UNIVERSIDADE DE VIGO

ESCOLA TÉCNICA SUPERIOR
DE ENXEÑEIROS DE TELECOMUNICACIÓN

Ph.D. Thesis

Submitted for the European Doctor mention

SIDE-INFORMED DATA HIDING: ROBUSTNESS AND SECURITY ANALYSIS

Author: Pedro Comesaña Alfaro
Advisor: Fernando Pérez-González

2006

Abstract

Whenever a researcher faces the design of a data hiding method, probably the two basic requirements that he/she will always take into account (besides perceptual imperceptibility) will be the robustness and security of that method. The core of this thesis is devoted to the analysis of these requirements for those data hiding methods with side information available at the embedder, i.e., where the embedder takes advantage of deterministically knowing the document to be watermarked. These methods, which currently constitute the state of the art of data hiding methods, have been extensively studied in the literature; there are however several important aspects which deserve attention and which have been analyzed in this thesis.

Concerning robustness, in this work we have studied the performance under additive noise of arguably the most extended side-informed method, i.e. Distortion-Compensated Dither Modulation (DC-DM), when uniform scalar quantizers and repetition coding are used; even though several theoretical results are available in the literature, they are only partial. In this thesis we have obtained for the first time the exact results, as well as some approximations and bounds which are shown to be useful for developing improved versions of DC-DM. A novel topic introduced in this thesis is the analysis of the decoding error probability of DC-DM with uniform scalar quantizers and repetition coding when the watermarked documents undergo a coarse quantization. Other side-informed method often used in the literature is the so-called Spread-Transform Dither Modulation (STDM), which is usually assumed to outperform DC-DM. Nevertheless, in this thesis we show that DC-DM with uniform scalar quantizers and repetition coding could be more robust against the cropping attacks, which have been studied for the first time for side-informed methods. Probably one of the most significant contributions of this work is the design of a novel sensitivity attack, named the Blind Newton Sensitivity Attack (BNSA), which was shown to be effective against a wide range of state-of-the-art data hiding methods. The robustness chapter ends with a game theoretic approach, where some optimized strategies are proposed for additive spread spectrum (Add-SS), DC-DM and STDM, and the exposition of some empirical results about the worst case (from an achievable rate point of view) additive attack for scalar DC-DM.

On the other hand, the security part is mainly focused on the security analysis of Add-SS data hiding methods and Costa's construction. Furthermore, the security of Add-SS and Costa's construction is compared with some existing results about the security of DC-DM. Our results have been obtained by following a novel approach.

Besides robustness and security of side-informed data hiding techniques, some results are introduced concerning the need of combining channel-coding and source-coding in order to achieve dirty paper channel capacity, and the adaptation of one of those schemes to hide data in real images. Finally, an application of a data hiding scheme to a video surveillance authentication system is discussed.

Resumen

Cuando un investigador afronta el diseño de un método de ocultación de datos, probablemente los dos requisitos básicos que siempre tendrá en cuenta (aparte de la imperceptibilidad) serán la robustez y la seguridad del método. El núcleo de esta tesis está dedicado al análisis de estos requisitos para los métodos de ocultación de datos con información lateral disponible en el codificador, es decir, donde el codificador aprovecha el hecho de conocer de forma determinista el documento a marcar. Estos métodos, que actualmente constituyen el estado del arte en ocultación de datos, han sido ampliamente estudiados en la literatura; de todos modos hay importantes aspectos que merecen atención y que han sido analizados en esta tesis.

En cuanto a la robustez se refiere, en este trabajo hemos estudiado las prestaciones del que probablemente es el método con información lateral más extendido, el *Distortion-Compensated Dither Modulation* (DC-DM), ante ruido aditivo, cuando se emplean cuantificadores escalares y codificación por repetición; aun cuando en la literatura hay varios resultados teóricos, éstos son únicamente parciales. En esta tesis hemos obtenido por primera vez los resultado exactos, así como algunas aproximaciones y cotas que demuestran ser útiles para desarrollar versiones mejoradas de DC-DM. Un novedoso tema presentado en esta tesis es el análisis de la probabilidad de error de decodificación de DC-DM con cuantificadores escalares uniformes y codificación por repetición cuando los documentos marcados sufren una cuantificación gruesa. Otro método con información lateral que se emplea frecuentemente en la literatura es el *Spread-Transform Dither Modulation* (STDM), que ha menudo se ha dado por hecho que mejoraba a DC-DM. Sin embargo, en esta tesis demostramos que DC-DM con cuantificadores escalares uniformes y codificación por repetición podría ser más robusto contra ataques de recortado, que aquí se han estudiado por primera vez para métodos con información lateral. Probablemente una de las contribuciones más significativas de este trabajo es el diseño de un novedoso ataque de sensibilidad, al que llamamos *Blind Newton Sensitivity Attack* (BNSA), y que ha mostrado su efectividad contra un amplio abanico de métodos dentro de los que pueden ser considerados como estado del arte en ocultación de datos. El capítulo de robustez termina con una aproximación a la teoría de juegos donde se proponen algunas estrategias optimizadas para *additive spread spectrum* (Add-SS), DC-DM y STDM, y

la exposición de algunos resultados empíricos acerca del ataque aditivo caso peor (desde un punto de vista de tasa alcanzable) para DC-DM escalar.

Por otro lado, la parte de seguridad se centra principalmente en el análisis de seguridad de los métodos de ocultación de información basados en Add-SS y la construcción de Costa. Además, se compara la seguridad de Add-SS con la de la construcción de Costa, y ambas con algunos resultados existentes sobre la seguridad de DC-DM. Nuestros resultados se han obtenido siguiendo un novedoso enfoque.

Además de la robustez y seguridad de las ténicas de ocultación de datos con información lateral, se presentan algunos resultados relacionados con la necesidad de combinar codificación de canal con codificación de fuente para alcanzar la capacidad de canal en el escenario de *dirty paper*, y uno de estos esquemas ha sido adaptado para ocultar datos en imágenes reales. Finalmente, se discute la aplicación de un esquema de ocultación de datos a un sistema de autenticación de imágenes de videovigilancia.

Agradecimientos

Creo sinceramente que es un ejercicio necesario cuando uno llega al final de una etapa de su vida, el echar la mirada hacia atrás y hacer balance de ese período: de lo que cree que ha sido positivo y negativo, de lo que ha aprendido, de cómo ha evolucionado en todas las facetas de su vida en esa etapa, y por supuesto de lo que debe mejorar. Y de lo primero que me he dado cuenta haciendo el análisis de estos casi cuatro años de tesis, es de la gran cantidad de gente a la que le debo agradecer su apoyo, consejo y ayuda a lo largo de estos años. Por tanto, no es para mí una mera prueba de cortesía, si no más bien una obligación moral, el mostrar mi agradecimiento a alguna de esa gente; no me atreveré a decir que están todos lo que son, porque mucho me temo que me olvidaré a alguno, pero sí puedo asegurar que todos lo que están, son.

No podría empezar esta lista de agradecimientos por otra persona que no fuese Fernando. Desde luego todo calificativo se queda corto acerca de su calidad profesional como docente e investigador. De lo que podemos dar fe un menor número de afortunados, es de su tremenda dedicación como tutor y de su calidad humana. Aparte de su generosidad proponiendo ideas, de su completa entrega, y de su espíritu de perfeccionamiento, me gustaría destacar sus dotes como psicólogo. Gracias por haberme dejado pertenecer a ese pequeño grupo de afortunados.

Durante gran parte del período de elaboración de esta tesis, estuve acompañado en mis cuitas por Luis, al que también deseo agradecerle todo su apoyo, y con el que he de confesar que ha sido muy agradable trabajar.

Durante el primer año de mi trabajo en esta tesis, recibí una incalculable ayuda por parte de Félix, desde los más pequeños detalles, hasta las más elaboradas discusiones. Desde entonces, y a pesar de la distancia, siempre he sentido su apoyo incondicional; siempre ha estado dispuesto a echar una mano cuando lo necesité, ya fuese por alguna pregunta técnica, ya para que me acompañase a hacer la mudanza. Está claro que los de Lugo están hechos de una pasta especial.

Me gustaría tener un recuerdo para las dos últimas incorporaciones al mundo del watermarking: Juan Ramón y Gabriel. Desde luego también le agradezco a Gabriel su buen trabajo en la aplicación de autenticación de imágenes de videovigilancia.

Agradezco a Carlos Mosquera, Roberto López-Valcarce y Nuria González-Prelcic la buena acogida que siempre me han dado en el Grupo de Procesado de Señal en Comunicaciones.

Por supuesto es de justicia tener también un momento de recuerdo para la Fundación Pedro Barrié de la Maza, y especialmente para Montserrat Ortas, por las facilidades prestadas para mi estancia en Eindhoven.

Special thanks are due to Dr. Frans Willems and Prof. Ton Kalker, who kindly received me in the Technische Universiteit Eindhoven, for their attention and enlightening discussions. Thanks are also due to Dr. Teddy Furon for his valuable comments on the security papers which constitute a part of this thesis, and to Prof. Ingemar J. Cox and Dr. Gwenaël Doërr who kindly provided us with the source code of their algorithm. I want to effusively thank to Prof. Slava Voloshynovskiy, Dr. Oleksiy Koval, Dr. José Emilio Vila Forcén, Emre Topak and Renato Villán their attentions when I visited them in Geneva.

I want also to thank to Joep Kierkels, Tanya Ignatenko, Harald van den Meerendonk and Suzanne Martens their warm hospitality for the six months I was in Eindhoven.

Gracias a todos los inquilinos del TSC-5 de los últimos 4 años. Seguro que me olvido a alguno, pero ahí van los nombres de los que me acuerdo: Fran Campillo (espero que sigas siendo igual de buena gente cuando vuelvas de Portland), Francisco Javier Diéguez Tirado (alias Conguito, JDT y chinito agitador... gracias por enseñarme tus trucos para conseguir billetes de avión más baratos), Francisco Javier Pazó (alias Abu; desde luego tienes mucha fuerza en los dedos, pero también tienes un corazón que no te cabe en el pecho), Fernando Piñeiro (alias Coqui; para aguantarnos hay que tener mucha paciencia... espera, que esto no funciona), Armando (desde que se nos fue el pollo, el gallinero ya no es lo mismo), Norberto (para ser de Barakaldo, pareces de Bilbao), Elisardo (English is easier with you), Enrique (alias Yogu; ciertas miradas no tienen precio), Gonzalo (detrás de cada hombre pequeño, hay una gran silla con un gran hombre), Marta Capdevila (a veces dudo de cómo nos puedes aguantar) y Brandán (siempre entre el vicio y la virtud). Me falta un individuo, al que tengo mucho aprecio, pero como no está en el TSC-5, lo tengo que poner aparte: Dani(el González Jiménez); los días que uno está de bajón, tú le alegras la hora de la comida (espero que nunca cambies).

Y para terminar, las más merecidas y sentidas gracias a Sissita, Paula, Cris y Ma(rité). Espero estar algún día a la altura de todo lo que me dáis.

Contents

List of Figures

Acronyms and Abbreviations

AC	Alternating Current
ACC	Accumulator
Add-SS	Additive Spread Spectrum
AM	Amplitude Modulation
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
BNSA	Blind Newton Sensitivity Attack
CLT	Central Limit Theorem
CM	Constant Message
CMA	Constant Message Attack
CND	Check Node Decoder
CNE	Check Node Encoder
CP-SNS	Continuous Periodic functions for Self-Noise Suppression
DC	Direct Current
DC-DM	Distortion-Compensated Dither Modulation
DC-QIM	Distortion-Compensated Quantization Index Modulation
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DICOM	Digital Imaging and Communications in Medicine
DVB-T	Digital Video Broadcasting-Terrestrial
DVD	Digital Versatile Disc
DWR	Document-to-Watermark Ratio
EOA	Estimated Original Attack
EWR	Estimation error to Watermark Ratio
FIM	Fisher Information Matrix
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
i.i.d.	Independent and Identically Distributed
IP	Internet Protocol
IPR	Intellectual Property Rights

JANIS	Just Another N-order Side-Informed Scheme
JND	Just Noticeable Difference
JPEG	Joint Photographic Experts Group
KLT	Karhunen-Loève Transform
KMA	Known Message Attack
KOA	Known Original Attack
LDPC	Low Density Parity Check
LMS	Least Mean Squares
MAP	Maximum A Posteriori
MER	Message Error Rate
MJPEG	Motion JPEG
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
PAM	Pulse-Amplitude Modulation
PCA	Principal Components Analysis
pdf	Probability Density Function
pmf	Probability Mass Function
PSC	Power Spectrum Condition
PSNR	Peak Signal to Noise Ratio
QAM	Quadrature Amplitude Modulation
QF	Quality Factor
QIM	Quantization Index Modulation
QP	Quantized Projection
QPD	Quantized Projection based Detection
ROC	Receiver Operating Characteristic
SBE	Scaled Bin Encoding
SCS	Scalar Costa Scheme
SNR	Signal to Noise Ratio
SSTDM	Scalar Spread-Transform Dither Modulation
STDM	Spread-Transform Dither Modulation
ST-SCS	Spread Transform-Scalar Costa Scheme
TCP	Transmission Control Protocol
UB	Union Bound
VND	Variable Node Decoder
VNE	Variable Node Encoder
VQ	Vector Quantizer
WCAA	Worst Case Additive Attack
WNR	Watermark-to-Noise Ratio

WOA Watermarked Only Attack

Notation

$\mathbb{Z}$	Set of integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^+$	Set of positive real numbers
$\mathbb{C}$	Set of complex numbers
j	Imaginary unit
$\mathbf{X}, \mathbf{x}$	Original host signal (random column vector, deterministic value)
$\mathbf{W}, \mathbf{w}$	Watermark signal (random column vector, deterministic value)
$\mathbf{Y}, \mathbf{y}$	Watermarked signal (random column vector, deterministic value)
$\mathbf{N}, \mathbf{n}$	Channel noise signal (random column vector, deterministic value)
$\mathbf{Z}, \mathbf{z}$	Received signal (random column vector, deterministic value)
$\mathbf{B}, \mathbf{b}$	Transmitted message (random column vector, deterministic value)
$\hat{b}_i$	Estimate of the i-th transmitted symbol
$\boldsymbol{\gamma}$	Perceptual mask
a_i	i-th component of vector $\mathbf{a}$
A_i	i-th component of random vector $\mathbf{A}$
$\sigma^2_{A_i}$	Power of random variable A_i
D_h	Average host signal power
D_w	Average watermark signal power
D_c	Average channel distortion power
L_1	Length of the original host signal
L_b	Length of the message to be conveyed
L_2	Number of samples of the original host signal per transmitted symbol
L_3	Dimensionality of the projected domain
L_4	Number of samples of the projected host signal per transmitted symbol
$\boldsymbol{\Theta}$	Secret key
P	Size of the alphabet of symbols
$\mathbf{s}$	Spreading sequence used by Add-SS
$\mathbf{S}$	Projecting matrix
$\mathbf{s}^T$	Transpose of the generic vector $\mathbf{s}$
$\mathbf{S}^T$	Transpose of the generic matrix $\mathbf{S}$
Λ	Used for denoting a generic lattice

$\mathcal{V}(\Lambda)$	Fundamental Voronoi region of Λ
$Q_\Lambda(\cdot)$	Minimum Euclidean distance quantizer using lattice Λ
$\mathbf{x} \mod \Lambda$	Modulo-Λ reduced version of the generic vector $\mathbf{x}$
$\mathbf{x}_{\text{mod}}$	Modulo-reduced version of the generic vector $\mathbf{x}$
$Q_b(\cdot)$	Quantizer corresponding to the b-th symbol
$\mathbf{E}$	Quantization error resulting from quantizing $\mathbf{X}$ with the quantizer $Q_b(\cdot)$
α	Distortion-compensation parameter
$\mathbf{v}(b)$	Dither vector
$\mathbf{d}$	Key-dependent part of the dither vector
$\arg\max_{x\in\mathcal{X}} f(\mathbf{x})$	It provides the element/s of the set $\mathcal{X}$ that maximizes $f(\cdot)$
$\arg\min_{x\in\mathcal{X}} f(\mathbf{x})$	It provides the element/s of the set $\mathcal{X}$ that minimizes $f(\cdot)$
$\max_{x\in\mathcal{X}} f(\mathbf{x})$	Maximum value of $f(\cdot)$ when it is evaluated over the elements of the set $\mathcal{X}$
$\min_{x\in\mathcal{X}} f(\mathbf{x})$	Minimum value of $f(\cdot)$ when it is evaluated over the elements of the set $\mathcal{X}$
$f_X(\cdot)$	Probability density function of the generic continuous random variable X
$f_{\mathbf{X}}(\cdot)$	Probability density function of the generic continuous random vector $\mathbf{X}$
$\mathbf{U}$	Modulo-reduced version of the normalized total noise
$\boldsymbol{\beta}$	Vector of weights intended to improve the decoding
$\mathbf{x}_p$	Projected version of the generic vector $\mathbf{x}$
$\mathbf{t}$	Attacking vector for the sensitivity attacks
$\mathcal{R}$	Acceptance or detection region
$\mathcal{R}^c$	Rejection region
$H(X)$	Entropy of the discrete random variable X
$H(X\|Y)$	Entropy of the discrete random variable X, given Y
$h(X)$	Differential entropy of the continuous random variable X
$h(X\|Y)$	Differential entropy of the continuous random variable X, given Y
$I(X;Y)$	Mutual information between X and Y
$I(X;Y\|Z)$	Mutual information between X and Y, given Z
$\mathbf{Y}^n$	n-th observation
N_o	Number of observations
$\hat{\mathbf{X}}$	Estimate of $\mathbf{X}$
$\tilde{\mathbf{X}}$	Estimation error of $\mathbf{X}$
$\mathcal{U}$	Codebook for Costa's construction
$\mathcal{U}_b$	Bin corresponding to message b

Chapter 1

Introduction

The mere nature of humankind makes us prone to hide information. For example, in our childhood we try to hide those behaviors that could be susceptible of being condemned by our parents, and we like to share *secrets* with that bunch of children that we consider to be our best friends. These basic structures of secrecy and trust get more complex, or at least we usually think so, as we get older. Nevertheless, even when we are adult these behaviors persist in several aspects of our lives. For example in Spanish it is typical to say that somebody is *playing poker* when he/she tries to hide his/her final intentions; in the same way, it is also usually said that if you meet a Galician person at the stairs, you will not be able to decide whether he/she is going up or down. We will not try to prove here the good *genetic predisposition* of Galician people for data hiding, since, of course, it is not true (or is it?).[1] In fact, the target of this first paragraph is just to show that data hiding can be considered as a natural human attitude, although sometimes we are encouraged to *hide* it.

Due to this natural tendency to hide information, data hiding has been used from the ancients to our days for different purposes (covert communications, intellectual property rights, traitor back tracing, etc.) as it is widely described in the literature (see, for example [95] and [136]). Therefore, it is not rare that with the growth of digital technologies in the last years, data hiding has been used to convey information in digital multimedia contents, such as still images, video, audio or electronic text, attracting a lot of attention from academia and industry. As a result of that interest, a huge number of digital data hiding methods have been proposed; before proceeding with the exposition of our results, we will introduce some of the applications of those methods, to get a general idea of what is under the wide term of digital data hiding.

[1] Dear reader, please take this just as a sample of Galician sense of humor. Of course we do not believe on any kind of *genetic predisposition*, and even less if it is mixed with *geographical* circumstances.

1.1. Applications

In this section some applications of data hiding in digital contents will be briefly described, comparing its choice with other alternative techniques:

- Copyright protection: in this application a signal, named *watermark*, is added to the original host signal, enabling the claim of ownership of the content. This gives rise to the term *(invisible) watermarking*, which is frequently used to replace *data hiding*, even when it is not applied to copyright protection; in fact, *watermark* is widely used to denote the signal added to the original host for all data hiding applications. In any case, one should be also aware that *watermarking* also encompasses the so-called *visible/perceptible watermarking*, where a signal is superimposed to the content to be protected, in such a way that the superimposed signal is perfectly identifiable; when the content is an image, that *perceptible watermark* is usually a logo. Since the superimposed image can be identified, we can not consider visible watermarking as part of data hiding.

  Copyright protection was one of the first digital data hiding applications, and it is also one of the most challenging, since the watermark should resist attacks of opponents aware of its existence; these attacks can range from typical signal processing (filtering, noise addition, lossy compression) to more sophisticated ones (e.g. geometrical and desynchronization attacks, private parameters leakage, etc.). Due to this, the task of designing a *good* watermarking technique seems to be a cumbersome problem, and in the last years copyright protection has no longer been the major data hiding application, being somehow relegated by some of the applications introduced below.

  Since it is just a piece of the architecture trying to preserve the rights of the actual holder and avoid the claim of the ownership by an impostor, data hiding for copyright protection should take into account the existing legal framework, as it could be considered as an evidence in a trial.

  Sometimes, a cryptographic approach has been proposed to avoid the access of bad-intentioned parties to the content, thus preventing the ownership claim. Nevertheless, this is not a complete solution, since a legal user should always be able to properly decode the content, and once he/she has access to the signal, a misuse is possible. Therefore, a second level of security should be established to enforce copyright protection. This second level should rely on information somehow inseparable from the content itself, in such a way that the proper use of the signal by a legal user does not limit the validity of the technique; this is the role to be played by data hiding. Then, data hiding and cryptography should not be regarded as *rival*, not even as *alternative* technologies, but as *complementary* and not mutually excluding solutions [10, 147, 90, 136, 59, 135, 56, 55, 64, 16, 120, 1].

- Fingerprinting: a serial number, which identifies the customer, is hidden in the host signal. In this way, if unauthorized copies were found, the owner of the copyright could trace back the identity of the customer that supplied the content to third parties and prosecute him/her. This application has been paid increasing attention in the last years, and the fact of identifying the traitor/s seems to be also really challenging; nevertheless, in some scenarios where the number of possible users is limited, even when a different watermark is associated with each of them, the complexity of the problem is reduced.

  A special attack to this application is the so-called *collusion attack*, where a number of users combine their respective watermarked versions to yield another content, assuming that this combination will make more difficult the estimation of the identity of any of them. On the other hand, the detector will try to identify as many *colluders* as possible, or, at least, to identify one of them, in such a way that this could prevent the users from *colluding*; on the other hand, the detector has to deal also with the danger of wrongly identify an innocent user. Recent references in the literature can be found about this attack [104, 61].

  Alternatives to this application could be the attachment of user identification files, or the usage of metadata fields conveying information about the user in some storage formats, e.g. JPEG-2000 or MPEG-4; nevertheless, these mechanisms can be easily removed, so a content embedded technique, as data hiding, is required [10, 90, 136, 59, 135, 56, 64, 14, 16, 120, 2, 1].

- Steganography (Covert communications): in this case the embedder does not just want to hide the conveyed information, but he/she tries to hide the process of communication itself, in such a way that an observer of the watermarked signal should not be able to state whether a watermark was hidden or not. Steganography can be used in military communications to avoid that the enemy knows that the communication is taking place. In that sense, steganography resembles the use of low-powered communication techniques, such as spread spectrum, since an observer unaware of the spreading sequence (equivalently, a secret parameter for steganography) will not be able to decide whether a communication is being performed.

  In [27] a steganography scheme is called *perfect* when the Kullback-Leibler distance between the probability distribution of the original host signal and its watermarked version is zero, i.e. when both signals are statistically undistinguishable.

  Alternative techniques to digital data hiding could be the broadcast of the aforementioned low-powered modulations, since they are usually unnoticed due to the their noise-like appearance; in fact, when the communication channel is constrained to be a multimedia content these modulations are frequently used as data hiding techniques. Further information about steganography can be found in [10, 147, 90, 136, 59, 120].

- Data authentication, integrity control or tamper proofing: due to the fast evolution of digital multimedia editing tools, the modification of original contents is nowadays easy and cheap. In order to realize the danger of this editing easiness, we can think of tampering a video surveillance system, replacing the identity of a person in a picture, or the modification of a text document.

  Data authentication, and similarly the related terms integrity control and tamper proofing, deals with this problem by changing the watermark detector output when the watermarked signal is modified, even after small changes, so those modifications on the watermarked content can be detected. Due to their sensitivity, these schemes are said to use *fragile watermarks*; they are employed, for example, in automatic video surveillance and authentication of drivers licenses.

  Cryptographic functions could be also used in this problem: a content provider could compute the hash of the sold signal, and then encrypt this hash with his/her private key, sending both the original host signal and the encrypted hash to the buyers. In this way the buyer will decrypt the hash using the public key of the provider, and compare it with the result of computing the hash of the received signal, so if both of them coincide, he/she can be sure that the received signal was not modified in the middleway. The validity of this scheme, which is usually termed *digital signature*, lies on the one-way nature of hash and encryption functions, meaning that it is computationally unfeasible to find another image which yield the same hash or to invert the encryption algorithm. Nevertheless, one must be aware that this alternative implies an increase in the payload, and that some typical signal processing, such as quantization for lossy compression, that are not intended for tampering with the signal, will be probably not allowed; these problems could be avoided by using data hiding, given that the watermark is conveyed by the host signal, so in general it does not increase the payload, and it could be robust against some kind of attacks, as lossy compression or other unintentional attacks [90, 136, 59, 135, 56, 55, 64, 16, 120, 1].

- Media Forensics: it goes a step further than authentication and tries to determine how the content was modified. For example, it could inform about which part of an image was modified [120], providing information about the attacker's intention to modify it.

- Reversible data-hiding: the target is not only to recover the embedded data, but also the restoration of the original host signal, just from its watermarked version [74, 98]. In this way, an allowed user could simultaneously extract the information from the received signal and improve the quality of the received content (compared with the original); on the other hand, a user without permission could not properly restore the received signal and his/her version would be poorer.

- Standards migration: Data hiding could help in the transition from one standard to another. Since the replacement of the complete infrastructure related to a given standard is extremely expensive, and this transit is not instantaneous, data hiding could be useful in the meantime to convey the information which makes possible the compatibility of the former and new standards; an example of this could be the MPEG transition [13] or the transmission of a digital signal within existing analog (Amplitude Modulation and/or Frequency Modulation) broadcast radio without interfering with conventional analog reception [32].

- Data tracking: a web-crawler tries to find copies of contents watermarked with a given watermark on the Internet or on the local network; the target is to find misuses of materials of a rights holder. A typical scenario is the search of copyrighted images in the websites of users who do not have the corresponding permissions, allowing their prosecution.

  One could think again in the use of metadata fields to convey information about the rights holder, making easier the detection of copyrighted signals; unfortunately, as it was previouly said, these fields are easily modifiable, so any bad-intentioned user could remove them [147, 90, 136, 135, 55, 2, 1].

- Data monitoring: the detector tries to verify the presence of a given watermark in the received signal. This is typically used for counting the times that a watermarked signal, e.g. an advertisement or a song, is played in a broadcasting media; in this way the advertiser or the owner of the rights of the song can get valuable information for billing. In the first case the radio station will not be interested in removing the watermark, but it could be interested in doing so in the second one, so data hiding is more reliable than metadata; furthermore, metadata fields could not be available, depending on the radio standard, e.g. amplitude modulation (AM), so data hiding seems to be a good choice [10, 147, 90, 59, 135, 56, 55, 64, 14, 120, 1].

- Embedded transmission of added-value services within multimedia data, as multilingual soundtracks or extra scenes: data hiding allows the sending of additional information without noticeably modifying the quality of the original content, neither increasing the bandwidth [147, 90, 14, 120].

- Communication of meta information within digital multimedia data [14, 2]: in some medical image formats, e.g. DICOM [4], some information, as the patient's name, goes in a separate file; with data hiding applications the medical safety could be improved by embedding such information in the image [10], as well as simplifying the file system and reducing the overhead.

- Tracking content creation, manipulation, and modification history: data hiding allows to track content creation, manipulation, and modification history without the overhead associated with creating a separate header or

history file. Data hiding could provide a useful way of storing the aforementioned information within the content itself. The embedding algorithm could be completely open, since this application does not seem to be susceptible of being attacked [147, 120].

- Providing different access levels to the data: for example, the amount of detail that can be seen in a given image can be controlled, in such a way that a user with a high access level can see details that another user with a lower access level would not see. This somehow resembles the hierarchical modulation used in Digital Video Broadcasting-Terrestrial (DVB-T), in the sense that the same signal can provide different quality levels, but whereas in DVB-T this is due to the available Signal to Noise Ratio (SNR), in data hiding it depends on the permissions the user has been granted. A possible partial alternative could be the use of hierarchical encryption [147, 59].

- Preventing unauthorized copying: related to the previous one. Data hiding was proposed to be implemented by the DVD (Digital Versatile Disc) video standard. Nevertheless, its success has been limited, since it increases the cost of the DVD recorder while reducing its value, meaning that users are usually interested in copying DVDs, even illegally, so hardly they are going pay in order to loose this feature.

  One could think of a cryptographic approach as a possible solution, but it is straightforward to see that it will not solve the problem, since, as it was said previously, as far as a legal user is able to decode the signal, he/she will be able to misuse it. Therefore, a signal embedded in the content, such as the watermark introduced by the data hiding algorithm, seems to be a good choice, since it can not be easily removed [25, 136, 135, 56, 64, 14, 16, 1].

- Labeling for user awareness: a warning encouraging the customer to buy the watermarked content could appear when he/she tries to save it [135]. It could be seen as an example of the old sentence: "Keep Honest People Honest", since somebody really trying to break the copyrights is not going to get discouraged by this kind of measure.

- Annotation and linking content to a web or database: the commercials including a typed web address where more information can be found are usual. The novelty of this application is that such information is embedded in the image/sound of the commercial itself, enabling to link the printed material or audio with their corresponding resources in the Internet or databases; this allows to complete the initial information, the sight of the advertisement of some product, or its purchase [10, 147, 136, 59, 55, 16, 120, 1].

- Device control: examples of this kind of application are available already for analog signals, allowing receivers to remove advertisments, or even synchronization of children's toys with live-broadcast or recorded video. The former function can be seen as similar to some headers in modern digital

standards, such as MPEG-4, where some *events*, e.g. the beginning or end of a film, the commercials, etc., are announced, enabling the recording and stopping of a digital video recorder. Using this application, the same could be done for analog video [147, 56, 55, 120, 2, 1].

- Estimate of the quality of multimedia communication links: digital fragile watermarking has been also used to blindly estimate the quality of multimedia communication links. This quality assessment system does not increase the bit rate and is based on the evaluation of the mean-squared-error between the estimated and the actual watermarks. The estimated quality of the received signal can be used by the service provider as a feedback information for billing purposes, to control feedback to the sending user, or by the operator to diagnose the effective status of the link [29].

We would like to remark again that in the last years the interest of researching data hiding community has moved from the initial copyright protection application to other, somehow less challenging, problems such as fingerprinting, steganography or authentication; this tendency can be noticed just by looking at the technical program of major data hiding conferences.

As a final comment, we emphasize that this overview does not try to be exhaustive, and that some of the former applications could be, and in fact are, overlapping.

1.2. Content types and formats

Digital data hiding is applicable to any kind of digital content such that a given instance is suitable for being represented in two (or more) different ways without noticeable differences; by choosing which of these representations is transmitted, the embedder is sending additional information which could be recovered at the decoder when an *a priori* protocol is established between them. This concept seems to be colliding with the source-coding principle, where the number of representations of the possible contents (codewords) is tried to be reduced as much as possible,[2] maintaining a given perceptual distortion. In this way, the size of the source-coded content is reduced compared with the original one, but another different instance of the source-coded content can be distinguished from the original signal (in other case, some redundancy is still available).

As it is well known, typical multimedia contents, as still images, audio or video, as far as they are not the outcome of perfect source coding, have some redundancy, so information can be hidden within them without perceptual distortion; in fact, a

[2]Strictly speaking, the encoder tries to minimize the entropy of the source-coded content, i.e. the entropy of the chosen codeword.

vast literature has been written about data hiding in those contents (an overview can be found in [147] or [90]). However, some other contents have been proposed for being watermarked, with special features and requirements:

- Watermarking of electronic text documents: usually this is a fingerprinting appplication, where the watermark is embedded by slightly increasing or decreasing the spaces between words or the distance between any two adjacent lines, according to the value of the corresponding watermark bit [7]. The noise sources are related with the irregular spacing between words before watermarking (zero for unjustified text), and printing and scanning noise. This technique is said to have been used by Margaret Thatcher to trace disloyal ministers who were leaking cabinet documents to the press [10].

- Semagrams: related with the previous one, in this case the information is hidden by very slight physical differences in appearance such as special fonts, punctuation marks, or very fine dots [147].

- Watermarking of software: the presence of the watermark must not change the functionality of the software. It is usually based on the order in which registers are pushed and popped, the automatic random replacement of code fragments with equivalent ones [10], the memory trace of an executing Java program [105], the topology of a dynamically built graph structure [35], or a program's control flow graph [153]. These techniques could be combined with *obfuscation* [36], where the program is transformed into an equivalent one, in order to hamper reverse engineering.

- Watermarking for digital cinema: since the number of cinemas where a film is exhibited is limited, a different watermark could be introduced in each copy (i.e. fingerprinting), in such a way that an illegal copy could be traced back to find the traitor. The system does not need to be unbreakable, but the attacker should spend a couple of months in breaking it, since in this period the value of the movie will have declined drastically. Further considerations about this subject are made in [102].

- Watermarking of natural language: it exploits the fact that the same idea can be expressed in different ways. For example, a sentence could be changed to its passive form without modifying its meaning, the order of some words could be changed, a word could be replaced by a synonymous, etc.. This topic was extensively studied in [149].

- Watermarking of 3D objects: some works proposed in the literature (e.g., see [9], [8] or [101]) are based on the slight modification of the mesh defining the 3D object.

- Watermarking based on automatic translation: due to the redundancy existent in natural language, there is a lot of choices for an automatic translator. Furthermore, the authors of [84] defend that the automatic translators have so many errors, that the inaccuracies due to their use to embed information would be somehow *masked*.

- Watermarking TCP/IP protocols: in [123] a method is proposed that embeds information in the *IP identification* and *TCP initial sequence number* headers.

- Watermarking of maps: an interesting utility of watermarking was introduced in [83], where the contour lines of maps are parametrized in order to hide information which enables a fingerprinting application.

1.3. Requirements

Once some of the applications of data hiding and the contents suitable for being watermarked have been explored, one can wonder which characteristics should be fulfilled by a good data hiding scheme. Fortunately, there seems to be some consensus in the literature [147, 90, 136, 135, 32, 56, 64, 11, 16], and following are the most representative of them:

- Payload: the number of bits to be hidden is lower-bounded for each application.

- Computational cost: the complexity of the algorithms should be limited, in order to be implementable. Depending on the application, this requirement will be more pressing in the embedder or in the decoder. Furthermore, some applications could have to be run in real-time.

- Imperceptibility: this is probably one of the most obvious and yet important. From the begining of digital data hiding, it was clear that the watermarked signal should be perceptually indistinguishable from the original one. In this way, the embedding process would not undermine the value of the watermarked product. This constraint is sometimes deliberately broken, e.g. when the watermark is inserted in thumbnail images, since it tries to allow the user to see the image before buying it, but not to provide a good enough quality, thus forcing the purchase [147].

- Robustness: it is usually understood as the ability of a data hiding scheme to survive attacks such as noise addition, compression, filtering, cropping, rotation, rescaling, amplitude modifications, etc.

- Security: it is usually related with the ability of a data hiding scheme to protect some secret parameter (sometimes the information-hiding code), in such a way that an attacker can not use it to modify the watermark or create falsely watermarked contents (forgeries).
- Detectability: in some applications, such as steganography, the presence of the watermark should not be detectable, so the watermarked signal is required to be statistically indistinguishable from the original [135, 120].

We would like to remark that these requirements are usually colliding [16, 120]. For example, if the payload is increased, the perceptibility of the watermark will be also increased or the robustness will decrease; similarly, for a given payload and imperceptibility constraints, the most robust schemes will be those using complex (so computationally expensive) channel-codes. Nevertheless, this is not always the case: if the attacker does not know the sent message and the payload is increased, the system will be generally more secure (see Chapter 4 for a further discussion about it).

Therefore, as a conclusion to this section we can say that the characteristics of the application will determine the balance among the general requirements which best suit our problem.

1.4. Data hiding terminology

In this section we will introduce some basic data hiding concepts which will help us to introduce the state-of-the-art schemes studied in this work.

1.4.1. Blind vs. Non-Blind Data Hiding

Some of the firstly proposed methods for digital data hiding were based on the availability of the original signal at the receiver; due to this characteristic, they are named *non-blind* algorithms. Obviously, they needed an extra storage capability which is not generally available, and implied a scalability and security problem (the database of original signals should be secure against attackers trying to access it).

Therefore, methods which did not required the original content, i.e. *blind* methods, are widely favored and studied in the literature, even when their performance is generally worse, since the original host signal is usually considered as an unknown interference by the decoder (although known by the embedder). Furthermore, the presence of the original signal could help in the design of countermeasures against geometric attacks.

1.4.2. Detection vs. Decoding

In most data hiding applications the output of the receiver is an estimate of the hidden message. Nevertheless, this is not always the case. For example, in ownership claim or authentication scenarios the output should be a measure of the certainty about the presence of the watermark, not an estimate of it. Therefore, this is a binary hypothesis problem (is the watermark present, or not?) usually named *detection problem*, as opposed to the aforementioned multiple hypothesis problem, i.e. when the hidden message has to be estimated, which is termed *decoding problem*. Following this nomenclature, the receiver is typically named *detector* or *decoder*, respectively. Due to their different nature, the measures used to quantify the goodness of data hiding schemes also differ for both scenarios.

In the detection problem, the figures of merit are the probability of false alarm P_{fa} and the probability of missed detection P_m, defined as the probability of deciding that the watermark is present when it is actually not, and the probability of deciding that the watermark is not present, when in fact it is, respectively; the joint representation of both quantities is the Receiver Operating Characteristic (ROC). Another figure of merit could be the Kullback-Leibler distance between the unwatermarked contents and the watermarked ones.

In the decoding problem the goodness of a scheme is measured by the probability of error, i.e. the probability of mistaking the sent symbol when estimating it at the decoder; when the symbols are binary the Bit Error Rate (BER) is defined. Some works in the literature, such as [115], focus on the probability of mistaking the sent message, defining the Message Error Rate (MER).

1.4.3. Spread-Spectrum vs. Side-Informed Data Hiding

The first attempts to consider the data hiding problem as a communications problem were based on the Spread-Spectrum principle, meaning the addition of a low-powered signal, termed *spreading sequence*, modulated by the information to be sent, in a number of features of the original content. Spread-Spectrum communication techniques cover a wide range of methods which are typically used in noisy channels; nevertheless, when talking about Spread-Spectrum in data hiding it is usually understood that the watermark does not try to cancel the interference due to the original host signal, but just survives it. Even when both additive and multiplicative Spread-Spectrum data hiding schemes were proposed in the literature (see for example [56] and [17]), in this work we will focus only on the additive ones (Add-SS). For Add-SS the decoder is usually based on the comparison of the correlation between the received signal and the spreading sequence with a threshold, although this strategy is optimal only in the Gaussian case (see [93] for a generalized version, based on the Generalized Gaussian distribution).

As it was previously explained, in state-of-the-art data hiding schemes, the receiver does not access the original non-watermarked content. If this lack of knowledge of the original is considered as noise, as it is the case for Spread-Spectrum methods, and given that the power of the watermark is significantly reduced compared with that of the original host, a huge host interference is produced, which leads to a very small capacity of the system, so spreading is necessary to tackle this interference.

Nevertheless, the complete original signal is known by the embedder before encoding the information and it knows how the decoding is going to be performed, so, *is it possible to take advantage of this knowledge by the encoder to improve the trasmission?*. This question was positively answered by Chen and Wornell [31, 32] and Cox et al. [57], stating that the data hiding problem can be interpreted as a communication with side information at the encoder; this paradigm is usually termed *side-informed data hiding*.

Chen and Wornell rescued an old result by Costa [50], who showed that for the case of i.i.d. Gaussian state channel (host signal) non-causally known by the encoder and unknown by the decoder, and i.i.d. Gaussian noise channel independent of the state channel, the capacity coincides with that obtained for the i.i.d. Gaussian noise channel without state channel (or equivalently, when this state channel is known by the decoder).

Despite of its evident importance, the main problem of Costa's scheme is that it is based on random coding, so its complexity makes its implementation infeasible; in fact, Chen and Wornell also proposed in their paper [32] the Distortion-Compensated Quantization Index Modulation (DC-QIM) method for canceling the host interference. The basic procedure of DC-QIM involves the quantization of a given host signal using a multidimensional quantizer selected from a finite set by the message to be embedded. A fundamental feature is that the watermarked signal is obtained by adding back to the quantized host signal the quantization error scaled depending on an optimizable parameter. This *distortion compensation* is what makes DC-QIM equivalent to Costa's scheme, as a proper choice of the parameter is known to yield the non-blind achievable rate under additive white Gaussian distortion independent of the host [32, 50]. Chen and Wornell also gave the first proposal to put DC-QIM into practice with Distortion-Compensated Dither Modulation (DC-DM), a particular case in which the set of quantizers are dithered (shifted) versions of a basic one. Due to the implementation and design issues associated to multidimensional quantizers, this basic quantizer usually relies on a lattice, which most of the times for practical implementations is the Cartesian product of scalar uniform quantizers. DC-DM based on uniform scalar quantization is straightforward to implement and more easily amenable to analysis than other more complex settings. A number of additional works have also aimed at building practical methods based on Costa's result. Among them we have the Scalar Costa Scheme (SCS) [64] and the Scaled Bin Encoding

(SBE) [106] —which are completely equivalent to DC-DM with uniform scalar quantizers—, the continuous periodic functions for self-noise suppression (CP-SNS) [139], and others. Due to their nature, these methods are also known as quantization-based and host-rejecting data hiding methods.

Finally, we would like to remark that in a recent paper Erez et al. [70] showed that the capacity of Additive White Gaussian Noise (AWGN) channels can be achieved using lattices and lattice encoding and decoding (as a less computationally demanding alternative to maximum likelihood decoding; read 2.5 for a further discussion on this topic) when the number of dimensions goes to infinity, independently of the distribution of the state channel, as far as the lattices verify some asymptotical conditions.

1.4.3.1. Intuitive insight

In this section we will try to provide an intuitive insight of the advantages of side-informed methods over those based on spread spectrum. Figure 1.1 plots the codewords related to different messages in the space of original and watermarked signals;[3] it shows one of the main differences between both of them: whereas spread spectrum just defines a codeword for each symbol, for side-informed methods a set of codewords is related to each symbol, in such a way that the embedder has more freedom to choose the codeword, equivalently the watermark, that better fits to the host signal.

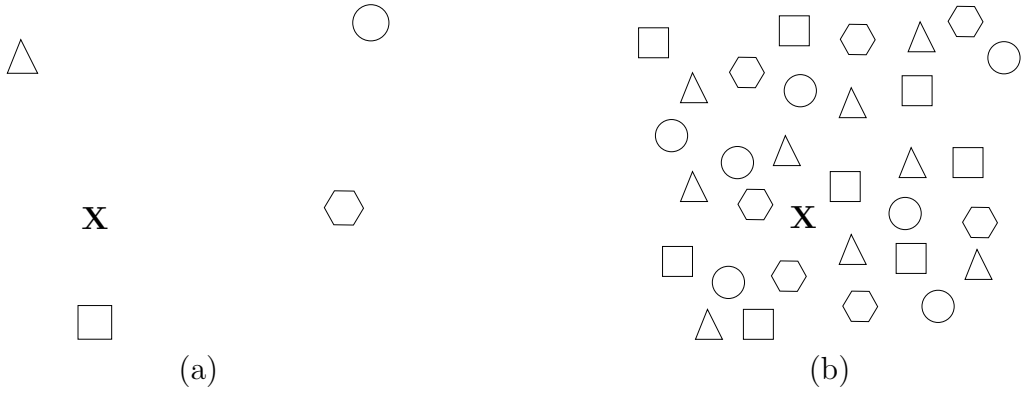


Figure 1.1: a) Spread-Spectrum data hiding codewords. There is just one codeword for each transmitted symbol (represented by different shapes). b) Side-Informed data hiding codewords. The embedder can choose the codeword used to convey a given symbol from a set of possible codewords (figures with the same shape). In both cases, the number of possible symbols is 4.

[3] Although the space of original host signals and the space of watermarked signals could differ, we will assume that they are the same. This assumption does imply any loss of generality, since a more general case including both spaces could be considered.

The advantages of side-informed methods are easily understood if one thinks of watermarking the original host signal $\mathbf{X}$ in Figure 1.1. For Spread-Spectrum the codewords (especially those represented by the circle and the hexagon) are really distant from the host signal. Given that the embedder *pushes* the host signal towards the desired codeword, and taking into account that the length of this push is upper-bounded due to the imperceptibility constraint, the watermarked signal will be still far from the codeword we want to transmit. In fact, it could happen that the watermarked signal were *closer* to another codeword than to that we wanted to transmit, yielding a decoding error even when the watermarked signal were not attacked; this effect is usually termed *host interference*. On the other hand, for side-informed methods the embedder can choose from the set of codewords related to the desired message that which is closer to the original host signal, so the imperceptibility constraint will not be typically a problem and the watermarked signal could be one of those codewords related to the desired message, i.e. a figure with any shape can be reached verifying imperceptibility constraint in Figure 1.1.[4] In this way the interference due to the original host signal, also known as *host interference*, is reduced. Summarizing, the host interference experienced by side-informed methods is lower than that corresponding to spread spectrum based ones, in such a way that the former techniques improve the performance of the latter.

Following the previous reasoning, one could think that the probability of mistaking the sent symbol of side-informed methods could be even larger than that of the spread-spectrum based ones for a certain level of channel noise, since the number of neighbors is increased. In this way, it was an old belief that scalar DC-DM with uniform quantizers (equivalent to SCS) could be outperformed by Add-SS when high levels of channel noise are considered [65]; nevertheless, it was recently showed by Pérez-Freire et al. that this is not the case when the DC-DM Maximum Likelihood decoder is taken into account [129].

1.4.4. Private vs. Public Key Data Hiding

Most data hiding schemes use a key to encode and decode the hidden information; usually this key is *private*,[5] i.e. only the embedder and decoder shares it, since its knowledge could make easier the design of attacks attempting to remove the watermark. Even when this problem could be avoided by the design of *public*

[4] As it will be discussed later, this strategy is not the optimal one in general, and the attacker could prefer to stay in the midway from the original host signal to the codeword in order to reduce the embedding distortion.

[5] Sometimes in the literature, private data hiding is more generally defined as those data hiding schemes where *"a mechanism is envisaged that makes it impossible for unauthorized people to extract the information hidden within the host signals"*, including non-blind techniques (extracted from [16]); nevertheless, in this case we preferred to follow the definition given in [89], to clearly distinguish private data hiding from non-blind data hiding.

data hiding schemes, where the decoding key is publicly available, these seem to be a utopia, and only some pioneering and simple works have been published in this field (see for example [89]).

1.4.5. Symmetric vs. Asymmetric Data Hiding

Most data hiding methods belong to the so-called category of *symmetric schemes*, i.e. those algorithms whose embedding and detection keys are the same; this means that disclosure of the detection key implies a total break of the system, allowing pirates to forge contents at their will. Moreover, the estimation of the detection key is possible by means of oracle-like attacks if a detector is available to the attacker, as it occurs in a great variety of data hiding applications. This is one of the reasons why researchers put their efforts on the design of *asymmetric* schemes, i.e. those schemes in which the embedding and detection keys do not need to be necessarily the same, in such a way that the impact of attacks revealing the detection key is reduced at a great extent. This and other advantages of asymmetric data hiding schemes are discussed in [78]: in this paper, four asymmetric watermarking methods are analyzed and unified in such a way that the detection function may always be written as a quadratic form. The main conclusion is that, although its robustness is worse than that of symmetric schemes, the security level can be improved because we are passing from linear detection functions to quadratic ones, thus increasing the complexity of oracle-like attacks (see [75] for a further discussion); this complexity may be progressively increased by the use of n-th order detection functions [77].

1.5. Outline

After presenting the data hiding problem, in the next chapter we will introduce the framework and state-of-the-art methods.

Robustness is the main topic of Chapter 3. There, the robustness of DC-DM against several kind of attacks is analyzed. A new version of the well-known sensitivy attack is introduced, showing its effectiveness not just for spread-spectrum methods, but also for a wide range of existing methods. Finally, a game-theoretic approach is proposed to analyze the performance of several methods when both the attacker and the decoder follow a smarter strategy than additive white Gaussian noise together with a minimum distance decoder; also related to this game-theoretic approach is the optimization of the worst case additive attack for scalar DC-DM, which concludes Chapter 3. The materials presented in this chapter can be found in [133], [46], [43], [40], [41], [42], [45], [44] and [130].

The other major subject of this thesis is security; to its analysis is completely devoted Chapter 4. After a brief overview of the historical evolution of watermarking security, some definitions and an information-theoretic measure of security are proposed. Using this measure, the security of spread spectrum and Costa's scheme is analyzed, and compared with that of DC-DM (which is not part of this thesis). The bulk of this chapter has been already published in [161], [39], [38] and [127].

A minor, but also interesting, part of this thesis is that exposed in Chapter 5. There, we delve in the way of achieving capacity in dirty paper schemes: it is not enough to just consider channel coding, but source coding must be also performed. A practical method (previously introduced in the literature) where this statement was considered in the design stage, was adapted in this thesis to a data hiding scenario, showing its gain over just channel-coding-based methods. The materials presented in this chapter are based on those in [47].

Although the previous chapters are mainly focused on theoretical questions, we have also paid attention to practical problems, and in Chapter 6 a watermarking application oriented to checking the integrity of the images of a video surveillance system is proposed.

Finally, in Chapter 7 we present the conclusions and future research lines of this thesis.

Chapter 2

Notation and Methods Description

2.1. Notation

We will denote scalar random variables with capital letters (e.g., X), and their outcomes with lowercase letters (e.g., x). The same notation criterion applies to random vectors and their outcomes, denoted in this case by bold letters (e.g., $\mathbf{X}$, $\mathbf{x}$). We assume without loss of generality that the host signal is modeled by a zero-mean random vector $\mathbf{X}^o = (X_1^o, \cdots, X_{L_1}^o)^T$. If necessary, these particulars can always be achieved by subtracting any non-zero mean from the host, and by using an arbitrary bijective transformation from the original arrangement of the host signal samples to a unidimensional one. Before embedding we apply a pseudorandom permutation $\Pi(\cdot)$ to $\mathbf{X}^o$; this permutation depends on a secret key $\mathbf{\Theta}$.[1] The permuted host $\mathbf{X} \triangleq \Pi(\mathbf{X}^o)$ is partitioned into L_b subvectors $\mathbf{X}_j \triangleq (X_{L_2\cdot(j-1)+1}, \cdots, X_{L_2\cdot(j-1)+L_2})^T$, for $j = 1, \cdots, L_b$, and assuming for notational simplicity that $L_2 \triangleq L_1/L_b$ is integer. Apart from the security increase due to the uncertainty that this permutation procedure causes to an attacker unaware of the key, an important side advantage is that of facilitating the analysis. This is due to the fact that the pseudorandom selection of the elements in each subvector $\mathbf{X}_j$ approximately grants their statistical independence. This hypothesis of approximate independence usually holds true for natural signals, as long as L_2 is not of the same order as L_1.

