to-Noise Ratio WNR $= 10\log_{10}(P/N)$. Bottom to top: $\mathbb{Z}$, A_2, D_4 and E_8 lattices. Solid: Capacity curves of DC-QIM. Dashed: AWGN capacity and asymptotic-limit. Dashed-dotted: Capacity curves of the Zero-Forcing approach.

For a given lattice Λ, the feasible capacity (3.6) is obtained by minimizing (over α) the entropy $h(\tilde{\mathbf{V}})$ of the equivalent noise $\tilde{\mathbf{V}} = \alpha\mathbf{V} - (1-\alpha)\mathbf{X}$ which is not strictly Gaussian as already mentioned above. Ignoring the non-Gaussianity of this equivalent noise, an approximation is obtained by considering the entropy of the equivalent Gaussian noise with the same variance. Computing the entropy of the restriction of this Gaussian noise to the Voronoï region $\mathcal{V}(\Lambda)$ is not straightforward, because it requires the computation of (3.8). One crucial point in computing (3.8) is to generate a sufficiently long random sequence (in the n-dimensional space) uniformly distributed over $\mathcal{V}(\Lambda)$. One possible solution is as follows:

(a) Generate n uniform random variables (in $[0,1)$) and map them via a generator matrix $\mathbf{G}(\Lambda)$ of the lattice to the fundamental region $\mathcal{V}(\Lambda)$.

(b) Take a lattice quantizer $\mathcal{Q}(\Lambda)$ and find the nearest lattice point to each generated vector. The quantization error is in the Voronoi region $\mathcal{V}(\Lambda)$ and is uniformly distributed over it.

The resulting capacity curves (in bits per dimension) are plotted in Fig.3.3. We observe that:

(i) Due to its small shaping gain, the integer lattice $\mathbb{Z}$ provides the lowest capacity. The gap to AWGN capacity is particularly large for low WNR. At low rates (below 0.1 bit/dimension), a gap of about 4 dB is observed. At high WNR, this gap is already partially bridged up using lattices A_2, D_4 and E_8.

(ii) The improvement due to the shaping gain $\gamma_s(\Lambda)$ of the lattice is particularly visible at high rates where the shaping gain (3.15) becomes significant. At low rates however, the shaping gain (3.15) is very small and the increase in capacity is marginal. Convergence toward the full AWGN capacity $C^{\max}$ is such that

$$0 \leq C^{\max} - C_\Lambda < \frac{1}{2}\log_2\left(2\pi e G(\Lambda)\right).$$

(iii) DC-QIM with optimal lattice encoding/decoding outperforms -as expected- the ZF approach. The gain is particularly large for low WNR. For rates above 2 bits/dimension, this gain is not significant. Also, the higher the lattice dimension n, the tighter are both the lower bound (3.13) and the approximation C^{approx}.

3.2.3 Error Probability Analysis

The analysis above has shown that increase in capacity due to high dimensional embedding is especially observed at high WNR. At low WNR however, the payload (capacity) does not matter much. Most important is the probability of error. In this section, this probability of error is discussed in the case of a ZF embedding approach. We will concentrate on the design of an efficient lattice codebook $\mathcal{C}_m$ that minimizes P_e.

3.2.3.1 An approximation

According to the binning interpretation given in Section 3.2.1, the decoder looks for the appropriate codeword among all possible codewords in the union of all cosets Λ_m, $m \in \mathcal{M}$. Hence the error probability depends mainly on the minimum distance between these cosets, given by

$$
\begin{aligned}
d_{min} &\triangleq \min_{1\leq i,j\leq M: i\neq j} \|\Lambda_i - \Lambda_j\|, \\
&= \min_{(i,j):i\neq j} \min_{(\boldsymbol{\lambda_i},\boldsymbol{\lambda_j})\in\Lambda_i\times\Lambda_j} \|\boldsymbol{\lambda_i} - \boldsymbol{\lambda_j}\|. \qquad (3.16)
\end{aligned}
$$

Suppose without loss of generality that $\mathbf{c}_1 = \mathbf{0}$ $(m = 0)$ is embedded into the cover signal $\mathbf{s}$, resulting in a received signal $\mathbf{y} = \mathcal{Q}_\Lambda(\mathbf{s} - \mathbf{c}_1) + \mathbf{c}_1 + \mathbf{v}$ (the key $\mathbf{k}$ is assumed to be null here). The decoder (3.3b) makes the correct decision if the nearest lattice point -among the set of all reconstruction points of all cosets- belongs to Λ_1. The error probability P_e can be expressed using the union bound. But noticing that this error probability is dominated by the two nearest cosets, P_e reads

$$
\begin{aligned}
P_e &\approx \Phi\left(\sqrt{\frac{d_{min}^2}{4N}}\right), \qquad (3.17)\\
&< \frac{1}{2}\exp\left(-\frac{d_{min}^2}{8N}\right), \qquad (3.18)
\end{aligned}
$$

where $\Phi(u) = \int_u^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u2/2} du$ is the tail probability of the Gaussian PDF. We consider the problem of selecting the optimal codebook $\mathcal{C}_m = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M\}$ for transmission over the channel in Fig.3.1. In a general setting, this problem is difficult. The optimal codebook should achieve very low error probability along with high transmission rates. Error probability and transmission rate are obviously conflicting requirements. The resulting problem can be formulated as a constrained optimization problem:

Given a certain minimum required transmission rate $R_{min} = \frac{1}{n}\log_2(M_{min})$ and a per-dimension distortion couple (P, N), select a lattice Λ of dimensionality n and a codebook $\mathcal{C}_m = \{\mathbf{c}_1, \ldots, \mathbf{c}_M\}$ with $\mathbf{c}_i \in \mathbb{R}^n$ so as to satisfy the following constraints:

$$
\begin{cases}
R \geq R_{min}, \ (\textit{i.e., a minimum guaranteed transmission rate}), \\
P_e \ \textit{as small as possible}, \ (\textit{i.e., reliable transmission}).
\end{cases}
$$

A simple measure of the extent by which a certain coding scheme $\{n,\Lambda,\mathcal{C}_m\}$ satisfies these conflicting requirements can be obtained using the parameter ν defined as

$$
\nu = \frac{1}{\sqrt{M}} \frac{d_{min}}{\sqrt{nP}} = 2^{-nR/2} \frac{d_{min}}{\sqrt{nP}}.
$$

Inequality (3.18) shows that minimizing P_e is equivalent to maximizing d_{min}, which amounts to maximizing ν. Satisfying the transmission rate requirement would require, as for it, to minimize ν. As a result, a proper choice of ν is based on a certain trade-off between rate transmission and reliability. Note that both d_{min} and $R = \frac{1}{n}\log_2 M$ depend on the choice of the codebook $\mathcal{C}_m$. Based on the geometrical structure of the lattice, we first address this problem through three examples. These examples correspond to different choices for coset leaders $\mathbf{c}_m, m \in \mathcal{M}$. A more general and stringent approach will be given in Section 3.3 within the context of source-channel coding.

(i) **Lattice holes:** Lattice holes have been introduced in Section 3.1. Two types of holes can be distinguished, *deep holes* and *shallow holes*. The former are the points of $\mathbb{R}^n$ that are furthest away from Λ, i.e., at a distance r_{cov} from it (see Fig.3.4). Moreover, it can be seen (as remarkably noticed in

[MGK04]) that $d_{min} \leq r_{cov}$. Thus, in order to maximize the inter-coset minimum distance d_{min}, these deep holes may be ideally used as coset leaders. However, two or more of these deep holes can generate the same coset, thus causing a decoding ambiguity at the receiver. Let N_h denote the number of these deep holes. To resolve any ties when a coset of Λ has more than one minimum-norm element, we choose a set of minimal vectors (somehow a basis, mathematically) $\mathbf{h}_1, \ldots, \mathbf{h}_{N_h^*}$ ($N_h^* \leq N_h$) such that

$$\bigcup_{i=1}^{N_h^*} \mathbf{h}_i + \Lambda = \bigcup_{i=1}^{N_h} \mathbf{h}_i + \Lambda. \tag{3.19}$$

In the following the lattice deep holes satisfying (3.19) are called *relevant deep holes.* The use of these relevant deep holes as codebook elements, i.e $\mathcal{C}_m = \{\mathbf{h}_1, \ldots, \mathbf{h}_{N_h^*}\}$, is optimal from a minimum distance d_{min} point-of-view, but not from a payload point-of-view, since it requires that $M \leq {N_h}^* + 1$, or equivalently that $R \leq \frac{1}{n}\log_2({N_h}^* + 1)$. Note that one could combine (relevant) deep holes and (relevant) shallow holes to form the codebook $\mathcal{C}_m$. However, while this surely increases the transmission rate, it inevitably leads to larger error probabilities. This is because the shallow holes are not as far away from Λ as are the deep holes.

(ii) **Kissing Points:** Kissing points defined in Section 3.1 are located at a distance equal to the packing radius $\rho(\Lambda)$ from the lattice Λ (see Fig.3.4). When used as coset leaders, ties can be resolved in exactly the same manner as for lattice holes. There exist then a set of minimum-norm vectors $\mathbf{k}_1, \ldots, \mathbf{k}_{N_k^*}$ ($N_k^* \leq N_k$) such that

$$\bigcup_{i=1}^{N_k^*} \mathbf{k}_i + \Lambda = \bigcup_{i=1}^{K(\Lambda)} \mathbf{k}_i + \Lambda. \tag{3.20}$$

Similarly to deep holes, the lattice kissing points satisfying (3.20) are called *relevant kissing points* in the following. Also, the choice $\mathcal{C}_m = \{\mathbf{k}_1, \ldots, \mathbf{k}_{N_k^*}\}$ gives a transmission rate R such that $R \leq \frac{1}{n}\log_2(N_k^* + 1)$.

(iii) **Construction A** A quite low complexity efficient method for increasing the transmission rate R with respect to the use of *relevant lattice holes* and *relevant kissing points* is Construction A [CS88]. Construction A provides means of constructing a lattice $\Lambda = C(n, k) + 2\mathbb{Z}^n$ with minimum distance

$$d_{min} = \min(2, \sqrt{d}), \tag{3.21}$$

from an appropriate linear code $C(n, k)$ of minimum Hamming distance d. In this chapter, design of transmission schemes based on Construction A consists in the following two steps:

(i) Choose N_a^* binary vectors $\mathbf{a}_1, \ldots, \mathbf{a}_{N_a^*}$ inside the Hamming ball centered at the origin $\mathbf{0}$ and of radius d. These vectors satisfy

$$dH(\mathbf{a}_i, \mathbf{c}) \leq d, \;\; \forall (i, \mathbf{c}) \in \{1, \ldots, N_a^*\} \times C(n, k),$$

where dH denotes the Hamming distance.

(ii) Map these vectors to N_a^* minimum-norm points $\mathbf{c}_1, \ldots, \mathbf{c}_{N_a^*}$ located inside $\mathcal{V}(\Lambda)$ by $\mathbf{c}_i = \mathbf{a}_i + 2\mathbf{z}, \mathbf{z} \in \mathbb{Z}^n$ and set the codebook $\mathcal{C}_m$ as $\mathcal{C}_m = \{\mathbf{c}_1, \ldots, \mathbf{c}_{N_a^*}\}$.

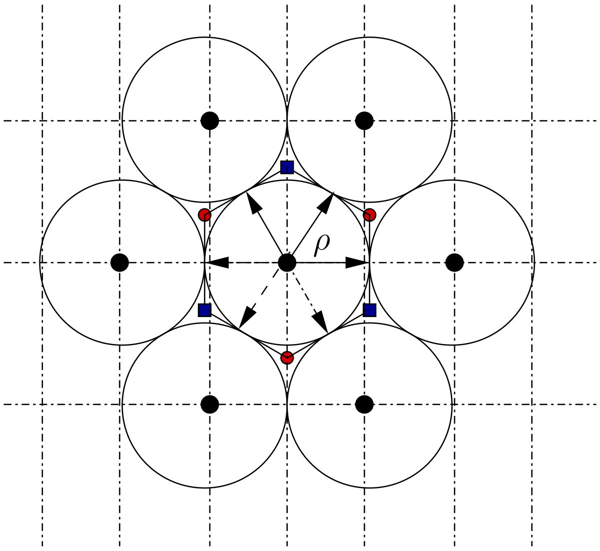
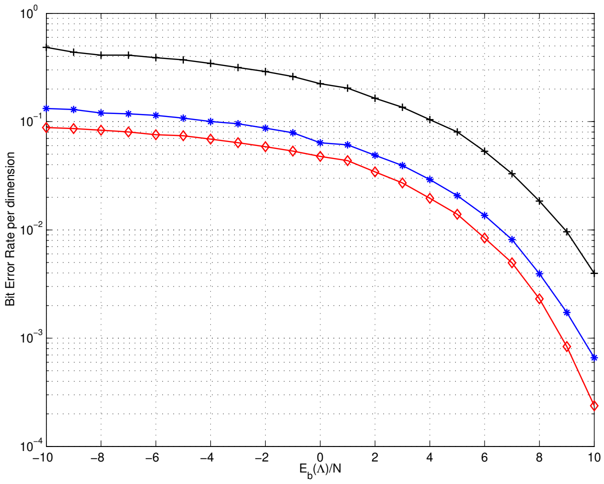


Figure 3.4: The hexagonal lattice A_2 in the plane. Lattice points are the centers of circles (of radius ρ). Deep holes- located at a distance r_{cov} from A_2- are indicated by (small) blue squares and red circles $(N_h, {N_h}^*) = (6, 2)$. Kissing points are pointed out with solid and dashed arrows $(K(\Lambda), {N_k}^*) = (6, 3)$.

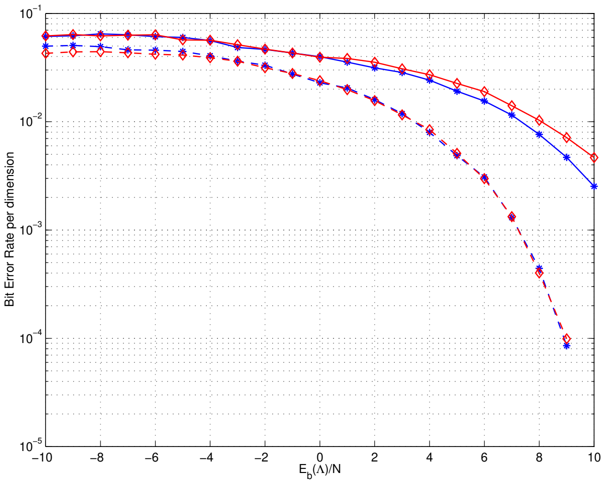
As we mentioned before, the use of Construction A enables transmission at larger rates R, with comparison to deep holes and kissing points. However the minimum distance d_{min}, as for it, is in general smaller, since it must satisfy (3.21). Several lattices with good packing and quantizing properties can be obtained with Construction A. For example, the lattice E_7 and the Gosset lattice E_8, which are the densest lattices in dimensions 7 and 8 respectively, can be obtained as $E_7 = (7, 3, 4) + 2\mathbb{Z}^7$ and $E_8 = (8, 4, 4) + 2\mathbb{Z}^8$. The binary linear code $(7, 3, 4)$ is the dual of the Hamming code $(7, 4, 3)$ and $(8, 4, 4)$ is the first order Reed-Muller code of length 8. Note that the goodness of these lattices inherits from that of the linear codes $(7, 3, 4)$ and $(8, 4, 4)$, which perform efficient error correction in their respective dimensions. The design of a good linear code to be used as a baseline for Construction A will be further discussed in Section 3.3.

3.2.3.2 Simulations

In the following, we provide Monte Carlo based simulation results corresponding to different choices of the codebook $\mathcal{C}_m$, taken from the examples discussed above. We retain deep holes and Construction A. Though $N_k^* \geq N_h^*$ for most of the lattices, deep holes are preferred to kissing points for low rate applications. Due to their large inter-cosets minimum distance ($r_{cov}(\Lambda) \geq \rho(\Lambda)$), the former are optimal for "Zero-Rate" embedding as remarkably noticed in [MGK04]. The latter, however, are more suitable for medium to high transmission rate applications. Also, the problem of resolving ties becoming too hard to solve when the dimensionality n of the lattice becomes large, we consider small dimensional lattices taken from Table 3.1. As a toy example, note that $(N_k^*, N_h^*) = (2, 1)$ for the square lattice $\mathbb{Z}^2$ and $(N_k^*, N_h^*) = (3, 2)$ for the hexagonal lattice A_2. The BER curves provided in the rest of this chapter correspond to embedding one symbol (index) per host sample. Fig.3.5 shows the per-dimension bit error probability[3] obtained with the coset leaders c_m, $m \in \{1, \ldots, N_h^* + 1\}$, mapped to the *relevant* deep holes of the lattices $\mathbb{Z}$, A_2 and D_4. The comparison of the different lattice-based transmissions is



(a) The overall gain brought by lattice-embedding



(b) The gain component due to the reduction in $G(\Lambda)$

Figure 3.5: Bit Error Probability v.s. the (per dimension per bit) Signal-to-Noise ratio $E_b(\Lambda)/N$ for DC-QIM based information embedding. (a) The curves correspond to use of the relevant deep holes of the lattices $\mathbb{Z}$ (plus sign), A_2 (asterisk) and D_4 (diamond). (b) BER using lattices E_7 (diamond) and E_8 (asterisk). The codebook $\mathcal{C}_m$ is obtained using relevant lattice holes (solid) and Construction A (dashed). The lattices E_7 and E_8 are obtained through construction A as $E_7 = (7,3,4) + 2\mathbb{Z}^7)$ and $E_8 = (8,4,4) + 2\mathbb{Z}^8)$. For deep holes, we considered the constructions $E_7 = A_7 \cup \left((-\frac{1}{2}^4, \frac{1}{2}^4) + A_7\right)$ and $E_8 = D_8 \cup \left(\frac{1}{2}^8 + D_8\right)$.

carried out as follows. In Fig.3.5(a), we are interested in comparing the error correction capability of the relevant holes of the different lattices. In Fig.3.5(b), we are interested in illustrating the trade-off between the transmission rate R and the bit-error probability P_e, by comparing Construction A to deep

[3]This per-dimension bit error probability can be approximated by the symbol error rate divided by $n \log_2(N_h^* + 1)$, for high signal-to-noise ratio.

holes. However, since the number of *relevant* deep holes (and thereby the transmission rate) as well as the per-dimension energy used for this transmission (given by $G(\Lambda)V(\Lambda)^{2/n}$) vary from lattice to lattice, a fair comparison of these lattices should assume the same energy used to transmit one bit of information per-dimension, in both Fig.3.5(a) and Fig.3.5(b) . Lattices must then be scaled accordingly. This (per-bit per-dimension) energy is given by

$$E_b(\Lambda) = \frac{G(\Lambda)V(\Lambda)^{2/n}}{\frac{1}{n}\log_2(M)}, \tag{3.22}$$

where $M = N_h^* + 1$ for simulations using relevant deep holes and $M = N_a^*$ for simulations using Construction A. In Fig.3.5(a), we use relevant deep holes of the lattices $\mathbb{Z}$, A_2 and D_4. We observe that the hexagonal lattice A_2 already provides significant improvements (bit error reductions) over the baseline cubic lattice $\mathbb{Z}^n$. Further enhancement is allowed by the use of the Checkerboard lattice D_4, which provides a gain of approximately 2.5 dB over the integer lattice (equivalent to SCS) at a bit error rate of about $P_e = 4 \times 10^{-3}$ bit per dimension. In the following (see section 6.5), it will be argued that this gain has two components: (i) a first component measuring the strength of the lattice holes as specific channel codewords and, (ii) a second component *intrinsic* to the lattice itself, not to the specific design of the channel codebook $\mathcal{C}_m$. *Intrinsic* refers to the reduction in the error probability due to the reduction in the second moment $G(\Lambda)$ of the lattice. The curves depicted in Fig.3.5(b) correspond to the use of the lattices E_7 and E_8. For each of them, two different transmission schemes are compared: (i) low rate transmission using the relevant deep holes of $E_7 = A_7 \cup \left((-\frac{1}{2}^4, \frac{1}{2}^4) + A_7\right)$ and $E_8 = D_8 \cup \left(\frac{1}{2}^8 + D_8\right)$ and (ii) high rate transmission using Construction A. In (ii) the lattices E_7 and E_8 are obtained from the Hamming code $(7,4,3)$ and the Reed-Muller code $(8,4,4)$ as mentioned before. The codewords $\mathbf{c}_m$ are chosen among the rows of the generator matrices of the binary linear codes $(7,3,4)$ and $(8,4,4)$. Geometrically, they correspond to the vertices of the quarter positive part of the unit cube at the origin. We observe that at $P_e \approx 3 \times 10^{-3}$ bit/dimension, the use of Construction A provides a gain of about 4 dB over deep holes. This is due to the fact that deep holes are optimal for the transmission of little information, only. Construction A however allow transmission at higher rates, more reliably. In addition, we observe from Fig.3.5(b) that the gain of Construction A over deep holes increases with the per-bit per-dimension SNR. This is due to the fact that at high SNR, channel noise is not strong enough to cause the transmitted signal to change from one coset to another. Hence, the need to have the different cosets far away from each other, something for which relevant deep holes are best suited, is of less importance.

3.2.4 Shaping for lattice bounded codebook

The problem of codebook selection partially addressed in Section 3.2.3 can be addressed through a more general approach. Let $\mathcal{R}$ denote the region from which the codebook elements $\mathbf{c}_m, m \in \mathcal{M}$, should be selected. $\mathcal{R}$ may be viewed as a support region (see definition in Section 3.1) for the union of all cosets $\Lambda_1 = \cup \Lambda_m$. Finding the optimal lattice codebook $\mathcal{C}_m$ amounts to finding the best support region $\mathcal{R}(\Lambda_1)$ since $\mathcal{C}_m = \Lambda_1 \cap \mathcal{R}$. Note that, the deep holes and kissing points considered above correspond to the particular choice of coset leaders as specific points from the set of points lying on the surface of the corresponding support regions. These support regions are given by the fundamental Voronoi region $\mathcal{V}(\Lambda)$ of the lattice for deep holes and the biggest ball centered at the origin and contained in $\mathcal{V}(\Lambda)$ for kissing

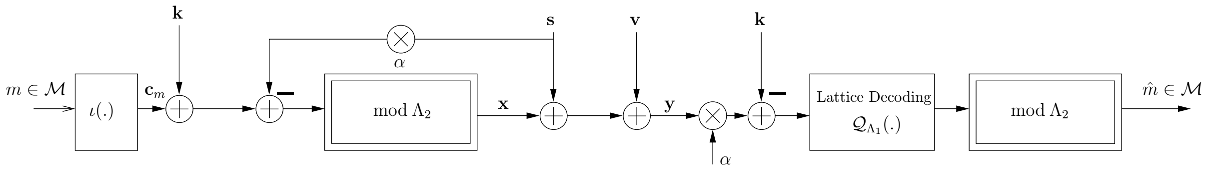


Figure 3.6: Nested encoding and decoding scheme for watermarking. The coarse lattice Λ_2 should be a good source-code and the fine lattice Λ_1 should be a good channel code.

points. The codebook $\mathcal{C}_m$ is bounded because the codewords are chosen inside the support region $\mathcal{R}$. In general Λ_1 does not need to have a lattice structure. However, the lattice structure makes modulo-reduction more feasible and it is preferred that $\Lambda_1 = \cup \Lambda_m$ be a lattice. In the examples considered above, this is the case. A slightly different approach consists in considering some (fine) lattice Λ_1 and finding the appropriate support region for it. An interesting choice for $\mathcal{R}$ is the Voronoi region of a larger scale *shaping lattice* Λ_2, i.e., $\mathcal{R} = \mathcal{V}(\Lambda_2)$. In this case the codebook $\mathcal{C}_m$ is a *Voronoi lattice codebook*[4] and the inter-cosets minimum distance (3.16) reduces to that of the lattice Λ_1. Hence the error probability (3.18) can be written as

$$\begin{aligned} P_e &\approx \Phi\left(\sqrt{\gamma_c(\Lambda_1)\frac{V(\Lambda_2)^{2/n}}{4N}}\right), \\ &\approx \Phi\left(\sqrt{3\gamma_c(\Lambda_1)\gamma_s(\Lambda_2)R\frac{E_b(\Lambda)}{N}}\right), \end{aligned} \tag{3.23}$$

where $\gamma_c(\Lambda_1) = \frac{d^2_{min}}{V(\Lambda_1)^{2/n}}$ is the packing (coding) gain of the lattice Λ_1, $\gamma_s(\Lambda_2) \triangleq \gamma_s(\mathcal{R})$ is the shaping gain of the support region $\mathcal{R}$ and $E_b(\Lambda)$ is the energy (per-dimension) needed to transmit one bit. From (3.23) we see that the total improvement (reduction) in the error probability is measured by the product $\gamma(\Lambda) = \gamma_c(\Lambda_1)\gamma_s(\Lambda_2)$. Thus, the lattice bounded codebook $\mathcal{C}_m = \Lambda_1 \cap \mathcal{V}(\Lambda_2)$ achieves about the same error probability as a scalar codebook at an SNR that is smaller by a factor of $\gamma(\Lambda)$. Equivalently, the rate $R = \frac{1}{n}\log_2|\mathcal{C}_m|$ can be increased by a factor of $\frac{1}{2}\log_2\gamma(\Lambda)$ at the same SNR, without increasing P_e. The strength of $\mathcal{C}_m$ is measured by the efficiency of Λ_1 in packing and that of Λ_2 in shaping.

In the following section, we give an interpretation of the shaping gain $\gamma_s(\Lambda_2)$ in terms of source-coding and argue that Costa-based lattice watermarking problem may be viewed as a Source-Channel coding problem. We also (partially) address the difficult problem of designing efficient practical codes.

3.3 Joint source-channel coding through nested-lattices

Consider the channel depicted in Fig.3.6 where two nested lattices (fine, Λ_1 and coarse, Λ_2) replace the single lattice in Fig.3.1. The lattices Λ_1 and Λ_2 are nested in the sense given in Section 3.1.2, with nesting ratio $\mu(\Lambda_1, \Lambda_2) = \sqrt[n]{V(\Lambda_2)/V(\Lambda_1)}$ and transmission rate $R = \frac{1}{n}\log_2\mu(\Lambda_1, \Lambda_2)$. The codebook $\mathcal{C}_m$ is given by $\mathcal{C}_m = \{\Lambda_1 \bmod \Lambda_2\} \triangleq \Lambda_1 \cap \mathcal{V}(\Lambda_2)$. For each $\mathbf{c}_m \in \mathcal{C}_m$, $\Lambda_2^{\mathbf{c}_m} = \mathbf{c}_m + \Lambda_2$ is a coset of Λ_2 relative to Λ_1. The indexes $m \in \mathcal{M}$ to be transmitted are arbitrarily associated by a one-to-one mapping function

[4] This denomination was first introduced by Forney in the context of lattice shaping for multidimensional constellations [GDF89].

$\iota(.)$ to the cosets $\mathbf{c}_m$ in the set $\mathcal{C}_m$. The union of these cosets forms the fine lattice Λ_1. According to [ZSE02], encoding and decoding for the channel in Fig.3.6 are given by

$$\mathbf{x}(\mathbf{s}; m, \Lambda_1, \Lambda_2) = (\mathbf{c}_m + \mathbf{k} - \alpha\mathbf{s}) \bmod \Lambda_2, \tag{3.24a}$$

$$\hat{\mathbf{c}_m}(\mathbf{y}; m, \Lambda_1, \Lambda_2) = \mathcal{Q}_{\Lambda_1}(\alpha\mathbf{y} - \mathbf{k}) \bmod \Lambda_2. \tag{3.24b}$$

Note that the transmitted signal $\mathbf{X}$ satisfies the average power constraint (3.4) and that the overall process resembles the one in (3.3), with the coarse lattice Λ_2 playing the role of the lattice Λ in (3.3). Besides, (3.24a) means that the transmitted signal is the error quantization between $\alpha\mathbf{s} - \mathbf{k}$ and the selected coset $\Lambda_2^{\mathbf{c}_m}$. Equation (3.24b), as for it, means that the overall decoding is performed through successive (layered) decoding: first use the fine lattice Λ_1 to find the quantizer representative $\mathcal{Q}_{\Lambda_1}(\alpha\mathbf{y}-\mathbf{k})$ of $\alpha\mathbf{y} - \mathbf{k}$. Next use the coarse lattice Λ_2 to quantize $\mathcal{Q}_{\Lambda_1}(\alpha\mathbf{y} - \mathbf{k})$ and reconstruct the message as the index of the unique coset containing $\mathcal{Q}_{\Lambda_1}(\alpha\mathbf{y} - \mathbf{k})$.

3.3.1 Performance

Noticing that the leader of the unique coset containing $\mathcal{Q}_{\Lambda_1}(\alpha\mathbf{y}-\mathbf{k})$ can be computed as $\hat{\mathbf{c}_m} = \mathcal{Q}_{\Lambda_1}(\mathbf{y}')\text{mod}\Lambda_2$, with $\mathbf{y}' = (\mathbf{c}_m + \alpha\mathbf{v} - (1-\alpha)\mathbf{x}) \bmod \Lambda_2$, we can easily show -in a straightforward manner to that in Section 3.2.2- that capacity is given by

$$C(\Lambda_1/\Lambda_2) = \max_{\alpha} \frac{1}{n}\left(\log_2(V_2) - h(\tilde{\mathbf{V}})\right) < \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right), \tag{3.25}$$

where the folded noise $\tilde{\mathbf{v}} = (\alpha\mathbf{v} - (1-\alpha)\mathbf{x}) \bmod \Lambda_2$ has, as we mentioned before, two components: a weighted Gaussian noise component $\alpha\mathbf{v}$ and a self-noise component $(1-\alpha)\mathbf{x}$.
Concentrate now on the error probability P_e. Here, this error probability is addressed with qualitative arguments in the case of finite dimension embedding. A more involved development about error exponents and asymptotic performance can be found in [LMK04]. From (3.24b) we see that decoding fails if either (i) reduction modulo Λ_1 fails to find the appropriate quantizer representative or (ii) reduction modulo Λ_2 fails to find the appropriate index. We denote the error probability related to the event (i) by $P_e^{(1)}$ and that related to the event (ii) by $P_e^{(2)}$. $P_e^{(1)}$ is measured by the coding gain $\gamma_c(\Lambda_1)$ of the fine lattice Λ_1 whereas $P_e^{(2)}$ is measured by the shaping gain $\gamma_s(\Lambda_2)$ of the coarse lattice Λ_2. Thus, minimizing P_e would require minimizing the probability of these two *error events*. Hence the problem amounts to finding lattices (or codes) with good shaping and packing (coding) properties. Two special regimes can be distinguished:

(i) At high coding rates (high nesting ratios), the self-noise tends to zero and the total equivalent noise becomes Gaussian. In this case, $P_e^{(2)}$ is negligible and the error probability P_e is principally determined by the channel code strength, i.e. the coding gain $\gamma_c(\Lambda_1)$ expressing the channel correction capability of the fine lattice Λ_1. This explains why such choices as Hamming, $(7, 4, 3)$ and Reed-Muller $(8, 4, 4)$ linear codes are quite efficient as basis for building fine lattices in Fig.3.5(b). This is also inline with the remark made in Section 3.2.3 regarding the non-influence of the second moment reduction (connected to the shaping gain through $\gamma_s(\Lambda) = 1/12G(\Lambda)$ at high SNR.

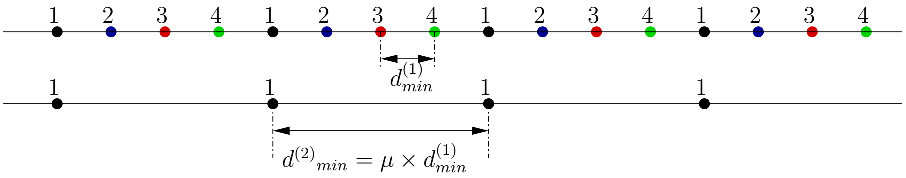


Figure 3.7: Algebraic binning based on 1-D lattice codes. The channel codewords represented by the set of all the spheres are divided into subsets or cosets (spheres indicated by the same number) of a single source code with larger minimum distance $d^{(2)}_{min}$ represented by the spheres indicated by the number 1. The minimum distances $d^{(1)}_{min}$ and $d^{(2)}_{min}$ characterize the channel and source codes strengths, respectively.

(ii) At low coding rates (low nesting ratios) however, the self-noise becomes a significant component of the total noise $\tilde{\mathbf{v}}$. The decoding error probability $P_e^{(2)}$ cannot be neglected and the error probability P_e is determined by both coding and shaping properties. Also in this case a good approximation for P_e is given by (3.23). This explains why lattice holes are appropriate for such rate range as previously mentioned.

3.3.2 Source-Channel coding in lattice watermarking

From a strict functional viewpoint, the watermarking problem depicted in Fig.3.6 is primarily a channel-coding problem, that is, for transmitting messages. However, the "power constraint" of the input of the communication channel is the quantization error of the side information. Hence from this point of view, side information $\mathbf{S}$ necessitates a good source coding in order to satisfy efficiently this power constraint. In other words, the encoding process (3.24a) satisfying the power constraint $\mathbb{E}_{\mathbf{X}}[\mathbf{X}^2] = P$ is basically a source coding problem. The only minor difference with respect to classic source coding quantization is that quantization is message-based (through a binning scheme). In addition, given that the power constraint P is equal to the norm of the quantization error of the side information $\mathbf{s}$, a good quantizer would be one that, for the same transmission rate R, minimizes this quantization error (thus allowing more information at the channel input for the same input power). So, in the watermarking problem shown in Fig.3.6, and broadly in the more general "Costa problem", source coding is used to design channel codewords that have the appropriate energy at the input of the channel. This is ensured by grouping channel codewords into (appropriate) cosets of (appropriate) source codes.

3.3.2.1 Binning interpretation

The basic concept of combined source-channel coding in lattice-based watermarking is inherently implicit in the original random binning coding argument for channels with state information [GP80]. "Binning" consists in randomly dividing the codebook entries into subsets (cosets or bins) such that the codewords are far apart as possible. Hence, the set of codewords in all cosets may be viewed as a set of channel codewords. The efficiency of this channel code is measured, for example, by its minimum distance. Large minimum distances are preferred. Moreover, to transmit a message $m \in \mathcal{M}$, a codeword that is D_E-distortion jointly typical with the state information $\mathbf{S}$ has to be found. This can be viewed as quantizing $\mathbf{s}$ to the nearest codeword in the bin identified by m. The set of codewords collapsed into the same bin m

may then be viewed as a set of source codewords. The efficiency of this source code is measured by the distortion introduced in quantizing $\mathbf{s}$. If a linear code is used for source coding, this distortion translates to its minimum distance. Small minimum distances are preferred. Thus the channel coding problem of Costa-based watermarking can also be understood as a source-channel coding problem, when considering that the watermark signal is obtained through message-depending quantization. In addition, the ratio of the minimum distance of the source code to that of the channel code has the significance of "*nesting ratio*" which determines the transmission rate R as in (3.1). An illustration of this principle based on one-dimensional scalar lattice codes is shown in Fig.3.7.

3.3.2.2 Nested lattices and source-channel coding

So far we have shown that the performance of nested lattices-based watermarking depend on channel coding properties, which are both the coding (packing) gain $\gamma_c(\Lambda_1)$ and the shaping gain $\gamma_s(\Lambda_2)$. We also argued that the watermarking problem can also be viewed as a source-channel coding problem. Moreover, similarly to the coding gain and the shaping gain in channel coding, source coding is characterized by granular gain and boundary gain [EF93]. Since the source code is nested inside the channel code (see the binning interpretation above), we may use the granular gain of the source code for "shaping" the channel code[5]. More precisely, in the nested lattices of Fig.3.6, the fine lattice Λ_1 should be used for channel coding and the coarse lattice Λ_2 should be used for source coding. This amounts to *shaping* Λ_1 by (the Voronoi region of) Λ_2 whose granular gain $\gamma_g(\Lambda_2) = 1/12G(\Lambda_2)$ translates to a shaping gain for Λ_1. The source and channel codes used for algebraic binning illustration in Fig.3.7 can be viewed as two 1-D nested (scalar) lattices. The coding lattice Λ_1 and the shaping lattice Λ_2 are both scaled versions of the one-dimensional integer lattice $\mathbb{Z}$. Namely, $\Lambda_1 = \alpha\mathbb{Z}$ and $\Lambda_2 = \alpha M\mathbb{Z}$ with $M = \lfloor d_{min}^{(2)}/d_{min}^{(1)} \rfloor$. The coding cells are the translates of the interval $(-\frac{\alpha}{2}, \frac{\alpha}{2}]$, which is the fundamental Voronoi region of Λ_1. The support region $\mathcal{R}$ is the interval $(-\frac{M\alpha}{2}, +\frac{M\alpha}{2}]$, which is the fundamental Voronoi region of Λ_2.

In this section, we first show, through simulations, that in finite dimensional embedding the two components of the total gain $\gamma(\Lambda)$ (i.e., the coding gain $\gamma_c(\Lambda)$ and the shaping gain $\gamma_s(\Lambda)$) are not decoupled but rather interacting, by opposition to the asymptotic case (in dimension n) where $\gamma(\Lambda) = \gamma_c(\Lambda) + \gamma_s(\Lambda)$ [dB] as it can be seen from (3.23). Next, we compare the importance of the two components in lowering the error probability. This gives insights into the design of good source-channel codes addressed in Section 3.3.3.

In Fig.3.9(a), the probability of error P_e is measured as in Fig.3.5(a), using the same codebooks too. However, we are interested, this time, in extracting the reduction in P_e due to the shaping gain γ_s only. Since $\gamma_s(\Lambda)$ depends only $G(\Lambda)$, and since $E_b(\Lambda)$ writes as in (3.22), the improvement due to the shaping gain $\gamma_s(\Lambda)$ can be brought out by scaling the lattices so as to have the same volume (e.g., that, $V(\mathbb{Z})$, of the lattice $\mathbb{Z}^n$). Reduction modulo a scaled lattice $\beta\Lambda$ where $\beta \in \mathbb{R}$ is some scale factor is such that

$$\mathcal{Q}_{\beta\Lambda}(\mathbf{x}) = \beta\mathcal{Q}_{\Lambda}(\frac{\mathbf{x}}{\beta}),\ \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{3.26}$$

We observe that if the cubic lattice $\mathbb{Z}^n$ is replaced by an hexagonal lattice A_2 with the same volume, but with smaller normalized second moment $G(A_2) = 5/36\sqrt{3} < G(\mathbb{Z}^n) = 1/12$, the resulting error

[5]It should be mentioned that the shaping gain and the granular gain of a lattice are given by the same formula $1/12G(\Lambda)$.

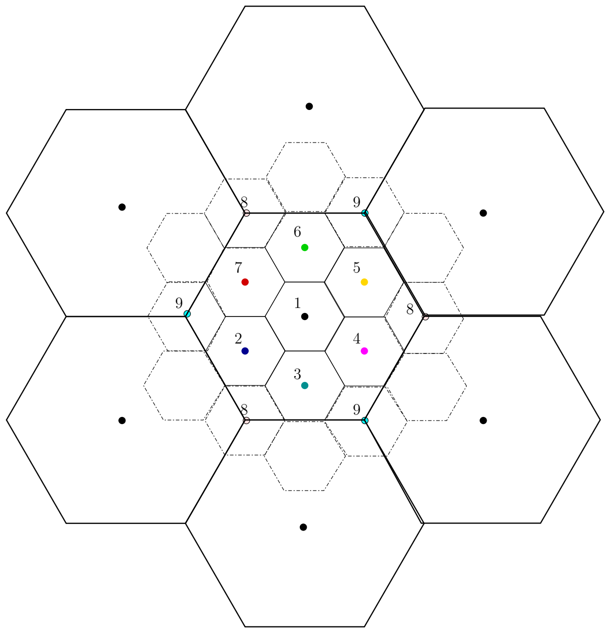


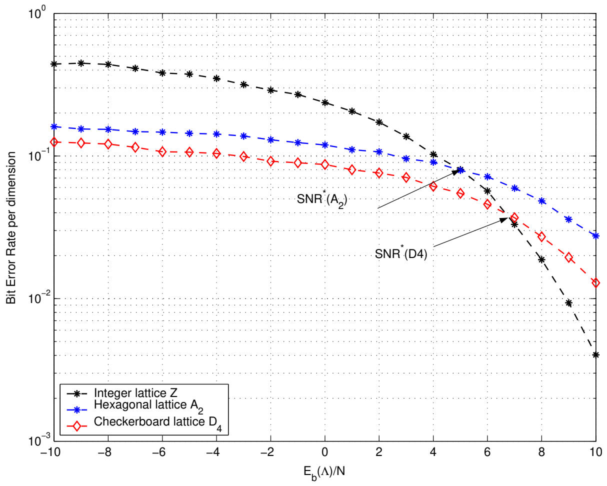
Figure 3.8: Shaping through nested lattices. The fine lattice code corresponds to the codewords represented by the centers of all small hexagonal cells. The coarse lattice includes only the codewords represented by the centers of big cells. All codewords having the same number correspond to a single coset code of this coarse lattice code.

probability is significantly reduced, at low SNR. Even lower BERs are obtained with the Checkerboard lattice D_4 for which $G(D_4) = 13/120\sqrt{2}$. Because of its large normalized second moment, the cubic lattice $\mathbb{Z}^n$ suffers performance loss mainly at low SNR. The reduction in the normalized second moment $G(\Lambda)$ translates to a reduction in the SNR by a factor equal to $1/12G(\Lambda)$. However, these gains (BER reductions) are visible only for the SNR range below some lattice-dependent SNR threshold $\text{SNR}^*(\Lambda)$. Upon this threshold, the improvement brought by the reduction in the second moment of the lattice does not counterbalance the loss caused by the decrease in the inter-cosets minimum distance d_{min} (at fixed volume). This is because, when scaled (with some scale factor $\beta(\Lambda)$) so as to have the same per-dimension volume $V(\mathbb{Z})$, a lattice Λ sees its inter-coset minimum distance d_{min} scaled accordingly. Namely, d_{min} decreases to βd_{min}, where

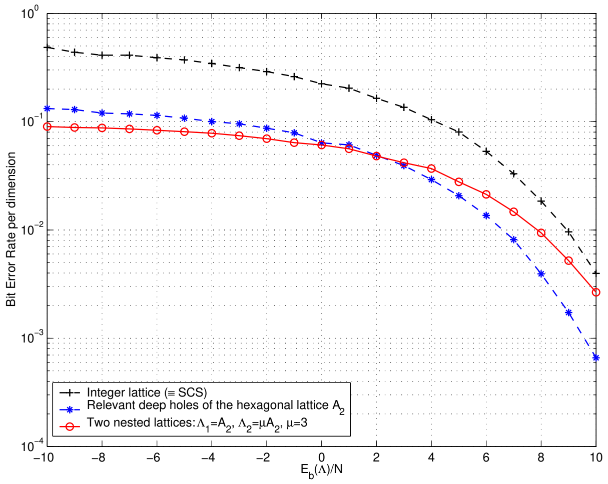
$$\beta = \frac{V(\mathbb{Z})}{V(\Lambda)^{1/n}}. \tag{3.27}$$

As $n \longrightarrow +\infty$, this loss in minimum distance (and thereby in coding gain) goes to zero. Thus, shaping and coding gains $\gamma_s(\Lambda)$ and $\gamma_c(\Lambda)$ are not decoupled but rather interacting. It is precisely this interaction that explains why the threshold SNR^* moves to the right as the shaping gain increases, in Fig.3.9(a).

Another important observation concerns the importance of shaping, depending on the SNR. More precisely, note the following. With respect to the simulations corresponding to the results shown in Fig.3.5(a), what we have done to get the curves in Fig.3.9(a) is nothing but fully exploiting the shaping gain (and hence, reducing the coding gain γ_c thereby). The resulting error probability is enlarged, mainly at high SNR, meaning that we get poor performance by diminishing the gain coding component. Thus, at high SNR, it is the coding gain $\gamma_c(\Lambda)$ that best determines the BER. This phenomenon can also be observed from Fig.3.9(b) where the BER corresponding to the 2-D nested hexagonal lattices in Fig.3.8



(a) The (per-dimension) volume $V(\Lambda)^{2/n}$ is kept fixed



(b) Shaping through nesting of lattices

Figure 3.9: Interaction Shaping/coding. (a) At low SNR, reduction in BER (with comparison to the integer lattice $\mathbb{Z}$) is due to the increase in the shaping gain $\gamma_s(\Lambda)$. At high SNR, the increase in shaping $\gamma_s(\Lambda)$ does not encompass the decrease in the coding gain $\gamma_c(\Lambda)$ caused by the decrease in the minimum distance d_{min}. (b) The effect of shaping through nesting of lattices (coarse $\Lambda_2 = 3A_2$, fine $\Lambda_1 = A_2$) is observed at low SNR. At high SNR reduction in error probability is principally determined by the coding gain $\gamma_c(\Lambda)$.

with $\mu = 3$ is depicted. At low SNR shaping plays an important role. This explains why shaping with the larger scale coarse lattice A_2 reduces the BER for SNR < 1 dB. At high SNR however, it is the coding gain $\gamma_c(\Lambda)$ that matters.

3.3.3 Practical design of good nested codes

So far we considered the use of Construction A for high rate information embedding. However the efficiency of a lattice issued from construction A naturally depends on that of the linear code used for the construction. Hence, a method for designing good channel codes is needed. In [ZSE02], Zamir et al.

proposed nested lattices as means of achieving capacity for efficient structured binning multi-terminal coding and tune the fine lattice to be *Poltyrev-good* and the coarse lattice to be at the same time *Poltyrev-good* and *Roger-good* [EZ04]. Nesting of good lattices has also been recently addressed in the context of distributed source coding for sensor networks as in [XLC04, XSC$^+$04, Ser04, CPR03]. However, very often performance are studied asymptotically, that is as the dimension of lattices goes to infinity. Thus, the resulting criteria are convenient only for theoretical analysis, not for practical implementations. Here, we use a less stringent, but more feasible approach. Namely, since shaping is important at low SNR only as explained in Section 6.5.3, we use a cubic lattice $\Lambda_2 = \mathbb{Z}^n$ as coarse lattice. With respect to the full shaping that would be obtained, asymptotically, with infinite dimensional spheres, this leaves only $\frac{1}{2}\log_2(\frac{\pi e}{6}) \approx 0.255$ bit per dimension unexploited. Also, we ask the fine code to be "good" enough in a minimum-distance sense.

3.3.3.1 RS codes and minimum distance criterion

The use of the *minimum distance* criterion is motivated by the fact that at high SNRs, the performance of a channel code depend almost only on its *minimum distance*. The remaining weight distribution does not much matter. So, we proceed as follows.

(a) Select a good fine code $\mathcal{C}_1$ according to the *minimum distance* criterion.

(b) Use Construction A[CS88] to build the corresponding fine lattice Λ_1 as described in Section 3.2.3.

(c) Finally, use a cubic lattice, as coarse lattice Λ_2, to *shape* for the fine lattice Λ_1.

An important class of codes having good (large) *minimum-distances* is that of Reed-Solomon (RS) codes [Pro01]. An (RS) code RS(N, K, D), $N = 2^m - 1$, is Maximum-Distance-Separable (MDS), meaning that it has the largest minimum-distance among all codes having the same length N. For instance, RS codes attain the singleton bound [Pro01], i.e. $D = N - K + 1$. However, an RS code being defined over a Galois-Field $GF(q)$ with $q = 2^m$, an equivalent binary representation (over $GF(2)$) should be found so as to make it possible to use it in conjunction with construction A, in building the fine lattice Λ_1. The RS code RS(N, K, D) over $GF(q)$ translates to the binary code $\mathcal{C}(n, k, d)$=$\mathcal{C}(mN, mK, d)$. Thus, (binary versions of) RS codes are good candidates for the fine lattice construction in the nested structure addressed above.

3.3.3.2 Example

We consider the RS code RS$(7, 5, 3)$ over $GF(8)$. We use the corresponding binary code $\mathcal{C}(21, 15)$ to build the fine lattice, by construction A. The way this binary code $\mathcal{C}(21, 15)$ is obtained from the RS code RS$(7, 5, 3)$ is undertaken below. We implemented a soft decision decoder based on the Euclidean distance. A sketch of the overall process is as follows. Lattices are constructed as $\Lambda_1 = \mathcal{C}(21, 15) + 4\mathbb{Z}^{21}$ and $\Lambda_2 = \mathbb{Z}^{21}$. The message to transmit is chosen from the alphabet $\mathcal{M} = \{1, \cdots, 16\}$. The codebook $\mathcal{C}_m$ is chosen such that the message $m \in \mathcal{M}$ is associated to the mth row vector $\mathbf{c}_m$ of the binary generator matrix G_{bin} ($m = 16$ is mapped to the zero vector, $\mathbf{0}$). Encoding and decoding functions are as follows.

Encoder: Given some index $m \in \{1, \cdots, 16\}$ to transmit, the encoder forms $\mathbf{x}$ as in (3.24a), (the key $\mathbf{k}$ is chosen to be zero).

Decoder: Given some received sequence $\mathbf{y}$, the decoder has to perform (3.24b). For that, he/she first searches for the closest point of the lattice Λ_1 to $\alpha\mathbf{y} - \mathbf{k}$. Since Λ_1 is obtained through construction A, this amounts to decoding the binary code $\mathcal{C}(21, 15)$ (see [CS88], Chapter 20, Section 5). The nearest codeword of $\mathcal{C}(21, 15)$ to some vector in the space $\mathbb{R}^{21}$ is the one that minimizes the Euclidean distance to that vector. Hence, an exhaustive research in the set of all possible $2^k = 2^{15}$ codewords is required. To obtain the list of these codewords, the binary generator matrix G_{bin} of the code $\mathcal{C}(21, 15)$ is needed. G_{bin} is constructed as the dual of the binary parity check matrix H_{bin}. H_{bin} is the binary representation of the parity check matrix H_q over $GF(8)$, given by

$$H_q = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 \\ 1 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha & \alpha^3 & \alpha^5 \end{pmatrix}.$$

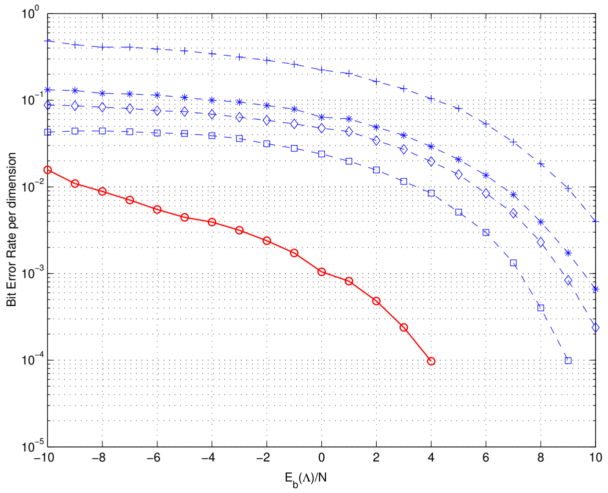


Figure 3.10: Bit Error Probability v.s. $E_b(\Lambda)/N$ for DC-QIM information embedding. Dashed: from bottom to top: lattices E_8 obtained from Construction A, D_4, A_2 and $\mathbb{Z}$. Solid: using the RS code $(7, 5, 3)$ for the design of the fine lattice with Construction A.

In Fig.3.10, the per-dimension bit error probability reduction that results from the use of RS$(7, 5, 3)$ in the construction of the fine lattice Λ_1 is compared to that using the Gosset lattice E_8 obtained from Construction A and also transmission with deep holes of lattices $\mathbb{Z}$, A_2 and D_4. We observe that the gain is particularly significant for low to medium SNR but may diminish for very high SNRs. The reason is that the minimum distance of the fine lattice is bounded by (3.21). Note that in Fig.3.10, the nesting ratio is such that the transmission rate corresponding to the use of the RS code $(7, 5, 3)$ (or, more precisely, its binary version $(21, 15)$) is respectively 3.75 times and 5 times those enabled by the use of the lattice D_4 and A_2. Further, the use of the RS code $(7, 5, 3)$, would enable much reduction in error probability if one relaxes the transmission rate. Hence, this example shows that reliable transmission, along with relatively high payload, is made possible. In addition, following Zamir et al. construction of family of codes that are asymptotically "good", RS codes represent a good starting point for a class of asymptotically (in dimension n) good channel codes. These are called *Justesen codes*. Justesen codes [Jus72] are good in the sense of both *Gilbert-Varshamov* and *McEliece-Rodemich-Rumsey-Welch* bounds. These bounds characterize channel codes for which both R and d/n remain bounded away from zero as n increases. These bounds are reminded here for information.

(a) *Gilbert-Varshamov lower-bound*: Let $\delta \in [0, \frac{1}{2}[$. There exist linear codes $\mathcal{C}(n,k,d)$ over $GF(q)$ with minimum distance d and rate k/n such that $d/n \geq \delta$ and $k/n \geq 1 - H_q(\delta) - \delta \log_q(q-1)\;\forall\, n$

(b) *McEliece-Rodemich-Rumsey-Welch upper bound*: For each linear code $\mathcal{C}(n,k,d)$ of minimum distance d, the rate k/n is such that $k/n \leq H_2\left(\frac{1}{2} - \sqrt{\frac{d}{n}(1-\frac{d}{n})}\right)$ for n sufficiently large.

Note that the codes with both R and d/n bounded away from zero are (asymptotically) good candidates for the problem of trade-off between transmission rate and reliability raised in Section 3.2.3. A Justesen code $\mathcal{C}(2N, 2K)$ may be obtained from the RS code RS(N,K) as follows: let α be a primitive element of $GF(q)$, i.e. $\alpha^N = 1$. If $\mathbf{c} = (c_1, \ldots, c_N)$, $c_i \in GF(q)$ is an arbitrary codeword of $\mathcal{C}(N,K)$, $\mathbf{c}' = (c_1, c_1, c_2, \alpha c_2, \ldots, c_N, \alpha^{N-1} c_N)$ and $\mathbf{c}''$ the corresponding binary m-tuple, the set of all codewords $\mathbf{c}''$ for $\mathbf{c} \in \mathcal{C}(N,K)$ forms a Justesen code $\mathcal{C}(2nN, mK)$. The minimum distance of this Justesen code satisfies [CS88]

$$\frac{d}{n} \gtrsim 0.11(1-2R). \tag{3.28}$$

Note that, though obviously dependent on D, the minimum-distance d of the binary code $\mathcal{C}(n,k,d)$ is not explicitly related to D. A loss in the relative distance may occur when transforming an RS code into its binary representation. But in most of the cases, large minimum-distance D over $GF(q)$ leads to sufficiently large minimum-distance d over $GF(2)$.

3.3.3.3 Discussion

In the example above, we considered an RS code as a starting point for building the fine lattice Λ_1. This choice may be non-optimal, but it already shows the gain achieved when channel codewords (fine lattice points) are carefully designed. For instance, one possible weakness concerns the minimum distance d of the binary representation of the RS code. Though obviously dependent on D, this minimum-distance d is not explicitly related to D. Hence, a certain loss in the minimum distance may occur when transforming an RS code into its binary equivalent code. Also, the use of construction A may be non-optimal for very high embedding dimensions. In this case, other more efficient constructions (constructions C and D [CS88] for instance) may be used instead. The principle described here remains unchanged however, at the cost of relatively higher modulo-reduction complexity of course. Finally, note that the proposed RS-based coding scheme is computationally simple. The performance shown in Fig.3.10 can be further enhanced by, for example, efficiently coding the message m prior to encoding. Also, substantial gain would be possible by changing the RS soft decoder so that it tales into account the *ordered statistics* of the bits at its input as in [FL98]. More sophisticated linear/Trellis codes can be considered. Some of these have already been used for the dual problem of distributed source coding. For example Low-Density-Parity-Check (LDPC) codes have been considered in [YCXZ03, RMZG03]. Also Turbo-based constructions have been proposed as in [CPR03, AG02, ZGF02, BM01]. In the context of data transmission, an important work is that of

Erez and ten Brink [EtB04]. In this work, lattice strategies are used in conjunction with MMSE scaling in order to perform efficient precoding. To this end, the authors rely on vector quantization together with iterative decoding techniques. They showed that a 2 dB improvement over scalar quantization techniques is achieved. An interesting implementation of Erez's scheme has been proposed in [CPGW05] in the context of information embedding. The schemes in [EtB04, CPGW05] are however too complex. Here, our main goal is to point out the source and channel coding problem in information embedding, to emphasize the interacting property of two shaping and coding (packing) gains and to give insights -through an example- into the proper design of the involved codes. The resulting construction has the advantage of enabling low error rates at relatively high payloads, thus showing that the trade-off between error probability and transmission rate mentioned above may have good solutions.

3.4 Summary

In this chapter we focused on lattice-based information embedding techniques for data hiding. Relying on recent results on Modulo channels, the gain achieved over scalar approaches (SCS and QIM) is illustrated by use of some finite dimensional lattices, with good quantizing and packing properties. Then we addressed the problem of codebook selection first through some relevant examples using the appealing algebraic structure of the lattice and then, through a more general approach inspired by shaping for multidimensional constellations. In the second part of the chapter, we used a binning interpretation to argue that the watermarking problem can also be seen as a source-channel coding problem, with source coding providing means of *shaping* the channel code. Interestingly enough a nested structure turns to be particularly useful for good source and channel codes design. This problem, though already solved from a theoretical point of view, still suffers from lack of feasible implementations. Here we proposed a simple *minimum distance*-based approach for selecting the fine channel codes. We also emphasized the interaction between shaping and coding, thus identifying the situations where shaping through nesting of lattices is most important. Analysis is supported by an illustrative example showing that reliable high rate transmission is possible if source and channel codes are carefully designed. Both Monte Carlo based integration (for capacity) and simulation (for BER) are provided for illustrations.

Chapter 4

Broadcast and MAC Aware Coding Strategies for Multiple User Information Embedding

The content of this chapter has been partially published in [ZPD04, ZP06, ZPD07].

Multiple user information embedding is concerned with embedding several messages into the same host signal. While emphasizing the tight relationship with conventional multiple user information theory, this chapter presents several implementable "Dirty Paper Coding" (DPC) based schemes for multiple user information embedding. 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 multi-user channels with state information available at the transmitter, for which recent theory is well developed. The focus is on the physically degraded Gaussian Broadcast Channel (BC) [see Appendix B for a brief review of BC] and the Gaussian Multiple Access Channel (MAC) [see Appendix C for a brief review of MAC]. For each of these channels, two practically feasible transmission schemes are compared. The first approach consists in a straightforward- rather intuitive- superimposition of Dirty Paper Coding schemes. The second consists in a joint design of these Dirty Paper Coding schemes. The joint approach is based on the ideal DPC for the corresponding channel. These results extend the practical implementations QIM and SCS that have been originally conceived for one user to the multiple user case. 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 (BER) and capacity region curves are provided for both methods, illustrating the improvements brought by a joint design.

4.1 Multiple User Information Embedding: A Prelude

During the last years, both QIM and SCS have been thoroughly studied and extended into different directions such as non-Gaussian channel noise [TBF+05], non uniform quantizers [LS04a] and lattice codebooks [MK04, ZD05c, ZD05a, ZD06a]. This chapter extends these schemes to another direction: multiuser information embedding. Multiuser inforrmation 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 multiuser inforrmation 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 to find the set of rates at which the different messages can be simultaneously embedded. Interestingly enough, this problem has tight relationship to conventional multiple user information theory. Consider for example watermark applications such as copy control, transaction tracking, broadcast monitoring and temper 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 context. The design and the optimization of algorithms for multiple information embedding applications should then benefit from recent advances and new findings in network information theory. For instance, in this chapter, we first argue that many multiple information embedding situations can be modeled as communication over either a degraded Broadcast Channel (BC) with state information at the transmitter or a Multiple Access Channel (MAC) with state information at the transmitters. Next, we rely heavily on the general theoretical solutions for these channels to devise efficient practical encoding schemes for the problem of multiple user information embedding. The resulting schemes consist, in essence, of applying the initial QIM or SCS as many times as the number of the watermarks to be embedded. While this is not surprising given the close-to-optimal performance of both QIM and SCS in the single user case, we show in this chapter that these schemes should be appropriately designed in 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 by comparison to the straightforward, rather intuitive, corresponding scheme obtained by super-imposing (i.e with no joint design) these SCSs (or DPCs for the ideal coding). This improvement is demonstrated through both capacity region and BER analysis. We finally show that these performance can further be made

closer to the theoretical limits by considering lattice-based codebooks. Some finite-dimensional lattice with good packing and quantization properties are considered for illustrative purposes.
The rest of the chapter is organized as follows. Two mathematical models corresponding to the multiuser inforrmation embedding problem viewed either as communication over a Broadcast Channel (BC) or as communication over a Multiple Access Channel (MAC) are provided in Section 4.2. Performance analysis corresponding to these two models are addressed in Sections 4.3 and 4.4, respectively. For each of them, analysis is carried out within the context of two watermarks using scalar codebooks first, and then extended to the more general case of an arbitrary number of watermarks and that of high dimensional lattice-based codebooks. Finally we give some concluding remarks in Section 4.5.

4.2 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 information embedding system may require either a 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 composite content. This is due to the fact that this composite content could have experienced different channel degradations 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 the 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 becoming the transmitters and vice-versa. Of course, though intentionally kept in its very general form, this model may not include some specific multiuser information embedding situations. This is due to the difficulty of finding a single unifying approach to all possible multiple user information embedding situations. Nevertheless, the model that we described is sufficiently general to involve the most important multiuser inforrmation embedding scenarios. For instance two classes of such scenariis that we will recognize as being equivalent to communicating over a degraded BC and a MAC in Sections 4.2.1 and 4.2.2 respectively, are worthy of deep investigations. To simplify the exposition, we first restrict our attention to a two-watermark scenario. Extension to the general case then follows.

4.2.1 A mathematical model for BC-like multiuser information embedding

Consider a 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 S. 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, should be of lesser-robustness or even fragile. This means that the watermark $\mathbf{X}_2$ must survive channel degradations up to some level N_2 larger than the level $N1$ up to which the watermark $\mathbf{X}_1$ could survive, 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. This means that the composite watermark $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ is constrained to have power P, i.e $\mathbb{E}_{\mathbf{X}}[\mathbf{X}^2] = P$. Assuming independent watermarks[1] $\mathbf{X}_1$ 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 arbitarily chosen to trade off power between the two watermarks. In practice, this multiuser information embedding scenario can be used to serve multiple

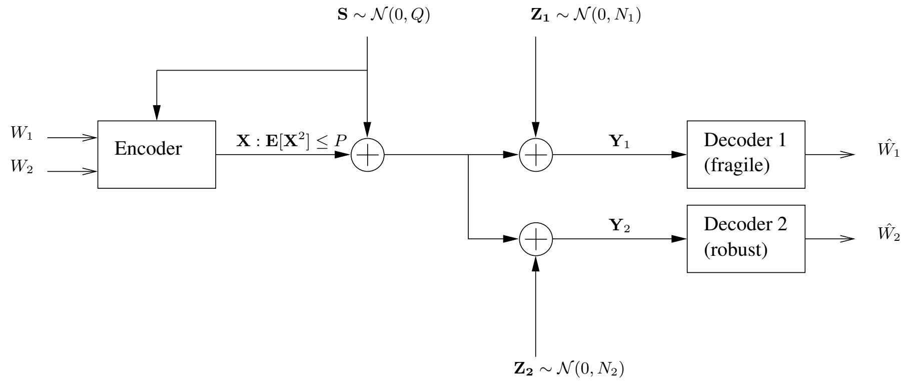


Figure 4.1: Two users information embedding viewed as communication over a two users Gaussian Degraded Broadcast Channel (GDBC).

purposes. In the scope of watermarking of medical images for example, we may wish 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. Data (think of software programs) should be watermarked so as it would be possible to reliably extract the information stored in different magnetic recording media, and hence having faced different alteration levels. The storage and the recording processes obviously introduce different alteration levels. Thus, the part of the data stored in the media with much alteration should be more robust than the remaining data. Of course many more 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.4.1. The 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 and declares an error if $\widehat{W_i} \neq W_i$. Functionally, this is the very transmission diagram of a two users Gaussian Degraded Broadcast Channel (GDBC) with state information available at the transmitter but not to 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

[1]A justification of this assumption will be given in Section 4.3.

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 degraded BC will be retained if, instead of just two watermarks, L watermarks are to be simultaneously embedded by the same so-called *trusted* authority.

4.2.2 A mathematical model for MAC-like multiuser information embedding

We now consider another watermarking situation. Again, the information embedding system aims at embedding two 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 and (ii) at the receiver, a single *trusted* authority having to check for the two watermarks. We assume no particular cooperation between the two embedding authorities, meaning that the watermarks $\mathbf{X}_1$ of power P_1 (carrying W_1) and $\mathbf{X}_2$ of power P_2 (carrying W_2) should be designed independently of each other. The composite watermark signal $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ must however satisfy -as before- the input-power constraint P, meaning that $P_1 + P_2 \leq P$. In practice, this multiuser informa-

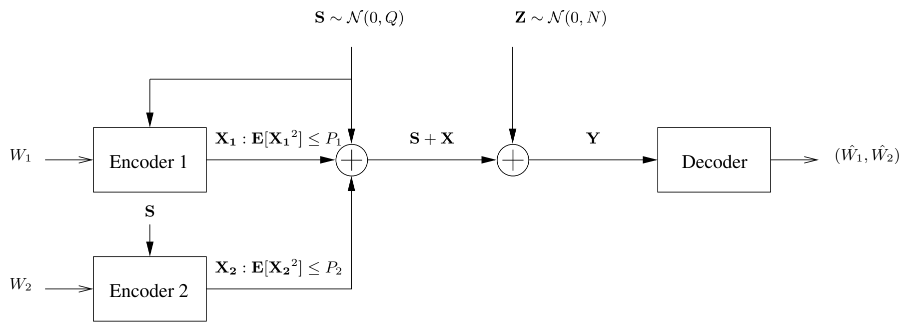


Figure 4.2: Two users information embedding viewed as communication over a (two users) Multiple Access Channel (MAC).

tion embedding scenario can be used to serve multiple purposes. Broadly speaking, every information embedding 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 watermarking 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 the two 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 [CW01]. 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 identification.

Assuming a Gaussian channel noise $\mathbf{Z} \sim \mathcal{N}(0, N)$ corrupting the composite signal $\mathbf{S} + \mathbf{X}$, a simplified diagram is shown in Fig.4.2. The encoder i, $i = 1, 2$, encodes W_i into $\mathbf{X}_i$ at rate R_i. The decoder outputs $(\widehat{W_1}, \widehat{W_2})$ and declares an error 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 multiuser information embedding and conventional multiple user communication. In Sections 4.3 and 4.4, we rely on recent findings in multi-user information theory to devise efficient implementable multiuser inforrmation embedding 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 composite content” and ”the receiver decoding the less noisy composite content”, respectively.

4.3 Watermarking over a Gaussian Broadcast Channel: Performance analysis

In this section, we are interested in designing efficient low-complexity multiuser information embedding schemes for the situation described in Section 4.2.1. We first present a straightforward rather intuitive method based on super-imposing two SCSs. This simple method can be thought as a broadcast-unaware strategy. Next, we use the similarity with a Gaussian degraded BC recognized above to design a more efficient multiuser inforrmation embedding scheme. The improvement brought by this broadcast-aware strategy is illustrated through both achievable capacity region and achievable Bit Error Rates (BER) enhancement. Finally, results are extended to both the L-watermark case and the high dimensional lattice-based codebooks case.

4.3.1 Broadcast-unaware coding for multiuser information embedding

A simple approach for designing a watermark system for the two users watermarking problem considered in Section 4.2.1 consists in using two independent single-user DPCs (or SCSs for the corresponding suboptimal practical implementation). In essence, the ideal coding is based on successive encoding at the transmitter as follows:

1. Use a first DPC (DPC1) taking into account the known state $\mathbf{S}$ and the unknown noise $\mathbf{Z}_2$ to form the most robust watermark $\mathbf{X}_2$ intended to the degraded user.

2. Use a second DPC (DPC2) 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 unknown noise $\mathbf{Z}_1$ to form the less robust watermark $\mathbf{X}_1$ intended to the better user.

3. 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.

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. Using (2.20), the ideal DPCs corresponding to the above steps are as follows.

1. Channel $\mathbf{Y}_2$ (DPC1): $\mathbf{X}_2 = \mathbf{U}_2 - \alpha_2\mathbf{S}$ where

$$\mathbf{U}_2 \sim \mathcal{N}\left(\alpha_2\mathbf{S}, (1-\gamma)P\right), \text{ with } \alpha_2 = \frac{(1-\gamma)P}{(1-\gamma)P + N_2}. \tag{4.1}$$

2. Channel $\mathbf{Y}_1$ (DPC2): $\mathbf{X}_1 = \mathbf{U}_1 - \alpha_1(\mathbf{S}+\mathbf{X}_2)$ where

$$\mathbf{U}_1 \sim \mathcal{N}\left(\alpha_1(\mathbf{S}+\mathbf{X}_2), \gamma P\right), \text{ with } \alpha_1 = \frac{\gamma P}{\gamma P + N_1}. \tag{4.2}$$

The theoretical rates R_1 and R_2 achievable by DPC1 and DPC2 are given by

$$R_1 = \frac{1}{2}\log_2\left(1 + \frac{\gamma P}{N_1}\right), \tag{4.3a}$$

$$R_2 = R(\alpha_2, (1-\gamma)P, Q, \gamma P + N_2), \tag{4.3b}$$

where $R(\alpha, P, Q, N) = \frac{1}{2}\log_2\left(\frac{P(P+Q+N)}{PQ(1-\alpha)^2 + N(P+\alpha^2 Q)}\right)$. Using (2.59) and following the way a single user SCS is derived from the theoretical single-user DPC, a suboptimal practical two-users scalar watermarking scheme can be derived by independently superimposing two SCSs (denoted by SCS1 and SCS2) taken as scalar versions of DPC1 and DPC2, respectively. This means that SCS1 and SCS2 should be applied successively, starting with SCS1 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 SCS2 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 appropriate uniform scalar quantizers $\mathcal{Q}_{\Delta_1}$ and $\mathcal{Q}_{\Delta_2}$ have step sizes $\Delta_1 = \frac{\sqrt{12\gamma P}}{\widetilde{\alpha_1}}$ and $\Delta_2 = \frac{\sqrt{12(1-\gamma)P}}{\widetilde{\alpha_2}}$, with

$$(\widetilde{\alpha_1}, \widetilde{\alpha_2}) = \left(\sqrt{\frac{\gamma P}{\gamma P + 2.71 N_1}}, \sqrt{\frac{(1-\gamma)P}{(1-\gamma)P + 2.71 N_2}}\right). \tag{4.4}$$

We denote by $(\widetilde{R_1}, \widetilde{R_2})$ the transmission rate pair practically achieved by this set-up. This pair has to be computed numerically. Results are shown in Fig.4.3 together with the theoretical rate pair (R_1, R_2). The performance of this first approach is worthy of some brief discussion.

(i) From (4.3a), we see that DPC2 -as given by (4.2)- 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 $\mathbf{Y}_1'$. DPC1 -as given by (4.2), as for it, is non optimal. The reason is as follows. The achievable rate R_2 given by (4.3a) is inferior to $\frac{1}{2}\log_2\left(1 + \frac{(1-\gamma)P}{\gamma P + N_2}\right)$, which is that of a watermark signal subject to the full interference penalty from both the cover signal $\mathbf{S}$ and the watermark $\mathbf{X}_1$.

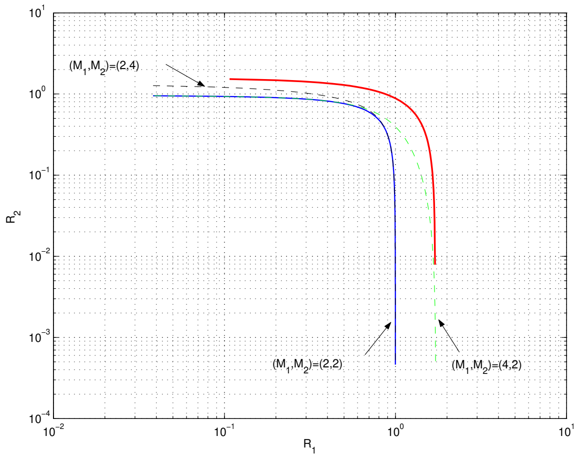


Figure 4.3: Theoretical and feasible transmission rates for broadcast-like multiple user information embedding. The upper curve corresponds to the rate pair (R_1, R_2) (4.3) of the double ideal DPC given by (4.1) and (4.2). The lower curve corresponds to the rate pair $(\widetilde{R_1}, \widetilde{R_2})$ of the two superimposed SCSs with quantization parameters given by (4.4). Dashed line corresponds to (2-ary,4-ary) and (4-ary,2-ary) transmissions. SNRs are such that $P/N_1 = 2P/N_2 = 16$.

(ii) SCS2 performs close to optimality. The scalar channel having the message W_1 as input and the quantization error as output is functionally equivalent to that from W_1 to $\mathbf{r}'_1 = \mathcal{Q}_{\Delta_1}(\mathbf{y}'_1) - \mathbf{y}'_1$, where $\mathbf{y}'_1$ is the single-user channel suffering no interference from the watermark $\mathbf{x}_2$[2]. The practical transmission rate over this channel is, as for a single-user SCS, given by the mutual information $I(r'_1, W_1)$, the maximum of which (i.e $\widetilde{R_1}$) is obtained with the choice (4.4) of $\widetilde{\alpha_1}$. However, SCS1 is non optimal, simply because DPC1 is not. The inflation parameter $\widetilde{\alpha_2}$ does not maximize the mutual information $I(r_2, W_2)$, with $\mathbf{r}_2 = \mathcal{Q}_{\Delta_2}(\mathbf{y}_2) - \mathbf{y}_2$. The practical rate $\widetilde{R_2}$ is the value of this mutual information taken at $\alpha_2 = \widetilde{\alpha_2}$, i.e $\widetilde{R_2} = I(r_2, W_2)|_{\alpha_2 = \widetilde{\alpha_2}}$.

In the following subsection we show that the encoding of W_2 can be improved so as to bring the rate $\widetilde{R_2}$ close to $R_2^{(\max)} = \frac{1}{2}\log_2\left(1 + \frac{(1-\gamma)P}{\gamma P + N_2}\right)$. The corresponding scheme, called "Joint scalar DPC", enhances the performance by making multiuser information embedding coding broadcast-aware.

4.3.2 Broadcast-aware coding for multiuser information embedding

In subsection 4.3.1, we have shown that the communication scenario depicted in Fig.4.1 is basically that of a degraded GBC with state information non-causally known to the transmitter but not to the receivers. In [KSS04], it has been shown (see Appendix B for the proof of the achievability) that the capacity region of this channel is given by

$$R_1 \leq \frac{1}{2}\log_2\left(1 + \frac{\gamma P}{N_1}\right), \tag{4.5a}$$

$$R_2 \leq \frac{1}{2}\log_2\left(1 + \frac{(1-\gamma)P}{\gamma P + N_2}\right), \tag{4.5b}$$

[2]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.

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 (DPC2) is still given by (4.2). For the encoding of W_2 however, the key point is to consider the unknown watermark $\mathbf{X}_1$ as noise, the gaussianity of which will be justified in Section 4.4.3. The resulting DPC (denoted by DPC1) uses the cover signal $\mathbf{S}$ as channel state and $\mathbf{Z}_2 + \mathbf{X}_1$ as total channel noise:

$$\mathbf{U}_2 \sim \mathcal{N}(\alpha_2 \mathbf{S}, (1-\gamma)P) \text{ with } \alpha_2 = \frac{(1-\gamma)P}{(1-\gamma)P + (N_2 + \gamma P)}, \tag{4.6a}$$

$$\mathbf{X}_2 = \mathbf{U}_2 - \alpha_2 \mathbf{S}. \tag{4.6b}$$

Obviously, this encoding does not remove the interference due to $\mathbf{X}_1$. Nevertheless, DPC2 is optimal in that it attains the maximal possible rate R_2^{max} at which W_2 can be sent together with W_1.

4.3.2.1 Joint scalar DPC and capacity region

Consider now a scalar implementation of this Joint DPC scheme consisting in two successive SCSs. DPC1 can be implemented by a scalar scheme SCS1, 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 Δ_1 the corresponding scale factor and quantization step size, respectively. DPC2 can be implemented by a scalar scheme SCS2, quantizing the newly made 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 Δ_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 set of the transmission rate pairs practically feasible by this practical coding is given by the convex hull of all rate pairs $(\widetilde{R_1}, \widetilde{R_2})$ simultaneously satisfying

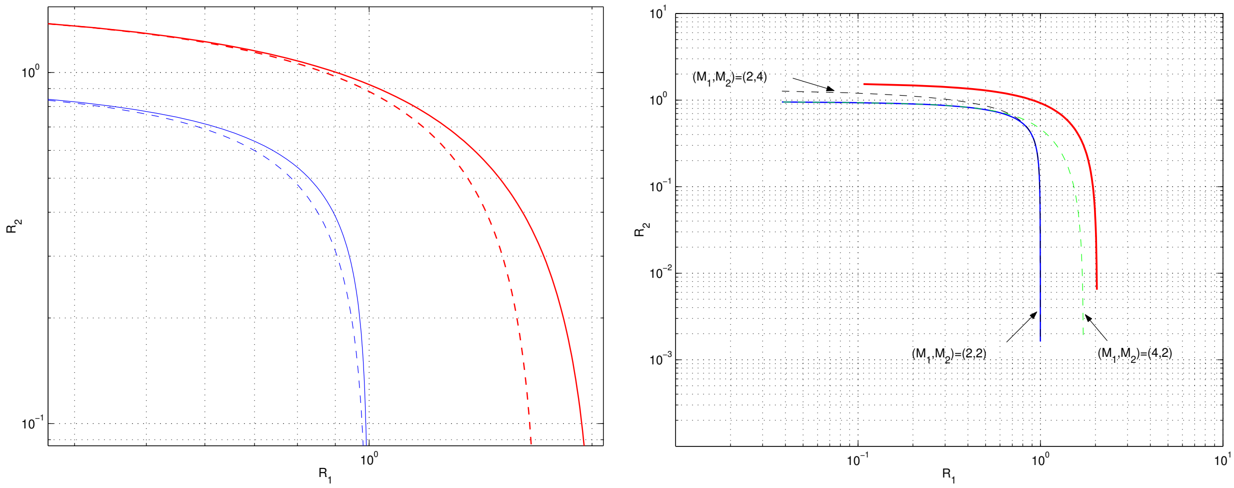
$$\widetilde{R_1} \leq \max_{\alpha_1} \; I(r'_1, W_1), \text{ with } \mathbf{r}'_1 = \mathcal{Q}_{\Delta_1}(\mathbf{y}'_1) - \mathbf{y}'_1, \tag{4.7a}$$

$$\widetilde{R_2} \leq \max_{\alpha_2} \; I(r_2, W_2), \text{ with } \mathbf{r}_2 = \mathcal{Q}_{\Delta_2}(\mathbf{y}_2) - \mathbf{y}_2. \tag{4.7b}$$

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 [EBTG03] 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 (4.7a) and (4.7b) is given by

$$(\widetilde{\alpha_1}, \widetilde{\alpha_2}) = \left(\sqrt{\frac{\gamma P}{\gamma P + 2.71 N_1}}, \sqrt{\frac{(1-\gamma)P}{(1-\gamma)P + 2.71(\gamma P + N_2)}} \right). \tag{4.8}$$

In Fig.4.4(a) the binary feasible capacity region (4.7), obtained through a Monte-Carlo based integration, is compared to the ideal DPC for BC given by (4.5). The (M_1-ary, M_2-ary) feasible capacity region for $(M_1, M_2) = (2, 4)$ and $(M_1, M_2) = (4, 2)$ is depicted in Fig.4.4(b). 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. Also, the curves in Fig.4.4(a) are obtained with the choice of the parameters P, N_1 and N_2 set to $P/N_1 = 2P/N_2 = 16$. Improvement over the "Double DPC" stated in Section 4.4.1 is made possible by increasing the rate R_2



(a) Improvement brought by the joint design of DPCs

(b) Capacity region for (M_1-ary,M_2-ary) alphabets

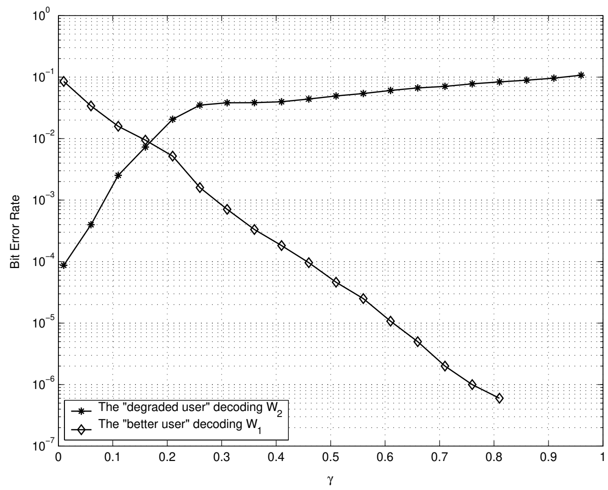
Figure 4.4: Achievable rates for Broadcast-like multiple user information embedding. SNRs are such that $P/N_1 = 2P/N_2 = 16$. (a): achievable rates of the Joint Scalar DPC with comparison to the two superimposed DPCs approach. Solid line corresponds to the capacity region of the joint design for both ideal (upper) and practical (lower) coding. Dashed line corresponds to the rate pair achieved by the Double DPC for both ideal (upper) and practical (lower) coding. (b): achievable rates of the Joint Scalar DPC for M_1-ary and M_2-ary alphabets $\mathcal{M}_1$ and $\mathcal{M}_2$.

at which the robust watermark can be sent. Also, we observe that the larger the alphabet sizes, the larger the feasible capacity region. For very large alphabet sizes $\mathcal{M}_1$ and $\mathcal{M}_2$, the practical joint Scalar DPC performs asymptotically close to the theoretical DPC derived from the broadcast solution as it can be seen in Fig.4.4(b).

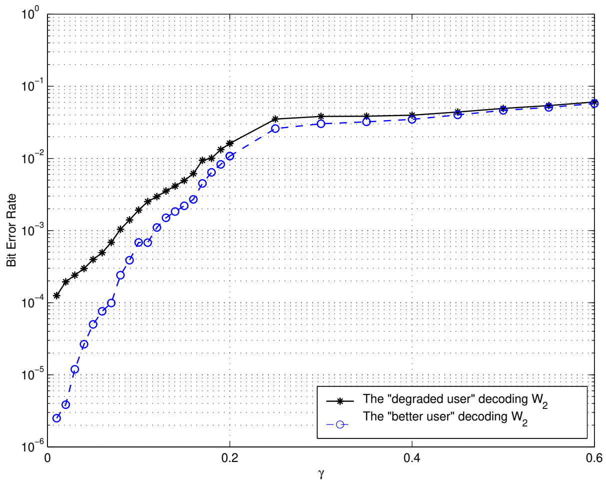
4.3.2.2 Bit Error Rate analysis and discussion

Another performance analysis is based on measured BERs for hard decision based decoding of binary scalar DPC. In Fig.4.5(a) BER curves are obtained with a Monte Carlo simulation. The signal-to-noise ratios are given by $\mathrm{SNR}_1 = 10\log_{10}\left(\frac{\gamma P}{N_1}\right) \in [-8, 12]$ dB and $\mathrm{SNR}_2 = 10\log_{10}\left(\frac{(1-\gamma)P}{\gamma P + N_2}\right) \in [-15, 9]$ dB. 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.4.5(a) are obtained with $(\rho_1, \rho_2) = (4, 4)$, meaning that W_1 and W_2 are being 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 SNR_1 and SNR_2, respectively. This conforms the assumption made above regarding the functionally equivalent channels $\mathbf{y}_1'$ and $\mathbf{y}_2$. The curves in Fig.4.5 also motivate the following discussion.

(i) In practical situations, ρ_1 and ρ_2 should be chosen in light of the desired transmission rates and



(a)



(b)

Figure 4.5: Broadcast-like multiple user information embedding. (a): Bit Error Rates for binary transmission using repetition coding. The messages W_1 and W_2 are repeated 4 times each, i.e. $(\rho_1, \rho_2) = (4, 4)$. Plotted curves correspond to $P/N_1 \approx 2P/N_2 = 16$. (b): the "better user" performs significantly better than the "degraded user" in decoding the information W_2 for small values of γ, but only slightly better as γ approaches unity.

robustness requirements. The choice $(\rho_1, \rho_2) = (4, 4)$ made above should be taken just as a baseline example. Channel 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 ρ_1 and ρ_2 should satisfy $\rho_2 \geq \rho_1$.

(ii) The scalar DPC considered here for multiuser information embedding is constructed using insights from coding for broadcast channels [Cov72, Cov88], as already 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) [see Appendix B]. In a data hiding 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 or peeling off technique (see Appendix B and C) at the "better user" (Decoder 1) [CT91], 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. Thus, the "better user" does not need to bother itself decoding the other message[3].

(iii) There could, however, be advantages and disadvantages for the described DPC-based scheme 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". On the other hand, 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 DPC-based scheme described above, to fulfill this additional requirement, one should focus on maximizing (over α_1) the conditional mutual information $I(r_1, W_1|W_2)$. This would however lead to a suboptimal choice $\widetilde{\alpha_1'}$ of the inflation parameter α_1 for the transmission of W_1, and consequently to a smaller transmission rate

$$\widetilde{R_1} = I(r_1', W_1)\big|_{\alpha_1 = \widetilde{\alpha_1'}}. \tag{4.9}$$

(iv) The present DPC-scheme, as is, partially satisfies this strategy. From Fig.4.5(b), we observe that the "better user" significantly outperforms the "degraded user" in the decoding of W_2 at small values of the parameter γ, but performs only slightly better as γ approaches unity. The advantage the "better user" has upon the "degraded user" in the decoding of the message W_2 depends mainly on the quotient N_2/N_1. This is due to the difference of SNR, $10\log_{10}(\frac{\gamma P+N_2}{\gamma P+N_1})$, which is clearly maximal at $\gamma = 0$. As γ increases, the power allocated to the transmission of W_2 diminishes and

[3]Note that in contrast to superposition coding, there is an important embedding ordering at the encoder. The benifit of such oredering is a decoupling of the receivers and hence a more scalable system. Each receiver need only know its own codebook to extract its message.

drops to zero for γ approaching unity, causing the two decoders to experience very bad SNRs $10\log_{10}(\frac{(1-\gamma)P}{\gamma P+N_1})$ and $10\log_{10}(\frac{(1-\gamma)P}{\gamma P+N_2})$, independently of the noise levels N_1 and N_2.

4.3.3 Extensions: L-watermarks and structured lattice-based codebooks

4.3.3.1 The L-watermark case

The results above 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 and $\sum_{i=1}^{L} P_i = P$. We consider a Gaussian degraded 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 (L-i+1)-th DPC (denoted here by DPCi) analogous to that in (2.20b). 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 gaussianity 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 [Cos83]. 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

$$R_i = \frac{1}{2}\log_2\left(1 + \frac{P_i}{N_i + \sum_{j=1}^{i-1} P_j}\right). \tag{4.10}$$

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 practical capacity region is given by the union of all rate L-tuples $(\widetilde{R_1}, \ldots, \widetilde{R_L})$ simultaneously satisfying

$$\widetilde{R_i} \leq \max_{\alpha_i} \; I(r'_i, W_i), \text{ with } \mathbf{r}'_i = \mathcal{Q}_{\Delta_i}(\mathbf{y}'_i) - \mathbf{y}'_i. \tag{4.11}$$

The union is taken over all power assignments $\{P_i\}$, $i = 1, 2, \ldots, L$, satisfying the average power constraint

$$\sum_{j=1}^{L} P_i = P. \tag{4.12}$$

The inflation parameter maximizing the right hand side term of (4.11) is

$$\widetilde{\alpha}_i = \sqrt{\frac{P_i}{P_i + 2.71\left(N_i + \sum_{j=1}^{i-1} P_j\right)}}. \tag{4.13}$$

4.3.3.2 Lattice-based codebooks for BC-based watermarking

The gap to the ideal capacity region of the sample-wise joint scalar DPC practical capacity region shown in Fig.4.4(a) can be partially bridged using structured finite-dimensional lattice-based codebooks. Lattices have already been considered in the context of single-user watermarking [MK04, ZD05a, ZD05c, ZD06a]. In the following, only the required ingredients are briefly reviewed. The reader may refer to [CS88] for a full discussion. Consider the transmission scheme depicted in Fig.4.6 where Λ is some n-dimensional lattice. This scheme is a generalization to the lattice codebook case of a slight variation of the one considered in Section 4.3.2[4]. The function $\iota_1(.)$ is used for arbitrary mapping the set of indexes $W_1 \in \{1, \ldots, M_1\}$ 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 $W_2 \in \{1, \ldots, M_2\}$. With respect to the scalar codebook case considered in Section 4.3.2, $\mathcal{C}_{w_i}$, $i = 1, 2$, is a lattice codebook whose entries must be appropriately chosen so as to maximize the encoding performance. For each $W_i \in \mathcal{M}_i$, with $i = 1, 2$, the codeword

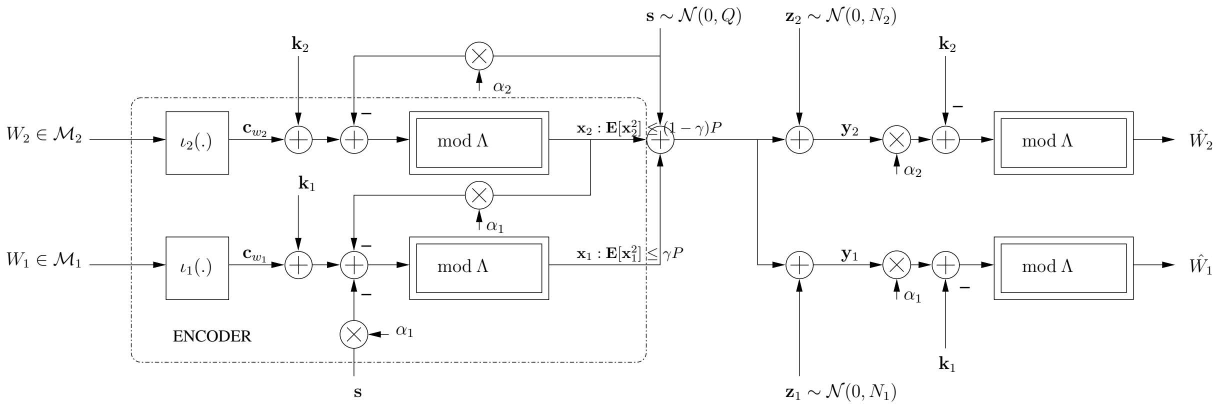


Figure 4.6: Lattice-based scheme for information embedding over a degraded Gaussian Broadcast Channel (GBC).

$\iota_i(W_i) = \mathbf{c}_{w_i}$ is the *coset leader* of the coset $\Lambda_{w_i} = \mathbf{c}_{w_i} + \Lambda$ relative to the lattice Λ. 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 Λ. 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.

[4]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 because it has very close performance to SCS as reported in Section 4.2.2.

In the following, we consider cover signal vectors (frames) of length n. Following (2.59), the encoding and decoding functions for the lattice-based joint DPC given by (4.2) and (4.6) can be written as

$$\mathbf{x}_2(\mathbf{s}; W_2, \Lambda) = (\mathbf{c}_{w_2} + \mathbf{k}_2 - \alpha_2 \mathbf{s}) \bmod \Lambda, \tag{4.14a}$$

$$\mathbf{x}_1(\mathbf{s}; W_1, \Lambda) = (\mathbf{c}_{w_1} + \mathbf{k}_1 - \alpha_1 (\mathbf{s} + \mathbf{x}_2)) \bmod \Lambda, \tag{4.14b}$$

$$\widehat{W}_i = \underset{W_i = 1, \ldots, M_i}{\operatorname{argmin}} \|(\alpha_i \mathbf{y}_i - \mathbf{k}_i - \mathbf{c}_{w_i}) \bmod \Lambda\|, \;\; i = 1, 2. \tag{4.14c}$$

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 capacity region achieved by (4.14). To this end, we rely on a previous work relative to practical achievable rates with lattice codebooks in the context of a single-user watermark [ZD05a, ZD05c]. Here, the situation is different since two watermarks are concerned, but the key ideas remain the same. Thus details are skipped and we only 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 [ZD05a, ZD05c, ZD05b], with however a different state information and a different channel noise. The establishment of the results below relies principally on the properties of a Modulo Lattice Additive Noise (MLAN) channel [FTC00] and on the following two important properties of the mod-Λ operation.

$$\text{(P1)} \;\; \forall (\boldsymbol{\lambda}, \mathbf{a}) \in \Lambda \times \mathbb{R}^n, (\mathbf{a} + \mathbf{v} + \boldsymbol{\lambda}) \bmod \Lambda \; = (\mathbf{a} + \mathbf{v}) \bmod \Lambda. \tag{4.15a}$$

$$\text{(P2)} \;\; \forall \, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2n}, ((\mathbf{x} \bmod \Lambda) + \mathbf{y}) \bmod \Lambda = (\mathbf{x} + \mathbf{y}) \bmod \Lambda. \tag{4.15b}$$

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 (P1) and (P2) and straightforward algebra calculations, it can be shown that

$$\mathbf{r}_1 = (\mathbf{c}_{w_1} + \alpha_1 \mathbf{z}_1 - (1 - \alpha_1)\mathbf{x}_1) \bmod \Lambda, \tag{4.16a}$$

$$\mathbf{r}_2 = (\mathbf{c}_{w_2} + \alpha_2 (\mathbf{z}_2 + \mathbf{x}_1) - (1 - \alpha_2)\mathbf{x}_2) \bmod \Lambda. \tag{4.16b}$$

Hence, the "degraded user" (more noisy composite content) sees the equivalent channel noise $\widetilde{\mathbf{V}}_2$ and the "better user" (less noisy composite content) sees the equivalent channel noise $\widetilde{\mathbf{V}}_1$, where

$$\widetilde{\mathbf{V}}_1 = (\alpha_1 \mathbf{Z}_1 - (1 - \alpha_1)\mathbf{X}_1) \bmod \Lambda, \tag{4.17a}$$

$$\widetilde{\mathbf{V}}_2 = (\alpha_2 (\mathbf{Z}_2 + \mathbf{X}_1) - (1 - \alpha_2)\mathbf{X}_2) \bmod \Lambda. \tag{4.17b}$$

Now, using the important *Inflated Lattice Lemma* reported in [ESZ00], $\mathbf{Y}_1$ 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 [FW89, GDF89]. It is shown that when modulo reduction is with respect to some lattice Λ and when the channel noise $\mathbf{V}$ is i.i.d Gaussian, capacity in bits per dimension can be written as

$$C(\Lambda) = \frac{1}{n} (\log_2(V(\Lambda)) - h(\mathbf{V})), \tag{4.18}$$

where $h(.)$ denotes differential entropy. Hence, the practically achievable rates $R_1(\Lambda)$ and $R_2(\Lambda)$ are given by (4.18), 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 α_1 and α_2, respectively. The corresponding practical capacity region is given by the convex hull of all rate pairs simultaneously satisfying

$$R_1(\Lambda) \leq \max_{\alpha_1} \frac{1}{n} \left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}}_1) \right) < \frac{1}{2} \log_2 \left(1 + \frac{\gamma P}{N_1} \right), \tag{4.19a}$$

$$R_2(\Lambda) \leq \max_{\alpha_2} \frac{1}{n} \left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}}_2) \right) < \frac{1}{2} \log_2 \left(1 + \frac{(1-\gamma)P}{N_2 + \gamma P} \right). \tag{4.19b}$$

The right hand side term of (4.19) is the full capacity region of a Gaussian degraded BC with state information at the encoder, achievable by the theoretical joint DPC scheme described before. In general no closed form of (4.19) can be derived and the optimal pair (α_1, α_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 α_1 and α_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 $(\alpha_1, \alpha_2) = (\frac{\gamma P}{\gamma P + N_1}, \frac{(1-\gamma)P}{N_2 + P})$ 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[5] $h(\widetilde{\mathbf{V}}_1) \leq \log(2\pi e \alpha_1 N_1)$ and $h(\widetilde{\mathbf{V}}_2) \leq \log 2\pi e \alpha_2 (N_2 + \gamma P)$, we get

$$R_1(\Lambda) \geq \frac{1}{n} \left(\frac{1}{2} \log(1 + \frac{\gamma P}{N_1}) - \frac{1}{2} \log 2\pi e G(\Lambda) \right), \tag{4.20a}$$

$$R_2(\Lambda) \geq \frac{1}{n} \left(\frac{1}{2} \log(1 + \frac{(1-\gamma)P}{N_2 + \gamma P}) - \frac{1}{2} \log 2\pi e G(\Lambda) \right). \tag{4.20b}$$

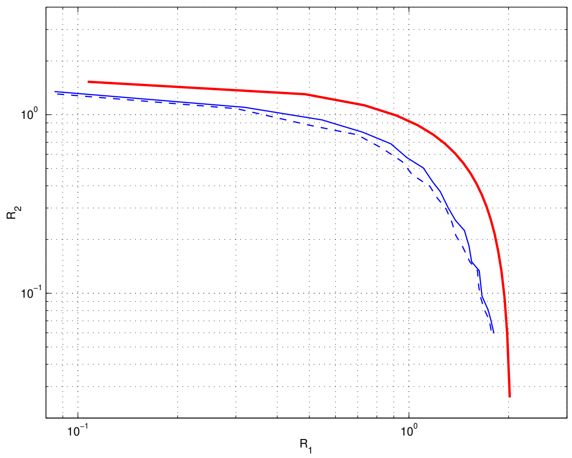
This means that using appropriate lattices for modulo-reduction, we are able to make the gap to the full theoretical capacity region smaller then $\log 2\pi e G(\Lambda)$. This can be achieved by selecting lattices that have good quantization properties. These are those for which the normalized second moment $G(\Lambda)$ approaches $\frac{1}{2\pi e}$.