The watermarked signal $\mathbf{Y}$ will be obtained from the host signal $\mathbf{X}$, the information message $\mathbf{b}$ to be conveyed and the secret key $\mathbf{\Theta}$. We will assume, without

[1] As it was previously explained in Section 1.4.5 the embedding and detection keys could be different, namely $\mathbf{\Theta}_e$ and $\mathbf{\Theta}_d$, for the asymmetric methods; nevertheless, given that most of the data hiding methods considered in this work are symmetric, and due to notational simplicity, we have decided to drop the secret key subindex. It will be explicitly written only if it is necessary.

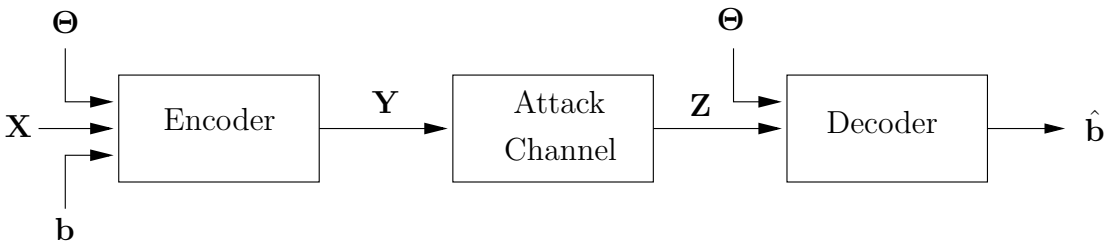


Figure 2.1: Data hiding scheme.

loss of generality, that $\mathbf{b} = (b_1, \cdots, b_{L_b})^T$ is a P-ary vector, with b_j taking values uniformly in $\{0, \cdots, P-1\}$ for $j = 1, \cdots, L_b$. A particular symbol b_j will be embedded using the subvector $\mathbf{X}_j$ to get $\mathbf{Y}_j$, i.e. $\mathbf{Y}_j = \mathbf{X}_j + \mathbf{W}_j$, where $\mathbf{W}_j$ is the watermark.

The imperceptibility of the differences between $\mathbf{X}$ and $\mathbf{Y}$ has to be guaranteed by means of a perceptual analysis of the host signal previous to the embedding operation. This procedure is intrinsically dependent on the type of host signal in question. Due to this fact, we will consider henceforth that the host is a multimedia signal given in a certain domain of interest. The only requirement is that the domain chosen is suited to compute a *perceptual mask* $\boldsymbol{\gamma}(\mathbf{X})$, taking into account human perceptual features. We assume in the following that the maximum energy for an unnoticeable modification of the corresponding host signal sample X_i is proportional to γ_i^2.

Decoding is accomplished by the receiver after the watermarked signal $\mathbf{Y}$ has undergone an attack channel, denoted by the addition of the vector $\mathbf{N}$, so the received signal will be given by $\mathbf{Z} = \mathbf{Y} + \mathbf{N}$. The attack channel is sometimes modeled by the transition probabilites $A(\mathbf{z}|\mathbf{y})$ [122]. Since the watermark depends on the perceptual mask, the decoding operation will typically need such mask. Nevertheless, in most cases the original host signal is not available at the decoder, so an estimate of $\boldsymbol{\gamma}(\mathbf{X})$ has to be performed, based on the received signal $\mathbf{Z}$, i.e. $\boldsymbol{\gamma}(\mathbf{Z}) \approx \boldsymbol{\gamma}(\mathbf{X})$. Generally, this estimate is quite good, since the received signal should be perceptually indistinguishable from the original one.

2.2. Distortion measures

Even when the perceptual mask is used, a distortion measure must be defined between the instance of the original host signal $\mathbf{x}$ and the corresponding watermarked signal $\mathbf{y}$, i.e., $\text{dist}(\mathbf{x}, \mathbf{y})$, to compare the distortion introduced by different schemes; this measure should be based on perceptual criteria (as those proposed in [160]), but most of times more manageable, although less perceptually suited, functions are used, such as the Mean Square Error (MSE). Furthermore, as all

subvectors are obtained in the same way, notice that we will only need to focus our attention on one arbitrary subvector for analytical purposes. In particular, note that the average host signal power in each partition will tend to be approximately constant as L_2 increases. Denoting this value as D_h, and using the intra-partition independence assumption, we can write

$$D_h \approx D_h^{(j)} = \frac{1}{L_2} \sum_{i=(j-1)\cdot L_2+1}^{j\cdot L_2} \sigma_{X_i}^2, \quad j = 1, \cdots, L_b, \tag{2.1}$$

where $\sigma_{X_i}^2 \triangleq \mathrm{Var}\{X_i\}$ and $D_h^{(j)}$ denotes the average host signal power in the j-th partition; similarly, for the watermark signal power

$$D_w \approx D_w^{(j)} = \frac{1}{L_2} \sum_{i=(j-1)\cdot L_2+1}^{j\cdot L_2} \sigma_{W_i}^2, \quad j = 1, \cdots, L_b. \tag{2.2}$$

In order to quantify the distortion introduced by the channel a distortion measure must be defined between $\mathbf{y}$ and $\mathbf{z}$, i.e., $\mathrm{dist}(\mathbf{y}, \mathbf{z})$. Following the aforementioned reasons a perceptual measure should be used, but most times MSE is preferred for the sake of simplicity. In this way, recalling that the elements of the L_2-length subvectors are pseudorandomly chosen through the permutation $\Pi(\cdot)$, we may also assume that the samples in $\mathbf{N}_j$ are mutually independent, with diagonal covariance matrix $\mathbf{\Gamma}_j = \mathrm{diag}(\sigma^2_{N_{(j-1)\cdot L_2+1}}, \cdots, \sigma^2_{N_{(j-1)\cdot L_2+L_2}}), \quad j = 1, \cdots, L_b$, so the *channel distortion* D_c can be then defined in a similar fashion as the embedding distortion, i.e.,

$$D_c \approx D_c^{(j)} = \frac{1}{L_2} \sum_{i=(j-1)\cdot L_2+1}^{j\cdot L_2} \sigma_{N_i}^2, \quad j = 1, \cdots, L_b. \tag{2.3}$$

Although the previous distortion measure quantifies the differences between $\mathbf{y}$ and $\mathbf{x}$, in the literature there are also examples of measures computing the distortion between $\mathbf{x}$ and $\mathbf{z}$ (see, for instance [64, 118]). The former approach was chosen for trying to distinguish more clearly the embedding distortion from that due to the attack; furthermore, the attacker does not know the original signal, so he/she is limited to use this kind of measures (between $\mathbf{y}$ and $\mathbf{z}$) when computing the distortion due to his/her attack.

Other MSE-based measure frequently used for images is the *Peak Signal to Noise Ratio* (PSNR), which compares signals $\mathbf{x}$ and $\mathbf{y}$ through

$$\mathrm{PSNR}(\mathbf{x}, \mathbf{y}) \triangleq 10 \log_{10} \left(\frac{L_1 \cdot \mathrm{Lum}_{\max}^2}{\sum_{i=1}^{L_1} (x_i - y_i)^2} \right), \tag{2.4}$$

where both $\mathbf{x}$ and $\mathbf{y}$ are given in the spatial domain, and $\mathrm{Lum}_{\max}$ is the maximum value achievable by x_i for $1 \leq i \leq L_1$, corresponding to white (in an 8-bit

representation, $\text{Lum}_{\max} = 255$). The PSNR could also be applied to measure the differences between $\mathbf{y}$ and $\mathbf{z}$ (PSNR($\mathbf{y}, \mathbf{z}$)), or $\mathbf{x}$ and $\mathbf{z}$ (PSNR($\mathbf{x}, \mathbf{z}$)). PSNR($\mathbf{x}, \mathbf{y}$) is frequently used in image watermarking applications, and for large values of L_2 it can be straightforwardly translated to D_w (and PSNR($\mathbf{y}, \mathbf{z}$) to D_c).

Last, we will find it useful to introduce the *watermark-to-noise ratio* as $\text{WNR} \triangleq \frac{D_w}{D_c}$, that relates the power of the embedding and channel distortion, establishing a working point similar to the signal-to-noise ratio (SNR) in communications, and similarly the *document-to-watermark ratio* as $\text{DWR} \triangleq \frac{D_h}{D_w}$.

2.2.1. On the MSE measures and perceptual masks

The distortion measures based on the mean square error (MSE) have some criticizable points; for example, this kind of measurement would in principle allow to either concentrate all the attacking distortion on a single sample of $\mathbf{Y}$ or spread it all over the vector. In fact, many authors advocate the use of more realistic, perceptual-based measures, for instance [80, 26, 162, 159].

In this section we will try to illustrate these problems by comparing different distorted versions of the image *Lena*, all of them with the same distortion power but different perceptual effect. In Figure 2.2 we can see the original image, a version distorted with noise shaped by a mask computed in the Discrete Cosine Transform (DCT) domain (based on [160] and [6]), another version distorted with noise shaped by a mask computed in the spatial domain (computed based on the gradient of the image [107]), and a freehand modified version. Since the last 3 verify $10 \log_{10}(\frac{||\mathbf{X}||^2}{||\mathbf{y}-\mathbf{X}||^2}) = 19.11$ dB, if the MSE were a good perceptual distortion measure, the modification should be similar for all them; nevertheless, their perceptual distortions are completely different, indicating that MSE measures are not suitable for setting up distortion embedding constraints. On the other hand, the result of computing the perceptual distortion measure proposed by Watson in [160] between the original host signal and the distorted ones is quite more meaningful: when the version distorted with noise shaped by the DCT mask is taken into account the result is 7.05, but when considering the spatial mask this increases to 30.46; finally, with the free hand modified version the distortion rises up to 60.20, showing the perceptual modifications.

In any case, it should be remarked that the Watson perceptual distortion measure is computed in the DCT domain. In that sense, the first distorted image (where the noise is shaped by a mask computed also in the DCT domain with a technique similar to that used by the measure) somehow has an *advantage*, as it produces a lower value of this perceptual distortion, over the version with spatially masked noise. This peculiarity is due to the fact of being based both Watson measure and the noise of the first distorted image on the same masking effects, while different masking criteria are being considered in the computation

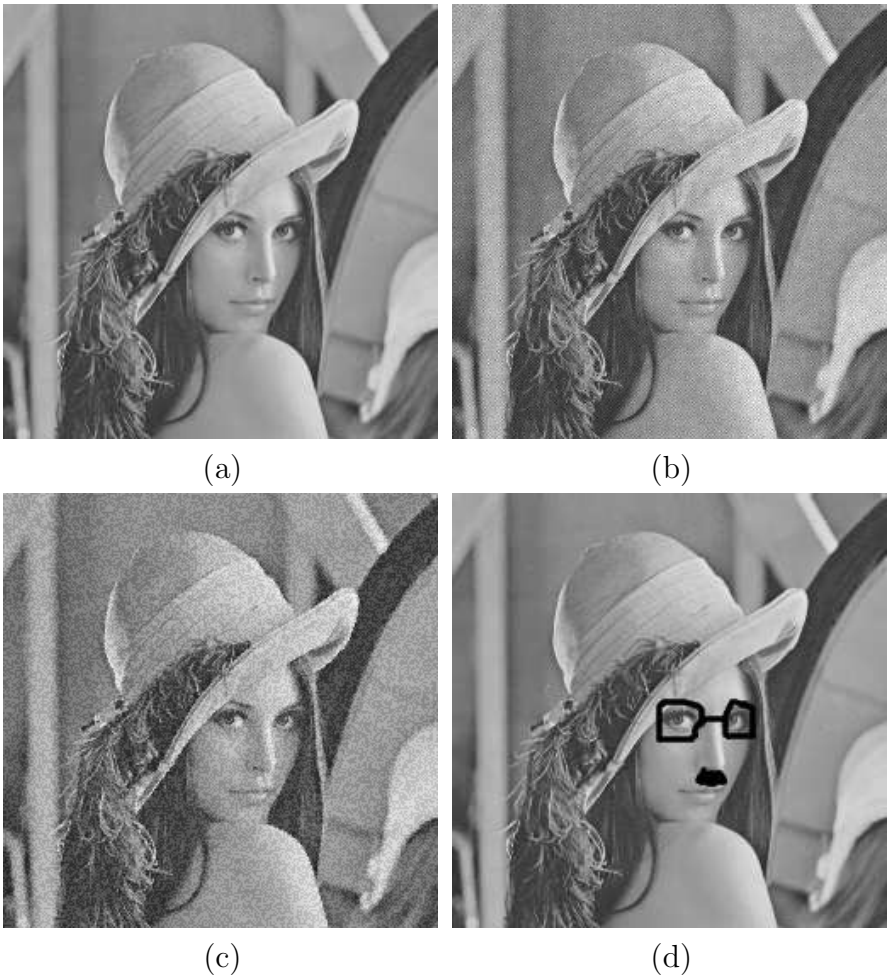
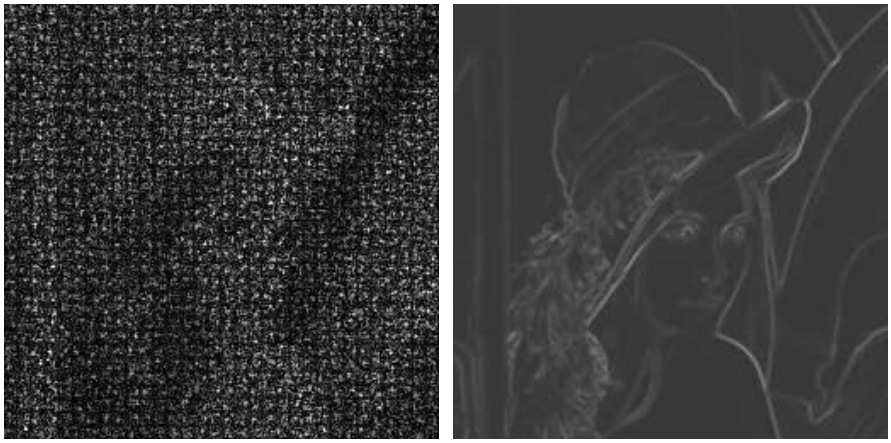


Figure 2.2: (a) Original image. (b) Image distorted by noise shaped with the DCT mask. (c) Image distorted by noise shaped with the spatial mask. (d) Image freehand distorted. In the last 3 cases $10\log_{10}(\frac{||\mathbf{x}||^2}{||\mathbf{y}-\mathbf{x}||^2}) = 19.11$ dB.

of the spatially masked noise. In order to clarify this last point, in Figure 2.3 enhanced versions of the masks are plotted for both cases, showing the differences between them. Finally, we would like to remark that the perceptual analysis of multimedia contents is still an open question, and considerable work remains to be done.

Despite the reasons exposed in this section, we will undertake all subsequent analyses using MSE, as this criterion has been the most employed in the literature so far for the sake of tractability and comparisons. In any case, both the embedder and attacker may try to partially relieve the intrinsic inconveniences of MSE in order to comply with the usual requirement of minimal perceptual impact. Assuming the adequacy of the perceptual mask, it is clear that one way to meet this condition is to *perceptually shape* the noise signal, such that its variance at



(a) (b)

Figure 2.3: Perceptual masks of Lena. (a) Computed in the DCT domain. (b) Computed in the spatial domain.

each dimension is proportional to the corresponding allowable perceptual energy.

2.3. Additive Spread-Spectrum Embedding and Decoding

In Add-SS the watermark does not try to cancel the interference due to the host signal, but it is just based on the addition of a pseudorandom spreading sequence to the original host signal. Most of times, the symbols $b_j, j = 1, \cdots, L_b$ belong to a binary alphabet, which is mapped to $\{-1, +1\}$, in such a way that

$$w_i = -(-1)^{b_j}\gamma_i s_i, \text{ for } i = L_2 \cdot (j-1) + 1, \cdots, L_2 \cdot j, \text{ and } j = 1, \cdots, L_b, \quad (2.5)$$

where $\mathbf{s}$ is the spreading sequence, which depends on the secret key $\boldsymbol{\Theta}$, and $\boldsymbol{\gamma}$ a non-negative mask. The decoding is usually based on the comparison with a threshold of the cross-correlation of the received signal $\mathbf{Z}_j$ and the corresponding spreading subsequence $\mathbf{s}_j$, i.e.

$$\hat{B}_j = \begin{cases} 0, & \text{if } \mathbf{s}_j^T \cdot \mathbf{Z}_j = \sum_{i=L_2\cdot(j-1)+1}^{L_2 \cdot j} s_i \cdot Z_i < 0 \\ 1, & \text{otherwise} \end{cases}, \quad (2.6)$$

where $\mathbf{s}_j^T$ denotes the transpose of $\mathbf{s}_j$. Be aware that this decoding strategy, which is based on the minimum distance criterion, is the optimal one only if both the host signal and the channel noise are Gaussian and independent and identically distributed (i.i.d.), $|s_i| = |s_k|$, and $\gamma_i = \gamma_k$, for all $i, k \in \{L_2 \cdot (j-1)+1, \cdots, L_2 \cdot j\}$,

with $j = 1, \cdots, L_b$. Following, the probability of mistaking the j-th bit is given by

$$P(\hat{B}_j \neq B_j) = \mathcal{Q}\left(\frac{\sum_{i=L_2\cdot(j-1)+1}^{L_2\cdot j} \gamma_i \cdot s_i^2}{\sqrt{\sum_{i=L_2\cdot(j-1)+1}^{L_2\cdot j}(\sigma_{X_i}^2 + \sigma_{N_i}^2)s_i^2}}\right) \tag{2.7}$$

where $\mathcal{Q}(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{\tau^2}{2}} d\tau$.

In order to reduce the loss in performance due to the host signal, a number of techniques were proposed aimed at reducing the interference due to the host signal by using linear filtering. For example, in the spatial domain case the received signal $\mathbf{z}$ could undergo a linear filtering operation as a way of reducing the host-interference power at the decoder. This can be represented by means of a space-varying, noise-independent filtering. Wiener filtering [92] is included in this category, since the host signal power usually is much greater than the noise power (at least if the attacked signal is to remain valuable); therefore, Wiener filter's coefficients will not be modified in a significant way by the addition of noise. We can represent this situation by a $L_1 \times L_1$ matrix that will be denoted by $\mathbf{H}$, so that the filtered host image would become $\mathbf{x}_f \triangleq \mathbf{Hx}$.

The performance achieved by the decoder described in (2.6) can be improved by considering some weights $\beta_i, i \in \{1, \cdots, L_1\}$, yielding

$$\hat{b}_j = \begin{cases} 0, & \text{if } \sum_{i=L_2\cdot(j-1)+1}^{L_2\cdot j} \beta_i \cdot s_i \cdot Z_i < 0 \\ 1, & \text{otherwise} \end{cases}. \tag{2.8}$$

The derivation of the optimal value of $\boldsymbol{\beta}$ will be given in Section 3.5.3.

On the other hand, for the detection problem the watermarked signal is written as

$$Y_i = X_i + \gamma_i s_i, \quad \text{for } i = 1, \cdots, L_1, \tag{2.9}$$

so the minimum distance based detection function (similarly to that shown in (2.6)), is now given by

$$l(\mathbf{Z}) = \sum_{i=1}^{L_1} s_i \cdot Z_i, \tag{2.10}$$

in such a way that a detector will determine that the watermark is present whenever $l(\mathbf{Z}) > \eta$, with η a threshold which will be fixed as a function of the false alarm or missed detection probabilities. Again, the minimum distance criterion will be the optimal one only if both the host signal and the channel noise are i.i.d. Gaussian distributed and $\gamma_i = \gamma_k$, for all $i, k \in \{1, \cdots, L_1\}$.

Nevertheless, some works in the literature modelled the host signals as following a given probability distribution in a certain domain, in such a way that

the performance of the system can be improved by optimizing the decoder for that distribution. This is the case of [93] and [17], where the host signal in the block Discrete Cosine Transform (DCT) domain and full-frame Discrete Fourier Transform (DFT) domain is modelled by a Generalized Gaussian and Weibull distribution respectively. The optimal detector in the first case, which was obtained in [93] and will be used in Section 3.4, is given by

$$l(\mathbf{Z}) = \sum_{i=1}^{L_1} \beta_i^{c_i} \left(|z_i|^{c_i} - |z_i - \gamma_i s_i|^{c_i} \right), \tag{2.11}$$

where c_i is the shape parameter of the Generalized Gaussian distribution corresponding to the i-th coefficient, $\beta_i = \frac{1}{\sigma_{X_i}}(\Gamma(3/c_i)/\Gamma(1/c_i))^{1/2}$ (both of them are estimated following a Maximum Likelihood (ML) criterion), and γ_i is the watermark strength for the aforementioned coefficient.

2.4. Basic concepts about lattices

In the next section the embedding and decoding of DC-DM based on lattices [49] will be studied. This calls for some basic concepts about these mathematical structures which are introduced next.

A lattice Λ is a discrete subgroup of the Euclidean space $\mathbb{R}^n$ determined by a set of basis vectors

$$\begin{aligned} \mathbf{v}_1 &= (v_{11}, v_{12}, \cdots, v_{1n}), \\ \mathbf{v}_2 &= (v_{21}, v_{22}, \cdots, v_{2n}), \\ &\vdots \\ \mathbf{v}_n &= (v_{n1}, v_{n2}, \cdots, v_{nn}), \end{aligned} \tag{2.12}$$

that can be arrayed in the so-called *generating matrix* $\mathbf{M}$,[2]

$$\mathbf{M} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix}. \tag{2.13}$$

[2] Although in this work we will assume that the number of vectors $\mathbf{v}_i$ will be n, examples are available in the literature where the number of composing vectors m is lower than n (see [49]). Nevertheless, the latter can be just considered as a particular case of our approach, where some components of $n - m$ vectors are set to infinity, i.e., just a degenerated version of the proposed approach.

Using this matrix, the lattice Λ is defined as

$$\Lambda \triangleq \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \boldsymbol{\xi}^T \mathbf{M}, \quad \text{for any } \boldsymbol{\xi} \in \mathbb{Z}^n\}, \tag{2.14}$$

and its Gram matrix as $\mathbf{A} \triangleq \mathbf{M}\mathbf{M}^T$, where T denotes transpose, verifying $\det(\Lambda) \triangleq \det(\mathbf{A})$. At sight of (2.13) and (2.14), one may realize that any linear code can be thought of as a lattice code.

A coset of Λ in $\mathbb{R}^n$ is any translated version of it, i.e., the set $\mathbf{x} + \Lambda$ is a coset of Λ for any $\mathbf{x} \in \mathbb{R}^n$ [70], where $\mathbf{x}$ is called *coset representative*. The fundamental Voronoi region of $\Lambda \subset \mathbb{R}^n$, denoted by $\mathcal{V}(\Lambda)$, is a set of minimum Euclidean norm coset representatives of the cosets of Λ, in such a way that every $\mathbf{x} \in \mathbb{R}^n$ can be uniquely written as

$$\mathbf{x} = Q_\Lambda(\mathbf{x}) + (\mathbf{x} \mod \Lambda), \tag{2.15}$$

where

$$Q_\Lambda(\mathbf{x}) \triangleq \arg\min_{\lambda \in \Lambda} ||\mathbf{x} - \lambda||, \tag{2.16}$$

$Q_\Lambda(\cdot)$ is the minimum Euclidean distance quantizer, and

$$\mathbf{x} \mod \Lambda \triangleq \mathbf{x} - Q_\Lambda(\mathbf{x}) \in \mathcal{V}(\Lambda), \tag{2.17}$$

is the modulo-Λ reduced version of $\mathbf{x}$;[3] therefore we can write $\mathcal{V}(\Lambda) \triangleq \mathbb{R}^n \mod \Lambda$. Taking into account the symmetries introduced by (2.14), is straightforward to see that $Q_\Lambda(-\mathbf{x}) = -Q_\Lambda(\mathbf{x})$ and $(-\mathbf{x}) \mod \Lambda = -[\mathbf{x} \mod \Lambda]$.

Some figures of merit of lattices are listed below:

- The volume $V(\Lambda)$

$$V(\Lambda) \triangleq \int_{\mathcal{V}(\Lambda)} d\mathbf{x}. \tag{2.18}$$

- The second moment per dimension $\sigma^2(\Lambda)$

$$\sigma^2(\Lambda) \triangleq \frac{1}{n} \frac{\int_{\mathcal{V}(\Lambda)} ||\mathbf{x}||^2 d\mathbf{x}}{V(\Lambda)}. \tag{2.19}$$

- The normalized second moment $G(\Lambda)$

$$G(\Lambda) \triangleq \frac{\sigma^2(\Lambda)}{V(\Lambda)^{2/n}} = \frac{1}{n} \frac{\int_{\mathcal{V}(\Lambda)} ||\mathbf{x}||^2 d\mathbf{x}}{V(\Lambda)^{1+2/n}}. \tag{2.20}$$

[3] Be aware that the quantization region boundaries, which are produced by the ties in (2.16), must be arbitrarily but systematically broken.

The normalized second moment $G(\Lambda)$, which is invariant to scaling and rotation of the lattice, is always greater than $\frac{1}{2\pi e}$, the normalized second moment of an infinite-dimensional sphere. Taking into account that $G(\Lambda)$ takes value $1/12$ when $\Lambda = \mathbb{Z}^n$, i.e. when the Voronoi region is a n dimensional hypercube, the *shaping gain* is defined as $g_s(\Lambda) = 10\log_{10}(1/(12 \cdot G(\Lambda)))$, whose maximum value is 1.53 dB.

It was shown in [165] that, as the dimension increases, lattices exist whose normalized second moment goes to $\frac{1}{2\pi e}$, i.e. their Voronoi region $\mathcal{V}(\Lambda)$ approaches a sphere. This result was used in [70] to show that the capacity of the Additive White Gaussian Noise (AWGN) channel can be achieved by lattice-based schemes, even for an arbitrarily distributed state channel, which is known to the encoder but not to the decoder.

2.5. Distortion-Compensated Dither Modulation Embedding and Decoding

We describe next the implementation of DC-DM based on lattices, generalizing Chen and Wornell's proposal [32] to account for perceptual constraints as done in [132]. We restrict our presentation to any of the L_2-dimensional subvectors inside which the host signal samples can be assumed independent, dropping the subindex j in the sequel for notational simplicity. Let us assume that the information symbol b is hidden using DC-DM inside the host $\mathbf{X}$. Then, we denote by

$$\mathbf{E} \triangleq Q_b(\mathbf{X}) - \mathbf{X}, \tag{2.21}$$

the quantization error resulting from quantizing $\mathbf{X}$ with the quantizer $Q_b(\cdot)$ corresponding to the b-th symbol, which is based on a minimum Euclidean distance criterion. The watermarked signal $\mathbf{Y}$ is then obtained as

$$\mathbf{Y} = \mathbf{X} + \alpha\mathbf{E} = Q_b(\mathbf{X}) - (1-\alpha)\mathbf{E}, \tag{2.22}$$

The distortion-compensation parameter α, $0 < \alpha \leq 1$, is an optimizable variable akin to the one in Costa's paper. The component $(1-\alpha)\mathbf{E}$ may be termed as self-noise, since it is caused by the watermarking process itself due to the distortion compensation. As we will see in Section 3.1.2.1, performance improvements are obtained by using $\alpha < 1$, i.e., allowing a certain degree of self-noise.

Dither modulation means that all the quantizers $Q_b(\cdot)$ are just shifted versions of a basic lattice quantizer $Q_\Lambda(\cdot)$. The offset for obtaining each of these quantizers is a dither vector $\mathbf{v}(b)$ that depends on both the secret key $\mathbf{\Theta}$ and the message to be sent b. Then, the quantizer $Q_b(\cdot)$ can be put as

$$Q_b(\mathbf{X}) = Q_\Lambda\left(\mathbf{X} - \mathbf{v}(b)\right) + \mathbf{v}(b). \tag{2.23}$$

When the minimum distance among the shifted versions of the basic lattice is maximized, the basic lattice can be seen as a shifted sublattice of the union of its shifted versions, yielding a nested lattices structure.

Without loss of generality, and assuming hereafter that all embedded symbols are equally likely, we will focus our analysis on any given symbol b. The optimal decoding criterion that minimizes the information symbol error rate is the Maximum Likelihood (ML) decision given by

$$\hat{b} = \underset{b\in\{0,\cdots,P-1\}}{\arg\max} f_{\mathbf{Z}|B}(\mathbf{z}|b), \tag{2.24}$$

where $f_{\mathbf{Z}|B}(\cdot|\cdot)$ denotes the pdf of the received signal $\mathbf{Z}$, when the transmitted symbol is B. The main drawback of this approach is that it requires prior knowledge about the host signal probability density function (pdf). Also, the ML approach to DC-DM decoding can be too costly since we have to take into account f_{X_i}, the sent symbol and the dither in order to compute f_{Y_i}. Therefore, simplifications to it are desirable. One such simplification is lattice decoding. Lattice decoding rules can be seen as operating over variables that are reduced modulo-Λ, in such a way that the best lattice decoding strategy, usually termed *ML lattice decoder* or also *noise-matched lattice decoder* [70], can be written as

$$\hat{b} = \underset{b\in\{0,\cdots,P-1\}}{\arg\max} f_{\mathbf{Z}_{\text{mod}}|B}(\mathbf{z}_{\text{mod}}|b), \tag{2.25}$$

where $\mathbf{Z}_{\text{mod}} = \mathbf{Z} \mod \Lambda$ is the modulo-reduced version of the received signal $\mathbf{Z}$. Nevertheless, in general this strategy still needs knowledge about the channel noise and host signal distributions, so even more simplified decoding strategies are used; in fact, probably the most used strategy is the so-called *Euclidean Lattice decoding* or *Minimum (Euclidean) Distance Lattice Decoding* [70], which is given by

$$\hat{b} = \underset{b\in\{0,\cdots,P-1\}}{\arg\min} ||\mathbf{Z} - Q_b(\mathbf{Z})|| = \underset{b\in\{0,\cdots,P-1\}}{\arg\min} ||\mathbf{Z}_{\text{mod}} - Q_b(\mathbf{Z}_{\text{mod}})||. \tag{2.26}$$

We would like to remark that both (2.25) and (2.26), are not optimal in general, since the modulo reduction operation is obviously information-lossy. Nevertheless, Erez and Zamir showed in [70] that (2.25) and even (2.26) can approach the capacity of a power constrained AWGN channel, as long as Λ verifies some constraints.

In any case, (2.26) is the strategy extensively used in the literature due to its simplicity (see for example [32, 64, 132], and [20] for an unidimensional projected version).

2.6. DC-DM with Uniform Scalar Quantizers and Repetition Coding

As mentioned above, the simplest and more widespread implementation of DC-DM is the one by means of uniform scalar quantizers [32, 137, 19, 64, 33]. In this case $Q_\Lambda(\cdot)$ may be defined as the quantizer whose quantization centroids are given by the points in the lattice $\Lambda = P\Lambda'$, with $\Lambda' \triangleq (\Delta_1\mathbb{Z}, \Delta_2\mathbb{Z}, \cdots, \Delta_{L_2}\mathbb{Z})^T$. We will impose the criterion that the dither vectors $\mathbf{v}(b)$ are such that the distance between the closest centroids of the quantizers corresponding to any two different symbols is maximized. This just means that, for instance, $\mathbf{v}(b) = b\cdot(\Delta_1, \cdots, \Delta_{L_2})^T + \mathbf{d}$, where $\mathbf{d}$ is a key-dependent vector deterministically known to both encoder and decoder. This strategy increases the robustness of the embedding by placing the centroids as far away as possible. Also, the resulting symmetry allows to assume an arbitrary embedded symbol b for the analysis, as we will see later.

Notice that, for $L_2 > 1$, this particular choice of the dither vectors amounts to using a repetition code. It is well known that, even though it is useful in many practical situations (e.g., see [137, 19, 33]), this channel coding strategy is not the optimal one. It is pertinent to note that an empirical study on the concatenation of repetition coding for SCS (DC-DM) with near-optimal turbo codes was given in [64]. From the results in that work, it is possible to conclude that the concatenation of turbo codes and repetition is quite close to the capacity limit for Gaussian channels at low embedding rates. Then, the appeal of this scheme lies in the fact that it presents evident advantages from the complexity point of view yet keeping quite a good performance. This result adds an interesting practical perspective to the analysis of DC-DM with repetition coding.

In order to keep the exposition simple we will only study the case $P = 2$ (i.e., binary), but the approach we will follow can be extended for arbitrary alphabet sizes. For the binary case, the quantization centroids for $Q_b(\cdot)$ will be given by the shifted lattice $\Lambda_b = 2\Lambda' + b\cdot(\Delta_1, \cdots, \Delta_{L_2})^T + \mathbf{d}$, for $b \in \{0, 1\}$.

The use of scalar lattices inherently introduces an amplitude-limited embedding distortion. Since we can write (2.21) as $\mathbf{E} = (\mathbf{v}(b) - \mathbf{X}) \mod \Lambda$, it follows that $\mathbf{E}$ will be uniformly distributed over $\mathcal{V}(\Lambda)$ when $\mathbf{v}(b)$ is a deterministic vector if and only if $\mathbf{X} \mod \Lambda$ satisfies the same condition (uniform condition). For typical continuous distributions this will be the case if $\sigma_{X_i} \gg \Delta_i$ for all i. Due to perceptual constraints, for most watermarking scenarios the uniform condition will approximately hold, and hence we will assume hereafter that $\sigma_{X_i} \gg \Delta_i$ for all i, and so that $E_i \sim U(-\Delta_i, \Delta_i]$.

Noticing that the watermark signal is given by $\mathbf{W} = \mathbf{Y} - \mathbf{X} = \alpha\mathbf{E}$, it is clear that its energy per dimension will be $\sigma^2_{W_i} = \mathrm{E}\{W_i^2\} = \alpha^2\Delta_i^2/3$. According

to the perceptual mask assumed in Section 2.1, we can achieve the maximum unnoticeable embedding distortion by choosing Δ_i to be proportional to γ_i.

According to the preceding exposition, the samples in $\mathbf{Z}$ can be assumed to be mutually independent, so we can expand (2.24) as

$$\begin{aligned}\hat{b} &= \arg\max_{b\in\{0,1\}} \prod_{i=1}^{L_2} f_{Z_i|B}(z_i|b) \\ &= \arg\max_{b\in\{0,1\}} \prod_{i=1}^{L_2} \int_{-\infty}^{\infty} f_{Y_i|B}(z_i - r_i|b) f_{N_i}(r_i) dr_i,\end{aligned}$$

where $f_{Y_i}(\cdot)$ and $f_{N_i}(\cdot)$ are the probability density functions (pdf) of the independent random variables Y_i and N_i, respectively. But, as it was said in the previous section, lattice decoding strategies are extensively used for the sake of simplicity. In our case, the decision will be based on the statistics $\tilde{z}_i \triangleq \frac{z_i - Q_0(z_i)}{\Delta_i}$, $i = 1, \cdots, L_2$, where $Q_0(z_i)$ is the i-th component of $Q_0(\mathbf{z})$. From the way it is constructed, it is clear that $\tilde{z}_i \in (-1, 1]$; this leads to considering modulo-$2\mathbb{Z}^{L_2}$ vector reductions, for which the result belongs to $(-1, 1]^{L_2}$. Also, the normalization in the definition of $\tilde{z}_i$ is reasonable if we assume that the channel noise is perceptually shaped, as in that case its variance will be roughly proportional to Δ_i^2.

Let $f_{\tilde{\mathbf{Z}}}(\tilde{\mathbf{z}})$ denote the pdf of $\tilde{\mathbf{Z}}$. Then, the ML lattice decoder will choose $\hat{b}$ according to the rule

$$\hat{b} = \arg\max_{b\in\{0,1\}} f_{\tilde{\mathbf{Z}}|B}(\tilde{\mathbf{z}}|b). \tag{2.27}$$

As we shall see in Section 3.1.1 a performance analysis requires to determine the distribution of the noise in the decision statistics, which we tackle next. Let us define the *total noise* (also known as *modulo-lattice equivalent noise*) random variable as

$$T_i \triangleq \frac{-(1-\alpha)E_i + N_i}{\Delta_i}. \tag{2.28}$$

Recalling that if X, Y are two random variables related by $Y = aX$, their pdfs satisfy $f_Y(y) = |a|^{-1} f_X(y/a)$, and that the pdf of the sum of two independent random variables is the convolution of the respective pdfs, we can write

$$f_{T_i}(t_i) = \frac{\Delta_i^2}{(1-\alpha)} \left[f_{N_i}(t_i\Delta_i) * f_{E_i}(t_i\Delta_i/(1-\alpha)) \right], \tag{2.29}$$

where $*$ denotes convolution, and $E_i \sim U(-\Delta_i, \Delta_i]$, for all $i = 1, \cdots, L_2$. Now, the *modular total noise* random variable $\mathbf{U}$ is simply defined as $\mathbf{U} \triangleq \mathbf{T}$

mod $2\mathbb{Z}^{L_2}$.[4] Consequently, the support of U_i will be the interval $(-1, 1]$, for all $i = 1, \cdots, L_2$.

Considering (2.28), the pdf of U_i can be written as

$$f_{U_i}(u_i) = \begin{cases} \sum_{l=-\infty}^{\infty} f_{T_i}(u_i - 2l), & \text{if } u_i \in (-1, +1] \\ 0, & \text{otherwise} \end{cases} . \tag{2.30}$$

Alternatively, $f_{U_i}(u_i)$ can be written as

$$f_{U_i}(u_i) = \frac{\Delta_i^2}{(1-\alpha)} \left[f_{N_i}(u_i \Delta_i) \circledast_2 f_{E_i}(u_i \Delta_i / (1-\alpha)) \right], \tag{2.31}$$

with $\circledast_2$ the circular convolution over $(-1, 1]$, which includes the aliasing effect evident in (2.30). For any two arbitrary pdfs $f_B(x)$ and $f_C(x)$ this operation is defined as

$$f_B(x) \circledast_2 f_C(x) \triangleq \begin{cases} \sum_{l=-\infty}^{\infty} \int_{-\infty}^{\infty} f_B(y - 2l) f_C(x - y) dy, & -1 < x \leq +1, \\ 0, & \text{otherwise} \end{cases} ,$$

A similar technique has been used in [64] to show the independence of the quantization error and the host signal when a uniform dither is used. In [64] the role of the circular convolution is played by the sampling of the characteristic function with period π. This sampling has an aliasing effect, since it is equivalent to the convolution in the time domain with an impulse train with period 2.

When the symbol b is embedded, it is clear from (2.22) that the decision statistics will take the form

$$\tilde{z}_i = \frac{Q_b(x_i) - (1-\alpha) e_i + n_i - Q_0(z_i)}{\Delta_i} \mod 2\mathbb{Z} = (u_i + b) \mod 2\mathbb{Z}, \tag{2.32}$$

for all $i = 1, \cdots, L_2$, or, in short, $\tilde{\mathbf{z}} = (\mathbf{u} + b\mathbf{1}) \mod 2\mathbb{Z}^{L_2}$, where $\mathbf{1}$ is a vector of L_2 ones. Equivalently, $\mathbf{u} = (\tilde{\mathbf{z}} - b\mathbf{1}) \mod 2\mathbb{Z}^{L_2}$. Then, the decision rule in (2.27) is equivalent to deciding $\hat{b} = 0$ whenever

$$f_{\mathbf{U}}(\tilde{\mathbf{z}}) > f_{\mathbf{U}}\left((\tilde{\mathbf{z}} - \mathbf{1}) \mod 2\mathbb{Z}^{L_2} \right), \tag{2.33}$$

and $\hat{b} = 1$ otherwise.

2.6.1. An Approximation to the ML Lattice Decoder

In [72] Forney et al. provided useful approximations to the pdf of a modulo-reduced —or aliased— Gaussian pdf. The same approach can be followed to

[4] Note that $\mathbf{U}$ is not the same as $\tilde{\mathbf{Z}}$, since the latter will depend on the sent bit.

write an approximation of the pdf of the *modular total noise* random variable U_i, defined in (2.30) and needed in (2.33) for ML lattice decoding.

Recall from (2.29) that, for $N_i \sim \mathcal{N}(0, \sigma_{N_i}^2)$, the pdf of T_i is just the convolution of a Gaussian with zero-mean and variance $\sigma_{N_i}^2/\Delta_i^2$, and a uniform pdf in $(-(1-\alpha), +(1-\alpha)]$. Pursuing the sort of approximations proposed in [72], it is possible to conclude that:

- For $\sigma_{T_i} \ll 1$, the contributions in the summation in (2.30) for $l \neq 0$ are negligible. The most significant part of the pdf of T_i is concentrated in the interval $(-1, +1]$, so the aliasing effect can be neglected. Therefore $f_{U_i}(u_i)$ can be well approximated by $f_{T_i}(t_i)$.
- For $\sigma_{T_i} \gg 1$, it is possible to consider that T_i follows a Gaussian distribution with $\sigma_{T_i}^2 = \sigma_{N_i}^2/\Delta_i^2 + (1-\alpha)^2/3$. Now the pdf $f_{U_i}(u_i)$ becomes nearly constant due to the strong aliasing. Observe that since $\sum_l f_{T_i}(u_i - 2l)$ in (2.30) is periodic if we do not restrict u_i to lie on $(-1, +1]$, it makes sense to expand it in terms of its Fourier series and then truncate it to this interval. Forney et al. suggested approximating this function by keeping the low-frequency terms of this expansion.

The computation of the Fourier series expansion of a periodic function on a lattice can be performed by using the dual of that lattice [49]. As in our case U_i is obtained by means of the lattice $2\mathbb{Z}$, the corresponding dual lattice is simply given by $\mathbb{Z}/2$, and so the desired pdf can be expanded as [72]

$$f_{U_i}(u_i) = \frac{1}{2} \sum_{k \in \mathbb{Z}} \exp\left(-\pi^2 \sigma_{T_i}^2 k^2/2\right) e^{j2\pi k u_i/2}, \quad -1 < u_i \leq 1. \tag{2.34}$$

The DC and fundamental frequency terms in this expansion correspond to $k = 0$ and $k = \pm 1$, respectively. Keeping just these two terms in (2.34), we can write $f_{U_i}(u_i) \approx \frac{1}{2}\left(1 + 2\mathrm{e}^{(-\pi^2\sigma_{T_i}^2)/2} \cos(\pi u_i)\right)$, for $-1 < u_i \leq 1$. The usefulness of this approximation is illustrated in Section 3.1.3, where a geometrical interpretation of lattice ML decision regions is provided.

2.6.2. Euclidean and Weighted Euclidean Distance-based Lattice Decoder

Despite the complexity reduction from ML decoding to ML lattice decoding and the further diminution brought about by Forney's approximation, it is desirable to seek even simpler decoding strategies. In this section we discuss lattice decoding based on the Euclidean distance.

When each dimension is normalized by its quantization step, Euclidean lattice decoding can be written as

$$\hat{b} = \underset{b\in\{0,1\}}{\arg\min} \|\mathbf{\Delta}^{-1}(\mathbf{z} - Q_b(\mathbf{z}))\|^2, \tag{2.35}$$

where $\mathbf{\Delta} \triangleq \text{diag}(\Delta_1, \cdots, \Delta_{L_2})$. This approximation is tantamount to choosing the b whose associated shifted lattice Λ_b yields the minimum normalized quantization error. Minimum distance decoding of DC-DM was in fact part of the original proposal of DC-DM in [32]. It is also used in [137, 33], and in [64] for the equivalent Scalar Costa Scheme (SCS).

To see the relationship between this decoding strategy and ML lattice decoding, let $\mathcal{S} \triangleq \{\pm 1\}^{L_2}$. Recalling the definition of $\tilde{\mathbf{z}}$, it is clear that for $b = 0$, $\|\mathbf{\Delta}^{-1}(\mathbf{z} - Q_b(\mathbf{z}))\| = \|\tilde{\mathbf{z}}\|$, while for $b = 1$, $\|\mathbf{\Delta}^{-1}(\mathbf{z} - Q_b(\mathbf{z}))\|$ becomes $\|\tilde{\mathbf{z}} - \mathbf{s}\|$, where $\mathbf{s} \in \mathcal{S}$ is such that $\|\tilde{\mathbf{z}} - \mathbf{s}\|$ is minimum. Putting this together, we can rephrase the decoding rule in (2.35) as deciding $\hat{b} = 0$ if

$$\|\tilde{\mathbf{z}}\|^2 < \min_{\mathbf{s}\in\mathcal{S}} \|\tilde{\mathbf{z}} - \mathbf{s}\|^2, \tag{2.36}$$

and $\hat{b} = 1$ otherwise.

But now, the vector $\mathbf{s}$ minimizing the norm in the right hand side of (2.36) is such that it also satisfies $(\tilde{\mathbf{z}} - \mathbf{s}) = (\tilde{\mathbf{z}} - \mathbf{s}) \mod 2\mathbb{Z}^{L_2}$. Thus, the parallelism between (2.36) and (2.33) is clear if one considers that the modulo-$2\mathbb{Z}^{L_2}$ operation maps the set $\mathcal{S}$ onto vector $\mathbf{1}$, i.e., for all $\mathbf{s} \in \mathcal{S}$, $(\tilde{\mathbf{z}} - \mathbf{s}) \mod 2\mathbb{Z}^{L_2} = (\tilde{\mathbf{z}} - \mathbf{1}) \mod 2\mathbb{Z}^{L_2}$.

In fact, the two decoding rules would be equivalent if the modular total noise $\mathbf{U}$ had Gaussian independent and identically distributed (i.i.d.) components. It is convenient to examine under which conditions this latter property would hold. First, in order to neglect overlaps of the shifted versions of f_{T_i} in the construction of f_{U_i} in (2.30), it is required that $\sigma_{T_i} \ll 1$ for all i. Second, for the T_i to be i.i.d. Gaussian, a necessary and sufficient condition would be that $\alpha = 1$ (i.e., there is no self-noise) and that the noise components N_i are independent Gaussian, with variances proportional to Δ_i^2 for all i. Notice that in general these conditions will not be satisfied, so Euclidean distance decoding will be suboptimal.

When those conditions are not met, it is useful to modify minimum distance decoding while still retaining a relatively simple decoding approach by comparison with ML lattice decoding. To this end, we introduce a weighted Euclidean distance, for which the decoding rule becomes

$$\hat{b} = \underset{b\in\{0,1\}}{\arg\min} \left\{ (\mathbf{z} - Q_b(\mathbf{z}))^T \mathbf{\Delta}^{-1}\mathbf{B}\mathbf{\Delta}^{-1} (\mathbf{z} - Q_b(\mathbf{z})) \right\}, \tag{2.37}$$

where the weighting matrix $\mathbf{B}$ is defined as

$$\mathbf{B} \triangleq \text{diag}(\beta_1, \cdots, \beta_{L_2}).$$

The purpose of these weights is to introduce additional degrees of freedom to improve decoding in practice when minimum distance decoding is just too far away from optimality. We will show in Section 3.1.2.2 how a proper design of the parameter vector $\boldsymbol{\beta}$ allows to improve decoding when additional information about the channel noise is available. Also, it should be taken into account that the normalization by Δ_i in (2.35) does not entail any loss of generalization or loss in performance, since its effect could be canceled by β_i in any case. Whenever no optimization is attempted, $\beta_i = 1$ will be set for all i. In this case (2.37) becomes equivalent to (2.35).

2.7. Spread Transform Dither Modulation Embedding and Decoding

In their seminal paper Chen and Wornell [32] also proposed the Spread-Transform Dither Modulation (STDM), which is based on the projection of a number of features of the original host signal $\mathbf{x}$ on a scalar domain. The projected signal x_{p_j} corresponding to the j-th subvector is simply obtained as $x_{p_j} = \mathbf{s}_j{}^T\mathbf{x}_j$, with $j \in \{0, \cdots, L_b\}$ where $\mathbf{s}_j$ is the j-th projecting vector, which depends on the secret key $\boldsymbol{\Theta}$. Then, when the symbol b_j is transmitted, the watermarked signal in the projected domain is given by $y_{p_j} = Q_{P\Delta_j\mathbb{Z}}(x_{p_j} - v_j(b_j)) + v_j(b_j)$ with $j \in \{0, \cdots, L_b\}$, i.e. the projected coefficient is watermarked using DM (without distortion compensation). Since an infinite number of vectors $\mathbf{y}_j$ exist verifying $\mathbf{s}_j^T\mathbf{y}_j = y_{p_j}$, it is customary to choose $\mathbf{y}_j$ such that the watermark $\mathbf{w}_j = \mathbf{y}_j - \mathbf{x}_j$ has minimum Euclidean norm; this results in $\mathbf{y}_j = \mathbf{x}_j + (y_{p_j} - x_{p_j}) \cdot \mathbf{s}_j / ||\mathbf{s}_j||^2$. Typically, the projection vector $\mathbf{s}_j$ is pseudorandomly generated in a key-dependent fashion; thus, it is no longer imperative to make the base dither d_0 pseudorandom. Therefore, in order to minimize the power of the watermark, the dither d_0 is usually restricted to take values in $\{-\Delta/2, +\Delta/2\}$, since in that case, for $L_2 \to \infty$, $D_w \to \Delta^2/(4 \cdot L_2 \cdot ||\mathbf{s}_j||^2)$, while if d_0 were uniformly distributed over $(-\Delta, +\Delta]$, $D_w = \Delta^2/(3 \cdot L_2 \cdot ||\mathbf{s}_j||^2)$ (see [132] and [20] for a further discussion on this topic). Hereafter we will denote this method as Scalar STDM (SSTDM), in order to distinguish it from the generalized versions that we will just introduce.

First of all, we can talk about the *Spread Transform-Scalar Costa Scheme* (ST-SCS) by Eggers and Girod [64], where the distortion compensation is performed. Interestingly, as discussed in [132] and [118], if the projection rate L_1/L_b is such that $L_1/L_b \gg 1$, as it often occurs in robust data hiding, then performing no distortion compensation is nearly optimal, i.e., the optimal α goes to 1, so ST-SCS using the optimal distortion compensation parameter would approach SSTDM.

Furthermore, Pérez-González et al. proposed their *Quantized Projection* (QP) method [132], which is based on the projection of the original host signal $\mathbf{x}$ to a lower dimensional domain, not necessarily the unidimensional one, where it is quantized using DC-DM, including the distortion compensation; a similar approach is studied in [120]. Therefore, we will define the projected host signal as $\mathbf{x}_p = \mathbf{S}^T\mathbf{x}$, where $\mathbf{S}$ is a projection matrix of size $L_1 \times L_3$, $L_3 \leq L_1$; now the L_3-dimensional vector $\mathbf{x}_p$ plays the same role that $\mathbf{x}$ played in DC-DM, i.e., it is divided in L_b subvectors of size L_4 (defined as $L_4 \triangleq L_3/L_b$), each of them conveying the j-th symbol, with $j \in \{1, \cdots, L_b\}$, in such a way that

$$\mathbf{y}_{p_j} = (1-\alpha_j)\mathbf{x}_{p_j} + \alpha_j[Q_{\Lambda_j}(\mathbf{x}_{p_j} - \mathbf{v}_j(b_j)) + \mathbf{v}_j(b_j)], \quad j \in \{1, \cdots, L_b\}. \tag{2.38}$$

Nevertheless, an infinite number of vectors $\mathbf{y}$ exists verifying $\mathbf{S}^T\mathbf{y} = \mathbf{y}_p$, so some additional criteria should be established to choose one of them; it is customary to select $\mathbf{y}$ such that the watermark $\mathbf{w} = \mathbf{y} - \mathbf{x}$ has minimum Euclidean norm, so using the Moore-Penrose pseudoinverse we obtain[5]

$$\mathbf{y} = \mathbf{x} + \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}(\mathbf{y}_p - \mathbf{x}_p). \tag{2.39}$$

Concerning decoding, the optimal strategy is given by

$$\hat{\mathbf{b}} = \underset{\mathbf{b}\in\{0,\cdots,P-1\}^{L_b}}{\arg\max} f_{\mathbf{Z}|\mathbf{B}}(\mathbf{z}|\mathbf{b}). \tag{2.40}$$

Nevertheless, as it was previously discussed for DC-DM, some simplifications are mandatory in order to obtain practical decoding strategies, even if they imply a loss in performance. The first one is to consider the projected version of the received signal, so decoding can be implemented as

$$\hat{b}_j = \underset{b_j\in\{0,\cdots,P-1\}}{\arg\max} f_{\mathbf{Z}_{p_j}|B_j}(\mathbf{z}_{p_j}|b_j), \quad j \in \{1, \cdots, L_b\}; \tag{2.41}$$

we remind that the projection operation could be information lossy, so considering only the projected received signal is not optimal in general. Nevertheless, in most cases this simplification is not enough, and a minimum Euclidean distance lattice decoding is usually followed, i.e.

$$\hat{b}_j = \underset{b_j\in\{0,P-1\}}{\arg\min} ||\mathbf{Z}_{p_j} - Q_{b_j}(\mathbf{Z}_{p_j})|| = \underset{b_j\in\{0,P-1\}}{\arg\min} ||\mathbf{Z}_{p_{\mathrm{mod}_j}} - Q_{b_j}(\mathbf{Z}_{p_{\mathrm{mod}_j}})||, \tag{2.42}$$

where $\mathbf{Z}_{p_{\mathrm{mod}}} = \mathbf{Z}_p \mod \Lambda$. As it was discussed in Section 2.5, the modulo reduction is information lossy as well, and the minimum distance criterium is not optimum in general, so (2.42) will be far from being the optimal strategy in most scenarios. Once again, due to its manageability, this strategy is the most used in practical implementations. In the same way, SSTDM is probably the most

[5]Making a proper choice of $\mathbf{v}_j(b_j)$ it is also possible to reduce the watermark power by 4/3 (see [132] and [20] for a further discussion on this topic).

popular of these methods (see [65], [28], [103], and references therein) due to its simplicity, and despite showing worse performance than the generalized versions; furthermore, SSTDM can be built upon an existing Add-SS scheme, since the projection (also referred to as spreading) stage is essentially identical for both.

Finally, be aware that DC-DM could be seen as a particular case of this generalized approach to STDM (hereafter just STDM), where $\mathbf{S} = \mathrm{I}_{L_1 \times L_1}$ is the identity matrix of size L_1.

2.7.1. Advantages of STDM

In order to recall some well-known advantages of STDM we will assume that the columns of $\mathbf{S}$ are orthonormal, i.e. $\mathbf{S}^T\mathbf{S} = \mathrm{I}_{L_3 \times L_3}$, $\mathbf{W}_p$ is i.i.d. with variance $\sigma^2_{W_p}$, and $\mathbf{N}$ is i.i.d. with variance σ^2_N; we will denote by σ^2_W the variance of the watermark in the original domain, and by $\sigma^2_{N_p}$ the noise variance in the projected one. Taking into account that $\sigma^2_W = \frac{1}{L_1}\mathrm{Tr}(\mathrm{E}\{\mathbf{W}\mathbf{W}^T\})$, where $\mathrm{Tr}(\cdot)$ is the trace operator, we can write

$$\begin{aligned}\sigma^2_W &= \frac{1}{L_1}\mathrm{Tr}(\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathrm{E}\{\mathbf{W}_p\mathbf{W}_p^T\}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T) \\ &= \frac{\sigma^2_{W_p}}{L_1}\mathrm{Tr}(\mathbf{S}\mathbf{S}^T) = \frac{\sigma^2_{W_p}}{L_1}\mathrm{Tr}(\mathbf{S}^T\mathbf{S}) = \frac{L_3\sigma^2_{W_p}}{L_1},\end{aligned}$$

and

$$\sigma^2_{N_p} = \frac{1}{L_3}\mathrm{Tr}(\mathbf{S}^T\mathrm{E}\{\mathbf{N}\mathbf{N}^T\}\mathbf{S}) = \frac{\sigma^2_N}{L_3}\mathrm{Tr}(\mathbf{S}^T\mathbf{S}) = \sigma^2_N,$$

in such a way that $\frac{\sigma^2_{W_p}}{\sigma^2_{N_p}} = \frac{L_1}{L_3}\frac{\sigma^2_W}{\sigma^2_N}$, i.e. the watermark to noise ratio is increased by a factor L_1/L_3 (the projection rate) in the projected domain, while keeping a small number of nearest neighbor codewords [32]. This gain afforded by spreading is the responsible for the robustness of STDM methods in presence of additive noise attacks [65], [132] and quantization attacks [20], and it is shared by Add-SS. Of course, owing to its host interference rejection properties, STDM performs much better than Add-SS in face of the previous attacks. Furthermore, this gain implies that the generalized version of STDM is more robust against additive noise attacks than both unidimensional DC-DM and DC-DM based on uniform scalar quantizers and repetition coding, as reported in [64] and [132].

However, as shown in [20], unlike Add-SS, SSTDM (and STDM in general) is very vulnerable to gain (i.e., amplitude scaling) attacks, a weakness that has sparked recent research in gain-invariant data hiding methods (see for example [134] and [12]).

In addition to this increase in the effective WNR, an interesting feature of STDM is that for large values of L_1/L_3 and a wide range of additive noise distributions, performance can be analyzed by considering an equivalent Gaussian noise with the same first and second order moments. This simplification, which is a consequence of the Central Limit Theorem (CLT), considerably facilitates the analysis and enables the optimization of some parameters which improve decoding, as investigated in Section 3.1.2.2. See [20] for an in-depth discussion upon the validity of the CLT approximation in this context.

A further advantage of STDM-like strategies is that it is more complicated to design attacks which render the attacked samples close to the decision boundary for them than for DC-DM. This said, there are other simple attacks which can be much more detrimental for STDM-like methods than for DC-DM, as it will be shown in Section 3.3.

Chapter 3

Robustness

In this chapter we will analyze the robustness of DC-DM with uniform scalar quantizers and repetition coding when the watermarked signal goes through some typical channels; for data hiding, these channels are, for example, additive noise, coarse quantization, or the cropping attack. We have chosen to study this simple version of DC-DM for several reasons:

- It is probably the most used side-informed scheme in practical implementations; nevertheless, despite of its extensive use, most results showed in this thesis are novel, since little has been said before about the robustness of that scheme.
- It is amenable to analytical expressions, which are not available for other schemes.
- STDM methods using uniform scalar quantizers and repetition coding (as the original scalar proposal of STDM [32]) can be analyzed by slightly modifying the proposed methodologies. In fact, when analyzing the cropping attack in Section 3.3 we will firstly focus on scalar STDM, and then will go to a more general framework (multidimensional STDM), trying to obtain a method robust to this attack.

Furthermore, due to the importance of oracle-like attacks, we will also study the robustness of some state-of-the-art data hiding detection methods against a novel version of these attacks: the Blind Newton Sensitivity Attack (BNSA). Among the analyzed methods is the Quantized Projection based Detection (QPD) [126], proposed by Pérez-Freire et al., which could be considered as the counterpart of DC-DM with uniform scalar quantizers and repetition coding for the detection problem. We will also discuss that the BNSA can be also applied to data hiding decoding schemes, by replacing the multiple hypothesis test by a binary one.

In order not to just consider *static* attackers and decoders, but also *active* and *smart* ones, some results showing the strategies that should be followed from a Game Theoretic point of view are introduced, for the cases of Add-SS, DC-DM and STDM methods. The performance achieved by following these strategies are somehow upperbounded by the mutual information between the received signal and the sent symbol; this justifies the importance of some experimental results [130] which compute that mutual information for the worst additive attack (from a Information Theoretic perspective), for the case of scalar DC-DM with uniform quantizers (equivalent to SCS); these experimental results also allow the computation of the optimal distortion compensation parameter α in that scenario, which will be compared with those obtained by Costa and Eggers.

3.1. Additive Noise

Although DC-DM with uniform scalar quantizers is a suboptimal side-informed scheme, it is well known that it has an achievable rate often acceptably close to the ideal limit [32, 64]. Nevertheless, performance analyses for the probability of decoding error of DC-DM are scarce, and usually either incomplete or inexact. Among previous attempts, we may cite first those ones devoted to determine the decoding performance of DM, i.e., without distortion compensation [137, 33, 132]. Also, upperbounding strategies to DC-DM with repetition coding were studied in [131], whereas an approximation to the bit error rate of generic DC-QIM methods is also given in [32]. The main objective of this section, which is mainly based on [46], is to provide a thorough analysis of DC-DM with uniform scalar quantizers and repetition coding, presenting accurate theoretical approximations and bounds to the bit error rate at the decoder. Building on our analysis, we also propose enhancements on this standard scheme, both by means of optimizable weights on the standard Euclidean-distance lattice decoder, and by introducing a novel vectorial structure for the distortion-compensation parameter.

As it was done in Section 2.5, we restrict our presentation to any of the L_2-dimensional subvectors inside which the host signal samples can be assumed independent, so the subindex j indicating the subvector will be dropped in the sequel for notational simplicity.

3.1.1. Performance Analysis

Next, we will analyze the performance of binary repetition DC-DM in terms of the bit error rate (BER) at the decoder output. In this section we will consider only the unweighted minimum Euclidean distance approach to DC-DM decoding,

i.e., $\mathbf{B} = \mathbf{I}_{L_2 \times L_2}$, where $\mathbf{B}$ was introduced in Section 2.6.2 and $I_{L_2 \times L_2}$ denotes the identity matrix of size L_2, leaving to Section 3.1.2.2 the study of the effect of the weights $\boldsymbol{\beta}$.

Costa's framework considers i.i.d. channel noise and signal (in our case watermark), which naturally induces the same distortion compensation parameter α for all dimensions. On the other hand, perceptual considerations motivate that in our scheme the variances of both the channel noise and the watermark be in general different for each of the dimensions. Although this setting suggests a vectorial distortion compensation parameter $\boldsymbol{\alpha}$ —i.e., dimension-dependent—, for the sake of simplicity we will only deal with a scalar α in this section, and explore the vectorial possibility in Section 3.1.2.1.

From (2.37) we can see that decoding is equivalent to quantizing $\mathbf{z}$ with both the shifted lattice Λ_0 and Λ_1 and then assigning the value of the bit that yields the smallest (in an Euclidean distance sense) normalized quantization error. Obviously, this is completely equivalent to quantizing $(\boldsymbol{\Delta}^{-1}\mathbf{z})$ with $(\boldsymbol{\Delta}^{-1}\Lambda_0) \cup (\boldsymbol{\Delta}^{-1}\Lambda_1)$ following also a minimum Euclidean distance criterion.

It can be readily seen that the probability of decoding error does not depend on the actual embedded bit. Let us assume then that $b = 0$ is sent, so $\tilde{\mathbf{Z}} = \mathbf{U}$. Hence, taking (2.36) into account, an error happens whenever

$$||\mathbf{u}||^2 > \min_{\mathbf{s} \in \mathcal{S}} ||\mathbf{u} - \mathbf{s}||^2. \tag{3.1}$$

The minimization in (3.1) is equivalent to seeking the closest centroid to $\mathbf{u}$ among the shifted lattice corresponding to $b = 1$. The decoding region given by (3.1) is a generalized octahedron [49] whose vertices are those vectors having only one non-zero component with value $\pm L_2/2$.

Therefore, a decoding error will happen if and only if $\mathbf{u}$ lies out of this generalized octahedron. Due to the symmetry of the octahedron in all the orthants with respect to the origin, it is reasonable to project the random variable $\mathbf{U}$ onto the positive one, to construct $U_i^+ \triangleq |U_i|, 1 \leq i \leq L_2$, and then proceed to determine the probability of being closer to the vertex $\mathbf{s}_1 \triangleq \mathbf{1} \in \mathcal{S}$, than to the origin. This probability is thus the probability of bit error, which can be written as

$$P_e = \Pr\{\|\mathbf{U}^+\|^2 > \|\mathbf{U}^+ - \mathbf{1}\|^2\} = \Pr\left\{\sum_{i=1}^{L_2} U_i^+ > L_2/2\right\}. \tag{3.2}$$

The evaluation of this expression requires the *probability density function* (pdf) of U_i^+, $1 \leq i \leq L_2$, which is just

$$f_{U_i^+}(u_i^+) \triangleq \begin{cases} \left[f_{U_i}(u_i^+) + f_{U_i}(-u_i^+)\right], & \text{if } 0 \leq u_i^+ \leq 1 \\ 0, & \text{otherwise} \end{cases}, \quad 1 \leq i \leq L_2. \tag{3.3}$$

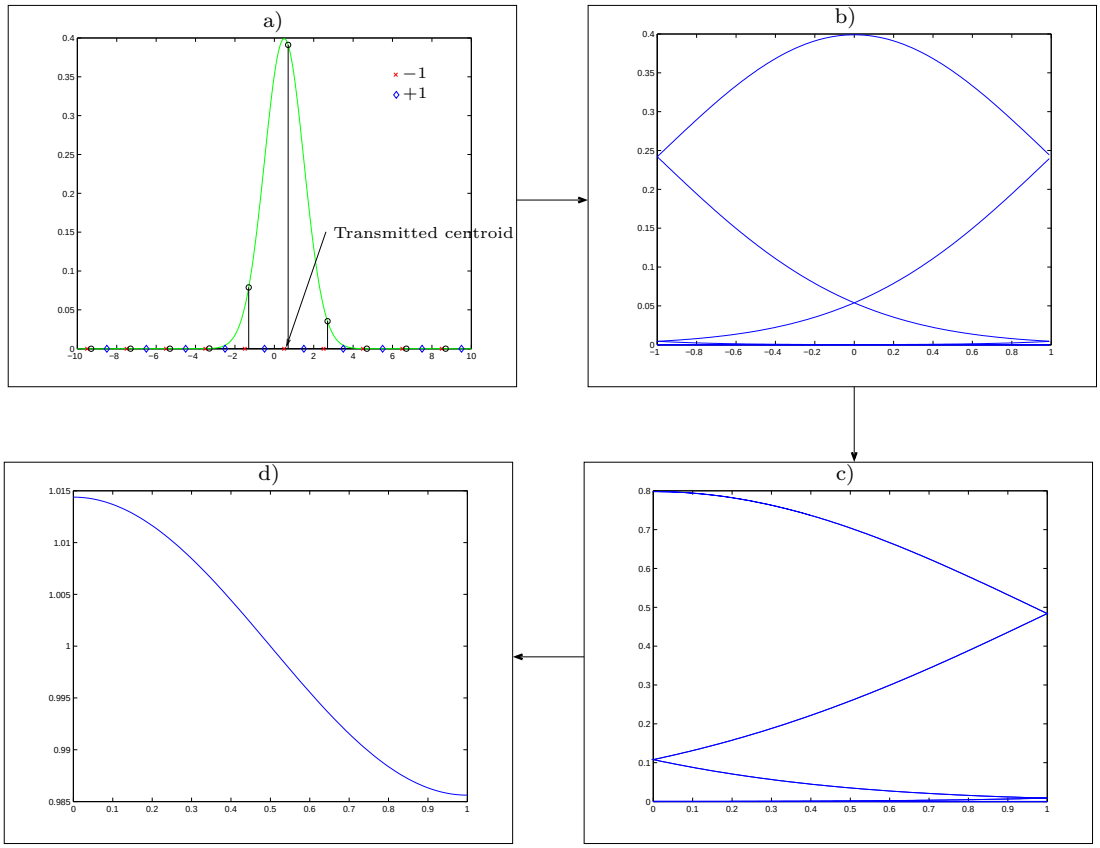


Figure 3.1: Gaussian noise modulo-lattice reduced for the unidimensional case. a) Histogram when a given centroid is transmitted. All the circles are *folded* to the same point after the modulo reduction. b) Noise's histogram after modulo reduction (before adding its components). c) Histogram of the noise absolute value (before adding its components). d) Histogram of the noise absolute value (after adding its components).

In Figure 3.1 the modulo-lattice reduction, and the projection onto the positive orthant is illustrated for the unidimensional case.

Therefore, if we define the variable

$$R \triangleq \sum_{i=1}^{L_2} U_i^+, \tag{3.4}$$

then from (3.2), the computation of P_e is equivalent to integrating the tail of the pdf of R from $L_2/2$ to L_2.

Even though formula (3.2) allows us to determine the *exact* probability of bit error, its computation is very expensive for large L_2. This motivates the proposal of two numerical approaches for its calculation, which are discussed in Sections 3.1.1.1 and 3.1.1.2. On the other hand, neither formula (3.2) nor these

practical methods provide closed-form expressions, making it difficult to extract conclusions of theoretical value. For this reason, Sections 3.1.1.3 and 3.1.1.4 are devoted to discussing analytical approximations and bounds respectively.

3.1.1.1. Beaulieu's Approach

In this section, we adapt a technique proposed by Beaulieu [21] for computing the tail probability of the summation of L_2 i.i.d. random variables, as it occurs in (3.2). This technique was already used in [132] to upperbound the bit error probability of DM. Let $\omega_l \triangleq \frac{2\pi l}{T}$ for any positive integer l, with T a large enough real number, and let $F_{U_i^+}(\omega)$ be the characteristic function of U_i^+, given by

$$F_{U_i^+}(\omega) = \int_0^1 \mathrm{e}^{j\omega u} f_{U_i^+}(u) du. \tag{3.5}$$

Then, the computation of P_e is made, following [21], as

$$P_e \approx \frac{1}{2} + \frac{2}{\pi} \sum_{\substack{l=1 \\ l \text{ odd}}}^{\infty} \frac{\prod_{i=1}^{L_2} |F_{U_i^+}(\omega_l)| \sin(\sum_{i=1}^{L_2} \phi_i(\omega_l))}{l}, \tag{3.6}$$

where $\phi_i(\omega)$ is defined as $\phi_i(\omega) \triangleq \arg\{F_{U_i^+}(\omega)\} - \omega/2$, with $\arg(\cdot)$ denoting the four-quadrant phase. The main drawback of this method is that it is rather computationally demanding, apart from the fact that it may present numerical problems due to the large values that could be involved in the summation of a truncated version of the series in (3.6). In Appendix B the expressions of the functions required for computing (3.6) for a Gaussian channel noise are derived.

3.1.1.2. DFT Method

Since the U_i^+ in (3.4) are independent random variables, the pdf of R is just the convolution of the pdfs of U_i^+, $1 \leq i \leq L_2$. This computation can be efficiently done in the Discrete Fourier Transform (DFT) domain. To that end, let $\Phi_{U_i^+} \triangleq \mathrm{DFT}_{L_2K}\left(K \cdot f_{U_i^+}(k/K)\right)$ be the $L_2 \times K$-point DFT of the sequence obtained by sampling $f_{U_i^+}(\tau)$ at $\tau = \frac{k}{K}$, with $k = 0, \cdots, K-1$. Using this definition it is straightforward to write $\Phi_R[m] = \prod_{i=1}^{L_2} \Phi_{U_i^+}[m], \quad m = 0, \cdots, L_2K - 1$. Finally, the discretized pdf of R is obtained using the Inverse Discrete Fourier Transform (IDFT) as $f_R = \mathrm{IDFT}_{L_2K}(\Phi_R)$, and (3.2) is computed as

$$P_e \approx \sum_{k=\left\lceil \frac{L_2(K-1)+1}{2} \right\rceil}^{L_2K-1} f_R[k],$$

where the limits of this summation stand for $R = L_2/2$ and $R = L_2$ in the corresponding integral. The accuracy of the computation can be increased by using a larger value of K, i.e., by sampling more finely the pdfs involved in the calculation.

This technique resembles Beaulieu's approach in that both of them work in a transform domain. Nevertheless, the DFT method presents a much lower computational cost, without any of the numerical problems shown by Beaulieu's approach. This fact makes the DFT method an enticing approach to assess the performance to any degree of accuracy required.

3.1.1.3. Central Limit Theorem-based Approximation

A third option consists in taking advantage of the independence of the random variables U_i^+ in the summation (3.4) to invoke the Central Limit Theorem (CLT). This result states that the distribution of R will tend to a Gaussian as $L_2 \to \infty$, in which case we may approximate the probability of error as

$$P_e \approx \mathcal{Q}\left(\frac{\frac{L_2}{2} - \sum_{i=1}^{L_2} \mathrm{E}\{U_i^+\}}{\sqrt{\sum_{i=1}^{L_2} \mathrm{Var}\{U_i^+\}}}\right), \tag{3.7}$$

where $\mathcal{Q}(\cdot)$ was already defined in Section 2.3 as $\mathcal{Q}(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{\tau^2}{2}} d\tau$.

The main advantage of CLT-based approximation is that it gives a closed expression for P_e, which can be exploited for analytical purposes (see Section 3.1.2.2). Although this method to compute P_e is much simpler than the previous ones, some remarks are due. First, for small values of L_2 it could entail problems in the convergence of R to a Gaussian. One factor that speeds up convergence is the similarity between the distributions of the summands. Also, and as discussed in [131], note that the process of building the one-sided distributions $f_{U_i^+}(u_i^+)$ may produce highly skewed pdfs whose convolution converges very slowly to a Gaussian distribution. If this is the case, the Gaussian approximation to P_e may underestimate the importance of the tails of $f_R(r)$.

Last, although analytical expressions of $\mathrm{E}\{U_i^+\}$ and $\mathrm{Var}\{U_i^+\}$ are available in closed form when N_i follows a uniform distribution [37], in general the explicit computation of these statistics may require numerical integration. Therefore, it is recommended to use the DFT method for obtaining numerical results, as it gives a higher degree of accuracy.

3.1.1.4. Bounds on P_e

In this section we discuss several other known bounds on the bit error probability.

3.1.1.4.1. Erez and Zamir's Bound

Erez and Zamir have recently proposed a method that can be accommodated to upperbound the probability of error of binary DC-DM when performing Euclidean lattice decoding —i.e., minimum distance decoding— under Additive White Gaussian Noise (AWGN) channel distortion [70]. Let $\mathcal{W} = \mathcal{V}\left(\{2\mathbb{Z}^{L_2}\} \cup \{2\mathbb{Z}^{L_2} + \mathbf{1}\}\right)$ denote the region associated to a right decision, i.e., $P_e = \Pr\{\mathbf{U} \notin \mathcal{W}\}$. Then, it is possible to write $P_e \leq \Pr\{\mathbf{T} \notin \mathcal{W}\}$. Erez and Zamir's procedure may be used to construct an upper bound on the latter probability that depends on D_w, D_c, Λ and L_2. In turn, this obviously upperbounds P_e. Unfortunately, the bound turns out to be rather loose for our particular problem (see Section 3.1.7); for this reason, we will omit the details of its implementation.

3.1.1.4.2. Union Bound and Nearest Neighbor Approximation

The classical *union bound* (UB) is based on adding the pairwise probability of mistaking the transmitted centroid with each of its nearest neighbors corresponding to a wrong decision. The possible overlaps of the error regions associated to each of these error events are disregarded in this computation, and this is the reason why it produces an upper bound. When the WNR is increased these overlaps diminish, and so the bound gets closer to the true value. As in our implementation of DC-DM we are using uniform scalar quantizers, there are 2^{L_2} nearest error neighbors. Thus, assuming that the pdf of the channel distortion is symmetric, the union bound may be computed as

$$\begin{aligned} P_e \leq P_{\text{union}} &= 2^{L_2} \cdot \Pr\{\|\mathbf{U}\|^2 > \|\mathbf{U} - \mathbf{1}\|^2\} \\ &= 2^{L_2} \cdot \Pr\left\{\sum_{i=1}^{L_2} U_i > L_2/2\right\}, \end{aligned}$$

where the last probability can be obtained by means of any of the methods in Sections 3.1.1.1–3.1.1.3, similarly to what is done with (3.2). Alternatively, for L_2 large enough, we can compute an approximation applying the Central Limit Theorem. To this end, we just need to compute the variance of the zero-mean random variable whose pdf is the circular convolution of the channel noise and the self-noise. Note that due to the approximation implicit in the CLT, we can no longer ensure that the result is a bound, but an approximation to the bound,

which will be asymptotically good as $L_2 \to \infty$. This approximation is given by

$$P_e \approx 2^{L_2} \cdot \mathcal{Q}\left(\frac{L_2}{2\sqrt{\sum_{i=1}^{L_2} \mathrm{Var}\{U_i\}}}\right). \tag{3.8}$$

In contrast to Section 3.1.1.3, if N_i is symmetric about the origin the involved pdfs (i.e., those of $f_{U_i}(u_i)$) are also symmetric, so their convolution will converge more quickly to a Gaussian distribution.

Following the previous guidelines for the union bound we may also approximate the bit error probability using the nearest neighbor distance sketched in [32]. The estimate therein assumes Quantization Index Modulation without distortion compensation and additive white Gaussian noise. This result may be improved by replacing the real Gaussian pdf with a Gaussian with variance the sum of those corresponding to the channel noise and the self-noise, what yields

$$P_e \sim \mathcal{Q}\left(\frac{L_2}{2\sqrt{\sum_{i=1}^{L_2} \mathrm{Var}\{U_i\}}}\right). \tag{3.9}$$

Following the discussion in [20] on the validity of the CLT, it is necessary to check against empirical results all the CLT-based approximations and bounds that we have given in Sections 3.1.1.3 and 3.1.1.4.2. This task is undertaken in Section 3.1.7.1.

3.1.2. Improvements on Standard DC-DM

In this section we introduce some improvements in the performance of the DC-DM scheme studied so far. Specifically, we will deal with the distortion compensation parameter as well as with the decoding weights.

3.1.2.1. Study of the Distortion Compensation Parameter

The distortion compensation parameter α, may be used in two equivalent ways. Namely, it may reduce the embedding power by a factor α^2 for a fixed lattice, or, alternatively, it may afford an expansion of the lattice by a factor $\frac{1}{\alpha}$ when the power of the watermark is kept constant. Interestingly, it can be shown that both lead to the same bit error probability for a given WNR when the power spectral density of the noise sequence is fixed, save for a multiplicative constant. Therefore, although throughout this work we are using a fixed lattice, we should be aware that, when the stated conditions are met, this is equivalent to the expansion of that lattice for a fixed D_w.

The determination of the distortion compensation parameter may be tackled under a number of different optimization criteria. Obviously, these criteria will in general lead to different values of α. Probably the simplest, but also one of the most used, is the minimum mean square error (MMSE) criterion (see [73]). This criterion was for instance used in [70]. We may also think of optimizing this parameter depending on the bit error rate. The problem in this case is the lack of closed-form expressions that would allow to face the optimization problem in an analytical way. Following MMSE, the initial intention would be to minimize $\sum_{i=1}^{L_2} \sigma_{U_i}^2$; however, due to the aliasing effect, this becomes an unsurmountable problem. Considering that for large WNRs and large values of α the modulo operation can be neglected, it is reasonable to address instead the minimization of

$$\begin{aligned}\varphi(\alpha) &\triangleq \sum_{i=1}^{L_2} \sigma_{T_i}^2 = \sum_{i=1}^{L_2} \left\{ \frac{\sigma_{N_i}^2}{\Delta_i^2} + \frac{(1-\alpha)^2}{3} \right\} \\ &= \sum_{i=1}^{L_2} \left\{ \frac{\alpha^2 \xi_i}{3} + \frac{(1-\alpha)^2}{3} \right\}, \end{aligned} \tag{3.10}$$

for a fixed $\xi_i \triangleq \sigma_{N_i}^2/\sigma_{W_i}^2$, $i = 1, \cdots, L_2$. Note that ξ_i can be regarded to as a *noise to watermark ratio* for the i-th dimension. Function $\varphi(\alpha)$ above can be easily seen to be minimized at

$$\alpha^* = \frac{1}{1 + \frac{1}{L_2} \sum_{i=1}^{L_2} \xi_i}.$$

Alternatively, one may also consider using a different value of α for each dimension. This yields a vector of distortion compensation parameters $\boldsymbol{\alpha} \triangleq (\alpha_1, \cdots, \alpha_{L_2})$, so (3.10) takes now the shape

$$\varphi(\boldsymbol{\alpha}) = \sum_{i=1}^{L_2} \left\{ \frac{\alpha_i^2 \xi_i}{3} + \frac{(1-\alpha_i)^2}{3} \right\}, \tag{3.11}$$

where, as above, the noise to watermark ratio in the i-th coefficient, ξ_i, is kept fixed. The vector of distortion compensation parameters that minimizes (3.11) is given now by

$$\alpha_i^* \triangleq \frac{1}{1+\xi_i},$$

for all $i = 1, \cdots, L_2$. Clearly, $\varphi(\boldsymbol{\alpha}^*) \leq \varphi(\alpha^*)$, since the first minimization is a particular case of the second constrained to a vector with equal components.

It is possible to regard the distortion compensation effect of the vector case as a Wiener filtering with matrix $\mathbf{A}^* \triangleq \mathrm{diag}(\boldsymbol{\alpha}^*)$. This is so because all the self-noise elements corresponding to the components of $\boldsymbol{\alpha}^*$ are mutually independent, what

implies a diagonal filter. In fact, similar solutions have been proposed by Yu et al. in [164] from an information-theoretic point of view.

Finally, we would like to make some remarks. The performance improvement achieved by replacing α with $\boldsymbol{\alpha}$ is compatible with the gain due to using the decoding weights in (2.37). Whereas $\boldsymbol{\alpha}$ modifies the pdfs independently at each dimension, we will see in the next section that $\boldsymbol{\beta}$ modifies the weighting of the dimensions when they are considered together. This fact will be duly shown in Section 3.1.7.2.

3.1.2.2. Derivation of the Improved Decoding Weights

We turn next our attention to the problem of optimizing the weights introduced in (2.37). Recall that the objective of this approach is to improve the performance of the minimum distance decoder using additional knowledge about the channel distortion eventually available at the decoder.

Adapting the method followed in Section 3.1.1 to the decoder in (2.37), it turns out that now P_e can be written as

$$P_e = \Pr\left\{\sum_{i=1}^{L_2} \beta_i U_i^+ > \frac{1}{2}\sum_{i=1}^{L_2} \beta_i\right\},$$

which obviously reduces to (3.2) for $\boldsymbol{\beta} = \mathbf{1}$. Taking into account that any analytical optimization of the weights requires the availability of a closed-form approximation to P_e, we will discuss here the minimization of (3.7) and (3.8) when weights are introduced. Starting with the CLT-based approximation, which we will see that it is very accurate for low values of WNR in Section 3.1.7, and under the same assumptions as in Section 3.1.1.3, it is possible to write

$$P_e \approx P_{s_1} = \mathcal{Q}\left(\frac{\frac{1}{2}\sum_{i=1}^{L_2}\beta_i - \sum_{i=1}^{L_2}\beta_i \mathrm{E}\{U_i^+\}}{\sqrt{\sum_{i=1}^{L_2}\beta_i^2 \mathrm{Var}\{U_i^+\}}}\right). \tag{3.12}$$

Recalling that the $\mathcal{Q}(\cdot)$ function is monotonically decreasing, it follows that P_{s_1} is minimized when its argument is maximized. Then, the improved decoding weights can be found by differentiating the argument of $\mathcal{Q}(\cdot)$ in (3.12) with respect to $\beta_i, 1 \leq i \leq L_2$. Then, the decoding weights minimizing P_{s_1} are

$$\beta_i^* = K \cdot \frac{\left(\frac{1}{2} - \mathrm{E}\{U_i^+\}\right)}{\mathrm{Var}\{U_i^+\}},\ 1 \leq i \leq L_2, \tag{3.13}$$

where K is an irrelevant positive real constant, since the weights vector can be scaled without any impact on performance. Also, it is very interesting to note that some of the β_i^* may be negative. This will happen when $\mathrm{E}\{U_i^+\} > 1/2$,

which may occur for large distortions. The effect of a negative weight can be interpreted as a swapping of the centroids assigned to each symbol.

As it can be inferred from (3.13), in order to compute the improved decoding weights, knowledge of $\mathrm{E}\{U_i^+\}$ and $\mathrm{Var}\{U_i^+\}$ is required. Note that due to the aliasing and truncation effects that show up in the construction of $\mathbf{U}^+$, this information is not directly derivable from the first and second order moments of the total noise random variable.

3.1.2.2.1. High WNR

As we will see in Section 3.1.7 (Figure 3.6), the CLT-based approximation moves away from the empirical results as the WNR increases. In this case we can consider to use the union bound (3.8) to compute the improved decoding weights, since it is a better approximation to the P_e in the present scenario. Accordingly, the function that we have to minimize now is

$$P_e \approx P_{s_2} = 2^{L_2} \cdot \mathcal{Q}\left(\frac{\sum_{i=1}^{L_2} \beta_i}{2\sqrt{\sum_{i=1}^{L_2} \beta_i^2 \mathrm{Var}\{U_i\}}}\right), \tag{3.14}$$

which can be shown to be equivalent to the minimization of $\sum_{i=1}^{L_2} \beta_i^2 \mathrm{Var}\{U_i\}$ constrained to $\sum_{i=1}^{L_2} \beta_i = G$, for some arbitrary G. Applying Lagrange multipliers we may write the optimization functional as

$$\varphi(\boldsymbol{\beta}) = \sum_{i=1}^{L_2} \beta_i^2 \mathrm{Var}\{U_i\} - \lambda\left(\sum_{i=1}^{L_2} \beta_i - G\right). \tag{3.15}$$

Differentiating it with respect to β_i and equating to zero it is straightforward to see that the minimum of (3.14) is obtained for

$$\beta_i^{**} = K\frac{1}{\mathrm{Var}\{U_i\}},$$

for $1 \leq i \leq L_2$ and any positive constant K. Interestingly, it is possible to show analytically that for large WNRs $\boldsymbol{\beta}^*$ will be nearly proportional to $\boldsymbol{\beta}^{**}$, which justifies the use of $\boldsymbol{\beta}^*$ also for large WNRs in spite of the lossy approximation employed for its computation.

Notice that, after the optimal weights for the CLT-based approximations have been obtained, it is possible to resort to a more accurate computation of P_e (such as the Beaulieu's method or the DFT approach) by slightly modifying it to take the weights into account. The improvements afforded by $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$ will be empirically shown in Section 3.1.7.2.

3.1.3. A Geometric Interpretation of the Decoding Strategies

Here we provide a geometric interpretation of the various decoding strategies we have discussed, which will help to understand the role of the decoding weights and the goodness of Forney's approximation. For pictorial reasons, the case $L_2 = 2$ is considered here. First of all, we derive the ML decision boundary based on Forney's approach when $\sigma^2_{T_i}$ is large. Noticing that from (2.33) the true ML lattice decoding boundary is the locus of the points $(u_1, u_2)^T$ for which $f_U(u_1, u_2) = f_U((u_1 - 1) \mod 2\mathbb{Z}, (u_2 - 1) \mod 2\mathbb{Z})$, and making use of the approximation in Section 2.6.1, we can conclude that in the positive quadrant this boundary is approximately given by

$$\begin{aligned}\phi = \Big\{ & (u_1, u_2)^T \in [0,1] \times [0,1] : \\ & \Big(1 + 2\mathrm{e}^{-\pi^2 \sigma^2_{T_1}/2} \cos(\pi u_1)\Big) \cdot \Big(1 + 2\mathrm{e}^{-\pi^2 \sigma^2_{T_2}/2} \cos(\pi u_2)\Big) \\ & = \Big(1 - 2\mathrm{e}^{-\pi^2 \sigma^2_{T_1}/2} \cos(\pi u_1)\Big) \cdot \\ & \Big(1 - 2\mathrm{e}^{-\pi^2 \sigma^2_{T_2}/2} \cos(\pi u_2)\Big) \Big\}, \end{aligned} \tag{3.16}$$

with straightforward extensions to all other quadrants.

Figure 3.2 shows for the positive quadrant the true ML lattice decoder decision region for $\hat{b} = 0$ (shaded area) and the approximate decision boundary given by (3.16). The parameters of this plot are: $\sigma_{N_1}/\Delta_1 = 0.4113$, $\sigma_{N_2}/\Delta_2 = 0.2530$ and $\alpha = 0.5$, so $\sigma_{T_1} = 0.5025$ and $\sigma_{T_2} = 0.3838$. As it can be perceived, Forney's approximation gives a very good estimate of the real boundary. Figure 3.2 also plots the decision boundaries that result using (2.37) with $\boldsymbol{\beta} = \mathbf{1}$, and $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, which with the above parameters becomes $\beta^*_1 = 1.5936$ and $\beta^*_2 = 3.9005$. Observe how the use of β^*_i leads to a linear approximation of the true lattice ML decision boundary. Note however, that the ultimate purpose of the weights $\boldsymbol{\beta}^*$ is not to yield the best linear approximation of this boundary but to minimize an approximation of the bit error probability.

3.1.4. Discussion about the Pseudorandom Choice of the Partitions

Throughout this section we have been assuming that the samples comprising the j-th host subvector $\mathbf{X}_j$ were pseudorandomly chosen. In this subsection we will try to theoretically justify the use of such pseudorandom assignment to minimize the overall P_e when L_2 is large enough, starting from our CLT-based approximations, and using the law of large numbers. Furthermore, a empirical justification will be provided.

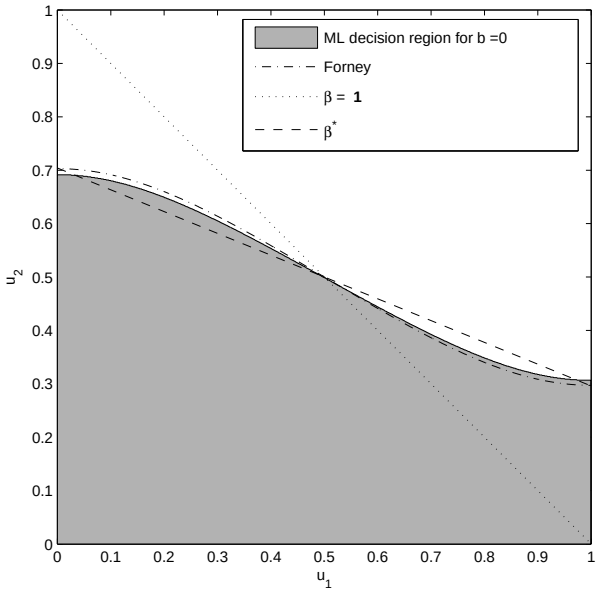


Figure 3.2: Comparison of the decision regions for DC-DM ($L_2 = 2$) obtained using Forney's approximation, the ML lattice decoder, and the Euclidean distance decoder with $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta} = \mathbf{1}$.

We will use in our proof the minimum weighted Euclidean distance lattice decoding, i.e. (2.37), when the improved decoding weights $\boldsymbol{\beta}^*$ computed in Section 3.1.2.2 are used; remember that we have shown that this $\boldsymbol{\beta}^*$ can improve the decoding. Therefore, using the CLT-based approximation and replacing (3.13) in (3.12), we can write the probability of decoding error for the j-th hidden bit hidden, P_e^j, as

$$P_e^j \approx \mathcal{Q}\left(\sqrt{\sum_{k=(j-1)\cdot L_2+1}^{j\cdot L_2} \frac{\left(\frac{1}{2} - \mathrm{E}\{U_k^+\}\right)^2}{\mathrm{Var}\{U_k^+\}}}\right), \quad j \in \{1, \cdots, L_b\}. \tag{3.17}$$

The overall probability of decoding error P_e will be just the average of all P_e^j, $j = 1, \cdots, L_b$, and it can be upperbounded using Chernoff bounding as

$$P_e \leq \frac{1}{L_b} \sum_{j=1}^{L_b} \mathrm{e}^{-\sum_{k=(j-1)\cdot L_2+1}^{j\cdot L_2} \gamma_k/2}, \tag{3.18}$$

with $\gamma_k \triangleq \left(\frac{1}{2} - \mathrm{E}\{U_k^+\}\right)^2 / \mathrm{Var}\{U_k^+\}$. Notice that, as $\gamma_k > 0$ for all k, the right-hand side of (3.18) will be always lower than 1. Since the γ_k's will be different for each dimension, they can be thought as realizations of a mean-ergodic random variable. Defining next $\theta_j \triangleq \sum_{k=(j-1)\cdot L_2+1}^{j\cdot L_2} \gamma_k/2$, the problem of choosing the best

partitions can be approximately formulated as the minimization of

$$\sum_{j=1}^{L_b} \mathrm{e}^{-\theta_j} \tag{3.19}$$

constrained to

$$\sum_{j=1}^{L_b} \theta_j = \sum_{j=1}^{L_b} \sum_{k=(j-1)\cdot L_2+1}^{j\cdot L_2} \gamma_k/2 = \sum_{k=1}^{L_b} \gamma_k/2 = K. \tag{3.20}$$

This constrained optimization problem can be solved by building the functional

$$\varphi(\boldsymbol{\theta}) \triangleq \sum_{j=1}^{L_b} \mathrm{e}^{-\theta_j} - \lambda \left(\sum_{j=1}^{L_b} \theta_j - K \right), \tag{3.21}$$

using the Lagrange multiplier λ. Differentiating (3.21) with respect to θ_j and equating to zero, we get

$$\frac{\partial \varphi(\boldsymbol{\theta})}{\partial \theta_j} = -\mathrm{e}^{-\theta_j} - \lambda = 0, \quad \text{for all } j = 1, \cdots, L_b, \tag{3.22}$$

and, consequently, the optimal θ_j has to be constant over all $j = 1, \cdots, L_b$; be aware that this result implies $\theta_j \geq 0$, so it is not neccesary to consider the Kuhn-Tucker multiplier corresponding to that constraint in (3.21).

Equivalently

$$\sum_{k=(j-1)\cdot L_2+1}^{j\cdot L_2} \gamma_k = C, \quad \text{for all } j = 1, \cdots, L_b, \tag{3.23}$$

for a given constant C. Invoking the law of large numbers, a simple way to achieve this condition for a large enough value of L_2 is by pseudorandomly choosing the indices in $\mathbf{x}_j$. Therefore, accepting the limitations of the CLT-based approximation and the Chernoff bounding, the pseudorandom selection of the subvectors seems to be a good choice for large values of L_2.

Nevertheless, taking into account the last reasons, the former justification is not so theoretically rigorous as it could be desirable, and an empirical justification will be also provided.

With this aim, we will consider the particular case of applying DC-DM watermarking to an image on the mid-frequencies of its 8×8-block DCT, the transform used in the *Joint Photographic Experts Group* (JPEG) standard [3]. Moreover, we will let the channel noise variance be proportional to the squared JPEG quantization step (quality factor QF = 80) in each dimension, being this noise uniform.

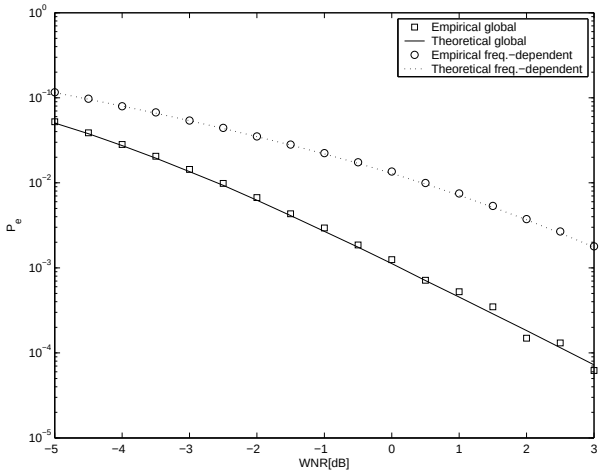


Figure 3.3: Empirical and theoretical performance obtained with global vs. frequency-dependent pseudorandom partitions, using DC-DM on the DCT domain with optimally weighted Euclidean distance decoding. $L_2 = 20$, $\alpha = 0.4$, uniform noise, host image Lena 256×256, payload $L_b = 1126$ bits.

This quality factor is a scalar ranging from 0 (poor quality) to 100 (high quality) used by some implementations of the JPEG compression algorithm to indicate the quantization table. We have chosen this attack because it is assumed to have a perceptually-based power distribution (as JPEG quantization steps stem from perceptual considerations), although it does not follow the same power allocation as the watermark. We will consider two cases for defining the subvectors $\mathbf{X}_j$: global pseudorandom partitions (i.e., all available coefficients in the same pool), and frequency-dependent pseudorandom partitions (i.e., each pool consists of those coefficients with the same frequency indices that come from different blocks); in both cases the permutation was changed at each realization of the experiment.

This last strategy resembles the one used by Ramkumar and Akansu in [138] as well as the parallel channels studied by Moulin [121] applied to the DCT domain. In the former work, the data hiding capacity of compressed images is analyzed by decomposing an image into L_b subbands using transform blocks, thus giving rise to L_b parallel subchannels. Then, each symbol is only transmitted through a specific subchannel. With that strategy, all the coefficients devoted to conveying a certain symbol can be assumed to have the same noise statistics, differently to what happens when the indices are chosen pseudorandomly.

In Figure 3.3 the improvement due to the use of global pseudorandom partitions is shown, choosing the mid-frequencies as in [93] and using the same perceptual mask and attack as in Section 3.1.7.2. The theoretical results were obtained using the DFT method. It is important however to note that a fixed

subvector length has been assumed in this comparison, which clearly puts the frequency-dependent scheme at disadvantage, because each subchannel will have different host and noise statistics and, thus, different SNR's. A solution to this is to use subvector lengths that are also frequency-dependent, at the price of needing additional knowledge about the channel at both embedder and decoder, something that is not required when global partitions are used. Additionally, global pseudorandom partitions may increase the entropy of the watermark and hence the security of the system.

3.1.5. Comparison with STDM

In this section we will compare the performance of DC-DM with repetition coding, with the STDM-like methods. As shown in [64] and later confirmed in [132], STDM-like methods show superior performance than DC-DM in AWGN channels as the repetition L_2 (and, equivalently, the spreading factor L_1/L_3, which coincides with L_2 when $L_3 = L_b$, i.e. when the embedding in the projected domain is performed by a scalar quantizer) increases; this was already briefly discussed in Section 2.7.1, showing the gain of L_1/L_3 in the effective WNR in the projected domain, and is now experimentally confirmed in Figure 3.4 using real images as host data. The watermark is embedded in the mid-frequency coefficients of the 8×8 block-DCT domain [93] with a fixed Peak Signal to Noise Ratio (PSNR) of 40 dB, and uniform noise is added with standard deviation proportional to the corresponding JPEG quantization step in each dimension (quality factor QF=80). The figure also shows theoretical results, obtained using the CLT method in Section 3.1.1.3 for DC-DM and [132] for SSTDM. We observe a large gap between both methods for high PSNRs, but it is necessary to take into account that $\alpha = 0.4$ used in the plot is not the optimal one when the PSNR of the attacked signal is close to 40 dB (large WNR). The optimal projection parameters $\boldsymbol{\beta}^*$ for SSTDM in Figure 3.4 will be derived in Section 3.5.5, even though other optimization strategies are available (see for instance [87]).

3.1.6. Performance under Unforeseen Attacks

An interesting problem is posed by the performance analysis of DC-DM when the attack is different than the one expected by both the embedder and the decoder. We remind that the available information about the attack is exploited by them to compute, respectively, the optimal distortion compensation parameter and the optimal decoding weights. The general problem should be addressed from a game-theoretic approach, trying to find the optimal attack and the optimal encoding/decoding strategies, using a bit error rate payoff in our case. A first approach to this problem will be introduced in Section 3.5.4, where the players

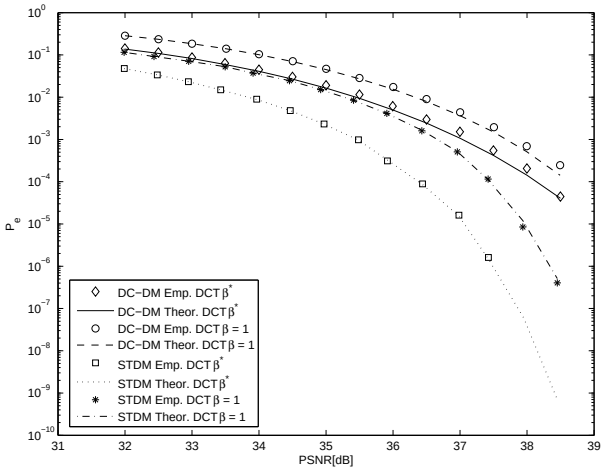


Figure 3.4: Empirical and theoretical performance of DC-DM ($\alpha = 0.4$) vs. scalar STDM, watermarking the DCT domain of real images and attacking with additive uniform noise; results averaged over twenty-two 256×256 images, with $L_2 = 20$ (payload $L_b = 1126$ bits).

of the game are only the attacker and the decoder, whereas the encoder strategy is fixed.

In any case, it is interesting to observe the performance degradation when there is a mismatch between the actual attack and the one considered when optimizing the method. In Figure 3.5 experimental results for this case are shown for a particular case in which DC-DM is applied to the Lena image in the spatial domain, and the embedder and decoder expect uniform noise in the same domain. However, the noise is added both in the spatial and DCT domains. In order to set realistic conditions, the uniform noise in the DCT domain has, at each coefficient, variances proportional to a squared perceptual mask computed following Watson [160]. Although it can be verified that the energy distribution of the corresponding inverse transformed noise in the spatial domain differs considerably from the spatial perceptual mask, we may see that there is only a small performance difference (in fact a gain) with respect to the ideal case where the noise follows the expected distribution.