The n-dimensional lattices considered for Monte-Carlo capacity region integration are summarized in table 4.1, together with their most important parameters. Capacity region curves in bits per dimension

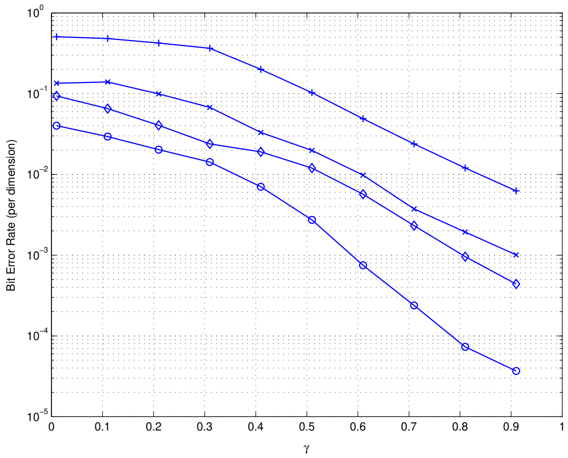
Lattice	Name	n	$G(\Lambda)$	$\gamma_s(\Lambda)$ [dB]	$\gamma_s(\Lambda)$ [bit per dimension]
$\mathbb{Z}$	Integer Lattice	1	$\frac{1}{12}$	0.00	0.000
A_2	Hexagonal Lattice	2	$\frac{5}{36\sqrt{3}}$	0.17	0.028
D_4	$4D$ Checkerboard L.	4	0.0766	0.37	0.061

Table 4.1: The considered lattices for multiple user information embedding

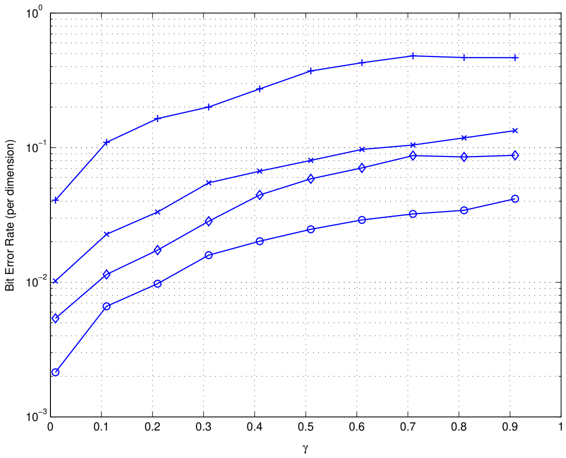
[5]This is because the normal distribution is the one that maximizes entropy for a given second moment.



(a) Feasible capacity region.



(b) The "better user" decoding W_1.



(c) The "degraded user" decoding W_2.

Figure 4.7: Performance improvement of broadcast-like multiple user information embedding due to the use of lattice codebooks. (a): feasible capacity region obtained with lattices $\mathbb{Z}$ (dashed line) , A_2 (solid point) and infinite dimensional hypershere (bold). SNRs are such that $P/N_1 = 2P/N_2 = 16$. (b) Bit Error Rate in decoding the first message W_1 obtained with lattices $\mathbb{Z}$ (plus sign), A_2 (cross) D_4 (diamond) and Gosset E_8 (circle). and (c) Bit Error Rate in decoding the second message W_2 obtained with lattices $\mathbb{Z}$ (plus sign), A_2 (cross) D_4 (diamond) and Gosset E_8 (circle).

are plotted in Fig.4.7(a) where 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 $\mathbb{Z}$. That of the lattice D_4 further enlarges it. Of course, this improvement goes along with a slight increase in computational cost.

4.4 Watermarking over a Gaussian Multiple Access Channel: Performance analysis

In this section we are interested in designing implementable multiuser information embedding schemes for the situation described in Section 4.4.2. Paralleling the development made in Section 4.3, we provide a performance analysis for two MAC-aware and unaware multiuser inforrmation embedding strategies. The former consists in super-imposing two SCSs. The latter uses the analogy with a Gaussian MAC with state information available at the transmitters recognized above to bring up performance. This improvement is illustrated through both achievable capacity region and achievable Bit Error Rates (BERs) enhancement. Finally, results are extended to both the K-watermark case and the high dimensional lattice-based codebooks case. Whenever the development closely follows the one we have stated in Section 4.3, details are skipped.

4.4.1 MAC-unaware coding for multiuser information embedding

The situation described in Section 4.2.2 corresponds in essence to two "Dirty Paper" 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 the other message. Suppose for example that encoder 2 uses a DPC (DPC1) taking into account the known state $\mathbf{S}$ and the 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

$$\mathbf{U}_2 \sim \mathcal{N}\left(\alpha_2\mathbf{S}, P_2\right), \text{ with } \alpha_2 = \frac{P_2}{P_2 + N}. \tag{4.21}$$

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. 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 (DPC2). For that, the watermark $\mathbf{X}_1$ is formed as $\mathbf{X}_1 = \mathbf{U}_1 - \alpha_1\mathbf{S}$, where

$$\mathbf{U}_1 \sim \mathcal{N}\left(\alpha_1\mathbf{S}, P_1\right), \text{ with } \alpha_1 = (1-\alpha_2)\frac{P_1}{P_1+N} = \frac{NP_1}{(P_1+N)(P_2+N)}. \tag{4.22}$$

The rates theoretically achievable by these two DPCs are those corresponding to the corner point (B1) of the diagram shown in Fig.4.8 and are given by

$$R_1(B1) = \frac{1}{2}\log_2\left(1 + \frac{P_1}{N}\right), \tag{4.23a}$$

$$R_2(B1) = \frac{1}{2}\log_2\left(\frac{P_2(P_2+Q+N+P_1)}{P_2Q(1-\alpha_2)^2 + (N+P_1)(P_2+\alpha^2Q)}\right). \tag{4.23b}$$

Following the same principles, 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 $\Delta_1 = \frac{\sqrt{12P_1}}{\widetilde{\alpha_1}}$ and $\Delta_2 = \frac{\sqrt{12P_2}}{\widetilde{\alpha_2}}$, with

$$(\widetilde{\alpha_1}, \widetilde{\alpha_2}) = \left((1-\alpha_2)\sqrt{\frac{P_1}{P_1 + 2.71N}}, \sqrt{\frac{P_2}{P_2 + 2.71N}} \right). \tag{4.24}$$

The feasible transmission rate pair achieved by this practical coding corresponds to the corner point (B1') in the diagram shown in Fig.4.8. As stated before, the point (C1') corresponds to the point (B1') with the roles of the watermarks $\mathbf{X}_1$ and $\mathbf{X}_2$ being reversed. The performance of this first approach,

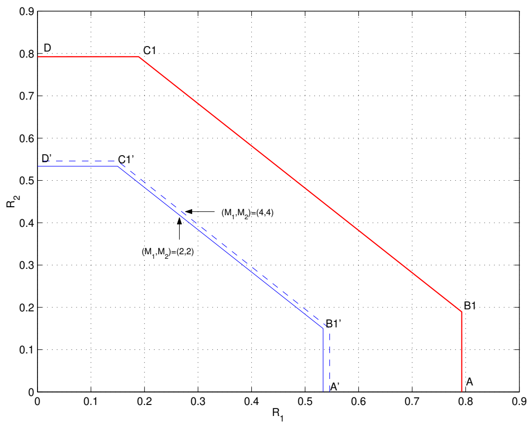


Figure 4.8: Theoretical and feasible transmission rates for MAC-like multiple user information embedding. The frontier with corner points (A), (B1), (C1), and (D) corresponds to the theoretical rate pair (R_1, R_2) of the double ideal DPC. The frontier with corner points (A'), (B1'), (C1'), and (D') 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. Numeric values are set to $Q/P = 100$, $P_1/N_1 = P_2/N_2 = 2$.

including both its theoretical and its practical settings, can be summarized as follows.

(i) From (4.23a), we see that DPC2 -as given by (4.22)- 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, DPC1 -as given by (4.21)- is non optimal. The reason is as follows. The achievable rate R_2 given by (4.23b) is inferior to $\frac{1}{2}\log_2\left(1 + \frac{P_2}{P_1+N}\right)$, which is that of a watermark subject to the full interference penalty from both the cover signal $\mathbf{S}$ and the watermark $\mathbf{X}_1$.

(ii) SCS2 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

$I(r_1, W_1)$, the maximum of which (i.e $\widetilde{R_1}$) is obtained with the choice (4.24) of $\widetilde{\alpha_1}$. However, SCS1 is non optimal, simply because DPC1 is not. The inflation parameter $\widetilde{\alpha_2}$ does not maximize the mutual information $I(r, W_2)$, with $\mathbf{r} = \mathcal{Q}_{\Delta_2}(\mathbf{y}) - \mathbf{y}$. The practical rate $\widetilde{R_2}$ is not maximal and corresponds to the value of this mutual information taken at $\alpha_2 = \widetilde{\alpha_2}$, i.e $\widetilde{R_2} = I(r, W_2)|_{\alpha_2 = \widetilde{\alpha_2}}$.

The encoding of W_2 can be improved so as to bring the practical 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 scalar DPC", enhances the performance by making multiuser information embedding coding MAC-aware.

4.4.2 Gaussian MAC-aware coding for multiuser information embedding

In subsection 4.2.2, we have argued that the communication scenario depicted in Fig.4.2 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 [KSS04], it is reported (see Appendix C for the proof of the achievability) that the capacity region of this channel is given by

$$R_1 \leq \frac{1}{2}\log_2\left(1 + \frac{P_1}{N}\right), \tag{4.25a}$$

$$R_2 \leq \frac{1}{2}\log_2\left(1 + \frac{P_2}{N}\right), \tag{4.25b}$$

$$R_1 + R_2 \leq \frac{1}{2}\log_2\left(1 + \frac{P_1 + P_2}{N}\right), \tag{4.25c}$$

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.4.9(a) 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 (4.22), recognized above to be optimal[6]. The encoding DPC1 of W_2 however should be changed so as to consider the watermark $\mathbf{X}_1$ as noise. The resulting DPC (again denoted by DPC1) uses the cover signal $\mathbf{S}$ as channel state and the signal $\mathbf{Z} + \mathbf{X}_1$ as total channel noise:

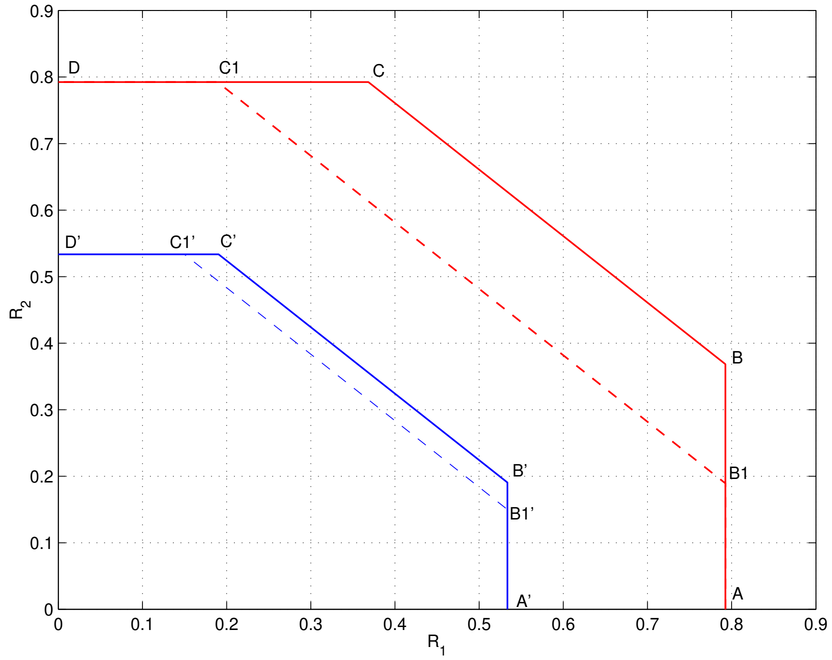
$$\mathbf{U}_2 \sim \mathcal{N}(\alpha_2 \mathbf{S}, P_2), \text{ with } \alpha_2 = \frac{P_2}{P_2 + (P_1 + N)}. \tag{4.26}$$

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.

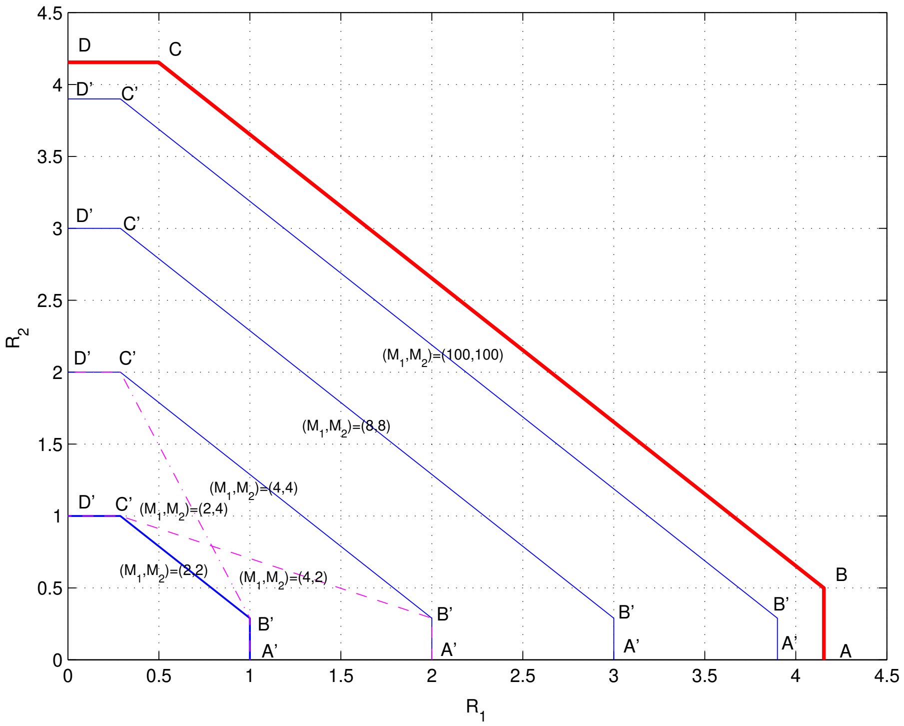
4.4.2.1 Joint scalar DPC and Capacity region

We consider now as practical implementation of this joint scheme 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 $\widetilde{R_2} = \max_{\alpha_2} \ I(r, W_2)$. However, the corresponding scale parameter α_2 is set this time to its optimal choice, i.e, $\widetilde{\alpha_2} = \sqrt{\frac{P_2}{P_2 + 2.71(N + P_1)}}$. The resulting transmission rate pair $(\widetilde{R_1}, \widetilde{R_2})$ is represented by the corner point (B') in Fig.4.9(a). Reversing the roles of the watermarks $\mathbf{X}_1$ and $\mathbf{X}_2$, the

[6]Note however that as α_1 depends on α_2, the optimal inflation parameter for DPC2 becomes $\alpha_1 = P_1/(P_1 + P_2 + N)$.



(a) Improvement brought by the joint design, $(M_1, M_2) = (2, 2)$



(b) Capacity region for $(M_1 - \text{ary}, M_2 - \text{ary})$ alphabets

Figure 4.9: MAC-like multiple user information embedding. (a): Achievable rates for the joint scalar DPC with comparison to the two superimposed DPCs approach. Solid line delineates the capacity region of both ideal (upper) and practical (lower) coding. Dashed line delineates the rate pair feasible with the Double DPC for both ideal (upper) and practical (lower) coding. (b): achievable rates with the joint scalar DPCs for M_1-ary and M_2-ary alphabets $\mathcal{M}_1$ and $\mathcal{M}_1$.

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 $\max_{\alpha_i} \; I(r_i, W_i)$, 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. The maximal feasible sum rate remains unchanged, however. Consequently, the practically feasible capacity region is given by the convex hull of all rate pairs $(\widetilde{R_1}, \widetilde{R_2})$ simultaneously satisfying

$$\widetilde{R_1} \leq \max_{\alpha_1} \; I(r_1, W_1), \text{ with } \mathbf{r}_1 = \mathcal{Q}_{\Delta_1}(\mathbf{y}_1) - \mathbf{y}_1, \tag{4.27a}$$

$$\widetilde{R_2} \leq \max_{\alpha_2} \; I(r_2, W_2), \text{ with } \mathbf{r}_2 = \mathcal{Q}_{\Delta_2}(\mathbf{y}_2) - \mathbf{y}_2, \tag{4.27b}$$

$$\widetilde{R_1} + \widetilde{R_2} \leq \max_{\alpha_1} \; I(r_1, W_1) + \max_{\alpha_2} \; I(r, W_2), \text{ with } \mathbf{r} = \mathcal{Q}_{\Delta_2}(\mathbf{y}) - \mathbf{y}. \tag{4.27c}$$

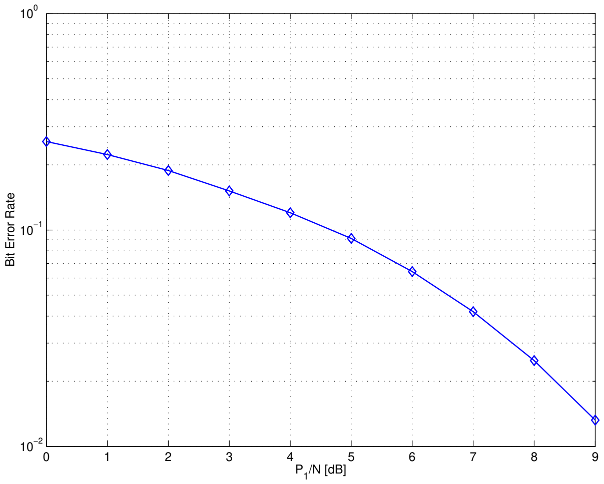
Fig.4.9(a) shows the feasible capacity region gain brought by the joint design of the DPCs in approaching the theoretical limit (4.25), with respect to the first method addressed above. Note that this improvement is especially visible in the situations where W_1 and W_2 are both transmitted with non-zero rate. 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. Equivalently, for a given transmission rate $\widetilde{R_1}$ of W_1, the maximal transmission rate at which W_2 can be sent is larger. Note also that the gap to the theoretical limit (4.25) can be reduced by use of sufficiently large size alphabets $\mathcal{M}_1$ and $\mathcal{M}_2$ as shown in Fig.4.9(b). Of course, this is achieved at the cost of a slight increase in encoding and decoding complexities.

4.4.2.2 BER analysis and discussion

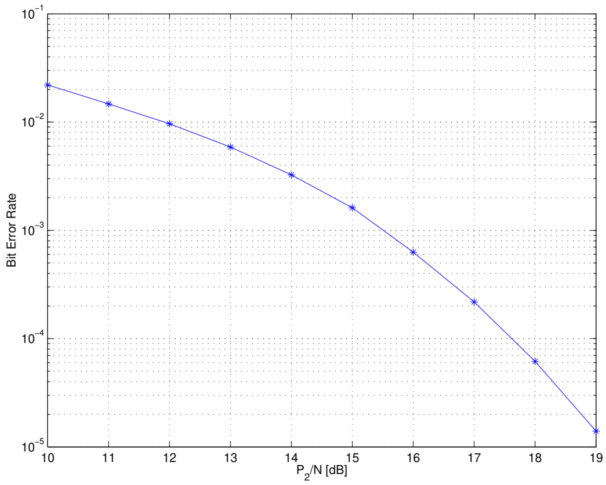
Consider the coding scheme given by (4.22) and (4.26). The key point is, as already mentioned, the *peeling off* technique. This technique aims to clean up the channel before decoding W_1, by subtracting the codeword $\mathbf{U}_2$. However, the transmission of W_2 suffers from the additional noise $\mathbf{x}_1$. The corresponding Signal-to-Noise Ratios (per-bit) SNR1 and SNR2 are given by SNR1 $= \frac{P_1}{R_1 N}$ [dB] and SNR2 $= \frac{P_2}{R_2(N+P_1)}$ [dB]. Thus, the BER curve corresponding to the transmission of W_2 can be obtained by translating to the right that of W_1, by

$$\beta(R_1, R_2) = \frac{R_1 P_2 N}{R_2 P_1 (N + P_1)} [\text{dB}]. \tag{4.28}$$

The upper curve in Fig.4.12 depicts the error probability relative to the transmission of W_1 using scalar codewords. We now pause to discuss the efficiency of the *peeling off* technique in practice. Such a strategy 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 $\mathbf{U}_2$. Theoretically, the $\widehat{\mathbf{U}}_2$ is determined as the (unique) codeword being typically joint with the received signal $\mathbf{Y}$. Of course, this is obtained by an estimation prosedure in practice. The accuracy of this estimation, and hence that of the decoding of the message W_1, depends on on SNR2. For instance, 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



(a) Decoding of W_1



(b) Decoding of W_2

Figure 4.10: 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.4.9. Upon reception, the decoder first cleans up the channel by decoding W_2 (b) and then decodes W_1 (a). Thus, the reliable the transmission of W_2, the more accurate the decoding of W_1.

additional noise source. Hence in this SNR2 range, the *peeling off* technique does not "properly" clean up the channel, as it is supposed to. Hence, decoding of W_1 is not necessarily improved. 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.4.10. Note for instance that at the same SNR, the decoding of the message W_1 is more accurate than that of W_2, though $P_2 = 10P_1$.

4.4.3 Extensions: K-users and structured lattice-based codebooks

4.4.3.1 The K-watermark case

The results above 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 Section 4.2.2. The capacity region of such a channel follows a straightforward generalization of (4.25). This region is given by the union of all rate K-tuples simultaneously satisfying

$$R_i \leq \frac{1}{2}\log_2\left(1 + \frac{P_i}{N}\right), \quad i = 1, 2, \ldots, K, \tag{4.29a}$$

$$\sum_{j=1}^{K} R_j \leq \frac{1}{2}\log_2\left(1 + \frac{\sum_{i=1}^{K} P_i}{N}\right), \tag{4.29b}$$

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 {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 $\mathbf{U}_i \sim \mathcal{N}(\alpha_i \mathbf{S}, P_i)$ 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 $R_i = \frac{1}{2}\log_2(1 + P_i/(\sum_{j=1}^{i-1} P_j + N))$. 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 (4.27). The corresponding practically feasible capacity region is given by the convex hull of all rate

K-tuples $(\widetilde{R_1}, \ldots, \widetilde{R_K})$ simultaneously satisfying

$$\widetilde{R_i} \leq \max_{\alpha_1} \; I(r_i, W_i), \;\text{ with } \mathbf{r}_i = \mathcal{Q}_{\Delta_i}(\mathbf{y}_i) - \mathbf{y}_i, \; i = 1, 2, \ldots, K, \tag{4.30a}$$

$$\sum_{j=1}^{K} \widetilde{R_j} \leq \sum_{j=1}^{K} \max_{\alpha_j} \; I(r_{j,b}, W_j), \;\text{ with } \mathbf{r}_{j,b} = \mathcal{Q}_{\Delta_j}(\mathbf{y}_{j,b}) - \mathbf{y}_{j,b}. \tag{4.30b}$$

The maximum of the mutual information $I(r_{i,b}, W_i)$ is attained with the optimal choice of α_i given by

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

4.4.3.2 Lattice-based codebooks for MAC-like watermarking

The gap to the ideal capacity region of the practical capacity region (4.27) shown in Fig.4.9 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.4.11 where Λ 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 in Section 4.4.3. 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 in subsection 4.3.2. Consider for example the

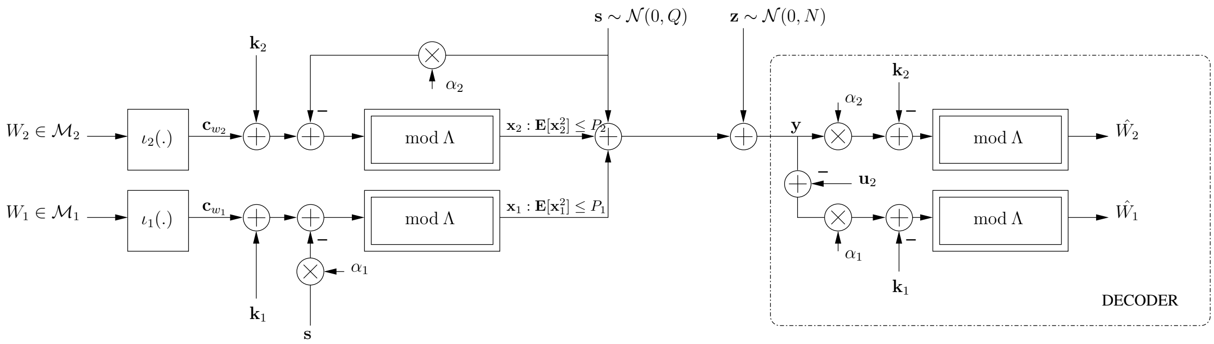


Figure 4.11: Lattice-based scheme for information embedding over a Gaussian Multiple Access Channel (GMAC).

corner point (B') of the capacity region shown in Fig..4.9. The encoding and decoding of W_1 and W_2 are performed according to

$$\mathbf{x}_1(\mathbf{s}; W_1, \Lambda) = (\mathbf{c}_{w_1} + \mathbf{k}_1 - \alpha_1(1 - \alpha_2)\mathbf{s}) \bmod \Lambda \tag{4.31a}$$

$$\mathbf{x}_2(\mathbf{s}; W_2, \Lambda) = (\mathbf{c}_{w_2} + \mathbf{k}_2 - \alpha_2\mathbf{s}) \bmod \Lambda, \tag{4.31b}$$

$$\widehat{W_1} = \underset{W_1 = 1, \ldots, M_1}{\operatorname{argmin}} \; \|(\alpha_1\mathbf{y}_1 - \mathbf{k}_1 - \mathbf{c}_{w_1}) \bmod \Lambda\|, \tag{4.31c}$$

$$\widehat{W_2} = \underset{W_2 = 1, \ldots, M_2}{\operatorname{argmin}} \; \|(\alpha_2\mathbf{y} - \mathbf{k}_2 - \mathbf{c}_{w_2}) \bmod \Lambda\|, \tag{4.31d}$$

where in (4.31c) $\mathbf{y}_1 = \mathbf{y} - (\mathbf{x}_2 + \alpha_2\mathbf{s})$. 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 in subsection 4.3.3, it can be shown that the signal $\mathbf{r}$ is given by $\mathbf{r} = (\mathbf{c}_{w_2} + \alpha_2(\mathbf{z} + \mathbf{x}_1) - (1 - \alpha_2)\mathbf{x}_2) \bmod \Lambda$. 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 $\widetilde{\mathbf{v}_1} = (\alpha_1\mathbf{z} - (1 - \alpha_1)\mathbf{x}_1) \bmod \Lambda$. Consequently, the practical transmission rate pair $(R_1(B'), R_2(B'))$ corresponding to the corner point (B') of the capacity region is given by

$$R_1(B') = \max_{\alpha_1} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}_1})\right) < \frac{1}{2}\log_2\left(1 + \frac{P_1}{N}\right), \tag{4.32a}$$

$$R_2(B') = \max_{\alpha_2} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}_2})\right) < \frac{1}{2}\log_2\left(1 + \frac{P_2}{N + P_1}\right). \tag{4.32b}$$

Similarly to the development made in subsection 4.4.3, the capacity region practically feasible by using the modulo reduction with respect to the lattice Λ straightforwardly generalizes (4.27) and is given by the set of all rate pairs $(R_1(\Lambda), R_2(\Lambda))$ simultaneously satisfying

$$R_1(\Lambda) \le \max_{\alpha_1} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}_1})\right), \tag{4.33a}$$

$$R_2(\Lambda) \le \max_{\alpha_2} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}_2})\right), \tag{4.33b}$$

$$R_1(\Lambda) + R_2(\Lambda) \le \max_{\alpha_1} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}_1})\right) + \max_{\alpha_2} \frac{1}{n}\left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}})\right), \tag{4.33c}$$

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.4.12 through the use of some finite dimensional lattices with good coding and quantizing properties.

Lattice codebooks (equivalent to multidimensional constellations in conventional communication) provide gains over scalar codebooks (equivalent to Pulse Amplitude Modulation (PAM) constellations) by improving the coding (coding gain $\gamma_c(\Lambda)$) and introducing the shaping (shaping gain $\gamma_s(\Lambda) = 1/12G(\Lambda)$). $G(\Lambda)$ is the second moment of the lattice. A full focus on lattices can be found in [CS88, FW89]. The n-dimensional lattices considered for Monte-Carlo capacity region integration are summarized in Table 4.1, together with their most important parameters.

4.5 Summary

In this chapter, we first investigated the tight relationship between multiple user information embedding and conventional multi-user information theory. For instance, two different situations of embedding several messages into one common cover signal are addressed. The first situation is recognized as being equivalent to communication over a degraded Gaussian Broadcast Channel (BC) with state information known to the transmitter but not the receivers. The second is argued as to be analog to communication over a Gaussian Multiple Access Channel with state information known to the transmitters but not the receiver. Next, based on this equivalence and relying on recent advances in network information

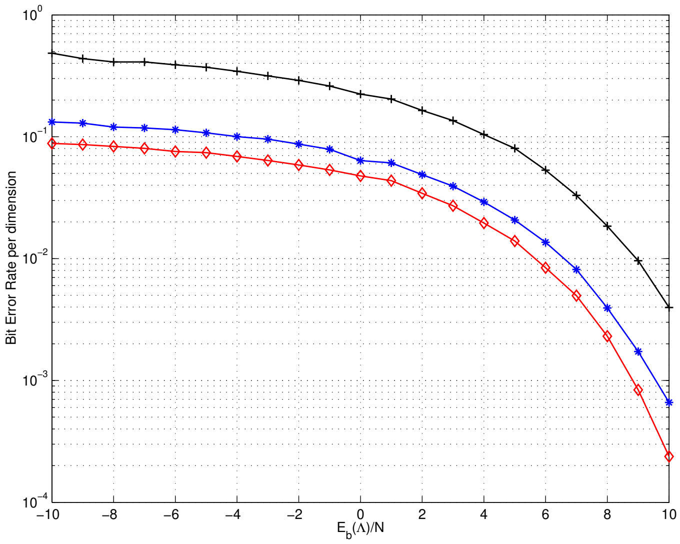


Figure 4.12: Bit Error Probability v.s. the (per-dimension per-bit) Signal-to-Noise Ratio SNR1 $= E_b(\Lambda)/N$ for QIM-embedding the message W_1. From bottom to top: lattices Checkerboard D_4, Hexagonal A_2 and Cubic $\mathbb{Z}$. Bit Error Probabilities (plotted v.s. SNR2) corresponding to transmitting W_2 are obtained by shifting these curves to the right according to (4.28).

theory, two practically feasible scalar schemes for simultaneously embedding two messages into the same host signal are proposed. These schemes turn to 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. The improvement brought by this joint design 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 capacity region and Bit Error Rates. 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 capacity region and BER curves obtained through Monte-Carlo integration and Monte-Carlo-simulation, respectively.

Chapter 5

On Channel Sensitivity to Partially Known Two-sided State Information

The content of this chapter has been partially published in [ZD05b, ZD06b, ZD06c].

In some information embedding situations, the encoder may not have perfect knowledge of the host signal. This is the case when, for security purposes, the encoder observes only a short description (say a quantized version) of the host signal, for example. In these situations, the overall transmission scheme is equivalent to communication over a channel with *partial* knowledge of the channel state at the transmitter. This state information may be viewed as the sum of a dominant (nominal) state information and a relatively weak perturbation. Obviously, the "uncertainty" about the channel state leads to a certain performance loss. This chapter is concerned with (i) evaluating capacity and rate losses and, (ii) providing insights into how efficiently use the "available knowledge" so as to increase system immunity to channel perturbations. As information embedding is a special case of communication over a channel with two-sided state information (see Section 2.1.3), we broaden our view to consider the general case of channel with an arbitrary pair of independent and identically distributed (i.i.d), possibly correlated, state information vector (S_1, S_2) available at the transmitter and at the receiver, respectively. We first analyze the decrease in capacity, or channel sensitivity to this perturbing noise. Both lower and upper bounds on this channel sensitivity are provided, using Fisher Information [CT91]. The lower bound turns out to be relatively tight, at low Signal-to-Noise-Ratio (SNR), in the Gaussian case, for which we provide closed form expression for channel capacity degradation. Next, we show that these results can be used to increase system immunity to noise, by adapting the encoder to the channel uncertainty. Finally, for illustration purposes, two possible applications are discussed.

5.1 Channel with Two-sided State Information

Consider the channel model shown in Fig.5.1. The vector (S_1^n, S_2^n) represents a pair of possibly correalted two-sided state information. The side information S_1^n is non-causally available at the transmitter and S_2^n is non-causally available at the receiver. This channel may model a variety of communications

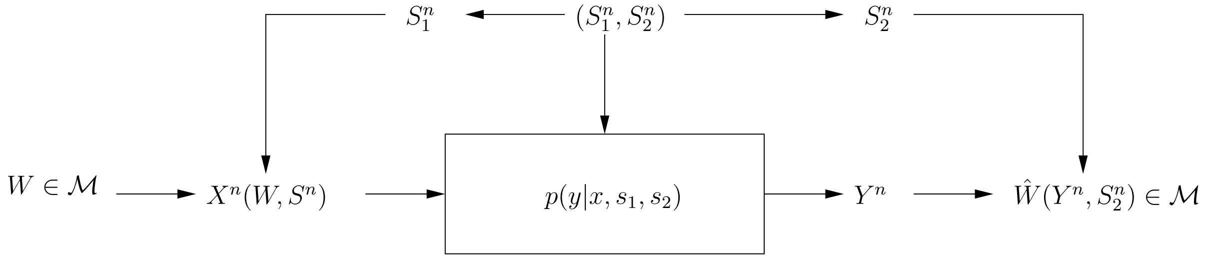


Figure 5.1: Channel with perfect knwoledge of a tow-sided state information pair (S_1^n, S_2^n). S_1^n is non-causally available at the transmitter and S_2^n is non-causally available at the receiver.

situations of practical usefulness. In the contex of information embedding, the blind embedding scheme considered in Chapter 2, 3 and 4 corresponds to the special case where $S_2^n = 0$. Non-blind embedding, on the other hand, corresponds to the special case where $S_1^n = S_2^n = S^n$. In the context of multiple-input multiple-output (MIMO) systems [Pro01], where the block-fading coefficients are usually obtained through a feedback channel, the transmitter and the receiver may see two different state information, due to feedback errors. When only the transmitter knows the state information, the channel capacity was provided by Gel'fand and Pinsker [GP80] and is given by (2.9). When only the receiver knows the state information, channel capacity was proved by Heegard and El Gamal in [HG83] to be

$$C_{01}^{mc} = \max_{p(x)} I(X; Y|S). \tag{5.1}$$

When both the transmitter and the receiver have access to possibly correlated two state information S_1^n and S_2^n, channel capacity expression was provided by Cover and Chiang in [CC02], using a unifying approach. They have shown that, in this case, channel capacity is given by a generalization of (2.9) and (5.1), as

$$C^{nc} = \max_{p(u,x|s_1)} \big\{ I(U; Y, S_2) - I(U; S_1) \big\}. \tag{5.2}$$

The auxiliary random variable U satisfies a Markov chain analog to that needed for establishing (2.9). Both (2.9) and (5.1) assume perfect knowledge of the state information at the transmitter and/or at the receiver. In some applications however, this state information may be known only with some uncertainty. Examples include partial state information at the encoder [RSS05]. This case received considerable

attention recently for its potential use, for example, in the context of multiuser systems (see for example [SH05] and references therein). In other situations, the uncertainty is located at the decoder [CDW01]. Many works focus on how partial state information, at the encoder [JSO02, LLC03] or at the decoder [DW04], can be utilized for improving system performance. In this chapter, we consider coding for communication over channels where the transmitter and the receiver may observe different versions (estimates) of channel conditions. This amounts to coding with noisy, or perturbed, state information at both sides of the channel. However, rather than focusing on designing transmission algorithms, we are interested in evaluating the loss due to the "uncertainty" or state information perturbation. Note that we also partially address the design of coding algorithms for the situation at hand. We assume imperfect knowledge of the two-sided state information. Namely, we consider the case where the sender has access to some part, S_1^n, of a noisy state information $\widetilde{S_1}^n = S_1^n + \theta_1 Z_1^n$. Similarly, we assume that the receiver has access to some part, S_2^n, of a noisy state information $\widetilde{S_2}^n = S_1^n + \theta_2 Z_2^n$. The pair (S_1^n, S_2^n) can be viewed as the dominant (so-called nominal) part of the two-sided state information $(\widetilde{S_1}^n, \widetilde{S_2}^n)$. The term $(\theta_1 Z_1^n, \theta_2 Z_2^n)$ is an unknown perturbing noise, independent of the two states S_1^n and S_2^n.
If $\theta \triangleq (\theta_1, \theta_2) \neq (0,0)$, the noise-like perturbation makes the nominal capacity $C^{nc}(0)$, given by (5.2), decrease to $C^{nc}(\theta)$, thus incurring the capacity loss given by $C^{nc}(0) - C^{nc}(\theta)$. In the following, we focus on this loss in channel capacity, or channel sensitivity to the perturbation. First, we use Fisher information to provide lower and upper bounds on this capacity degradation. The key ingredient for deriving these bounds is an expression of the entropy of a variable slightly contaminated by another, as provided in [Pha05]. We also consider the Gaussian case, for which explicit expression of capacity loss exist, and show that the encoder should *adapt* to the imperfect knowledge of the channel at the receiver. Finally two illustrative applications in the causal and the non-causal case are discussed.
The remainder of this chapter is organized as follows. In Section 5.2, we address channel capacity degradation in the presence of two-sided noisy state information. Section 5.3 considers the Gaussian case for which we provide closed expressions for channel sensitivity and discuss the tightness as well as the usefulness of the bounds on channel sensitivity, obtained in the general case. In Section 5.4, we reconsider the particular case of channel sensitivity due to small perturbations of the state information available at the encoder only, and re-establish the channel sensitivity expression. In Sections 5.5 and 5.6, we provide straightforward extension of these results and illustrative applications, respectively. Final concluding remarks are given in Section 5.7.

5.2 Channel Sensitivity to Small Perturbation of the Two-Sided State Information

Consider the channel $\tilde{Y} = X + \widetilde{S_1} + \widetilde{S}_2 + V$ depicted in Fig.5.2. The pair (S_1, S_2) is a pair of strong (nominal) two-sided state information. Variable S_1 is i.i.d, non-causally known to the transmitter and S_2 is i.i.d, non-causally known to the receiver. (Z_1, Z_2) is a pair of noise perturbations to the nominal state information pair (S_1, S_2). These noise terms, Z_1 and Z_2, are assumed to be independent of each other and also, independent of the two nominal states S_1 and S_2. We write $\widetilde{S}_i = S_i + \theta_i Z_i$, $i = 1, 2$. Also, we suppose without loss of generality that $\mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0$ and denote by θ and Z the vectors

$\theta = (\theta_1, \theta_2)^T$ and $Z = (Z_1, Z_2)^T$, where $\cdot^T$ denotes the transpose operation. With these notations, the

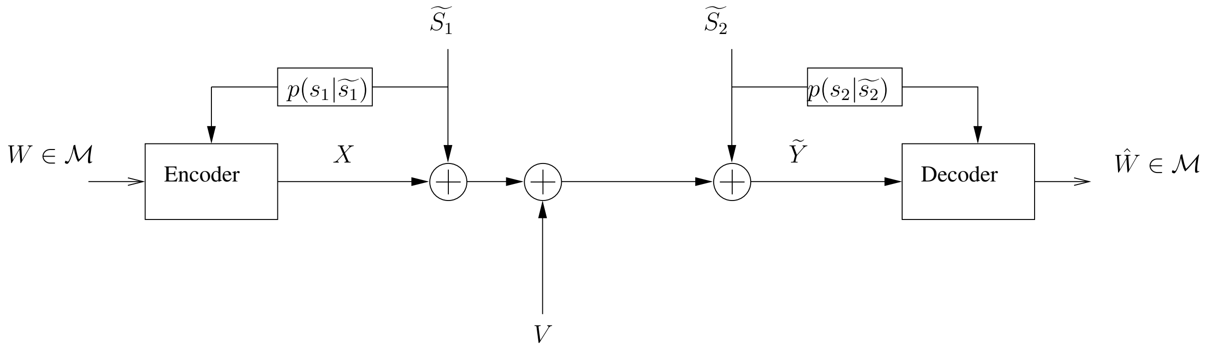


Figure 5.2: Channel with a two-sided state information pair $(\widetilde{S_1}, \widetilde{S_2})$. $\widetilde{S_1}$ is known only partially at the transmitter and $\widetilde{S_2}$ is known only partially at the receiver.

receiver sees the signal $\tilde{Y} = Y + \theta^T Z$. In the classical case where the perturbation is zero, the received signal is $Y = X + S_1 + S_2 + V$. In this case, the capacity of the channel of input W and received signal Y, in presence of the two-sided state information pair (S_1, S_2), has been expressed by Cover et al. in [CC02]. They have shown that channel capacity is given by

$$C^{nc}(0) = \max_{p(u,x|s_1)} \big\{ I(U; S_2, Y) - I(U; S_1) \big\}, \tag{5.3}$$

where the auxiliary variable U satisfies conditions of (5.2). If $\theta \neq 0$, channel capacity can be expressed by a slight variation of (5.2),

$$C^{nc}(\theta) = \max_{p(u,x|s_1)} \big\{ I(U; S_2, \tilde{Y}) - I(U; S_1) \big\}. \tag{5.4}$$

The state information perturbation $\theta^T Z$ results in a capacity loss, with respect to the nominal capacity $C^{nc}(0)$, which can be evaluated by $C^{nc}(0) - C^{nc}(\theta)$. Since $C^{nc}(\theta) \leq C_{nc}(0)$, it makes sense to consider the sensitivity of channel capacity to the perturbation as

$$\gamma \triangleq \lim_{\|\theta\| \to 0} \frac{C^{nc}(0) - C^{nc}(\theta)}{\mathrm{Var}(\theta^T Z)}, \tag{5.5}$$

where $\mathrm{Var}(\theta^T Z)$ is the variance of the perturbation. Note that equation (5.5) means that for small values of $\|\theta\|$, we have

$$C^{nc}(\theta) = C^{nc}(0) - \gamma \mathbb{E}[(\theta^T Z)^2] + o(\|\theta\|^2). \tag{5.6}$$

Note that a similar definition of sensitivity (5.5) has already been used in [PPV95b] to address the case of non-Gaussian contaminating noise in an AWGN channel. Now, using standard properties of max(.) and min(.) functions, we obtain:

$$C^{nc}(0) - C^{nc}(\theta) \leq \max_{p(u,x|s_1)} \big\{ I(U; S_2, Y) - I(U; S_2, \tilde{Y}) \big\}, \tag{5.7a}$$

$$C^{nc}(0) - C^{nc}(\theta) \geq \min_{p(u,x|s_1)} \big\{ I(U; S_2, Y) - I(U; S_2, \tilde{Y}) \big\}. \tag{5.7b}$$

Interestingly, the received signal being $\tilde{Y} = Y + \theta^T Z$, the mutual information difference $I(U; S_2, Y) - I(U; S_2, \tilde{Y}) = I(U, Y|S_2) - I(U, \tilde{Y}|S_2)$ characterizes the loss (due to the perturbation $\theta^T Z$) in the information conveyed by the auxiliary random variable U about the received signal. This "information loss" can be related to entropy using the "Information Chain Rule" and the "Entropy Chain Rule" [CT91], as follows:

$$\begin{aligned} I(U;S_2,Y) - I(U;S_2,\tilde{Y}) &= I(U,Y|S_2) - I(U,\tilde{Y}|S_2) \\ &= H(Y|S_2) - H(\tilde{Y}|S_2) \\ &\quad + H(\tilde{Y}|U,S_2) - H(Y|U,S_2). \end{aligned} \tag{5.8}$$

Noticing that $H(Y|S_2) - H(\tilde{Y}|S_2) = H(Y,S_2) - H(\tilde{Y},S_2)$ and that

$$H(\tilde{Y}|U,S_2) - H(Y|U,S_2) = H(\tilde{Y};U,S_2) - H(Y;U,S_2),$$

equation (5.8) can be rewritten as

$$I(U;S_2,Y) - I(U;S_2,\tilde{Y}) \;=\; H(Y;S_2) - H(\tilde{Y};S_2) + H(\tilde{Y};U,S_2) - H(Y;U,S_2). \tag{5.9}$$

The idea in the following is to use the Taylor expansion formula so as to expand, as a function of θ, the two differential entropy quantities $H(\tilde{Y},S_2)$ and $H(\tilde{Y};U,S_2)$. For that, we rely heavily on recent results in [Pha05] where the author provides an informal derivation of the conditions under which this Taylor expansion applies. In our case, we write $(\tilde{Y},S_2) = (Y,S_2) + \theta^T Z^{(1)}$ and $(\tilde{Y};U,S_2) = (Y;U,S_2) + \theta^T Z^{(2)}$, where $Z^{(1)} = \big((Z_1,Z_2)^T,(0,0)^T\big)$ and $Z^{(2)} = \big(Z^{(1)},(0,0)^T\big)$. Noticing that $Z^{(1)}$ and $Z^{(2)}$ are respectively independent of (Y,S_2) and (Y,U,S_2), we only need the joint distributions $p_{(Y,S_2)}(y,s_2)$ and $p_{(Y,U,S_2)}(y,u,s_2)$ be "well behaved". Namely, to expand $H(\tilde{Y},S_2)$ in θ, we only need that $p_{(Y,S_2)}\log p_{(Y,S_2)}$, $p'_{(Y,S_2)}\log p_{(Y,S_2)}$ and $p'_{(Y,S_2)}$ converge to zero at infinity, where $p'_{(Y,S_2)}$ denotes the derivative of $p_{(Y,S_2)}$. Similarly, to expand $H(\tilde{Y};U,S_2)$ as a function of θ, we only need that $p_{(Y,U,S_2)}\log p_{(Y,U,S_2)}$, $p'_{(Y,U,S_2)}\log p_{(Y,U,S_2)}$ and $p'_{(Y,U,S_2)}$ converge to zero at infinity. Assuming these reasonable assumptions to hold, we get

$$H(\tilde{Y},S_2) = H(Y,S_2) + \frac{1}{2}\mathrm{Tr}\{J(Y,S_2)\}\mathrm{Var}(\theta^T Z) + o(\|\theta\|^2), \tag{5.10a}$$

$$H(\tilde{Y};U,S_2) = H(Y;U,S_2) + \frac{1}{2}\mathrm{Tr}\{J(Y;U,S_2)\}\mathrm{Var}(\theta^T Z) + o(\|\theta\|^2). \tag{5.10b}$$

Tr(.) denotes the trace operator and $J(.)$ the Fisher information. Combining (5.9) and (5.10), we come out with the decrease in the transmission rate $R(0) - R(\theta) = I(U;S_2,Y) - I(U;S_2,\tilde{Y})$, due to the perturbation $\theta^T Z$, as

$$R(0) - R(\theta) = \frac{1}{2}\mathrm{Tr}\{J(Y;U,S_2) - J(Y,S_2)\}\mathrm{Var}(\theta^T Z) + o(\|\theta\|^2). \tag{5.11}$$

This expression can be used to predict the decrease in the transmission rate that a variation of transmission conditions of the channel in Fig.5.2 would cause, and thus, act accordingly. Further, lower and upper bounds on channel sensitivity to these state information variations, defined as in (5.5), can be obtained by using (5.11) and inverting the limit and max-min operations, as

$$\min_{p(u,x|s_1)} \mathrm{Tr}\{J(Y;U,S_2) - J(Y,S_2)\} \leq 2\gamma, \tag{5.12a}$$

$$2\gamma \leq \max_{p(u,x|s_1)} \mathrm{Tr}\{J(Y;U,S_2) - J(Y,S_2)\}. \tag{5.12b}$$

Note that these bounds on channel sensitivity γ translate to lower and upper bounds on channel capacity loss by equality (5.6), respectively. We now pause to briefly discuss the implications and usefulness of the bounds in (5.12).

1. Generally speaking (though, not always, as we will see in the Gaussian case), channel sensitivity depends on (i) the encoder strategy, through the codebook U, (ii) the nominal state information at the transmitter, through both Y and U and (iii) the nominal state information at the receiver.

2. In the classical case, i.e., when the two-sided state information pair is known perfectly, the codebook U is generally designed so as to maximize the transmission rate (or equivalently, channel capacity). This is ensured by maximizing the "information" conveyed about the received signal Y. However, this unfortunately thereby increases the Fisher information in the left hand side term of (5.12a), thus making the system more sensitive to noise. Hence, in the presence of channel state perturbations, the optimal codebook should still maximize the transmission rate, on one hand, but should also minimize channel sensitivity to these perturbations, on the other hand. In particular, one would, in some non-demanding rate applications (as in information embedding), voluntarily lower the system requirements in terms of transmission rate so as to increase its immunity to small state information perturbations. This is particularly possible when closed form expression of the sensitivity γ exists.

3. The bounds in (5.12), though may be not tight, are useful, especially in the situations where no closed form expression of the channel sensitivity γ is available. For instance, the upper bound shows how one may (at least) limit the sensitivity by devising the codebook U so as to minimize the right hand side term of (5.12b). On the other hand, the lower bound (5.12a) gives the "minimum unavoidable" loss in the transmission rate, due to the uncertainty in both the encoder and the decoder, thus permitting to predict the system performance in the "most favorable" channel condition case.

5.3 Gaussian Noise and Gaussian State

In this section we make the additional assumption that the state informations S_1 and S_2 are i.i.d normally distributed with zero means and variances Q_1 and Q_2, respectively. We also assume that the ambient Gaussian channel noise $V \sim \mathcal{N}(0, N)$ is statistically independent of S_1 and S_2. Prior to dealing with channel capacity loss, we pause to consider the non-noisy case, i.e., $\theta = 0$, and consider the nominal capacity $C^{nc}(0)$. Using straightforward algebra calculation, it can be easily shown that, in this case (i.i.d Gaussian variables and channel input satisfying an average power constraint $\frac{1}{n}\mathbb{E}[X^2] \leq P$), the AWGN capacity limit $C = 1/2 \log(1 + P/N)$ can be attained by choosing the auxiliary random variable U in the form $U = X + \alpha S_1$, where $\alpha = P/(P+N)$. Note that, as expected, both the optimal codebook U and the optimal inflation parameter α have the same expressions as the Dirty Paper Codebook (DPC) and the "Costa parameter", provided in [Cos83] in the case where only the state information S_1 is available at the encoder, i.e., no state information is available at the decoder ($S_2 = 0$). This is because, if S_2 is non zero, the receiver should simply subtract-off it before proceeding to decoding. We now turn to the case where the two-sided state information pair is subject to small perturbation, i.e., $(\widetilde{S_1}, \widetilde{S_2}) = (S_1, S_2) + \theta^T Z$. In this case, channel capacity (5.4) simplifies to

$$C^{nc}(\theta) = \frac{1}{2}\log\left(1 + \frac{P}{N + \mathrm{Var}(\theta^T Z)}\right), \tag{5.13}$$

where $\mathrm{Var}(\theta^T Z)$ represents the additional (channel) noise due to the imperfect knowledge of the state information. The optimal inflation parameter α that allows to attain this capacity is $\alpha = P/(P+N+\mathrm{Var}(\theta^T Z))$. Also, channel sensitivity to the perturbation $\theta^T Z$ can be obtained by a simple derivation of the Gaussian capacity (5.13). We obtain

$$\gamma(C) = \frac{P}{2N(P+N)}. \tag{5.14}$$

Note that this sensitivity coefficient is, in this case, independent of the two-sided state information (S_1, S_2). However, the generation of the codebook U necessitates the knowledge of the deviation θ, through α. In some situations however, the encoder may not have access to θ, or may even completely ignore the presence of the noise term $\theta^T Z$. This may occur, for example, in the situations where the channel is "upgraded" a certain time after digital communication architecture deployment. Another example is provided in Section 5.5. In these situations, the codeword X, tailored for the classical case (perturbation-free channel), faces state perturbations in the channel. Thus, in this case, it is more reasonable to evaluate the loss in the transmission rate $R(\alpha,\theta) = I(U; Y+\theta^T Z, S_2) - I(U, S_1)$, due to the perturbing noise, instead of that of capacity. Straightforward calculation gives

$$R(\alpha,\theta) = \frac{1}{2}\log\left(\frac{P(P+Q_1+N+\mathrm{Var}(\theta^T Z))}{PQ_1(1-\alpha)^2 + (N+\mathrm{Var}(\theta^T Z))(P+\alpha^2 Q_1)}\right). \tag{5.15}$$

Again, a simple derivative of $R(\alpha,\theta)$ yields $R(\alpha,\theta) = R(\alpha,0) - \gamma\theta^2 + o(\theta^2)$, where

$$\gamma(R) = \frac{1}{2}\frac{(P+\alpha Q_1)^2}{(P+Q_1+N)(PQ_1(1-\alpha)^2 + N(P+\alpha^2 Q_1))}, \tag{5.16}$$

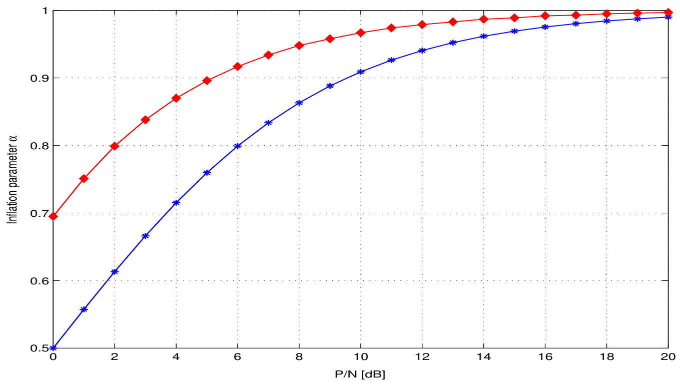
is, by opposition to $\gamma(C)$, naturally dependent on the state-information S_1 (through Q_1). The non-dependence on the decoder state-information S_2 follows the same argument as for capacity achievement. Note however that if $Q_1 \gg P$ and $Q_1 \gg N$ (which is relevant in applications such as information embedding), we have $\gamma(R) \approx P/2N(P+N) = \gamma(C)$. Hence, the channel sensitivity is maximized for high SNR $= P/N$ [dB] and the degradation is of less importance for small SNR, as illustrated in Fig.5.3(b).

We now discuss the tightness of the bounds (5.12) by evaluating them in the Gaussian case and comparing them to the explicit expressions (5.14) and (5.16). For that, we need the Fisher information matrices of the vectors involved. Recall that the Fisher information matrix $J(T)$ of a Gaussian vector T is given by the inverse of its covariance matrix, i.e., $J(T) = \mathrm{Cov}^{-1}(T)$. Hence, straightforward calculation gives

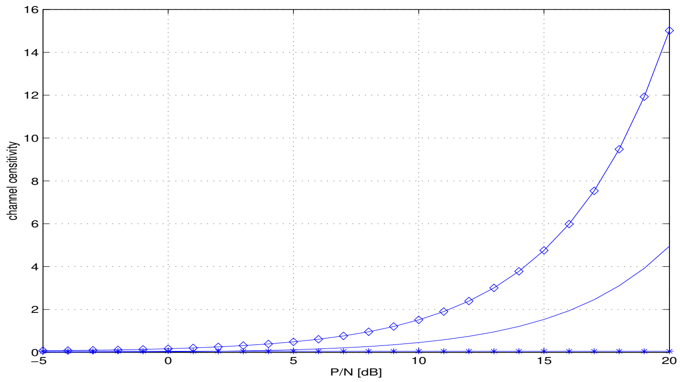
$$\mathrm{Tr}\{J(Y;U,S_2)\} = \frac{2(P+\alpha^2 Q_1) + P + Q_1 + N}{(P+Q_1+N)(P+\alpha^2 Q_1) - (P+\alpha Q_1)^2} + \frac{1}{Q_2}, \tag{5.17a}$$

$$\mathrm{Tr}\{J(Y;S_2)\} = \frac{2}{P+Q_1+N} + \frac{1}{Q_2}. \tag{5.17b}$$

Moreover, optimization over the joint distribution $p(u,x|s_1)$ in (5.12) reduces, in this case, to an optimization over the sole parameter α. In Fig.5.3(a), the capacity-achieving Costa parameter $\alpha = P/(P+N)$ is compared to the one that maximizes channel sensitivity. We observe that the Costa-parameter, which is the one that the encoder may intentionally use (cf., the encoder is aware of the perturbation, but does not take it into account) or non-intentionally use (cf., the encoder ignores the presence of the perturbation) in the generation of the codebook $U = X + \alpha S_1$, causes high sensitivity at large SNR. The



(a) Encoding parameter α.



(b) Channel sensitivity coefficient γ.

Figure 5.3: (a) Capacity-achieving parameter α (Asterisk) compared to the encoding parameter α that maximizes sensitivity to noise (Diamond). (b) Gaussian channel sensitivity (5.14) compared to lower (Asterisk) and upper (Diamond) bounds given by (5.12a): the lower bound is relatively tight at small SNR (SNR ≤ 5 [dB]), but becomes coarse at high SNR.

sensitivity-maximizing parameter in Fig.5.3(a) is obtained by maximizing the right hand side term of (5.12b). Yet, it is not that, in general, the optimal DPC strategy $U = X + \alpha S_1$ and/or the optimal Costa-parameter $\alpha = P/(P+N)$ should be changed, but rather, that increasing the transmission rate by, for example, increasing the transmission power P, inevitably increases sensitivity to state information perturbations. This (un)-avoidable "negative effect" is illustrated in Fig.5.4 and Fig.5.5, where both the incurred capacity and rate losses are plotted v.s the SNR.

We now discuss the usefulness of the bounds (5.12) in this special case, where channel sensitivity has closed form expression. Suppose that the transmitter has some (partial) knowledge of the perturbation, obtained using some "extra" mechanism (through estimation, for example). Assume, for example, that this mechanism provides an upper bound on the deviation θ, e.g., $\mathbb{E}[\theta^2] \leq N_\theta$. If the bound on $\mathbb{E}[\theta^2]$ is tight enough, the encoder should *adapt* its encoding strategy to the newly being available channel knowledge. In fact, it can be easily shown that, with respect to the approach consisting in completely ignoring the perturbation θZ, *adapting* is better if and only if

$$N_\theta \leq 2\mathbb{E}[\theta^2].$$

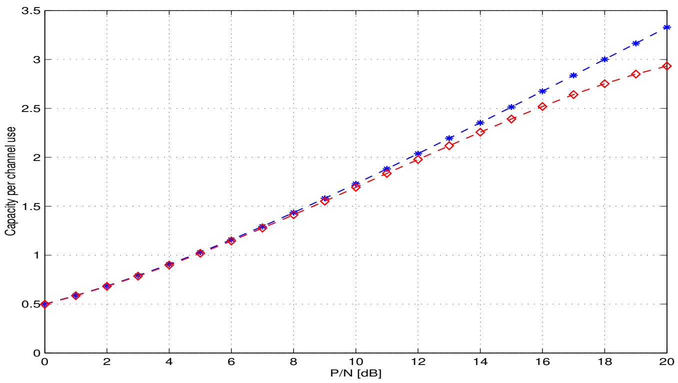


Figure 5.4: Channel capacity loss, in the Gaussian case, due to a weak contaminating perturbation of the two-sided state information. The curves represent the nominal capacity $C^{nc}(0)$ with a perfect knowledge of the two sided state information (Asterisk) and the capacity with imperfect knowledge of two sided state information (Diamond). Perturbation parameters are set to $\mathbb{E}[Z^2] = Q_1/100$ and $\theta = (0.1, 0.1)^T$. Capacity loss is independent of the two-sided state information.

In this case, substantial gain is provided by setting the inflation parameter α to its (new) optimal choice $\alpha_{opt} = P/(P + N + N_\theta)$. Note that here we have assumed that Z has unit variance. Consequently, the encoder may, likewise, limit the loss due the channel uncertainty.

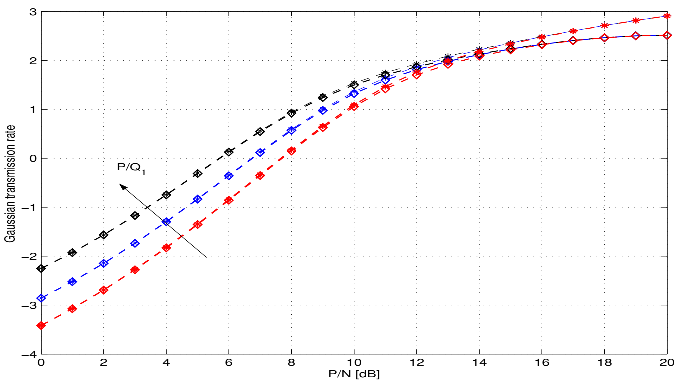


Figure 5.5: Transmission rate loss, in the Gaussian case, due to a weak contaminating perturbation of the two-sided state information. The curves represent the transmission rate with perfect (Asterisk) and imperfect (Diamond) knowledge of the two-sided state information. The stronger the state information S_1 at the encoder, the larger the loss in the transmission rate.

5.4 State Information at the Encoder

In this section, we consider the case of one-sided state information S_1, available at the transmitter. As special case of the general channel model considered in Section 5.2, both channel capacity and transmission rate sensitivities, considered earlier in [ZD05b], can be obtained by omitting the terms S_2 and Q_2 in (5.12) and (5.17). Alternatively, the same results can be obtained through a different approach. The basic steps are outlined hereafter. The received signal is $\tilde{Y} = X + S_1 + \theta Z + V = Y + \theta Z$, where, in this case, the parameter θ is a scalar variable ($\theta = \theta_1$) and $Z = Z_1$. Also, we assume, for convenience, that the variance of the perturbation Z is normalized to unity. Following the same principles as in Section 5.2 and defining the channel sensitivity γ as in (5.5), we get, after straightforward calculation,

$$\min_{p(u,x|s_1)} \left\{ \lim_{\theta \to 0} \frac{H(Y) - H(\tilde{Y})}{\theta^2} - \lim_{\theta \to 0} \frac{H(Y,U) - H(\tilde{Y},U)}{\theta^2} \right\} \leq \gamma, \tag{5.18a}$$

$$\gamma \leq \max_{p(u,x|s_1)} \left\{ \lim_{\theta \to 0} \frac{H(Y) - H(\tilde{Y})}{\theta^2} - \lim_{\theta \to 0} \frac{H(Y,U) - H(\tilde{Y},U)}{\theta^2} \right\}. \tag{5.18b}$$

Now, recall De Bruijn identity which relates entropy $H(.)$ to Fisher information $J(.)$.

Lemma 2 *(De Bruijn Identity [CT91])*
Let X be any random variable with finite variance and density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance. Then

$$\frac{\partial}{\partial t} H(X + \sqrt{t}Z) = \frac{1}{2} J(X + \sqrt{t}Z). \tag{5.19}$$

Using (5.19), bounds in (5.18) reduce to

$$\min_{p(u,x|s_1)} \mathrm{Tr}\{J(Y,U) - J(Y)\} \leq 2\gamma, \tag{5.20a}$$

$$2\gamma \leq \max_{p(u,x|s_1)} \mathrm{Tr}\{J(Y,U) - J(Y)\}, \tag{5.20b}$$

which is, as already mentioned, a special case of (5.12).

5.5 Extension: Causal State Information

In this section, we show that the results above can be extended to the case where the state information is known only causally to the transmitter. This case is relevant for applications where the state information is obtained through a feedback strategy (e.g., fading coefficients), for example. In such situations, the capacity (2.9) becomes [Sha58]

$$C_{10}^{c}(\theta) = \max_{p(u)p(x|u,s_1)} I(U; \tilde{Y}). \tag{5.21}$$

The auxiliary random variable U satisfies $|\mathcal{U}| \leq |\mathcal{Y}|$ and the joint distribution $p(u, x, s_1, \tilde{y})$ is such that

$$p(u, x, s_1, \tilde{y}) = \begin{cases} p(s_1)p(u)p(\tilde{y}|x, s_1) & \text{if } x = f(u, s_1) \\ \\ 0 & \text{otherwise} \end{cases}. \tag{5.22}$$

In order to see how the bounds (5.20) on channel sensitivity translate to the causal case, we first notice that the random variables U and S_1 are independent (and hence $I(U; S_1) = 0$) under the joint distribution (5.22). Thus the expression to be maximized is the same as in the non-causal case (5.4). Namely, $I(U, \tilde{Y}) = I(U, \tilde{Y}) - I(U, S_1)$. The only minor difference is that we are maximizing over a smaller set of distributions. Consequently, defining the (causal) channel sensitivity as in (5.5) results in the same expressions for the sensitivity bounds. Only the set of admissible distributions over which these expressions are minimized or maximized is changed (reduced in the causal-case). The resulting bounds are given by

$$\min_{p(u)p(x|u,s_1)} \mathrm{Tr}\{J(Y,U) - J(Y)\} \leq 2\gamma, \tag{5.23a}$$

$$2\gamma \leq \max_{p(u)p(x|u,s_1)} \mathrm{Tr}\{J(Y,U) - J(Y)\}. \tag{5.23b}$$

5.6 Applications and Practical Usefulness

We consider two applications. The first application is from conventional communication, not from information embedding. However, this is considered here to strengthen the analogy between information embedding and classical communication that has been followed up throughout this thesis. In this application, the state information is known causally and the vector channel is equivalent (by nature) to an instance of that in Fig.5.2 (no state information at the decoder). The second application is from information embedding. The state information is known non-causally. These two applications are not deeply investigated. The aim is to just illustrate the principles discussed in the sections above.

5.6.1 Communication over channels with fading

Consider the transmission from a base station (BS) with N_T transmit antennas to $K \leq N_T$ users. Using the vector/matrix notation, the received signal can be written as $\tilde{\mathbf{y}} = \mathbf{H}\mathbf{x} + \mathbf{v}$ with $\mathbf{x} = [x_1, \ldots, x_{N_T}]^T$, $\mathbf{v} = [v_1, \ldots, v_{N_T}]^T$ and $\mathbf{H} = [h_{i,j}]$. Obviously, the channel coefficients $h_{i,j}$ are time-dependent. We suppose that channel state variation is indicated by the superscript i. We write $\tilde{\mathbf{y}} = \mathbf{H}^{(i)}\mathbf{x} + \mathbf{v}$ to denote the received signal under the channel state $\mathbf{H}^{(i)}$. Suppose that the receiver first performs an estimate $\widehat{\mathbf{H}^{(i)}}$ of the current realization $\mathbf{H}^{(i)}$ of channel state and then transmits it to the transmitter using some backward feedback loop. The transmitter uses this newly available (partial) channel knowledge as state information to combat channel interference. If channel variation is slow, the nominal current state $\mathbf{H}^{(i)}$ can be written as the sum of the estimation $\widehat{\mathbf{H}^{(i)}}$ and a noise-like estimation error $\theta^T\tilde{\mathbf{H}}^{(i)}$,

$$\mathbf{H}^{(i)} = \widehat{\mathbf{H}^{(i)}} + \theta^T\tilde{\mathbf{H}}^{(i)}.$$

Thus, the overall vector channel is equivalent to

$$\tilde{\mathbf{y}} = \widehat{\mathbf{H}^{(i)}}\mathbf{x} + \mathbf{v} + \theta^T\tilde{\mathbf{H}}^{(i)}\mathbf{x},$$

which is in the form discussed above, with nominal state information $S_1 = \widehat{\mathbf{H}^{(i)}}$ and small perturbation $\theta^T Z = \theta^T\tilde{\mathbf{H}}^{(i)}\mathbf{x}$. If, due to channel fluctuations or any other perturbing phenomenon at the receiver, the accuracy of the estimation varies, the causal channel capacity varies accordingly as $C^c(\theta) = C^c(0) - \gamma\theta^2 + o(\theta^2)$, where $C^c(0)$ corresponds to the case where the estimation is error-free. If the perturbation $\theta^T\tilde{\mathbf{H}}^{(i)}\mathbf{x}$ is somehow controlled, the encoder should adapt to the situation by changing its strategy, accordingly. The bounds provided in (5.23) on the sensitivity γ to the estimation error $\theta^T\tilde{\mathbf{H}}^{(i)}$ give, as already mentioned, means of predicting system performance in the "most accurate" and the "least accurate" channel estimation cases.

5.6.2 Information Embedding under channel desynchronization

Consider the channel in Fig.5.6 where a message $m \in \mathcal{M}$ has to be sent to some receiver over a noisy channel (a digital watermark channel, for example). In the classical case, the transmitter has full access to the cover (called also host) signal and uses it as state information at the encoder. In some other situations of interest, the transmitter has access to a simplified version of this state, only. This occurs, for example, in high secure transmissions. Here, we consider the case where only a quantized version $\hat{\mathbf{s}}$ of the cover signal $\mathbf{s}$ is made available at the transmitter. In this situation, the encoder forms the channel codeword $\mathbf{x}$, based (only) on the nominal state $\hat{\mathbf{s}}$. The full state $\mathbf{s}$ can be viewed as the sum of $\hat{\mathbf{s}}$ and some quantization error $\mathbf{e}$. In addition, we consider the general case where the composite signal $\mathbf{x}+\mathbf{s}$ is subject to some time-delay desynchronization, in the channel. The resulting desynchronized transmission may be modeled by

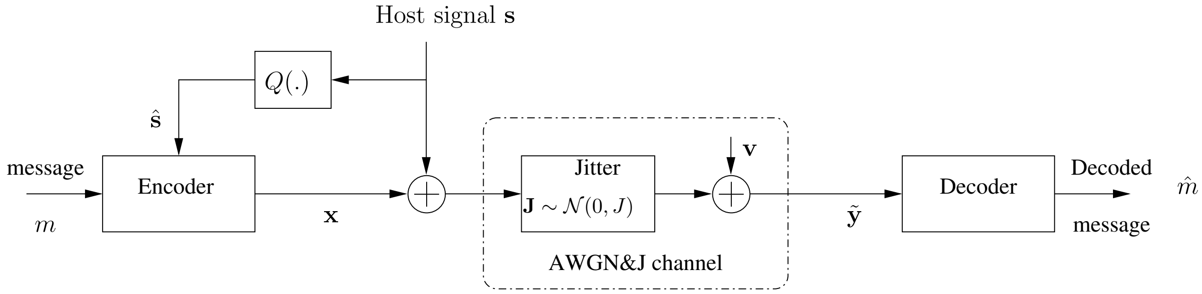


Figure 5.6: Information embedding using quantized (version of) host signal over an AWGN and Jitter (AWGN&J) channel.

transmission over an Additive White Gaussian Noise and Jitter channel (AWGN&J) [**?**]. An AWGN&J channel is an AWGN channel in which the signal $\mathbf{s}+\mathbf{x}$ and the i.i.d Gaussian noise $\mathbf{v}$ are randomly sampled. More precisely, the received signal at time nT is $\tilde{y}[n] \triangleq \tilde{y}[nT] = x[nT+\tau] + s[nT+\tau] + v[nT]$ where the delay $\tau = \delta T$ is a fraction of the sampling period T. The deviation $\delta \in [0,1]$ is a realization of some random process (jitter), assumed to be Gaussian $\mathbf{J} \sim \mathcal{N}(0, J)$, at time nT. Writing the received signal as $\tilde{y}[n] = x[n]+\hat{s}[n]+v[n]+e[n]+\tau\frac{d}{dt}(x+s)(t)|_{t=nT}$, $\tilde{\mathbf{y}}$ yields the form discussed above, with $S_1 = \hat{\mathbf{s}}$, $\theta = \tau$ and $Z_1 = \frac{\mathbf{e}}{\tau} + \dot{\mathbf{x}} + \dot{\mathbf{s}}$. These two phenomena (partial knowledge of the host signal at the transmitter and perturbation of nominal sampling instants in the channel) lead then to a (total) capacity loss due to the (total) contaminating noise-like term $\mathbf{e}+\tau(\dot{\mathbf{x}}+\dot{\mathbf{s}})$. The effect of the jitter in reducing channel capacity is depicted in Fig.5.7 where the quantization error is not considered. Depending on the jitter strength, the resulting information embedding capacity $C^{nc}(J)$ decreases as $C^{nc}(J) = C^{nc}(0) - \gamma \mathrm{Var}(Z_1) + o(J)$, where γ is given by (5.14). Here again, if, by means of some process, the jitter is (partially) controlled, the encoder should adapt to the situation.

5.7 Summary

In this chapter, we studied the influence of a weak noise-like perturbation on a pair of two-sided nominal state information (S_1, S_2), made available at the transmitter and at the receiver, respectively. We first considered the non-causal case and use Fisher information to provide both lower and upper bounds on

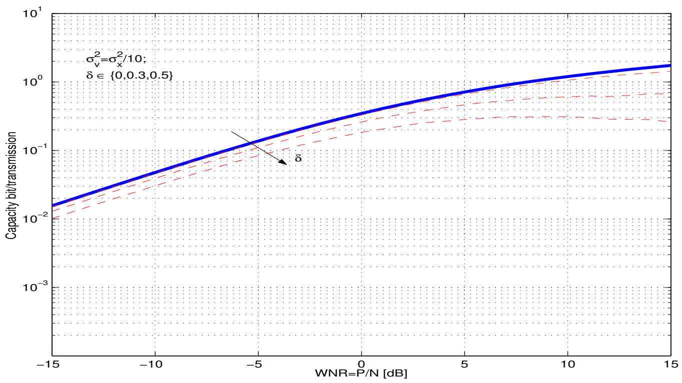


Figure 5.7: Capacity loss of DPC-based information embedding under the influence of a time delay desynchronization $\tau = \delta T$ and additive noise $\mathbf{v}$. The variance of the embedded signal, σ_x^2, varies according to $\sigma_x^2 = 10^{\mathrm{SNR}/10}\sigma_v^2$. The upper curve represents the capacity $C^{nc}(0) = 1/2\log_2(1+P/N)$ of the perfectly synchronized channel.

channel sensitivity to the weak contaminating noise. We also discussed the tightness and the usefulness of these bounds through comparison with the Gaussian case, for which we gave closed form expression. In particular, we showed that, in some situations, the encoder should *adapt* to the imperfect-knowledge of the channel, by changing its encoding strategy, so as to increase system immunity to noise. Next, we extended these results the case of channel with causal channel state at the transmitter. Finally, two illustrative applications have been discussed.

Chapter 6

Information Embedding over an AWGN&J Channel

The content of this chapter has been partially published in [ZBD04, ZBD05, **?***].*

In the previous chapters, we concentrate on the design of efficient coding and decoding strategies for the problem of information embedding. Efficient coding techniques are those which approach channel capacity, in a point-to-point (single user) communication, or capacity region frontier in a multi-user environment. Until now, we used the classical definition of channel capacity, i.e., the supremum of all achievable rates. However, channel capacity has an alternative definition based on a min-max optimization problem. As information embedding is basically a communication problem, capacity can be defined in a similar way. Further, this alternative definition of channel capacity makes more sense in this case, for the channel may comprises an embedder (encoder) and an attacker (channel) whose goals are opposite, by nature. However, while the distortion constraint involved in the min-max problem can be expressed simply by the norm of the channel noise in classical transmissions situations, a more accurate distortion measure has to be found, in information embedding. This is a *perceived* distortion measure, meaning a distortion measure that characterizes the perceptual loss in quality due to transmission over the channel. In this chapter, we first provide a simple measure of this perceived distortion. Next, we use this perceived distortion measure in assessing the impact that a scale plus additive noise channel attack, modeled by transmission over an AWGN&J channel, has in reducing channel capacity. Channel capacity as well as optimal "embedder" and "attacker" strategies are provided in a min-max game theory context.

6.1 Min-max and Max-min in Classical Communication

Max-min and min-max optimization problems have large use in classical communication. This can be seen by simply saying that, usually, the transmitter wants a guaranteed rate of reliable transmission under any channel distortion. In fact, even in the simple situation where all signals are Gaussian, a max-min optimization problem exists, but is somewhat implicit. Consider the vector channel depicted in Fig.6.1 where the input $\mathbf{s}$, the output $\mathbf{y}$ and the channel noise $\mathbf{z}$ are vector-valued Gaussian signals with covariance matrices K_{xx}, K_{yy} and K_{zz}. The input must satisfy a power constraint $\mathbb{E}[\mathbf{X}\mathbf{X}^T] \leq P$, which translates to a constraint on the input covariance matrix

$$\mathrm{Tr}(K_{xx}) \leq P. \tag{6.1}$$

Maximization is explicitly present, by virtue of Shannon's noisy channel coding theorem for a discrete

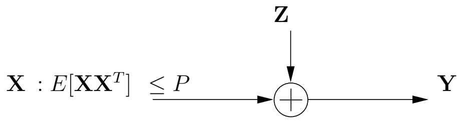


Figure 6.1: An abstract communication model for constrained classical transmission

memoryless channel. This implies that $C = max\ I(\mathbf{X};\mathbf{Y})$, where maximization is over all input distributions that satisfy the power constraint (6.1). The maximum mutual information is achieved with Gaussian inputs, and in this case the channel capacity can be evaluated as

$$C = \max_{K_{xx}} \frac{1}{2} \log \frac{|K_{xx} + K_{zz}|}{|K_{zz}|}, \tag{6.2}$$

where $|\cdot|$ denotes matrix determinant and the covaiance matrix K_{xx} is such that (6.1) is satisfied. Now, note that for any non Gaussian noise $\mathbf{Z}'$, $I(\mathbf{X}, \mathbf{X}+\mathbf{Z}) \leq I(\mathbf{X}, \mathbf{X}+\mathbf{Z}')$. This means that channel capacity (6.2) can be re-written as

$$\begin{aligned} C &= \max_{p(\mathbf{s})} I(\mathbf{X}; \mathbf{X}+\mathbf{Z}) \\ &= \max_{p(\mathbf{s})} \min_{p(\mathbf{z}')} I(\mathbf{X}; \mathbf{X}+\mathbf{Z}'), \end{aligned} \tag{6.3}$$

where the problem of achieving capacity can be viewed as a game between the encoder (through the distribution of the input) and the channel (through the distribution of the noise). More involved use of game-theory in traditional communication can be found in, for example, competitive and coopertative multi-user environments. For instance, the search for saddlepoints in broadcast channels relies on game-theory [Yu02]. In information embedding, the usage of game theory is more extensive, because of the nature of the channel.

6.2 Min-max and Max-min in Information Embedding

Consider a generic information embedding system aiming at embedding a message $m \in \mathcal{M}$ into a host signal $\mathbf{s} = (s_1, \cdots, s_n)$ as shown in Fig.6.2. Based on the host signal $\mathbf{s} \in \mathcal{S}^n$, the message $m \in \mathcal{M}$ and

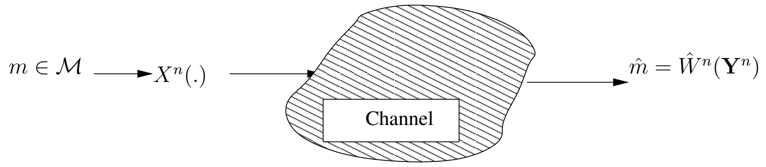


Figure 6.2: An abstract communication model for information embedding

eventually some private key $\mathbf{k} \in \mathcal{K}^n$, the embedder designs an encoding function $X^n : \mathcal{S}^n \times \mathcal{M} \times \mathcal{K}^n$ and a decoding function $\hat{W} : \mathcal{Y}^n \times \mathcal{K}^n$ such that:

1. The embedded signal $\mathbf{x} = X^n(\mathbf{s}, m, \mathbf{k})$ satisfies the embedding distortion
$$\sum_{\mathbf{s}\in\mathcal{S}^n} \sum_{\mathbf{k}\in\mathcal{K}^n} \sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{M}|} d_e^n\left(\mathbf{s}, X^n(\mathbf{s}, m, \mathbf{k})\right) \leq D_E, \tag{6.4}$$
where the non-negative function $d_e : \mathcal{S} \times \mathcal{X} \longrightarrow \mathbb{R}_+$ denotes the distortion function for the embedder.

2. Embedding performance are maximized, meaning that a certain *payoff* function $\mathcal{F}(.)$ should be maximized.

Conversely, an eventual channel attacker, subject to distortion D_a, processes the composite signal $\mathbf{c} \triangleq \mathbf{s} + \mathbf{x}$ so as to fool the receiver. This amounts to designing a sequence of conditional probability mass functions $\Theta^n(\mathbf{y}|\mathbf{x})$ from $\mathcal{X}^n$ to $\mathcal{Y}^n$ such that:

1. The induced distortion does not exceed D_a, i.e.,
$$\sum_{\mathbf{x}\in\mathcal{X}^n} \sum_{\mathbf{y}\in\mathcal{Y}^n} d_a^n\left(\mathbf{x}, \mathbf{y}\right) \Theta^n(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \leq D_a, \tag{6.5}$$

2. Embedding performace are minimized, meaning that the payoff function should be minimized.

Hence, information embedding can be thought of as a game between two cooperative players (the encoder and the decoder) and an opponent (the attacker). A natural choice for this payoff function would be the probability of error $P_e(\hat{m} \neq m)$, averaged over $m \in \mathcal{M}$, i.e.,

$$P_e = \frac{1}{|\mathcal{M}|} \sum_{m'\in\mathcal{M}} \Pr[\hat{W}(\mathbf{y}, \mathbf{k}) \neq m' | m = m']. \tag{6.6}$$

Another classical choice would be the maximum achievable rate of reliable transmission $R = \frac{1}{n} \log |\mathcal{M}|$. A *Nash equilibrium* [MO03] $(X^{*n}, \Theta^{*n}, \hat{W}^*)$ of the game is obtained if and only if

$$\mathcal{F}\left(X^n, \Theta^{*n}, \hat{W}\right) \leq \mathcal{F}\left(X^n, \Theta^n, \hat{W}\right) \leq \mathcal{F}\left(X^{*n}, \Theta^n, \hat{W}^*\right) \tag{6.7}$$

for all admissible strategies $(X^n, \Theta^n, \hat{W})$. Under some conditions, a Nash equilibrium is also a saddle-point. The falue of the game in this case is $\mathcal{F}\left(X^{*n}, \Theta^{*n}, \hat{W}^*\right)$. If however, which is more secure, the embedder assumes that the decoder will be unable to learn the attack $\Theta^n(.)$, the value of the game is

$$f^* = \max_{X^n(.), \hat{W}(.)} \min_{\Theta^n(.)} \mathcal{F}\left(X^n, \Theta^n, \hat{W}\right) \tag{6.8}$$

In the rest of this chapter, we first provide means of evaluating channel distortions involved in the optimization problem. We then evaluate the capacity loss of common information embedding systems when facing an important class of channel attaks, amplitude scaling plus additive noise. Analysis is specialized to the situation when communication can be modeled by transmission over an Additive White Guassian Noise and Jitter (AWGN&J) channel. The second part of this chapter concentrates on finding optimal embedder (encoder) and attacker (channel) strategies. The payoff function is the detection probability and embedding is based on Spread Spectrum. The embedder wants to reliably transmit information, under any distortion constrained channel attack strategy. Conversely, the attacker wants to impair this transmission for any power constrained information embedding strategy.

6.3 The Watermark Channel and Its Model

The classical communication channels (BSC, AWGN, Rayleigh,...) are not likely to accurately model a watermark channel in real world scenarios. A better understanding of the watermark channel can be achieved by considering attacks not through their nature but through their impact on the composite signal: attacks on the cover signal can in general be modeled easily by filtering plus additive noise. In a general setting, a straightforward model may involve a signal dependent noise. That is, the noise may be highly correlated with the cover signal. This chapter studies a special case of this filtering plus noise channel, and provides some tools for increasing its usefulness (through a noise decorrelation process). The proposed approach is then used to focus on desynchronization attacks. Research to assess the impact of desynchronization attacks in digital watermarking has been carried out in two different directions:

(i) Some watermarking methods attempt to overcome desynchronization attacks by embedding the watermark in an "*invariant domain*" as in [OP97] and [Kut97],

(ii) Other schemes are based on an estimation of the attack parameters followed by a compensation as in [KM02b].

The AWGN&J channel was introduced by Baggen [Bag93] in the context of data storage applications to study the effect of timing jitter in the capacity of magnetic recording media. Insights from this model are used in this chapter to investigate the effect of desynchronization attacks on several watermarking schemes.

6.3.1 A distortion model for a watermark attack

Let m be the message to be transmitted. m is usually first encoded into a watermark $\mathbf{x}$ and then embedded into the cover signal $\mathbf{s}$. The resulting composite (composite) signal is $\mathbf{c} = \mathbf{s}+\mathbf{x}$. Consider then a general attack $\mathcal{A}$ over the watermark channel. In an attempt to fool the receiver, the attacker may use a set of admissible attack parameters $\{\theta_1, \theta_2, \ldots, \theta_p\}$ from some finite domain Θ. The attacker processes the signal $\mathbf{c}$ in such a way that the received signal $\mathbf{y}$ is given by $\mathbf{y} = \mathcal{A}_{(\theta_1,\theta_2,\ldots,\theta_p)}(\mathbf{c})$. Equivalently, the received (attacked) signal $\mathbf{y}$ can be written as the sum of the composite signal $\mathbf{c}$ and an interfering signal $\mathbf{z} = \mathbf{y} - \mathbf{c}$. Of course, $\mathbf{z}$ is $(\theta_1, \theta_2, \ldots, \theta_p)$- dependent and it fully characterizes the attack $\mathcal{A}$, i.e,

$$\mathbf{z} = \mathcal{A}_{(\theta_1,\theta_2,\ldots,\theta_p)}(\mathbf{c}) - \mathbf{c}. \tag{6.9}$$

Thus, the watermarking system can be modeled as depicted in Fig.6.3. The distortion resulting from the channel attack is generally measured by

$$D_a \triangleq \|\mathbf{y} - \mathbf{c}\| = \|\mathbf{z}\|. \tag{6.10}$$

After the channel attack, the composite signal must remain of sufficient quality. Thus, the channel attack $\mathcal{A}(\theta_1, \theta_2, \ldots, \theta_p)$ has to be upper bounded by a maximum distortion D_{amax}. Clearly, "sufficient quality" should correspond to a perceived distortion, but is often measured by (6.10). This results in

$$\|\mathbf{z}\| \leq D_{amax}. \tag{6.11}$$

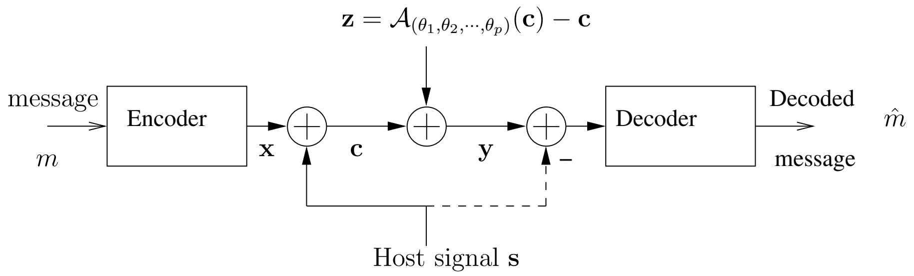


Figure 6.3: An abstract communication model for blind (solid line) and non-blind (dashed line) watermarking.

6.3.2 Outline of our approach

There are two problems with the classical channel description using the difference signal as given in Section 6.3.1. First, denoting this difference signal $\mathbf{z}$ as "noise" is not always accurate: $\mathbf{z}$ may contain parts of the composite signal $\mathbf{c}$. In such a situation, $\mathbf{z}$ should not be treated as independent noise. Also, "useful" components of $\mathbf{z}$, i.e those which are highly correlated with the desired signal $\mathbf{c}$ must not be counted as noise and should be considered as "useful". Second, the distortion measure D_a does not perceptually characterize the attack $\mathcal{A}$ effect on the composite signal $\mathbf{c}$. To cope with these problems, one can note that the attacker effect, that is the additional signal $\mathbf{z}$, can be decomposed into two parts: one which is correlated with the desired signal $\mathbf{c}$ and one which is not. The first part is somehow useful and should be "included" in the desired signal $\mathbf{c}$. The second being decorrelated with $\mathbf{c}$ can be reasonably considered as noise and will be denoted as "attacker noise" hereafter. The overall approach is equivalent to removing from the signal $\mathbf{z}$ the part that is correlated with the composite signal and characterizing the attack $\mathcal{A}$ by the remaining part only, i.e the attacker noise. One straightforward advantage is that the attacker induced perceived distortion is, likewise, readily measured by the energy (or power) of the "noise" part. This decorrelation-based approach was used previously to model quantization noise (when the high resolution assumption is not valid). More formally, our proposal is to use a "scale plus additive noise" channel model, and impose the noise to be uncorrelated with the host signal:

$$\mathbf{y} \quad = \quad k_z\mathbf{c} + \mathbf{n}_z \text{ under the constraint that } E(\mathbf{c}\bar{\mathbf{n}}_z) = 0. \tag{6.12}$$

Coefficient k_z is easily obtained by imposing $E(\mathbf{n}_z\bar{\mathbf{c}}) = k_z E(\mathbf{c}\bar{\mathbf{c}}) + E(\mathbf{n}_z\bar{\mathbf{c}}) = 0$, which gives

$$k_z = \frac{E(\mathbf{y}\bar{\mathbf{c}})}{E(\mathbf{c}\bar{\mathbf{c}})}. \tag{6.13}$$

The residual noise $\mathbf{n}_z$ is then given by

$$\mathbf{n}_z = \mathbf{y} - \frac{E(\mathbf{y}\bar{\mathbf{c}})}{E(\mathbf{c}\bar{\mathbf{c}})}\mathbf{c}. \tag{6.14}$$

Note that, disregarding the value-metric scaling coefficient k_z, the resulting model (6.12) is additive -just like that given by (6.9), $\mathbf{y} = \mathbf{c} + \mathbf{z}$. The main difference however consists in the fact that unlike signal $\mathbf{z}$, $\mathbf{n}_z$ is uncorrelated with $\mathbf{c}$. Also, in contrast to some recent watermarking related works where specific attacks are addressed as in [EBG02a], [MI03] and [CL02a], we proceed differently here: given a general attack $\mathcal{A}$ which processes the signal $\mathbf{c}$ in such a way that the received signal is $\mathbf{y} = \mathcal{A}_{(\theta_1,\theta_2,\ldots,\theta_p)}(\mathbf{c})$, we begin writing this received signal as $\mathbf{y} = \mathbf{c} + \mathbf{z}$. Next, we derive coefficient k_z and signal $\mathbf{n}_z$ according to (6.13) and (6.14) such that the constraint in (6.12) is satisfied. As a result, the decorrelation process results in a model, (6.12), that is apparently common at first glance (i.e. of the form $\mathbf{y} = A\mathbf{c} + \mathbf{v}$). However, important differences are that (i) parameters k_z and $\mathbf{n}_z$ are not "explicit" in the channel attack and, (ii) they depend on the transmitted signal $\mathbf{c}$ itself. Another fundamental difference comes from the fact that if ever the signal $\mathbf{v}$ involves a part that is correlated with $\mathbf{c}$, the communication model will remove it and include it with composite signal $\mathbf{c}$. The above model will be shown to be particularly useful with desynchronization attacks. Note also that the subscript z in k_z and $\mathbf{n}_z$ is used to point out the model parameters dependency on the attack $\mathbf{z}$ as clearly shown by (6.13) and (6.14). For convenience, we will simply use k and $\mathbf{n}$ to characterize the channel attack every time no ambiguity is possible.

6.3.3 Objective and perceived distortion measure

A very simple computation allows the computation of the "error signal" variance in terms of the "noise signal" variance, which in most circumstances is much smaller. Due to decorrelation between $\mathbf{c}$ and $\mathbf{n}$, the objective distortion defined by (6.10) becomes

$$D_a \triangleq \|\mathbf{y} - \mathbf{c}\| = |k-1|^2\sigma_c^2 + \sigma_n^2. \tag{6.15}$$

Thus, the communication model (6.12) shows a scale factor k (a luminance change for images, a sound level change for audio signals) and an additive noise $\mathbf{n}$. Both the scaling and the noise inhibit reliable detection of the watermark at the receiver side. But only the noise $\mathbf{n}$ should be considered in evaluating host signal quality loss. Consequently, we assume in this chapter that, for the perspective of a perceived distortion measure, the scale factor does not contribute to the distortion. Hence, rather than assuming that the MSE (the norm of the error signal) is a good model for the perceived distortion, we shall use σ_n^2 to cope with the perceived distortion. Obviously, more accurate models exist, involving human perception models, but the model (6.12) seems to be a good trade-off between accuracy and tractability. Simulation results based on real audio signals in the presence of desynchronization attacks show its accuracy.

6.4 AWGN&J Channel Classical Model

In this section, after a short presentation of the AWGN&J channel, AWGN&J desynchronization effects are investigated differently: (i) using common Inter Symbol Interference (ISI) assumptions commonly known the communication theory in Section 6.4.1, and (ii) using the model (6.12) in Section 6.4.2. Both approaches are finally compared. In other words, we will compare the distortions resulting from the two writings of the jittered signal[1]. This comparison will confirm the accuracy of the model (6.12) and underline its particular usefulness for desynchronization attacks characterization.