3.1.7. Empirical Results

In this section we will check the validity of our theoretical developments, comparing the analytical results with empirical ones.

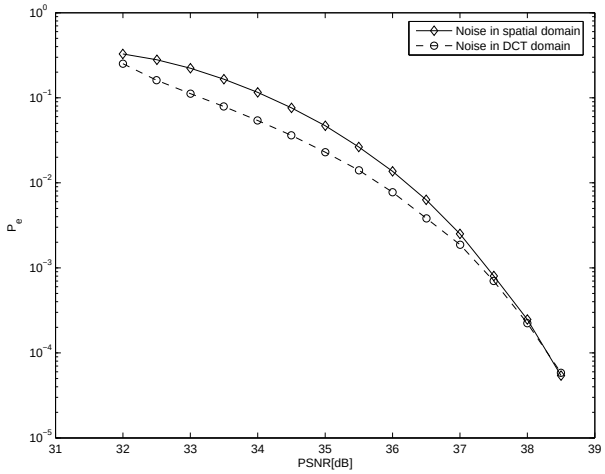


Figure 3.5: Experimental performance of DC-DM in the spatial domain under uniform additive noise applied in the spatial and DCT domain, with $\alpha = 0.4$, $L_2 = 20$ and optimal decoding weight computed taking into account the noise in the spatial domain (payload $L_b = 1126$ bits); results averaged over twenty-two 256×256 images.

3.1.7.1. Comparison of the Approximations and Bounds

Figure 3.6 shows the approximations and bounds in Section 3.1.1 versus the outcomes of i.i.d. Montecarlo simulations. In this plot channel noise is additive zero-mean Gaussian, the components of $\mathbf{X}$ and $\mathbf{N}$ are i.i.d., $L_2 = 10$ and α is optimized following Costa's formula, i.e., $\alpha = \alpha_c \triangleq D_w/(D_w + D_c)$. We may verify that the accuracy of the approximations given in Sections 3.1.1.1 and 3.1.1.2 is remarkable. The CLT-based approximation is excellent for low values of the WNR, but, as the WNR is increased, it gets away from the true probability of error. As it was explained in Section 3.1.1.3, this is due to the support of $f_{U_i^+}(u_i^+)$ being only positive, to the small value of L_2 used in the experiment, and to the increase in the skew-effect of the resulting pdf for large values of the WNR. Since this approach underrates the importance of the tails of $f_R(r)$, the approximation produces overly optimistic results.

On the other hand, the union bound gets closer to the empirical results when the WNR increases. This is a consequence of the reduction of the probability corresponding to the overlapped decision regions when the WNR grows. We also have plotted the results of applying the CLT to compute the probability of error with only one neighbor and then using the union bound, as described in Section 3.1.1.4.2. In this case the pdf involved in the computation is symmetric about the origin, so convergence to the Gaussian distribution is unaffected when the WNR is increased. Note that both bounds approach the true probability of bit error asymptotically as the WNR increases. The values predicted by the

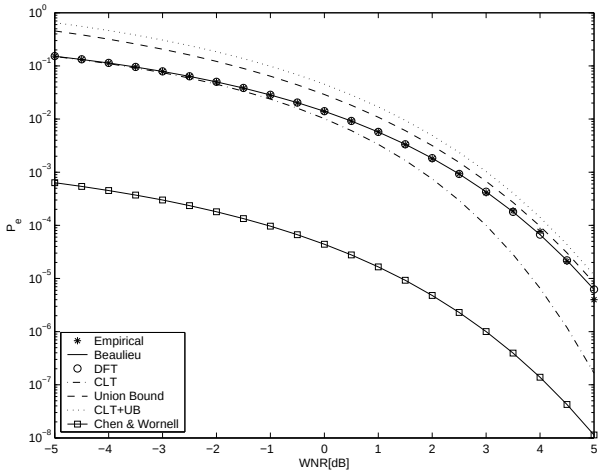


Figure 3.6: Comparison of the empirical BER vs. the different analytical and numerical approximations and bounds for DC-DM under Gaussian noise. $L_2 = 10$, $\alpha = \alpha_c$. Synthetic host data.

approximation of Chen and Wornell are obviously parallel to those obtained when both the union bound and the CLT are used (see Section 3.1.1.4.2). As it should be expected, those values are clearly lower than the empirical results, since only the probability of mistaking two neighbors is taken into account.

Finally, the bound by Erez and Zamir is not shown in Figure 3.6 because its value is around 10^3. It is pertinent to remark here that even though this bound is valid for any pair of nested lattices, it was designed to show the capacity-achieving property of lattice decoding. Nevertheless, for that purpose, it is necessary that the pair of nested lattices verify certain properties which fall short of being true for the lattices used by DC-DM. This explains why such large values arise and demonstrates how information-theoretic results cannot always be effortlessly extrapolated to practical schemes.

3.1.7.2. Optimized Distortion Compensation Parameter and Improved Decoding Weights

The next set of experiments were carried out by watermarking the image *Lena* 256×256 in the DCT domain, using a perceptual mask proportional to the perceptual thresholds proposed in [160] and [6]. The attack is uniformly distributed with amplitude proportional in each dimension to the corresponding JPEG quantization step for QF=80.

Figure 3.7 shows the performance improvements due to the use of the weights $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$ in the Euclidean distance decoder. The plot depicts the WNR needed

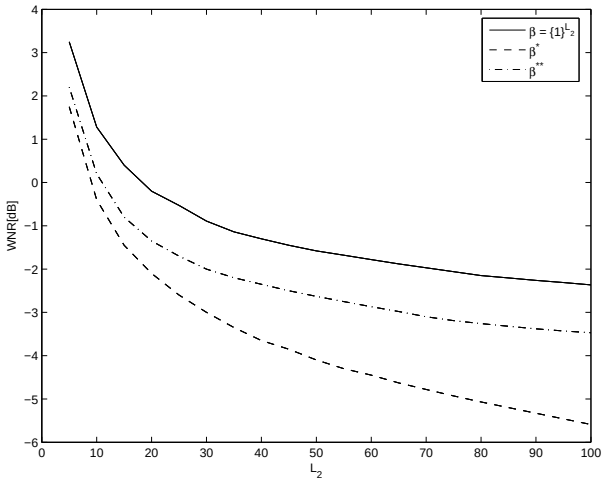


Figure 3.7: DC-DM watermarking of the Lena 256×256 host signal in the DCT domain with uniformly distributed additive attack. WNR needed to achieve $P_e = 0.01$ vs. L_2 for $\alpha = 0.7$, with different weightings on the Euclidean distance decoder.

to achieve $P_e = 0.01$ with L_2 ranging from 5 to 100, and clearly shows the improvement obtained when $\boldsymbol{\beta}^*$ is used. The performance gain is already large at $L_2 = 100$, but the gap keeps increasing with L_2. Nevertheless, the improvement is not so large when $\boldsymbol{\beta}^{**}$ is used. In order to explain this effect, consider that the WNR's studied are rather negative, and therefore that the CLT-based approximation used for the computation of $\boldsymbol{\beta}^*$ is clearly better than the union bound plus CLT expression used for the computation of $\boldsymbol{\beta}^{**}$ (see Section 3.1.1.4.2 and cf. Figure 3.6).

Figure 3.8 shows the results obtained when α^* and $\boldsymbol{\alpha}^*$ are used in conjunction with $\boldsymbol{\beta} = \mathbf{1}$ (i.e., no weighting), $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$, for the case $L_2 = 10$. A considerable gain is achieved by using a vectorial distortion compensation parameter $\boldsymbol{\alpha}^*$ instead of a scalar one, α^*. The improvement due to using $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$ compared to no weighting is also apparent. Note also that the weighting strategy $\boldsymbol{\beta}^*$ yields the best results for the whole range considered in this case. Finally, as we have pointed out in Section 3.1.2.1, the use of a distortion compensation vector is compatible with the improved decoding weights, so the combination offers improvements of about 2 dB over the standard embedding/decoding strategy.

In Figure 3.9 we compare ML lattice decoding versus Euclidean distance decoding weighted by $\boldsymbol{\beta}^*$. The theoretical results for $\boldsymbol{\beta}^*$ in that figure were computed employing the DFT method. This plot clearly shows the near-optimality of performing Euclidean distance decoding with our optimal weighting strategy, since the results obtained are virtually the same than those obtained with ML lattice decoding. This result can be explained (at least for small values of WNR,

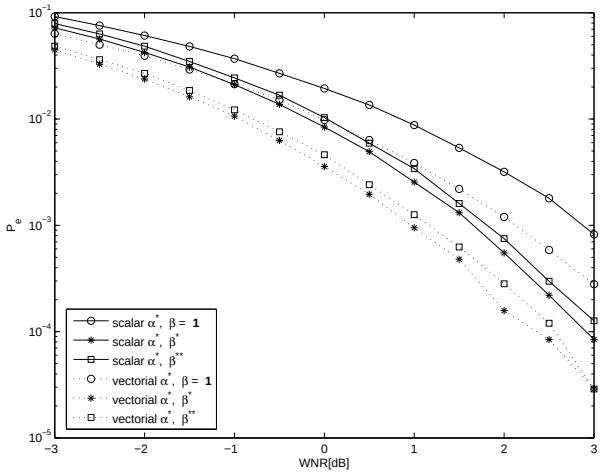


Figure 3.8: DC-DM watermarking of the Lena 256×256 host signal in the DCT domain with uniformly distributed additive attack. Comparison of the empirical BER obtained when α^* or $\boldsymbol{\alpha}^*$ are used in conjunction with $\boldsymbol{\beta} = \mathbf{1}$, $\boldsymbol{\beta}^*$, and $\boldsymbol{\beta}^{**}$. $L_2 = 10$, payload $L_b = 2252$ bits.

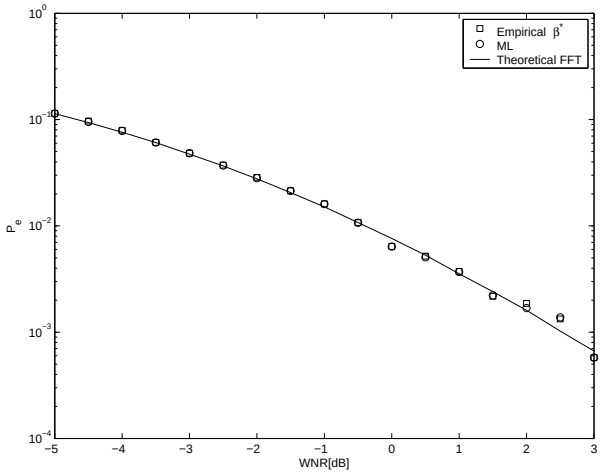


Figure 3.9: Performance of DC-DM using ML lattice decoding vs. using Euclidean distance decoding using the optimal weights $\boldsymbol{\beta}^*$. Gaussian noise with variance proportional to the squared JPEG quantization step (QF = 80), $\alpha = 0.5$, $L_2 = 10$, payload $L_b = 2252$ bits, host image Lena 256×256.

where Forney's approximation is valid) in view of the resemblance between the decision regions used by these two decoders (see Section 3.1.3). It is interesting to remark that, as the variance of the host signal is much larger than that of the watermark, adjacent DC-DM centroids have similar probabilities, and then ML lattice decoding approaches ML decoding.

3.1.7.3. Comparison with Miller et al.'s Trellis-based Embedding

We compare next DC-DM to the side-informed algorithm based on trellis quantization presented in [115]. In order to undertake the comparison, we encoded DC-DM using the cascade of an outer code, given by two serially-concatenated codes [22] with global rate 1/4, with an inner 1/3 repetition code, obtaining the same overall coding rate 1/12 used in [115]. The use of channel coding is necessary in order to make a fair comparison, since the method in [115] inherently includes an involved (source) code. Admittedly, the comparison will be dependent on the particular codes used in each case, but we may get in this way an acceptable perspective of the relative performance of both methods.

In order to set the same test conditions, DC-DM embedding is performed with the same image and using the same DCT coefficients as in [115], and hence the payload is also $L_b = 1380$ bits. Similarly, the same Watson-based perceptual constraints [160] are taken into account, and the Watson measure due to the DC-DM watermark is fixed to 27.20 as in [115]. Our experiments show that $P_e \approx 10^{-3}$ for DC-DM when the standard deviation of the additive noise is 8.5, marking the region of the turbo-cliff in the iteratively decoded DC-DM scheme. For the same noise power, Miller et al.'s method yields $P_e \approx 3.3 \times 10^{-3}$. Thus, both techniques exhibit similar performance under this very specific scenario.

3.2. DC-DM Performance under Coarse Quantization

In this section we will analyze the performance of DC-DM when the watermarked signal $\mathbf{Y}$ undergoes coarse quantization, which is quite a common unintentional attack. Notice that we cannot deal with this particular attack using the generic methods presented in Section 3.1, as in this case we cannot assume the independence of the channel noise (actually the coarse quantization error). Furthermore, our analysis will serve to show how to improve the performance of DC-DM under this particular attack.

We assume next that a coarse quantizer with centroids given by the lattice $\delta_i\mathbb{Z}$ is applied to y_i for all $1 \leq i \leq L_2$. The computation of the probability of decoding error relies on knowing the probability mass function (pmf) of Z_i. Notice that this pmf will not only depend on the pdf of the host image, but also on that of the watermark, which in turn depends on the transmitted bit b and on the dither d_i. In order to obtain the desired probability we need the upper and lower limits of the k-th coarse-quantization bin, which will be denoted by $\theta_{i_k}^+ \triangleq k\delta_i + \delta_i/2$ and $\theta_{i_k}^- \triangleq k\delta_i - \delta_i/2$, respectively. So, the probability that Z_i is equal to the k-th

coarse-quantization centroid conditioned to the transmission of b is

$$\begin{aligned}\Pr\{Z_i = k\delta_i \mid b\} &= \Pr\{Y_i \in (\theta_{i_k}^-, \theta_{i_k}^+] \mid b\} \\ &= \int_{\theta_{i_k}^-}^{\theta_{i_k}^+} f_{Y_i}(y_i|b)dy_i. \end{aligned} \tag{3.24}$$

We are interested in reformulating this integral in terms of X_i, what requires a change of variable affecting the integration limits of the expression. This change of variable is not evident, but it can be obtained in a straightforward manner. First, notice that the DC-DM centroid corresponding to the symbol b and closest to the upper limit $\theta_{i_k}^+$ of the integral (3.24) is just $Q_b(\theta_{i_k}^+)$, with $Q_b(\cdot)$ defined in (2.23). Then, considering the offset $\rho_y(\theta_{i_k}^+, b) \triangleq \theta_{i_k}^+ - Q_b(\theta_{i_k}^+)$, it can be shown that the corresponding offset with respect to $Q_b(\theta_{i_k}^+)$ from the point of view of X_i is

$$\rho_x(\theta_{i_k}^+, b) \triangleq \frac{\min\{\max[\rho_y(\theta_{i_k}^+, b), -(1-\alpha)\Delta_i], (1-\alpha)\Delta_i\}}{(1-\alpha)}. \tag{3.25}$$

Therefore, the upper limit when the integral in (3.24) is evaluated using $f_{X_i}(x_i)$ is just $\gamma_{i_k}^+(b) \triangleq Q_b(\theta_{i_k}^+) + \rho_x(\theta_{i_k}^+, b)$. The lower limit $\gamma_{i_k}^-$ can be obtained similarly, and then the desired probability can be put as

$$\Pr\{Z_i = k\delta_i \mid b\} = \int_{\gamma_{i_k}^-(b)}^{\gamma_{i_k}^+(b)} f_{X_i}(x_i)dx_i. \tag{3.26}$$

This pmf plays a similar role as the pdf $f_{T_i}(\cdot)$ in (2.30). Hence, the probability of decoding error under coarse quantization can be obtained by applying to this pmf the same modular strategy used in Section 3.1.1. Unfortunately, the resulting expression is quite involved and it has to be computed numerically in practice.

Notice that the probability of error thus obtained will be in general dependent on b. A side-effect of this dependence is that the weights optimization in Section 3.1.2.2 is not valid for coarse quantization in general. Actually, the improved decoding weights β_i^* will only be valid for symmetric settings. In section 3.2.2 we will compare the performance under coarse quantization using two kinds of dithers. For the first one we choose $d_i \in \{\pm\Delta_i/2\}$, for all $i = 1, \cdots, L_2$. Due to symmetry, in this case the statistics for each dimension are independent of the embedded bit, and the procedure to compute the decoding weights can still be used. For the second one, $d_i \in \{0, \Delta_i\}$ for all $i = 1, \cdots, L_2$, which does not give a symmetric setting. With this choice, the statistics in each dimension do depend on the embedded bit, thus making it impossible to derive the aforementioned weights. Be also aware that the watermark power in both cases has not to be the same. In fact, in the asymptotic case, when the host distribution goes to a Dirac's delta, the power of the watermark in the first case is $\Delta^2/4$, but in the last one $\Delta^2/2$; in any case, with more realistic (in the sense of smoother) host distributions the difference is not going to be so large.

3.2.1. JPEG Compression

We may particularize the expression (3.26) for a real coarse quantization case such as the one induced by the popular JPEG standard for image compression [3]. Accordingly, let us assume throughout this subsection that the host signal is given in the 8×8 block-DCT domain where JPEG works. As discussed in [93] the AC coefficients of the DCT can be reasonably modeled by zero-mean generalized Gaussian pdfs, given by the expression

$$f_X(x) = A\mathrm{e}^{-|\eta x|^c}. \tag{3.27}$$

The parameters A and η can be expressed as a function of the shape parameter c and the standard deviation σ_X. We refer the reader to [93] for the details on how to tackle in practice the issue of their estimation. Taking into account the model (3.27), and assuming that its parameters are estimated adaptively for each dimension, we may rewrite (3.26) as $\Pr\{Z_i = k\delta_i \mid b\} = \Pr\{X_i \leq \gamma_{i_k}^+(b)\} - \Pr\{X_i \leq \gamma_{i_k}^-(b)\}$, with

$$\Pr(X_i \leq \tau) = \begin{cases} \frac{A_i}{\eta_i c_i}\Gamma(1/c_i, |\eta_i\tau|^{c_i}), & \text{if } \tau \leq 0 \\ 1 - \frac{A_i}{\eta_i c_i}\Gamma(1/c_i, |\eta_i\tau|^{c_i}), & \text{if } \tau > 0 \end{cases},$$

where $\Gamma(\cdot,\cdot)$ is the incomplete Gamma function[1].

Figure 3.10 tries to provide an intuitive idea of the reasoning followed in this section.

3.2.2. Empirical Results

We compare next in Figure 3.11 the performance of DC-DM under the coarse quantization attack given by JPEG compression, using the symmetric and asymmetric dithers discussed in Section 3.2. In the plot, the probability of bit error is plotted versus the quality factor QF used to compress the watermarked signal *Lena* using JPEG. Embedding takes place in the 8×8 block-DCT domain. In order to obtain the theoretical results we have used the CLT-based approximation and assumed a Laplacian distribution for the host signal, which corresponds to $c = 1$ in (3.27). This approximation explains the small discrepancies between the theoretical and empirical results, which are more evident for $\boldsymbol{\beta}^*$ as the convergence of the decision statistic to a Gaussian is slower with weighting. As it can be seen, the use of an asymmetric dither yields superior performance, even considering that it is not possible to use the optimal weights in this case.

[1] $\Gamma(a,z) \triangleq \int_z^\infty t^{a-1}\mathrm{e}^{-t}dt$.

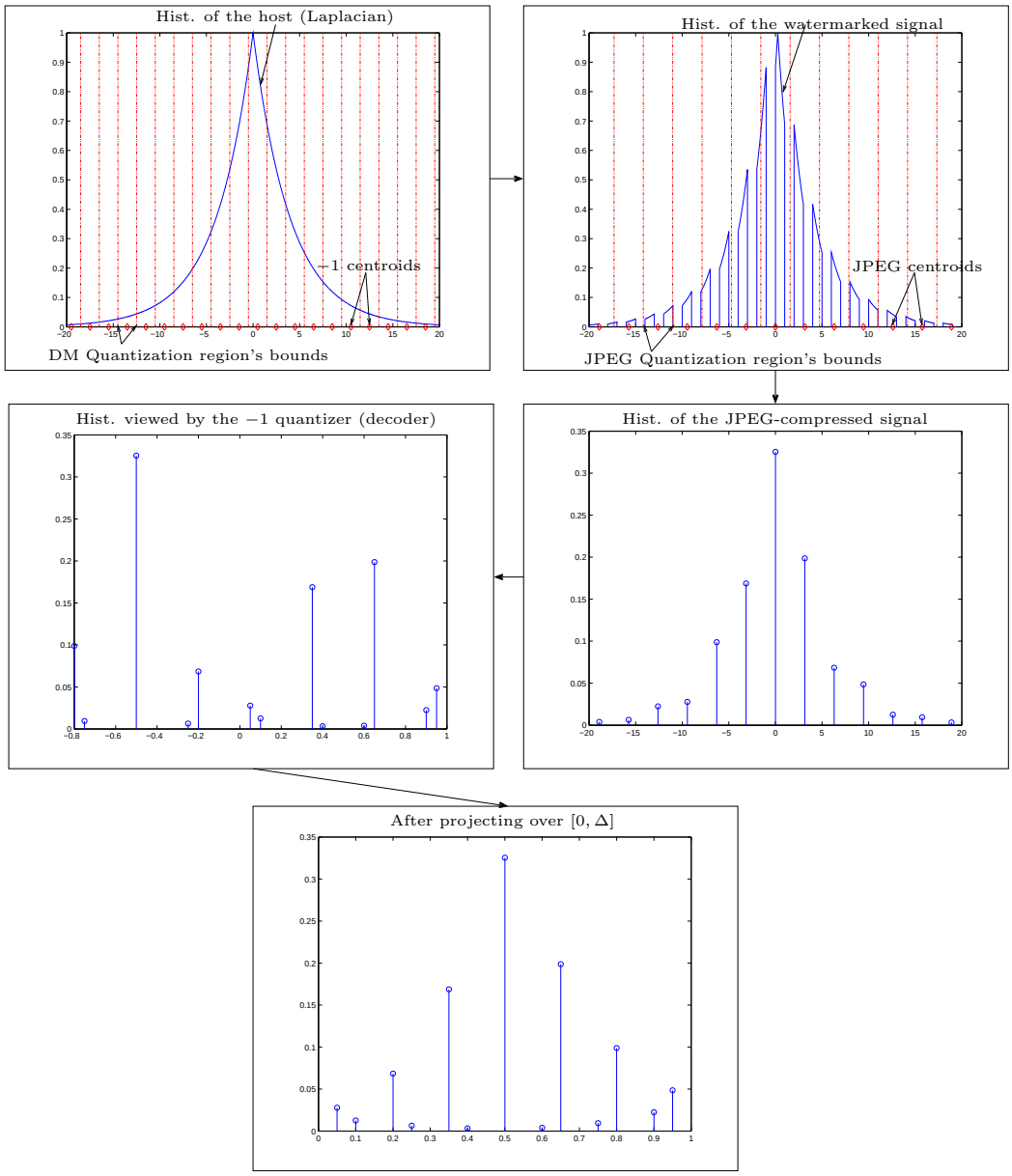


Figure 3.10: Computation of the JPEG-equivalent noise when a -1 is hidden.

3.3. Cropping Attack

As it was discussed in Section 2.7.1, SSTDM achieves the same spreading gain as Add-SS by relying on the projection stage of the latter, so that both become robust to additive attacks. In this section, we will show that the similarities between Add-SS and SSTDM do not extend to the popular cropping attack (i.e., the removal of some components of the watermarked signal), since when the removed area increases, performance degrades smoothly for the former but steeply for the latter. One arrives at this conclusion after comparing the analytical expressions

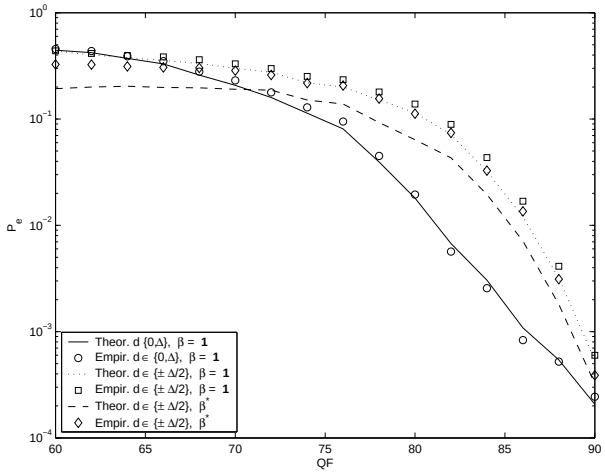


Figure 3.11: Empirical and theoretical performance when *Lena* is watermarked with DC-DM in the DCT domain and JPEG-compressed with quality factor QF, for symmetric and asymmetric dithers. $L_2 = 10$, payload $L_b = 2252$ bits, $\alpha = 0.5$.

derived for both methods. Increasing the size of the projected subspace and using a cubic lattice with repetition coding is shown to be a possible solution to cope with cropping while retaining host interference cancellation. Our main results, presented in Section 3.3.1, also apply to other relatively similar impairments, such as line-removal and block-replacement attacks.

3.3.1. Describing the Cropping Attack

For the sake of notational simplicity, our presentation in this subsection will be restricted to any of the L_2-dimensional subvectors $\mathbf{x}_j$, dropping the subindex j. The cropping attack removes some components of the watermarked image $\mathbf{y}$, reducing the size of the received signal. Let $\mathcal{I}$ and $\bar{\mathcal{I}}$ respectively denote the set of removed components indices and its complement, i.e., $\bar{\mathcal{I}} = \{1, \cdots, L_2\} \backslash \mathcal{I}$. The decoder receives the $|\bar{\mathcal{I}}|$-length signal $\mathbf{z}$ from which it obtains an L_2-length estimate $\hat{\mathbf{y}}$ of the watermarked signal[2] in such a way that $\hat{\mathbf{y}} = \mathbf{y} + \mathbf{n}$, where $\mathbf{n}$ is a noise vector (which can be seen as realization of random vector $\mathbf{N}$). The decoding process will not take into account the received incomplete signal $\mathbf{z}$, but its *restored* version, i.e. the estimate of the watermarked signal $\hat{\mathbf{y}}$, so the decoding strategy will be typically given by

$$\hat{b} = \underset{b \in \{0, P-1\}}{\arg\min} \, ||\hat{Y}_p - Q_b(\hat{Y}_p)||,$$

where $\hat{Y}_p = \mathbf{s}^T \hat{\mathbf{Y}}$.

[2]This estimation makes sense whenever parts of the watermarked signal are completely lost, as it is the case for the cropping atttack.

In spite of producing a large distortion in a mean squared error sense, cropping does not result in any perceptual distortion provided that significant parts of the host signal are not removed, e.g. by removing a thin frame from an image and leaving the rest unaltered. Taking this into account, a meaningful measure of distortion is to consider only those coefficients that have survived cropping; hence, it is convenient to define $D_{c_{\text{remain}}} = \frac{1}{|\bar{\mathcal{I}}|}\sum_{i\in\bar{\mathcal{I}}} \mathrm{E}\{(Z_i - Y_i)^2\}$, so the Watermark to Noise Ratio (WNR) is redefined as $\mathrm{WNR}_{\text{remain}} = D_w / D_{c_{\text{remain}}}$.

Notice that cropping also causes a synchronization problem; however, this is not only characteristic of SSTDM, but of *any* data hiding scheme. For this reason, we will assume that proper measures have been taken to ensure synchronization; further discussions about this topic can be found in [15], [114] and [148]. Hence, under perfect synchronization, the main complication caused by cropping is the loss of some coefficients of the watermarked signal, so the decoder will have to decide which symbol was embedded by taking into account the available prior information about those coefficients. In this way, the decoder will see the cropping attack as the addition of the random variable $N_j = \hat{Y}_j - Y_j$, $j \in \mathcal{I}$, where $\hat{y}_j$ is an estimate of the unkown watermarked coefficient y_j. For those indices $j \in \bar{\mathcal{I}}$, we will take $\hat{y}_j = z_j$, as it is typically done, so $N_j = Z_j - Y_j$. Notice that, as long as a stochastic characterization is available, this model also encompasses any kind of restoration attempt based on extrapolation from the available coefficients.

Next, we analyze the impact of the cropping attack on Add-SS and SSTDM, taking the probability of error P_e as a measure of performance. For the sake of simplicity, and in order to obtain interpretable results, throughout this section we make the following assumptions: 1) binary (i.e., $P = 2$) signaling; 2) projection vector $\mathbf{s}$ is such that $s_i \in \{\pm 1\}$, for all $1 \le i \le L_2$; 3) the attack is limited to cropping, so $z_i = y_i$, for all $i \in \bar{\mathcal{I}}$; and 4) $\hat{\mathbf{Y}}$ is an unbiased estimate of $\mathbf{Y}$, and the components of both $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are mutually independent. As it was discussed in Section 2.1, the latter assumption is reasonable if the projected host features are pseudorandomly chosen over the full image, as this choice practically eliminates local dependencies.

3.3.1.1. Performance Analysis of Cropping Attack on Add-SS

When host samples are modeled by a zero-mean independent Gaussian process and $\boldsymbol{\gamma} = \mathbf{1}$[3], the usual decoder is given by

$$\hat{b} = \begin{cases} 0, & \text{whenever } \mathbf{s}^T \cdot \hat{\mathbf{y}} < 0 \\ 1, & \text{whenever } \mathbf{s}^T \cdot \hat{\mathbf{y}} > 0 \end{cases},$$

[3]The last hypothesis, i.e. $\boldsymbol{\gamma} = \mathbf{1}$, is considered in this section for the sake of notational simplicity of the obtained results.

then taking $\hat{y}_i = 0, i \in \mathcal{I}$, the probability of error when $|\mathcal{I}|$ samples are cropped, can be written as

$$P_e(|\mathcal{I}|) = \mathcal{Q}\left(\frac{|\bar{\mathcal{I}}|}{\sqrt{\sum_{j\in\bar{\mathcal{I}}}\sigma^2_{X_j}}}\right).$$

Therefore, if we assume that any coefficient has a probability ξ of being removed, and for the case $\sigma^2_{X_i} = \sigma^2_X$, $i = 1, \cdots, L_2$, we can average over the set of possible croppings, so

$$P_e = \sum_{j=0}^{L_2}\binom{L_2}{j}\xi^j(1-\xi)^{L_2-j}\mathcal{Q}\left(\frac{\sqrt{L_2-j}}{\sigma_X}\right),$$

which, according to the Law of Large Numbers, for large values of L_2 can be approximated by

$$P_e \approx \mathcal{Q}\left(\frac{\sqrt{L_2}\sqrt{1-\xi}}{\sigma_X}\right). \tag{3.28}$$

3.3.1.2. Performance Analysis of Cropping Attack on SSTDM

In [132] the probability of error for SSTDM was shown to be

$$P_e = 1 - \sum_{k=-\infty}^{\infty}\int_{-\Delta/2+2k\Delta}^{+\Delta/2+2k\Delta} f_{N_p}(\tau)d\tau, \tag{3.29}$$

where $N_p \triangleq \mathbf{s}^T \cdot \mathbf{N}$ denotes the projected noise random variable. An upper bound is obtained by considering just the two nearest centroids to the desired one [132]:

$$P_e \leq P_{e,u} = 1 - \int_{-\Delta/2}^{+\Delta/2} f_{N_p}(\tau)d\tau$$

When N_p is Gaussian (see Section 2.7.1) the upper bound evaluates to

$$P_{e,u} = 2\mathcal{Q}\left(\frac{\Delta/2}{\sqrt{\sum_{i=1}^{L_2}\mathrm{Var}\{N_i\}}}\right). \tag{3.30}$$

When the random variables N_i are i.i.d. with variance σ^2_N for $i \in \mathcal{I}$, and zero otherwise, equation (3.30) must be averaged over the possible croppings:

$$P_{e,u} = 2\sum_{j=1}^{L_2}\binom{L_2}{j}\xi^j(1-\xi)^{L_2-j}\mathcal{Q}\left(\frac{\Delta/2}{\sqrt{j\cdot\sigma^2_N}}\right). \tag{3.31}$$

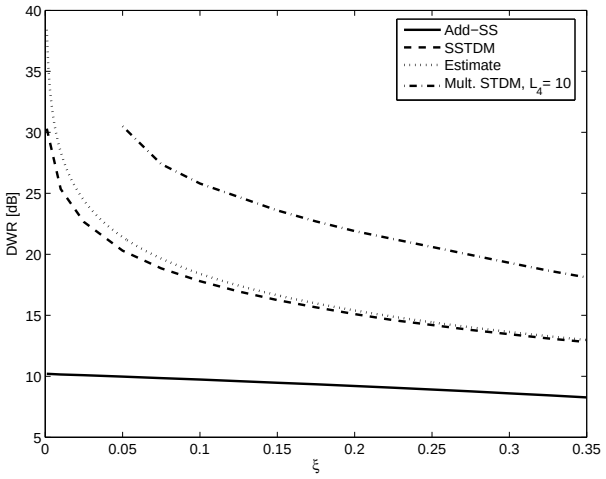


Figure 3.12: Comparison between Add-SS, SSTDM and multidimensional STDM. The plot shows the DWR empirically needed to achieve $P_e = 10^{-3}$ versus the fraction of cropped coefficients ξ, taking $\mathrm{Var}\{N_i\} = \mathrm{Var}\{Y_i\}$ for all $i \in \mathcal{I}$. $L_2 = 100$. $L_4 = 10$ for multidimensional STDM.

A further simplification results for large values of L_2, as the number of cropped coefficients is roughly $L_2\xi$. Moreover, for $\mathrm{Var}\{X_p\} \gg \Delta^2$, i.e., for a high resolution quantization, we have $D_w = \frac{\Delta^2}{3\cdot(L_2)^2}$. Therefore, equation (3.31) can be approximated by

$$P_{e,u} \approx 2\mathcal{Q}\left(\sqrt{\frac{3 \cdot L_2 \cdot D_w}{4 \cdot \xi \cdot \sigma_N^2}}\right).$$

From this expression it is immediate to estimate the fraction ξ of pixels that can be randomly removed in order to achieve a given probability of error P_e:

$$\xi = \frac{D_w \cdot 3 \cdot L_2/4}{\eta^2 \cdot \sigma_N^2}, \tag{3.32}$$

where $\eta \triangleq \mathcal{Q}^{-1}(P_e/2)$.

Figure 3.12 plots the DWR empirically needed to achieve $P_e = 10^{-3}$, versus the cropping ratio ξ for Add-SS, which virtually coincides with the results obtained from (3.28), and SSTDM, when $\mathrm{Var}\{N_i\} = \mathrm{Var}\{Y_i\}$, for all $i \in \mathcal{I}$. It is worth mentioning that this case corresponds to the decoder being unable to estimate the removed coefficients from the available ones, so it sets $\hat{y}_i = 0$, for all $i \in \mathcal{I}$. Note that this can be considered as a worst case, because in practice it may be possible to estimate some of the lost coefficients from their available neighbors. In any event, Figure 3.12 shows the sharp degradation of SSTDM as opposed to the graceful one corresponding to Add-SS when cropping is applied, and illustrates as well the goodness of the estimate given in (3.32).

3.3.2. Possible solutions

In this section we will justify why the multidimensional version of STDM using the Cartesian product of scalar uniform quantizers (equivalently, ST-SCS with repetition and $\alpha = 1$) [65], [132] turns out to be a good solution against cropping, while still showing a satisfactory behavior under additive noise attacks.[4]

In order to have similar noise power in all the projected components, we propose to choose an orthogonal $\mathbf{S}$, such that $\mathbf{S}^T \cdot \mathbf{S} = (L_1/L_3) \cdot \mathbf{I}_{L_3 \times L_3}$, where $\mathbf{I}_{L_3 \times L_3}$ is the identity matrix of size L_3. Furthermore, the set of indices with non-zero values should be disjoint for each pair of columns, in such a way that a single cropped coefficient will affect just one component in the projected domain. This condition is not fulfilled by watermarking systems which embed the watermark in the full-block DCT or DFT domains, because a single removed pixel will change many components in the transformed domain.

It can be also seen that the dimensionality of the projected subspace should not be too small; in this way, it will be possible to find components in such subspace that have not been distorted by cropping. Note, however, that there is a trade-off between robustness to cropping (which would require a large value of L_3), and to additive noise attacks (for which a small L_3 is preferred). Finally, it is desirable that quantization of a projected component should be done independently of the other ones. In such case, the cropping-induced distortion in one projected component will not leak into other components, which then can still be correctly decoded. This requirement is fulfilled by cubic lattices ($\Lambda = K\mathbb{Z}^{L_4}$, with $K \in \mathbb{R}$) with repetition coding.

With all these considerations, we propose the following projection matrix $\mathbf{S}$ to study the trade-off between robustness to cropping and to additive noise attacks:

$$s_{ij} = \begin{cases} \pm\frac{\sqrt{L_1/L_3}}{\sqrt{K_1}}, & \text{if } (j-1)K_1 < i \leq jK_1, \\ & 1 \leq j \leq L_3 - K_2 \\ \pm\frac{\sqrt{L_1/L_3}}{\sqrt{K_1+1}}, & \text{if } (j-1)(K_1+1) - L_3 + K_2 < i \\ & \leq j(K_1+1) - L_3 + K_2, \\ & L_3 - K_2 + 1 \leq j \leq L_3 \\ 0, & \text{otherwise} \end{cases}$$

where $K_1 = \lfloor L_1/L_3 \rfloor$ and $K_2 = L_1 - K_1 \cdot L_3$, and the sign of the non-zero elements is pseudorandomly generated.

Figure 3.12 also plots the DWR empirically needed for this scheme to achieve $P_e = 10^{-3}$ when $L_2 = 100$ and $L_4 = 10$, showing a better behavior of the multidi-

[4] Be aware that, as it was mentioned, the decoding rule is based on the estimate of the watermarked signal $\hat{\mathbf{Y}}$, i.e. $\hat{b}_j = \arg\min_{b_j \in \{0,P-1\}} ||\hat{\mathbf{Y}}_{p_j} - Q_{b_j}(\hat{\mathbf{Y}}_{p_j})|| = \arg\min_{b_j \in \{0,P-1\}} ||\hat{\mathbf{Y}}_{p_{\mathrm{mod}_j}} - Q_{b_j}(\hat{\mathbf{Y}}_{p_{\mathrm{mod}_j}})||$, where $\hat{\mathbf{Y}}_p = \mathbf{S}^T\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}_{p_{\mathrm{mod}}} = \hat{\mathbf{Y}}_p \mod \Lambda$.

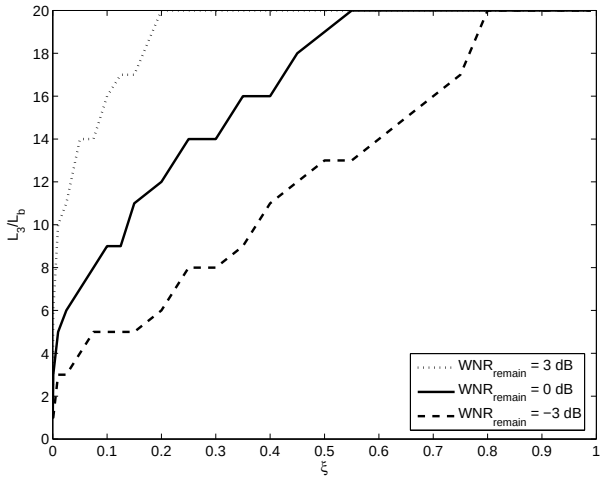


Figure 3.13: Comparison of the optimal dimensionality of the projected domain per symbol L_3/L_b as a function of the cropping ratio ξ, when DWR = 30 dB, taking $\mathrm{Var}\{N_i\} = \mathrm{Var}\{X_i\}$ for all $i \in \mathcal{I}$ and $\mathrm{WNR}_{\mathrm{remain}} = -3, 0$ and 3 dB. $L_1/L_b = 20$.

mensional version in all the range of cropping ratios; nevertheless, be aware that in order to make a fair comparison between SSTDM and the former multidimensional version, the channel noise should be taken into account. In Figure 3.13 the value of L_3/L_b minimizing P_e for different $\mathrm{WNR}_{\mathrm{remain}}$s is plotted as a function of the cropping ratio ξ, showing the aforementioned trade-off between robustness to cropping and additive noise attacks. Decoding was performed disregarding those projected components which depend on cropped coefficients, because their large variance would likely confuse the decoder. This strategy is equivalent to weighting those components by 0 when computing Euclidean distances, see [46]. Whenever the cropping ratio is close to 0, the predominant effect is due to the additive noise, so the best results are obtained for small values of L_3/L_b, that is, strategies close to SSTDM ($L_4 = L_3/L_b = 1$); however, when the ratio ξ is raised, cropping becomes the main problem and, as discussed above, increasing L_3/L_b is a good solution. Recall that by increasing L_3 we are approaching a non-projected scheme; in fact, the case $L_3 = L_1$ corresponds to SCS with repetition coding.

Finally, Figure 3.14 compares the performance of DC-DM with uniform scalar quantizers and repetition coding with that obtained for the SSTDM case, after removing an 8-pixels-wide outer frame; the original *Lena*, and its cropped version are plotted in Figure 3.15. The data are hidden in the mid-frequency coefficients of the 8×8 block-DCT domain, as it was done in Section 3.1.5. The results shown in Figure 3.14, jointly with those previously plotted in Figure 3.4, confirm that a combination of DC-DM and SSTDM is a good choice towards a truly robust moderate-rate scheme.

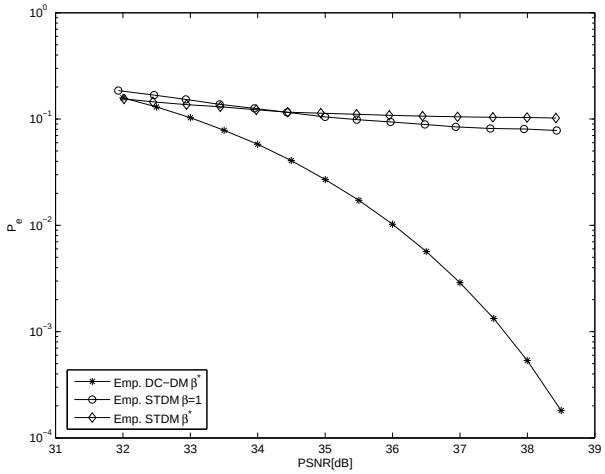


Figure 3.14: Experimental results of DC-DM ($\alpha = 0.4$) vs. SSTDM, watermarking the DCT domain of actual images, with additive uniform noise after cropping an 8-pixels-wide frame; results averaged over twenty-two 256×256 images, with $L_2 = 20$ (payload $L_b = 1126$ bits).

(a) (b) (c)

Figure 3.15: (a) Original *Lena*. (b) 8×8-pixels-wide outer frame is drawn in black. (c) Resulting cropped image.

3.3.3. Conclusions

Even though data hiding schemes that work by projecting and quantizing on a scalar domain achieve the same spreading gain as Add-SS, care should be taken before extrapolating the favorable properties of the latter to the former. For instance, it is known that while binary Add-SS is impervious to amplitude scalings, this is no longer the case of standard quantization-based methods. Here we have shown that for cropping attacks both methods behave quite differently, since the degradation observed in SSTDM is much sharper than for Add-SS. By following some design rules for the projection matrix, which entail increasing the

dimensionality of the projected subspace, it is possible to derive a quantization-based method robust both to cropping and to additive noise attacks. Due to the simplicity and potential impact of cropping, future benchmarks for Quantization Index Modulation (QIM) like methods should always consider this attack.

3.4. Sensitivity Attack

Until now, the sensitivity attack was considered as a serious threat to the robustness and security of spread-spectrum-based schemes, since it provides a practical method of removing watermarks with minimum attacking distortion. Nevertheless, it had not been used to remove the watermark from other watermarking algorithms, such as those which use side-information. Furthermore, the sensitivity attack has never been used to obtain falsely watermarked contents, also known as forgeries. In this section an overview of previous research on this subject is presented and a new version of the sensitivity attack based on a general formulation is proposed; this method does not require any knowledge about the detection function nor any other system parameter, but just the binary output of the detector, being suitable for attacking most known watermarking methods. The new approach is validated with experiments.

3.4.1. Scenario

As it was said in Section 1.1, digital data hiding was conceived in its early years as a potential solution to the problems of illegal copy control and intellectual property rights (IPR) protection. Perhaps for this reason and the analogies commonly made to the field of cryptography, watermarking was declared as synonymous to security [96]. However, watermarking research until now has much more to do with *robustness* than with *security*: roughly speaking, watermarking security may be related to attacks which try to gain knowledge about certain secret parameters of the watermarking system, whereas robustness is more concerned with attacks whose aim is to degrade the performance of the watermarking system [38]; see Section 4.2 for a further discussion on this topic.

In watermarking for IPR protection and copy control, the aim is to distinguish whether the digital media at hand contains a certain watermark or not. This is a *watermark detection* problem, which is commonly modeled as a binary hypothesis testing problem, which can be written as

$$\begin{aligned} H_0 &: \mathbf{y} = \mathbf{x} \\ H_1 &: \mathbf{y} = \mathbf{x} + \mathbf{w}. \end{aligned}$$

Recall that $\mathbf{w}$ may be made key-dependent in order to improve the security of the system. Since the watermarked signal could have been attacked, the detector

should adapt this test to take the attack into account. As it will be explained, the sensitivity attack is iteratively computed, so in order to distinguish it from the typical *one-attempt* attacks, we will model it as the addition of a vector $\mathbf{t}$, yielding a signal $\mathbf{z} = \mathbf{y} + \mathbf{t}$. The optimal solution to the hypothesis test is given by the likelihood ratio test, i.e.,

$$l(\mathbf{z}) = \frac{f_{\mathbf{Z}|H_1}(\mathbf{z}|H_1)}{f_{\mathbf{Z}|H_0}(\mathbf{z}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3.33}$$

where $f_{\mathbf{Z}|H_i}(\mathbf{z}|H_i)$ is the pdf of $\mathbf{Z}$ conditioned on hypothesis H_i and η is a threshold which can be adjusted so as to optimize a certain criterion (Neyman-Pearson, Bayes, etc.). The output of the detector will be denoted by $D \in \mathcal{H} = \{H_0, H_1\}$. The detection function given by (3.33) divides the subspace $\mathbb{R}^{L_1}$ in two disjoint regions, $\mathcal{R}$ and $\mathcal{R}^c$, termed *acceptance* or *detection region* and *rejection region*, respectively, such that $\mathbb{R}^{L_1} = \mathcal{R} \cup \mathcal{R}^c$. These regions are defined as

$$\mathcal{R} = \{\mathbf{z} \in \mathbb{R}^{L_1} : l(\mathbf{z}) > \eta\}; \mathcal{R}^c = \{\mathbf{z} \in \mathbb{R}^{L_1} : l(\mathbf{z}) \leq \eta\}.$$

Unfortunately, an analytical derivation of the likelihood ratio test is not always feasible, so we will consider instead a more general family of detection functions. Thus, the test performed by the detector is

$$f(\mathbf{z}, \boldsymbol{\theta}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta,$$

where $\boldsymbol{\theta}$ is the secret key used in the detection process. Of course, the resulting detector will be optimal only when $f(\mathbf{z}, \boldsymbol{\theta})$ coincides with the likelihood ratio $l(\mathbf{z})$.

In the considered scenarios, i.e. copyright protection and copy control, the watermark detector is often made public, generally in the form of a tamper-proof black box which only provides binary outputs, in such a way that an observer can check whether $f(\mathbf{z}, \boldsymbol{\theta})$ is larger or smaller than η, but he/she can not know its actual value. This scenario gives rise to the so-called *oracle attacks*, where the attacker uses the detector outputs to some selected inputs in order to gain knowledge about secret information used in the detection process (for instance, the detection key). Intuitively speaking, the detector acts as an oracle, responding *yes* or *no* to the inputs provided by the attacker. A block-diagram of the oracle attacks can be found in Figure 3.16.

The most popular oracle attack is the so-called *sensitivity attack*, introduced for the first time in [53]. At the time this attack was proposed, *additive spread*

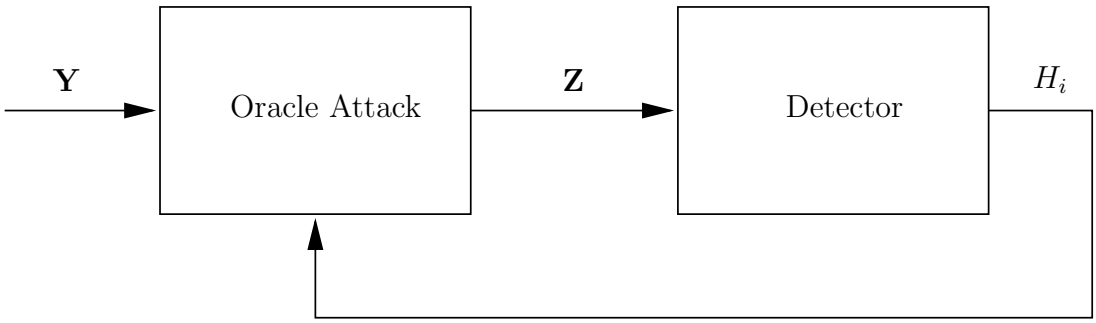


Figure 3.16: Block diagram of oracle attacks.

spectrum (Add-SS) methods [52] constituted the state of the art in digital watermarking, so this attack was suited to this particular scenario. For Add-SS under the assumption of an i.i.d. Gaussian host, the likelihood ratio is given by the linear correlation $l(\mathbf{z}) = \mathbf{z}^T\mathbf{w}$, so the optimal detector in this case must apply the following test:

$$\mathbf{z}^T\mathbf{w} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3.34}$$

Detectors that implement the test given by (3.34) are termed *linear correlation detectors*. Essentially, the sensitivity attack (specialized to the case of digital images) for this kind of detectors consists of the following steps [53]:

1. The algorithm starts from a watermarked image $\mathbf{y}$ of dimension L_1. The first step is the modification of $\mathbf{y}$ so as to obtain a new image $\mathbf{z}$ near the boundary of $\mathcal{R}$, which according to (3.34) is a hyperplane in an L_1-dimensional subspace, perpendicular to $\mathbf{w}$.

2. For the i-th pixel of $\mathbf{z}$, a vector $\mathbf{t}^i = (0, \cdots, 0, t_i, 0, \cdots, 0)^T$ is added to $\mathbf{z}$ observing how the sign of t_i affects the outputs of the detector and, hence, gaining knowledge about the polarity of the watermark in each pixel. Since $\mathbf{z}$ is near the detection boundary, small changes are likely to toggle the detector response. This procedure is repeated for all $i = 1, \cdots, L_1$.