An AWGN&J channel is an AWGN channel in which the signal $\mathbf{c}$ is, in addition so the i.i.d Gaussian noise $\mathbf{v}$, are sampled randomly. More precisely, the receiver has to decide on the presence of the watermark based on $c_J[n] + v[n] = c[nT + \tau] + v[n]$ rather than $c[n] + v[n] = c[nT] + v[n]$. The delay τ can be larger than one sampling period T. But, in most cases, the receiver can compensate for any time shifts multiple of T with relatively easy re-synchronization procedures. A very easy method will be described in Section 6.5.1. In the following, we assume that $\tau = \delta T$ is a fraction of the sampling period T, i.e. $\delta \in [0, 1]$. The deviation δ is a realization of the process $\mathbf{J}$ at time nT and $\mathbf{J}$ is assumed to be Gaussian, $\mathbf{J} \sim \mathcal{N}(0, J)$. Depending on the desynchronization (constant shift or random sampling), δ can either be random or constant. Both cases are addressed hereafter. The resulting watermarking communication

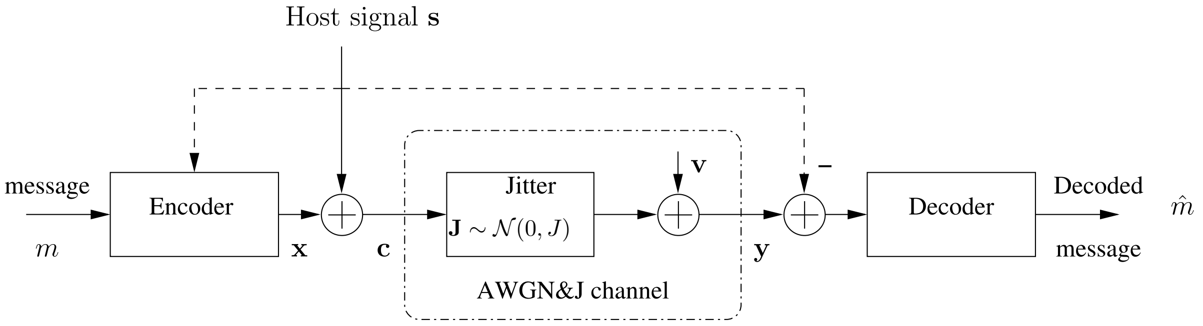


Figure 6.4: Additive White Gaussian Noise and Jitter channel AWGN&J.

over an AWGN&J channel is similar to that described in Section 6.3 except that, this time, the composite signal $\mathbf{c}$ is replaced by $\mathbf{c}_J$ such that $\mathbf{y} = \mathbf{c}_J + \mathbf{v}$. The jittered signal $\mathbf{c}_J$ will be denoted by $\mathbf{c}_f$ in case of a constant scaling and by $\mathbf{c}_r$ in case of random sampling.

Some studies of desynchronization attacks using the AWGN&J channel model, [LOJPG03, LOJH03, PGD04] or not [BEH02], already exist. However, in these works, the *desynchronization noise* is expressed using the Inter-Symbol-Interference (ISI) term and is assumed to be uncorrelated with the composite signal. This assumption, while valid in a traditional communication context, cannot hold in the context of watermarking due to the correlation of signals. Instead, this ISI term must first be processed to remove from it the part that is correlated with the composite signal $\mathbf{c}$. Only after that, the remaining part can be assumed to be noise-like. This is a straightforward application of the model (6.12) above. Using this model will shed light on AWGN&J desynchronization and will highlight the inaccuracy of the classical

[1]Note that a plain comparison consists in comparing the distortions resulting from writing the received signal as (i) $\mathbf{y} = \mathbf{c}_J + \mathbf{v}$ and (ii) $\mathbf{y} = k\mathbf{c} + \mathbf{n} + \mathbf{v}$, which amounts to comparing $\mathbf{c}_J$ to $\mathbf{y} = k\mathbf{c} + \mathbf{n}$.

ISI approach. The latter approach is described in the following Section.

6.4.1 An ISI approach to AWGN&J channel desynchronization

Under appropriate band-limited assumptions, the time-continuous signal $c(t)$ can be reconstructed without error from the sequence $\{c[n]\}_{n\in\mathbb{Z}}$ according to *Shannon-Nyquist interpolation* :

$$c(t) = \sum_{n\in\mathbb{Z}} c[n]\mathrm{sinc}(\frac{t}{T} - n). \tag{6.16}$$

This expression will be used to derive expressions for the desynchronization noise and induced distortions in presence of a jitter. Whenever required, indexes f and y will refer respectively to fixed and random jitters. Eq. (6.16) can be put in the form

$$c_J[n] = \mathrm{sinc}(\delta)c[n] + \sum_{k\in\mathbb{Z}\setminus\{n\}} c[k]\mathrm{sinc}(n - k + \delta). \tag{6.17}$$

This equality shows that introducing a constant time shift is equivalent to filtering the composite signal or alternatively, to first, attenuating the composite signal $c(t)$ and then adding a signal dependent noise $z_f(t)$ given by

$$z_f(t) = \sum_{k\in\mathbb{Z}\setminus\{n\}} c[k]\mathrm{sinc}(n - k + \delta). \tag{6.18}$$

The signal $z_f(t)$ can be seen in the context of digital communication as the Inter-Symbol Interference (ISI) term. Moreover, in case of a constant shift, scaling does not change the overall energy of the signal $c(t)$. Thus, under the uncorrelation hypothesis assumed in [BEH02] and using (6.17), the distortion due to adding the signal $z_f(t)$ can be written as

$$\sigma^2_{z_f} = (1 - \mathrm{sinc}(\delta)^2)\sigma^2_c. \tag{6.19}$$

In the case of random re-sampling, the variable δ is random. The corresponding distortion can be expressed as in (6.19) with an additional expectation over all possible values of δ. A much simpler alternative expression of $s_r[n]$ can be obtained by using the *Taylor-Young* series expansion around τ. At first order, $c_r(t) = c(t) + \tau\frac{d}{dt}c(t)$. The effect of the jitter can then be viewed as the introduction of an additional signal $z_r(t)$ given by

$$z_r(t) = \tau\frac{d}{dt}c(t). \tag{6.20}$$

Clearly, the signal $z_r(t)$ depends on the compositesignal $c(t)$. The corresponding distortion is given by

$$\sigma^2_{z_r} = JE\left([\frac{d}{dt}s(t) + \frac{d}{dt}x(t)]^2)\right). \tag{6.21}$$

Note that the ISI signals $\mathbf{z}_f$ and $\mathbf{z}_r$ arise directly from interpolation in case of constant shift and random re-sampling respectively. Hence, a priori, these terms are not necessarily decorrelated from $\mathbf{c}$. An additional *decorrelation process* (as described in the model (6.12)) is needed to extract the corresponding *noise* parts[2]. However, in this section, we forget for a while the correlation with $\mathbf{c}$ and derive insights

[2]These noise parts given by model (6.12) will be denoted by $\mathbf{n}_f$ and $\mathbf{n}_r$ respectively. Fig.6.5(b) shows that $\sigma^2_{n_f} \ll \sigma^2_{z_f}$. So, the signal $\mathbf{z}_f$ contains parts of $\mathbf{c}$ that have to be removed from it in order to get the noise part (i.e. $\mathbf{n}_f$). Note also that simulation results with real audio signals support the fact that $\mathbf{z}_f$ and $\mathbf{c}$ are highly correlated.

into the AWGN&J channel using conventional ISI assumptions[3]. In this case, based on (6.19) and (6.21) one can already give some specificities of watermarking channels including jitter:

(a) The influence of the jitter depends on the composite signal power $\sigma_c^2 = \sigma_s^2 + \sigma_x^2$. Hence, the well known embedder strategy consisting in increasing the watermark power σ_x^2 to improve detector performance in case of AWGN attacks is no longer the optimum strategy, since at the same time, it enforces the impact of the desynchronization attack by increasing the attack distortion (see Fig.6.6).

(b) Since the jitter noise is somehow proportional to the original signal, embedding the watermark into a *transform domain* where the original data is less powerful may alleviate the effect of the jitter.

In the following section, the AWGN&J channel is characterized using the model (6.12). The goal is, as stated before, to compare the resulting distortions to those being derived using the ISI approach and given by (6.19) and (6.21).

6.4.2 The AWGN&J channel in light of model (6.12)

Expressing differently the jittered signal $\mathbf{c}_J$, (i) using the model (6.12) and, (ii) using (6.17), we get: $k\mathbf{c} + \mathbf{n} = \mathbf{c}_J$. Constant and random time shifts are treated separately.

6.4.2.1 Constant time shift

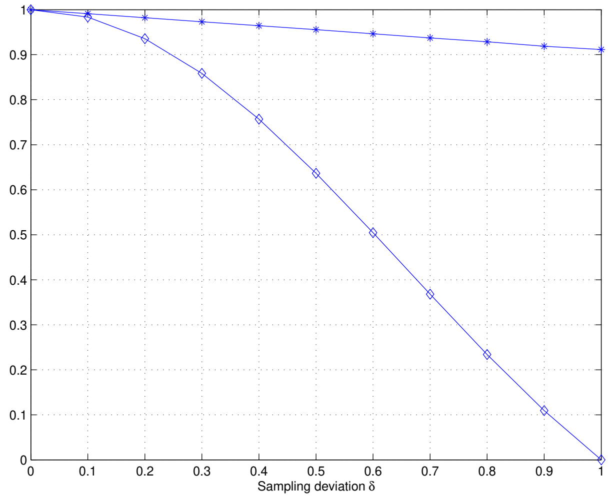
As mentioned before, the scaling does not change the overall energy of the signal $\mathbf{c}$ in case of a constant time shift. This can be shown to result in $k \in [-1, 1]$ and $\sigma_n^2 = (1 - k^2)\sigma_c^2$. Also, using the model (6.12), it follows that

$$k_f = \mathrm{sinc}(\delta) + \frac{< \mathbf{z}_f, \mathbf{c} >}{\|\mathbf{c}\|}, \tag{6.22a}$$

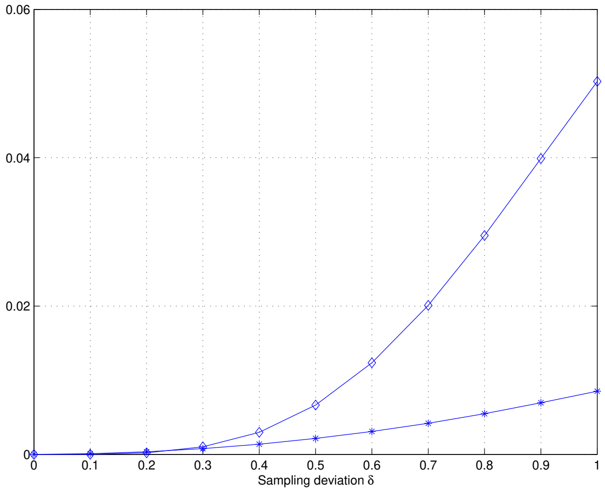
$$\mathbf{n}_f = \mathbf{z}_f - \frac{< \mathbf{z}_f, \mathbf{c} >}{\|\mathbf{c}\|}\mathbf{c}. \tag{6.22b}$$

Fig.6.5(a) depicts the dependency of the equivalent scaling factor k_f on the sampling deviation δ. It can be seen that k_f decreases with δ, but has a much smaller dependency on δ than the factor $sinc(\delta)$ corresponding to the ISI approach. Note that curves in Fig.6.5(a) correspond to a Document-to-Watermark Ratio (DWR $= 10\log_{10}(\frac{\sigma_s^2}{\sigma_x^2})$) of 20 dB and a Watermark-to-Noise Ratio (SNR $= 10\log_{10}(\frac{\sigma_x^2}{\sigma_v^2})$) of 0 dB, which are typical values in watermarking systems. Smaller values of k_f can be obtained with stronger composite signals. Noteworthy, the model parameter k_f given by (6.22a) is larger than the scaling factor sinc(δ) of expression (6.17) obtained with the ISI approach. In order to further outline the accuracy of the model (6.12), we compare the power $\sigma_{n_f}^2$ of the noise $\mathbf{n}_f$ to that of the ISI term $\mathbf{z}_f$. The result is depicted in Fig.6.5(b). We see that $\sigma_{n_f}^2$ naturally increases with the shift δ. However, unlike the scaling factor, $\sigma_{n_f}^2$ is smaller than $\sigma_{z_f}^2$. As stated before, writing the jittered signal $\mathbf{c}_f$ as the sum of two signals, one which is proportional (highly correlated) to it and another which is decorrelated from it, permits the extraction of the noise part $\mathbf{n}_f$. Since $\sigma_{n_f}^2$ is smaller than $\sigma_{z_f}^2$ and $\sigma_{n_f}^2$ is the power of the exact noise term in $\mathbf{y}_f$, it follows that $\mathbf{z}_f$ should not be totally accounted for as noise. The difference $\sigma_{z_f}^2 - \sigma_{n_f}^2$ corresponds to the power of the part of $\mathbf{z}_f$ that is falsely attributed to noise in the ISI approach.

[3]The ISI term is uncorrelated with the signal $c(t)$ being interpolated.



(a) scaling factor



(b) desynchronization noise

Figure 6.5: The effect of a constant time scaling $\Delta = \delta T$ is investigated differently (i) as an additional noise of power $\sigma_{z_f}^2$ resulting from the ISI term (diamond) as considered in [BEH02] and (ii) using the proposed model (asterisk). Corresponding scale factors and desynchronization noises are compared. (a): diagram of dependency of the scale factor k_f on the deviation δ with respect to $\mathrm{sinc}(\delta)$. (b): the equivalent white noise power $\sigma_{n_f}^2$ with respect to $\sigma_{z_f}^2$ stemming from the plain model. Results are obtained with DWR = 20 dB.

6.4.2.2 Random time shift

Consider now the random jitter case. Again, we have

$$\mathbf{c} + \mathbf{z}_r = k_r \mathbf{c} + \mathbf{n}_r \tag{6.23}$$

with $\mathbf{n}_r$ uncorrelated with $\mathbf{c}$. Parameters k_r and $\mathbf{n}_r$ can be derived in a way similar to the constant shift case. Intuitively however, unlike a constant shift, the random variable τ in $z_r(t)$ ensures enough randomness, this time, so that the objective error may be reasonably considered as uncorrelated with the composite signal $c(t)$ (this is checked below by simulation, see Fig.6.6). $z_r(t)$ can hence be assimilated to a signal-dependent noise which is approximately decorrelated from $\mathbf{c}$. Therefore, it follows that

$$k_r \approx 1, \tag{6.24a}$$

$$\mathbf{n}_r \approx \mathbf{z}_r. \tag{6.24b}$$

Desynchronization experiments including real audio signals sampled at $f_e = 44.1$ kHz show that k_r is most of the time very close to unity and that for a jitter square deviation $J \in [0, 1]$, we have $k_r \geq 0.97$. Also, these tests show that the embedded watermark is inaudible as long as $J \leq 0.04$. Of course, this threshold depends on the signal used and should not be taken for granted, but it already gives an idea about the jitter square deviation range of interest. For this range, simulations show that $k_r \geq 0.99$. The uncorrelation assumption is, unlike the constant time shift, approximately valid for practically all relevant jitter attacks. The jitter acts then as an additive noise of power $\sigma_{n_r}^2 = \sigma_{z_r}^2$. However, this noise is dependent on the composite signal $\mathbf{c} = \mathbf{s} + \mathbf{x}$. Fig.6.6 illustrates this dependency: here, the cover signal power σ_x^2 is maintained fixed. That (σ_x^2) of the watermark $\mathbf{x}$ varies according to DWR $= 10\log_{10}(\sigma_s^2/\sigma_x^2) \in \{10, 15, 25, 30\}$ dB. Also, the additive Gaussian noise power $\sigma_v^2 = \sigma_x^2 10^{-3}$ is fixed. We see that: (i) the effect of the jitter (strength of desynchronization noise $\mathbf{n}_r$) increases with the jitter square deviation J (the dependency is approximately linear). In addition, (ii) as the power σ_x^2 increases (DWR decreases), the jitter becomes stronger. This illustrates the remark above: increasing the watermark power for more reliable detection in an AWGN&J channel, increases at the same time the effect of the jitter.

In light of the comparison stated above, we conclude that:

(a) The decorrelation hypothesis between $\mathbf{z}_f$ and $\mathbf{c}$ is in general not accurate. The ISI term $\mathbf{z}_f$ is highly correlated with $\mathbf{c}$.

(b) Removing from the ISI term the signal-like term results in a more accurate characterization of the attack where the *real* scaling factor k_f is larger than $\mathrm{sinc}(\delta)$ and the *real* equivalent additive noise $\mathbf{n}_f$ is much weaker than the ISI term $\mathbf{z}_f$.

(c) The objective distortion induced by the scaling attack is

$$D_{af} = |k_f - 1|^2 \sigma_c^2 + \sigma_{n_f}^2. \tag{6.25}$$

The perceived distortion is given by $\sigma_{n_f}^2$ which is much smaller than that rising directly from the plain model.

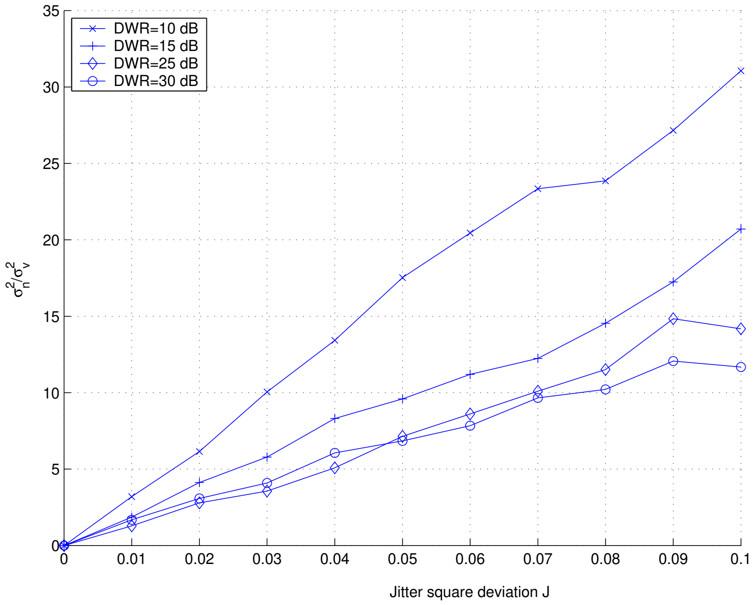


Figure 6.6: Diagram of dependency of the desynchronization noise $\sigma_{n_r}^2$ on the jitter square deviation J. Dependency on the composite signal $\mathbf{c} = \mathbf{s}+\mathbf{x}$ is illustrated through that on the Document-to-Watermark Ratio DWR: the jitter becomes stronger with strong watermarks (lower DWRs).

(d) The *random* jitter has an additive signal-dependent like behavior.

The AWGN&J channel has been characterized in terms of jitter induced distortions. Since the capacity of any watermarking scheme depends mainly on these distortions[4], one important point is to evaluate the performances of this scheme over an AWGN&J channel. The distortions expressed above will help estimate the real performances loss. To that end, two watermarking schemes taken respectively from the interference-rejecting and non-rejecting watermarking methods are considered. For the former, a brief overview of communication with state information at the encoder is given.

6.5 Optimal and Suboptimal Information Embedding over an AWGN&J Channel

We assume watermarking of an independent identically distributed (IID) Gaussian original signal $\mathbf{s} \sim \mathcal{N}(0, \sigma_x^2)$ over a watermark channel characterized by its attack $\mathcal{A}$ such that $\mathbf{y} = k\mathbf{c} + \mathbf{n}$. Such a channel may represent the traditional AWGN channel, the SAWGN channel investigated in [EBG02a], the AWGN&J depicted above or any other watermarking channel (attack). Only the pair $(k, \mathbf{n})$ would vary accordingly. The receiver compensates for the scaling by dividing $\mathbf{y}$ by k to produce the pre-processed signal

$$\mathbf{y}' = \mathbf{s} + \mathbf{x} + \frac{\mathbf{n}}{k}. \tag{6.26}$$

[4]It also depends on the embedding distortion $D_E = \sigma_x^2$. In addition, for blind Spread-Spectrum embedding, the host signal itself accounts for self noise and must be included in the channel distortion as shown by (6.32a).

Thus, the watermark receiver sees an AWN channel with the *effective* noise $\mathbf{n}' = \frac{\mathbf{n}}{k}$, with variance σ_n^2/k^2. The watermark capacity for communicating over this effective channel depends only on the cover signal $\mathbf{s}$ and the ratio of the embedding distortion $D_E = \|\mathbf{c} - \mathbf{s}\| = \sigma_x^2$ by the effective channel noise σ_n^2/k^2. The noise power σ_n^2 is related to D_a by (6.15) which enables the computation of the ratio $\frac{k^2 D_E}{\sigma_n^2}$ as

$$\frac{k^2 D_E}{\sigma_n^2} = \frac{k^2 D_E}{D_a - (k-1)^2(\sigma_s^2 + D_E)}. \tag{6.27}$$

6.5.1 The Ideal Costa Scheme ICS

Rather than considering watermarking as communication over a very noisy channel where the host signal $\mathbf{s}$ acts as self-interference (as in SS), it has recently been realized [CW99, CMM99] that blind watermarking can be viewed as *communication with side information at the encoder*. The relevant work is the initial Costa "Writing on Dirty Paper" [Cos83]. Fig.6.7 depicts a block diagram of blind watermark communication over the channel (6.12) where the encoder exploits the side-information about the host signal. The scheme originally conceived by Costa is called "*Ideal Costa Scheme*" and emerges as a

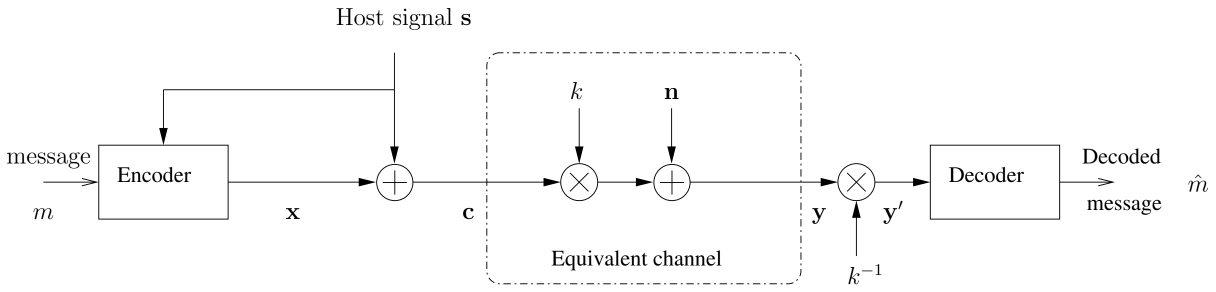


Figure 6.7: Blind watermarking as Writing on Dirty Paper over channel (6.12).

universally good encoding strategy for coding with side information available at the encoder. Based on a huge random codebook, Costa showed that optimal transmitter encodes its message "*in the direction*" of the interfering signal $\mathbf{s}$ such that the latter does not affect the capacity of the channel, achieving thus the standard Gaussian channel capacity. In our case, the effective watermark-to-noise power ratio $\frac{\sigma_x^2}{\sigma_{n'}^2}$ is given by (6.27). Hence, the communication rate under an attack of the form (6.12) can be written as

$$\mathcal{R}_{\mathrm{ICS}}^{\mathcal{A}} = \frac{1}{2}\log_2\left(1 + \frac{k^2 D_E}{D_a - (k-1)^2\sigma_c^2}\right). \tag{6.28}$$

For given D_E, D_a and σ_s^2, capacity is defined as the supremum of all achievable rates. Alternatively, Moulin et al. showed in [MO03] that the *hiding capacity* may be formulated as a min-max problem between the information hider and the attacker. The information hider wants a guaranteed rate of reliable transmission under any attack that satisfies a upper-bound constraint on D_a. Conversely, the attacker wants to minimize this rate for any information hiding strategy that satisfies an upper-bound constraint on the embedding distortion D_E. Later, in [MML00], Moulin et al., using a different distortion measure, have shown that the optimum attack over all possible attacks is a specific Scale plus Additive White Gaussian Noise (SAWGN). For the channel (6.12) investigated here, capacity is then obtained by minimizing (6.28) over all possible attacks $k \in [1 - \sqrt{\frac{D_a}{\sigma_s^2 + D_E}}, 1 + \sqrt{\frac{D_a}{\sigma_s^2 + D_E}}]$. The constraint on the admissible scaling factor set corresponds to the expression inside the function log(.) in (6.28) strictly larger than unity. Otherwise, capacity would be negative and the watermarking system design becomes meaningless. Details of the resolution are skipped here since a very similar game, where the objective function is the detection probability, will be throughly studied in Section 6.5. The resolution gives $k_{opt} = 1 - \frac{D_a}{\sigma_s^2 + D_E}$ and

$$\mathcal{C}^{\mathcal{A}}_{\text{ICS}} = \frac{1}{2}\log_2\left(1 + \frac{D_E(\sigma_c^2 - D_a)}{\sigma_c^2 D_a}\right) < \frac{1}{2}\log_2\left(1 + \frac{D_E}{D_a}\right). \tag{6.29}$$

Note that in general $D_a \ll \sigma_s^2 + D_E$ such that $k_{opt} \in [1 - \sqrt{\frac{D_a}{\sigma_s^2 + D_E}}, 1 + \sqrt{\frac{D_a}{\sigma_s^2 + D_E}}]$ is satisfied. Also, the term on the right hand of (6.29) is the achievable capacity if there were no attack (which is that of an AWGN channel with signal-to-noise ratio D_E/D_a).

6.5.2 Traditional Spread-Spectrum

A simplified diagram of basic SS-based watermarking over the channel (6.12) is shown in Fig.6.8. Blind

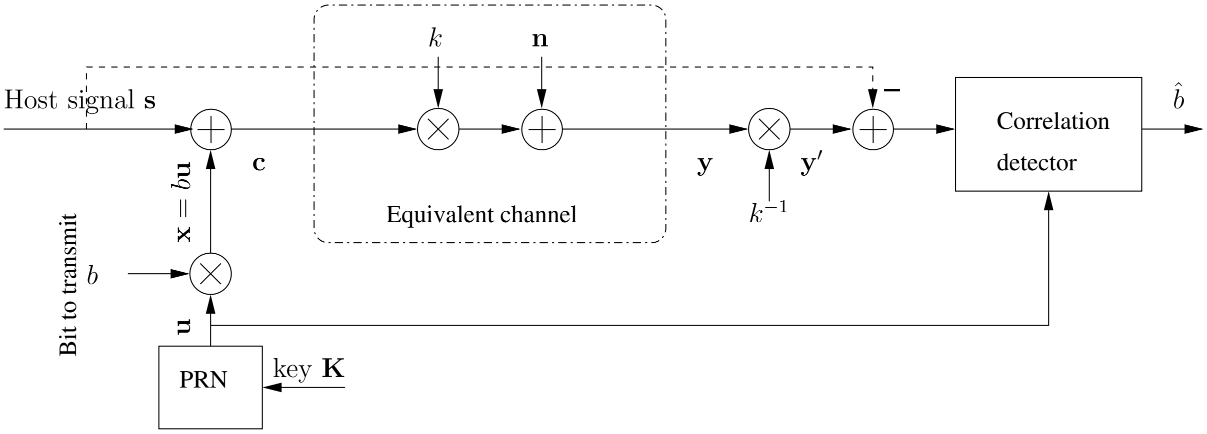


Figure 6.8: Blind (solid line) and non-blind (dashed line) spread-spectrum-based watermarking over channel (6.12).

and non-blind reception refer to the fact of having or not access to the cover signal $\mathbf{s}$ at the receiver side. If the decoder has access to $\mathbf{s}$, the decoder subtracts the cover signal $\mathbf{s}$ from the received signal $\mathbf{y}'$ prior to decoding. If not, the decoder performance suffers greatly from host signal interference. Blind and non-blind communication rates under an attack of the form (6.12) can be written as

$$\mathcal{R}^{\mathcal{A}}_{\text{Blind SS}} = \frac{1}{2}\log_2\left(1 + \frac{k^2 D_E}{k^2\sigma_s^2 + D_a - (k-1)^2\sigma_c^2}\right), \tag{6.30a}$$

$$\mathcal{R}^{\mathcal{A}}_{\text{Non-blind SS}} = \frac{1}{2}\log_2\left(1 + \frac{k^2 D_E}{D_a - (k-1)^2\sigma_c^2}\right). \tag{6.30b}$$

Again, capacity is obtained through a min-max problem resolution. The set of admissible scaling factors for the non-blind case is the same as before. For blind SS, it is given by

$$k \in \left[\frac{(\sigma_s^2 + D_E) - \sqrt{(\sigma_s^2 + D_E)^2 - (\sigma_s^2 + D_E - D_a)D_E}}{D_E}, \frac{(\sigma_s^2 + D_E) + \sqrt{(\sigma_s^2 + D_E)^2 - (\sigma_s^2 + D_E - D_a)D_E}}{D_E}\right]. \tag{6.31}$$

The optimization results in the same saddle-point $k_{opt} = 1 - \frac{D_a}{\sigma_s^2 + D_E}$ as before which satisfies (6.31) and for which transmission rates are given by

$$\mathcal{C}^{\mathcal{A}}_{\text{Blind SS}} = \frac{1}{2}\log_2\left(1 + \frac{D_E(\sigma_c^2 - D_a)}{\sigma_c^2 D_a + (\sigma_c^2 - D_a)\sigma_s^2}\right), \tag{6.32a}$$

$$\mathcal{C}^{\mathcal{A}}_{\text{Non-blind SS}} = \frac{1}{2}\log_2\left(1 + \frac{D_E(\sigma_c^2 - D_a)}{\sigma_c^2 D_a}\right). \tag{6.32b}$$

Capacity loss for both ICS and SS is depicted in Fig.6.9. As shown by (6.29) and (6.32a), the attack (6.12) results in significant capacity loss especially for very low watermark-to-noise ratios D_E/D_a. As for the AWGN channel, ICS outperforms SS for almost all values of D_E/D_a. Note however that ICS-capacity reduction is larger than that for SS: ICS is less robust than SS facing attacks of the form (6.12). This fact will be supported by simulations over an AWGN&J channel (see Figs. 6.10(a) and 6.10(b) below). Also, in case of very strong attacks, ICS and SS capacities fall to the same values and ICS presents no gain over SS. These attacks are however sufficiently strong to practically impair any communication and are, consequently, not relevant in real applications. For reasonable watermark-to-noise ratios ($10\log_{10}(D_E/D_a) > -16$ dB), ICS remains more efficient.

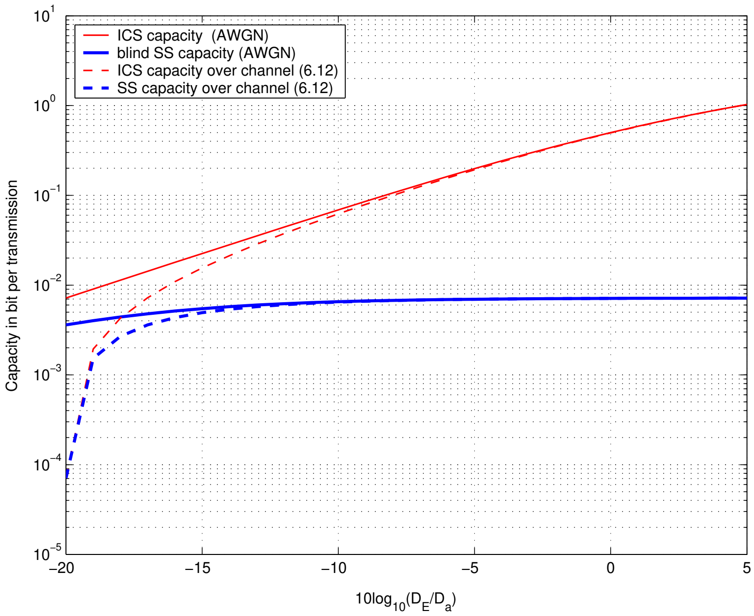


Figure 6.9: Capacity loss of both Ideal Costa Scheme (ICS) and Blind Spread Spectrum Scheme under the influence of an attack of the form (6.12). The result is depicted for DWR $= 10\log_{10}(\frac{\sigma_s^2}{D_E}) = 20$ dB. For strong attacks, ICS and SS capacities fall to same values. ICS becomes more sensitive than SS.

Now, focus on the special case of an AWGN&J channel. This channel has been shown to be a special case of attacks of the form (6.12), with parameters k and $\mathbf{n}$ given by (6.22a), (6.22b), (6.24a) and (6.24b). Hence, ICS and blind SS capacities over an AWGN&J channel are readily given by (6.29) and (6.32a), respectively and are shown in Fig.6.9. However, since these capacities are obtained through a min-max resolution, they correspond to the achievable rate under the optimum attack (k_{opt}). More insights can be obtained using achievable rates (6.28) and (6.30a) instead of capacities as stated below.

6.5.3 Application to AWGN&J channels

Here, unlike capacity which is derived analytically using k_{opt}, we want to see how transmission rates $\mathcal{R}_{\text{ICS}}^{\text{AWGN\&J}}$ and $\mathcal{R}_{\text{blind-SS}}^{\text{AWGN\&J}}$ degrade in presence of a jitter $\mathbf{J}$. That is, we are interested in the observed jitter and not the optimal one as in capacity analysis. Simulations are required to compute parameter k and noise $\mathbf{n}$ in (6.28) and (6.30a). We proceed as follows: given some jitter (shift δ), the composite signal $\mathbf{c}$ is interpolated resulting in $\mathbf{c}_J$. Next, the equivalent attack (scaling k and noise $\mathbf{n}$) is derived and white Gaussian noise $\mathbf{v}$ is added. The received signal is $\mathbf{y} = k\mathbf{c} + \mathbf{v}_{eq}$, where $\mathbf{v}_{eq} = \mathbf{v} + \mathbf{n}$ is the overall channel noise.

6.5.3.1 Rate loss under constant time shift attacks

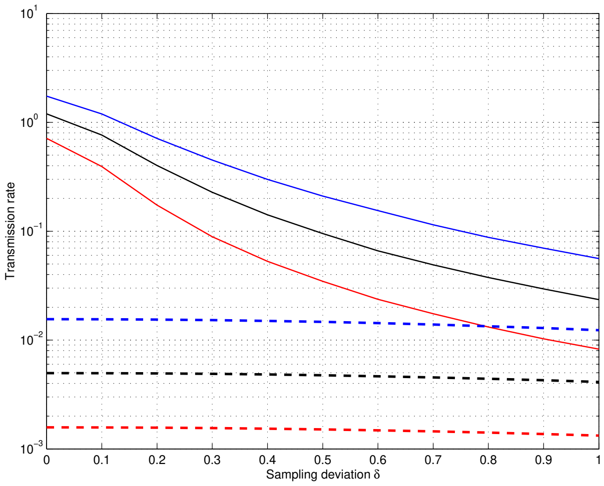
Fig.6.10(a) depicts transmission rates given by (6.28) and (6.30a) using expressions given by (6.22a) and (6.22b) for the scaling factor k and the noise $\mathbf{n}$. For both ICS and SS, these are shown for three values of $\text{DWR} = 10\log_{10}(D_E/D_a)$. We observe that ICS transmission rate drastically decreases if the sampling deviation δ increases. This illustrates the loss in ICS-capacity already shown in Fig.6.9 and particularly apparent for low $\text{SNR} = 10\log_{10}(D_E/D_a)$: as δ is close to unity, the jitter induced distortion D_a is large and SNR is low. Note that, ICS rate degradation reveals a more general setting: almost all quantization-based embedding schemes are highly sensitive to scaling. When scaled, the received signal is rounded to a bad quantization cell center. Blind SS, however, is almost insensitive to scaling, but performs far below ICS. This is particularly useful for the design of watermarking systems in situations where the transmitted signal may be scaled by the channel: ICS should be preferred to SS for applications where a great amount of information is to be transmitted. However, SS may be used for applications where transmission rate is not the main issue and where robustness against scaling is highly appreciated. The latter applications are referred to as ”one bit watermarking” problems in digital watermarking. Another important remark arises from comparing the transmission rates corresponding to the same shift δ but different values of DWR. It can be seen that the higher the DWR, the larger the rate loss. This does not contradict (6.19) because the embedding distortion D_E is reduced as well so that the transmission rate broadly decreases for large DWR[5].

6.5.3.2 Rate loss under random jitter attacks

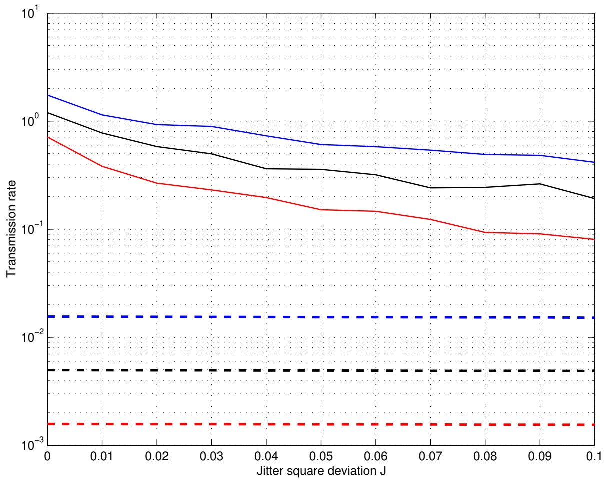
The effect of a random jitter $\mathbf{J} \sim \mathcal{N}(0, J)$ combined with an additive white Gaussian noise $\mathbf{v}$ attack on a composite signal $\mathbf{c} = \mathbf{s} + \mathbf{x}$ is depicted in Fig.6.10(b). For the same reason as above, we concentrate on attacks with jitter square deviation $J < 0.04$. Again, we use (6.28) and (6.30a) where k_r and $\mathbf{n}_r$ are replaced by (6.24a) and (6.24b) respectively. As for the constant time shift case, we observe that ICS rate reduction is larger than that of SS, which is almost insensitive to the jitter. Also, though large DWRs result in small distortions as previously shown by (6.21), the decrease in the embedding distortion D_E cause the transmission rate to degrade.

Now compare the ICS rate loss to that which results from a constant shift attack. Note that, the rate loss is larger when facing constant shifting. This is not completely surprising: remember that the random

[5] Note that small increasing the DWR is obtained through decreasing the embedding distortion $D_E = \sigma_x^2$. Cover signal and Gaussian noise powers σ_x^2 and σ_v^2 are maintained fixed as above.



(a) constant scaling



(b) random jitter

Figure 6.10: Transmission rate loss of both Ideal Costa Scheme (solid line) and Blind Spread Spectrum (dashed line) over an AWGN&J channel. Gaussian noise $\mathbf{v}$ is such that $\sigma_v^2 = 10^{-3}\sigma_s^2$. (a): the composite signal $\mathbf{c}$ is scaled with $\Delta = \delta T$. (b): $\mathbf{c}$ is randomly re-sampled using the jitter $\mathbf{J} \sim \mathcal{N}(0, J)$. With both schemes and under both attacks, transmission rate degrade with DWR. From bottom to top: $\mathrm{DWR} = 25, 20$ and 15 dB.

jitter attack has been shown to behave like additive noise. With ICS, whose practical implementations are forms of quantization, scaling is more harmful than adding noise. This fact will be supported by game theory resolution in Section 6.5.

The remaining of the chapter is devoted to providing insights into both the *optimum attack* and the *optimum defense*. By "*optimum*", we mean "the best strategy" in a *game theory* context. The Watermarking Game does not have universal solutions and both attacker and defender should *adapt* to each other . Here, the game is first briefly reviewed and then solved in case of an AWGN&J attack and blind spread spectrum. We also provide a simple means of circumventing constant time shift attacks.

6.6 A Game Theory Approach to AWGN&J Channels

In a robust watermarking transmission context, the embedder must design his embedding scheme so that the watermark survives the worst possible attack. Conversely, the attacker has to perform the optimal attack that best impairs the watermark, for a given distortion budget. The resulting optimization problem (game theory problem) is often formulated as a max-min (or min-max) problem. The criterion to be optimized is the detection (or equivalently, error) probability in case of *one-bit watermarking* and, the watermarking capacity in case of *data hiding*. Since capacity has already been optimized in Section 6.4 and since for many watermarking applications, the most significant criterion is reliable detection, we concentrate on the one-bit watermarking. We consider the criterion of detection probability. The watermarking game has been thoroughly studied in the case of an AWGN channel [CL02a, CL01, SA80]. In [MI03], Moulin and al. discussed the case of attacks by filtering and additive noise. In [EBG02a], Eggers et al. considered attacks by amplitude scaling and additive noise. But, in contrast to the following, only objective distortions were used and were evaluated with respect to the original host signal, not to the composite signal.

In an AWGN&J channel, the attacker can desynchronize the signal and add noise as well. In this section, we answer the following three questions:

(i) If ever the attacker has a perceived distortion budget, with two ways of using it: either by introducing jitter, or by using additive noise or any combination of both, what is his best strategy?

(ii) Conversely, what is the best tuning for the defender knowing the best potential attack?

(iii) Is there means for the defender to find countermeasures to the attacker strategy (put some limits to the efficiency of its optimal strategy)?

We begin by answering (iii). The main difficulty in synchronizing a randomly scaled received signal stems, as stated above, from the fact that random time scaling the composite signal broadly behaves like adding noise (disregarding the fact that this noise is signal-dependent as given by (6.20)). In case of a constant time-shift however, the receiver should be able to reverse the effect of scaling. The main solutions that have been proposed can be divided into two categories: (i) embedding of a pilot sequence as classically used in traditional communication and (ii) using a correlation-based alignment algorithm [SK04]. While pilot sequences present an additional source of weakness if ever intercepted by an attacker, the algorithm in [SK04], has good matching properties but requires the availability of (a copy of) the original signal at the receiver side[6]. This algorithm consists in computing the maximum normalized correlation between the pirated (attacked) signal and the original. Here, we propose a cross-correlation based matching process that we denote by "multiple correlation test". This procedure is similar to that in [SK04], but does not require knowledge of the original content at the receiver. Blind re-synchronization is made possible by using the watermark instead of the original signal for correlation computation. Having access to the watermark $\mathbf{x}$ at the receiver side is commonly assumed in a "one bit watermarking" context. The aim is to *mark* users specific contents with the same small watermark.

[6] In [SK04], Schonberg et al. proposed this algorithm in the context of *fingerprinting*, which is indeed an application where availability of the original signal is usually assumed. Here, we focus on detecting the same watermark embedded in several different contents.

6.6.1 Preventing constant shift

Suppose the attacker performs, in addition to the additive white Gaussian noise $\mathbf{v}$, a time shift $\Delta = \delta T$ with $\delta \in [0,1]$. The restriction $\delta \in [0,1]$ is due to the fact that, with a cross-correlation based re-synchronization procedure, the defender can compensate for any T-multiple time shift: the receiver searches for the maximum cross-correlation between the received (attacked) signal $\mathbf{y}$ and the watermark $\mathbf{x}$ and realign $\mathbf{y}$ before proceeding to detection so that he gets rid of any T-multiple scaling. We concentrate then on the case $\delta \in [0,1]$. As stated above, the received composite (and attacked) signal is given by $\tilde{y}(t) = y(t+\Delta) = \tilde{s}(t) + \tilde{x}(t) + v(t)$ where $\tilde{s}(t)$, $\tilde{x}(t)$ and $\tilde{y}(t)$ are respectively desynchronized signals $s(t)$, $x(t)$ and $y(t)$. The analysis below shows that desynchronization is much more harmful than white noise. Therefore, it is very important for the defender to maintain this part of the attack to a reasonable level. One possible way is to interpolate the received signal, so that the receiver performs several shifts of the watermark $x(t)$ along the time axis, and proceed to correlation tests with $\tilde{y}$ for each of these shifted versions. Depending on the number of correlation tests the decoder can perform, the receiver can maintain the maximum time shift to a desired bounded value. A large number of tests at the receiver side, however, increases the computational complexity. Therefore, there will be a trade-off between computational complexity and optimality. Suppose the receiver is able to perform $M(\geq 2)$ tests. Let $\tilde{\mathbf{x}}^{(k)}$ denote the watermark signal shifted by $\frac{kT}{M}$, $k \in [1:M-1]$, that is $\tilde{x}^{(k)}[n] = \tilde{x}[nT + \frac{kT}{M}]$ and $\rho^{(k)} = \frac{<\tilde{\mathbf{y}},\tilde{\mathbf{x}}^{(k)}>}{\sqrt{\|\tilde{\mathbf{y}}\|\|\tilde{\mathbf{x}}^{(k)}\|}}$ the correlation coefficient between the received signal and $\tilde{x}^{(k)}(t)$. $\rho^{(0)} \triangleq \frac{<\tilde{\mathbf{y}},\mathbf{x}>}{\sqrt{\|\tilde{\mathbf{y}}\|\|\mathbf{x}\|}}$. In order to bound Δ within an interval of length $\frac{T}{M}$, the receiver determines $k_0 \in [1:M-1]$ according to

$$k_0 = \underset{k \in [1:M-1]}{\arg\max} \quad \rho^{(k)}. \tag{6.33}$$

k_0 represents the location index for which $\tilde{y}(t)$ optimally matches the signal $y(t)$ when scaled back by $\frac{k_0 T}{M}$. Next, the receiver proceeds to detection using the aligned signal $\tilde{y}(t - \frac{k_0 T}{M})$. Likewise, the residual desynchronization is smaller than $\frac{T}{M}$ and the attacker should not waste energy in further desynchronizing the composite signal $c(t)$. The cost the receiver has to pay in order to maintain the maximum time shift to a (small) bounded value is the computation of M correlations. From the analysis outlined in the first part of the chapter, this bounding (a parameter of the transmitter) is absolutely required, otherwise desynchronization induced noise would increase to very large values, resulting in very poor detection performances.

6.6.2 Game theoretical formulation

We consider the embedding of one bit of information $b \in \{0,1\}$ into an original data $\mathbf{s}$ of length N, assumed to be Gaussian, $\mathbf{s} \sim \mathcal{N}(0, \sigma_s^2)$. The watermark signal is given by $\mathbf{x} = b\mathbf{u}$ where chips x_i are mutually independent with respect to $\mathbf{s}$. The sequence $\mathbf{u}$ is produced by a Pseudo Random Number generator (PRN) using a secret key $\mathbf{k} \in K$. Its elements are equal to $+\sigma_u$ or $-\sigma_u$ (see Fig.6.8). Also, according to Kerkhoff's principle, we suppose the attacker knows the watermarking scheme that we used. The embedder, however, not having access to the attacker scale factor k, does not normalize the received signal. In this context, the attacker may either add white Gaussian noise, desynchronize the composite signal or perform both operations as long as the overall attack distortion D_a is upper bounded by a certain tolerance level $D_{a_{max}}$. On the other side, the embedder chooses the appropriate length N of original data, the number M of correlations to be performed, and watermark power σ_x^2 subject to a certain maximum embedding distortion $D_{E_{max}}$.

6.6.2.1 Detection probability

Detection is based on the sign of $y = \frac{<\mathbf{y},\mathbf{u}>}{\|\mathbf{u}\|}$ where the received signal is $\mathbf{y} = k\mathbf{c} + \mathbf{n} + \mathbf{v}$ and $y = kb + ks + n + v$, with

$$\begin{cases} s = \frac{<\mathbf{s},\mathbf{u}>}{\|\mathbf{u}\|} \sim \mathcal{N}(0, \frac{\sigma_s^2}{N\sigma_x^2}), \\ v = \frac{<\mathbf{v},\mathbf{u}>}{\|\mathbf{u}\|} \sim \mathcal{N}(0, \frac{\sigma_v^2}{N\sigma_x^2}), \\ n = \frac{<\mathbf{n},\mathbf{u}>}{\|\mathbf{u}\|} \sim \mathcal{N}(0, \frac{\sigma_n^2}{N\sigma_x^2}). \end{cases}$$

So, the PDF of y is given by $y \sim \mathcal{N}(kb, \frac{k^2\sigma_s^2+\sigma_n^2+\sigma_v^2}{N\sigma_x^2})$. Natural performances measure for the one-bit watermarking problem are probability of false positive (false alarm) P_{FA}, false negative (miss detection) P_{MD} and probability of detection P_{d}. The watermark detection problem can be formulated as a hypothesis test:

$$\begin{cases} H_0 : b = 0 : \text{no watermark}, \\ H_1 : b = 1 : \text{watermark found}. \end{cases}$$

The detector decides that a watermark is present if $y > \nu$, where ν is some detection threshold that controls the trade-off between false positive and false negative decisions. These probabilities are given by

$$P_{\text{FA}} = P(y > \nu | H_0) = \frac{1}{2}\text{erfc}\left(\nu\sqrt{\frac{N\sigma_x^2}{2(k^2\sigma_s^2 + \sigma_v^2 + \sigma_n^2)}}\right), \tag{6.34a}$$

$$P_{\text{MD}} = P(y < \nu | H_1) = \frac{1}{2}\text{erfc}\left((k-\nu)\sqrt{\frac{N\sigma_x^2}{2(k^2\sigma_s^2 + \sigma_v^2 + \sigma_n^2)}}\right), \tag{6.34b}$$

$$P_{\text{d}} = P(y > \nu | H_1) = \frac{1}{2}\text{erfc}\left((\nu-k)\sqrt{\frac{N\sigma_x^2}{2(k^2\sigma_s^2 + \sigma_v^2 + \sigma_n^2)}}\right). \tag{6.34c}$$

The parameter k is unknown. Non coherent detection theory provides several techniques to solve detection problems with unknown parameters. Below *Neyman-Pearson approach* is first reviewed and then applied to derive consistent choice of parameter ν. More details about such a choice can be found in [PBbC98].

(i) **Neyman-Pearson criterion for threshold selection** *Subject to a constraint on the maximum acceptable probability of false positive (false alarm), the test consists in minimizing the probability of false negative (miss-detection).*

For example, a maximum allowable probability of false alarm $P_{\text{FA}_{\max}} = 10^{-6}$ leads to a threshold $\nu \approx 3.3\sqrt{2\sigma_y^2}$. In [OP97], it is stated that to improve the characteristics of robustness against attacks, the new threshold should be evaluated directly on the composite and possibly attacked signal $\mathbf{y}$. This results in the following choice:

$$\nu \approx \begin{cases} 3.3\sqrt{\frac{2(k_f^2\sigma_s^2+\sigma_v^2+\sigma_{n_f}^2)}{N\sigma_u^2}} & \text{for a constant shift} \\ 3.3\sqrt{\frac{2(\sigma_s^2+\sigma_v^2+JE[(\frac{d}{dt}c(t))^2])}{N\sigma_u^2}} & \text{for a random shift.} \end{cases}$$

(ii) **Max-min criterion** Over an AWGN&J channel, the detection probability, denoted by $P_d^{AWGN\&J}$,is given by (6.34c). Also, we assume as stated in [BEH02], that meaningful embedding and attack distortions should satisfy

$$0 \leq D_{Emax} \leq D_{amax} \leq \sigma_s^2. \tag{6.35}$$

Taking into account the defender ability to perform the *multiple correlation* test described above, the scaling must satisfy $\delta \leq \frac{1}{M}$. Due to the cost of computing the correlations, we assume that $M \leq M_{max}$, where M is a parameter of the defender. For the same reason (correlation based detection cost), very large values of the signal length N are not allowed ($N \leq N_{max}$). As for the bound on M, that on N should ensure good compromise between detection performance and computing complexity. The embedder wants to maximize $P_d^{AWGN\&J}$ and the attacker wants to minimize it under constraints pair (D_{Emax}, D_{amax}). The problem is then naturally formulated as a game between the embedder and the attacker, and can be written as

$$\max_{D_E \leq D_{Emax}} \min_{D_a \leq D_{amax}} P_d^{AWGN\&J}. \tag{6.36}$$

This optimization problem is solved in the following section for both constant and random scalings.

6.6.3 Solving the watermarking game

The attack distortion has been shown above to be given by $D_a = |k-1|^2\sigma_c^2 + \sigma_n^2 + \sigma_v^2$. Let us first determine the part of the distortion budget that the attacker should allocate to noise and that to allocate to jitter, so that the detection performance is maximally reduced. By considering the proposed model $k\mathbf{c} + \mathbf{n} + \mathbf{v}$ in which $\mathbf{n}$ and $\mathbf{c}$ are uncorrelated as required by model (6.12), there is a priori no difference in nature between $\mathbf{n}$ and $\mathbf{v}$ (disregarding the dependency of $\mathbf{n}$ on the signal $\mathbf{c}$). Hence, we divide the global attack distortion D_a into two parts: D_k due to scaling and D_v due to the additive noise:

$$D_v = \sigma_v^2 + \sigma_n^2 = \alpha D_a, \tag{6.37a}$$

$$D_k = (k-1)^2(\sigma_s^2 + \sigma_x^2) = (1-\alpha)D_a, \tag{6.37b}$$

where parameter $\alpha \in [0, 1]$ characterizes the trade-off between the two components of the global distortion. We will refer to the case $\alpha = 1$ as the *all noise* case since the overall attack is equivalent to that of adding the noise quantity $\mathbf{n} + \mathbf{v}$. Similarly, we will refer to the case $\alpha = 0$ as the *all desynchronization* case since it corresponds to a channel attack by time axis scaling only. Any other attack with $\alpha \in]0, 1[$ will be termed as *mixed* since both adding noise and scaling are required.

6.6.3.1 Case of a constant time shift attack

Prior to revealing the optimum attacker and defender strategies, we assume that proper resynchronization procedure investigated above is used and reformulate the optimization problem. The *multiple correlation* *test* does not change the criterion to be optimized. However the ranges of the optimization variables M, N and α are modified (as it will shown by (6.41)). Intuitively, this follows from the fact that counter-measurements performed by the defender naturally reduce the set of admissible parameters for the attacker. In our case, a lower bound on δ can be shown to result in a lower bound on the attack scale factor k: let $h(.)$ be the function relating k_f to δ. For an explicit expression of $k_f = h(\delta)$, we need to combine equations (6.22a) and (6.18). Namely, we need invert (6.18) to get δ and replace it in (6.22a). Unfortunately, no explicit formula for parameter δ can be derived from (6.18). However, the dependence of parameter k on δ is depicted in Fig.6.5(a). Using this curve will be shown to be sufficient to bypass the difficulty raised above[7]. Of course, this results in an approximate solution but it is already enough to answer questions raised while formulating the game. Also, the curve depicted in Fig.6.5(a) corresponds to specific values of SNR = 0 dB and DWR = 20 dB but this would not change the concluding remarks related to relative noise and desynchronization effects and stated at the end of this section. In other words, not knowing the watermark power σ_x^2 does not matter since we only use the monotonously decreasing property of $h(.)$ in solving the game. This property implies a lower-bound on the set of admissible values for the scaling factor k. The constraint $\Delta \leq \frac{T}{M}$ gives δ in $[0, \frac{1}{M}]$. There exists then a lower bound on k_f, say $k_{min} \in [0, 1]$ such that $k_{min} = h(\frac{1}{M})$ and $k_f \geq k_{\min}\ \forall\, \delta \in [0, \frac{1}{M}]$. Using (6.37b), we obtain:

$$k_f = 1 - \sqrt{\frac{(1-\alpha)D_a}{\sigma_x^2 + \sigma_s^2}}. \tag{6.38}$$

Similarly, lower bound constraint $k_{\min}$ on k_f implies a similar constraint on α: there exists $\alpha_{\min} \in [0, 1]$ such that $\alpha \in [\alpha_{min}, 1]\ \forall\, \delta \in [0, \frac{1}{M}]$, which when combined with (6.38), gives

$$\alpha_{min} = 1 - \frac{(1-k_{min})^2(\sigma_s^2 + \sigma_x^2)}{D_a}. \tag{6.39}$$

Furthermore, inequality $\alpha_{min} \geq 0$ gives $D_k \leq D_{k\,max} = (1-\alpha_{min})D_a$. The latter upper bound on D_k can be understood this way: a part of the overall distortion D_a must be allocated to noise. Noteworthy, scenarios corresponding to $\alpha = \alpha_{min}$ and $\alpha = 1$ are worthy of some discussion.

- $\alpha = \alpha_{min}$

  With respect to cases $\alpha = 0$ (*all desynchronization*) and $\alpha = 1$ (*all noise*), this case corresponds to a mixed situation where the attacker should both add noise and desynchronize the signal. The global objective distortion D_a results then from both (i) an attack by amplitude scaling causing an objective distortion $D_k = (k_{\min} - 1)^2\sigma_c^2$ and (ii) an attack by additive noise of power $D_v = D_a - (k_{\min} - 1)^2\sigma_c^2$. With regard to these distortions, one can remark that

  (a) Increasing the composite signal power σ_c^2 enforces the distortion D_k due to the scaling factor with respect to that of the equivalent noise $\mathbf{v}_{eq} = \mathbf{v} + \mathbf{n}_z$.

  (b) Increasing the admissible set of correlations M causes the distortion D_k to decrease. Conversely, the additive noise distortion D_v increases.

[7]It will be shown that for the final solution, we need just bounds on the value-metric scaling parameter k. These can already be obtained from Fig.6.5(a).

Imposing a lower bound on α gives $D_k \leq (1-\alpha_{min})D_a$ and prevents the receiver from the *all desynchronization* attack. However, this is achieved at the cost of a certain signal processing complexity at the receiver side implicitly shown here through the defender parameter M.

- $\alpha = 1$

  This is the case of an attenuating additive noise $\mathbf{v}$. The attack is of type AWGN and traditional watermarking game solutions apply. Most prominent examples of these can be found in [MI03] and [CL02a].

We now rewrite the detection probability (6.34c) with k_f and $\sigma_v^2 + \sigma_{n_z}^2$ expressed by (6.38) and (6.37a) respectively. The resulting formula can be expressed as a function of both the setting of defender parameters $\{N, M, \sigma_x^2\}$ and that of the attacker $\{\alpha, D_a\}$, as

$$P_d^f\left([N, M, \sigma_x^2], [\alpha, D_a]\right) = \frac{1}{2}\text{erfc}\left((\nu - k_f)\sqrt{\frac{N}{2}\frac{\sigma_x^2}{(1-\sqrt{\frac{(1-\alpha)D_a}{\sigma_s^2+\sigma_x^2}})^2\sigma_s^2 + \alpha D_a}}\right), \tag{6.40}$$

where $\nu_f = 3.3\sqrt{\frac{2}{N}\frac{k_f^2\sigma_s^2+\alpha D_a}{\sigma_x^2}}$. Note that P_d^f depends on M through the admissible set values of parameter α. Consequently, the max-min problem (6.36) specializes as

$$\max_{[N,M,\sigma_x^2]} \min_{[\alpha,D_a]} P_d^f\left([N, M, \sigma_x^2], [\alpha, D_a]\right), \tag{6.41}$$

where

$$\begin{cases} N \leq N_{max}, \\ M \in [0, M_{max}], \\ 0 < \sigma_x^2 \leq D_{Emax}, \\ \alpha \geq 1 - \frac{(1-k_{min})^2(\sigma_s^2+\sigma_x^2)}{D_a}, \\ D_a \leq D_{a max}. \end{cases}$$

We now turn to the attacker and defender optimum strategies.

***(i)* Optimum attack** For a given set of defender parameters $\{N, M, \sigma_x^2\}$, the detection probability $P_d^f\left([N, M, \sigma_x^2], [\alpha, D_a]\right)$ can be written as a function of the attacker parameters pair (D_a,α). A 2D plot of this function is shown in Fig.6.11(a). We see that the detection probability decreases with D_a. The optimal attack corresponds, as intuitively expected, to a maximized global distortion, $D_a = D_{a max}$. Minimizing then P_d^f over $\alpha \in [\alpha_{min}, 1]$ for given values of N, M and σ_x^2 provides the optimal scaling attack. Fig.6.11(b) clearly shows that the detection probability is maximally reduced for $\alpha = \alpha_{min}$. This corresponds to a *mixed* attack and refers to the fact that desynchronization is much more efficient than noise in impairing the detection probability for a given distortion budget. Note that without the multiple correlation procedure, the optimal solution would be $\alpha = 0$, is the so called *all desynchronization* attack. In summary :

When given the possibilities of adding white noise, desynchronizing the composite signal by constant scaling or performing both operations, desynchronization turns out to be optimal. However, to cope with appropriate defender counter-measurements (the multiple correlation test described above), the attacker is constrained to a maximum allowable attack distortion budget D_a. Thus its best strategy is first to

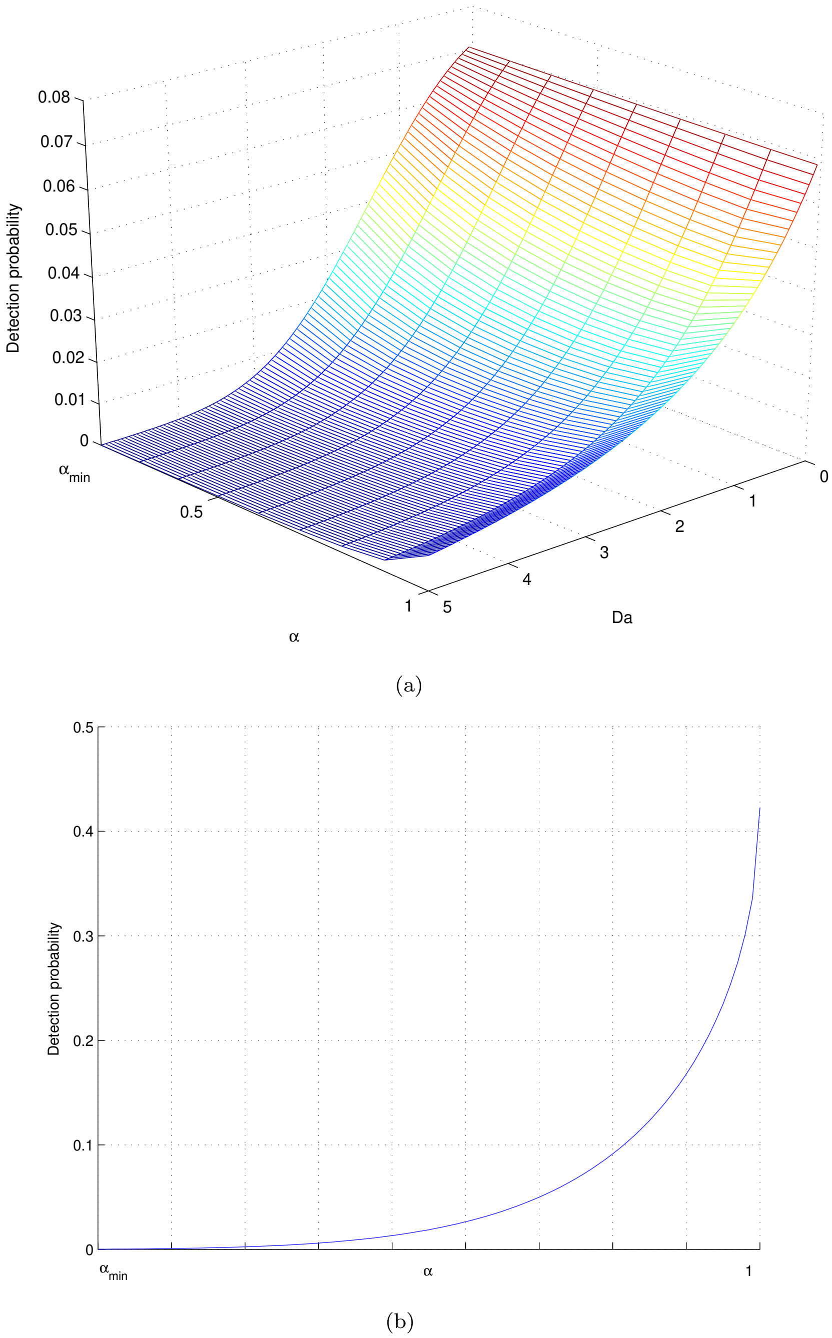


Figure 6.11: Optimum attack: detection probability has to be minimized over the set of attacker parameters $\{\alpha, D_a\}$. The 2D plot (top) shows that maximally reduced detection is obtained with large attack distortion $D_a = D_{amax}$. The corresponding detection probability $P_{\text{d}}(D_a = D_{amax})$ plotted over $\alpha \in [\alpha_{min}, 1]$ (bottom) shows that the optimal attack is *mixed*, $\alpha_{opt} = \alpha_{min}$.

desynchronize the signal, and then fulfill the remainder of the distortion budget by adding the appropriate noise amount.

As a result, the received signal corresponding to the worst attack can be expressed as

$$\mathbf{y} = \left(1 - \sqrt{\frac{(1-\alpha_{min})D_a}{\sigma_s^2 + \sigma_x^2}}\right)\mathbf{c} + \mathbf{v}_{eq}, \tag{6.42}$$

where $\mathbf{v}_{eq}$ is such that $\sigma^2_{v_{eq}} = \alpha_{min} D_a$.

Recalling that in real world scenarios, the attacker choses the parameters Δ and σ_v^2 (not k_f and α). The optimal attack turns to correspond to the combination of the following single attacks:

- a time shift $\Delta = \frac{T}{M}$,
- an additive noise of power $\sigma_v^2 = D_a - \sigma^2_{c_J}$ where $c_J(t) = c(t + \frac{T}{M})$.

(ii) **Optimized defense:** We now turn to characterize the optimized defense that best prevents the defender from the worst attack ($\alpha = \alpha_{min}$). After replacing attacker parameters by corresponding optimum values derived above, the detection probability (6.40), depending only on N, M and σ_x^2, can be wriiten as

$$P_d^f([N, M, \sigma_x^2]) = \frac{1}{2}\mathrm{erfc}\left((\nu_f - k_{min})\sqrt{\frac{N}{2}\frac{\sigma_x^2}{k^2_{min}\sigma_s^2 + \alpha D_a}}\right). \tag{6.43}$$

The aim of the defender is to maximize this worst detection probability (6.43). With qualitative considerations, we can already determine the optimum value of M : the detection probability depends on M through k_{min}. Larger values of k_{min}, corresponding to a tight range of δ, are better for the defender. Then, disregarding computational complexity, M should be maximized. An optimum defender choice would then intuitively correspond to $M = M_{\mathrm{max}}$. The resulting P_d^f depicted in Fig.6.12(a) shows that the watermark embedding power should be maximized, namely $\sigma^2_{x\,opt} = D_{Emax}$. Also, the parameter N should have the largest possible value, i.e. $N = N_{max}$. Hence, the optimum defense corresponds to the set of defender parameters chosen to be maximal ($N = N_{max}, M = M_{max}$ and $\sigma^2_{x\,opt} = D_{Emax}$). This is not surprising and is rather consolidating. One important issue, however, is to compare the robustness of the optimized defense against the *mixed attack* (shown to be optimal) to that facing the *all noise* attack. Fig.6.12(b) depicts the detection probability (6.43) for different values of the watermark power σ_x^2. We observe that:

(a) For the same watermark power σ_x^2, we have $P_d^f(\alpha = \alpha_{min}) < P_d^f(\alpha = 1)$, (the *mixed* attack is stronger than the *all noise* attack). In other words, to achieve the same detection probability, the embedding distortion of a watermark facing the mixed attack must be larger than that of a watermark facing the *all noise* attack.

(b) The slope of the detection probability curve in case of the *all noise* attack is larger than that of the mixed attack: a part of the watermark power σ_x^2 enhances the attack impact in the latter case. This fact has already been outlined in Section 6.4.1 with a non-optimized defense. Unfortunately, it remains valid with an optimized defense too.

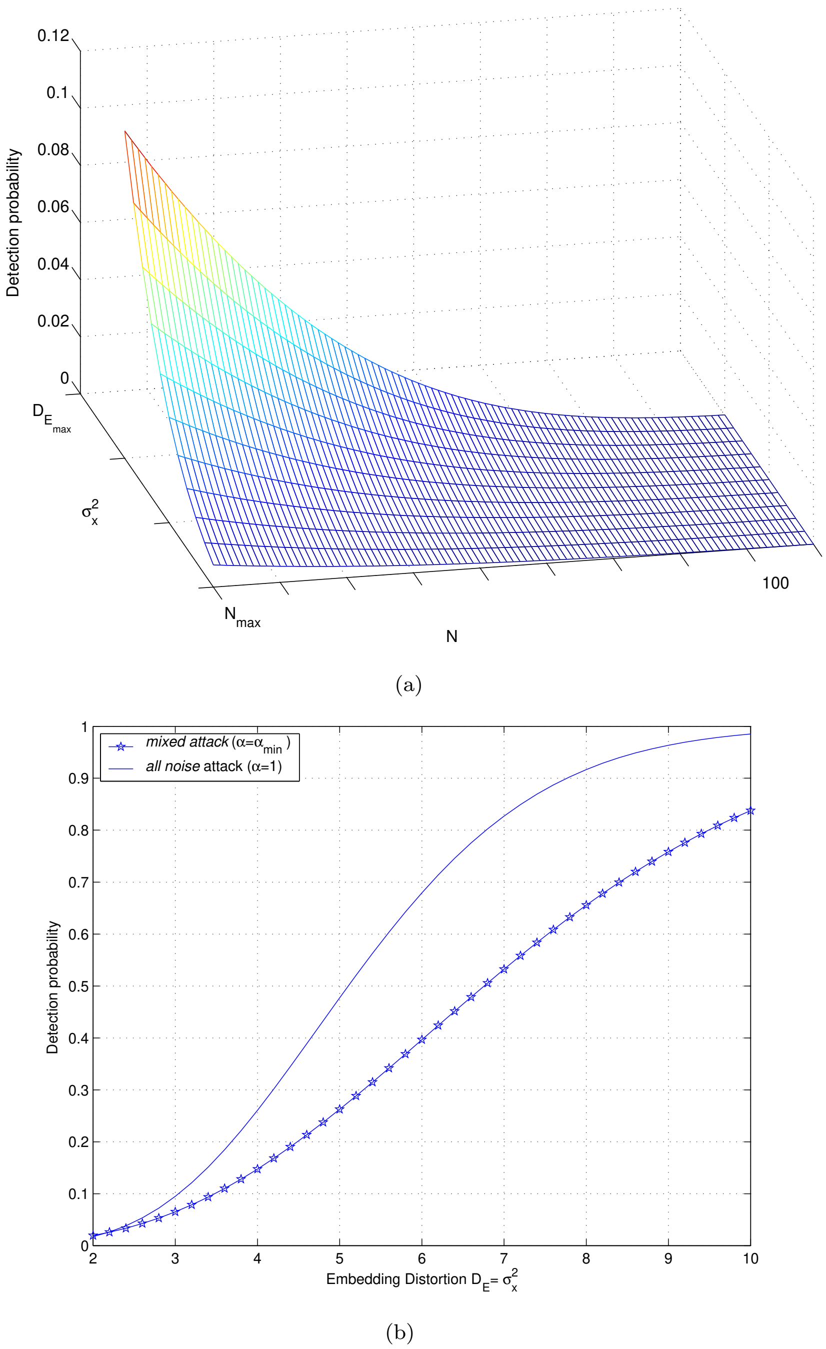


Figure 6.12: Optimum defense: detection probability has to be maximized over the set of defender parameters $\{N, D_E\}$. The 2D plot (top) shows that reliable detection is obtained with large embedding distortion $D_E = D_{Emax}$ and $N = N_{max}$. Bottom: the detection probability resulting from solving the game (*mixed* attack) is compared to that of the *all noise* attack. For the same embedding distortion, $\mathrm{Pd}(\alpha = \alpha_{min})$ is smaller than $\mathrm{Pd}(\alpha = 1)$.

6.6.3.2 Case of random Jitter

This attack has been shown to be equivalent to an additive noise $\mathbf{v}_{eq} = \mathbf{n}_r + \mathbf{v}$. Again, suppose $D_v = \sigma_v^2 = \alpha D_a$ and $D_{J_r} = \sigma_{n_r}^2 = (1-\alpha)D_a$. The resulting detection probability is

$$P_d^r = \frac{1}{2}\mathrm{erfc}\left((\nu_r - 1)\sqrt{\frac{N}{2}\frac{\sigma_x^2}{\sigma_s^2 + D_a}}\right), \tag{6.44}$$

with $\nu_r = 3.3\sqrt{\frac{2}{N}\frac{\sigma_s^2+D_a}{\sigma_x^2}}$. The threshold ν_r depends only on the global attack distortion D_a. This would suggest that, from a strict theoretical game-solving point of view, the situation is equivalent to that of the Gaussian watermarking game [CL02a] (under the hypothesis of a Gaussian jitter noise $\mathbf{n}_r$). One can see, however, that from the defender point of view fighting against a random jitter attack is more difficult than that of facing a Gaussian noise. At least, the perceived quality degradation will be greater with the jitter. This means that jittering the composite signal $\mathbf{c}$ would remain optimal from the attacker point of view. This claim is enforced by the fact that, unlike the Gaussian noise $\mathbf{v}$, the jitter noise $\mathbf{n}_r$ depends on the composite signal (as suggested by (6.21)) and is, consequently, significantly increased whenever the defender wants to combat it by increasing the watermark power σ_x^2. In addition, the host signal contributes itself to enforce the jitter effect through σ_s^2 in (6.21)). Then, attributing the hole distortion to the jitter noise ($\sigma_{n_r}^2 = D_a$) and using (6.21), the optimal jitter square deviation J must satisfy $J = \frac{D_a}{E[(\frac{d}{dt}c(t))^2]}$. The optimum defense, again, corresponds to $\sigma_x^2 = D_{Emax}$.

6.6.3.3 Discussion

Results following from the analysis above can be summarized as follows:

(i) Facing AWGN attacks, increasing the watermark power is always positive from the embedder point-of-view;

(ii) Under constant scaling attacks, two contradicting effects related to deliberately increasing the watermark power appear:

   – *a positive effect*: increasing the watermark power results in a more reliable detection.
   – *a negative effect*: increasing the watermark power enforces the desynchronization attack.

(iii) From the optimized defense analysis, one can see that even in the worst case, that is "the mixed attack", increasing the watermark power remains optimal. Expressed differently: the so called *positive effect* always overcomes the *negative effect* under constant time shift attacks.

(iv) The *multi-correlation* test alleviates the impact of the (*all desynchronization*) attack (optimum when no counter-measure is taken).

(v) Even if the random jitter behavior is noise-like, its dependency on both the host signal and the watermark makes it optimal from an attacker point of view.

6.7 Summary

In this chapter we first investigated the general watermarking channel $\mathcal{A}$. Our main motivation was to evaluate the perceived impact an attacker has on a composite signal. Our approach consists in removing from the equivalent additive signal $\mathbf{z}=\mathbf{y}-\mathbf{c}$, very often assumed to be uncorrelated with the composite signal $\mathbf{c}$, the part that is signal-like. The *equivalent* attack turns to be a particular case of well studied channel attack: attacks by filtering and additive noise. This additive noise referred to as *the desynchronization noise* has been shown to more accurately characterize the attack impact on the original composite signal quality loss. Our approach has then been applied to the desynchronization attacks modeled by attacks by jitter plus noise, the AWGN&J channel. Performance loss of the most common watermarking schemes in presence of such attacks have then been derived. Finally, we investigated optimal attacker and defender strategies in a game watermarking theory context. Results outline a somewhat intuitive result: desynchronization attacks is much more harmful than additive noise. This was the motivation for providing means to the defender to limit this contribution. Finally, the best strategies for the defender and attacker were described.

Chapter 7

Application: Secured Information Embedding in a Cellular Network System

Some of the results in this chapter have been obtained within the context of the RNRT project SDMO: Secured Diffusion of Music on mObiles in a cellular network system. The author thanks the RNRT project SDMO for funding.

In this chapter, we heavily rely on the materials stated in the previous chapters (mainly Chapter 3, 4 and 5) to efficiently implement a two-messages information embedding system. The first message is required to be fragile and is used for tamper detection. The second message is required to be robust and is used to convey ownership information. The host signal is chosen to be an audio content[1]. In the second part of this chapter, we consider more stringent security constraints and assume that the encoder has access to only a shorter description of this host (a quantized version) in analyzing system performance. The choice of the lattice, the design of the codebooks as well as the coding and decoding functions follow from the results in Chapter 3 and Chapter 4.

[1]This however, does not restrict the results and principles herein to audio contents. Embedding information into still images, text or video contents can be carried out in a straightforward manner.

7.1 SDMO Context

The project SDMO is concerned with Secured Diffusion of Music on mObiles in a cellular system. The application targets third generation (3G)-like cellular system and aims at providing efficient tools for ensuring ownership protection. In this section, we give a brief description of the system. The system is designed to track any illegal use and/or distribution of audio contents in a mobile network. In addition, the system should be able to resolve multiple ownership problems. Of course, copyright-protection using information embedding techniques is not new, for it is historically the very first targeted application. However, the novelty of the SDMO system treated in this Chapter is the use of embedded codes in a real time full industrial context. Of course, this imposes additional constraints on the design of the system.
The global architecture of the SDMO system is relatively complex and involves a large variety of technical and architectural issues such as networking, digital right management, encryption, information embedding and compression. In the following, we restrict ourselves to the information embedding part. The underlying strategy consists in embedding a watermark to identify the owner of the audio content. If, somewhere in the network (at a checking node, for example), an illegal copy is found, the owner can prove his/her paternity thanks to the embedded watermark and, potentially, can sue the illegal user in court. This perfect scenario is however likely to be disturbed by malicious users in the real world. For instance, if an attacker removes the watermark, he/she can either use or distribute the watermark-free audio signal, without any restriction. Further, he/she can even add a second watermark into the audio content and therefore claim ownership, exactly as the original owner does. Hence, in order to inhibit these malicious attackers to defeat the purpose of using information embedding as means of ownership protection, two solutions are possible: (i) allow each content owner to both embed and check for his/her own watermark and (ii) allow one single *trusted authority* to track for the illegal use and/or distribution of all the audio contents involved in the network. While (i) is potentially more secure since impairing or removing the watermark from the content of one user does not weaken, by any means, those of the other users, it is more complex to put in practice, partially because of the problem of multiple watermark claims. The problem with (ii) is that the single watermark used to mark the different contents of the different users should be very robust. Otherwise, the overall process can be defeated by (simply) breaking the system security at one point of the network. This very strong robustness is also required for another reason: to discriminate the non-watermarked contents (i.e., not copyright-protected) from the contents that have been originally protected but from which the watermark has been removed by a malicious attacker. If the watermark is robust enough, the watermark can not be removed without impairing the original content. This latter problem can be bypassed by assuming that all the contents involved in the network traffic are copyright-protected. But, this assumption is somehow limiting.
In this chapter, it is the solution (ii) which is retained. The very robust watermark is referred to as "SDMO label". This denomination refers to the fact of identifying all the protected audio contents by the same watermark, specific to the so-called third-authority. The problem of multiple watermark claims is solved by using an additional owner-specific watermark in conjunction with the so-called "SDMO label". This is because the "SDMO label", by itself, does not (uniquely) identify the content owner. Also, this is because any eventual attacker who, instead of removing the "SDMO label" adds his/her own watermark, can claim ownership. By opposition to the "SDMO label", this second watermark is asked to be fragile

and carries much more data. These data are used to convey information about the name, the affiliation of the owner and possibly a short description of the content. Note that when taken separately, neither the robust "SDMO label" nor the owner-identifying watermark could be self-sufficient in assessing security. It is precisely the aggregation of the two that allows a reasonable level of security in the cellular network.

7.2 Proposed Framework

The transmission scheme of interest is depicted in Fig.7.1. We want to embed two messages m_1 and m_2 into the same cover signal, with different robustness requirements.

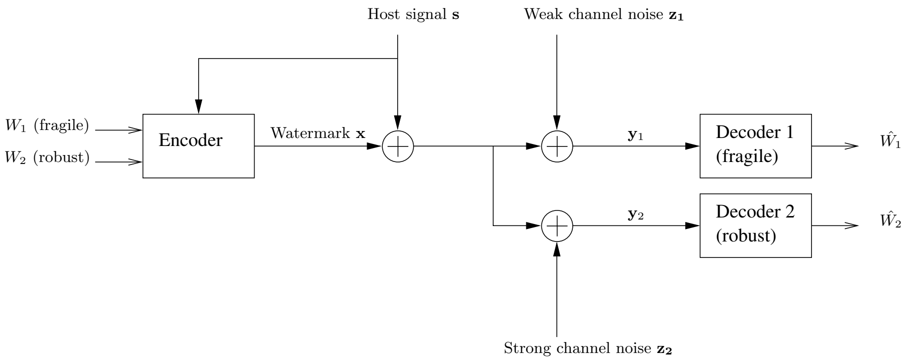


Figure 7.1: Mathematical model of the considered information embedding system.

The message m_2, representing the so-called "SDMO label", is used for detecting tampering and carries little information. Typically, m_2 comprises a few bits used to detect eventual attacks in the channel and/or to identify the marked contents. Hence, the watermark $\mathbf{x}_2$ associated with m_2 is embedded at low rate R_2 and is designed to be *sufficiently robust* to channel degradations. We denote by N_2 the per-dimension channel distortion up to which the watermark $\mathbf{x}_2$ could survive and by $\mathbf{Z}_2$ the Gaussian noise of variance N_2, i.e., $\mathbf{Z}_2 \sim \mathcal{N}(0, N_2)$. Message m_1 carries the information used to identify the owner of the audio content. Hence, m_1 is generally embedded at high rate R_1 and the watermark $\mathbf{x}_1$ associated with it is designed so as to be *fragile* or *semi-fragile*. We denote by N_1 the per-dimension channel distortion above which the watermark $\mathbf{x}_1$ should be removed and by $\mathbf{Z}_1 \sim \mathcal{N}(0, N_1)$ the Gaussian noise of variance N_1, with $N_1 \ll N_2$. In this fragile/robust framework, the decoder aims at reliably recovering the two messages, if channel conditions are not too severe. Otherwise, the decoder aims at (at least) recovering the sole message m_2 designed to be robust.

Note that here we restrict ourselves to low-rate R_2. However, in the more general setting depicted in Fig.7.1, m_2 can carry as much information as does the message m_1. Only the channel conditions for transmitting the two messages differ. In the rest of this chapter, we assume that the channel noise $\mathbf{Z}_2$ is stronger than $\mathbf{Z}_1$. Under this assumption, the information embedding channel looks like a Degraded Broadcast Channel (DBC) (see Chapter 4 for details). In the following, the system design is addressed in the very general framework where the two messages can have arbitrary rates. Monte-Carlo simulations however, are (almost all) carried out in the SDMO context, i.e., a low-rate very robust watermark $\mathbf{x}_2$ along with a high-rate fragile watermark $\mathbf{x}_2$.

7.3 System Design

Prior to encoding, the L_1-length message m_1 and the L_2-length message m_2 are mapped to two sequences of M_1-ary and M_2-ary indexes, respectively. We write $m_1 \equiv W_1^1 W_1^2 \cdots W_1^{L_1}$ and $m_2 \equiv W_2^1 W_2^2 \cdots W_2^{L_2}$, with $W_1^i \in \mathcal{M}_1 \triangleq \{1,2,\ldots,M_1\}$, $i = 1,2,\ldots,L_1$ and $W_2^j \in \mathcal{M}_2 \triangleq \{1,2,\ldots,M_2\}$, $j = 1,2,\ldots,L_2$. Encoding and decoding are lattice-based as shown in Fig.7.2. The choice of the n-dimensional lattice Λ is undertaken in the sequel. One block transmission consists in transmitting L_1 indexes W_1 and L_2 indexes W_2 within a L-length sequence $\mathbf{s}$ of the host signal. Embedding is power-constrained, by virtue of the *transparency* requirement. This means that the two watermarks $\mathbf{x}_1$ and $\mathbf{x}_2$ put on top of each other must satisfy the per-dimension power constraint P. Assuming independent watermarks, we can suppose without loss of generality that we have (per-dimension)

$$\mathbb{E}[\mathbf{X}_1^2] = (1-\gamma)P, \tag{7.1a}$$

$$\mathbb{E}[\mathbf{X}_2^2] = \gamma P. \tag{7.1b}$$

Decoder 1 receives $\mathbf{y}_1 = \mathbf{x}+\mathbf{s}+\mathbf{z}_1$ and outputs an estimate $\widehat{W_1}$ of W_1. Decoder 2 receives $\mathbf{y}_2 = \mathbf{x}+\mathbf{s}+\mathbf{z}_2$ and outputs an estimate $\widehat{W_2}$ of W_2. Performance is measured by the set of the transmission rate pairs (in bits per host sample per dimension) at which the pair of messages (m_1, m_2) (or equivalently, the pair of indexes (W_1, W_2)) is reliably recovered. The transmission rate pair (R_1, R_2) is given by

$$(R_1, R_2) = \left(\frac{L_1}{nL}\log_2(M_1), \frac{L_2}{nL}\log_2(M_2)\right). \tag{7.2}$$

Reliable recovering means recovering with sufficiently low probabilities of error $P_e^{(1)} \triangleq \Pr(\widehat{W_1} \neq W_1)$ and $P_e^{(2)} \triangleq \Pr(\widehat{W_2} \neq W_2)$. Since m_2 is required to be more robust than m_1, it is the watermark $\mathbf{x}_2$ that must be formed first. Encoding is performed according to

$$\mathbf{x}_2(\mathbf{s}; W_2, \Lambda) = (\mathbf{c}_{w_2} + \mathbf{k}_2 - \alpha_2 \mathbf{s}) \bmod \Lambda, \tag{7.3a}$$

$$\mathbf{x}_1(\mathbf{s}; W_1, \Lambda) = (\mathbf{c}_{w_1} + \mathbf{k}_1 - \alpha_1(\mathbf{s} + \mathbf{x}_2)) \bmod \Lambda. \tag{7.3b}$$

Decoding is performed according to

$$\widehat{W_i} = \underset{W_i = 1,\ldots,M_i}{\operatorname{argmin}} \|(\alpha_i \mathbf{y}_i - \mathbf{k}_i - \mathbf{c}_{w_i}) \bmod \Lambda\|, \;\; i = 1,2. \tag{7.4}$$

The set of codewords $\{\mathbf{c}_{w_i}\}$, $i = 1,2$, has cardinality $M_i = |\mathcal{M}_i|$ and forms the codebook $\mathcal{C}_{w_i}$. The goal of the following is to give insights into how efficiently design the system in Fig.7.2 such that the message m_2, which undergoes stronger channel noise, has the required robustness.
The overall problem can be formulated as follow. Given the channel conditions (N_1, N_2),

(i) select n and an n-dimensional lattice Λ with good coding and quantizing properties.

(ii) choose sequence length L_1, alphabet size M_1 and a codebook $\mathcal{C}_{w_1}$ so as to maximize the transmission rate R_1 at reasonable probability of error $P_e^{(1)}$.

(iii) choose sequence length L_2, alphabet size M_2 and a codebook $\mathcal{C}_{w_2}$ so as to minimize the probability of error $P_e^{(2)}$ at low transmission rate R_2.

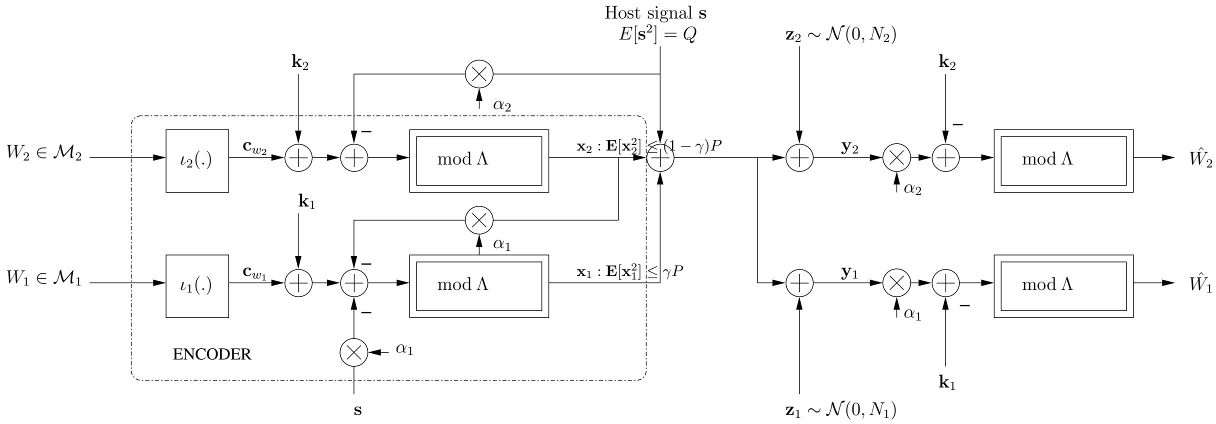


Figure 7.2: A two-users information embedding system. The message $m_2 \equiv \{W_2\}$ should be more robust than the message m_1.

We first address the choice of the lattice Λ, among the set of well known finite-dimensional lattices with reasonable quantizing complexity. For this, the two messages are embedded at the same rate.

7.3.1 Choice of lattice Λ: $R_2 = R_1$

In this section, we consider the integer lattice $\mathbb{Z}$, the hexagonal lattice A_2, the checkerboard lattice D_4 and the Gosset lattice E_8. The lattice E_8 is obtained through construction A using the first-order Reed-Muller linear code $(8, 4, 4)$ as $E_8 = (8, 4, 4) + 2\mathbb{Z}^8$. For each of the lattices $\mathbb{Z}$, A_2 and D_4, the two codebooks $\mathcal{C}_{w_1}$ and $\mathcal{C}_{w_2}$ are formed by the *relevant* deep holes of each of these lattices (see Chapter 3). This makes the transmission rate pair (R_1, R_2) vary from lattice to lattice. However, for a given lattice, the two messages m_1 and m_2 are sent at the same rate $R_1 = R_2 = R = \frac{1}{n}\log_2(M)$, where $M = N_h^* + 1$ and N_h^* is the number of relevant holes. This is ensured by setting $M_1 = M_2 = M$ and $L_1 = L_2 = L$. For the Gosset lattice E_8, the codewords $\{\mathbf{c}_{w_i}\}$, $i = 1, 2$ are chosen among the vertices of the quarter positive part of the unit cube at the origin. The curves in Fig.7.3 are obtained by setting the variance N_2 of the i.i.d. channel noise $\mathbf{Z}_2$ corrupting the transmission of m_2 to

$$N_2 = \frac{G(\Lambda)V(\Lambda)^{2/n}}{R}10^{-\mathrm{SNR}/10},$$

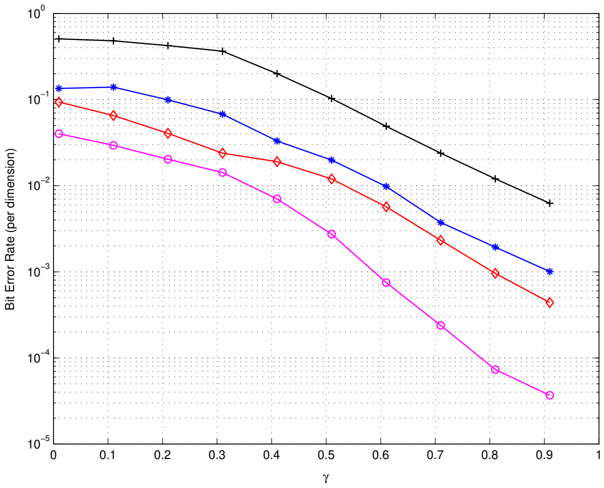
where SNR $= 7$ dB. The channel noise $\mathbf{Z}_1$ corrupting the transmission of m_1 is i.i.d Gaussian and has variance $N_1 = N_2/2$. The host signal is an audio content of length $L = 42572$ elements sampled at $F_e = 22.05$ KHz.

The values of SNR (in bit per dimension) seen respectively at "Decoder 2" (SNR_2) and "Decoder 1" (SNR_1) are given by

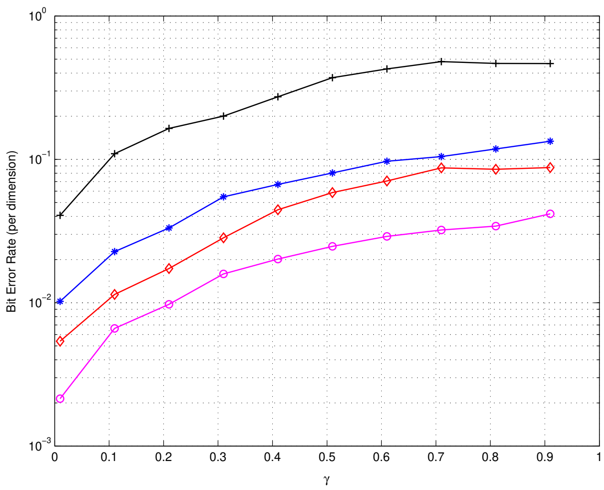
$$\mathrm{SNR}_2 = \frac{(1-\gamma)P}{R(N_2 + \gamma P)} \text{ in [dB]}, \tag{7.5a}$$

$$\mathrm{SNR}_1 = \frac{\gamma P}{RN_1} \text{ in [dB]}. \tag{7.5b}$$

As γ increases from 0 to 1, SNR_2 decreases from SNR $= 7$ dB to $-\infty$ and SNR_1 increases from $-\infty$ to SNR $+ 3 = 10$ dB. The first observation is about the overall behavior of the curves in Fig.7.3. We



(a) Error Probability $P_e^{(1)} \triangleq \Pr(\widehat{W_1} \neq W_1)$.



(b) Error Probability $P_e^{(2)} \triangleq \Pr(\widehat{W_2} \neq W_2)$.

Figure 7.3: Bit Error Probability v.s. the (per-bit per-dimension) SNR for 2-user information embedding. The curves correspond to the use of the relevant deep holes of the lattices $\mathbb{Z}$ (plus sign), A_2 (asterisk), D_4 (diamond) and E_8 (circle). (a) Error probability $P_e^{(1)}$ in decoding the message m_1. (b) Error probability $P_e^{(2)}$ in decoding the message m_2.

observe that in the range of small values of γ, the message m_2 is more reliably decoded than the message m_1, as it is expected intuitively. This is because, in this case ($R_1 = R_2$), "Decoder 2" sees a larger SNR than "Decoder 1" iff

$$0 \leq \gamma \leq \frac{-(N_1+N_2)+\sqrt{(N_1+N_2)^2+4N_1P}}{2P}.$$

As γ increases, the power allocated to the transmission of m_1 increases and that allocated to the transmission of m_2 decreases, causing SNR_1 to increase and SNR_2 to decrease. We want to assess the robustness of the two messages m_1 and m_2 to channel degradations. First, note that even in this special case where m_2 is embedded at the same rate R as m_1, the message m_2 is -by construction- more robust than m_1, simply because it is embedded first. Also, even in the extreme case where the channel is not degraded (i.e., when $N_1 = N_2$ and $\gamma = 0.5$), m_2 is slightly more robust than m_1. This is a way of saying that m_2 can not be removed without removing m_1 which is embedded on top of m_2. This is illustrated by the simulation results shown in Fig.7.3. We can see from the BER curves that the message m_2 can be decoded with the same error probability as message m_1 at lower SNR. Consider for example the BER curve corresponding to the Gosset lattice E_8: for $\gamma = 0.51$, the message m_1 is decoded with probability of error $P_e^{(1)} \approx 2.8\times10^{-3}$ at $\text{SNR}_1 \approx 7.1$ dB. The message m_2, on the other hand, is decoded with slightly lower error probability $P_e^{(2)} \approx 2.1\times10^{-3}$, obtained at slightly smaller SNR ($\text{SNR}_2 \approx 6.9$ dB). Similarly, for $\gamma = 0.11$, we have $\text{SNR}_1 \approx 0.42$ dB and $P_e^{(1)} \approx 2.9\times10^{-2}$. In transmitting the message m_2 however, the same BER is obtained with $\text{SNR}_2 \approx 0.15$ dB, only.