3. At the end of the previous step, by combining the results for all pixels, the attacker has a rough estimate $\hat{\mathbf{w}}$ of the watermark vector and, thus, of the detection boundary, which in the considered case is perpendicular to $\mathbf{w}$.

According to the classification introduced at the beginning of this section, the sensitivity attack clearly falls into the category of attacks to security for Add-SS methods, since the attacker is trying to disclose the spreading sequence (which is supposed to be secret to unauthorized users), or equivalently the boundary of

the detection region. Nevertheless, when more involved detection regions are considered, the attacker can try just to obtain the closest point to the watermarked signal onto the decision boundary, and this will not almost provide him/her information about the shape of the boundary (at most, just a local estimation), or the secret key, so it can be considered to be a robustness attack more than a security one; summarizing, the sensitivity can be thought of being in the midway between security and robustness attacks, or even being a previous step to other attack. For example, once the attacker has estimated this boundary, he/she can use his/her knowledge to devise smart attacks against watermarked contents: for instance, once the estimate $\hat{\mathbf{w}}$ has been obtained, the attacker can generate an attacked image $\mathbf{z}$ with small distortion, capable of fooling the detector, just by subtracting a suitably scaled version of $\hat{\mathbf{w}}$. Before the sensitivity attack was proposed, it was believed that the complexity of an attack disclosing the watermark was $O(2^{L_1})$ (by means of a *brute force* approach), but the proposed strategy showed that for Add-SS it would be feasible in a number of iterations which is linear with the dimensionality of the watermarked image, i.e., the complexity of the attack was reduced to $O(L_1)$. Hence, it is easy to realize that this attack represented a serious threat to any watermarking scheme with a public detector available, and it raised up the problem of security in watermarking.

This section is concerned with the generalization of the sensitivity attack, providing a formulation that encompasses most known watermark detection schemes with parameterizable and differentiable (but unknown to the attacker) detection boundaries. Furthermore, we will also use the attack for generating false positives, i.e. from a watermarked signal and an original unwatermarked different signal, a watermarked version of the latter will be constructed. The rest of the section is organized as follows: Section 3.4.2 provides an overview of previous works dealing with this attack and the countermeasures proposed to increase the security of a watermarking system where public detectors are available. In Section 3.4.3, our new formulation of the problem is presented, and its application to some examples is given in Section 3.4.4. Finally, Section 3.4.5 deals with the computational complexity of the proposed method, and some final remarks are introduced in Section 3.4.6.

3.4.2. Previous work and improvements

The sensitivity attack for detectors based on linear correlation, i.e., those given by (3.34), was extensively studied in [109] and [97]. Starting from the formulation of the attack given in [53], which was explained in Section 3.4.1, the work in [109] proposes a first countermeasure based on the randomization of the detection boundary: the basic idea is to define a region around the points that satisfy $\mathbf{z}^T\mathbf{w} = \eta$, where the decision of the detector is made random, in order to reduce the sensitivity of the detector to small changes in its inputs. Thus, the

detection function is modified as follows:

$$D = \begin{cases} H_1, & \text{if } \mathbf{z}^T\mathbf{w} > \eta_2 \\ H_0, & \text{if } \mathbf{z}^T\mathbf{w} < \eta_1 \\ H_1 \text{ with probability } p(\mathbf{z}^T\mathbf{w}), & \text{if } \eta_1 \leq \mathbf{z}^T\mathbf{w} \leq \eta_2 \end{cases}, \tag{3.35}$$

where the two new thresholds η_1 and η_2 must be close to η so as not to degrade significantly the performance of the detector, and $p(r)$ verifies $p(\eta_1) = 0$ and $p(\eta_2) = 1$. The internal behavior of the detector is such that its outputs are deterministic, i.e., the response of the detector is always the same for a fixed input signal $\mathbf{z}$, in order to avoid the estimation of $p(r)$ simply by feeding the same $\mathbf{z}$ to the detector repeatedly. Anyway, estimation of the watermark is still possible. Let $\mathbf{z}'$ be a vector such that $\eta_1 \leq (\mathbf{z}')^T\mathbf{w} \leq \eta_2$, and $\boldsymbol{\epsilon}$ a random vector. For sufficiently small $\epsilon_i, i = 1, \cdots, L_1$, and $\mathbf{z} = \mathbf{z}' + \boldsymbol{\epsilon}$, we have that

$$p(\mathbf{z}^T\mathbf{w}) = p((\mathbf{z}')^T\mathbf{w} + \boldsymbol{\epsilon}^T\mathbf{w}) \approx p((\mathbf{z}')^T\mathbf{w}), \tag{3.36}$$

so after trying a sufficiently large number of different vectors $\boldsymbol{\epsilon}$, the value of $p((\mathbf{z}')^T\mathbf{w})$ can be estimated simply by counting the number of outcomes that yield $D = H_1$. Similarly, for $\mathbf{t}^i = (0, \cdots, 0, t_i, 0, \cdots, 0)^T$ and $\mathbf{z}^i = \mathbf{z}' + \mathbf{t}^i + \boldsymbol{\epsilon}$, we have

$$(\mathbf{z}^i)^T\mathbf{w} = (\mathbf{z}')^T\mathbf{w} + t_i w_i + \boldsymbol{\epsilon}^T\mathbf{w} \approx (\mathbf{z}')^T\mathbf{w} + t_i w_i = (\mathbf{z}')^T\mathbf{w} \pm t_i\delta, \tag{3.37}$$

where in the last equality we have assumed that $w_i \in \{\pm\delta\}$. By means of a first order approximation, and assuming that $p(r)$ is differentiable, we can write

$$p((\mathbf{z}^i)^T\mathbf{w}) \approx p((\mathbf{z}')^T\mathbf{w} \pm t_i\delta) \approx p((\mathbf{z}')^T\mathbf{w}) \pm t_i\delta p'((\mathbf{z}')^T\mathbf{w}), \tag{3.38}$$

where $p'(r) \triangleq \frac{\partial p(r)}{\partial r}$ is the derivative of $p(r)$. Again, using enough different vectors $\boldsymbol{\epsilon}$, an estimate of $p((\mathbf{z}^i)^T\mathbf{w})$ can be obtained. By comparing this estimate to the previous estimate of $p(\mathbf{y}^T\mathbf{w})$, the sign of w_i can be inferred (as long as $p(r)$ is a monotically increasing function). In [109], the information leakage about the watermark provided by the detector outputs is quantified in an information-theoretic sense, and the shape of the optimum function $p(r)$ for $\eta_1 \leq r \leq \eta_2$ that minimizes the information leakage is given. It is easy to see that this countermeasure complicates the sensitivity attack, but its complexity still remains linear with the dimensionality of the images. In fact, a practical method for estimating the watermark in this framework was devised in [97]. The method basically consists of the following steps:

1. Starting from a valid watermarked image $\mathbf{y}$, an image $\mathbf{z}'$ which yields $\eta_1 \leq (\mathbf{z}')^T\mathbf{w} \leq \eta_2$ is constructed by iteratively degrading $\mathbf{y}$.

2. The image $\mathbf{z}'$ is perturbed by the addition of zero-mean random vectors $\mathbf{t}$ with $t_i = \{\pm\delta\}$ (where δ has not to take the same value that in (3.37)). If $\mathbf{w}$ and $\mathbf{t}$ are positively correlated, the detector will return $D = H_1$ with higher probability, so $\mathbf{t}$ will be taken as an approximation of $\mathbf{w}$; otherwise, if $D = H_0$, then $-\mathbf{t}$ will be taken as an estimate of $\mathbf{w}$.

3. By averaging the estimates obtained in the previous step, an approximation of $\mathbf{w}$ is obtained.

Following this approach it is possible to obtain reliable estimates of $\mathbf{w}$ in a number of iterations which is a small multiple of L_1, as it was shown in [97].

Another approach for performing a successful sensitivity attack was presented in [111]. The method is able to estimate the boundary of the acceptance region by modeling the attack as a classical adaptive filtering problem: it is easy to realize that the linear detection function given in (3.34) for additive spread spectrum can be thought of in terms of filtering $\mathbf{z}$ with a filter $\tilde{\mathbf{w}}$ such that $\tilde{w}_i = w_{L_1+1-i} \; \forall \; i = 1, \cdots, L_1$; furthermore, the attacker knows that $\mathbf{z} * \tilde{\mathbf{w}} = f(\mathbf{z}, \boldsymbol{\theta})$, where $*$ denotes the convolution operator, so if he/she can access the values of $f(\mathbf{z}, \boldsymbol{\theta})$, then using this signal as reference he/she can manage to construct an estimate of $\tilde{\mathbf{w}}$. The authors propose in [111] the use of the Least Mean Squares (LMS) algorithm in order to iteratively construct these estimates. Let $\hat{\mathbf{w}}_k$ be the estimate of $\tilde{\mathbf{w}}$ in the k-th iteration and $\{\mathbf{z}_k\}$ a set of vectors near the detection boundary; each iteration of the LMS algorithm consists of the following steps:

$$
\begin{array}{ll}
1. & r_k = \mathbf{z}_k * \tilde{\mathbf{w}}, \\
2. & e_k = f(\mathbf{z}_k, \boldsymbol{\theta}) - r_k, \\
3. & \mathbf{w}_{k+1} = \mathbf{w}_k + \mu e_k \mathbf{z}_k,
\end{array}
$$

where μ is the step-length. In a more realistic situation, the attacker only has access to the detector output, D, so the authors introduce some modifications in the algorithm to work with just the binary output of the system. In this situation, the attacker must restrict the set $\{\mathbf{z}_k\}$ to those vectors lying near the detection boundary, because he/she still knows that $f(\mathbf{z}_k, \boldsymbol{\theta}) \approx \eta$; thus, the algorithm is complicated by the fact of computing the appropriate set $\{\mathbf{z}_k\}$. The authors also propose some modifications in order to cope with the countermeasure introduced in [109], which was explained in the above paragraph.

In view of the security flaws presented by traditional spread spectrum methods under sensitivity-like attacks, researchers put their effort in the design of *asymmetric* schemes [78]. One of the advantages offered by asymmetric schemes against sensitivity attacks is the fact that the embedding and detection keys are different, thus the impact of a successful attack revealing the detection boundary is minimized (recall that disclosure of the watermark in traditional Add-SS methods allows to unwatermark legal contents, as well as generating forged illegal documents). The other advantage of asymmetric watermarking is the use of more involved detection regions, complicating the description of the detection boundary; for instance, in [78], four asymmetric methods are analyzed under a unified framework, showing that the detection function can be written in terms

of a quadratic form in $\mathbb{R}^{L_1}$ for all cases, i.e.

$$\frac{\mathbf{z}^T\mathbf{A}\mathbf{z}}{L_1} \underset{H_0}{\overset{H_1}{\gtrless}} \eta,$$

where $\mathbf{A}$ is a matrix which depends on the secret key $\boldsymbol{\theta}$.

The idea of increasing the security of the system against sensitivity attacks by complicating the detection region is exploited by the family of detection functions called JANIS (Just Another N-order Side-Informed Scheme) [77], which use N-th order polynomial detection functions, i.e.

$$f(\mathbf{z}, \boldsymbol{\theta}) = \frac{1}{L_1} \sum_{k=1}^{L_1/N} \prod_{j=1}^{N} z_{p[(k-1)\cdot N+j](\boldsymbol{\theta})} \cdot a_{p[(k-1)\cdot N+j](\boldsymbol{\theta})}(\boldsymbol{\theta}),$$

where $\mathbf{a}$ is a pseudorandom ± 1 vector and $\mathbf{p}$ is a pseudorandom permutation vector, both of them depending on the secret key $\boldsymbol{\theta}$. Based on this detection function the watermark is obtained as $\mathbf{w} = \gamma \nabla f(\mathbf{x}, \boldsymbol{\theta})$, where γ is a parameter to adjust the embedding distortion and $\nabla(\cdot)$ is the gradient operator.[5] Indeed it makes much more difficult the sensitivity attack, but obviously this is not the ultimate solution: for example, an N-th order detection boundary can still be described by estimating $(L_1)^N$ points on such boundary. This point was addressed in [111], showing that the LMS attack can be properly modified in order to cope with this kind of detection boundaries. A possible solution to this problem was proposed also in [111] by means of non-parametric decision boundaries, i.e., by using decision boundaries that can not be described by a finite number of parameters. An example of such decision boundaries are those given by fractal curves like the Peano curve, which is used in [111] to replace the original linear detection boundary in a spread spectrum scheme. With a proper design, the authors defend that the proposed method can invalidate sensitivity attacks with slight degradations in robustness.

A different approach is followed in [154] which proposes the use of a decision function which is randomized whenever a signal with a similar hash was previously input to the detector; in this way the attacker could not rely on the outputs of the detector when he/she inputs similar signals, as it is the case of the sensitivity attack. Furthermore, [154] proposed the combination of the latter strategy with a response delay as a possible countermeasure against oracle attacks.

Recently, a rigorous formulation of the sensitivity attack was presented in [67]: first, the convergence of the algorithm proposed in [97] is proved, using the law

[5] Be aware that $\mathbf{a}$ and $\mathbf{p}$ are needed by the encoding algorithm, so JANIS is not an asymmetric method.

of large numbers; thereafter, a new non-iterative sensitivity attack for detectors based on linear correlation is presented. This new algorithm is also suitable for estimating continuous-valued watermarks, whereas the algorithms previously proposed in [109] and [97] assumed that the watermark could only take discrete values. The main steps of this new algorithm are outlined in the following:

1. As in the former algorithms, the first step is the construction of a signal $\mathbf{z}'$ near the boundary of the detection region.

2. Now consider the set of vectors $\{\mathbf{t}^i\}, \quad i = 1, \cdots, L_1$, defined by the canonical basis of $\mathbb{R}^{L_1}$. For each $\mathbf{t}^i$, a signal $\mathbf{z}'' = \mathbf{z}' + \nu_i \mathbf{t}^i$ on the detection boundary is constructed, by properly selecting the scaling factor ν_i. The search for this value of ν_i must be accomplished by means of some numerical algorithm, so it will be surely the most costly part.

3. For the detector under consideration, it holds that $(\mathbf{z}'')^T\mathbf{w} = \sum_{k=1}^{L_1} z'_k w_k + \nu_i w_i = \eta, i = 1, \cdots, L_1$ where η is the detection threshold, and $w_i = (\mathbf{t}^i)^T\mathbf{w}$. Thus, a linear system with L_1 equations and L_1 unknowns has been defined. By taking into account the special structure of this system, it is easy to show that it can be solved in $L_1 + 1$ elemental operations.

Another remarkable contribution of [67] is the extension of the sensitivity attack in order to work with a more generic family of detection functions of the form $g(\mathbf{y}, \mathbf{w})$; furthermore, this method has the advantage of returning an estimate of the watermark. Nevertheless, this approach presents several drawbacks: the attacker needs to know the detection function and even the inverse of the gradient of the detection function. Thus, the need for a new formulation which overcomes these problems is justified; in the next section we will try to solve this problem, achieving a solution which will be shown to work with a wider range of detection functions. The method proposed has the following characteristics:

- It does not require knowledge about the detection function; it just needs to know the binary output of the detection function for a given input. Due to this, our method is indeed able to deal with watermarking methods which use a secret detection key (different from the embedding key), in such a way that the attacker has no access to the decoding function. This is the case, for example, of asymmetric watermarking and JANIS.

- The gradient of the detection function does not need to be inverted. As it was said in the previous point, sometimes the detection function will not be known by the attacker, so he/she will not be able to invert its gradient.

3.4.3. The Blind Newton Sensitivity Attack (BNSA)

Focusing on watermark detection, we will describe the detector output through the function $f_{\text{binary}} : \mathbb{R}^{L_1} \rightarrow \mathcal{H}$, with $\mathcal{H} = \{H_0, H_1\}$. Without loss of generality, we can define the following functions

$$\begin{aligned} f &: \mathbb{R}^{L_1} \rightarrow \mathbb{R}^{L_3} \text{ and} \\ g_{\text{binary}} &: \mathbb{R}^{L_3} \rightarrow \mathcal{H}, \end{aligned} \tag{3.39}$$

with $L_3 \leq L_1$, in such a way that $f_{\text{binary}} = g_{\text{binary}} \circ f$, and f is parameterized by the secret key $\boldsymbol{\theta}$.[6] This decomposition will be shown to be useful in the next sections, since some of the most popular watermarking techniques perform embedding/detection in a projected domain so f can be seen as the projection function.[7] Furthermore, in the schemes studied in this work the output of g_{binary} will be based on the output of a real function g and a threshold η, in such a way that

$$g_{\text{binary}}(\mathbf{x}) = \begin{cases} H_0, & \text{if } g(\mathbf{x}) \leq \eta \\ H_1, & \text{if } g(\mathbf{x}) > \eta \end{cases}, \tag{3.40}$$

with $g : \mathbb{R}^{L_3} \rightarrow \mathbb{R}$.

On the other hand, a distortion measure has to be defined in order to quantify the impact of the attacking signal, termed $\mathbf{t}$, on the watermarked signal $\mathbf{y}$: [8]

$$\begin{aligned} d_{\mathbf{y}} : \mathbb{R}^{L_1} &\rightarrow \mathbb{R}^{+} \\ \mathbf{t} &\rightarrow d_{\mathbf{y}}(\mathbf{t}). \end{aligned}$$

This distortion measure should be based on perceptual criteria (depending on the nature of the host signal), although very often, and for the sake of simplicity, the squared Euclidean norm of $\mathbf{t}$ is chosen (i.e., $d_{\mathbf{y}}(\mathbf{t}) = ||\mathbf{t}||_2^2$).

Recalling that the attacker tries to find the vector $\mathbf{t}$ which yields a "no watermark" decision (i.e., $f_{\text{binary}}(\mathbf{y}+\mathbf{t}) = H_0$) while minimizing the distortion measure $d_{\mathbf{y}}(\mathbf{t})$, his/her target can be formalized as

$$\arg \min_{\mathbf{t}: g \circ f(\mathbf{y}+\mathbf{t}) \leq \eta} d_{\mathbf{y}}(\mathbf{t}). \tag{3.41}$$

Let us assume that $d_{\mathbf{y}}(\mathbf{t})$ is a continuous and convex function of $\mathbf{t}$ (for a given watermarked signal $\mathbf{y}$), which achieves its global minimum value at $\mathbf{t}_0$ (the

[6] Be aware that f is generalized with respect to its definition in the previous sections, in order to provide L_3-dimensional outputs, and the dependency with $\boldsymbol{\theta}$ is not explicitly shown for the sake of notational simplicity.

[7] We have used L_3 to denote the dimensionality of the projected domain, following the notation introduced in the description of STDM methods (Section 2.7), even when in this case the projection could be nonlinear.

[8] Ideally, this measure should quantify the differences between the original host signal and its attacked version; nevertheless, the attacker will have to design his/her strategy taking into account the watermarked signal, since he/she has not access to the original one.

squared Euclidean norm obviously fulfills these conditions), a vector that belongs to the set of attacking vectors yielding H_1 (which we will denote by $\mathcal{B}$),[9] i.e., $\mathbf{t}_0 \in \mathcal{B} \triangleq \{\mathbf{t} : g \circ f(\mathbf{t} + \mathbf{y}) > \eta\}$. Note that $\mathcal{B} \triangleq \mathcal{R} - \mathbf{y}$. Then, replacing $\mathcal{B}$ in (3.41), and denoting by $\partial\mathcal{B}$ its boundary and by $\mathcal{B}^c$ its complement, it is straightforward to show that $\arg\min_{\mathbf{t}\in\mathcal{B}^c} d_{\mathbf{y}}(\mathbf{t}) \in \partial\mathcal{B}$, so (3.41) is equivalent to

$$\arg \min_{\mathbf{t}:g\circ f(\mathbf{y}+\mathbf{t})=\eta} d_{\mathbf{y}}(\mathbf{t}). \tag{3.42}$$

This is a typical Lagrange multipliers problem, so the attacker could find a theoretical solution if both $d_{\mathbf{y}}$ and $g \circ f$ were known by him/her; nevertheless, this is not the case, since the latter depends on the secret key, which is unknown for the attacker. Actually, he/she will have only access to the binary output of the detector. In Appendix C we show that (3.42) is equivalent to

$$\arg \min_{\mathbf{s}\in\mathbb{R}^{L_1}} d^{\star}_{\mathbf{y}}(h_{\mathbf{y}}(\mathbf{s})), \tag{3.43}$$

where $d^{\star}_{\mathbf{y}}$ is the restriction of $d_{\mathbf{y}}$ to $\partial\mathcal{B}$, and $h_{\mathbf{y}}$ is a function verifying that: a) $h_{\mathbf{y}}$ is a surjection which maps $\mathbb{R}^{L_1}$ onto $\partial\mathcal{B}$, i.e. $h_{\mathbf{y}}(\mathbb{R}^{L_1}) = \partial\mathcal{B}$; b) $h_{\mathbf{y}}(\mathbf{s}) = \mathbf{s}$, for all $\mathbf{s} \in \partial\mathcal{B}$; c) $h_{\mathbf{y}}(\mathbf{s}) \in C^2$, i.e., its second derivative is continuous in a neighborhood of $\mathbf{s}$; and d) $h_{\mathbf{y}}(\mathbf{s})$ is estimated based on the binary output of the detector, without any other knowledge about the detection function.

Since theoretical solutions to (3.43) are not possible in general due to the lack of knowledge of the boundary of the decision region, numerical iterative methods should be applied by the attacker in order to find a solution. Due to this, in most cases the attacker will have to be satisfied with computing a local minimum of the function in (3.43), since the achievement of global minima is only ensured for convex problems, which need also the convexity of $\mathcal{B}^c$ [24]; nevertheless, in Section 3.4.4 the experimental results will show that this suboptimal solution is usually good enough. Concerning the iterative method, in this work we will use an adaptation of Newton's method [124], where the updated vector in the $(k+1)$-th iteration is computed as

$$\mathbf{s}_{k+1} = \mathbf{s}_k - \xi_k \cdot \left[\nabla^2(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)\right]^{-1} \cdot \nabla(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k), \tag{3.44}$$

where $\xi_k \in \mathbb{R}^+$ is the step-length, whose computation requires (in general) a line search [124, 24]: a small value of ξ_k will imply a slow convergence, but with a large one convergence cannot be assured.

It is straightforward to see that $\nabla(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)$ and $\nabla^2(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)$ cannot be obtained in an analytic way, therefore they must be numerically approximated

[9] Be aware that in most cases it is reasonable to consider that $\mathbf{t}_0 = \mathbf{0}$, since in that case the attacked signal will be the watermarked one, so the distortion is minimized; furthermore $\mathbf{t}_0$ is in $\mathcal{B}$, since $g \circ f(\mathbf{y})$ will yield H_1.

by taking into account that

$$\frac{\partial(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})}{\partial s_i}(\mathbf{s}) = \frac{(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_i) - (d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s})}{\delta} + O(\delta), \text{ and} \quad (3.45)$$

$$\frac{\partial^2(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})}{\partial s_i \partial s_j}(\mathbf{s}) = \frac{(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_i + \delta \mathbf{e}_j) - (d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_i)}{\delta^2} + \frac{-(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_j) + (d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s})}{\delta^2} + O(\delta), \quad (3.46)$$

with $\mathbf{e}_i$ the i-th vector of the canonical basis, and $\delta > 0$ the step size used by the approximation.

An alternative strategy to (3.44), which is especially suitable for large-scale problems, is based on replacing the Hessian by a diagonal matrix just keeping the diagonal elements; in this way, an iteration of the algorithm just requires $(2 \cdot L_1 + 1)$ evaluations of $(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s})$ and (3.44) is computed with L_1 scalar divisions (if the complete matrix were used, a linear system with L_1 equations and L_1 variables should be solved). Another updating algorithm is given by

$$\mathbf{s}_{k+1} = \mathbf{s}_k - \xi_k \cdot \hat{\nabla}(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k), \quad (3.47)$$

where $\hat{\nabla}(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)$ is the estimation of the gradient of $(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})$ based on (3.45), which only requires $L_1 + 1$ evaluations; for a small enough ξ_k, equation (3.47) will guarantee a decrease in $(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})$ as far as $(\hat{\nabla}(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k))^T \cdot \nabla(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k) > 0$, where the last condition is based on the Taylor series expansion of the objective function.

Although $h_{\mathbf{y}}(\mathbf{s})$ could be any function verifying the properties introduced above, in the experimental part of this work we will use $h_{\mathbf{y}}(\mathbf{s}) = \nu^* \cdot \mathbf{s}$, where $\nu^* = \arg\min_{\nu \in \mathbb{R}: \nu \cdot \mathbf{s} \in \partial\mathcal{B}} |\nu|$. This computation of $h_{\mathbf{y}}(\cdot)$ is based on the fact that $\mathbf{0} \in \mathcal{B}$, since $\mathbf{y}$ is a watermarked signal, and also that for most known watermarking methods $\beta \cdot \mathbf{s} \in \mathcal{B}^c$ for large values of β, so ν^* (equivalently $h_{\mathbf{y}}$) can be estimated by a bisection algorithm, where just the binary output of the detector is needed to determine what will be the extremes of the interval at the next iteration. Given that this method is based on the binary output of the detector, without any other knowledge about the detection function, the algorithm is said to be *blind*. An example of an iteration of the algorithm is plotted in Figure 3.17, where the reduction from $||\nu_k \mathbf{s}_k||^2$ to $||\nu_{k+1} \mathbf{s}_{k+1}||^2$ is represented by the decrease of the radius of two spheres centered at $\mathbf{y}$. It is important to remark that the blindness of the proposed attack comes at the price of its iterative nature and thus typically requires many more calls to the oracle than other algorithms, e.g. [53], [97], [67], for which the attacker is assumed to perfectly know the shape of the detection function.

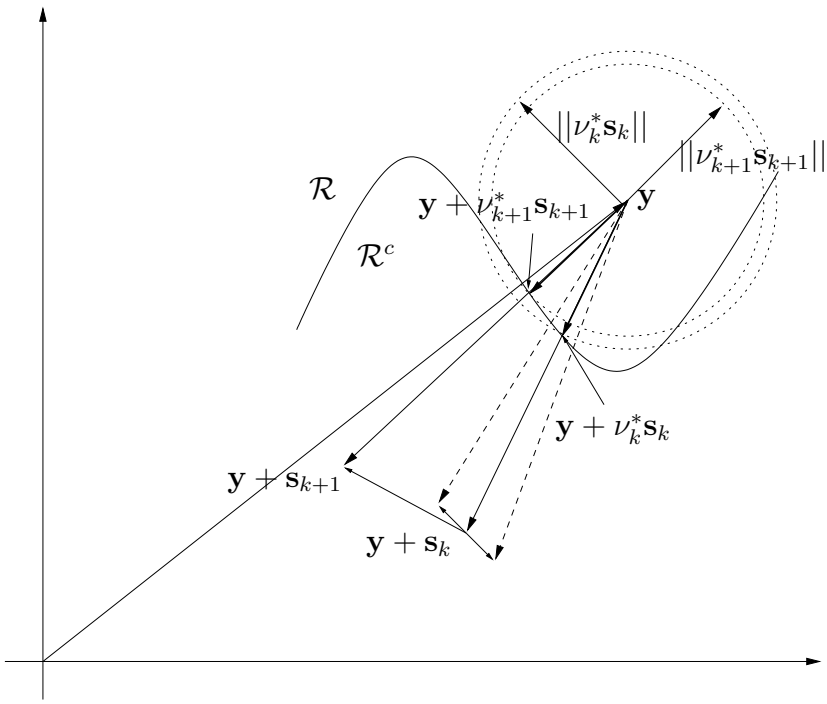


Figure 3.17: Example of an iteration of the algorithm. Given a watermarked signal $\mathbf{y}$ and the attacking vector in the k-th iteration $\mathbf{s}_k$, the last one is slightly modified to estimate the gradient and Hessian of $d^\star_{\mathbf{y}} \circ h_{\mathbf{y}}(\mathbf{s}_k)$. Once the descent direction and the step-length have been computed, $\mathbf{s}_{k+1}$ is obtained. It can be seen that $\mathbf{y} + \nu^*_{k+1}\mathbf{s}_{k+1}$ is closer to $\mathbf{y}$ than $\mathbf{y} + \nu^*_k\mathbf{s}_k$, i.e. $||\nu^*_{k+1}\mathbf{s}_{k+1}||^2 < ||\nu^*_k\mathbf{s}_k||^2$.

3.4.3.1. Implementation

In this section, the different steps of BNSA are enumerated, and some considerations on its implementation are made.

1. Initialization: in order to start to iterate, a first point on the boundary of the detection region is needed. We propose to use $\mathbf{z} = \gamma^* \cdot \mathbf{y}$, where $\gamma^* = \arg\max_{\gamma \in \mathbb{R}^+ : g \circ f(\gamma \mathbf{y}) = \eta}(\gamma)$. This choice is based on the fact that for most known watermarking methods $\mathbf{0}$ is in $\mathcal{R}^c$. Were this not the case, we assume that the attacker can compute a signal yielding in $\mathcal{R}^c$, and hence a document on the decision boundary can still be computed by simple convex combination of this signal and the watermarked one. The factor γ^* can be computed using a bisection algorithm.

   Therefore, the initialization value of $\mathbf{s}_k$ is given by $\mathbf{s}_0 = (\gamma^* - 1)\mathbf{y}$.

2. Updating: the update of the algorithm has the generic form

$$\mathbf{s}_{k+1} = \mathbf{s}_k - \xi_k \cdot {\mathbf{B}_k}^{-1} \cdot \hat{\nabla}(d^\star_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k),$$

where $\hat{\nabla}(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)$ is the estimate of the gradient of $(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)$, and its i-th component is computed as

$$[\hat{\nabla}(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)]_i = \frac{(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_i) - (d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s})}{\delta}.$$

This computation requires also to estimate $h_{\mathbf{y}}(\cdot)$, which is not known by the attacker. As it was mentioned above, the proposed strategy is to use $h_{\mathbf{y}}(\mathbf{s}) = \nu^* \cdot \mathbf{s}$, where $\nu^* = \arg\min_{\nu \in \mathbb{R}: \nu \cdot \mathbf{s} \in \partial \mathcal{B}} |\nu|$, although any function verifying the conditions described above could be used. Again, in the proposed implementation ν^* is computed using a bisection algorithm.

Concerning the matrix $\mathbf{B}_k$, different possibilities can be considered; they are chosen by the attacker based on his/her computational resources. First of all, we can consider an approximation to the Hessian, i.e. $\mathbf{B}_k = \hat{\nabla}^2(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s}_k)$, which yields an approximation to the classical Newton's method. The (i, j)-th component of that matrix is computed as

$$\begin{aligned}[\hat{\nabla}^2(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s})]_{i,j} = & \frac{(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_i + \delta \mathbf{e}_j) - (d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_i)}{\delta^2} \\ & + \frac{-(d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s} + \delta \mathbf{e}_j) + (d^{\star}_{\mathbf{y}} \circ h_{\mathbf{y}})(\mathbf{s})}{\delta^2}.\end{aligned}$$

This alternative is really computationally expensive, since it requires $O(L_1^2)$ calls to the detector per update of the algorithm.

Finally, the computation of ξ_k, which is done just once per iteration of the algorithm, can be calculated using one off-the-shelf line search algorithm (see [124, 24] for further references). In our implementation we have followed Armijo's rule, due to its simplicity.

3. Stopping: the stopping condition could be based on different criteria. For example, the attacker could stop when he/she computes a non-watermarked signal with a sufficiently good quality, or when the distortion measure function is not significantly reduced from one iteration to the next. Another alternative is to constrain the number of calls to the detector.

3.4.3.2. Computing forgeries

Let us suppose now that the attacker has access to a watermarked signal $\mathbf{y}$ and an original unwatermarked content $\mathbf{x}$, i.e. $f_{\text{binary}}(\mathbf{x}) = H_0$. A possible application of the proposed method is the computation of a signal $\mathbf{z} = \mathbf{x} + \hat{\mathbf{w}}$ which yields $f_{\text{binary}}(\mathbf{z}) = H_1$ while minimizing the distortion when it is compared with $\mathbf{x}$, i.e., we can create a watermarked version of a signal $\mathbf{x}$ whenever a watermarked content $\mathbf{y}$ and a detector are available.

Following the approach introduced in the previous section, this problem can be formalized as

$$\arg \min_{\mathbf{t}: g \circ f(\mathbf{y}+\mathbf{t}) > \eta} d_{\mathbf{X}}(\mathbf{y}+\mathbf{t}-\mathbf{x}), \tag{3.48}$$

i.e., we are trying to find a vector $\mathbf{t}$ such that when it is added to $\mathbf{y}$ yields a watermarked signal which is as similar as possible to $\mathbf{x}$, so $\mathbf{z} = \mathbf{x} + \hat{\mathbf{w}} = \mathbf{y} + \mathbf{t}$. Assuming again that $d_{\mathbf{X}}(\mathbf{t})$ is a continuous and convex function of $\mathbf{t}$, the solution $\mathbf{t}^*$ to (3.48) will verify $g \circ f(\mathbf{y} + \mathbf{t}^*) = \eta + \epsilon$, where $\epsilon > 0$ is arbitrarily small. Therefore, from a practical point of view, we can think of $\mathbf{t}^*$ as being on $\partial\mathcal{B}$. Proceeding in a similar way to (3.43), we can rewrite (3.48) as

$$\arg \min_{\mathbf{s} \in \mathbb{R}^{L_1}} d_{\mathbf{X}}(\mathbf{y} + h_{\mathbf{y}}(\mathbf{s}) - \mathbf{x}). \tag{3.49}$$

The solution to the last problem will be based on numerical iterative methods, so in the general case of non-convex decision regions just a local minimum will be achieved, as it happened with the watermark removal problem.

Taking into account that the probability of randomly finding a point in the detection region is given by the probability of false alarm P_{fa},[10] which typically is really small due to design criteria, the attacker still needs to be given a watermarked signal $\mathbf{y}$. This signal, which can be completely different of the document to be forged $\mathbf{x}$, is necessary in order to define the projecting function $h_{\mathbf{y}}(\cdot)$, which will allow the attacker to compute documents on the detection boundary. Assuming that he/she has access to a watermarked signal $\mathbf{y}$, then $h_{\mathbf{y}}(\cdot)$ can be given by the expression proposed in Section 3.4.3.1, i.e. $h_{\mathbf{y}}(\mathbf{s}) = \nu^* \cdot \mathbf{s}$, where $\nu^* = \arg\min_{\nu \in \mathbb{R}: \nu \cdot \mathbf{s} \in \partial\mathcal{B}} |\nu|$.

Finally, one must be aware that the decision regions could depend on the statistics of the input signal, as it is the case for the ML detector for a Generalized Gaussian distributed host, where the decision region depends on the shape parameter of the received signal. This dependency has a limited impact when the initial signal is not substantially modified, since in that case the statistics of that initial signal will be almost unaltered; this will be usually the case when the attacker tries to unwatermark a signal, but *the statistics* could dramatically change the decision regions when the attacker tries to *move* from one image to another, possibly frustrating an attack.

3.4.4. Application to real methods

In this section we will particularize the proposed algorithm to some of the most popular watermarking methods, showing the practical usefulness of this new attack and comparing the performance of the different schemes.

[10] The probability of false alarm P_{fa} is defined as $\Pr\{g_{\text{binary}} \circ f(\mathbf{x}+\mathbf{t}) = H_1 | H_0\}$. On the other hand, the probability of missed detection P_m is defined as $\Pr\{g_{\text{binary}} \circ f(\mathbf{x}+\mathbf{w}+\mathbf{t}) = H_0 | H_1\}$.

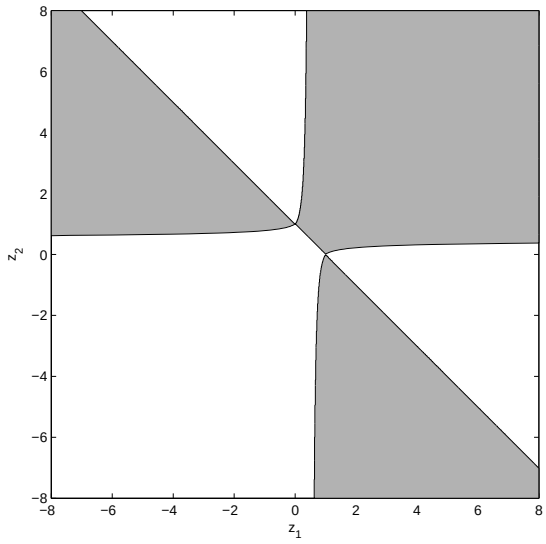


Figure 3.18: Decision regions obtained taking into account a l_c-norm when $c_1 = c_2 = 0.5$.

3.4.4.1. Spread Spectrum

Detection of standard Add-SS methods is based on the correlation between the received signal $\mathbf{z}$ and the watermark $\mathbf{w}$. Therefore, the function f, defined in (3.39), projects $\mathbf{z}$ onto a one-dimensional domain ($L_3 = 1$), i.e. $f(\mathbf{z}) = \mathbf{z}^T \cdot \mathbf{w}$, and g in (3.40) will be the identity function ($g(x) = x$, for all $x \in \mathbb{R}$), so the detection function is given by

$$\mathbf{z}^T \cdot \mathbf{w} \underset{H_0}{\overset{H_1}{\gtrless}} \eta,$$

in such a way that the boundary of the decision region will be a hyperplane. We will denote this case by *SS-corr*.

Another alternative for the detection function is that proposed by Cox *et al.* in [56]; in that case, the embedding is still given by (2.5), and f quantifies the angle between the received signal $\mathbf{z}$ and the watermark vector $\mathbf{w}$, i.e. $f(\mathbf{z}) = \frac{\mathbf{z}^T \cdot \mathbf{w}}{||\mathbf{z}|| \cdot ||\mathbf{w}||}$, and g is again the identity function, yielding a decision region $\mathcal{B}$ which is an L_1-dimensional cone. This method will be named *SS-angle*, as it really measures the angle between the received signal and the watermark.

As a countermeasure against BNSA, one could design detection functions for which component-wise modifications produce bounded increments, since for this kind of functions the task of finding vectors on the boundary of the detection function is considerably complicated. Interestingly, the ML detection function

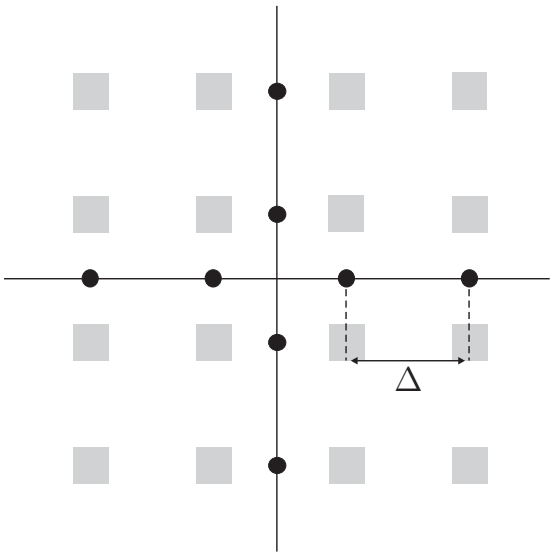


Figure 3.19: AND Region for QPD.

for Generalized Gaussian distributed hosts [93], i.e. equation (2.11), denoted by *SS-GG*, fulfills this requirement whenever the shape parameter c_i is such that $c_i < 1$. An example of the resulting decision region is plotted in Figure 3.18.

3.4.4.2. Side-informed methods

In Section 3.4.2 the JANIS methods were introduced. In order to make a comparison with the other existing methods, we have fixed the order of the detection function to $N = 4$, so

$$f(\mathbf{z}) = \frac{1}{L_1} \sum_{k=1}^{L_1/4} \prod_{j=1}^{4} z_{p[(k-1)\cdot N+j]} \cdot a_{p[(k-1)\cdot N+j]}.$$

Quantization-based methods have been shown to be useful for data hiding applications; nevertheless, and despite of their success in that application, very little has been said about their use in detection scenarios. To the best of our knowledge, the first work addressing the problem from this point of view is [66], where the Scalar Costa Scheme is adapted to authentication purposes by embedding a fixed message, yielding the detection function $g_{\text{SCS}}(\mathbf{z}, \boldsymbol{\theta}) = \frac{f_{\mathbf{Y}}(\mathbf{z})}{f_{\mathbf{X}}(\mathbf{z})}$. Note that in this case the sensitivity attack is straightforward, since it can be done componentwise.

On the other hand, in [110] the received signal $\mathbf{z}$ is quantized with a lattice Λ and the decision is made upon the squared norm of the quantization error.

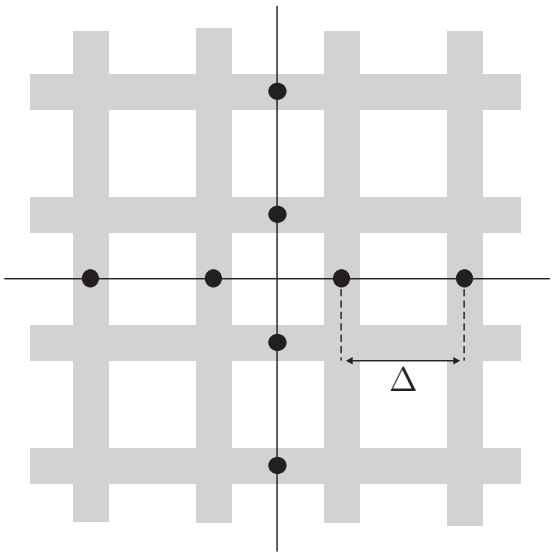


Figure 3.20: OR Region for QPD.

Formalizing it, we can write $f(\mathbf{z}) = ||\mathbf{z} \mod \Lambda||^2$, and g is the identity function again. In this way, the acceptance region is the union of L_1-dimensional hyperspheres centered at the centroids of Λ. From the point of view of attacking such a system, this decision region assures that the attacker can produce a signal yielding H_0 by adding *any* noise vector with a given variance, as far as that noise vector is independent of the self noise. Therefore, a sensitivity attack is not really necessary in this case.

Another approach to this problem is Quantized Projection based Detection (QPD) [126], where uniform scalar quantizers are used to quantify an L_3-dimensional projected version of the received signal $\mathbf{z}$ and the detection function depends on the quantization error, introducing two different strategies: the AND and OR detection regions, which can be formalized as

$$\begin{aligned} f_i(\mathbf{z}) &= \sum_{j=1}^{L_1} a_{ij} z_j,\ 1 \leq i \leq L_3, \\ g_{\text{AND}}(f(\mathbf{z})) &= \max_{1 \leq i \leq L_3} |(f_i(\mathbf{z}) \mod \Delta) - \Delta/2|, \text{and} \\ g_{\text{OR}}(f(\mathbf{z})) &= \min_{1 \leq i \leq L_3} |(f_i(\mathbf{z}) \mod \Delta) - \Delta/2|, \end{aligned}$$

where Δ is the quantization step, a_{ij} are the secret projection matrix coefficients and L_3 the dimensionality of the projected subspace. The resulting methods are denoted as *QPD-AND* and *QPD-OR*, and the obtained decision regions are plotted in Figure 3.19 and Figure 3.20 for $L_3 = 2$. The convergence of the algorithm introduced in Section 3.4.3 for finding the optimal attacking vector will be very much slower for the OR region, since the cost function has its minimum

value at a non-differentiable point of g_{OR}. In fact, in such case we will follow a different strategy in which we try to estimate the L_3 projecting vectors; this implies the complete disclosure of the secret key, and the optimal attacking vector can be computed as the sum of those vectors.

3.4.4.3. Comparison

This section shows the robustness of the watermarking methods introduced above to the BNSA attack. Experiments were carried out for both synthetic and real images, considering two different measures of robustness:

1. The ratio ρ (in dB) between the power needed to achieve an unwatermarked signal and the embedding power. If the watermarked image is $\mathbf{y} = \mathbf{x} + \mathbf{w}$ (with $\mathbf{x}$ the host and $\mathbf{w}$ the watermark) and the attacked image is $\mathbf{z} = \mathbf{y} + \mathbf{t}$ (with $\mathbf{t}$ the output of BNSA), then

$$\rho = 10 \log_{10} \left(\frac{||\mathbf{t}||^2}{||\mathbf{w}||^2} \right).$$

   The average of ρ over a large number of realizations is a good indication of the robustness of each method against BNSA when used to remove the watermark. This will be the robustness measure used for synthetic images.

2. The Peak Signal to Noise Ratio (PSNR). In the case of watermark removal, we will measure PSNR$(\mathbf{y}, \mathbf{z})$, whereas for the case of creating forgeries the measure will be PSNR$(\mathbf{x}, \mathbf{z})$.[11] This will be the robustness measure used for real images.

3.4.4.4. Synthetic images

In this case, the host images are random vectors generated according to a Gaussian distribution, except when the ML detector for the Generalized Gaussian host is used, where we have chosen the shape parameter to be 0.5. In order to make a fair comparison, the value of the probability of false alarm P_{fa} has been fixed to 10^{-4}, $L_1 = 2048$ and the document to watermark ratio to 16 dB (with $\sigma_W^2 = 1$) in order to ensure a reasonable probability of missed detection for all the studied methods. We have used the version of BNSA with the diagonal estimate of the Hessian.

[11] Be aware that in both cases the PSNR is given by $10 \log_{10} \left(\frac{(L_1 \cdot 255)^2}{||\mathbf{v}||^2} \right)$, where $\mathbf{v}$ is the vector added to the considered image, i.e. in the watermarking removal problem we have $\mathbf{v} = h_{\mathbf{y}}(\mathbf{s})$, whereas in the problem of computing forgeries, $\mathbf{v} = \mathbf{y} + h_{\mathbf{y}}(\mathbf{s}) - \mathbf{x}$.

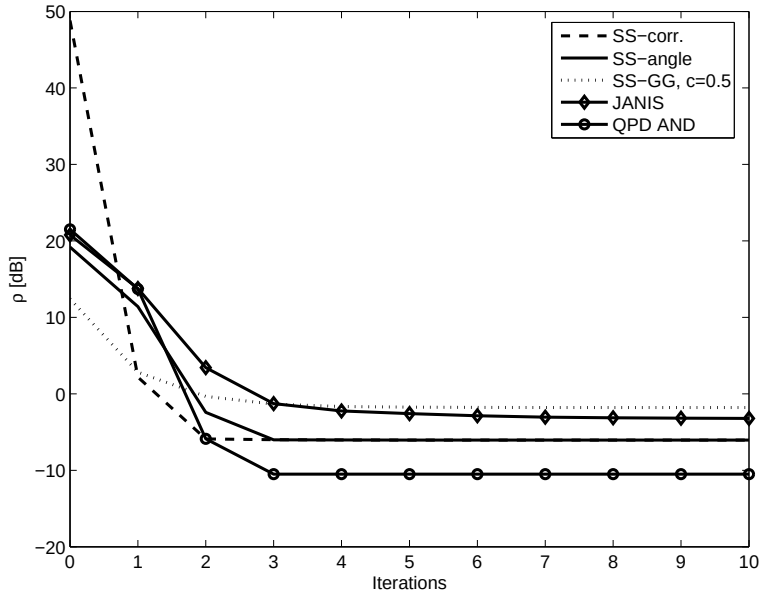


Figure 3.21: ρ averaged over 100 watermarked Gaussian vectors as a function of the number of iterations, for different decision regions: Add-SS based on a hyperplane (SS-corr), Add-SS based on the angle (SS-angle), Add-SS ML detector for Generalized Gaussian distributed hosts with shaping parameter $c = 0.5$ (SS-GG), JANIS and QPD for AND regions (QPD-AND). Iteration 0 corresponds to random attacking vectors (without applying the proposed algorithm).

Figure 3.21 shows the robustness measure ρ versus the number of iterations of BNSA, for each of the considered watermarking methods. Iteration 0 means that no BNSA attack was applied, and as such it represents the value of ρ when the attacking vectors are randomly generated.

It can be seen that at iteration 0 SS-corr is much more robust than SS-angle, but the robustness of both methods converge to the same value when the number of iterations is increased. The most robust method against BNSA turns out to be SS-GG with shape parameter $c_k = 0.5$, even when the power required for producing an unwatermarked signal is reduced in 11.8 dB after just 3 iterations, achieving its minimum at -1.79 dB. Close to this result are the -3.21 dB needed by JANIS, for which the power required to produce an unwatermarked signal is reduced in 24 dB after 10 iterations.

For QPD-AND with $L_3 = 10$, as soon as one of the projecting vectors has been estimated, its robustness against BNSA is significantly smaller than that of the methods commented above. Finally, QPD-OR with $L_3 = 10$ (not plotted in Figure 3.21) shows the smallest robustness among the considered methods, since the value of ρ after 10 iterations is only -38 dB.

3.4.4.5. Real images

In order to reduce the computational cost, the updating algorithm described in (3.47) was implemented in this case. Once again, the probability of false alarm P_{fa} was fixed to 10^{-4} for all watermarking methods, for the sake of fairness.

3.4.4.5.1. Watermark removal.

Image Lena 256×256, i.e. $L_1 = 65536$, was watermarked using the above described methods with an embedding PSNR, i.e. PSNR$(\mathbf{x}, \mathbf{y})$, of 38.58 dB (with $\sigma_W^2 = 9$). The results are summarized in Table 3.1, which shows the value of PSNR$(\mathbf{y}, \mathbf{z})$ for one iteration of BNSA. It can be seen the similar behavior of SS-corr and SS-angle, and on the other hand JANIS appears to be most robust method, supporting the conclusions drawn in Section 3.4.4.4. For illustration purposes, the image watermarked with JANIS is plotted in Figure 3.22, whereas the unwatermarked result of applying BNSA can be seen in Figure 3.23. Indeed, one can observe that the quality of the unwatermarked image obtained by BNSA is better than the quality of the watermarked one, meaning that BNSA is successful in estimating the actual embedded watermark, i.e. the attacked image is closer to the original host image than its watermarked version. This is clearly reflected on PSNR$(\mathbf{x}, \mathbf{z})$, which takes a value of 58.48 dB.

Figure 3.24 is included in order to provide an intuitive idea of the robustness of JANIS against blind attacks, showing the result of applying to the watermarked signal Additive White Gaussian Noise (AWGN) with the power necessary to push the watermarked signal out of the detection region. By comparing this figure to Figure 3.23, one can realize that strong robustness against additive noise attacks does not imply at all robustness in the case of smarter attacks, as those represented by BNSA. The method labeled as *Trellis-based [115]* in Table 3.1 is considered here for similar reasons: it stands for the Trellis-based side-informed method proposed in [115], well known for its robustness against AWGN attacks; however, it is somewhat surprising to see that an unwatermarked version of Lena was obtained with a PSNR$(\mathbf{y}, \mathbf{z}) = 52.17$, just after one BNSA iteration. In any case, one must be aware that this method was firstly proposed for decoding scenarios, not for detection; we have adapted it to detection by comparing the decoded message with a reference one. Furthermore, the reported results could vary depending on the chosen parameters: in this case, the parameters were fixed to a spreading factor of 4, 16 states, and 8 arcs per state.