Discussion:

The results above show that, even if m_2 is embedded at the same transmission rate as m_1, the message m_2 is, by construction, more robust to channel degradations. In practice however, m_1 and m_2 should be embedded at different rates depending on their intended usage. For instance, since the message m_1 is required to carry a large amount of information, the primary goal in designing the codebook $\mathcal{C}_{w_1}$ is to maximize its cardinality M_1, without increasing too much the error probability $P_e^{(1)}$. On the other hand, since the message m_2 is required to survive strong channel degradations, the primary goal in designing the codebook $\mathcal{C}_{w_2}$ is to sufficiently lower the error probability $P_e^{(2)}$, without completely nullifying the transmission rate R_2. An interesting solution based on the algebraic structure of the lattice is as follows. Construct the codebook $\mathcal{C}_{w_2}$ by one-to-one mapping the codewords in $\mathcal{C}_{w_2}$ to (all or a part of) the set of the *relevant* deep holes of the lattice (see chapter 3 for details) and, construct the codebook $\mathcal{C}_{w_1}$ by one-to-one mapping the codewords in $\mathcal{C}_{w_1}$ to (all or a part of) the set of the *relevant* kissing points of the lattice. Of course, this simple solution obtained by just exploiting the structure of the lattice may require few changes to allow large rate R_1. For instance, if the number of *relevant* kissing points (denoted by N_k^* in Chapter 3) is such that

$$N_k^* \leq 2^{nR_1L/L_1} - 1,$$

this simple solution can be improved by *carefully* choosing additional *coset-leader* codewords $\mathbf{c}_{w_1}$ inside the Voronoi region $\mathcal{V}(\Lambda)$. In general, the *careful* design of the two codebooks $\mathcal{C}_{w_1}$ and $\mathcal{C}_{w_2}$ must be such that:

1. The codewords (i.e., the *coset leaders*) are selected in such a way that (3.19) and (3.20) are satisfied. This is a way of saying that each codeword must identify a unique coset of the lattice and, conversely, that each coset is identified by a unique codeword in the codebook (see Chapter 3).

2. The inter-cosets minimum-distance d_{min} defined as in (3.16) is large enough (see Chapter 3).

3. The targeted transmission rate pair (R_1, R_2) is in the feasible capacity region[2]. Otherwise, the error probability pair $(P_e^{(1)}, P_e^{(2)})$ could not be sufficiently lowered, even if powerful channel coding techniques are used.

The problem of codebook design briefly invoked here is an instance of the more general and more complex problem of *rate/power allocation* in multiuser environments. A fundamental question is that of how *optimally* partition the power budget P available at the transmitter so as to *optimally* allocate the amount of information that could be reliably sent to each user. In a general setting, this is a very difficult task. In practice, the *optimal* allocation depends on the targeted application. In the application considered in this chapter, optimally allocating the power amounts to finding the optimal choice of the parameter γ in (7.1). Another important question is that of how efficiently allocate the different transmission rates, through for example using different channel coding techniques. Here, appropriately allocating the two transmission rates R_1 and R_2 amounts to appropriately choosing the parameters L_1, L_2, M_1 and M_2. Basically, the problem of rate/power allocation is closely related to the classical trade-off problem of transmission-rate/probability-of-error or equivalently, to that of *payload/robustness*. In order to be consistent with the SDMO context, we will restrict our attention in the following section to the design of the codebook $\mathcal{C}_{w_2}$ so as to make the transmission of the message m_2 very robust. Also, we will retain the Gosset lattice for the modulo-reduction, for it provides the smallest error probability pair $(P_e^{(1)}, P_e^{(2)})$ as it can be seen from Fig.7.3.

7.3.2 Design of codebook $\mathcal{C}_{w_2}$ for $R_2 \ll R_1$

Since the codebook $\mathcal{C}_{w_2}$ must be designed such that the signal $\mathbf{x}_2$ is very robust to channel degradations, relevant lattice holes are good candidates for the choice of the coset leaders (see Chapter 3). However, in order to lower the transmission rate R_2, only few holes among all lattice holes should be selected. Typically, the codebook $\mathcal{C}_{w_2}$ contains just two elements (i.e., one single bit used to detect the presence/absence of the watermark) chosen as far apart as possible. Fig.7.4(a) depicts the BER curve corresponding to the uncoded binary transmission of the message m_2 using the most far-way (in a hamming distance sense) relevant deep holes of the Gosset lattice E_8. This BER curve is plotted versus the parameter γ. The values of SNR_2 are not indicated but can be found using (7.5a). For example, observe that for $\gamma \approx 0.01$, we have $\text{SNR}_2 \approx 10.5$ dB and the probability of error is $P_e^{(2)} \approx 1.1 \times 10^{-4}$, which is significantly lower than the values that could be obtained by for example the integer lattice. However, for a transmission rate of $R_2 = \frac{1}{8}$ bit per host-sample per dimension (n, M and L_2 are set to $n = 8$, $M = 2$ and $L_2 = L$ in (7.2), this error probability is not sufficiently low to ensure the required level of robustness. To further reduce this error probability while keeping R_2 reasonable , one should select $M > 2$ appropriate codewords from the set of relevant holes of the lattice and then rely on powerful channel coding techniques. However, powerful channel coding techniques have high computing complexity in their non-binary form. For example, the attempt to use powerful non-binary Turbo codes and non-binary LDPC codes is contrasted

[2]This has been taken into account in the simulations carried out in this chapter. Refer to Chapeter 4 for the feasible capacity region of the resulting Degraded BC.

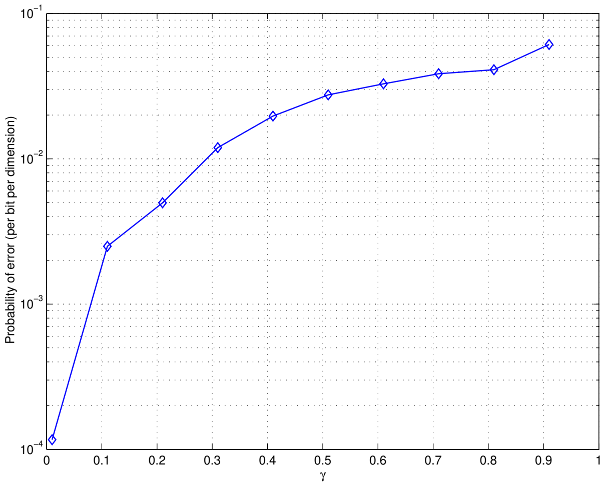
with their huge coding and decoding complexities [BJDK01] in their non-binary form. The simplest coding technique consists in repeating each index $W_2 \in \{1, 2, \ldots, M_2\}$ several times (say ρ times, for example). Of course the transmission rate is divided by ρ but the transmission is strengthened since each "symbol" is transmitted ρ times. Also, repetition coding is retained due to its efficiency at very-low to low SNRs. BER curves corresponding to non-binary transmission using different values of the repetition factor ρ are depicted in Fig.7.4(b) where $\mathcal{M}_2 = \{1, 2, \ldots, 5\}$. We observe that lower error probability $P_e^{(2)}$ (in comparison to the uncoded binary transmission stated above) is made possible. For instance, for $\gamma = 0.01$, we obtain $P_e^{(2)} \approx 7.5 \times 10^{-5}$.

The problem of large coding complexity raised in the non-binary case is no longer of concern if a binary alphabet $\mathcal{M}_2$ is used (but alas, the transmission rate R_2 is maximally reduced). The curves depicted in Fig.7.4(c) are obtained by combining turbo coding and repetition coding. The use of turbo coding is motivated by the nature of the information embedding channel which may be modelled (under certain circumstances) with a time-varying fading channel. For instance, "localized" channel attacks where some "parts" of the embedded signal may undergo more degradations than other parts of the same signal, are fading-like. In Fig.7.4(c), we use a rate 1/4 Recursive Serial Concatenated (RSC) turbo code followed by a repetition code with repetition factor $\rho = 4$. It can be seen that for $\gamma = 0.01$, we have $P_e^{(2)} \approx 4.5 \times 10^{-6}$ which can be considered as sufficiently small to allow a high level of robustness against channel degradations. Even smaller error probabilities can be obtained by means of stronger channel coding techniques. However, this is in general achieved at the cost of higher coding complexity.

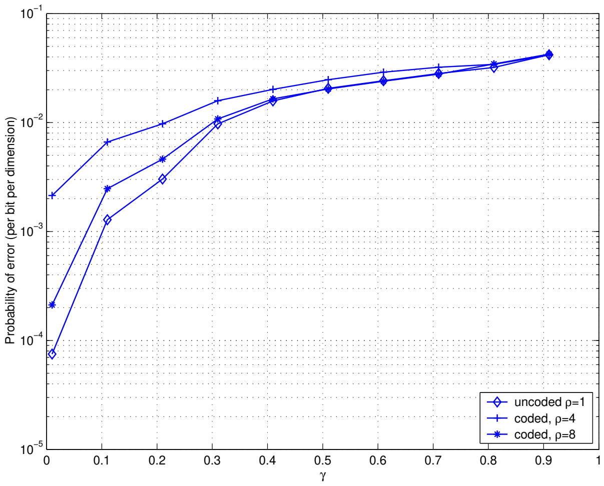
7.4 Embedding Using a Short Description of The Host

So far, we assumed that the encoder has full knowledge of the host signal in the encoding process. For instance, we assumed that the so-called *trusted* authority has perfect knowledge of the host signal $\mathbf{s}$ while embedding the "SDMO label". Also, we assumed that the content owner has perfect knowledge of the composite signal $\mathbf{s} + \mathbf{x}_2$ while embedding his/her specific identifying mark. While the first assumption (regarding the embedding "SDMO label") is reasonable, the second is not, simply because the so-called *trusted* authority may not want the content owner to have full access to the watermark $\mathbf{x}_2$. In fact, in applications where security is a central issue and where embedding is carried out by two different entities, these two entities which (together) form the channel encoder, may not fully-cooperate. In the SDMO context considered in this chapter, it is precisely the mobile phone operator which has to embed and check for the robust watermark, to ensure that the transmission over its deployed network is not being illegally exploited by any malicious attacker. The operator should perform this so as to ensure the content providers as for the safety of the proposed service.

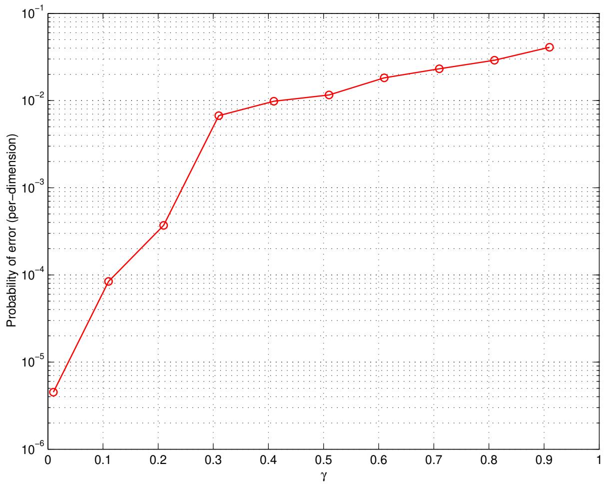
The content owner, which embeds his/her watermark last, needs not (and must not) have access to the "exact" composite signal $\mathbf{c} = \mathbf{s} + \mathbf{x}_2$. Otherwise, he/she needs only subtract his/her host signal $\mathbf{s}$ to get the watermark $\mathbf{x}_2$. Of course, even in the case where the content owner knows $\mathbf{x}_2$, he/she can not easily get the message m_2 or the key $\mathbf{k}_2$. However, this may be possible under certain specific circumstances (by processing collusion attacks, for example). Hence, one way to circumvent this problem is to assume that the content owner (generally, the one who embeds last) has access to only a *short description* of



(a) Rate $R_2 = \frac{1}{n}$ bit per host-sample per dimension.



(b) Rate $R_2 = \frac{1}{n\rho} \log_2(M_2)$ bit per host-sample per dimension. $M_2 = 5$.



(c) Rate $R_2 = \frac{R_c}{n\rho} \log_2(M_2)$ bit per host-sample per dimension. $R_c = 1/4$, $\rho = 4$ and $M_2 = 5$.

Figure 7.4: Error Probability $P_e^{(2)}$. The role of channel coding in strengthening the transmission of the robust watermark. (a) uncoded binary transmission. (b) Non-binary transmission with repetition coding. (c) Non-binary transmission with combined turbo-coding and repetition coding.

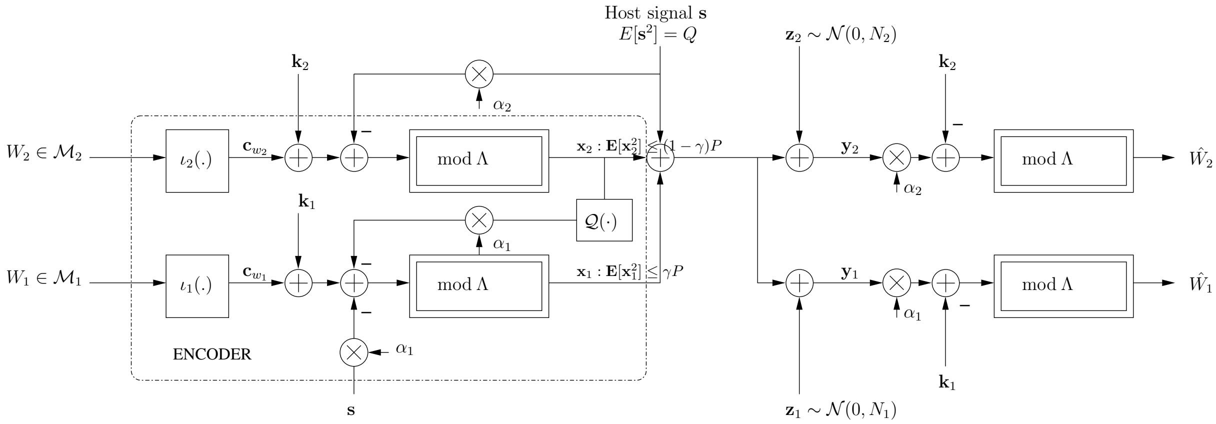


Figure 7.5: A two-users information embedding system with partial host at the encoder.

the already watermarked signal $\mathbf{c} = \mathbf{s} + \mathbf{x}_2$. The term "short description" refers to the partial knowledge of the signal $\mathbf{x}_2$. This short description can be, for example, a *quantized version* $\widetilde{\mathbf{x}_2} \triangleq \mathcal{Q}(\mathbf{x}_2)$ of the watermark signal $\mathbf{x}_2$, where $\mathcal{Q}(\cdot)$ is some quantizer unknown to the content owner. From a transmission point-of-view, this partial knowledge of the host (signal $\mathbf{s}+\mathbf{x}_2$) by the content owner can be interpreted as an additional noise-like *uncertainty* upon the channel. From a side-information communication point-of-view, this uncertainty can be interpreted as a partial knowledge of the state information at the encoder. Viewed as such, this causes the system performance (mainly, the transmission rate R_1) to decrease, as previously mentioned in Chapter 5. This is addressed in the next section, after a brief discussion of the general assumptions made above.

Discussion:

The problem raised above related to "who embeds what" or "who has access to what" is a problem of Digital Right Management (DRM) and is, thus, outside the scope of this work. However, from a communication point-of-view, it is important to take it into account in the design of the system. For example, it is precisely based on this preliminary qualitative study that embedding is recognized either as being broadcast-like (BC-like) or rather, as being MAC-like. The embedding and decoding functions are also designed based on that preliminary studies. In the situation at hand, taking into account the partial cooperation between the two embedding entities at the encoder (for security purposes) has wrongly made the encoding process look as is if it were MAC-like. This seemingly MAC-like transmission is in fact rather broadcast-like. The reason is as follows. First, checking for the two watermarks (the decoding process) is not performed at the same "node point" of the cellular network and hence, can not be performed in a joint manner as it is the case in MAC scenarios. Second, even though the encoding procedure is carried out in two steps by two different entities, it can always be viewed as a single "big" encoder composed of two interacting *sub-encoders*, as it is common in more conventional multi-antenna broadcast systems.

7.4.1 Performance analysis

The (new) communication model is depicted in Fig.7.5. We denote by $\widetilde{\mathbf{x}_2}$ the short description (quantized version) of the signal $\mathbf{x}_2$. This short description of $\mathbf{x}_2$ is used in embedding the message $m_1 \equiv W_1^1 W_1^1 \cdots W_1^{L_1}$, where $W_1^i \in \mathcal{M}_1$ and $i = 1, 2, \ldots, M_1$. Let $\mathbf{e}_2 \triangleq \mathbf{x}_2 - Q(\mathbf{x}_2) = \mathbf{x}_2 - \widetilde{\mathbf{x}_2}$ denote the error incurred in quantizing the embedded signal $\mathbf{x}_2$. Under appropriate assumptions (high resolution quantization), the error $\mathbf{e}_2$ is statistically independent of $\mathbf{x}_2$ and can be viewed as a small perturbation to the true host[3] signal $\mathbf{s} + \mathbf{x}_2$, in the sense mentioned in Chapter 5. Under this condition, the encoding functions (7.3) become

$$\mathbf{x}_2(\mathbf{s}; W_2, \Lambda) = (\mathbf{c}_{w_2} + \mathbf{k}_2 - \alpha_2 \mathbf{s}) \bmod \Lambda, \tag{7.6a}$$

$$\mathbf{x}_1(\mathbf{s}; W_1, \Lambda) = (\mathbf{c}_{w_1} + \mathbf{k}_1 - \alpha_1(\mathbf{s} + \widetilde{\mathbf{x}_2})) \bmod \Lambda. \tag{7.6b}$$

The decoding functions remain unchanged, i.e., given by (7.4). Due to the above mentioned channel uncertainty, the theoretically maximal feasible transmission rate R_1 in the rate pair $(R_1(\Lambda), R_2(\Lambda))$ given by (4.19) drops to

$$R_1(\Lambda)^{(\max)} = \max_{\alpha_1} \frac{1}{n} \left(\log_2(V(\Lambda)) - h(\widetilde{\mathbf{V}_1}) \right) < \frac{1}{2} \log_2 \left(1 + \frac{\gamma P}{N_1 + \sigma_{e_2}^2} \right), \tag{7.7}$$

where $\widetilde{\mathbf{V}_1} = (\alpha_1 \mathbf{Z}_1 + \alpha_1 \mathbf{E}_2 - (1 - \alpha_1)\mathbf{X}_1) \bmod \Lambda$ is the new equivalent noise in the equivalent modulo channel. The maximal feasible transmission rate R_2, as for it, remains unchanged, i.e, given by (4.19). In contrast to the previous section where we concentrated on the transmission of the message m_2 and evaluated both the transmission rate R_2 and the error probability $P_e^{(2)}$, we concentrate in this section on the transmission of the message m_1. We want to analyze the decrease in the transmission rate R_1 or equivalently the increase in the probability of error $P_e^{(1)}$, due to the new situation. Fig.7.6 depicts the increase in the error probability caused by the partial knowledge of the host at the encoder for the encoding of m_1, for the cubic lattice $\mathbb{Z}^n$ and the Gosset lattice E_8. Naturally, the stronger the perturbation (i.e., the larger $\sigma_{e_2}^2$), the worse the decoding capability. Conversely, the more accurate the host description provided to the encoder (i.e., the smaller $\sigma_{e_2}^2$), the better the system performance. We observe also from Fig.7.6 that for a given level of accuracy in the knowledge of the host (i.e., for given $\sigma_{e_2}^2$), the decrease in the error probability $P_e^{(1)}$ (i.e, the gap to the performance obtained in the case where the encoder has perfect knowledge of the host) increases with SNR_1. The reason is as follows. The influence of the so-called perturbation of the host can be viewed as the introduction of an additional noise term $\mathbf{e}_2$ as it can be seen from (7.7). As the parameter γ increases, SNR_1 increases and the power N_1 of the ambient noise $\mathbf{z}_1$ decreases. Thus, the contribution of the perturbation term $\mathbf{e}_2$ to the total noise $\mathbf{e}_2 + \mathbf{z}_1$ increases, comparatively.

7.5 Summary

In this chapter, we provided a lattice implementation of a two-users information embedding scheme directed to be used in a real-time digital watermarking system. The application that we considered

[3]Note that the already watermarked signal $\mathbf{s} + \mathbf{x}_2$ is considered as host (side information) in embedding the message m_1.

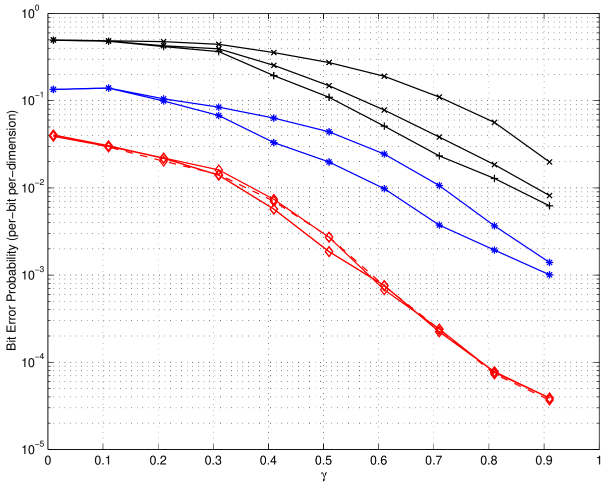


Figure 7.6: Decrease in the transmission rate R_1 due to the partial knowledge of the host at the encoder. Error probabilities corresponding to the use of the cubic lattice $\mathbb{Z}^n$ (plus sign), the Hexagonal lattice A_2 (Asterisk) and the Gosset lattice E_8 (diamond) are measured in the case when the encoder has access to only a quantized version of the host signal, in encoding the message m_1. The error probability $P_e^{(1)}$ increases with the strength of the perturbation: from bottom to top, $\sigma_{e_2}^2 = 0, (1-\gamma)P/1000, (1-\gamma)P/100$ and $(1-\gamma)P/10$.

consists is Secure Diffusion of Music on mObiles (RNRT project SDMO) in a cellular network system. Two watermarks are simultaneously embedded into the same host content. The first watermark carries much information and is used for ownership identification. It is required to be fragile. The second watermark (referred to as "SDMO label") carries little information used for tamper detection and is required to be very robust. The overall transmission scheme is broadcast like and guidelines from Chapters 3 and 4 are used for system design. Monte-Carlo based BER simulations for the resulting scheme pointed out an interesting problem of *rate/power allocation* that may be encountered in all broadcast situations, in real world scenarios. First, there is the problem of rate assignment. This amounts to choosing the appropriate functioning "point" in the capacity region of the corresponding two-users broadcast scheme. Second, there is the problem of codebook selection. In general, this is a difficult task. In the proposed framework, this is addressed through some examples. Third, there is the problem of the choice of the appropriate channel coding strategy. Again, this is addressed by applying some sub-optimal well known codes. This is because powerful codes like LDPC and Turbo codes have large complexity in their non-binary form. The fourth problem is information-embedding specific. It concerns security issues: in security-demanding applications, the encoder should have access to only a partial knowledge of the host.

Chapter 8

Conclusion and Future Work

We conclude this thesis by briefly summarizing some of the main results and commenting on several promising directions for future research.

About ten years after its infancy, information embedding is still considered as a young technology. Also, as any young technology, it still has its (many ?) own weaknesses. But, it also has its own strengths. The primary goal of this thesis is to go one step further in solving some of these weaknesses. To this end, many guidelines from data transmission and data compression are used to solve information embedding problems. This highlights the potential use of the strong background of these well developed conventional area in information embedding. The results obtained in this context are summarized in Section 8.1. The second goal of this work was to illustrate the potential use of information embedding techniques and principles in solving problems raised in conventional data transmission and data compression. This is summarized in Section 8.2.

8.1 Concluding Summary

Information embedding plays a key role in addressing a major challenge that arose from the widespread distribution of multimedia content over digital communication network: secure transmission of data. In this two-sided challenge, two antagonist problems have to be solved: security and data transmission. Of course, for data to be securely transmitted, there must be little information. Conversely, for these same data to be transmitted at high rate, the security level must be lowered. This thesis addresses these two problems from the communication and information-theoretic point-of-views. It is shown that solutions to the security problem can be found while dealing with the transmission problem, in the form of multiple-user information embedding for example. The main results can be summarized as follows.

1. Based on the well known Quantization Index Modulation (QIM) technique as well as on the famous Scalar Costa Scheme (SCS), the first part of this thesis focuses on the design of lattice-based algebraic codes for the problem of information embedding. We designed the resulting codebooks to be modulo-reduction-based and showed that these should be carefully designed. The careful design concerns the lattice selection, the parameters setting as well as the choice of the codewords (i.e. *coset leaders*) for the construction of the *cosets*. This problem is first addressed through some examples using the appealing algebraic structure of the lattice and then, through a more general approach using insights from *shaping* for multidimensional constellations in conventional communication. It is recognized that Costa-based information embedding is a joint source-channel coding problem. We then used insights from Erez, Shamai and Zamir's work on nested lattices to design good source and channel codes. For instance, we evaluated the system performance obtained with a finite-dimensional nested-lattices and showed that in this nested structure, by opposition to infinite-dimensional coding, the two components of the overall gain provided by lattice coding (the *shaping gain* and the *coding gain*) are not-decoupled but rather interact. This was illustrated by some sub-optimal constructions of nested-codes using important results from coding theory.

2. The second part of this thesis extended the initial QIM and SCS schemes to the multi-user case, using guidelines from coding for Broadcast and MAC channels with side information at the encoder(s). Multiple information embedding is recognized as being equivalent to one of two channels for which recent theory is well developed: the Degraded Broadcast Channel (DBC) and Multiple Access Channel (MAC). The problem is first addressed using scalar codebooks and then using more involved lattice-based codebooks. For instance, it is shown that *appropriately* designed, embedded lattice codes allow simultaneous reliable transmissions. The appropriate design involves the *embedding order* as well as the construction of the codebooks.

3. The third part of this thesis deals with coding with partial side information. The general framework of channel sensitivity to partially known two-sided state information is addressed by evaluating the loss in channel capacity (or transmission rate) due to some small noise-like perturbation. Also, lower and upper bounds on this channel sensitivity are provided using the famous *De-Bruijn identity*. Then, particular emphasis is put on the special case of one-sided state information which is partially known to the encoder. We showed that, in certain circumstances, the encoder should *adapt* to the situation by, eventually, changing its coding strategy. The resulting scheme is more robust to noise and overall performance is improved.

4. The fourth part of this thesis is more application-oriented and is concerned with side information coding analysis when the encoder and the decoder are not *fully* synchronized. The situation is modelled with side-information transmission over an Additive White Gaussian Noise and Jitter (AWGN&J) channel. Using a game theoretic analysis, it is shown that desynchronization is more *harmful* to reliable transmission. From an encoder (embedder) point-of-view, this means that

combating channel desynchronization should be considered first. To this end, we provided a simple correlation-based algorithm for enforcing Spread-Spectrum (SS) based information embedding against this type of channel degradations.

8.2 Extensions and Future Work

In this section, we comment on several directions for extending the results of this thesis and discuss possible approaches to tackling some remaining open problems.

8.2.1 Extensions

In both single-user and multi-user cases, the design of the codebook has assumed transmission over a Gaussian channel. While this is widely assumed in communication theory, it is not likely to hold in certain practical situations. In fact, in real life information embedding, channel noise is not strictly Gaussian. More involved channel models already exist. For these, the design of the lattice-based codebook may be slightly changed. Possible extensions may concern the study of the optimal strategy to adapt to this situation. For instance, how do the system performance vary according to a non-Gaussian channel noise? and, how could the encoding/decoding strategies be adapted accordingly?

8.2.2 Use for conventional data transmission

The problems raised in this thesis, and specially those in the chapters 3, 4 and 5 are closely related to those that may be encountered in real Broadcast and MAC situations. For these channels, the problems of *power allocation*, *rate allocation* and *codebook design* are of primary concern. Different conventional techniques have been used in the past, like the Decision-Feedback-Equalizer (DFE) presented in Chapter 2 for example. However, some of the problems mentioned above have not been solved yet. Embedded codes (i.e., Dirty Paper Codes, DPC) have the potential to cope with these situations. This is specially due to the DPC-based *successive encoding* at the encoder in the BC and the DPC-based *successive decoding* or *peeling-off technique* at the decoder in the MAC. Though slightly different (the state information is not strictly non-causal and it is also possibly non-Gaussian), the lattice-based constructions designed in this thesis could be slightly changed so as to be used in real conventional multi-user environments.

8.2.3 Use for conventional data compression

Another possible area of future exploration is the information-theoretic duality between information embedding and so-called Wyner-Ziv source coding, which is lossy source coding with side information at the decoder. For example, the information-embedding capacity has a rate-distortion counterpart in the Wyner-Ziv problem. Similarly, the lossless version of Wyner-Ziv source coding, called Slepian-Wolf source coding, is the dual of the noise-free information embedding problem.

As a result of this duality, one can use insights from the design and analysis of information embedding systems to better design and analyze Wyner-Ziv source coding systems, and vice versa. For instance, the nested code structure provided in Chapter 3 has potential use in designing a good Wyner-Ziv system, by swapping the roles of the encoder and the decoder. Also, the problem of multi-user transmission with side information at the encoder considered in Chapter 4 has potential use in its counter-part problem of *sensor-networking* where a single remote source has to jointly decode information gathered from separate encoding source nodes.

Appendix A

Short Review of Strong Typical Sequences

Let (X_i, Y_i) be drawn i.i.d. according to a joint probability mass function $p(x, y)$. Let $\mathcal{X}$ and $\mathcal{Y}$ the corresponding sets and $p(x)$ and $p(y)$ the marginals of X and Y respectively.

Definition 1 (Typical and Strongly Typical Set) *Let $\mathcal{X}$ be a finite set and $p_X(x)$ a given probability distribution over this set. Let $a \in \mathcal{X}$, $i \in \{1, 2, \ldots, n\}$ and $x^n = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ and $N(a|x^n) = |\{i : x_i = a\}|$. The set $\mathcal{T}_X^n(\delta)$ of typical sequences is defined as:*

$$\mathcal{T}_X^n(\delta) \triangleq \left\{x^n \in \mathcal{X}^n \,:\, |n^{-1}N(a|x^n) - p_X(a)| < \delta, \textit{for all}\, a \in \mathcal{X}\right\}. \tag{A.1}$$

If, in addition, for all $a \in \mathcal{X}$ with $p_X(a) = 0$ we have $N(a|x^n) = 0$, the sequences in $\mathcal{T}_X^n(\delta)$ are called δ-strongly typical.

Lemma 3 ([CT91]) *The typical set $\mathcal{T}_X^n(\delta)$ with respect to $p_X(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the following property*

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

Lemma 3 is a way of saying that for sufficiently large n, all n-sequences are "almost equally surprising", each with probability $p(X_1, X_2, \ldots, X_n) \approx \exp -n(H \pm \epsilon)$. Roughly, the number of these n-sequences in the set $A_\epsilon^{(n)}$ is $|A_\epsilon^{(n)}| \approx \exp -n(H \mp \epsilon)$. The typical sequences, with respect to $p(x)$, can be understood as being the most probable sequences, among all possible n-sequences. Most of the attention in information theory is on such sequences. Any property that is proved for the typical sequences will then be true, with high probability, and will determine the average behavior of a large sample.

Definition 2 (Strongly Typical Set of a Pair of Sequences) *Let $(a, b) \in \mathcal{X} \times \mathcal{Y})$, $(i, j) \in \{1, \ldots, n\} \times \{1, 2, \ldots, n\}$ $N(ab|x^n y^n) = |\{(i, j) \,:\, (x_i, y_j) = (a, b)\}|$. The set $\mathcal{T}_{XY}^n(\delta)$ of jointly typical sequences is defined as:*

$$\mathcal{T}_{XY}^n(\delta) \triangleq \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \,:\, |n^{-1}N(ab|x^n y^n) - p_{XY}(a, b)| < \delta, \textit{for all}\, (a, b) \in \mathcal{X} \times \mathcal{Y}\right\}. \tag{A.2}$$

If, in addition, for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p_{XY}(a,b) = 0$ we have $N(ab|x^n y^n) = 0$, x^n and y^n are called jointly δ-strongly typical.

Lemma 4 ([CT91]) *The set $\mathcal{T}^n_{XY}(\delta)$ of jointly typical sequences (x^n, y^n) with respect to the distribution $p(x,y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies, i.e.,*

$$\begin{aligned} \mathcal{T}^n_{XY}(\delta) &= \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : & \text{(A.3)} \\ & |-\frac{1}{n}\log p(x^n) - H(X)| < \epsilon & \text{(A.4)} \\ & |-\frac{1}{n}\log p(y^n) - H(Y)| < \epsilon & \text{(A.5)} \\ & |-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)| < \epsilon\}, & \text{(A.6)} \end{aligned}$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$

Lemma 4 is a way of saying that the number of sequences X and Y that are jointly typical is about $\exp\{nH(X;Y)\}$. Hence, since there are about $\exp\{nH(X)\}$ typical X sequences, about $\exp\{nH(Y)\}$ typical Y sequences and only about $\exp\{nH(X;Y)\}$ jointly typical sequences, the probability that any randomly chosen pair is jointly typical is about

$$\frac{\exp\{nH(X;Y)\}}{\exp\{nH(X)\} \times \exp\{nH(Y)\}} = \exp\{-nI(X;Y)\}. \tag{A.7}$$

Definition 3 (Strongly Conditionally Typical Set) *Let $W : \mathcal{X} \longrightarrow \mathcal{Y}$ be a conditional distribution. Let x^n be δ-strongly typical sequence. For $\delta_1 \geq \delta > 0$ the set $\mathcal{T}^n_W(x^n, \delta_1)$ of conditionally typical sequences given x^n is defined as:*

$$\mathcal{T}^n_W(x^n, \delta_1) \triangleq \left\{ y^n \in \mathcal{Y}^n \times \mathcal{Y}^n \ : \ |n^{-1}N(ab|x^n y^n) - n^{-1}N(a|x^n)W(b|a)| < \delta_1, \textit{for all}\, (a,b) \in \mathcal{X} \times \mathcal{Y} \right\}. \tag{A.8}$$

If, in addition, for all for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $W(b|a) = 0$ we have $N(ab|x^n y^n) = 0$, x^n and y^n x^n and y^n are called δ-strongly conditionally typical.

Appendix B

Some Results on Broadcast Channels

This appendix provides a brief review of some results on the discrete memoryless Broadcast Channel (BC), the Gaussian BC, the physically degraded BC and the the physically degraded BC with state information at the encoder. The key ideas for establishing the capacity regions are outlined. The complete proof of the capacity region of the Gaussian BC can be found in [CT91]. The capacity region of the Gaussian BC with state information at the transmitter has been established in [KSS04].

B.1 Broadcast Channel (BC) and Degraded BC

A two-users broadcast channel is illustrated in Fig.B.1(a). The transmitter has power P and wishes to send independent messages W_1 (at rate R_1) and W_2 (at rates R_2) to two distant receivers Y_1 and Y_2. The received sequences are $\mathbf{Y}_1 = \mathbf{X} + \mathbf{Z}_1$ and $\mathbf{Y}_2 = \mathbf{X} + \mathbf{Z}_2$, where $\mathbf{Z}_1$ (of power N_1) is the channel noise corrupting the transmission of W_1 and $\mathbf{Z}_2$ (of power N_2) is the channel noise corrupting the transmission of W_2. Without loss of generality, we assume that $N_1 < N_2$. Thus, receiver Y_1 is less noisy than receiver Y_2. Formally, a broadcast channel is defined as follows.

Definition 4 *A broadcast channel consists of an input alphabet $\mathcal{X}$, two output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ and a probability transition function $p(y_1, y_2|x)$. The broadcast channel will be said to be memoryless if $p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{x}) = \Pi_{i=1}^{n} p(y_{1i}, y_{2i}|x_i)$, where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})$, $i = 1, 2$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.*

A $((2^{nR_1}, 2^{nR_2}), n)$ code for the broadcast channel with independent information consists of an encoder

$$\mathbf{X} \; : \; (\{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\}) \longrightarrow \mathcal{X}^n, \tag{B.1}$$

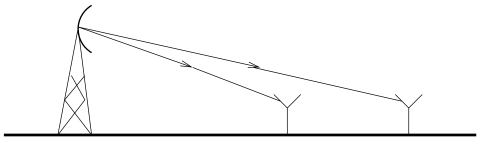
and two decoders

$$g_1 \; : \; \mathcal{Y}_1{}^n \longrightarrow \{1, 2, \ldots, 2^{nR_1}\}, \tag{B.2a}$$

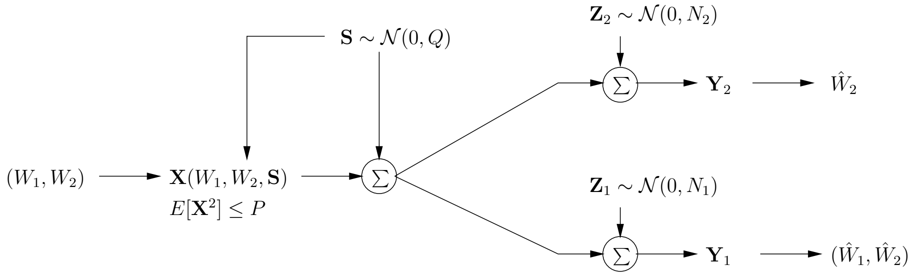
$$g_2 \; : \; \mathcal{Y}_2{}^n \longrightarrow \{1, 2, \ldots, 2^{nR_2}\}. \tag{B.2b}$$

The average probability of error is defined as

$$P_e^{(n)} = \Pr\left(g_1(\mathbf{y}_1) \neq W_1\right) \text{ or } g_2(\mathbf{y}_2) \neq W_2\right), \tag{B.3}$$



(a) A Broadcast Channel (BC).



(b) A Gaussian BC with State Information at the Transmitter.

where (W_1, W_2) are assumed to be uniformly distributed over $\{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\}$.

Definition 5 *A rate pair (R_1, R_2) is said to be achievable for the broadcast channel if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \longrightarrow 0$.*

Definition 6 *The capacity region of the broadcast channel is the closure of the set of all achievable rates.*

Definition 7 *A broadcast channel is said to be physically degraded if $p(y_1, y_2|x) = p(y_1|x)p(y_2|x)$.*

B.2 Capacity Region of a Degraded BC

We now consider sending independent information over a degraded BC at rate R_1 to Y_1 and at rate R_2 to Y_2.

Theorem 1 ([CT91]) *The capacity region for sending independent information over the degraded channel $X \longrightarrow Y_1 \longrightarrow Y_2$ is the closure of the convex hull of the set of all rate pairs (R_1, R_2) satisfying*

$$R_2 \leq I(U; Y_2), \tag{B.4a}$$

$$R_1 \leq I(X; Y_1|U), \tag{B.4b}$$

for some joint distribution $p(u)p(x|u)p(y, z|x)$, where the auxiliary random variable U has cardinality bounded by $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\}$.

Proof 3 *(Achievability of the capacity region)*
The cardinality bounds for the auxiliary random variable U can be derived using standard methods from convex set theory. A rough argument for the proof is as follows. The auxiliary random variable U serves as a cloud center that can be distinguished by both receivers Y_1 and Y_2. Each cloud consists in 2^{nR_1} codewords $\mathbf{X}^n$ distinguishable by the receiver Y_1. The worst receiver can only see the clouds, while the better receiver can see the individual codewords within the clouds.
The formal proof of the achievability of this region uses a random coding argument. Fix $p(u)$ and $p(x|u)$.
Random codebook generation: *Generate 2^{nR_2} independent codewords of length n, $\mathbf{U}(w_2)$, $w_2 \in$*

$\{1, 2, \ldots, 2^{nR_2}\}$ *according to* $\Pi_{i=1}^{n} p(u_i)$. *For each codeword* $\mathbf{U}(w_2)$, *generate* 2^{nR_1} *independent codewords* $\mathbf{X}(w_1, w_2)$ *according to* $\Pi_{i=1}^{n} p(x_i|u_i(w_2))$[1]. *The codeword* $\mathbf{u}(i)$ *plays the role of the cloud center understandable to both* Y_1 *and* Y_2, *while* $\mathbf{x}(i, j)$ *is the* j*-th satellite codeword in the* i*-th cloud.*

Encoding: *To send the pair* (W_1, W_2), *transmit the corresponding codeword* $\mathbf{X}(W_1, W_2)$.

Decoding: *Receiver* 2 *determines the unique* $\hat{\hat{W}}_2$ *such that* $(\mathbf{U}(\hat{\hat{W}}_2), \mathbf{Y}_2)$ *are jointly typical. If there are none such or more than one such, an error is declared. Receiver* 1 *looks for the unique* $(\hat{W}_1, \hat{W}_2)$ *such that* $(\mathbf{U}(\hat{W}_2), \mathbf{X}(\hat{W}_1, \hat{W}_2), \mathbf{Y}_1)$ *are jointly typical. If there are none such or more than one such, an error is declared.*

Analysis of the probability of error: *Assume that the message* $(W_1, W_2) = (1, 1)$ *was sent. Let* $P(\cdot)$ *denote the conditional probability at an event given that* $(1, 1)$ *was sent and* $\mathcal{T}^n(\epsilon)$ *the set of jointly typical sequences. The channel from* $\mathbf{U}$ *to* $\mathbf{Y}_2$ *is basically a single user channel. Hence, we will be able to decode the* $\mathbf{U}$ *codewords with low probability of error if* $R_2 < I(U; Y_2)$. *Define the error events*

$$E^{(1)}{}_{Yi} \triangleq \{(\mathbf{U}(i), \mathbf{Y}_1) \in \mathcal{T}^n(\epsilon)\}, \tag{B.5a}$$

$$E^{(1)}{}_{Yij} \triangleq \{(\mathbf{U}(i), \mathbf{X}(i, j), \mathbf{Y}_1) \in \mathcal{T}^n(\epsilon)\}, \tag{B.5b}$$

$$E^{(2)}{}_{Yi} \triangleq \{(\mathbf{U}(i), \mathbf{Y}_2) \in \mathcal{T}^n(\epsilon)\}. \tag{B.5c}$$

Then the probability of error at receiver 2 *is*

$$\begin{aligned} P_e^n(2) &= P\left(\bar{E}_{Y1}^{(2)} \cup \cup_{i\neq 1} E^{(2)}{}_{Yi}\right), \\ &\leq P(\bar{E}_{Y1}^{(2)}) + \sum_{i\neq 1} P(E^{(2)}{}_{Yi}), \\ &\leq \epsilon + 2^{nR_2} 2^{-n(I(U;Y_2)-2\epsilon)}, \\ &\leq 2\epsilon. \end{aligned} \tag{B.6}$$

(B.6) follows if n *is large enough and* $R_2 < I(U; Y_2)$. *Similarly, for decoding for receiver* 1, *we have*

$$\begin{aligned} P_e^n(1) &= P\left(\bar{E}_{Y1}^{(1)} \cup \cup_{i\neq 1} E^{(1)}{}_{Yi} \cup \cup_{j\neq 1} E^{(1)}{}_{Y1j}\right), \\ &\leq P(\bar{E}_{Y1}^{(1)}) + \sum_{i\neq 1} P(E^{(1)}{}_{Yi}) + \sum_{j\neq 1} P(E^{(1)}{}_{Y1j}). \end{aligned} \tag{B.7}$$

The term $P(\bar{E}_{Y1}^{(1)})$ *is upper bounded by* ϵ. *The second term can be bounded as* $\sum_{i\neq 1} P(E^{(1)}{}_{Yi}) \leq 2^{nR_2} 2^{-n(I(U;Y_1)-2\epsilon)}$ *and goes to* 0 *because* $R_2 < I(U; Y_2) \leq I(U; Y_1)$. *The second inequality follows since the channel is degraded. The third term can be bounded as follows.*

$$\begin{aligned} P(E^{(1)}{}_{Y1j}) &= P\left((\mathbf{U}(1), \mathbf{X}(1, j), \mathbf{Y}_1) \in \mathcal{T}^n(\epsilon)\right), \\ &= \sum_{(\mathbf{u},\mathbf{x},\mathbf{y}_1)\in\mathcal{T}^n(\epsilon)} P\left(\mathbf{U}(1), \mathbf{X}(1, j), \mathbf{Y}_1\right), \\ &= \sum_{(\mathbf{u},\mathbf{x},\mathbf{y}_1)\in\mathcal{T}^n(\epsilon)} P(\mathbf{U}(1)) P(\mathbf{X}(1, j)|\mathbf{U}(1)) P(\mathbf{Y}_1|\mathbf{U}(1)), \\ &\leq \sum_{(\mathbf{u},\mathbf{x},\mathbf{y}_1)\in\mathcal{T}^n(\epsilon)} 2^{-n(H(U)-\epsilon)} 2^{-n(H(X|U)-\epsilon)} 2^{-n(H(Y_1|U)-\epsilon)}, \\ &\leq 2^{-n(H(U,X,Y_1)+\epsilon)} 2^{-n(H(U)-\epsilon)} 2^{-n(H(X|U)-\epsilon)} 2^{-n(H(Y_1|U)-\epsilon)}, \\ &= 2^{-n(I(X,Y_1|U)-4\epsilon)}. \end{aligned} \tag{B.8}$$

[1]This is the superposition coding referred to in Chapter 4.

Hence, we have $\sum_{j\neq 1} P(E^{(1)}{}_{Y1j}) \leq 2^{nR_1}2^{-n(I(X,Y_1|U)-4\epsilon)}$ *which goes to* 0 *if* $R_1 < I(X;Y_1|U)$*. Finally, the probability of error* $P_e^n(1)$ *is bounded as*

$$P_e^n(1) \leq \epsilon + 2^{nR_2}2^{-n(I(U,Y_1)-3\epsilon)} + 2^{nR_1}2^{-n(I(X,Y_1|U)-4\epsilon)}, \tag{B.9}$$

and goes to 0 *if* n *is large enough and* $R_2 < I(U,Y_1)$ *and* $R_1 < I(X,Y_1|U)$*. Hence, there exists a sequence of good* $((2^{nR_1}, 2^{nR_2}), n)$ *codes with probability of error going to zero as* n *becomes large enough.*

B.3 The Gaussian BC

We begin this section by noticing that a Gaussian BC is degraded [CT91]. This is because a Gaussian BC where $\mathbf{Y}_1 = \mathbf{X} + \mathbf{Z}_1$ and $\mathbf{Y}_2 = \mathbf{X} + \mathbf{Z}_2$, with $\mathbf{Z}_i \sim \mathcal{N}(0, N_i)$, $i = 1, 2$, is equivalent to the channel

$$\mathbf{Y}_1 = \mathbf{X} + \mathbf{Z}_1, \tag{B.10a}$$

$$\mathbf{Y}_2 = \mathbf{X} + \mathbf{Z}_2 = \mathbf{Y}_1 + \mathbf{Z}_2', \tag{B.10b}$$

where $\mathbf{Z}_2' \sim \mathcal{N}(0, N_2 - N_1)$.

Theorem 2 ([CT91]) *The capacity region of the Gaussian broadcast channel defined by* $\mathbf{Y}_1 = \mathbf{X} + \mathbf{Z}_1$ *and* $\mathbf{Y}_2 = \mathbf{X} + \mathbf{Z}_2$ *is given by the convex hull of all rate pairs* (R_1, R_2) *satisfying*

$$\begin{array}{l} R_1 < \frac{1}{2}\log(1 + \frac{\gamma P}{N_1}) \\ R_2 < \frac{1}{2}\log(1 + \frac{(1-\gamma)P}{\gamma P + N_2}) \end{array} \tag{B.11}$$

where γ *may be arbitrarily chosen* $(0 \leq \gamma \leq 1)$*.*

Proof 4 *(Achievability of capacity region of the Gaussian BC)*
Encoder: *To encode the messages, the encoder generates two codebooks, one with power* γP *at rate* R_1 *and the other with power* $(1-\gamma)P$ *at rate* R_2*. Then to send the pair* $(i,j) \in \{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\}$*, the transmitter transmits the sum* $\mathbf{X}_1(i) + \mathbf{X}_2(j)$*, where* $\mathbf{X}_1(i)$ *belongs to the first codebook* $\mathbf{X}_1$ *and* $\mathbf{X}_2(j)$ *belongs to the second codebook* $\mathbf{X}_2$*.*
Decoder: *The bad receiver (receiver* 2*) looks through the second codebook to find the closest codeword to the received sequence* $\mathbf{Y}_2$ *(i.e., the one that is jointly typical with* $\mathbf{Y}_2$*). He sees the effective* $SNR_2 = (1-\gamma)P/(\gamma P + N_2)$ *since the codeword directed receiver* 1 *acts as noise. The good receiver first decodes the message directed ti receiver* 2 *(he can accomplish this because his is less noisy*[2]*). He then subtracts the codeword* $\hat{\mathbf{X}}_2$ *from the received sequence* $\mathbf{Y}_1$ *and looks through the first codebook to the closest codeword to* $\mathbf{Y} - \hat{\mathbf{X}}_2$*. Receiver* 1 *sees the an SNR of* $SNR_1 = \gamma P/N_1$*).*
Analysis of the probability of error: *The channel can be divided into two fictitious channels: the one from* $\mathbf{X}_2$ *to* $\mathbf{Y}_2$ *and the one from* $\mathbf{X}_1$ *to* $\mathbf{Y} - \hat{\mathbf{X}}_2$*. Hence, the resulting probability of error can be made as low as desired.*

[2]This is a nice dividend for degraded BC in that the better receiver always knows the message intended for the worse receiver.

B.4 The GBC With State Information at the Transmitter

The (physically Degraded) GBC with state information at the encoder is shown in Fig.B.1(b). Here we have

$$\mathbf{Y}_1 = \mathbf{X} + \mathbf{S} + \mathbf{Z}_1, \tag{B.12a}$$

$$\mathbf{Y}_2 = \mathbf{X} + \mathbf{S} + \mathbf{Z}_2. \tag{B.12b}$$

When the state $\mathbf{S}$ is available everywhere -at the transmitter and at both receivers-, the receivers can simply subtract $\mathbf{S}$ to reduce the channel to the case without additive state and attain the same region as given by (B.11). When only the transmitter knows the state $\mathbf{S}$, the capacity region can be obtained as in the following theorem.

Theorem 3 ([KSS04]) *The capacity region of the Gaussian BC (B.12) with state information non-causally available at the transmitter is given by the standard capacity region (B.11).*

Proceeding similarly to Costa's approach, we need only proof the achievability of the region. We use the following result on the discrete memoryless physically DBC with state information non-causally available at the transmitter.

Lemma 5 *The capacity region of a discrete memoryless physically DBC $p(y_1, y_2|x, s) = p(y_1|x, s)p(y_2|y_1)$ with state information non-causally available at the transmitter contains the convex hull of all rate pairs (R_1, R_2) satisfying*

$$R_1 \leq I(U_1; Y_1|U_2) - I(U_1; S|U_2) \tag{B.13a}$$

$$R_2 \leq I(U_2; Y_2) - I(U_2; S) \tag{B.13b}$$

for some joint distribution $p(s)p(u_1, u_2, x|s)p(y_1|x, s)p(y_2|y_1)$ where U_1 and U_2 are auxiliary random variables with finite cardinality.

Proof 5 *(Achievability of the region (B.11)) First note that that the optimality of (B.13) is yet to be proved in general. But, "for the Gaussian BC with state information at the encoder, the region (B.13) turns out to be the capacity region" [KSS04]. This can be seen through evaluating the region (B.13) using the choice of the joint distribution $p(u_1, u_2, x, s)$ given by*

$$\mathbf{U}_1 \sim \mathcal{N}(\alpha_1\mathbf{S}, \gamma P) \tag{B.14a}$$

$$\mathbf{U}_2 \sim \mathcal{N}(\alpha_2\mathbf{S}, (1-\gamma)P) \tag{B.14b}$$

$$\mathbf{X} = \mathbf{U}_1 + \mathbf{U}_2 - (\alpha_1 + \alpha_2)\mathbf{S}, \tag{B.14c}$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are conditionally independent given $\mathbf{S}$ and

$$\alpha_2 = \frac{(1-\gamma)P}{P + N_2}, \tag{B.15a}$$

$$\alpha_1 = (1-\alpha_2)\frac{\gamma P}{\gamma P + N_1}. \tag{B.15b}$$

This evaluation results in the region (B.11). Since the capacity region of the GBC with state information available at the transmitter (B.12) can not exceed (B.11), it turns out that (B.11) is indeed the required capacity region.

Appendix C

Some Results on Multiple Access Channels

This appendix provides a brief review of some results on the discrete memoryless Multiple Access Channel (MAC), the Gaussian MAC, the GMAC with state information available at the transmitters. The key ideas for establishing the capacity regions are outlined. The complete proof of the capacity region of the Gaussian MAC can be found in [CT91]. The capacity region of the Gaussian MAC with state information known to the transmitters has been established in [KSS04].

C.1 Multiple Access Channel (MAC)

A MAC channel consists in several transmitters sending information to one distant receiver. An example is shown in Fig.C.1(c) where three ground stations wish to communicate with a common satellite. We assume that two transmitters have average power P_1 and P_2 respectively and wish to send independent messages W_1 (at rate R_1) and W_2 (at rates R_2) to a distant receiver Y. This receiver sees the two transmitted sequences $\mathbf{X}_1$ and $\mathbf{X}_2$ added together with the noise $\mathbf{Z}$ (of power N). The MAC can be formally defined as follows.

Definition 8 *A discrete memoryless Multiple Access Channel consists of three alphabets $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathbf{Y}$, and a probability transition matrix $p(y|x_1, x_2)$.*

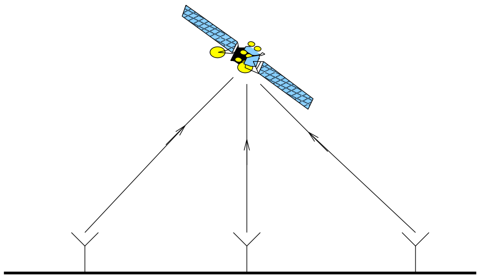
A $((2^{nR_1}, 2^{nR_2}), n)$ code for the MAC with independent information consists of two sets of integers $\mathcal{W}_1 = \{1, 2, \ldots, 2^{nR_1}\}$ and $\mathcal{W}_2 = \{1, 2, \ldots, 2^{nR_2}\}$ called the message sets, two encoding functions

$$\mathbf{X}_1 : \mathcal{W}_1 \longrightarrow {\mathcal{X}_1}^n, \tag{C.1a}$$

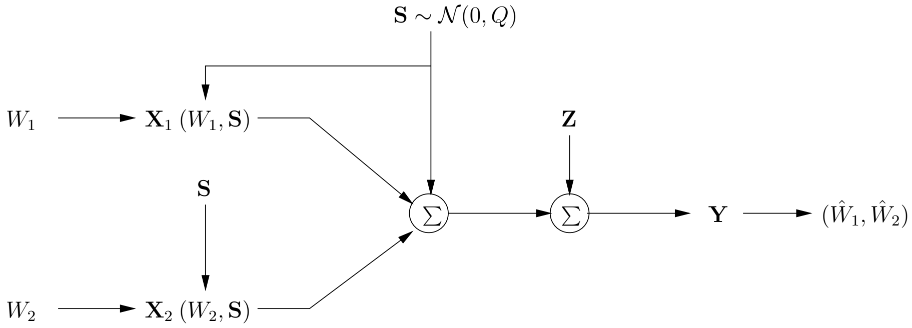
$$\mathbf{X}_2 : \mathcal{W}_2 \longrightarrow {\mathcal{X}_2}^n, \tag{C.1b}$$

and a decoding function

$$g : \mathcal{Y}^n \longrightarrow \mathcal{W}_1 \times \mathcal{W}_2. \tag{C.2}$$



(c) A Multiple Access Channel (MAC).



(d) A Gaussian MAC with State Information at the Transmitters.

Assuming the messages are independent and equally likely, the average probability of error is defined as

$$P_e^{(n)} = \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2)\in\mathcal{W}_1\times\mathcal{W}_2} \Pr\{g(\mathbf{y}) \neq (w_1, w_2) \,|\, (w_1, w_2) \text{ sent } \}. \tag{C.3}$$

Definition 9 *A rate pair (R_1, R_2) is said to be achievable for the MAC if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \longrightarrow 0$.*

Definition 10 *The capacity region of the MAC is the closure of the set of achievable rates.*

C.2 Capacity Region for the MAC

We state the capacity region of the MAC in the form of a theorem.

Theorem 1 ([CT91]) *The capacity region of a MAC $(\mathcal{X}_1\times\mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})$ is the closure of the convex hull of the set of all rate pairs (R_1, R_2) satisfying*

$$R_1 \leq I(X_1; Y|X_2), \tag{C.4a}$$

$$R_2 \leq I(X_2; Y|X_1), \tag{C.4b}$$

$$R_1 + R_2 \leq I(X_1, X_2; Y) \tag{C.4c}$$

for some product distribution $p_1(x_1)p_2(x_2)$ on $\mathcal{X}_1\times\mathcal{X}_2$.

We only give the proof of the achievability part of Theorem 1. The converse can be found in [CT91].

Proof 6 *(Achievability of the capacity region (C.4))*
Fix $p(x_1, x_2) = p_1(x_1)p_2(x_2)$.
Random codebook generation: *Generate* 2^{nR_1} *independent codewords of length* n, $\mathbf{X}_1(i), i \in \{1, 2, \ldots, 2^{nR_1}\}$ *according to* $\Pi_{i=1}^n p_1(x_{1i})$. *Similarly, generate* 2^{nR_2} *independent codewords of length* n, $\mathbf{X}_2(j)$, $j \in \{1, 2, \ldots, 2^{nR_2}\}$ *according to* $\Pi_{i=1}^n p_1(x_{1i})$. *The set of all these codewords form the codebook which is revealed to the senders and the receiver.*
Encoding: *To send index* i*, transmitter* 1 *sends the codeword* $\mathbf{X}_1(i)$*. Similarly, to send* j*, transmitter* 2 *sends the codeword* $\mathbf{X}_2(j)$.
Decoding: *The Receiver determines the pair* (i, j) *such that* $(\mathbf{x}_1(i), \mathbf{x}_2(j), \mathbf{y})$ *is jointly typical. If there are none such or more than one such, an error is declared.*
Analysis of the probability of error: *Assume that the message* $(W_1, W_2) = (1, 1)$ *was sent. Let* $P(\cdot)$ *denote the conditional probability at an event given that* $(1, 1)$ *was sent and* $\mathcal{T}^n(\epsilon)$ *the set of jointly typical sequences. Define the error event*

$$E_{ij} \triangleq \{(\mathbf{X}_1(i), \mathbf{X}_2(j), \mathbf{Y}) \in \mathcal{T}^n(\epsilon)\}. \tag{C.5}$$

Then

$$\begin{aligned} P_e^n &= P\left(\bar{E}_{11} \cup \cup_{(i,j)\neq(1,1)} E_{ij}\right), \\ &\leq P(\bar{E}_{11}) + \sum_{i\neq 1, j=1} P(E_{i1}) + \sum_{i=1, j\neq 1} P(E_{1j}) + \sum_{i\neq 1, j\neq 1} P(E_{ij}). \end{aligned} \tag{C.6}$$

Or

$$\begin{aligned} P(E_{i1}) &= P\{(\mathbf{X}_1(i), \mathbf{X}_2(1), \mathbf{Y}) \in \mathcal{T}^n(\epsilon)\}, \\ &= \sum_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in \mathcal{T}^n(\epsilon)} p(\mathbf{x}_1)p(\mathbf{x}_2, \mathbf{y}), \\ &\leq |\mathcal{T}^n(\epsilon)| 2^{-n(H(X_1)-\epsilon)} 2^{-n(H(X_2,Y)-\epsilon)}, \\ &\leq 2^{-n(H(X_1)+H(X_2,Y)-H(X_1,X_2,Y)-3\epsilon)}, \\ &= 2^{-n(I(X_1;X_2,Y)-3\epsilon)}, \\ &= 2^{-n(I(X_1;Y|X_2)-3\epsilon)}. \end{aligned} \tag{C.7}$$

(C.7) follows since $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and hence $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y|X_2) = I(X_1; Y|X_2)$. Similarly, we have

$$P(E_{1j}) \leq 2^{-n(I(X_2;Y|X_1)-3\epsilon)} \text{ for } j \neq 1, \tag{C.8a}$$

$$P(E_{ij}) \leq 2^{-n(I(X_1,X_2;Y)-4\epsilon)} \text{ for } i \neq 1, j \neq 1. \tag{C.8b}$$

It follows that

$$P_e^n \leq \epsilon + 2^{nR_1} 2^{-n(I(X_1;Y|X_2)-3\epsilon)} + 2^{nR_2} 2^{-n(I(X_2;Y|X_1)-3\epsilon)} + 2^{n(R_1+R_2)} 2^{-n(I(X_1,X_2;Y)-4\epsilon)}, \tag{C.9}$$

which tends to 0 under the conditions of the theorem.

C.3 The Gaussian MAC

Consider two senders sending to a single receiver over an i.i.d Gaussian channel $\mathbf{Z} \sim \mathcal{N}(0, N)$. The receiver sees the sequence $\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{Z}$. Assume that there is a power constraint P_j on sender j, $j = 1, 2$.

Theorem 2 ([CT91]) *The capacity region of the Gaussian MAC defined by $\mathbf{Y}_1 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{Z}$ where the channel noise $\mathbf{Z} \sim \mathcal{N}(0, N)$ is i.i.d. Gaussian, is given by the closure of the convex hull of the set of all rate pairs (R_1, R_2) satisfying*

$$\begin{aligned} R_1 &\leq \tfrac{1}{2}\log(1 + \tfrac{P_1}{N}), \\ R_2 &\leq \tfrac{1}{2}\log(1 + \tfrac{P_2}{N}), \\ R_1 + R_2 &\leq \tfrac{1}{2}\log(1 + \tfrac{P_1+P_2}{N}). \end{aligned} \tag{C.10}$$

Proof 7 *The proof of the capacity region of the discrete memoryless MAC can be extended to the the Gaussian MAC. The converse also can be extended similarly. So, the capacity region of the GMAC is also given by (C.4), with, this time, an additional constraint on channel inputs in the form $\mathbb{E}[\mathbf{X}_1^2] \leq P_1$ and $\mathbb{E}[\mathbf{X}_2^2] \leq P_2$. Next, expanding the mutual information in terms of relative entropy we get: $I(X_i; Y|X_j) \leq \frac{1}{2}\log_2(1 + P_i/N)$, $i, j = 1, 2$ and $i \neq j$.*

C.4 GMAC with State Information at the Transmitters

The GMAC with state information $\mathbf{S}$ non-causally known to the transmitters is shown in Fig.C.1(d). Here the channel output is given by $\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{S} + \mathbf{Z}$. The channel state $\mathbf{S}$ is distributed according to $\mathcal{N}(0, Q)$ and is independent of the channel noise $\mathbf{Z}$. When the state $\mathbf{S}$ is available everywhere -at both transmitters and at the receiver-, the receiver can simply subtract out $\mathbf{S}$ to reduce the channel to the case without additive state and attain the same region as given by (C.10). When only the transmitters know the state $\mathbf{S}$, the capacity region can be obtained as in the following theorem.

Theorem 3 ([KSS04]) *The capacity region of the Gaussian MAC with state information non-causally available at the transmitters is given by the standard capacity region (C.10).*

Proceeding similarly to Costa's approach, we need only proof the achievability of the region. We use the following result on the discrete memoryless MAC with state information non-causally available at the transmitters.

Lemma 6 *The capacity region of a discrete memoryless MAC $p(y|x_1, x_2, s)$ with state information non-causally available at the transmitters contains the convex hull of all rate pairs (R_1, R_2) satisfying*

$$R_1 \leq I(U_1; Y|U_2) - I(U_1; S|U_2), \tag{C.11a}$$

$$R_2 \leq I(U_2; Y|U_1) - I(U_2; S|U_1), \tag{C.11b}$$

$$R_1 + R_2 \leq I(U_1, U_2; Y) - I(U_1, U_2; S) \tag{C.11c}$$

for some joint distribution $p(s)p(u_1, x_1|s)p(u_2, x_2|s)p(y|x_1, x_2, s)$ where U_1 and U_2 are auxiliary random variables with finite cardinality.

Proof 8 *(Achievability of the region (C.10))*
First note that that the optimality of (C.11) is yet to be proved in general. But, "for the Gaussian MAC with state information at the transmitters, the region (C.11) turns out to be the capacity region" [KSS04]. This can be seen through evaluating the region (C.11) using the choice of the joint distribution $p(u_1, u_2, x, s)$ *given by*

$$\mathbf{U}_1 \sim \mathcal{N}(\alpha_1 \mathbf{S}, P_1), \tag{C.12a}$$

$$\mathbf{U}_2 \sim \mathcal{N}(\alpha_2 \mathbf{S}, (1-\gamma)P) \tag{C.12b}$$

$$\mathbf{X}_1 = \mathbf{U}_1 - \alpha_1 \mathbf{S}, \tag{C.12c}$$

$$\mathbf{X}_2 = \mathbf{U}_2 - \alpha_2 \mathbf{S}, \tag{C.12d}$$

$$\mathbf{X} = \mathbf{U}_1 + \mathbf{U}_2 - (\alpha_1 + \alpha_2)\mathbf{S}, \tag{C.12e}$$

where $\mathbf{U}_1$ *and* $\mathbf{U}_2$ *are conditionally independent given* $\mathbf{S}$ *and*

$$\alpha_2 = \frac{P_2}{P_1 + P_2 + N}, \tag{C.13a}$$

$$\alpha_1 = \frac{P_1}{P_1 + P_2 + N}. \tag{C.13b}$$

This evaluation results in the region (C.10). Since the capacity region of the GMAC with state information available at the transmitters can not exceed (C.10), it turns out that (C.10) is indeed the required capacity region.

Bibliography

General Bibliography

[AEV02] E. Agrell, T. Eriksson, and A. Vardy, *Closest point search in lattices*, IEEE Trans. on IT **IT-48** (2002), 2201–2214.

[AG02] A. Aaron and B. Girod, *Compression with side information using using turbo codes*, proc. of DCC'02 (Snowbird, UT), April 2002.

[Bag93] C. M. M. J. Baggen, *An information theoretic approach to timing jitter*, Ph.D Thesis, San Diego, USA, 1993.

[BBCS05] A. Bennatan, D. Burshtein, G. Caire, and S. Shamai, *Superposition coding for side-information channels*, IEEE Transactions on Information Theory (submitted) (2005).

[BEH02] R. Bäuml, J. J. Eggers, and J. Huber, *A channel model for watermarks subject to desynchronization attacks*, International ITG Conference on Source and Channel Coding (Berlin), January 2002, pp. 28–30.

[BJDK01] C. Berrou, M. Jézéquel, C. Douillard, and S. Kerouédan, *The advantages of non-binary turbo codes*, Proc. IEEE Iformation Theory Workshop ITW (Caims, Australia), September 2-7 2001.

[BM01] J. Bajcsy and P. Mitran, *Coding for the slepian-wolf problem with turbo codes*, Proc. GlobeCom (San Antonio, TX), November 2001.

[BPN01] P. Bassia, I. Pitas, and N. Nikolaidis, *Robust audio watermarking in the time domain*, IEEE Transactions on Multimedia **3** (2001), 232–241.

[BTH96] L. Boney, A. H. Tewfik, and K. N. Hamdy, *Digital watermarks for audio signals*, Proc. EUSIPCO, vol. III, September 1996, pp. 1697–1700.

[CC02] T. M. Cover and M. Chiang, *Duality between channel capacity and rate distortion with two-sided state information*, IEEE Trans. on IT **IT-48** (2002), 1629–1638.

[CDW01] B. Chen, Stark C. Draper, and G. Wornell, *Information embedding and related problems: Recent results and applications*, Allerton Conference (USA), 2001.

[CKLS97] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, *Secure spread spectrum watermarking for multimedia*, IEEE Transactions on image Processing **51** (1997), 1673–1678.

[CL01] A. S. Cohen and A. Lapidoth, *The capacity of the vector gaussian watermarking game*, IEEE Proc ISIT (Washington, DC), 2001, p. 5.

[CL02a] ______, *The gaussian watermarking game*, IEEE Transactions on Information Theory **48** (2002), 1639–1667.

[CL02b] ______, *Generalized writing on dirty paper*, Proc. ISIT 2002 (Lausanne-Switzerland), July 2002.

[CMM99] I. Cox, M. Miller, and A. McKellips, *Watermarking as communication with side information*, Proc. Int. Conference on Multimedia Computing and Systems, July 1999, pp. 1127–1141.

[CMM01] ______, *Electronic watermarking: the first 50 years*, Proc. Int. Workshop on Multimedia Signal Processing, 2001, pp. 225–230.

[Cos83] MAX H. M. Costa, *Writing on dirty papers*, IEEE Trans. on IT **IT-29** (1983), 439–441.

[Cov72] T. M. Cover, *Broadcast channels*, IEEE Transactions on Information Theory **IT-18** (1972), 2–14.

[Cov88] ______, *Comments on broadcast channels*, IEEE Transactions on Information Theory **IT-44** (1988), 2524–2530.

[CPGW05] P. Comesaña, F. Pérez-González, and F. M. J. Willems, *Applying erez and ten brinks dirty paper codes to data-hiding.*, Proc. Security, Steganography, and Watermarking of Multimedia Contents VI, SPIE (CA, USA), January 2005.

[CPR99] J. Chou, S. S. Pradhan, and K. Ramchandran, *On the duality between source coding and data hiding*, IEEE 33rd Asilomar Conference, vol. Pacific Grove, CA, November 1999.

[CPR03] ______, *Turbo and trellis-based constructions for source coding with side information*, proc. of DCC'03, March 2003.

[CRS82] J. H. Conway, E. M. Rains, and N. J. A. Sloane, *Voronoi regions of lattices, second moments of polytops, and quantization*, IEEE Transactions on Information Theory **IT-28** (1982), 211–226.

[CRS99] ______, *On the existence of similar sublattices*, Canadian Journal of Mathematics **IT-51** (1999), 1300–1306.

[CS88] J. H. Conway and N. J. A. Sloane, *Sphere packing, lattices and groups*, third edition, John Willey & Sons INC., New York, 1988.

[CS98] B. Chen and C. E. W. Sundberg, *Broadcasting data in the fm band by means of adaptive contiguous band insertion and precanceling techniques*, Proc. IEEE International Conference on Communications (BC-canada), vol. 2, June 1998, pp. 823–827.

[CS03] G. Caire and S. Shamai (Shitz), *On the throughput of a multi-antenna gaussian broadcast channel*, IEEE Transactions on Information Theory **IT-49** (2003), 1691–1706.

[CT91] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Willey & Sons INC., New York, 1991.

[CW98] B. Chen and G. Wornell, *Digital watermarking and information embedding using dither modulation*, Proc. Workshop on Multimedia Signal Processing (1998), 273–278.

[CW99] ______, *Achievable performance of digital watermarking systems*, Proc. Int. Conference on Multimedia Computing and Systems (Florence, Italy), vol. 87, June 1999, pp. 13–18.

[CW00] ______, *Preprocessed and postprocessed quantization index modulation methods for digital watermarking*, Proc. SPIE Security and Watermarking of Multimedia Contents II (san Joze, CA, USA), vol. 3971, January 2000, pp. 48–59.

[CW01] ______, *Quantization index modulation: a class of provably good methods for digital watermarking and information embedding*, IEEE Transactions on Information Theory **47** (2001), 1423–1443.

[DW04] S. C. Draper and G. Wornell, *Side information aware coding strategies for sensor networks*, IEEE Journal on Selected Areas in Comm. **22** (2004), 966–976.

[EBG02a] J. J. Eggers, R. Bäuml, and B. Girod, *Digital watermarking facing attacks by amplitude scaling and additive white noise*, International ITG Conference on Source and Channel Coding (Berlin), January 2002, pp. 28–30.

[EBG02b] ______, *Estimation of amplitude modifications before scs watermark detection*, Proc. of SPIE (San Jose CA), vol. 4675, 2002, pp. 387–398.

[EBTG03] J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, *Scalar costa scheme for information embedding*, IEEE Transactions on Signal Processing (2003), 1003–10019.

[EF92] M. Vedat Eyboğlu and G. D. Forney, *Trellis precoding: Combining coding, precoding and shaping for intersymbol interference channels*, IEEE Trans. on IT **IT-38** (1992), 301–314.

[EF93] ______, *Lattice and trellis quantization with lattice- and trellis-bounded codebooks- high-rate theory for memoryless sources*, IEEE Trans. on IT **IT-39** (1993), 46–59.

[EG02] J. J. Eggers and B. Girod, *Informed watermarking*, Kluwer, Boston, MA, 2002.

[Egg01] J. J. Eggers, *Information embedding and digital watermarking as communication with side information*, PhD dissertation, University of Erlangen-Nurnberg, Erlang Germany, November 2001.

[ESG00] J. J. Eggers, J. K. Su, and B. Girod, *A blind watermarking scheme based on structured codebooks*, International IEE Colloquim Secure Image and Image Authentication (London, UK), April 2000, pp. 1–6.

[ESG01] ______, *Performance of a practical watermarking scheme*, Proceedings SPIE Security and Watermarking of Multimedia Contents III (San Jose, CA, USA), vol. 4314, January 2001, pp. 28–30.

[ESZ00] U. Erez, S. Shamai, and R. Zamir, *Capacity and lattice strategies for cancelling known interference*, Int. Symps. on IT and Its Applications, ISITA (Honolulu, Hawaii), 2000, pp. 681–684.

[EtB04] U. Erez and S. ten Brink, *A close-to-capacity dirty paper coding scheme*, IEEE Transactions on Information Theory (submitted) (2004).

[EZ04] U. Erez and R. Zamir, *Achieving* $\frac{1}{2}\log_2(1 + SNR)$ *on the AWGN channel with lattice encoding and decoding*, IEEE Transactions on Information Theory **IT-50** (2004), 1–23.

[Fis04] R. F. H. Fischer, *The modulo-lattice channel: The key feature in precoding schemes*, accepted for publication in IEEE Trans. on Electronic Communication (2004).

[FL98] M. Fossorier and S. Lin, *Soft-input soft-output decoding of linear block codes based on ordered statistics*, Proc. IEEE GLOBOCOM, vol. 5, 1998, pp. 2828–2833.

[For88] G. D. Forney, *Coset-codes: Part* I: Introduction and geometrical classification, IEEE Trans. on IT **IT-34** (1988), 1123–1151.

[FTB04] R. F. H. Fischer, R. Tzschoppe, and R. Bäuml, *Lattice costa schemes using subspace projection for digital watermarking*, Proc. ITG Conference on Source and Channel Coding, 2004.

[FTC00] G. D. Forney, M. D. Trott, and S. Y. Chung, *Sphere-bound-achieving cosets codes and multilevel coset codes*, IEEE Trans. on IT **IT-46** (2000), 820–850.

[FW89] G. D. Forney and L. F. Wei, *Multidimensional constellations- part* I: Introductions figures of merit, and generalized crosss constellations, IEEE J. Select. Areas Commun. **7** (1989), 877–892.

[Gal68] R. G. Gallager, *Information theory and reliable communication*, John Willey, New York, 1968.

[GC80] A. A. El Gamal and T. M. Cover, *Multiple user information theory*, IEEE Transactions on Information Theory **IT-68** (1980), 1466–1483.

[GDF89] Jr. G. D. Forney, *Multidimensional constellations- part* II: Voronoi constellations, IEEE J. Select. Areas Commun. **7** (1989), 941–958.

[GP80] S. I. Gel'fand and M. S. Pinsker, *Coding for channel with random parameters*, Problems of Control and IT. **9** (1980), 19–31.

[GV97] A. Goldsmith and Varaiya, *Capacity of fading channels with channel side information*, IEEE Transactions on Information Theory **IT-43** (1997), 1986–1992.

[HG83] C. D. Heegard and A. A. El Gamal, *On the capacity of computer memory with defects*, IEEE Transactions on Information Theory **IT-29** (1983), 731–739.

[HK99] F. Hartung and M. Kutter, *Multimedia watermarking techniques*, Proc. IEEE, vol. 87, 1999, pp. 1079–1107.

[HM72] H. Harashima and H. Miyakawa, *Matched-transmission technique for channels with intersymbol interference*, IEEE Trans. on Communication **COM-20** (1972), 774–780.

[HSG99] F. Hartung, J. K. Su, and B. Girod, *Spread spectrum watermarking: malicious attacks and counterattacks*, Proc. SPIE (Salt Lake City UT), vol. 3657, January 1999, pp. 147–158.

[JSO02] C. Jongren, M. Skoglund, and B. Ottersten, *Combining beamforming and orthogonal space-time block coding*, IEEE Transactions on Information Theory **IT-48** (2002), 611–627.

[Jus72] J. Justesen, *A class of constructive asymptotically good algebraic codes*, PGIT, vol. 18, 1972, pp. 652–656.

[KM01] D. Kirovski and H. Malvar, *Robust spread-spectrum audio watermarking*, Proc. International Conference Acoustic Speech, Signal Processing (Salt Lake City UT), May 2001.

[KM02a] ———, *Embedding and detecting spread spectrum watermarks under estimation attacks*, IEEE International, vol. 3657, January 2002.

[KM02b] ———, *Spread spectrum watermarking of audio signals*, IEEE Transactions on Signal Processing: Special ISSUE on Data Hiding (2002).

[KP93] F. R. Kschischang and S. Pasupathy, *Optimal nonuniform signaling for gaussian channels*, IEEE Trans. on Inf. Theory (1993), 913–929.

[KSS04] Y. H. Kim, A. Sutivong, and S. Sigurjonsson, *Multiple user writing on dirty paper*, Proc. ISIT 2004 (Chicago-USA), June 2004, p. 534.

[Kut97] M. Kutter, *Watermarking resisting to translation, rotation and scaling*, Proc. SPIE, 1997.

[LLC03] K. N. Lau, Y. Liu, and T. A. Chen, *Optimal partial feedback design for mimo block fading channels with causal noiseless feedback*, Proc. IEEE Int. Symp. on Information Theory (Yokohama), Jun./Jul. 2003, p. 65.

[LMK04] T. Liu, P. Moulin, and R. Koetter, *On error exponents of nested lattice codes for the awgn channel*, Submitted IEEE Transactions on Information Theory (2004).

[LOJH03] V. Licks, F. Ourique, R. Jordan, and G. Heileman, *Performance of dirty-paper codes for additive white gaussian noise*, IEEE Workshop of Statistical Signal Processing, Proceedings of WSSP03. (St. Louis.), 2003.

[LOJPG03] V. Licks, F. Ourique, R. Jordan, and F. Pérez-González, *The effect of the random jitter attack on the bit error rate. performance of spatial domain image watermarking*, Proceedings of the IEEE Int. Conference on Image Processing, ICIP (Barcelona, Espana), vol. 2, 2003, pp. 28–30.

[Lor93] R. Loroia, *Coding for intersymbol interference channels. combined coding and precoding*, IEEE Trans. on IT **IT-42** (1993), 1053–1061.

[LS04a] N. Liu and K. P. Subbalakshmi, *Non-uniform quantizer design for image data hiding*, Proc. of IEEE Int. Conf. on Image Processing, ICIP (Singapore), vol. 4, October 2004, pp. 2179– 2182.

[LS04b] Y. W. Liu and J. O. Smith, *Multiple watermarking: is power sharing better than time sharing?*, Proc. ICME (Taiepi-Taïwan), June 2004.

[LTF93] R. Loroia, S. A. Tretter, and N. Fardavin, *A simple and effective precoding scheme for noise whitening on intersymbol interference channels*, IEEE Trans. on IT **IT-41** (1993), 1460–1463.

[MF02] H. S. Malvar and D. A. F. Florêncio, *An improved spread spectrum technique for robust watermarking*, Proc. IEEE, 2002.

[MF03] ______, *Improved spread spectrum: A new modulation technique for robust watermarking*, IEEE Transactions on Signal Processing **51** (2003), 898–905.

[MGK04] P. Moulin, Anik K. Goteti, and R. Koetter, *Optimal sparse-QIM codes for zero-rate blind watermarking*, Proceedings of ICASSP (Canada), May 2004.

[MH69] H. Miyakawa and H. Harashima, *Information transmission rate in matched transmission systems with peak transmitting power limitation*, Nat. Conf. Rec. Inst. Electron. Inform. Commun. Eng. of Japan, August 1969, p. 1268.

[MI03] P. Moulin and A. Ivanovic, *The zero-rate spread spectrum watermarking game*, IEEE Transactions on Signal Processing **51** (2003), 1098–1117.

[MK04] P. Moulin and R. Koetter, *Data-hiding codes*, IEEE Int. Conference on Image Processing (Singapore), October 2004.

[MML00] P. Moulin, M. Kivanç Mihçak, and G. I. Lin, *An information-theoretic model for image watermarking and data hiding*, Proceedings of the IEEE Int. Conference on Image Processing (Vancouver, Canada), September 2000.

[MO03] P. Moulin and J. A. O'Sullivan, *Information-theoretic analysis of information hiding*, IEEE Transactions on Information Theory **49** (2003).

[myb]

[Ner00] N. Nerhav, *On random coding error exponents of watermarking systems*, IEEE Transactions on Information Theory **46** (2000), 420–430.

[Nyq28] H. Nyquist, *Certain topics in telegraph transmission theory*, IEEE Transactions AIEE **47** (1928), 617–644.

[OP97] J. J. K. ORuanaidh and T. Pun, *Rotation, scale and translation invariant digital image watermarking*, Proc. IEEE International Conference on Image Processing, vol. 1, 1997, pp. 536–539.

[PAK99] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, *Information hiding– a survey*, Proc. IEEE, vol. 87, July 1999, pp. 1062–1078.

[PBbC98] A. Piva, M. Barni, F. bartolini, and V. Capellini, *Threshold selection for correlation-based watermark detetion*, Proc. European Signal Processing Conference EUSIPCO (Island of Rhodes, Greece), September 1998, pp. 337–355.

[PCR03] S. S. Pradhan, J. Chou, and K. Ramchandran, *Duality between source coding and channel coding and its extension to the side information case*, IEEE Trans. on IT **IT-49** (2003), 1181–1203.

[PGBAM04] F. Perez-Gonzalez, M. Barni, A. Abrardo, and C. Mosquera, *Rational dither modulation: A novel data-hiding method robust to value-metric scaling attacks*, Proc. IEEE MMSP, October 2004.

[PGD04] S. Pateux, G. Le Guelvouit, and J. Delhumeau, *Capacity of data-hiding system subject to desynchronization*, Proceedings of SPIE (San Jose, California, USA), January 2004.

[Pha05] D. T. Pham, *Entropy of a variable slightly contaminated with another*, IEEE Signal Processing Letters **12** (2005), 536–539.

[PPV95a] M. S. Pinsker, Y. V. Prelov, and S. Verdú, *Sensitivity of channel capacity*, IEEE Trans. on IT. **41** (1995), 1877–1888.

[PPV95b] ______, *Sensitivity of gaussian channel capacity and rate-distortion function to nongaussian contaminating*, Proc. of IEEE Sympo. on IT., 1995.

[PR00] S. S. Pradhan and K. Ramchandran, *On the duality between distributed source coding and channel coding with side information*, submitted to IEEE Transactions on Information Theory (2000).

[Pra91] W. K. Pratt, *Digital image processing*, John Willey & Sons INC., New York, 1991.

[Pro01] J. G. Proakis, *Digital communications*, McGraw-Hill, 4th edition, 2001.

[PS98] H. C. Papadopoulos and C. E. W. Sundberg, *Simultaneous broadcasting of analog fm and digital audio signals by means of adaptive precanceling techniques*, IEEE Transactions on Communications **46** (1998), 1233–1242.

[PSM82] R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, *Theory of spread spectrum communications - a tutorial*, IEEE Transactions on Communications **30** (1982), 855–884.

[Ram98] M. Ramkumar, *Data hiding in multimedia: Theory and applications*, Ph.D thesis, Kearny, NJ, USA, November 1998.

[RMZG03] D. Rebollo-Monedero, R. Zhang, and B. Girod, *Design of optimal quantizers for distributed source coding*, proc. of DCC'03 (Snowbird, UT), 2003.

[RSS05] A. Rosenzweig, Y. Steinberg, and S. Shami, *On channels with partial channel state at the transmitter*, IEEE Transactions on Information Theory **IT-51** (2005), 1817–1830.

[SA80] K. Shimizu and E. Aiyoshi, *Necessary conditions for min-max problems and algorithms by a relaxation procedure*, IEEE Transactions on Automatic Control **AC-25** (1980), 62–66.

[SEG00a] J. Su, J. J. Eggers, and B. Girod, *Illustration of the duality between channel coding and rate distortion with side information*, Proc. of the 34th Asilomar Conf. on Signals, Systems and Computers (CA, USA), November 2000.

[SEG00b] J. K. Su, Jochim J. Eggers, and B. Girod, *Optimum attack on digital watermarks and its defence*, IEEE International, 2000.

[Ser04] S. D. Servetto, *Lattice quantization with side information*, proc. of DCC (Snowbird, UT), March 2004.

[SH05] M. Sharif and B. Hassibi, *On the capacity of mimo broadcast channels with partial side information*, IEEE Transactions on Information Theory **IT-51** (2005), 506–522.

[Sha49] C. E. Shannon, *Communication in the presence of noise*, Proc. Institute of Radio Engineers, vol. 37, January 1949, pp. 10–21.

[Sha58] ———, *Channels with side information at the transmitter*, IBM journal of Research and Development **2** (1958), 289–293.

[SHG99] J. K. Su, F. Hartung, and B. Girod, *A channel model for a watermark attack*, Proc. SPIE Security and Watermarking of Multimedia Contents (San Jose, CA, USA), January 1999, pp. 159–170.

[SK04] D. Schonberg and D. Kirovski, *Fingerprinting and forensic analysis of multimedia*, Multimedia and Security Workshop (Magdeburg, Germany), September 2004.

[SKT98] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, *Multimedia data-embedding and watermarking technologies*, Proc. IEEE International Conference on Communications (BC-canada), vol. 2, June 1998, pp. 823–827.

[SVZ96] S. Shamai, S. Verdú, and R. Zamir, *Systematic lossy source/channel coding*, Proc. Int. Symp. Information Theory and its Applications (ISITA) (Victoria, BC, Canada), September, 17-20 1996, pp. 513–516.

[SVZ97] ———, *Information theoretic aspects of systematic coding*, Proc. Symp. Turbo Codes and Related Topics (ENST de Bretagne, Brest, France), September 1997, pp. 40–46.

[SVZ98] ———, *Systematic lossy source/channel coding*, IEEE Transactions on Information Theory **IT-44** (1998), 564–579.

[Sze79] W. Szepanski, *A signal theoretic method for creating forgery-proof documents for automatic verification*, Proc. Carnahan Conference on crime counter-measures (Lexington, KY), May 1979, pp. 101–109.

[SZT96] M. D. Swanson, B. Zhu, and A. H. Tewfik, *Robust data hiding for images*, in Proc. IEEE Digital Signal Processing Workshop (Loen, Norway), September 1996, pp. 37–40.

[SZTB98] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, *Robust audio watermarking using perceptual masking*, Elsevier Signal Processing, Special Issue on Copyright protection and Access Control, vol. 66, 1998, pp. 337–355.

[TBF+05] R. Tzschoppe, R. Bäuml, R. Fischer, J. Huber, and A. Kaup, *Additive non-gaussian noise attacks on the scalar costa scheme (scs)*, Proc. of SPIE & IST (San Joze, CA, USA), January 2005, pp. 114–123.

[TOH98] A. Z. Tirkel, C. F. Osborne, and T. E. Hall, *Image and watermark registration*, Signal Processing **66** (1998), 373–383.

[Tom71] M. Tomlinson, *New automatic equalizer employing modulo arithmetic*, IEEE Electonic Letter **7** (1971), 138–139.

[TOvS96] A. Z. Tirkel, C. F. Osborne, and R. G. van Schyndel, *Image watermarking- a spread spectrum application*, Proc. IEEE 4th International Symp. Spread Spectrum Technical Applications (Mainz, Germany), May 1996, pp. 785–789.

[VJG03] S. Viswanath, N. Jindal, and A. Goldsmith, *Duality, achievable rates and sum rate capacity of gaussian mimo broadcast channel*, IEEE Transactions on Information Theory **IT-49** (2003), 2658–2668.

[VKM+04] S. Voloshynovskiy, O. Koval, K. Mihcak, F. Pérez-Gonzáles, and T. Pun, *Data-hiding with host state at the encoder and partial side information at the decoder*, Submitted to IEEE Trans. Signal Processing (2004).

[VT03] P. Viswanath and D. N. Tse, *Sum capacity of the vector gaussian mimo broadcast channel*, IEEE Transactions on Information Theory **IT-49** (2003), 1912–1921.

[WSS04] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), *The capacity region of the gaussian mimo broadcast channel*, Proc. of IEEE Int. Symp. on Information Theory (Chicago, IL), June/July 2004, p. 174.

[WZ76] A. D. Wyner and J. Ziv, *The rate-distortion function for source coding with side information at the decoder*, IEEE Trans. on Inf. Theory **22** (1976), 1–10.

[XLC04] Z. X., A. D. Liveris, and S. Cheng, *Distributed source coding for sensor networks*, 04-th issue of the IEEE Signal Processing Magazine (2004).

[XSC+04] Z. Xiong, V. Stanković, S. Cheng, A. D. Liveris, and Y. Sun, *Source-channel coding for algebraic multiterminal binning*, IEEE Information Theory Workshop, ITW (USA), 2004.

[YC01] W. Yu and J. Cioffi, *Sum capacity of gaussian vector broadcast channels*, submitted to IEEE Transactions on Information Theory (2001).

[YCXZ03] Y. Yang, S. Cheng, Z. Xiong, and W. Zhao, *Wyner-ziv coding based on tcq and ldpc codes*, Proc. of Asilomar Conf. on Signals, Systems and Computers (Pacific Grove, CA, USA), November 2003.

[YSJ+01] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, *Writing on colored paper*, Proc. IEEE ISIT (Washington D.C.), June 2001, p. 302.

[Yu02] W. Yu, *Competition and cooperation in multi-user communication environments*, PhD thesis, Stanford University, 2002.

[ZGF02] Y. Zhao and J. Garcia-Frias, *Data compression of correlated non binary sources using punctured turbo codes*, proc. of DCC'02 (Snowbird, UT), April 2002.

[ZS98] R. Zamir and S. Shamai, *Nested linear/lattice codes for wyner-ziv encoding*, IEEE Information Theory Workshop (Killarney, Ireland), 1998.

[ZSE02] R. Zamir, S. Shamai, and U. Erez, *Nested linear/lattice codes for structured multiterminal binning*, IEEE Transactions on Information Theory **IT-48** (2002), 1250–1276.

Publications of the author

[myb]

[ZBD04] A. Zaidi, R. Boyer, and P. Duhamel, *A scaled signal plus noise model for digital watermarking, application to time jitter*, Proc. of Asilomar Conf. on Signals, Systems and Computers (CA, USA), vol. 2, November 2004, pp. 2165 – 2169.

[ZBD05] ———, *Time jitter vs. additive noise in a game theory context.*, Proc. of SPIE Conf. on Security, Steganography and Watermarking of Multimedia Content (CA, USA), January 2005.

[ZBD06] ———, *Audio watermarking under desynchronization and additive noise attacks*, IEEE Trans. Signal Processing **54** (2006), 570–584.

[ZD05a] A. Zaidi and P. Duhamel, *Modulo lattice additive noise channel for QIM watermarking*, proc of Int. Conf. Image Processing ICIP (Genova, Italy), September 2005, pp. 993–996.

[ZD05b] ———, *On coding with a partial knowledge of the state information*, Asilomar Conf. on Signals, Systems and Computers (Monterey, CA, USA), October 2005, pp. 657–661.

[ZD05c] ———, *Source-channel coding for lattice watermarking*, proc. of European Signal Processing Conf. EUSIPCO (Antalya, Turkey), september 2005.

[ZD06a] ———, *On capacity and bit error rate computation in finite-dimentional lattice coding for dirty paper coding*, Submitted to IEEE Trans. Information Theory (2006).

[ZD06b] ———, *On channel sensitivity to partially known two-sided state information.*, Proc. of the IEEE International Conference on Communications, ICC. (Istanbul, Turkey), September 2006, pp. 1520–1525.

[ZD06c] ———, *On coding for the gaussian broadcast channel with estimation error*, Submitted to the IEEE Globecom Conf. (San-Francisco, CA, USA), November 2006.

[ZP06] A. Zaidi and P. Piantanida, *Mac-aware coding strategy for multiple user information embedding*, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP. (Toulouse, France), May 2006, pp. 393–396.

[ZPD04] A. Zaidi, J. P. Piantanida, and P. Duhamel, *Scalar scheme for multiple user information embedding*, proc of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP (Philadelphia, USA), March 18-23 2004, pp. 5–8.

[ZPD07] ———, *Broadcast- and mac-aware coding strategies for multiple user information embedding*, IEEE Transactions on Signal Processing. (2007), 2974–2992.

Résumé Le problème de codage avec information adjacente (CCSI) est une technique récente d'annulation d'interférences en transmission et en compression de données. Ceci concerne les situations où l'émetteur est informé (par une voie retour par exemple) d'une partie de l'interférence canal. L'objectif est alors d' utiliser cette connaissance afin de concevoir un codage efficace. Une application étroitement liée à la transmission et à la compression de données est le "marquage de l'information" (information embedding). Potentiellement prometteur, le marquage d'information pose de nombreux défis dans différents domaines de recherche, allant de l'étude des limites théoriques de performance d'un point de vue théorie de l'information aux aspects liés à leurs implémentation d'un point de vue traitement de signal, en passant par la conception de code d'un point de vue communication numérique et codage.

Dans cette thèse nous considérons la problématique de marquage de l'information sous ses trois aspects: de théorie de l'information, de codage et communication et de traitement de signal. Le travail effectué dans le cadre de cette thèse peut être structurée en quatre parties. Dans la première partie, nous formalisons le problème de construction de dictionnaire comme un problème de conception de constellation. En particulier, nous montrons que le problème de codage avec information adjacente disponible à l'encodeur est fondamentalement un problème de codage conjoint source-canal. Ensuite, nous nous basons sur les réseaux de points imbriquées (nested lattices) pour la construction de bons codes algebriques à complexité réduite. Dans la deuxième partie, nous considérons le problème de marquage multiple comme un problème de communication multi-utilisateurs et nous construisons des stratégies de codage qui permettent d'approcher au mieux les limites théoriques de performances. La troisième partie traite le problème de sensibilité à l'information adjacente. Nous y évaluons la dégradation des performances due à une petite perturbation additive de l'information adjacente et nous y montrons que, dans certaines conditions, l'encodeur doit s'adapter à la perturbation en, éventuellement, changer sa stratégie de codage. La quatrième partie traite les performances du CCSI sur un canal AWGN avec jitter (AWGN&J) d'un point de vue théorie de jeux.

Mots clés Information adjacente, codage conjoint source-canal, canal de broadcast (BC), canaux à accès multiples (MAC), capacité et région de capacité, réseaux de points, théorie des jeux.

Abstract The problem of coding with state information (CCSI) is a new interference cancellation technique for both data transmission and data compression. It concerns all the situations where the transmitter knows a part of the interference in the channel (via a feedback loop, for example). The goal is then to use this knowledge about the channel in order to conceive an efficient coding scheme. One potentially promising application, at the cross-road of both data transmission and data compression, is information embedding. The embedding of information poses many challenges in a variety of reseach areas. This involves information theory for assessing the theoretic limits of performance, signal processing for implementation issues and communication theory for code design.

In this thesis, we consider the problem of information embedding in its three aspects. The work can be structured into four parts. In the first part, information embedding is mathematically formalized as a joint source-channel coding problem. For instance, we show that the problem of CCSI available at the transmitter is basically a joint source-channel coding probelem. Next, we use nested lattices for the design of good low-complexity algebraic-based codes. In the second part, we consider the problem of multiple user information embedding (recognized as a multi-user communication problem) and conceive structured codebooks and appropriate coding strategies that closely approache the theoretic limits. The third part concerns channel sensitivity to little perturbations of the state information. We evaluate the loss in performance due to a weak additive contaminating state information and show that, under certain circumstances, the transmitter must adapt to the available knowledge about the channel, by (eventually) changing its coding strategy. The fourth part determines the performance of CCSI over an AWGN channel with jitter (AWGN&J) in a game theory context.

Key words State information, joint source-channel coding, broadcast channel (BC), multiple access channel (MAC), capacity and capacity region, lattices, game theory.