3.4.4.5.2. Generation of forgeries.

For the creation of forgeries, the original unwatermarked signal $\mathbf{x}$ was chosen to be Baboon (Figure 3.25), whereas the reference watermarked signal is Lena.

Figure 3.22: *Lena* watermarked by JANIS. $\mathrm{PSNR}(\mathbf{x}, \mathbf{y}) = 38.58$ dB.

Figure 3.23: *Lena* watermarked by JANIS and attacked by BNSA. $\mathrm{PSNR}(\mathbf{y}, \mathbf{z}) = 38.31$ dB.

The results shown in this section were obtained after 10 BNSA iterations, and are summarized in Table 3.2. We want to highlight here the result obtained for SS-GG, for which the detection regions are modified when the host statistics change, complicating the optimization problem: the resulting forgery can be found in Figure 3.26. Likewise, the forged Baboon for JANIS detector is represented in Figure 3.27.

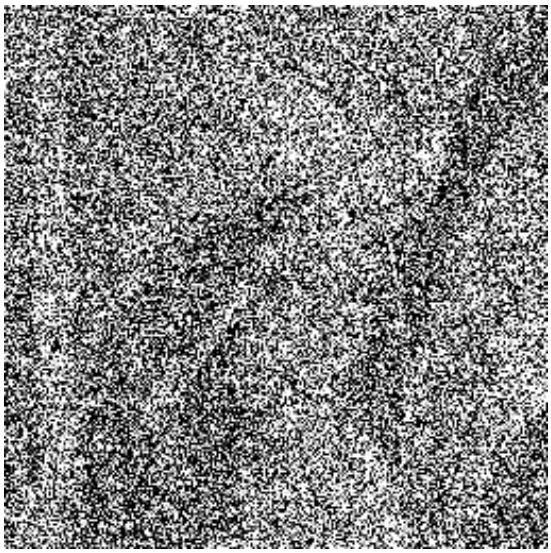


Figure 3.24: *Lena* watermarked by JANIS and attacked by AWGN.

	SS-corr	SS-angle	JANIS	Trellis-based [115]
PSNR($\mathbf{y}, \mathbf{z}$) (dB)	40.25	40.25	38.31	52.17

Table 3.1: Values of PSNR($\mathbf{y}, \mathbf{z}$) for Lena image, after one iteration of BSNA for different watermarking methods.

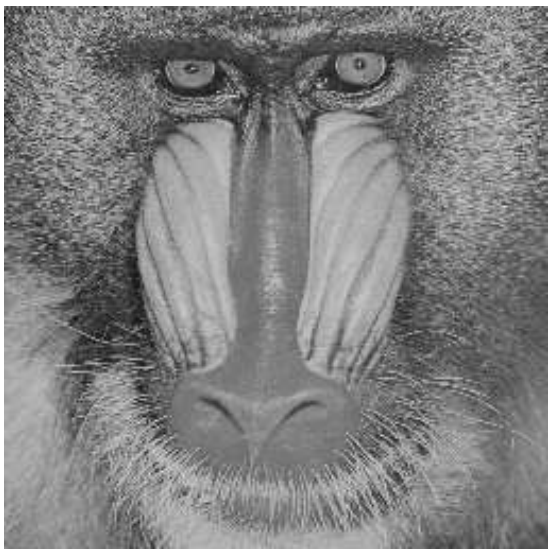


Figure 3.25: Original *Baboon* 256×256.

3.4.5. Computational complexity

One question that needs to be answered is the amount of calls to the detector that are required to achieve the results presented so far. With this aim, the num-

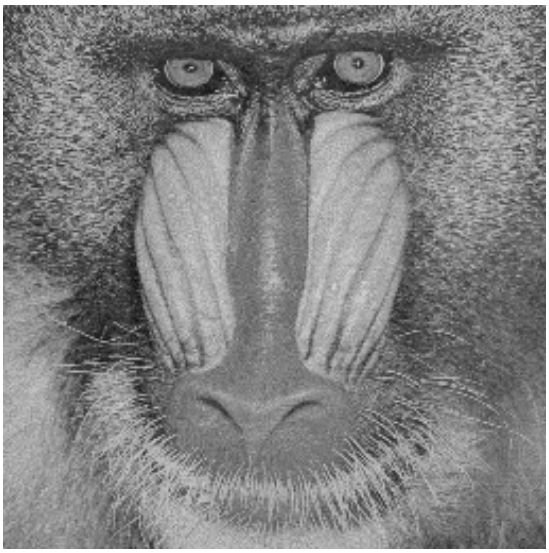


Figure 3.26: Forgery of *Baboon* for the ML detector of Generalized Gaussian distributed hosts. $\mathrm{PSNR}(\mathbf{x}, \mathbf{z}) = 32.03$ dB.

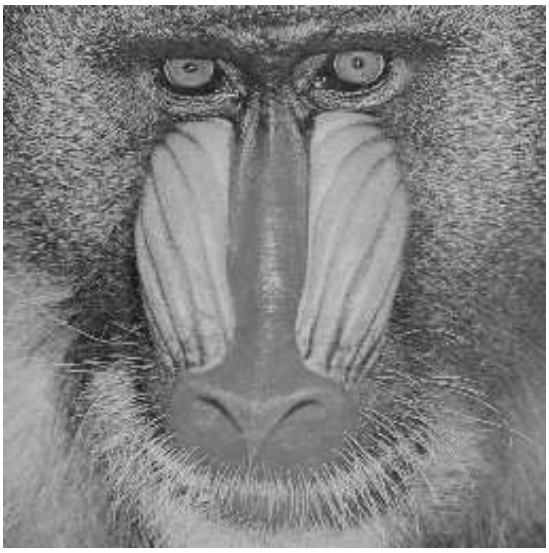


Figure 3.27: Forgery of *Baboon* for JANIS. $\mathrm{PSNR}(\mathbf{x}, \mathbf{z}) = 50.69$ dB.

ber of calls per dimension needed to perform one iteration of BNSA was averaged for all the images considered in Section 3.4.4.5. Interestingly, this number was about 58 for all the methods analyzed in this section, independently of whether the attack consisted in removing the watermark or forging watermarked images. In this computation we used the initialization step and the definition of $h_{\mathbf{y}}(\cdot)$ proposed in Section 3.4.3.1, whereas matrix $\mathbf{B}_k$ was set to the identity and Armijo's rule was chosen for the line search. The tolerance for the bisection algorithm

	SS-corr	SS-angle	SS-GG	JANIS
PSNR($\mathbf{x}, \mathbf{z}$) (dB)	55.73	55.23	32.03	50.69

Table 3.2: Values of PSNR($\mathbf{x}, \mathbf{z}$) for the forgeries of Baboon, after 10 iterations of BSNA for different watermarking methods.

was set to 10^{-12}. This implies that each iteration of BNSA for our test images needs about 3.8×10^6 calls to the detector. Although this large number could be frightening at first sight, the attack is still practical, as it is confirmed by the experimental part of this work. In any case, note that the number of calls to the detector required by the non-iterative method proposed in [67] will be similar to that needed by one iteration of BNSA, since in both cases L_1 boundary points must be located using a binary search (bisection algorithm).

As a final remark, note that our conclusions do not necessarily generalize to other schemes; in fact, one possible countermeasure against BNSA could be based on the design of detection regions which need a large number of iterations of BNSA to converge to an acceptable solution.

3.4.6. Final remarks

Following are some guidelines on how to measure the robustness of watermarking methods against BNSA, and the application of BNSA to new scenarios:

- Although ρ can be seen as a measure of the robustness of a watermarking method against BNSA, note that this measure does not provide full information on the behavior of a particular method; for instance, QPD methods, which have been shown here to be quite weak against BNSA, have a very good Receiver Operating Characteristic in AWGN channels (see [126] for a comparison with SS-corr).
- Taking into account that it just needs the binary output of the detector, the BNSA is also suitable for detectors based on zero-knowledge protocols [5], where, at the end, regardless of the domain where the detection function is computed, the detector will output a binary decision which can be used by the proposed algorithm to estimate the underlying detection region. The only difficulty that could emerge, is related to the bandwidth and computational cost required by these methods; but this is their limitation, not a problem intrinsic to the BNSA.
- The detection regions based on fractal curves proposed in [111] are also suitable for being attacked by BNSA. Although the decision boundaries are non-parametric, the attacker could try to estimate their envelope, since

this estimation will be usually enough to remove the watermark or create forgeries. From a practical point of view, the envelope can be estimated by computing (3.45) and (3.46) with a large value of δ.

- As a final remark, the approach presented in this section can be also used in the case of data-hiding systems, since the decoding process is nothing but a multiple hypothesis test. In this case, any change of the decoder output should be interpreted as if it were done by a change in the detector output; this is equivalent to have the following binary hypothesis: a) the decoded message is changed; b) the decoded message is unaltered.

3.5. Game Theoretic Approach

In this section Add-SS, DC-DM with uniform scalar quantizers and repetition coding and scalar STDM are analyzed from a game-theoretic point of view, using the probability of bit error as the payoff. The theoretical expressions for the BER obtained in the previous sections are optimized to derive the strategies for both the attacker and the decoder, assuming that the embedder simply follows point-wise constraints given by the perceptual mask. Experimental results supporting our analyses are also shown, with examples of watermarking in the spatial domain as well as the block DCT domain.

3.5.1. State-of-the-art

In the literature there is a number of works dealing with watermarking from a game-theoretic approach, e.g. [34, 122, 121, 119, 142, 143]. In this section we will recall the main results of these works, as well as the framework and assumptions they are based on.

3.5.1.1. The Gaussian Watermarking Game

One of the most relevant works is [34], where the embedding and decoding functions, as well as the distributions of the original host signal and of the secret key are assumed to be public. Since the embedding and decoding functions are fixed before the attack, the Maximum-Likelihood decoder, which requires knowledge of the attack, is not included in this scenario; therefore, it can be considered a conservative approach. On the other hand, attacking noise is modeled as a deterministic mapping which depends on an attacker's key and the watermarked signal; the attacking distortion is measured with respect to the watermarked signal. Both the blind and non-blind watermarking cases (denoted respectively

as public and private) are studied in different frameworks. The payoff function used throughout most of the paper is the so-called *Coding Capacity*, which is defined as the supremum of all achievable rates (of reliable transmission); the main conclusions of this paper are:

- Reliable transmission is not possible in either version of the game if average distortion constraints are taken into account, instead of *almost surely* distortion contraints.
- A Gaussian host signal yields the highest coding capacity: the embedder takes advantage of the uncertainty of the host signal to transmit the watermark, and it is well-known that the Gaussian is the distribution with the highest entropy for a given variance.
- When the original host signal is an i.i.d. Gaussian sequence with zero mean (Gaussian watermarking case), the capacity of the game is achieved for both the blind and non-blind scenarios.
- Costa's *"Writing on dirty paper"* [50] can be regarded as a particular case of the Gaussian watermarking game, where the attacking noise is an i.i.d. Gaussian sequence independent of $\mathbf{Y}$.
- The *additive attack watermarking game*, where the attacking noise sequence is independent of the watermarked signal, is shown to be suboptimal for the attacker. The authors showed that the capacity of Costa's *"Writing on dirty paper"* can be achieved for both the blind and non-blind scenarios, independently of the distribution of the noise, when the host signal is Gaussian.[12] This result can be viewed as an extension of Costa's result, since the noise sequence distribution is arbitrary, instead of being an i.i.d. Gaussian sequence, and its distribution is unknown to the embedder and decoder; nevertheless, it also shows that the most harmful additive attack for the watermarking game is an i.i.d. Gaussian sequence.
- The payoff function could also be some mutual information based measure, yielding the so-called mutual information games; in this case, the game is played between the embedder and attacker, having the last one full knowledge of the strategy followed by the former.
- As long as the original host signal is a power-constrained ergodic process noncausally known to the encoder, and the channel noise is a stationary Gaussian process not known to either the embedder or the decoder, being both of them independent and independent of the watermark, Costa's result is applicable.

[12] Be aware that this is a sufficient condition, since Erez et al. later showed that this rate can be achieved in the blind scenario independently of both host and noise distributions [70].

3.5.1.2. Information-Theoretic Analysis of Information Hiding

Another outstanding paper dealing with watermarking from a game theoretic point of view is [122]; in this work the authors state the information-hiding game for finite alphabets, and afterwards they generalize it for the infinite case. Similarly to the approach followed in [34], the attacker is assumed to know the embedding function and the distribution of all the random variables, but not the secret key; nevertheless, in this case the decoder will be designed taking into account not just the embedding function, but also the attacking strategy, so the ML decoder can be considered. Other characteristics of their approach are:

- Similarly to [34] the attacks are constrained by taking into account the distortion they introduce with respect to the watermarked signal, although the game using constraints over the distortion introduced with respect to the original host signal is sketched too. Due to the former constraint, the set of feasible attacks depends on the embedding function, whereas the set of feasible decoding functions does not depend on the choice of the embedding function nor on the attacking strategy. Throughout the paper, just memoryless attacks are considered, except for a final section where a blockwise memoryless information hiding problem is explored. Concerning the host signal, it is assumed to be an i.i.d. sequence, and the paper studies both the non-blind and blind scenarios.

- In the proposed information hiding game two cooperative players (embedder and decoder) try to maximize a payoff function which the attacker tries to minimize. This payoff function could be related with the probability of error or the maximum achievable rate of reliable transmision; the last one was chosen by the authors. The results obtained for the unidimensional case and continuous alphabets coincide with those in [34], achieving the maximum for Gaussian hosts, and they are generalized for the multidimensional case using a sphere-packing argument. If the host signal is zero-mean non-Gaussian distributed with a given variance, then the hiding capacity is upper-bounded by that obtained for the previous (Gaussian) scenario. Following this approach both Add-SS and DC-DM information-hiding capacities are analyzed.

- An interesting by-product of this analysis (performed for continuous alphabets in the unidimensional domain) are the expressions of both the optimal embedding and the attacking strategies when the host is Gaussian and the distortion measure is the squared Euclidean norm. For the studied unidimensional case, when the host is available at the decoder, the embedder consists in scaling the original host signal and adding a Gaussian signal independent of the original host, whereas the optimal attack is the so-called Gaussian test channel, from rate-distortion theory [51], which also scales

the watermarked signal (with a different scaling factor) and adds a Gaussian signal independent of the watermarked signal. In the blind data hiding scenario both embedding and attacking strategies coincide with those described above, with a change on the scaling factor used by the embedder; in any case, this does not affect the capacity of the system.

- Surprisingly, in the small-distortion regime, i.e. when both the allowed embedding and attacking distortions are much smaller that the host variance, the Additive White Gaussian Noise (AWGN) attack is asymptotically optimal; furthermore, the hiding capacity is shown to be asymptotically independent of the distribution (including the variance) of the host signal.
- A final conclusion is that the attack trying to recover the original host signal using the maximum a posteriori estimation rule, which yields a Wiener filtering, does not reduce at all the hiding capacity, given that the optimal decoder is used.

3.5.1.3. The Parallel-Gaussian Watermarking Game

In [121] the authors follow an approach similar to that in [122], extending some of the results presented therein. Some of the main results are:

- Perhaps the most apparent difference with respect to [122], is that in all the results exposed in [121] the attacking distortion is measured with respect to the original host signal, instead of taking into account the watermarked signal. Despite of this difference, the embedding and attacking strategies, as well as the hiding capacities, are similar to those in [122].
- Trying to generalize the previous results, the authors propose to model the multidimensional host signals as independent parallel-Gaussian channels (also named sources). They show that given the embedding and attacking powers devoted to each channel, the optimal strategy for a given channel is that obtained in the previous unidimensional case, and independent of the strategies followed on the other channels. Therefore, the problem is simplified to just finding the optimal allocation of the embedding and attacking powers; to this end, a numerical optimization algorithm is provided. Furthermore, the asymptotical behavior of channels with both large and small host signal powers is studied.

  If the correlation matrix of the host signal is not diagonal, i.e. if its dimensions are correlated, the solution to the watermarking game is achieved by diagonalizing the correlation matrix with the Karhunen-Loève transform, in such a way that the problem is converted again to that of independent parallel Gaussian channels described above. Furthermore, if the host signal is non-Gaussian, its hiding capacity is upper bounded by the hiding capacity

obtained for a Gaussian host with the same correlation matrix. Finally, it is shown that the data-hiding capacity for Gaussian hosts is strictly reduced by correlation.

- If the host signal can be modeled as a stationary Gaussian process with bounded and continuous spectral density, then the watermarking game is redefined as a stationary-Gaussian watermarking game. The solution to this problem is just the extension of the previous results for parallel-Gaussian channels to the continuous case, i.e. the role previously played by the power of a given channel, is now played by the spectral density of the host signal. Taking into account this difference, the power allocation (i.e. the spectral densities of both the watermark and the attack) is obtained following the methodology used for the parallel-Gaussian channels.

3.5.1.4. The Zero-Rate Spread-Spectrum Watermarking Game

The two previous works ([122] and [121]) have a clear influence on [119], where the authors analyze the game among embedder, attacker and decoder/detector for spread-spectrum watermarking schemes. Next we have described some of the peculiarities and main results of their analysis:

- In this work, both the attacker and the decoder/detector know the distributions of the original host signal, the message, and the secret key. As in [122] and [121], the decoder/detector is assumed to know the attacking strategy; this enables the study of the scenario where the decoder/detector uses the Maximum a Posteriori (MAP) decision rule, which is the optimal strategy for the decoder/detector. Therefore, the game is played again just between the attacker and the embedder; the first one tries to maximize a pay-off function (in this case the probability of error) whereas the second tries to minimize it. The strategies of embedder and attacker have to verify a power constraint, which in the case of the attack will take into account the distortion introduced with respect to the original host signal (similarly to [121]).

- The authors obtained the solution for the scenario with the following characteristics: a) zero-mean Gaussian host signal, b) the watermarked signal is computed as the result of filtering the addition of the host signal and a zero-mean Gaussian random vector independent of the host signal, and c) the attacked signal is computed as the filtering of the watermarked signal and the subsequent addition of a zero-mean Gaussian random vector independent of the watermarked signal. Under those assumptions the MAP decoder is derived, yielding a generalized version of the widely extended correlation-based decoder. Taking into account this analysis, the authors

study the detection problem,[13] replacing the probability of error by the averaged distance as pay-off function. For the scalar case, both the embedding and attacking solutions are the above introduced Gaussian test channel, which is a non-additive strategy. For the vector case, the authors restrict their attention to diagonal processors on the Karhunen-Loève Transform (KLT) coefficients; a numerical method is provided to compute the solution.

- As a special case of the previous scenario, the use of additive watermarks and attacks is studied. The obtained solution is a *waterfilling* strategy [51], where the attacker spends his/her power in ruining the components with a small host interference, since those with a large host interference are already almost not useful before attacking them. On the other hand, the embedder will spend his/her power in those components with a smaller host interference. The authors conclude that the additive noise attack is much less harmful than that combining filtering and additive noise, although in small-distortion regime both the additive watermarks and attacks asymptotically approach the optimal strategy; this result agrees with a similar one obtained in [122].

3.5.1.5. Works by Somekh-Baruch and Merhav

In [142] and [143] Somekh-Baruch and Merhav studied the games of private and public watermarking respectively. Some of their results are summarized below:

- In [142] two pay-off functions are considered, yielding to two different games: the error exponent and the coding capacity. In both cases the host takes values from a finite-alphabet (discrete) memoryless stationary source, contrarily to previous works where the host was assumed to be Gaussian [34]; the authors named *private game* to the fact of being the host available to the decoder, i.e. non-blind decoding. The distortions are measured with a general function, not necessarily the squared Euclidean distance; furthermore, the attacking distortion is measured with respect to the watermarked signal. The probability of non verifying the constraint on the attacking distortion decreases exponentially with the dimensionality (the authors named it *large deviations distortion constraint*); the almost-sure constraint of Cohen and Lapidoth's approach [34] can be seen as a particular case of this new formulation. Concerning the game itself, the attacker is aware of embedder

[13] The decoding problem, i.e. multiple hypothesis testing, is considered at the end as a generalization of the detection analysis, taking into account the union bound. The main conclusion is that most of errors will be due to the unwatermarked-watermarked decision, not to mistaking the embedded message.

and decoder strategy, but the decoder does not know attacker's strategy; this obviously constitutes a pessimistic approach. In this framework, the structure of the embedder, worst case attack and decoder solutions to the error exponent game are introduced, showing that the order the game is played does not affect to its result (i.e., a saddle point is achieved). Finally, the capacity of the watermarking game in the described scenario is found.

- The same authors studied the problem of *public* (i.e., blind) watermarking in [143]. The scenario is basicly that studied in [142], but for the fact that the host now is not availabe to the decoder, and the *large deviations distortion constraint* is replaced by the almost-sure one. One of the novelties of this paper compared with the previous ones in the literature is the lack of an upper bound on the cardinality of the alphabet of the auxiliary random variable the transmission is based on (i.e., the **U** of Costa and Gel'fand-Pinsker). Following authors' explanation, this is due to the game between attacker and embedder-decoder studied in this framework, oppositely to Gel'fan-Pinsker model, where the channel is a known fixed one. Comparing this result with other ones in the literature where the aforementioned bound does appear (see [122]), the authors claimed that that paper has a mistake.

3.5.1.6. Works by Le Guelvouit, Pateux and Guillemot

A game theoretic approach was also followed by Le Guelvouit, Pateux and Guillemot in some of their works, which we will briefly described here:

- In [86] the authors studied Add-SS from a game theoretic point of view. In this work, the authors modeled the host signal as a vector of independent non identically distributed Gaussian random variables. In the considered framework the attacker, who is aware of the embedding strategy, is constrained to just scale the watermarked signal and to add it AWGN; the attacking distortion is measured as a weighted squared Euclidean norm of the difference between the original host signal and the attacked one. The pay-off of the game is the probability of error when the MAP detector is used; therefore, the detector is out of the game. In the obtained solution, three regions are differentiated, corresponding to different attacks: the *erase attack*, where the signal is set to **0**; the *Wiener filtering attack*, where no noise is added, but the watermarked image is filtered using Wiener filter; and finally, the *intermediate attack*, which is composed of the filtering and addition of additive Gaussian noise. Taking this result into account, the optimal watermark power allocation is computed. One of the main novelties of this work is that the watermarked signal is proposed to undergo a Wiener filter after embedding, in order to reduce the embedding distortion, showing the improvement of that strategy.

- This framework was generalized in [87], allowing to consider the reduction of the host signal interference, as well as the reduction of the interference due to carriers conveying other symbols; an important consequence of this generalization is that STDM-like methods can be analyzed. The structure of the decoder is also modified; now, a linear correlator is considered, and the authors optimize the weight of each dimension. The same three attack regions are obtained for this modified framework.
- These ideas are also the fundamental of [125], where further discussions about the dimensionality of the projected domain of STDM, and a geometrical interpretation of watermarking with side information are also introduced.
- Finally, in [85] the authors particularize the results obtained in [87] to the case of no interference due to the host signal, neither to the other carriers. A dirty paper trellis based code, that recalls that presented in [115], is proposed to reduce the host signal interference; similarly, some strategies are introduced aimed at eliminating the interference due to other carriers.

3.5.1.7. Works by Su, Eggers and Girod

In [145] and [144] Su, Eggers and Girod dealt with the best distribution for the watermark, as well as with the optimization of certain kind of attacks. Some characteristics of their approach are:

- In [145], the authors study the problem of watermark detection (no decoding) based on linear correlation; in the proposed scenario, the detector is fixed (out of the game), so it does not compensate the attack. The target of the attacker is to minimize the MSE distortion (measured with respect to the original host signal) of the attack needed to obtain a signal with a correlation value below a given threshold. An important characteristic of their approach is that the attacker designs his/her strategy taking into account the power spectrum of the watermark, meaning that the attacker has the final word (this is a *maxmin* problem). Furthermore, Su and Girod just consider attacks which try to remove the watemark by linear, shift-invariant filtering, and additive noise. The authors show that the optimal attack consists of estimating the watermark using the Wiener filter, and then subtracting a scaled version of this estimate from the watermarked signal. The optimal attack should not introduce any additive noise; this occurs because the fixed correlation detector does not change trying to compensate the attack. Finally, as a countermeasure, the watermark power spectrum that maximizes the target distorsion is computed. The result is nothing but a scaled version of the power spectrum of the original host signal, yielding the so-called *power-spectrum condition* (PSC).

- In a later work [144],[14] Su, Eggers and Girod study a problem similar to that in [145], but considering the decoding scenario. In the followed approach the decoder also plays the game, and it is assumed to perfectly know the attacker strategy. In turn, the attacker is assumed to perfectly know the embedder strategy; therefore, this is a *max-min-max* problem. In the proposed approach, the target of the attacker is to minimize the channel capacity constrained to a given attacking distortion, whereas both the embedder and the decoder try to maximize that capacity verifying a constraint on the embedding distortion. These distortions are measured using MSE and frequency-weighted MSE, in both cases with respect to the original host signal. Moreover, similarly to [145], the attacker is constrained to use linear, shift-invariant filtering and Additive Colored Gaussian Noise. Concerning the interference due to the host, the two extremes are studied: conventional blind decoding (where the system does not take advantage of the knowledge of the original host), and optimal blind decoding (where the host interference is completely rejected). The obtained optimum attack can be roughly characterized as adding noise at low attacking distortions, and discarding frequency components at high attacking distortions. On the other hand, due to the active role of the decoder, the PSC previously introduced does not longer provide the optimal power spectrum distribution of the watermark. Therefore, the authors use iterative numerical methods to compute this optimal distribution, alternately re-optimizing the watermark power spectrum and the optimum attack. The main conclusion of these optimizations for the case of optimal blind decoding is that at low attack distortions white watermarks are nearly optimal, while at high attack distortions, PSC-compliant watermarks are almost optimal. On the other hand, for the conventional blind decoding a trade-off between the robustness to the attack and the robustness against the interference due to the host signal is observed. In this way, for low distortions the optimal watermark designing strategy does not allocate power on those frequencies with largest values of power spectrum of the host, but as the attacking distortion becomes larger, the watermark becomes more PSC-compliant.

3.5.2. Our approach

The main purpose of this section will be to obtain improved decoding and attacking strategies (since attacker and decoder will be the only agents involved) for three of the main data hiding methods, namely, Add-SS, DC-DM with uniform scalar quantizers and repetition coding, and scalar STDM, using the symbol decoding error probability as pay-off function. We have chosen those algorithms because of their widespread use.

[14] Although [144] was published in 2001 and [145] in 2002, they were respectively submitted in 2000 and 1999.

The main differences between our approach and those described above are:

- The pay-off function typically studied in the literature is the information hiding capacity, with the remarkable exception of [119], where the probabilities of detection and decoding errors are used.
- The agents involved: in [34] embedder, attacker and decoder played the game, whereas in [122], [121] and [119] just the embedder and the attacker take part on it, since the decoder is assumed to be the optimal one, following a ML or MAP rule.

This last point could be someway criticizable, given that the decoder will be usually constrained to have reduced complexity and cost, so the decoding strategy should be as simple as possible; this requirement obviously collides with the exact estimate of the attack channel needed for the aforementioned decoding strategies. In this sense, the target of this section will be the optimization of some decoding parameters which are intended to improve the decoding performance at the cost of just a reduced complexity increase. With these decoding parameters, we are trying to modify the usual decoders the least possible. This is feasible because the proposed decoders have the same structure as the usual ones, and their improvement is just based on such those optimized parameters, that will simply weight the decoding criterion (usually the squared Euclidean distance between the attacked signal and the codewords) at each dimension depending only on the variance of the attack channel, the host signal, and the watermark. Obviously, the resulting strategy will not be the optimal one, but in any case the improvement achieved by this novel approach is shown to be substantial compared with the strategies typically used so far. In fact, in order to distinguish the obtained strategies from other suboptimal ones, we will termed them *optimal*, but the reader should be aware of this semantic licence.

On the other hand, the embedder, which in the previous works always played an active role, is kept out of the game in our approach (an exception was made for the scalar STDM analysis, as it will be described below). All the above described works in the literature use the squared Euclidean distance computed over the complete signals to constrain the embedding strategies. Nevertheless, recalling the discussion in Section 2.2.1, the MSE measures do not have a perceptual meaning, so we have preferred to establish a coefficient-wise constraint for the embedder, in such a way that the watermark power corresponding to the i-th coefficient will verify $\sigma^2_{W_i} \leq \gamma_i^2$, with $1 \leq i \leq L_1$. Taking the last inequality into account, we will assume that the embedder will just allocate as much power as possible in all the signal coefficients, i.e. $\sigma^2_{W_i} = \gamma_i^2$ for $1 \leq i \leq L_1$, independently of the strategies followed by attacker and/or decoder; therefore, the embedder is left out of the game.

One could think of following the same strategy with the attacker, i.e. allocating as much power as possible in each coefficient. Nevertheless, we have discarded

the study of such an option because it would yield to a trivial solution. Moreover, the distortion constraint seems to be weaker for the attacker than for the embedder, so a more relaxed constraint, as the MSE one, could make sense for the attack. This last consideration is based on the fact that the watermarked signal should have a high quality, as the buyer/user is going to pay for watching/listening it, and he/she will not usually allow a reduction in the quality of the content; contrarily, the attacker could sell the attacked media to less exigent clients, who would allow a reduction in the quality, if it implied a reduction in its price.

Taking this discussion into account, we have set a double power constraint for the attacking vector. On one hand, the MSE introduced by the complete attacking vector will be constrained, as it was done in the above referenced works; this constraint can be written as $\frac{1}{L_1}\sum_{i=1}^{L_1}\sigma_{N_i}^2 \leq D_c$. One the other hand, trying to overcome the lack of perceptual meaning of the MSE measures (see the discussion in Section 2.2.1), we will avoid that the attacking power devoted to a given coefficient were larger that the mean attacking power corresponding to the set of coefficients which convey a symbol, i.e. $\sigma_{N_i}^2 \leq L_2 D_c$, for $1 \leq i \leq L_1$. Furthermore, we will constrain the set of feasible attacks to the set of additive random vectors, with mutually independent components and also independent of the watermarked signal. In this way, the attacker will just have freedom to choose the noise distribution and to allocate the attacking power to each dimension, constrained to the above introduced conditions. In general, this will be a suboptimal choice for the attacker. Nevertheless, our interest in studying this attacking strategy is justified by its reduced complexity, the wide use of additive noise attacks in the literature as a tool to measure watermarking methods robustness, as well as by the results in [122] and [119], where the optimal attack in small-distortion regime was shown to be the additive Gaussian one, although the actual scenario is clearly different of those studied in the mentioned works.

Summarizing, the game consists in the maximization/minimization of the probability of decoding error P_e, which is a function of the attack power allocation $\boldsymbol{\sigma}_N^2$ and the decoding weights $\boldsymbol{\beta}$, by respectively the attacker and the decoder, i.e.

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\sigma}_N} P_e(\boldsymbol{\beta}, \boldsymbol{\sigma}_N), \quad \text{or} \tag{3.50}$$

$$\max_{\boldsymbol{\sigma}_N} \min_{\boldsymbol{\beta}} P_e(\boldsymbol{\beta}, \boldsymbol{\sigma}_N). \tag{3.51}$$

The problem in (3.50) is said to be a *minimax* problem, and in that case the decoder plays first than the attacker, that is, the last one knows the strategy followed by the first one when he/she designs his/her strategy. In a certain sense, we can say that the attacker has the final word. On the other had, (3.51) is said to be a *maximin* problem, where the attacker plays before, and the decoder manages to know the attack power allocation, taking it into account when designing the decoding strategy. It can be said that the decoder has now the final word.

The game has a pure (deterministic) equilibrium if the minmax solution equals the maxmin one at a given P_e value (called the value of the game) for some deterministic optimal values $\boldsymbol{\sigma}_N^*$ and $\boldsymbol{\beta}^*$. Then, the payoff function is said to have a saddle-point at $(\boldsymbol{\sigma}_N^*, \boldsymbol{\beta}^*)$. If this happens, the order in which the agents play the game is irrelevant as neither the attacker nor the decoder want to deviate from the most conservative option marked by the saddle-point. Nevertheless, the order is relevant if there does not exist at least one saddle-point.

Finally, we would like to recall that the indices of the coefficients devoted to transmit the i-th symbol will depend on the secret key θ through the permutation $\Pi(\cdot)$ described in Section 2.1. Given that the attacker has not access to θ, he/she will not know which coefficients are used to convey a given bit. Therefore, the attacker will have to renounce to compute the exact probability of error, and afterwards maximize it; due to this, in our analyses we will assume that the attacker will just take into account the averaged channel, and not the particular partition due to a certain value of θ. For the sake of notation, each one of the possible partitions of the host signal will be denoted as $\mathcal{T}$, and $\mathcal{D}$ will be the set containing them.

3.5.3. Additive Spread Spectrum

As it was said in Section 2.3, the most popular Add-SS decoder is that based on the cross-correlation (see equation (2.6)) between the received signal and the spreading sequence, even when this strategy is the optimal one only if both the host signal and the channel noise are i.i.d. Gaussian distributed, $|s_i| = |s_k|$, and $\gamma_i = \gamma_k$, for all $i, k \in \{L_2 \cdot (j-1)+1, \cdots, L_2 \cdot j\}$, with $j = 1, \cdots, L_b$. Furthermore, the generalized version of this decoder using the weighting parameters $\boldsymbol{\beta}$ was given in equation (2.8).

If some technique trying to reduce the interference due to the host signal by using linear filtering were used (as it was suggested in Section 2.3), the correlation value for the j-th subvector, i.e. $r_j \triangleq \mathbf{s}_j^T \cdot \mathbf{z}_j$, with $j = 1, \cdots, L_b$, when $\mathcal{T}$ is known can be modeled as the output of an AWGN channel (see [92]), $r_{i|\mathcal{T}} = a_{i|\mathcal{T}} b_i + u_{i|\mathcal{T}}$, $i \in \{1, \cdots, L_b\}$, where

$$a_{i|\mathcal{T}} = \sum_{k=(i-1)\cdot L_2+1}^{i \cdot L_2} \beta_k h_{k,k} \gamma_k, \; i = 1, \cdots, L_b \tag{3.52}$$

and $u_{1|\mathcal{T}}, \cdots, u_{L_b|\mathcal{T}}$ are samples of an i.i.d. zero-mean Gaussian random process with variance

$$\sigma^2_{u_{i|\mathcal{T}}} = \sum_{k=(i-1)\cdot L_2+1}^{i \cdot L_2} \beta_k^2 \left[x_{f_k}^2 + \sum_{l=1}^{L_1} h_{k,l}^2 \left(\gamma_l^2 + \sigma_{N_l}^2 \right) - h_{k,k}^2 \gamma_k^2 \right], i = 1, \cdots, L_b. \tag{3.53}$$

Since $\mathcal{T}$ is generated by θ, we will assume the attacker does not know it. Therefore, assuming $s_i \in \{-1, +1\}$ for all $1 \leq i \leq L_1$, he/she will try to maximize the probability of error considering the averaged channel, whose statistics for the case of uniform partitions are

$$a = \sum_{\forall \mathcal{T} \in \mathcal{D}} \mathrm{E}(r_i|\mathcal{T})\mathrm{Pr}(\mathcal{T}) = \frac{1}{L_b} \sum_{k=1}^{L_1} \beta_k h_{k,k} \gamma_k \tag{3.54}$$

$$\begin{aligned} \sigma_u^2 &= \sum_{\forall \mathcal{T} \in \mathcal{D}} \mathrm{Var}(r_i|\mathcal{T})\mathrm{Pr}(\mathcal{T}) + \sum_{\forall \mathcal{T} \in \mathcal{D}} \mathrm{E}^2(r_i|\mathcal{T})\mathrm{Pr}(\mathcal{T}) - \left(\sum_{\forall \mathcal{T} \in \mathcal{D}} \mathrm{E}(r_i|\mathcal{T})\mathrm{Pr}(\mathcal{T}) \right)^2 \\ &= \frac{1}{L_b} \sum_{k=1}^{L_1} \beta_k^2 \left[x_{f_k}^2 + \sum_{l=1}^{L_1} h_{k,l}^2 \left(\gamma_l^2 + \sigma_{N_l}^2 \right) - h_{k,k}^2 \gamma_k^2 \right] \\ &\quad + \frac{L_b - 1}{L_b^2} \sum_{k=1}^{L_1} \beta_k^2 h_{k,k}^2 \gamma_k^2 \end{aligned} \tag{3.55}$$

and since L_b will be typically large, $(L_b - 1)/L_b^2$ can be replaced by $1/L_b$. Note that even when the covariance matrix of the averaged channel is no longer diagonal, [92] shows that the cross-covariance terms will be small compared with the diagonal terms, so they can be neglected.

From (2.7), and recalling we are considering the averaged channel

$$\overline{P_e} = \mathcal{Q}\left(\frac{a}{\sigma_u} \right), \tag{3.56}$$

from the attacking point of view, the objective will be to maximize the partition-averaged signal-to-noise ratio given by

$$\overline{\mathrm{SNR}} \triangleq \frac{a^2}{\sigma_u^2} \tag{3.57}$$

while from the decoding point of view, the objective will be to maximize the signal to noise ratio corresponding to the i-th symbol, given to be

$$\mathrm{SNR}_i \triangleq \frac{a_{i|\mathcal{T}}^2}{\sigma_{u_{i|\mathcal{T}}}^2} \tag{3.58}$$

for all $i \in \{1, \cdots, L_b\}$, since the decoder knows the partition which is being used, so he/she knows the probability of error for this partition is

$$P_e = \frac{1}{L_b} \sum_{i=1}^{L_b} \mathcal{Q}\left(\frac{a_{i|\mathcal{T}}}{\sigma_{u_{i|\mathcal{T}}}} \right) \tag{3.59}$$

3.5.3.1. Optimal Decoding Weights for a Known Attack Distribution.

First, we will consider the case in which the attacking-noise distribution is known and determine the optimal decoding weights vector $\boldsymbol{\beta}^*$ that minimizes the BER in (3.59). Substituting (3.52) and (3.53) into (3.58) and inverting the result, we obtain the noise-to-signal ratio for the i-th bit, that the decoder should minimize:

$$\mathrm{NSR}_i = \frac{\sum_{j=(i-1)\cdot L_2+1}^{i\cdot L_2} \beta_j^2 \left[x_{f_j}^2 + \sum_{l=1}^{L_1} h_{j,l}^2 \left(\gamma_l^2 + \sigma_{N_l}^2\right) - h_{j,j}^2 \gamma_j^2 \right]}{\left(\sum_{j=(i-1)\cdot L_2+1}^{i\cdot L_2} \beta_j h_{j,j} \gamma_j \right)^2}, \quad \forall\, i = 1, \cdots, L_1. \tag{3.60}$$

The problem can be solved in a general form to yield the following optimal weights

$$\beta_j^* = \frac{K h_{j,j} \gamma_j}{x_{f_j}^2 + \sum_{l=1}^{L_1} h_{j,l}^2 \left(\gamma_l^2 + \sigma_{N_l}^2\right) - h_{j,j}^2 \gamma_j^2}, \; 1 \le j \le L_1 \tag{3.61}$$

with K any positive constant. Be aware that β_j^* does not depend on the indices of the other coefficients devoted to transmit the i-th symbol, i.e. it does not depend on the chosen partition. This means that the attacker could also compute β_j^* in spite of not knowing the used partition.

3.5.3.2. Optimal Attack for Known Decoding Weights.

In the case that the attacker knows the decoding weights vector $\boldsymbol{\beta}$, his/her problem becomes that of maximizing the $\overline{\mathrm{NSR}}$ in (3.57) subject to an imperceptibility constraint. It can be proven that for a MSE distortion constraint the optimal attack would imply concentrating all the distortion in those coefficients with the largest values of $\tau_j = \sum_{k=1}^{L_1} \beta_k^2 h_{k,j}^2$. Note that this strategy will likely produce visible results and clearly shows that constraining just the MSE may lead to impractical attacks.

3.5.3.3. Optimal Attack When the Decoder Follows the Optimal Strategy.

Now, suppose that the decoder knows which distribution the attacker is using, so that he/she employs the optimal strategy derived in Section 3.5.3.1. In this case, the best an attacker can do is to minimize (3.57) after replacing β_j with (3.61), while satisfying a certain distortion constraint. Therefore, making the

assignments $p_j^2 = x_{f_j}^2 + \sum_{l=1}^{L_1} h_{j,l}^2 \left(\gamma_l^2 + \sigma_{N_l}^2\right) - h_{j,j}^2 \gamma_j^2$, and $q_j = h_{j,j}\gamma_j$ the attacker has to minimize

$$\overline{\mathrm{SNR}} = \frac{\left(\sum_{k=1}^{L_1} \beta_k q_k\right)^2}{L_b \left[\sum_{k=1}^{L_1} \beta_k^2 p_k^2 + \beta_k^2 q_k^2\right]} = \frac{\left(\sum_{k=1}^{L_1} \frac{q_k^2}{p_k^2}\right)^2}{L_b \left[\sum_{k=1}^{L_1} \frac{q_k^2}{p_k^2} + \frac{q_k^4}{p_k^4}\right]}. \tag{3.62}$$

Since $p_j^2 \gg q_j^2$ we may neglect the second term in the denominator, so we can reformulate the problem as the minimization of

$$\varphi \triangleq \sum_{k=1}^{L_1} \frac{q_k^2}{p_k^2} = \sum_{k=1}^{L_1} \frac{h_{k,k}^2 \gamma_k^2}{m_k^2 + \sum_{l=1}^{L_1} h_{k,l}^2 \sigma_{N_l}^2}, \tag{3.63}$$

where $m_k \triangleq x_{f_k}^2 + \sum_{l=1}^{L_1} h_{k,l}^2 \gamma_l^2 - h_{k,k}^2 \gamma_k^2$. Unfortunately, a close look at (3.63) reveals that each particular noise sample exerts influence on several terms of the sum, thus making it difficult the interpretation of the solution. Aiming at producing meaningful results, for the remaining of this section we will make the simplification $\mathbf{H} = \mathrm{diag}(h_{1,1}, \cdots, h_{L_1,L_1})$ which is reasonable in many practical situations: as an example we have closely studied Wiener filtering and made the whole numerical optimization taking into account all the values of $h_{k,l}$ [152], [94]. The results are virtually the same as those we obtained with the proposed simplification. The explanation is based on the fact that the central element of the filter is much larger than the others, so the influence of the latter on the optimization is very small. Therefore, (3.63) becomes $\varphi = \sum_{k=1}^{L_1} \frac{\gamma_k^2}{\frac{x_{f_k}^2}{h_{k,k}^2} + \sigma_{N_k}^2}$ and (3.61) simplifies to $\beta_i = \frac{k\gamma_i h_{i,i}}{x_{f_i}^2 + h_{i,i}^2 \sigma_{N_i}^2}$. As in the previous section, the attack is constrained to meet a condition for the maximum allowed distortion introduced in the image, that is, $D_c \geq \frac{1}{L_1} \sum_{j=1}^{L_1} \sigma_{N_j}^2$ and it must also verify $\sigma_{N_j}^2 \leq L_2 \cdot D_c$. As it was previously explained, this last condition tries to avoid the effect of assigning all the power to a few coefficients. One host image coefficient should not be assigned more power than the average power dedicated to each bit. In this case it can be shown that the optimal attacking distribution is

$$\sigma_{N_i}^{*\,2} = \min\left[L_2 \cdot D_c, \left(\xi\gamma_i - \frac{x_{f_i}^2}{h_{i,i}^2}\right)^+\right], \text{ for all } 1 \leq i \leq L_1 \tag{3.64}$$

where $(x)^+ \triangleq \max\{x, 0\}$, and ξ is a suitably chosen parameter so that

$$\frac{1}{L_1} \sum_{i=1}^{L_1} \min\left[L_2 \cdot D_c, \left(\xi\gamma_i - \frac{x_{f_i}^2}{h_{i,i}^2}\right)^+\right] = D_c. \tag{3.65}$$

This strategy is closely related with the so-called *waterfilling*, which is the solution achieved in the capacity analysis of channels with colored Gaussian noise (see [51]).

Although the analyzed problems are very different, this result is quite similar to the expression obtained in [119], where after the diagonalization by the KLT, the eigenvalues of the noise covariance matrix are

$$\sigma_{N_i}^{*\,2} = \left(\xi_2\gamma_i - \sigma_{X_i}^2\right)^+, \tag{3.66}$$

where $\sigma_{X_i}^2$ is the variance of X_i and ξ_2 a constant such that

$$\frac{1}{L_1}\sum_{i=1}^{L_1}\left(\xi_2\gamma_i - \sigma_{X_i}^2\right)^+ = D_c. \tag{3.67}$$

3.5.4. DC-DM with uniform quantizers and repetition coding

The analysis of DC-DM is clearly more difficult than the previous one. This difficulty is intrinsically due to the non-linear nature of quantization based methods. Nevertheless, in Section 3.1.2.2 analytical values of optimal decoding weights were derived;[15] these values ($\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$) were based on the CLT-based approximations to the probability of decoding error, and took into account the statistics of $\mathbf{U}^+$ or $\mathbf{U}$.

Unfortunately, such an analysis is not possible for the other two scenarios studied in the previous section. For example, when one tries to compute the *optimal attack for known decoding weights*, it is easy to verify that the procedure of building the Lagrangian and equating its derivatives to zero leads to a system of nonlinear equations, which requires numerical methods for solving it. Since this does not shed any light on the strategy that the attacker should follow, we will not develop it any further.

In order to illustrate the complexity of the game-theoretic analysis of this scheme, we will focus on the computation of the *optimal attack when the decoder uses the optimal decoding weights* $\boldsymbol{\beta}^*$, given by (3.13). This problem is rather difficult to solve even in the simplest cases. In fact, in order to obtain analytical results we have limited our analysis to the case where the attacker knows the partition $\mathcal{T}$ (which can be regarded to as a worst case scenario for the decoder), the attack consists on uniform noise with distribution for the i-th dimension $[-\eta_i\Delta_i, \eta_i\Delta_i]$ ($[-\eta_i, \eta_i]$ once it has been normalized by Δ_i), with η_i the parameter to be optimized, for all $1 \leq i \leq L_1$, and there is no distortion compensation (pure DM case); furthermore, the previous MSE and component-wise constraints will be replaced by the simpler one $\sum_{k=(i-1)\cdot L_2+1}^{i\cdot L_2} \frac{\Delta_k^2\eta_k^2}{3} \leq D_c(i)$, $i = 1, \cdots, L_b$. Therefore,

[15] As it was discussed before, the strategies introduced in this section are suboptimal. In Section 3.1.2.2 we prefered to termed those weights *improved* in order to make clear that they are computed to improve the decoding performance, but they are not necessarily optimal in a strict sense.

replacing $\boldsymbol{\beta}^*$ by its optimal value in the argument of (3.12), the attacker has to minimize

$$\mathrm{SNR}_i = \sum_{k=(i-1)\cdot L_2+1}^{i\cdot L_2} \frac{\left(\frac{1}{2} - \mathrm{E}\{U_k^+\}\right)^2}{\mathrm{Var}\{U_k^+\}} = \sum_{k=(i-1)\cdot L_2+1}^{i\cdot L_2} \frac{3\left(1-\eta_k\right)^2}{\eta_k^2},\ i = 1, \cdots, L_b$$

constrained to $\sum_{k=(i-1)\cdot L_2+1}^{i\cdot L_2} \frac{\Delta_k^2 \eta_k^2}{3} \leq D_c(i),\ i = 1, \cdots, L_b$.

Using the Lagrange multipliers technique, we may proceed to differentiate the unconstrained functional with respect to η_j and equate to zero to get

$$\frac{\left(\eta_j - 1\right)\eta_j^2 - \left(\eta_j - 1\right)^2 \eta_j}{\eta_j^4} + \lambda_i \eta_j \Delta_j^2 = 0,$$
$$\text{for all } \left((i-1)\cdot L_2 + 1\right) \leq j \leq i \cdot L_2, \quad i = 1, \cdots, L_b.$$

So even in this simple case, the following fourth order equation has to be solved for every η_j, $\left((i-1)\cdot L_2 + 1\right) \leq j \leq i \cdot L_2$,

$$\lambda_i \eta_j^4 \Delta_j^2 + \eta_j - 1 = 0 \tag{3.68}$$

Equation (3.68) gives a hint on the complexity of the problem for DC-DM, because in such case the noise due to distortion compensation (self-noise) is combined with the additive noise from the attacker.

3.5.5. Scalar STDM

In this case, the decoder will use the weighting vector $\boldsymbol{\beta}$ in order to yield the projected received signal $\mathbf{z}_p$, i.e., $z_{p_j} = \sum_{i=(j-1)\cdot L_2+1}^{j\cdot L_2} \beta_i s_i y_i$. By doing so, he/she will be able to provide more importance in the decoding process to those dimensions with a lower level of relative noise. A problem with this approach is that the embedder should compute $\mathbf{x}_p$ using the same vector $\boldsymbol{\beta}$ that the decoder, i.e. $x_{p_j} = \sum_{i=(j-1)\cdot L_2+1}^{j\cdot L_2} \beta_i s_i x_i$, in order to obtain the decoding centroids at the same locations that the embedding ones. This implies that the embedder should be able in some way to estimate the attacking power in order to compute $\boldsymbol{\beta}$. This constraint seems to be rather restrictive; nevertheless, one must take into account that similar estimates of the channel at the embedder are also needed for the computation of the distortion compensation parameter α for DC-DM (and equivalently SCS). Fortunately, in Section 3.5.5.3 it will be shown that such estimation is not needed when the attacker is assumed to know the permutation vector, i.e. when he/she knows the coefficients devoted to convey a given symbol.

Following the procedure in [132] it is straightforward to show that the probability of decoding error for the i-th bit, i.e. $P_e(i)$, can be approximated by

$$P_e(i) \approx 2\mathcal{Q}\left(\frac{\Delta_i}{2\sigma_{N_{p_i}}}\right) = 2\mathcal{Q}\left(\frac{\tau_i\left(\sum_{j=(i-1)\cdot L_2+1}^{i\cdot L_2}\gamma_j\beta_j\right)}{2\sqrt{\sum_{j=(i-1)\cdot L_2+1}^{i\cdot L_2}\sigma_{N_j}^2\beta_j^2}}\right), \quad i \in \{1,\cdots,L_b\} \tag{3.69}$$

where $\tau_i \in [\sqrt{3},2]$ is a function that depends on the ratio $\frac{\sigma_{X_{p_i}}}{\Delta_i}$, and consequently also on $\boldsymbol{\beta}$, although in a weaker way (a further discussion about τ_i can be found both in [132] and [20]). Therefore, as $\mathcal{Q}(\cdot)$ is monotonic, the attacker (decoder) has to minimize (maximize) the argument of this function in (3.69).

3.5.5.1. Optimal Decoding Weights for a Known Attack Distribution.

If we assume that τ_i does not depend on $\boldsymbol{\beta}$ (in fact, there is only a weak dependence), it can be proven that the optimal weights become

$$\beta_j^* = \frac{K\gamma_j}{\sigma_{N_j}^2}, \text{ for all } (i-1)\cdot L_2+1 \le j \le i\cdot L_2, \;\; i \in \{1,\cdots,L_b\} \tag{3.70}$$

being K any positive constant.

3.5.5.2. Optimal Attack for Known Decoding Weights.

In this case, we are in the same situation as in Section 3.5.3.2, so all the considerations made there are perfectly valid here. All the attacking power will be concentrated in those coefficients with the largest values of β_k^2.

3.5.5.3. Optimal Attack When the Decoder Follows the Optimal Strategy.

If we follow a strategy similar to the one described in Section 3.5.3.3, assuming that the attacker does not know the actual partition, we obtain an expression like (3.62), where now $p_j = \sigma_{N_j}$, $q_j = \gamma_j$, $t_j = 0$. In this case it is not so clear that $p_j \gg q_j$. In fact, for WNR> 0, $q_j > p_j$. Therefore, the same simplification as in (3.63) cannot be done and the problem requires to be solved by numerical optimization. In order to be able to compare the results in previous sections with some theoretical results on scalar STDM, we have also analyzed the case when the attacker knows the partition; this could be seen as a pessimistic scenario for the system. In that case, when the decoder follows the optimum strategy, the

probability of wrongly decoding a given bit (whose index has been removed for the sake of simplicity) can be approximated as

$$P_e \approx 2\mathcal{Q}\left(\frac{\tau(\sum_{j=1}^{L_2}\gamma_j^2/\sigma_{N_j}^2)}{2\sqrt{\sum_{j=1}^{L_2}\gamma_j^2/\sigma_{N_j}^2}}\right) = 2\mathcal{Q}\left(\frac{\tau}{2}\sqrt{\sum_{j=1}^{L_2}\frac{\gamma_j^2}{\sigma_{N_j}^2}}\right), \tag{3.71}$$

so the attacker will be willing to minimize $\sum_{j=1}^{L_2}\gamma_j^2/\sigma_{N_j}^2$, constrained to $\sum_{j=1}^{L_2}\sigma_{N_j}^2 \leq D_c$. Using Lagrange multipliers, and differentiating with respect to $\sigma_{N_j}^2$, one obtains

$$-\frac{\gamma_j^2}{\sigma_{N_j}^4} + \lambda = 0, \tag{3.72}$$

where λ is the corresponding Lagrange multiplier. This yields

$$(\sigma_{N_j}^2)^* = \xi\gamma_j, \text{ for all } j \in \{1, \cdots, L_2\}, \tag{3.73}$$

with $\xi = L_2 \cdot D_c / \left(\sum_{j=1}^{L_2}\gamma_j\right)$.

Finally, we would like to remark that this attacking power allocation leads to $\beta_i = K_j$, for all $i \in \{(j-1)\cdot L_2 + 1, \cdots, j \cdot L_2\}$ and $j \in \{1, \cdots, L_b\}$. This clearly obviates the need for estimating the attack at the embedder, as it was discussed in the introductory part of Section 3.5.5.

3.5.6. Experimental Results

We show next the results of applying the strategies derived along the previous sections to real data. In the figures that follow, symbols refer to empirical (Monte Carlo) simulations, while lines show theoretical results. Empirical data come from the gray-scale *Lena* image (256×256), for which the spatial perceptual mask $\boldsymbol{\gamma}$ has been computed using the edge detection method described in [107], except for the DC-DM scheme where, for illustrative purposes, we have chosen to work in the DCT domain, using the perceptual mask proposed by Watson in [160].

First, in Figure 3.28 the P_e's resulting when different strategies are considered for Add-SS (Section 3.5.3) are shown. Watermarking has been performed in the spatial domain with Wiener filtering prior to decoding and 50 pixels per bit ($L_2 = 50$) have been used. Three cases are analyzed: first, the noise variance $\sigma_{N_j}^2$ at each sample is made proportional to γ_j^2 and $\boldsymbol{\beta} = K\boldsymbol{\gamma}$, with K any positive constant; second, the attack is the same as in the previous case but the optimal decoding weights $\boldsymbol{\beta}^*$ are employed; finally, the plot labeled as "worst attack" refers to the case where the attacker follows his/her optimal strategy knowing that the

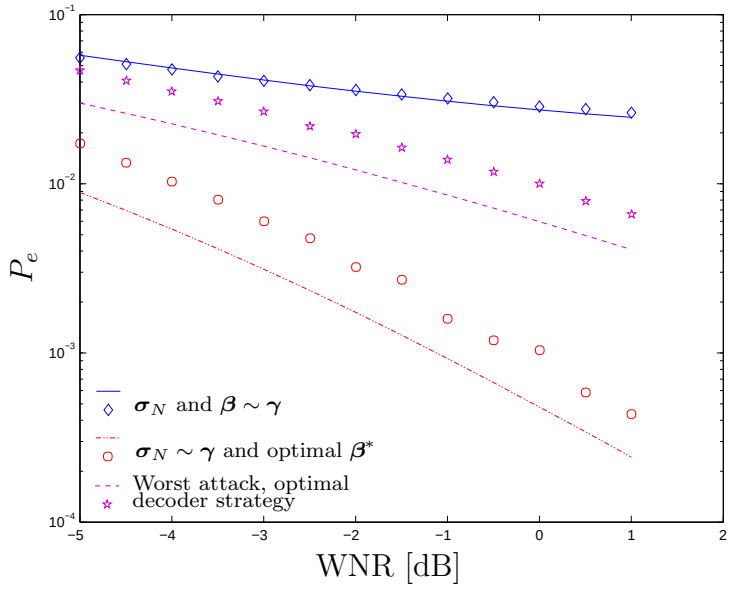


Figure 3.28: BER versus WNR for Add-SS ($L_2 = 50$) showing three different attacking/decoding strategies.

decoder also uses the optimal decoding weights. In all cases, the theoretical results lie close to the empirical ones, although for those where the optimal $\boldsymbol{\beta}^*$ is used the difference is larger.

The cases depicted in Figure 3.29 correspond to the binary DC-DM method where, as mentioned, watermarking is done in the DCT domain. The distortion compensating parameter α is set to 0.7. In order to establish a meaningful case for the experiments, we have selected uniform noise proportional to the quantization step that results when a JPEG quality factor of 80 is selected. Two scenarios are depicted in Figure 3.29: in the first case, each sample, say the j-th, is scaled by Δ_j at the decoder but no further weighting (i.e., $\beta_j = 1$) is considered; in the second plot, the optimal $\boldsymbol{\beta}^*$ that follows from applying the results from Section 3.1.2.2 is used.[16] For both Figures we have set $L_2 = 10$. The theoretical approximations are based on the results introduced in [131], so they were computed as

$$P_{approx.} = \Pr\{(\mathbf{T}^+)^T \mathbf{B}(\mathbf{T}^+) > (\mathbf{T}^+ - \mathbf{1})^T \mathbf{B}(\mathbf{T}^+ - \mathbf{1})\}, \tag{3.74}$$

where $T_i^+ = |T_i|$, for $i \in \{1, \cdots, L_2\}$, with $\mathbf{T}$ defined in (2.28).

The fact that in the second case the empirical results lie above the theoretical ones may be surprising at first sight, since the latter was said in [131] (where

[16] For the sake of computational simplicity, the optimal weights $\boldsymbol{\beta}^*$ used in Figure 3.29 were computed taking into account the statistics of the random variables $|T_i|$, $1 \leq i \leq L_2$, not their modulo-lattice reduced versions U_i^+, as it should have been done following the results of Section 3.1.2.2. In any case, Figure 3.29 shows the gain of using the weighting parameters.

$\mathbf{B} = \mathbf{I}_{L_2 \times L_2}$) to be an upper bound to P_e. The explanation to this phenomenon is that in such case some β_j^* take negative values, affecting the validity as an upper bound of the CLT approximation based on the non-modulo reduced variables. Note that as we have less noise (i.e., the WNR increases), it becomes more unlikely to have negative values of β^* (since $\mathrm{E}\{U_i^+\}$, $1 \leq i \leq L_2$, decreases), so the theoretical curve and the empirical results get much closer. In any case, be aware that the exact theoretical values could be also computed using the results introduced in Section 3.1 (their accuracy can be checked, for example, in Figures 3.3 and 3.9).

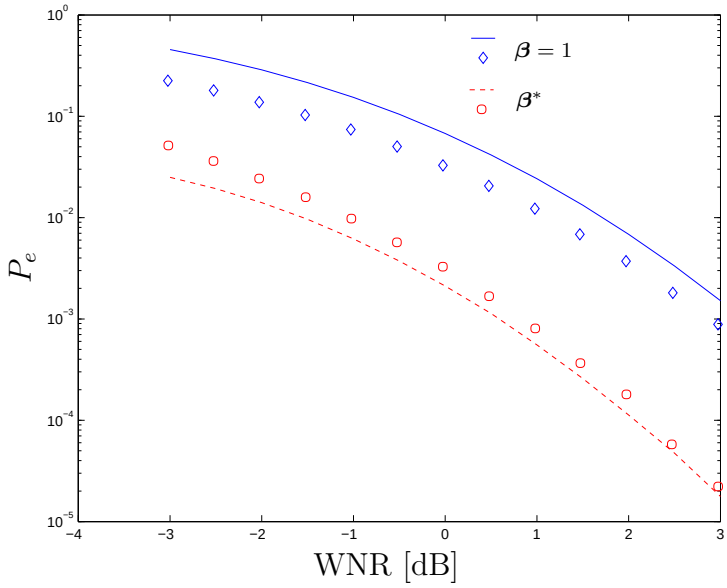


Figure 3.29: BER versus WNR for DC-DM with uniform scalar quantizers and repetition coding ($L_2 = 10$, $\alpha = 0.7$), for uniform noise proportional to JPEG quantization step ($QF = 80$) when no weights are used, and for the optimal weighting.

Finally figure 3.30 shows a similar comparison for the case considered in Section 3.5.5.2. The decoding weights are set so that $\beta = \gamma$, and the optimal attack for this case is compared to an attack consisting in using noise variances $\sigma_{N_k}^2$ proportional to γ_k.

3.5.7. Conclusions

As a conclusion of this section, one aspect that clearly requires further study is that of distortion constraints and their relationship with optimal strategies. For instance, as it can be checked in Sections 3.5.3.2 and 3.5.5.2, the optimal attack will likely end up in a visible attacked image. Whether this image keeps

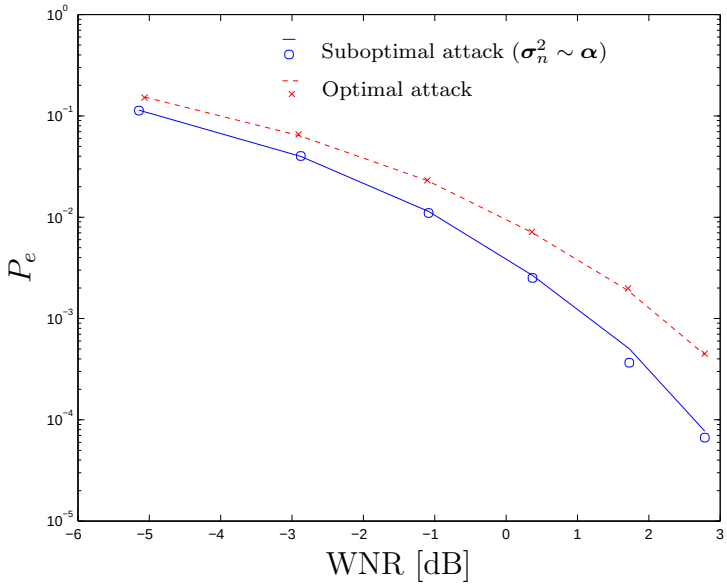


Figure 3.30: BER versus WNR corresponding to the suboptimal and optimal attacks for scalar STDM when the attacker knows the decoder weights ($L_2 = 10$).

some of its original value is a moot question that largely depends on the final application scenario.

Related to this, we can think of the problem where the embedder has an active role (as we have already done in scalar STDM), and does not just generate the watermarked image independently of the possible attacks. In any way, the distortion introduced by the embedder has to be extremely small; in that regard, we can assume that the attacker has always more freedom to make it difficult the decoding process.

3.6. Worst Additve Attack for scalar DC-DM

As it was explained in Section 2.5, the lattice decoding approach, where the decoder operates over variables that are reduced modulo-lattice, is typically used to decode DC-DM. Its use is based on a complexity reduction, given that a really reduced number of centroids (only one per possible symbol) must be taken into account. Furthermore, for very large values of DWR and the most common distributions of the host, the modulo-lattice reduced version of the host will be asymptotically uniform over the Voronoi region of the lattice; therefore, the original host signal distribution has not to be considered in order to perform the decoding. Nevertheless, the reader should be aware that, in general, the modulo-lattice reduction operation could be information lossy.

Despite of this possible information loss, lattice decoding is extensively used in practical implementations of DC-DM, so the analysis of its maximum reliable transmission rate is fully justified. The objective of this section will be the study of this maximum reliable rate for the case of power-constrained additive noise channels. This problem can be interpreted as a game between embedder, who must choose the watermarking code in order to maximize the reliable rate, and the attacker, who tries to find the power-constrained additive noise distribution which minimizes the rate. In this way, the embedder will choose the lattice Λ and the distortion compensation parameter α maximizing the reliable rate. On the other hand, the attacker will be assumed to know Λ and α (this can be regarded as a pessimistic approach), and will try to find the power-constrained additive noise distribution $f_{\mathbf{N}}(\mathbf{n})$ minimizing the reliable rate (a.k.a. worst power-constrained additive noise).

3.6.1. Computation of the worst additive noise in the literature

The problem of finding the worst case additive noise for a given constellation has been widely studied in the literature. Below, we will summarize some of these results:

- A pioneering paper in this field is [141], where the worst power-constrained additive noise is studied for the scenario of binary input channels ($\{-1, +1\}$) with continuous-valued outputs. In this paper the receiver is assumed to know the noise statistics, so the ML detector is considered. The p.d.f. maximizing the probability of error is shown to be a train of Dirac's delta functions located at the integers $\{-K, \cdots, K\}$, with K depending on the signal-to-noise ratio, and the probability of each delta on whether the related integer is even or odd.

  When the target function to be minimized is the capacity, the optimal noise is shown to be located on a subset of a lattice. In fact, at least for the cases considered by the authors, the noise is located at $2\mathbb{Z}$, with a probability distribution approaching a Gaussian shape for low SNRs.

  These results can be generalized for the case of finite input constellations that take values on a finite subset of a lattice. The worst power-constrained additive noise is then shown to be a mixture of two distributions on lattices that are shifted versions of the input lattice. In any case, be aware that this framework is far from being the one we are interested in, due to the lack of the modulo reduction.

- The previous work was extended in [112], where the noise is constrained not only in power but it is also required to be similar to a nominal distribution;

this similarity is measured with the Kullback-Leibler distance between the pdf of the studied noise and that of the nominal distribution. The worst case additive noises when just the similarity constraint is considered, and when the power constraint is also taken into account are studied; the target function is in both cases the probability of error of the zero-threshold detector, paying special attention to the case of Gaussian nominals. When the similarity constraint is relaxed (allowing for a large difference), the worst noise approachs a three-point distribution (recalling the result in [141]). Finally, asymptotical results are also provided for the case of small differences allowed between the noise pdf and the nominal one.

- A framework similar to [112] was later analyzed by the same authors in [113], where the ML detector is considered. Again, special attention is paid to Gaussian nominals, and asymptotic behaviors are analyzed.

The computation of DC-DM power-constrained case worst additive attack (WCAA) obviously differs from the former scenarios in several points:

- The input signal, i.e. the watermarked signal, is not assumed to be distributed on a finite subset.
- The modulo-lattice reduction must be considered.
- The uniform noise due to the distortion compensation (the *self-noise*) has to be taken into account.

These points make the optimization of DC-DM worst additive attack quite particular. In fact, the problem was not solved even for the simplest case of uniform scalar quantizers, until [130].

3.6.2. Theoretical Analysis

Given that the maximum reliable rate is given by the mutual information between the observed signal and the message, this problem can be formulated as

$$\max_{\Lambda,\alpha,\mathbf{v}(\mathbf{B})} \min_{f_{\mathbf{N}(\mathbf{n})}} I(\mathbf{Z}_{\text{mod}}; \mathbf{B}), \tag{3.75}$$

which shows the attacker knowledge about the watermarking code on the order the game is played. The minimization in (3.75) has to be carried out over the set of possible additive channels verifying $\mathrm{E}\{||\mathbf{Z}-\mathbf{Y}||^2\} = \mathrm{E}\{||\mathbf{N}||^2\} \leq L_2 \cdot D_c$; therefore, (3.75) is obviously equivalent to

$$\max_{\Lambda,\alpha,\mathbf{v}(\mathbf{B})} \min_{f_{\mathbf{N}(\mathbf{n})}:\mathrm{E}\{||\mathbf{N}||^2\}\leq L_2\cdot D_c} I([\mathbf{v}(\mathbf{B}) - (1-\alpha)\mathbf{E} + \mathbf{N}] \mod \Lambda; \mathbf{B}), \tag{3.76}$$

where $f_{\mathbf{N}}(\mathbf{n})$ is a pdf, so $\int f_{\mathbf{N}}(\mathbf{n})d\mathbf{n} = 1$ and $f_{\mathbf{N}}(\mathbf{n}) \geq 0$, for all $\mathbf{n} \in \mathbb{R}^{L_2}$. As it was previously explained, in most of practical scenarios $\mathbf{X} \mod \Lambda \sim U(\mathcal{V}(\Lambda))$, so $\mathbf{E} \sim U(\mathcal{V}(\Lambda))$. Moreover, it is straightforward to see that the capacity will be maximized for a uniform distribution of $\mathbf{v}(\mathbf{B})$ over $\mathcal{V}(\Lambda)$ (see [70] for a further discussion on this topic); since an infinite number of vectors $\mathbf{v}(\mathbf{B})$ must be available to obtain such distribution, also an infinite number of messages is required, i.e. $P \to \infty$.

Furthermore, as it was introduced in [130], the worst additive attack will no longer be the Gaussian one, as it could be thought, since there will exist other distributions whose modulo reduced version will have the same pdf than the modulo reduced version of the Gaussian one, but with lower variance. Considering this effect, the *worst case* distribution has to be computed in order to know the maximum reliable rate. In order to do so, we will assume that all the messages have the same probability, and all the codewords are equidistant; be aware that these hypotheses are not so restricting as they might seem, as they are verified in most practical scenarios. Taking into account the structure of the resulting constellation and denoting $\mathbf{U}' \triangleq [-(1-\alpha)\mathbf{E}+\mathbf{N}] \mod \Lambda$, the mutual information in (3.76) was shown to be [130]

$$I([\mathbf{v}(\mathbf{B}) - (1-\alpha)\mathbf{E} + \mathbf{N}] \mod \Lambda; \mathbf{B}) = D\left(f_{\mathbf{U}'}(\mathbf{x}) || f_{\mathbf{U}''}(\mathbf{x})\right)$$

where $\mathbf{U}'' \triangleq [\mathbf{v}(\mathbf{B}) + \mathbf{U}' - \mathbf{v}(\mathbf{0})] \mod \Lambda$, so $f_{\mathbf{U}''}(\mathbf{x}) = \frac{1}{P^{L_b}} \sum_{\mathbf{b} \in \{0,\cdots,P-1\}^{L_b}} f_{\mathbf{U}'}([\mathbf{v}(\mathbf{b}) + \mathbf{x} - \mathbf{v}(\mathbf{0})] \mod \Lambda)$, for $\mathbf{x} \in \mathcal{V}(\Lambda)$. Therefore, the mutual information can be seen as the Kullback-Leibler distance between the pdf of the total noise ($f_{\mathbf{U}'}$), and the average pdf obtained when this noise is shifted by the dither vector related to each message.

In the next section, the maximum reliable rate in the above described framework will be studied for the case of DC-DM with uniform scalar quantizers, or equivalently SCS, (i.e., $\Lambda = K \cdot \mathbb{Z}$, $L_2 = 1$), both for the binary message (i.e., $P = 2$) and the continuous (infinite) approximation (i.e., $P \to \infty$); therefore, in each of these two cases the only parameter the encoder can play with will be the distortion compensation parameter α.

3.6.3. Numerical Optimization Results

In this section the results of the numerical optimization introduced so far are presented; in all of them, we will assume that the host signal verifies the flat-host assumption, i.e. $X \mod \Lambda \sim U(\mathcal{V}(\Lambda))$ and all the elements of Λ have the same probability of being chosen as the quantized value of the host. Furthermore, given that we will just study the scalar case, we will assume for representation purposes, and without loss of generality, that $\Lambda = 2\mathbb{Z}$.

First, we will consider the binary case. In Figure 3.31 we can see the maximum achievable rate, for this scenario. Obviously, the fact of just having two codewords is constraining the rate to be lower or equal to 1 bit per channel use. Given that we are interested in computing the maximum of the achievable rates, a maximization has been performed over the only parameter the embedder can play with: the distortion compensation parameter α. As it was expected, the achievable rate for a fixed WNR is always lower for the WCAA (so the worst case attack distribution is clearly not the Gaussian one), although for some ranges of WNRs this distance is evidently reduced; this is the case for very low WNRs, and around 2.5 dB. It is especially remarkable the large difference that can be observed when the WNR is around 10 dB. But perhaps more striking is to analyze the optimal value of α obtained for the WCAA. In Figure 3.32 we can compare the value of α obtained in different scenarios; we can see the optimal value of α numerically computed for the WCAA (so the solution to the maximim game described above), as well as the optimal α, also numerically computed, in presence of Gaussian noise. Finally, both results can be compared with the values of α proposed by Costa [50], Eggers [65], and Pérez-González in [130]. It is remarkable that the optimal value of α for the WCAA does not follow at all the value obtained for any of the other strategies; while α clearly goes to 1 when the WNR is increased in all the other cases, here it is increased until it achieves a value of 0.63, where it stays (except for some fluctuations around that value due to numerical errors). This means that the increase in the power of the attacking noise that we would have by increasing α is not compensated by the corresponding reduction in the self-noise, i.e. the power that we allowed the attacker to use is more harmful than that due to the self-noise. Finally, in Figure 3.33, we can see the pdfs of the worst case additive attack, when α is set to maximize the achievable rate, for different WNRs.

On the other hand, the maximum achievable rate under both Gaussian noise and the WCAA, when the input is uniformly distributed over the Voronoi region of the quantizing lattice (or equivalently $P \to \infty$) can be seen in Figure 3.34. At first sight, we can see one of the main differences with the previous binary case; now, the achievable rate can be made as large as desired by increasing the WNR.[17] Furthermore, the achievable rate for the WCAA is just slightly lower than that obtained with Gaussian noise. In Figure 3.34 we can also see the capacity of Costa's scheme, given by $\frac{1}{2}\log_2(1 + \text{WNR})$. The asymptotic difference between the capacity and the achievable rate for both the WCAA and the Gaussian noise is given by 1.53 dB, i.e. the *shaping gain* of a hypersphere of infinite dimensions.[18] Consequently, if the system designer wanted to close this gap (or at least reduce

[17] As a matter of fact, the achievable rate would be upper-bounded by $\log_2(P)$, but $P \to \infty$.

[18] The shaping gain is defined as the ratio between the normalized second moment of a hypercube and that of the analyzed lattice, i.e. $g_s(\Lambda) = -10\log_{10}(12G(\Lambda))$, with $G(\Lambda)$ the normalized second moment of the considered lattice. The region with the smallest normalized second moment for any dimensionality L_1 is the L_1-sphere, and that value is lower-bounded by $\frac{1}{2\pi e}$, achieving it just when $L_1 \to \infty$. Therefore, the maximum of the shaping gain is given by $10\log_{10}\left(\frac{2\pi e}{12}\right) \approx 1.53$ dB.

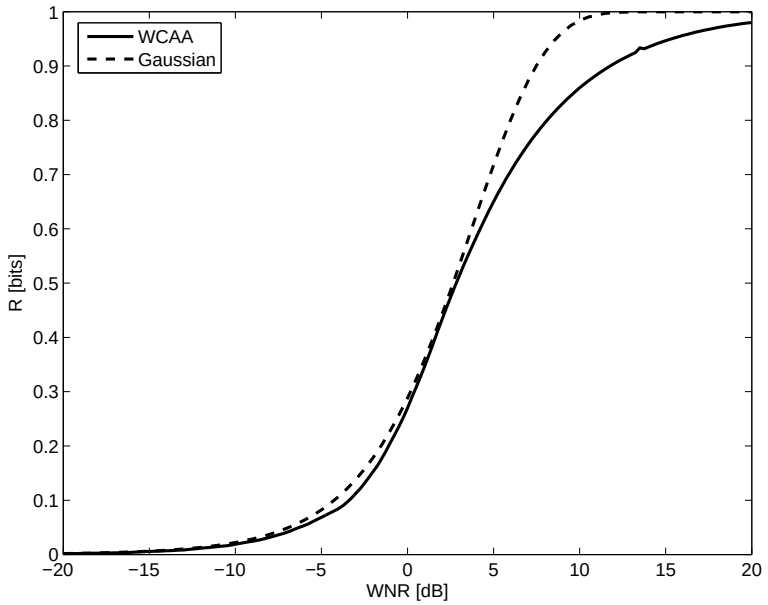


Figure 3.31: Maximum achievable rate, for the case of binary message ($P = 2$), and Gaussian and worst case additive attack noise. The optimization was performed sampling the Voronoi region, i.e. $(-1, 1]$ at 256 points, and α at 100.

it), a form of source coding should be performed. This will be further explained in Chapter 5, where a method combining both channel coding and source coding is discussed, showing the advantage of following such an approach. Concerning the optimization of the distortion compensation parameter, in Figure 3.35 we can compare the different strategies. On the one hand, the optimized α for the WCAA, and the value of α maximizing the achievable rate for the Gaussian case; on other hand the previously explained values proposed by Costa, Eggers, and Pérez-González are plotted. Note that despite the step-like aspect and the large variability of the value of α obtained for the WCAA (due to the finite precision of the optimization algorithm), there is a high resemblance with the other plots. Finally, in Figure 3.36, we can see the pdfs of the worst case additive attack, when α is chosen to maximize the achievable rate, for different WNRs. Note that as long as the WNR is increased, the WCAA seems to approach a Gaussian distribution.

3.6.4. Subsequent works on the worst additive attack for DC-DM

After the worst power-constrained additive attack for DC-DM was for the first time introduced in [130], several other works have dealt with this subject. Next, we will summarize some of them:

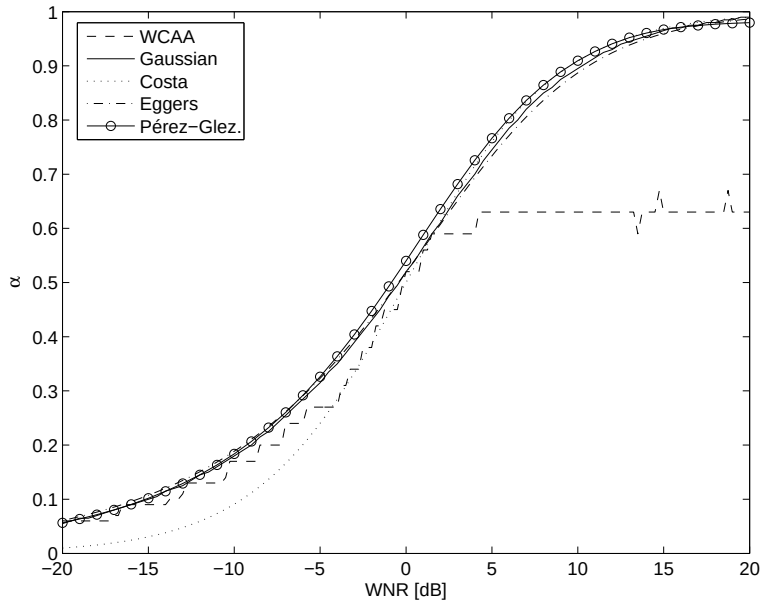


Figure 3.32: Optimal distortion compensation parameter α, for the case of binary message ($P = 2$), and Gaussian and worst case additive attack noise. These results, which were obtained by numerical optimization, are compared with the optimal value of α by Costa [50], $\alpha = \sigma_W^2/(\sigma_W^2 + \sigma_N^2)$, the approximation given by Eggers [65], $\alpha = \sqrt{\sigma_W^2/(\sigma_W^2 + 2.71\sigma_N^2)}$, and the value proposed by Pérez-González in [130]. The optimization was performed sampling the Voronoi region, i.e. $(-1, 1]$ at 256 points, and α at 100.

- In [158] the authors addressed the problem of finding the noise distribution maximizing the probability of error for the minimum distance decoder; the noise is constrained to be additive, power-constrained and similar to a *target* distribution, resembling the approach followed in [112]. Considering the obtained results, they proposed to follow a 3 Dirac's deltas distribution, parameterized by the distance from the two extreme deltas to the origin and their probability. An interesting by-product of this analysis is the proposal of a near-optimal distortion compensation parameter $\alpha = 2/3$, for which the probability of error can be upperbounded for a given WNR.

  A similar approach was followed in [157] and [155], where the mutual information problem is also studied; the mutual information values obtained for Gaussian and uniform attacks are compared with those obtained by the worst case additive attack. The same subject is retaken in [156], where the authors show that the 3 delta attack asymptotically produces the same probability of error than the real worst additive attack.

- The problem of finding the worst additive attack was also studied for the non-blind Add-SS watermarking in [82]. In this work, both watermark and attack are constrained in amplitude due to the Just Noticeable Difference

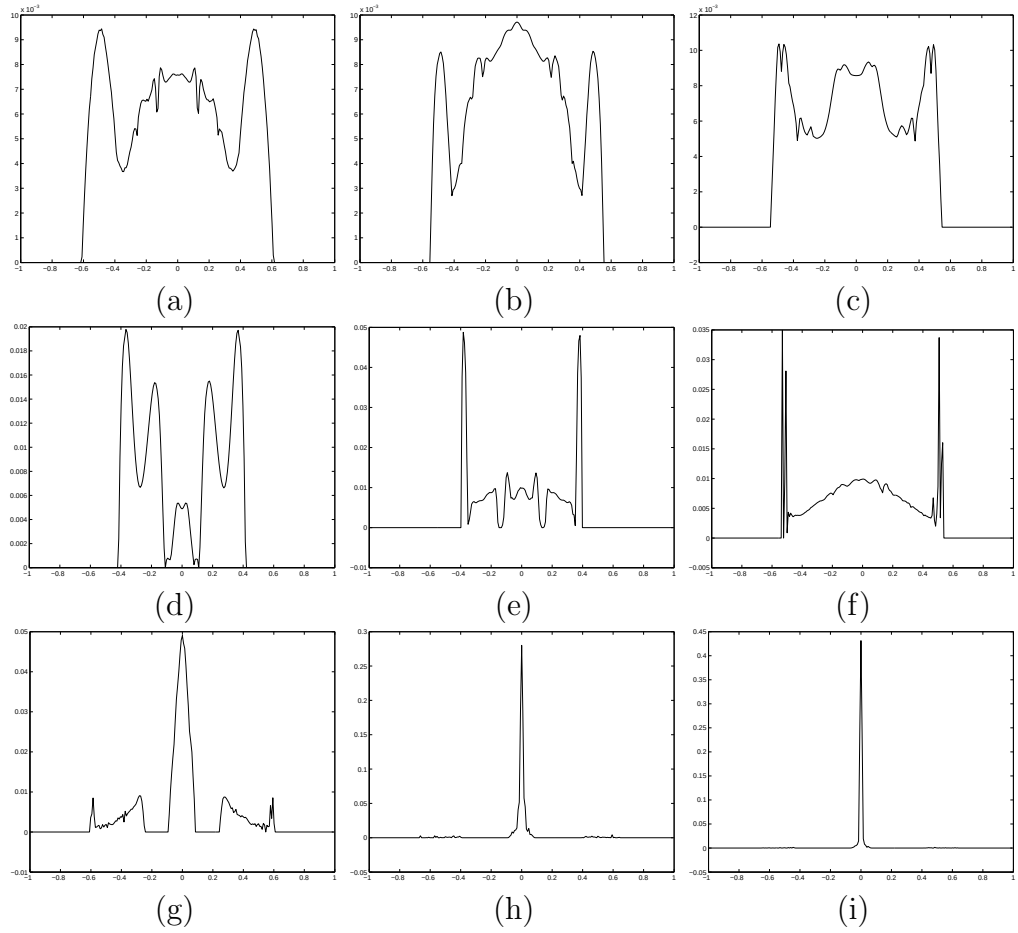


Figure 3.33: Worst case additive attack pdfs for a binary message and different WNRs, with α optimized for every WNR: (a) -20 dB, (b) -15 dB, (c) -10 dB, (d) -5 dB, (e) -2 dB, (f) 0 dB, (g) 3 dB, (h) 10 dB and (i) 15 dB. The optimization was performed sampling the Voronoi region, i.e. $(-1, 1]$ at 256 points, and α at 100.

(JND) level; binary and quaternary input alphabets are analyzed. The payoff functions were chosen to be the probability of decoding error and the capacity of the system. For both of them the optimal distribution of the attack and the watermark are computed.

The same authors considered in [81] a minimax problem, where the embedder computes the value of α minimizing the probability of error of the ML detector, whereas the attacker computes the distribution of the memoryless attack which maximizes such probability of error. The obtained results are based on the Bhattacharyya bound. In order to gain robustness, the authors proposed the use of randomly rotated lattices; in that case, the worst attack pdf is shown to be a radial one, so it has memory.

A similar strategy (i.e. ML decoder, minimax problem, Bhattacharrya

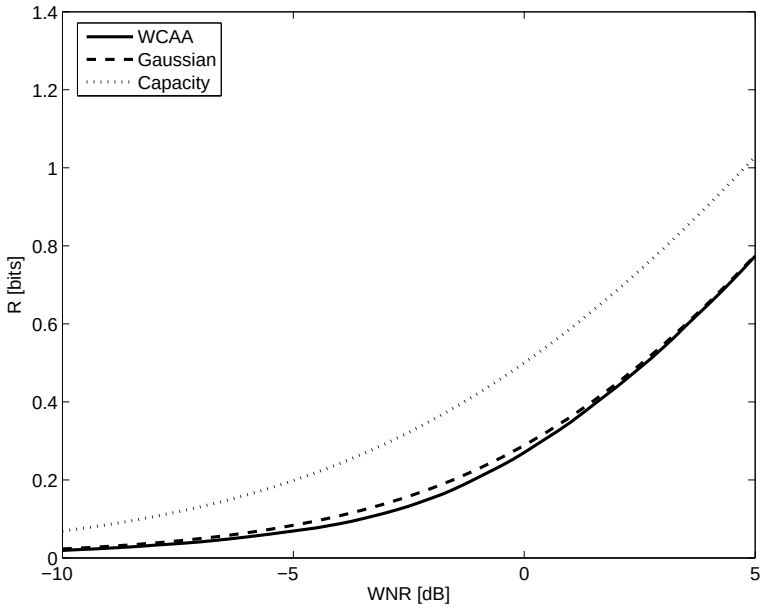


Figure 3.34: Maximum achievable rate, for the case of uniform input ($P \to \infty$), and Gaussian and worst case additive attack noise. The optimization was performed sampling the Voronoi region, i.e. $(-1, 1]$ at 256 points, and α at 100.

bound and randomized rotation) is also followed in [117], where DC-DM is combined with repetition coding in order to improve the decoding performance. Worst additive noise distributions are provided for the cases of a hexagonal lattice and a randomly rotated cubic lattice (radial noise). Another contribution of this paper is the consideration of a more general target function: this function is a generic upper bound on the probability of error, in such a way that both Bhattacharrya and Chernoff bounds can be considered as particular cases of it.

- Finally, in [150] the mutual information game for SCS (or equivalently DC-DM with uniform scalar quantizers) is considered. In the proposed framework the embedder is allowed to play with the distribution of the input and the distortion compensation parameter α, whereas the attacker, who is assumed to know the embedder strategy, is restricted to additive attacks. Since inputs uniformly distributed over the Voronoi region of the uniform scalar quantizer maximize the mutual information, this will be the chosen distribution. The resulting problem is solved by using the Blahut-Arimoto algorithm. Obviously, the obtained results are equivalent to those in [130] when $P \to \infty$, since in that case the resulting input pdf is also (asymptotically) uniform over the Voronoi region of the uniform scalar quantizers, and the optimization is just performed over $f_{\mathbf{N}}(\cdot)$ and α.

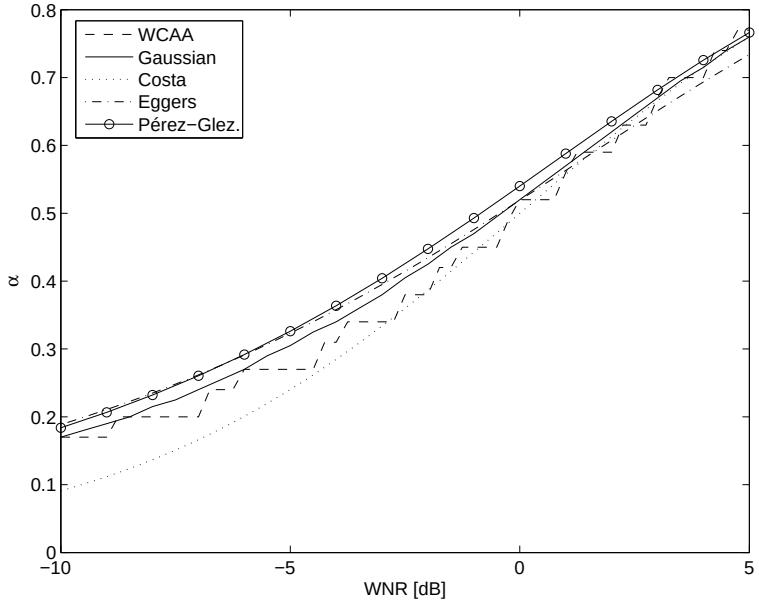


Figure 3.35: Optimal distortion compensation parameter α, for the case of uniform input ($P \to \infty$), and Gaussian and worst case additive attack noise. These results, which were obtained by numerical optimization, are compared with the optimal value of α by Costa [50], $\alpha = \sigma_W^2/(\sigma_W^2 + \sigma_N^2)$, the approximation given by Eggers [65], $\alpha = \sqrt{\sigma_W^2/(\sigma_W^2 + 2.71\sigma_N^2)}$, and the value proposed by Pérez-González in [130]. The optimization was performed sampling the Voronoi region, i.e. $(-1, 1]$ at 256 points, and α at 100.

3.7. Conclusions

In this chapter the performance of the most relevant state-of-the-art methods (paying special attention to DC-DM) has been analyzed under a wide range of attacks: from the classically considered additive noise, or the also typical (although rarely analyzed) coarse quantization attack, to the cropping attack, studied here to show an important weakness of the extensively used STDM-like methods. Another interesting contribution is the BNSA, a generalized version of the sensitivity attack, which was shown to be suitable for attacking several of the most popular watermarking schemes. Nevertheless, one could also wonder what is the optimal strategy for the attacker and the decoder/detector, when the former does not have access to an instance of the latter. This is the question that we tried to solve in our game-theoretic approach, constraining the decoder to have a simplified structure. Finally, we have also dealt with the problem of computing the noise distribution that minimizes the achievable rate for scalar DC-DM and a given attacking distortion; the obtained results provide an enlightening comparison with those obtained for the Gaussian noise, including a comparison of the different values of the optimal distortion compensation parameter in each scenario.

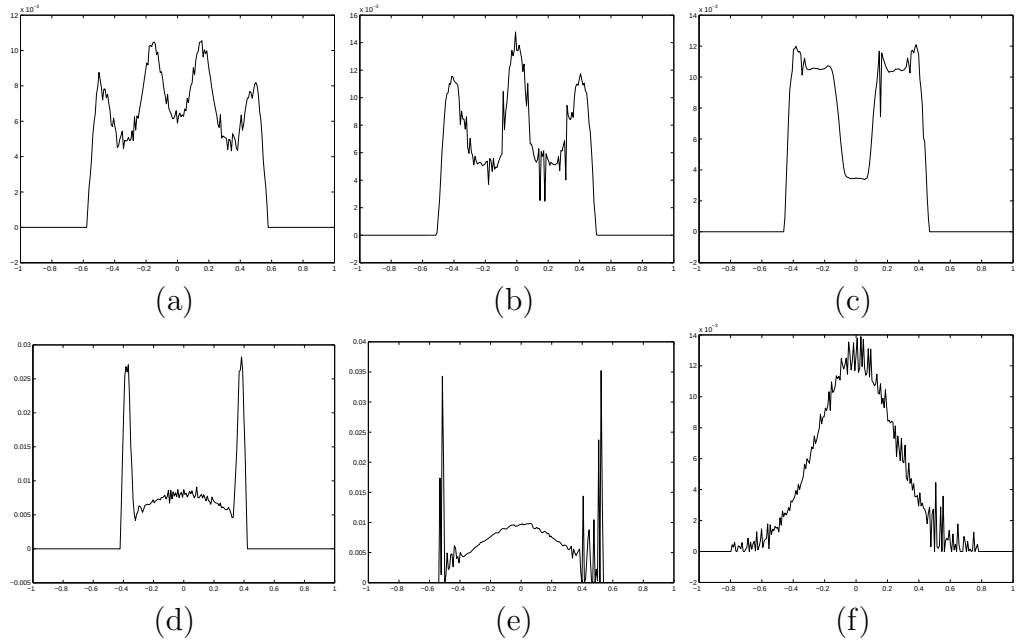


Figure 3.36: Worst case additive attack pdfs for uniform input and different WNRs, with α optimized for every WNR: (a) -10 dB, (b) -8 dB, (c) -5 dB, (d) -2 dB, (e) 0 dB, (f) 5 dB. The optimization was performed sampling the Voronoi region, i.e. $(-1, 1]$ at 256 points, and α at 100.

Chapter 4

Security

In this chapter we try to distinguish the concepts of *security* and *robustness* in watermarking. In order to do so, first we make a brief historical overview of the separation of these concepts in the literature. Taking into account the key ideas of some of those previous works, an information-theoretic approach is followed to study data-hiding and watermarking security. In this approach, the security is measured by the mutual information that quantifies the information about the secret key that leaks from the observation of watermarked documents. This framework is applied to the analysis of Add-SS and Costa's data-hiding schemes in different scenarios. For Add-SS some interesting links are shown between a measure used in previous works in the literature, which is based on the Fisher Information Matrix, and our proposed measure. Furthermore, the results for both Add-SS and Costa's scheme are compared with those obtained for scalar DC-DM with uniform scalar quantizers (SCS).

4.1. Historical Overview

During the first years of digital data-hiding research focused mainly on the analysis of the different proposed methods and their behavior against attacks. Those attacks were usually divided in *intentional* and *non-intentional*, or some similar classification. For example, in [54] the authors distinguished between *signal transformations* and *itentional attacks*, where the latter including the so-called *statistical averaging attack* (related to the *collusion attack* [100]) and the *sensitivity attack* [53], which was extensively explained in Section 3.4.

But it is in [116], which was inspired by [27] and [167], where for the first time a theoretical framework for analysing watermarking security is proposed. The author measures the secrecy of the system as the mutual information between the embedded message $\mathbf{B}$ and the watermarked signal $\mathbf{Y}$, i.e. $I(\mathbf{B};\mathbf{Y})$, and the

robustness as the mutual information between the embedded message $\mathbf{B}$ and the received signal $\mathbf{Z}$ given the decoding key $\mathbf{\Theta}_d$, i.e. $I(\mathbf{B};\mathbf{Z}|\mathbf{\Theta}_d)$, clearly separating *secrecy* from *robustness*. In this way, the system is said to achieve *perfect secrecy* when the corresponding mutual information is 0, resembling the concept introduced by Shannon in his paper about cryptanalysis [140]. Nevertheless, probably the most criticizable point of [116] is that the author does not consider the information leakage about the secret key that could appear when several watermarked contents are available.

As it was discussed in Section 3.4.2, *asymmetric schemes* try to reduce the risk of attacks aimed at estimating the detection/decoding key; in *symmetric schemes*, this estimate could be also used for easily forging falsely watermarked signals, whereas this is not so straightforward for *asymmetric schemes*. The main conclusion of [78], where four asymmetric methods were unified as quadratic forms, is that one can gain security by increasing the number of parameters to be estimated (which in this case is achieved by increasing the order of the decoding/detection function), at the cost of reducing robustness. The proposed approach has some similarities with that followed in [16].

Nevertheless, although security was starting to be a *hot topic* in watermarking, there was still a lack of a proper definition. In [96], Kalker took significant steps towards such definition. First of all, he defined *robust watermarking* as *"a mechanism to create a communication channel that is multiplexed into original content"*, and whose capacity *"degrades as a smooth function of the degradation of the marked content"*. On the other hand, *"security refers to the inability by unauthorized users to have access to the raw watermarking channel"*. That access includes *"removing, detecting and estimating, writing and modifying the raw watermarking bits"*. Nevertheless, this definition does not reflect some crucial aspects as, for example, the intentionality of the attacks. In that sense intentionality and robustness/security can be regarded as independent concepts, being feasible the four possible combinations of them. Therefore, following Kalker's definitions, both intentional and non-intentional attacks may result in a threat to security.

In [76] the differences between security and robustness are emphasized: security is not just related to the removal of the watermark, but also to the embedding and detection by unauthorized parties. Furthermore, some aspects of the definitions proposed by the authors somehow collide with, or at least evolved from, those introduced by Kalker in [96]. For example in [76], the authors claim that *"security deals only with intentional attacks, whereas robustness measures the impact of classical content transformations on the detectability of the watermark"* being *"inmaterial"* for robustness *"whether such transformations are intentional or not"*. Moreover, in the proposed framework robustness attacks are characterized by the lack of knowledge of the watermarking scheme by the attacker (those are usually termed *blind* attacks), whereas in attacks to security the attacker does

have knowledge of the system. This clearly resembles Kerckhoffs' principle [99] in cryptography, which in fact was translated to watermarking by the authors; this principle states that the security of a system must rely just on a secret key that is not known by the attacker, considering that the rest of the parameters of the system, including embedding and detection/decoding functions, are perfectly known by the attacker. Taking into account Kerckhoffs' work the security level is defined as *"the effort (complexity, time, money, ...) the attacker requires to disclose the secret key"*. The authors, inspired by another outstanding work in cryptography by Diffie and Hellman [60], introduce a classification of attacks:

- Only watermarked content attack: just some watermarked contents are available to the attacker.
- Watermarked content pair attack: pairs of original contents and their corresponding watermarked versions are available.
- Chosen original content attack: a watermark embedder is available.
- Chosen watermarked content attack: a watermark detector is available (*oracle attack*).

Finally, the authors adapt the approach proposed by Shannon in cryptography to measure the uncertainty about the secret key when some encrypted messages are available to the decoder; in his work [140], Shannon measure this ignorance as the entropy of the key given the encrypted messages. In [76], the uncertainty about the secret key when some watermarked contents are available is similarly measured as the entropy of the key given the watermarked contents. Unfortunately, some problems appear when the continuous case is studied.

Another interesting work about watermarking security is [18]. Probably the main innovation of this work is that watermarking is considered as a game with some rules, which determine the information publicly available. If the attacker uses only this information, the attack is said to be *fair*; if he/she tries to learn more information about the system, the attack is said to be *unfair*. The information publicly available can range from *no knowledge*, that clearly collides with Kerckhoffs' principle, *knowledge of embedding and detection algoritms*, *knowledge of the detection key* (for asymmetric schemes), to *knowledge of both embedding and detection keys, and the algorithms*. Similarly to [76], the mutual information is used to measure the knowledge gained by the attacker. Finally, a definition of *security level*, also similar to that in [76], is introduced: *"is the amount of observations, the complexity, the amount of time, or the work that the attacker needs to gather in order to hack a system"*.

One of the most recent and outstanding works on watermarking security is [30]. In this paper, that was the main inspiration of our subsequent theoretical

analysis, the authors claim that security and robustness are *"neighboring concepts, which are hardly perceived as different"*. According to the authors, *"the intentionality behind the attack is not enough to make a clear cut between these two concepts"*; furthermore the definitions of security and robustness by Kalker [96] are clarified, establishing that robustness is related to *"a classical content processing (compression, low filtering, noise addition, geometric attack...)"*, whereas security is related to attacks *"whose aims are not only the removal of the watermark signal, excluding those already encompassed in the robustness category"*. Following an approach similar to that introduced in [76], that in turn was based on Diffie-Hellman's work [60], a classification of attacks is proposed:

- Watermarked Only Attack (WOA): only watermarked documents are available.
- Known Message Attack (KMA): the attacker can access to watermarked documents and the corresponding messages.
- Known Original Attack (KOA): the attacker can access to the original host signal and their watermarked versions.

The authors also continue with the adaptation of Shannon's concepts about cryptography to watermarking, translating the concept *perfect secrecy* to *perfect covering*, meaning the situation where the observation of watermarked contents does not provide any information about the secret key. For the case of discrete variables, the measure proposed to quantify the uncertainty about the secret key given N_o observations is the conditional entropy

$$H(\mathbf{\Theta}|\mathbf{Y}^1, \mathbf{Y}^2, \cdots, \mathbf{Y}^{N_o}) = H(\mathbf{\Theta}) - I(\mathbf{\Theta}; \mathbf{Y}^1, \mathbf{Y}^2, \cdots, \mathbf{Y}^{N_o}), \tag{4.1}$$

so the information leakage is proposed to be measured by the mutual information between the observations and the secret key. Nevertheless, some problems appear when continuous variables are analyzed; the authors defend that *"the entropy (or the conditional entropy) of a continuous random variable does not measure a quantity of information*, since, *for instance, the equivocation can take positive or non positive values"*. This argument can be criticizable, given that differential entropy of continuous random variables is just related to the volume of their typical sets [51]; although it can yield negative values,[1] its results can be still really insightful.

Due to these problems with the entropy of continuous random variables, in [30] the Fisher measure is proposed to quantify the information leakage. The Fisher Information Matrix [71], can be used jointly with the Cramér-Rao theorem [58] to

[1] In fact, the differential entropy of a deterministic variable is $-\infty$.

provide a lower bound on the covariance matrix of an unbiased estimator of a parameter (in watermarking, the secret key). Specifically, the authors characterize the information leakage using the variable

$$N_o^\star = N_o \mathrm{tr}(\mathrm{FIM}(\boldsymbol{\Theta})^{-1}), \tag{4.2}$$

where N_o is the number of observations, and $\mathrm{FIM}(\boldsymbol{\Theta})$ is the Fisher Information Matrix of the secret key. The larger $N_o^\star$ is, the more secure the system will be, in such a way that the security level can be measured as $O(N_o^\star)$. Nevertheless, this measure neglects some important parameters as the uncertainty in the secret key or the watermarked signal, as it will be shown in Section 4.4.2.

The framework proposed in [30] is particularized for the security analysis of Add-SS; in Add-SS the secret key determines the spreading sequences, so the parameters the attacker would like to estimate, and therefore those involved in the FIM computation, are such spreading sequences. The analysis is performed for the three attacks described above, and the main conclusions are:

- For KMA the information leakage is linear with the number of observation N_o, whereas $N_o^\star = O(\mathrm{DWR})$.
- The KOA case is related to a blind source separation in a noisy environment, and the spreading sequences can be identified up to a signed permutation ambiguity.
- For the WOA case the embedded messages play the role of nuisance parameters, that make more difficult the estimate of the spreading sequences.

Finally, the authors of [30] also propose algorithms based on Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to get access to the watermarking channel. We would like to cite another simultaneous work in the literature [62], where PCA is also used to disclose the watermarking channel.

4.2. Definitions and measures

Based on the works described in the previous section, fundamental definitions and theoretical measures of security have been proposed in [38], and are summarized next.

Firstly, a definition of robustness is proposed, focusing on the fact that attacks to robustness are oriented to increase the probability of error of the data-hiding channel. On the other hand, in attacks to security the attacker is interested in gaining knowledge about the secret key; in this sense, if this knowledge is

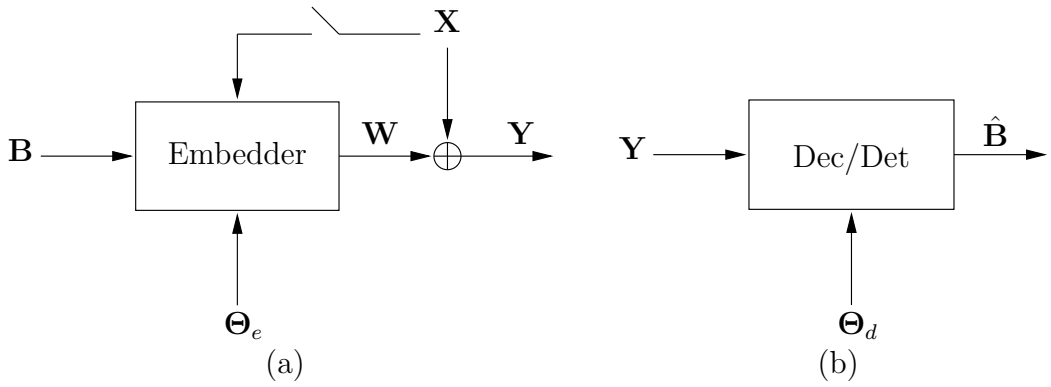


Figure 4.1: General model for security analysis: embedding (a) and decoding/detection (b).

later used to increase the probability of error, security attacks could be regarded as a previous step to attacks to robustness. Since the attacker is aware that he/she is attacking the system, and he/she is assumed to know all the details of the data-hiding system except for the secret key (following Kerckhoffs' principle [99]), all attacks to security are intentional and non-blind. Some other interesting considerations and relationships between robustness and security can be found in [38].

In order to define a security measure, Shannon's cryptographic approach [140] was translated to data-hiding; this measure was already foreseen for watermarking by Hernández and Pérez-González in [91]. Nevertheless, some differences must be taken into account when continuous random variables are considered. In fact, the entropy in the discrete case, is replaced by the differential entropy in the continuous case, so even though the entropy of a deterministic discrete variable is 0, the differential entropy of a deterministic continuous variable is $-\infty$.

Depending on what side of the data-hiding system is considered, security can be analyzed in two different scenarios, which are depicted in Figure 4.1:

1. For the scenario depicted in Figure 4.1-a, informaton leakage is measured by the mutual information between the observations $\mathbf{Y}$ and the secret key $\mathbf{\Theta}$

$$\begin{aligned} I(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta}) &= h(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}) - h(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o} | \mathbf{\Theta}) \\ &= h(\mathbf{\Theta}) - h(\mathbf{\Theta} | \mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}), \end{aligned} \tag{4.3}$$

   where $h(\cdot)$ denotes the differential entropy, and $\mathbf{Y}^n$ the n-th observation.[2] Equivocation is defined as the remaining uncertainty about the key after

[2]The observations are produced from independent signals watermarked with the same secret key $\mathbf{\Theta}$.

the observations:

$$h(\mathbf{\Theta}|\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}) = h(\mathbf{\Theta}) - I(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta}). \tag{4.4}$$

This scenario encompasses attacks concerning the observation of watermarked signals, where it is possible that additional parameters like the embedded message $\mathbf{B}$ or the host $\mathbf{X}$ are also known by the attacker. The model is valid for either side-informed and non-side-informed watermarking/data-hiding schemes.

2. In the scenario depicted in Figure 4.1-b the attacker tries to gain knowledge about the secret key $\mathbf{\Theta}$ by observing the outputs $\hat{\mathbf{B}}$ of the detector/decoder corresponding to some chosen inputs $\mathbf{Y}$; it includes *oracle attacks*, and the information leakage is measured by

$$I(\hat{\mathbf{B}}^1, \ldots, \hat{\mathbf{B}}^{N_o}, \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta}),$$

where, in this case, the $\mathbf{Y}^n$ are not necessarily watermarked objects but any arbitrary signal, for instance the result of the iterations of an attacking algorithm.

In both cases, relationships can be established between the residual entropy and the variance of the estimation error (namely σ^2_{EE}). For example, for the first scenario, it is possible to write

$$\sigma^2_{EE} \geq \frac{1}{2\pi e} e^{2h(\mathbf{\Theta}|\mathbf{Y})}. \tag{4.5}$$

This estimation error variance could be related to the probability of success of an attack, so (4.5) would enable the computation of the minimum number of observations needed to achieve a given estimation variance, which ensures certain probability of success of the attack; this minimum number of observations could be also considered as a measure of the security of the system. Nevertheless, the relation between the probability of success of an attack and the variance of the estimation error is not straightforward, so in the subsequent analyses we will use the information theoretic measures described above. Finally, and concerning those measures, we would like to note that from the three involved quantities (i.e., $h(\mathbf{\Theta})$, $I(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta})$, and $h(\mathbf{\Theta}|\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o})$ for the first scenario, and $h(\mathbf{\Theta})$, $I(\hat{\mathbf{B}}^1, \ldots, \hat{\mathbf{B}}^{N_o}, \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta})$, and $h(\mathbf{\Theta}|\hat{\mathbf{B}}^1, \ldots, \hat{\mathbf{B}}^{N_o}, \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o})$ for the second one), at least two[3] must be provided in order to have a complete picture of the security of the scheme:

- $h(\mathbf{\Theta})$ just provides information on the a priori uncertainty about the key, and does not depend on the analyzed watermarking scheme.

[3]The third measure can be straightforwardly computed from the other two.

- $I(\cdot;\mathbf{\Theta})$ just provides information about the information leakage, i.e. the knowledge the attacker can gain, but it does not help in knowing what was the a priori uncertainty, or what is the residual (given the observations) entropy of the secret key. This is an important issue, since the mutual information could be really small, but if the a priori uncertainty is also reduced, the system can hardly be thought of as being secure.

- $h(\mathbf{\Theta}|\cdot)$ reflects the uncertainty about the value of the secret key when the available information is considered, but it does not provide any information about the a priori uncertainty.

4.3. Analyzed attacks

We have performed the security analyses of the scenario depicted in Figure 4.1-a under several attacks. These attacks are basicly those introduced in [30], although some modifications were introduced, as it will explained later. Depending on the considered attack, the information theoretic security measures are given by:

- Known Message Attack (KMA): the mutual information between the received signal and the secret key, when the sent message is known by the attacker, is computed as

$$\begin{aligned} I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{\Theta}|\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}) &= h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}) \\ &\quad - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}), \end{aligned}$$

  so the residual entropy will be

$$\begin{aligned} h(\mathbf{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}) &= h(\mathbf{\Theta}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}) \\ &\quad + h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}). \end{aligned}$$

- Watermarked Only Attack (WOA): the mutual information between the observations and the secret key is

$$I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{\Theta}) = h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta})$$

  and the residual entropy will be

$$\begin{aligned} h(\mathbf{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) &= h(\mathbf{\Theta}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) + I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}|\mathbf{\Theta}) \\ &\quad + h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}). \end{aligned}$$

- Estimated Original Attack (EOA): In this case the following will be computed

$$I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{\Theta}|\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}) = h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}), \quad (4.6)$$

  where $\hat{\mathbf{X}}^i \triangleq \mathbf{X}^i + \tilde{\mathbf{X}}^i$ is an estimate of the original host signal for the i-th observation $\mathbf{X}^i$ and $\tilde{\mathbf{X}}^i$ is the estimation error; $\tilde{\mathbf{X}}^i$ is assumed to have power σ_E^2 and to be independent of $\mathbf{X}^i$. The Known Original Attack (KOA) proposed in [30] can be regarded to as a particular case of EOA, where the variance of the original host estimation error is set to 0. On the other hand, when the original host estimation error is σ_X^2, we are in the WOA case, so it can be also seen as particular case of EOA. The attacker could obtain this estimate by averaging several versions of the same host watermarked with different keys, but in order to ensure independence between the key and the estimate, the watermarked version with the to-be-estimated key should not be included in the averaging. Other alternative could be to filter the watermarked signal to compute the estimate of the original host (assuming the resulting signal is independent of the watermark).

  Taking into account (4.6), it is possible to write

$$h(\mathbf{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}) = h(\mathbf{\Theta}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}) + h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}).$$

- Constant Message Attack (CMA): The attacker does not know which message is embedded in each observation, but he/she does know that it is the same for all of them. This attack makes sense for applications such as fingerprinting or copyright, where the attacker can have access to different documents, or even different blocks of the same document (this is the case of video sequences, for example) which are watermarked with the same secret key and the same message, in order to facilitate synchronization. In this case we will denote the mutual information as:

$$I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{\Theta}|\mathrm{CM}) = h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathrm{CM}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\mathrm{CM}),$$

  where CM stands for *Constant Message*. Therefore, the residual entropy will be

$$h(\mathbf{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\mathrm{CM}) = h(\mathbf{\Theta}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathrm{CM}) + h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{\Theta},\mathrm{CM}).$$

When $N_o = 1$, the superscript denoting the observation will be obviated for notation simplicity.

Finally, note that, depending on the method, the secret key could be related to the watermarking scheme parameters (i.e. the spreading sequence in spread-spectrum, the dither sequence or lattice rotation parameters in DC-DM, or the codebook in Costa's scheme with random codebook) through a deterministic function, constructing a Markov chain, in such a way that the attacker could be interested in just estimating the result of this function and not in the secret key itself; furthermore, even if he/she wants to estimate the secret key, this estimate will be based on the aforementioned parameters. Therefore, in the following sections we will study the mutual information and/or residual entropy of these scheme parameters, i.e. spreading sequences for Add-SS, codebook for Costa's scheme, and dither vector for DC-DM, instead of the corresponding to the *actual* secret key.

4.4. Security Analysis of Add-SS Watermarking

In order to make a fair comparison with [30], the definition of Add-SS given at Section 2.3 is now changed to remove the assumption that the coefficients devoted to convey a given symbol constitute a subvector whose components are taken from a subset which is disjoint from the subsets corresponding to other symbols. Now, L_b random vectors $\mathbf{S}_i$ (i.e., the spreading sequences), one for each symbol to be hidden, are generated depending on the secret key $\mathbf{\Theta}$. In this way, the embedding function can be written as:

$$\mathbf{Y}^j = \mathbf{X}^j - \frac{1}{\sqrt{L_b}} \sum_{i=1}^{L_b} \mathbf{S}_i (-1)^{B_i^j}, \quad 1 \leq j \leq N_o, \tag{4.7}$$

with the watermarked signal $\mathbf{Y}^j$, the original host signal $\mathbf{X}^j$ and the spreading sequences $\mathbf{S}_i$ L_1-dimensional vectors, where $S_{i,j}$ is the j-th component of the i-th spreading sequence. The host is modeled as an i.i.d. Gaussian process, $\mathbf{X}^j \sim \mathcal{N}(0, \sigma_X^2 \mathbf{I}_{L_1})$, and the message letters $B_i^j \in \{0, 1\}$, being $Pr\{B_i^j = 0\} = Pr\{B_i^j = +1\} = 1/2$. All of these quantities are assumed to be mutually independent. Since (4.7) is related to the secret key $\mathbf{\Theta}$ only through the spreading sequences $\mathbf{S}_i$'s, we will measure the security with respect to the $\mathbf{S}_i$'s.

4.4.1. Known Message Attack

First of all, we will consider the case with only one observation, that is, $N_o = 1$. In this case, the information leakage can be computed as $I(\mathbf{Y}; \mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{L_b} | \mathbf{B})$, so for a generic distribution of $\mathbf{S}_i$ numerical integration must be used. In Figure 4.2 and Figure 4.3 the results of this numerical integration are shown for the

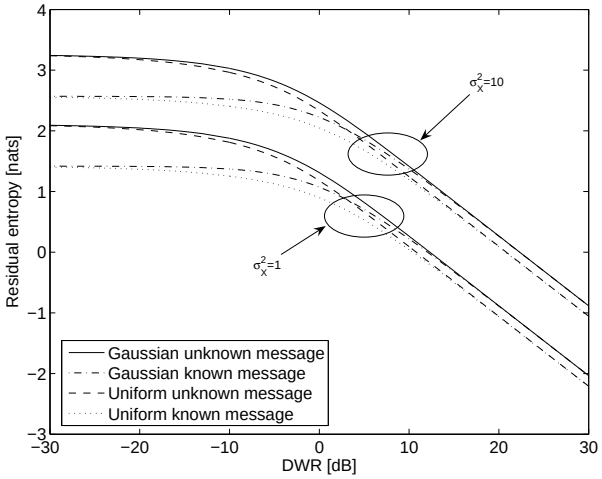


Figure 4.2: Results of numerical integration for the equivocation $h(\mathbf{S}_1|\mathbf{Y})$ and $h(\mathbf{S}_1|B,\mathbf{Y})$ in Add-SS for Gaussian and uniform distributions of $\mathbf{S}_1$. $L_1 = 1$ and $L_b = 1$.

case of one transmitted symbol $L_b = 1$ and both Gaussian and uniform distributions of $\mathbf{S}_1$ in the scalar case. Those figures show that the information the attacker can not learn (i.e., $h(\mathbf{S}_1|\mathbf{B},\mathbf{Y})$) is larger if $\mathbf{S}_1$ is chosen to be Gaussian. Taking this into account, we will focus on the case $\mathbf{S}_i \sim \mathcal{N}(0, \sigma_S^2 \mathbf{I}_{L_1})$. When the sent symbol is known to the attacker, the following result is derived in Appendix D.1 for $L_1 > 1$, $L_b > 1$ and $N_o = 1$,

$$I(\mathbf{Y}; \mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{L_b}|\mathbf{B}) = \frac{L_1}{2} \log\left(1 + \frac{\sigma_S^2}{\sigma_X^2}\right), \tag{4.8}$$

yielding

$$h(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{L_b}|\mathbf{Y}, \mathbf{B}) = \frac{L_1}{2} \log\left[\left(2\pi e \frac{\sigma_S^2}{L_b}\right)^{L_b} \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_S^2}\right].$$

The result in (4.8) says that the information that an attacker can obtain is the same whatever the number of carriers, although the entropy of the key is a linear function of this parameter (this result applies to a great variety of pdfs for the key, since by the central limit theorem, the sum of the carriers tends to a Gaussian). This result is also a consequence of the power normalization performed in (4.7); independently of the number of carriers, the power of the watermark stays constant.

In Appendix D.2, we analyze the case of one sent bit ($L_b = 1$), $L_1 = 1$, when there are several available observations ($N_o > 1$), all of them watermarked with the same secret key. If $L_1 > 1$ and the components are independent, the result is also valid, after multiplying it by L_1, so we can write

$$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{S}_1|\mathbf{B}^1, \cdots, \mathbf{B}^{N_o}) = \frac{L_1}{2} \log\left(1 + \frac{N_o \sigma_S^2}{\sigma_X^2}\right), \tag{4.9}$$

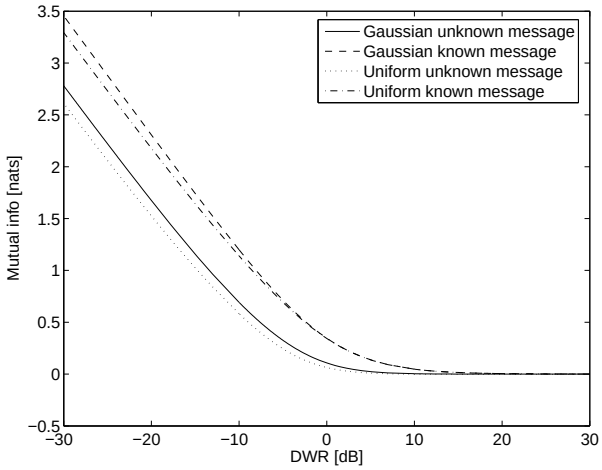


Figure 4.3: Results of numerical integration for $I(\mathbf{Y};\mathbf{S}_1)$ and $I(\mathbf{Y};\mathbf{S}_1|B)$ in Add-SS for Gaussian and uniform distribution of $\mathbf{S}_1$. $L_1 = 1$ and $L_b = 1$.

which yields

$$h(\mathbf{S}_1|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\mathbf{B}^1,\cdots,\mathbf{B}^{N_o}) = \frac{L_1}{2}\log\left(2\pi e\frac{\sigma_S^2\sigma_X^2}{N_o\sigma_S^2+\sigma_X^2}\right). \tag{4.10}$$

This result shows that $I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{S}_1|\mathbf{B}^1,\cdots,\mathbf{B}^{N_o})$ grows non-linearly with the number of observations, although for large Document to Watermark Ratios (DWR $>>$ 1) and low values of N_o the growth is almost linear. Moreover, (4.9) coincides with the capacity of a Gaussian channel with signal power σ_S^2 and noise power σ_X^2/N_o. This suggests that the best method the attacker should follow for estimating $\mathbf{S}_1$ is just to average the observations $\mathbf{Y}^i$ (at least this is the case when both the host signal and the watermark are Gaussian distributed). In Figure 4.4 the mutual information is compared with an upper-bound (which is based on the linear approximation for small values of N_o) when DWR = 30 dB.

4.4.2. Comparison with the result in [30]

In [30], the security level is defined as $O(N_o^\star)$, where $N_o^\star \triangleq N_o\text{tr}(\text{FIM}(\boldsymbol{\theta})^{-1})$ with FIM($\boldsymbol{\theta}$) the Fisher Information Matrix of $\boldsymbol{\theta}$. In this section we try to link the result obtained in that paper with the one given here for Add-SS KMA when only one symbol is transmitted, i.e. $L_b = 1$.

It is shown in Appendix E that the FIM obtained when a constant multiple (i.e., vector) parameter is estimated in the presence of i.i.d. Gaussian noise, taking into account N_o independent observations in the estimate, is $\frac{N_o}{\sigma_X^2}\mathbf{I}_{L_1}$, where σ_X^2 is the power of the interfering signal (the original host in our case). This

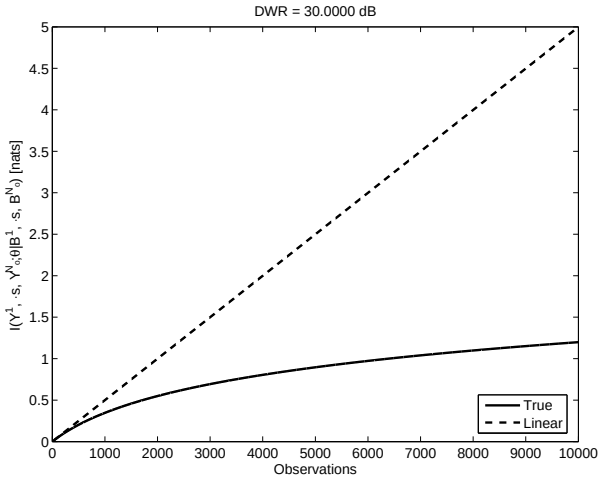


Figure 4.4: $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{\Theta} | \mathbf{B}^1, \cdots, \mathbf{B}^{N_o})$ for Add-SS and Known Message Attack. DWR = 30 dB. $L_b = 1$.

is the only term considered in [30]. Nevertheless, an additional term should be included, due to the random nature of the secret key (see [151]):

$$J_{P_{ij}} = \mathrm{E}\left[\frac{\partial \log f_{\mathbf{S}_1}(\mathbf{s}_1)}{\partial s_{1,i}} \cdot \frac{\partial \log f_{\mathbf{S}_1}(\mathbf{s}_1)}{\partial s_{1,j}}\right]. \tag{4.11}$$

If $\mathbf{S}_1$ is an i.i.d. Gaussian vector, it is easy to prove that $\mathbf{J}_P = \frac{1}{\sigma_S^2}\mathbf{I}_{L_1}$, so $\mathrm{FIM}(\mathbf{S}_1) = \left(\frac{N_o}{\sigma_X^2} + \frac{1}{\sigma_S^2}\right)\mathbf{I}_{L_1}$, yielding

$$N_o^\star = L_1 \frac{\sigma_X^2 \sigma_S^2}{\sigma_S^2 + \sigma_X^2/N_o},$$

which is obviously related to the proposed information-theoretic approach, since (4.10) is the differential entropy of a i.i.d. Gaussian random vector with covariance matrix $N_o^\star/(N_o L_1)\mathbf{I}_{L_1}$.

On the other hand, if we had considered only the FIM obtained when estimating a constant multiple parameter, the obtained $N_o^\star$ would be $L_1\sigma_X^2$, which is obviously related to $h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathbf{S}_1, \mathbf{B}^1, \cdots, \mathbf{B}^{N_o}) = \frac{L_1 N_o}{2}\log(2\pi e \sigma_X^2)$; this was the methodology followed in [30]. Therefore, it does not take into account the entropy of the secret key neither the entropy of the watermarked signal. As stated in Section 4.2, both terms are relevant for the analysis of the system, so they should be considered. In fact, $h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathbf{S}_1, \mathbf{B}^1, \cdots, \mathbf{B}^{N_o})$ for the KMA case grows linearly with the number of observations, while the mutual information will not increase linearly due to the dependence between observations. The linear approximation is actually an upper-bound; the larger the number of observations, the worse this approximation is.

4.4.3. Watermarked Only Attack

Due to the symmetry of the pdfs, it is possible to conclude that the components of the watermarked vector $\mathbf{Y}$ are still mutually independent, so for $L_b = 1$ and a single observation, we can write

$$I(\mathbf{Y};\mathbf{S}_1) = L_1 I(Y_i;S_{1,i}) = L_1\left(h(Y_i) - h(Y_i|S_{1,i})\right) \tag{4.12}$$

$$= L_1\left(h(Y_i|\mathbf{B}=\mathbf{0}) - h(Y_i|S_{1,i})\right). \tag{4.13}$$

In order to determine this for a generic distribution of the spreading sequence $\mathbf{S}_1$, numerical integration should be used, whose results are plotted in Figure 4.2. Once again, the information the attacker can not learn (i.e., $h(\mathbf{S}_1|\mathbf{Y})$) is larger for the shown cases when $\mathbf{S}_1$ is chosen to be Gaussian. Therefore, assuming $\mathbf{S}_1$ to be Gaussian, we can write

$$I(\mathbf{Y};\mathbf{S}_1) = L_1\left(\frac{1}{2}\log\left(2\pi e(\sigma_X^2+\sigma_S^2)\right) - h(Y_i|S_{1,i})\right). \tag{4.14}$$

The rightmost term of (4.14) must still be numerically computed. When DWR $<<$ 1 we can easily analyze the asymptotic behavior of the mutual information taking into account that $h(\mathbf{Y}) \approx h(\mathbf{S}_1)$ and $h(\mathbf{Y}|\mathbf{S}_1) \approx h(\mathbf{X}) + \log(2)$, yielding

$$I(\mathbf{Y};\mathbf{S}_1) \approx h(\mathbf{S}_1) - h(\mathbf{X}) - \log(2), \tag{4.15}$$

$$I(\mathbf{Y};\mathbf{S}_1|\mathbf{B}) \approx h(\mathbf{S}_1) - h(\mathbf{X}). \tag{4.16}$$

This explains and quantifies the asymptotic gap between the WOA and KMA cases, which is exactly $\log(2) = 0.69$ nats. Nevertheless, note that a very small DWR is not practical, since it would yield unuseful watermarked images. This case has been introduced here only to shed some light into the general behavior of the mutual informations. On the other hand, to compute the gap between a Gaussian and a uniform distribution for $\mathbf{S}_1$, $h(\mathbf{S}_1)$ will be determined in both cases for a constant variance σ_S^2,

$$h(\mathbf{S}_{Gauss}) - h(\mathbf{S}_{unif}) = \frac{1}{2}\log(2\pi e\sigma_S^2) - \frac{1}{2}\log(12\sigma_S^2) = \frac{1}{2}\log\left(\frac{\pi e}{6}\right) = 0.1765 \text{ [nats]},$$

which will be the asymptotic gap (in residual entropy terms) between the Gaussian and uniform cases for both known and unknown messages (see Figure 4.2) when DWR $>>$ 1, since for a large DWR both $I(\mathbf{Y};\mathbf{S}_1)$ and $I(\mathbf{Y};\mathbf{S}_1|\mathbf{B})$ are approximately 0.

For L_b carriers and one observation, i.e. $N_o = 1$, we have, similarly to the KMA case, the following mutual information:

$$\begin{aligned} I(\mathbf{Y};\mathbf{S}_1,\mathbf{S}_2,\ldots,\mathbf{S}_{L_b}) &= L_1 I(Y_i;S_{1,i},S_{2,i},\ldots,S_{L_b,i}) \\ &= L_1\left(h(Y_i) - h(Y_i|S_{1,i},\ldots,S_{L_b,i})\right) \\ &= L_1\left[\frac{1}{2}\log(2\pi e(\sigma_X^2+\sigma_S^2)) - h(Y_i|S_{1,i},\ldots,S_{L_b,i})\right], \end{aligned}$$

where the second term of the last equality must be numerically computed again.

The case of one sent bit ($L_b = 1$), unidimensional host $L_1 = 1$, and several available observations ($N_o > 1$) needs very expensive numerical computations. Practical computations demand the reduction of the number of available observations to a very small value; in that case, the mutual information will be in the linear region, so no knowledge is available about the growth of the mutual information for large values of N_o. However, it is obvious that the mutual information in this scenario will be upper-bounded by that obtained for KMA.

4.4.4. Estimated Original Attack

In this case, the attacker will have access to an estimate of the original host signal, with some estimation error denoted by $\tilde{\mathbf{X}}$, which is assumed to be i.i.d. Gaussian with variance σ_E^2, in such a way that for $N_o = 1$ we can write $I(\mathbf{Y}; \mathbf{S}_1, \cdots, \mathbf{S}_{L_b} | \mathbf{X} + \tilde{\mathbf{X}}) = L_1 \Big[h(Y_i | X_i + \tilde{X}_i) - h(Y_i | X_i + \tilde{X}_i, S_{1,i}, \cdots, S_{L_b,i}) \Big]$. Assuming that $\sigma_X^2 >> \sigma_E^2$, $\tilde{X}_i$ will be almost orthogonal (and therefore independent) to $X_i + \tilde{X}_i$, so

$$I(\mathbf{Y}; \mathbf{S}_1, \cdots, \mathbf{S}_{L_b} | \mathbf{X} + \tilde{\mathbf{X}}) \approx L_1 \left\{ h\left(\frac{-1}{\sqrt{L_b}} \sum_{j=1}^{L_b} S_{j,i} (-1)^{B_j} - \tilde{X}_i \right) - h\left(\frac{-1}{\sqrt{L_b}} \sum_{j=1}^{L_b} S_{j,i} (-1)^{B_j} - \tilde{X}_i | S_{1,i}, \cdots, S_{L_b,i} \right) \right\}.$$

This situation is equivalent to that described in 4.4.3, but replacing σ_X^2 by σ_E^2, so when $L_b = 1$ it is possible to use Figure 4.2 for obtaining numerical results, using the *Estimation error to Watermark Ratio* (EWR), defined as $\frac{\sigma_E^2}{\sigma_S^2}$, instead of the DWR, in the horizontal axis. When the estimate is perfect, i.e. $\sigma_E^2 = 0$, the mutual information approaches infinity.

4.4.5. Constant Message Attack

For the case of just one observation available, i.e. $N_o = 1$, the CMA is equivalent to the WOA (see Section 4.4.3), since the attacker can not take advantage of knowing that the same message has been embedded in all the observations. On the other hand, when several observations are available, we can write

$$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{S}_1, \cdots, \mathbf{S}_{L_b} | \mathrm{CM}) =$$
$$h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathrm{CM}) - h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathbf{S}_1, \cdots, \mathbf{S}_{L_b}, \mathrm{CM}) =$$
$$h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathrm{CM}) - \textstyle\sum_{i=1}^{N_o} h(\mathbf{Y}^i | \mathbf{S}_1, \cdots, \mathbf{S}_{L_b}, \mathbf{Y}^1, \cdots, \mathbf{Y}^{i-1}, \mathrm{CM}),$$

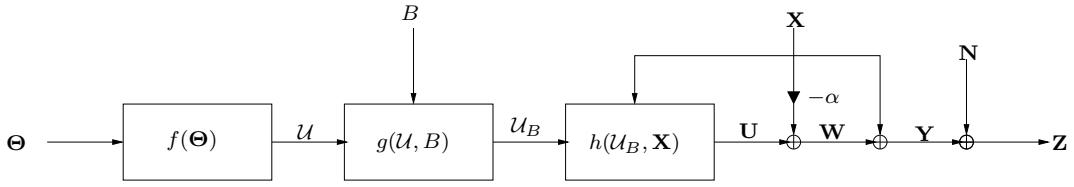


Figure 4.5: Block diagram of Costa's schemes.

where the leftmost term was already computed in Appendix D.2, and the rightmost one requires numerical computation. Again, as it happened in Section 4.4.3 for WOA, practical computations demand the reduction of the number of available observations to a very small value; in that case, the mutual information will be in the linear region, and the behavior of the mutual information for large value of N_o can not be predicted.

4.5. Security Analysis of Costa's construction (Random codebooks)

As it is well-known, one of the main advantages of lattice-based implementations of DC-DM is their highly structured nature. This structure makes easier the embedding and the decoding; but on the other hand, it could be also used by the attacker to gain knowledge about the complete codebook, making easier a security attack, so it can be also seen as a security flaw of these methods. We will talk about the security of lattice-based DC-DM methods in Section 4.6, but in this section we will try to analyze the performance of side-informed methods when this structure is removed. In this way we could compare the two extreme cases: completely structured codebooks, and codebooks without structure at all; this comparison could provide some insight on the security performance of intermediate situations. Obviously, a codebook without structure can be achieved by choosing a random codebook. This is the case of Costa's construction, where the codebook is random by definition.

In Figure 4.5 the considered framework is represented. The randomness can be parameterized by a secret key $\mathbf{\Theta}$, resulting in a codebook $\mathcal{U} = f(\mathbf{\Theta})$. This codebook is partitioned in as many bins as possible messages. Depending on the sent message b, the corresponding bin in the codebook will be chosen, namely $\mathcal{U}_b = g(\mathcal{U}, b)$. Taking into account the host signal $\mathbf{X}$ and the distortion compensation parameter α (which belongs to the interval [0,1]) the encoder will look for a sequence $\mathbf{U} = h(\mathcal{U}_b, \mathbf{X})$ belonging to $\mathcal{U}_b$ such that $|(\mathbf{U} - \alpha\mathbf{X})^t\mathbf{X}| \leq \delta$, for some arbitrarily small δ. The watermark signal will be $\mathbf{W} = \mathbf{U} - \alpha\mathbf{X}$, and the watermarked signal $\mathbf{Y} = \mathbf{X} + \mathbf{W}$. Finally, the decoder will observe $\mathbf{Z} = \mathbf{X} + \mathbf{W} + \mathbf{N}$, where $\mathbf{N}$ is the channel noise, independent of both $\mathbf{X}$ and $\mathbf{W}$. The random vec-

tors $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{N}$ are also i.i.d., with distributions $\mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_{L_1})$, $\mathcal{N}(\mathbf{0}, \sigma_W^2 \mathbf{I}_{L_1})$ and $\mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I}_{L_1})$, respectively, where $\mathbf{I}_{L_1}$ denotes the L_1-th order identity matrix.

4.5.1. Known Message Attack

4.5.1.1. One available observation ($N_o = 1$)

Since knowledge of the codebook and the sent symbol implies knowledge of the bin in the codebook (i.e., $\mathcal{U}_b$), we can write

$$I(\mathbf{Y};\mathcal{U}|B) = h(\mathbf{Y}) - I(\mathbf{Y};B) - h(\mathbf{Y}|\mathcal{U}_B).$$

In Appendix F.1, we show that if $\alpha > 0.2$, then

$$I(\mathbf{Y};\mathcal{U}|B) = \frac{L_1}{2} \log \left[\frac{\sigma_W^2 + \sigma_X^2}{(1-\alpha)^2 \sigma_X^2} \right],$$

so

$$h(\mathcal{U}|\mathbf{Y}, B) = h(\mathcal{U}) - \frac{L_1}{2} \log \left[\frac{\sigma_W^2 + \sigma_X^2}{(1-\alpha)^2 \sigma_X^2} \right]. \tag{4.17}$$

Since each component of each sequence $\mathbf{U}$ follows a Gaussian distribution with power $\sigma_W^2 + \alpha^2 \sigma_X^2$, and all of them are mutually independent, it follows that

$$h(\mathcal{U}) = \frac{|\mathcal{U}| L_1}{2} \log \left[2\pi e (\sigma_W^2 + \alpha^2 \sigma_X^2) \right],$$

where $|\mathcal{U}| = e^{I(\mathbf{U};\mathbf{Z})} = \left(\frac{[\sigma_W^2 + \sigma_X^2 + \sigma_N^2][\sigma_W^2 + \alpha^2 \sigma_X^2]}{\sigma_W^2 \sigma_X^2 (1-\alpha)^2 + \sigma_N^2 (\sigma_W^2 + \alpha^2 \sigma_X^2)} \right)^{L_1/2}$.

Equation (4.17) shows that the higher the DWR is, the higher the residual entropy becomes, because the host signal is making difficult the estimation of the secret key. On the other hand, the larger α, the smaller the residual entropy, since the self-noise is reduced and the estimation becomes easier. In Figures 4.6 and 4.7, the theoretical results are plotted for different values of α and the DWR.

4.5.1.2. Multiple observations ($N_o \geq 1$)

In order to have a first approximation, we will assume that the attacker knows the index of the codeword related to each observation, which we will denote by the random vector $\mathcal{J}$. Taking this into account the residual entropy can be written as

$$\begin{aligned} h(\mathcal{U}|\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}, B^1, \cdots, B^{N_o}, \mathcal{J}) &= h(\mathcal{U}|B^1, \cdots, B^{N_o}, \mathcal{J}) \\ &\quad - I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o};\mathcal{U}|B^1, \cdots, B^{N_o}, \mathcal{J}) \\ &\leq h(\mathcal{U}|\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}, B^1, \cdots, B^{N_o}). \end{aligned}$$

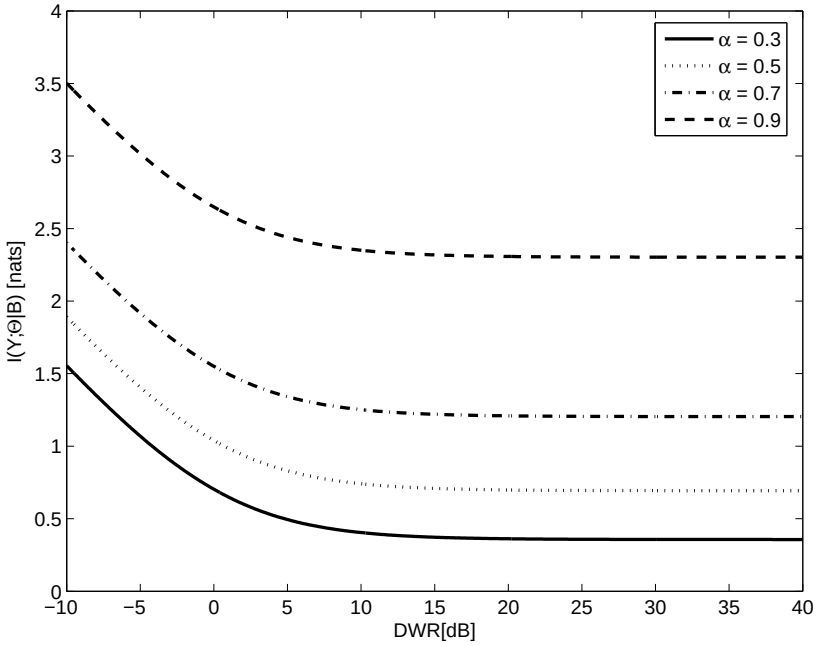


Figure 4.6: $I(\mathbf{Y};\mathcal{U}|B)$ for Costa in nats vs. DWR, for different values of α and $L_1 = 1$.

Since the message and the index of the observed codeword are independent of the codebook, we have that $h(\mathcal{U}|B^1,\cdots,B^{N_o},\mathcal{J}) = h(\mathcal{U})$. On the other hand,

$$I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U}|B^1,\cdots,B^{N_o},\mathcal{J}) =$$
$$h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|B^1,\cdots,B^{N_o},\mathcal{J}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathcal{U},B^1,\cdots,B^{N_o},\mathcal{J}) =$$
$$\textstyle\sum_{i=1}^{N_o} h(\mathbf{Y}^i|B^1,\cdots,B^{N_o},\mathcal{J},\mathbf{Y}^1,\cdots,\mathbf{Y}^{i-1}) - h(\mathbf{Y}^i|\mathcal{U},B^1,\cdots,B^{N_o},\mathcal{J},\mathbf{Y}^1,\cdots,\mathbf{Y}^{i-1}),$$

but the i-th observation is independent of the previous observations which are not related to the same codeword. Furthermore, we can arrange the codewords of the j-th message in such a way that if we denote by $L_{i,j}$ the number of observations related to the i-th codeword of the j-th message, $L_{i_1,j} \geq L_{i_2,j}$ for all $i_2 > i_1$ with $1 \leq i_1, i_2 \leq |\mathcal{U}_j|$, without modifying the entropy. Therefore, for a given realization of the sequence of messages $\{B^1,\cdots,B^{N_o}\}$ and $\mathcal{J}$, the previous equation yields

$$\sum_{j=0}^{P-1}\sum_{i=1}^{|\mathcal{U}_j|}\sum_{k=1}^{L_{i,j}} h(\mathbf{Y}_{j,i,k}|\mathbf{Y}_{j,i}^{k-1}) - h(\mathbf{Y}_{j,i,k}|\mathcal{U}_j,\mathbf{Y}_{j,i}^{k-1},\mathbf{Y}_{j,i-1}^{L_{i-1,j}},\cdots,\mathbf{Y}_{j,1}^{L_{1,j}}), \quad (4.18)$$

where $\mathbf{Y}_{j,i,k}$ is the k-th observation related to the i-th codeword of the j-th message, and $\mathbf{Y}_{j,i}^k$ is the vector containing the k first observations related to such codeword (the notation has been changed in order to clarify our exposition).

The leftmost term in (4.18) can be developed as

$$\begin{aligned} h(\mathbf{Y}_{j,i,k}|\mathbf{Y}_{j,i}^{k-1}) &= h(\mathbf{Y}_{j,i,k}|\mathbf{Y}_{j,i}^{k-1},\mathbf{U}_{j,i}) + I(\mathbf{Y}_{j,i,k};\mathbf{U}_{j,i}|\mathbf{Y}_{j,i}^{k-1}) \\ &= h(\mathbf{Y}_{j,i,k}|\mathbf{U}_{j,i}) + h(\mathbf{U}_{j,i}|\mathbf{Y}_{j,i}^{k-1}) - h(\mathbf{U}_{j,i}|\mathbf{Y}_{j,i}^{k-1},\mathbf{Y}_{j,i,k}) \\ &= h(\mathbf{Y}_{j,i,k}|\mathbf{U}_{j,i}) + h(\mathbf{U}_{j,i}|\mathbf{Y}_{j,i}^{k-1}) - h(\mathbf{U}_{j,i}|\mathbf{Y}_{j,i}^{k}), \end{aligned}$$

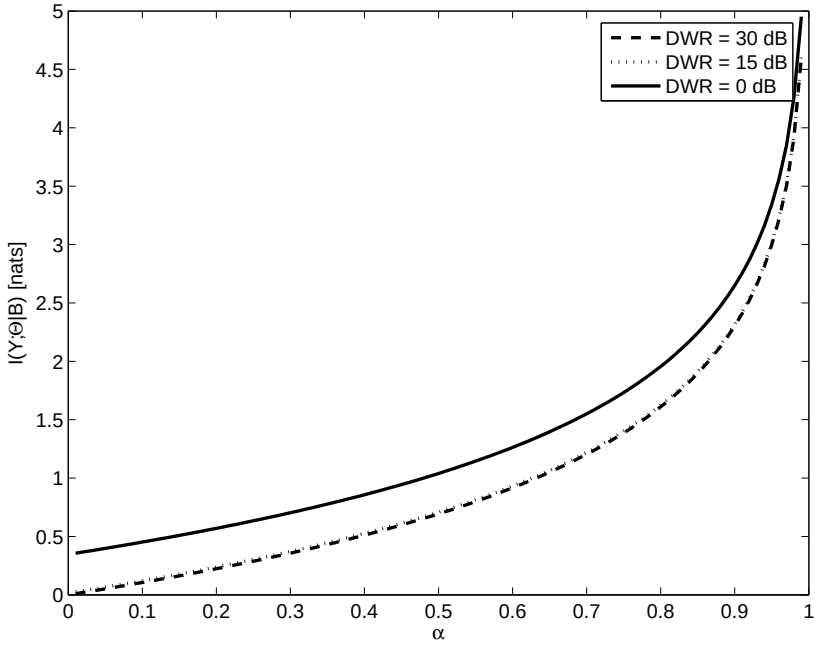


Figure 4.7: $I(\mathbf{Y};\mathcal{U}|B)$ for Costa in nats vs. α, for different values of DWR and $L_1 = 1$.

where we have taken into account that $\mathbf{Y}_{j,i,k}$ given $\mathbf{U}_{j,i}$ is independent of $\mathbf{Y}_{j,i}^{k-1}$, so $h(\mathbf{Y}_{j,i,k}|\mathbf{Y}_{j,i}^{k-1}, \mathbf{U}_{j,i}) = h(\mathbf{Y}_{j,i,k}|\mathbf{U}_{j,i})$. The uncertainty about the value of one observation when one knows the codeword related to that observation, i.e. $h(\mathbf{Y}_{j,i,k}|\mathbf{U}_{j,i})$, is identical for all codewords and for all the observations. Therefore, and for the sake of simplicity, we will remove the subscripts; in that case, recalling that

$$\begin{aligned}\mathbf{U} &= \mathbf{W} + \alpha\mathbf{X}\\ \mathbf{Y} &= \mathbf{X} + \mathbf{W},\end{aligned}$$

we know that the covariance matrix of $\mathbf{Y}$ given $\mathbf{U}$ can be written as $\mathrm{Cov}\{\mathbf{Y}|\mathbf{U}\} = \frac{(1-\alpha)^2\sigma_W^2\sigma_X^2}{\sigma_W^2+\alpha^2\sigma_X^2}\mathbf{I}_{L_1}$ (see Appendix F.1); furthermore $h(\mathbf{U}_{j,i}|\mathbf{Y}_{j,i}^k)$ is the same for all the codewords, so we can write $\mathbf{U} = d\mathbf{Y} + \mathbf{Y}^\perp$, where $\mathbf{Y}^\perp$ is the component of $\mathbf{U}$ which is orthogonal to $\mathbf{Y}$, so $\mathbf{Y}^\perp = \mathbf{U} - d\mathbf{Y} = (\alpha - d)\mathbf{X} + (1-d)\mathbf{W}$; due to the orthogonality of $\mathbf{Y}$ and $\mathbf{Y}^\perp$,

$$\begin{aligned}||\mathbf{U}||^2 &= L_1(\sigma_W^2 + \alpha\sigma_X^2) = L_1(d^2\sigma_X^2 + d^2\sigma_W^2 + (1-d)^2\sigma_W^2 + (\alpha-d)^2\sigma_X^2)\\ &= d^2||\mathbf{Y}||^2 + ||\mathbf{Y}^\perp||^2,\end{aligned}$$

yielding $d = \frac{\sigma_W^2+\alpha\sigma_X^2}{\sigma_W^2+\sigma_X^2}$, and $\mathrm{Cov}\{\mathbf{U}|\mathbf{Y}\} = \mathrm{Cov}\{\mathbf{Y}^\perp\} = \frac{(1-\alpha)^2\sigma_W^2\sigma_X^2}{\sigma_W^2+\sigma_X^2}\mathbf{I}_{L_1}$. It is also straightforward to see that $\mathrm{Cov}\{\mathbf{U}|\mathbf{Y}_{j,i}^k\} = \frac{(1-\alpha)^2\sigma_W^2\sigma_X^2}{k(\sigma_W^2+\sigma_X^2)}\mathbf{I}_{L_1}$.

Finally, the rightmost term in Equation (4.18) can be shown to be

$$h(\mathbf{Y}_{j,i,k}|\mathcal{U}_j, \mathbf{Y}_{j,i}^{k-1}, \mathbf{Y}_{j,i-1}^{L_{i-1,j}}, \cdots, \mathbf{Y}_{j,1}^{L_{1,j}})$$
$$= \begin{cases} \log(|\mathcal{U}_j| - i + 1) + h(\mathbf{Y}|\mathbf{U}), & \text{if } k = 1 \\ h(\mathbf{Y}|\mathbf{U}), & \text{if } k > 1 \end{cases}.$$

In order to obtain $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o}, \mathcal{J})$, (4.18) has to be averaged over all possible realizations of $\{B^1, \cdots, B^{N_o}\}$ and $\mathcal{J}$. We will show now that the probability of those realizations with $L_{i,j} > 1$, for any $(i,j) \in \{1, \cdots, |\mathcal{U}_j|\} \times \{0, \cdots, P-1\}$ goes to 0, when L_1 goes to infinity, as long as the number of observations N_o verifies a constraint depending on L_1. First of all, we will upperbound the probability of having 2 or more observations related to a codeword; a possible upperbound is given by

$$P\big\{(i,j) \in \{1, \cdots, |\mathcal{U}_j|\} \times \{0, \cdots, P-1\} : L_{i,j} \geq 2\big\} \leq [1 - (1-p)^{N_o}]N_o \triangleq P_{u,\text{KMA}},$$

where

$$p \triangleq \frac{1}{|\mathcal{U}|} = \left(\frac{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_N^2(\sigma_W^2 + \alpha^2\sigma_X^2)}{(\sigma_W^2 + \sigma_X^2 + \sigma_N^2)(\sigma_W^2 + \alpha^2\sigma_X^2)}\right)^{L_1/2}.$$

Then, it can be shown that if N_o is such that $N_o \leq (1/p)^{L_1/8}$, then

$$\lim_{L_1 \to \infty} P_{u,\text{KMA}} = p^{L_1/4},$$

i.e., even for the above described exponential increase in the number of observations, the probability of having 2 or more observations related to the same codeword decreases exponentially.

Finally, for large values of L_1, if $|\mathcal{U}_j| >> N_o$, which is reasonable since $|\mathcal{U}_j|$ increases also exponentially with L_1, then $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o})$ can be accurately approximated as

$$\begin{aligned} I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o}) &\approx I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o}, \mathcal{J}) \\ &\approx N_o\big[h(\mathbf{Y}|\mathbf{U}) + h(\mathbf{U}) - h(\mathbf{U}|\mathbf{Y}) \\ &\quad - \log(|\mathcal{U}_b|) - h(\mathbf{Y}|\mathbf{U})\big] \\ &= N_o\big[h(\mathbf{U}) - h(\mathbf{U}|\mathbf{Y}) - \log(|\mathcal{U}_b|)\big], \end{aligned}$$

since the information provided by $\mathcal{J}$ does not change the value of (4.18), given that, as we have shown, the probability of having more than one observation related to the same codeword decreases exponentially with L_1. In this way, $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o})$ for Costa's scheme linearly increases with the number of observations, contrarily to the observed behaviour for Add-SS or DC-DM (see [127] for this last case, where the dither is the parameter to be estimated). This can be explained because in those schemes the codewords are repeated for all the observations, whereas for Costa's schemes the number of different codewords is huge, thus reducing significantly the probability of observing twice the same one.

4.5.2. Watermarked Only Attack

4.5.2.1. One available observation ($N_o = 1$)

Again, knowledge of the codebook and the sent symbol implies knowledge of the bin in the codebook (i.e., $\mathcal{U}_b$). Therefore, we can write

$$I(\mathbf{Y};\mathcal{U}) = h(\mathbf{Y}) - I(\mathbf{Y};B|\mathcal{U}) - h(\mathbf{Y}|\mathcal{U}_B). \tag{4.19}$$

In Appendix F.2, it is shown that if $\alpha > 0.2$

$$I(\mathbf{Y};\mathcal{U}) = \frac{L_1}{2}\log\left[\frac{(\sigma_W^2+\sigma_X^2)\,(\sigma_W^2\sigma_X^2(1-\alpha)^2+\sigma_N^2(\sigma_W^2+\alpha^2\sigma_X^2))}{\sigma_W^2(\sigma_W^2+\sigma_X^2+\sigma_N^2)(1-\alpha)^2\sigma_X^2}\right]. \tag{4.20}$$

Be aware that we are assuming that the embedder transmits at the maximum reliable rate allowed, thus the expected power of the channel noise will affect the information leakage (this is further explained in Appendix F.2). For instance, when $\sigma_N^2 = 0$, the supremum of the maximum reliable rates is achieved, so the uncertainty about the sent symbol is also maximum, which complicates the attacker's work, yielding in this case $I(\mathbf{Y};\mathcal{U}) = 0$ (*perfect secrecy* in the Shannon's sense [140]). In any case, using (4.20) we can write

$$h(\mathcal{U}|\mathbf{Y}) = h(\mathcal{U}) - \frac{L_1}{2}\log\left[\frac{(\sigma_W^2+\sigma_X^2)\,(\sigma_W^2\sigma_X^2(1-\alpha)^2+\sigma_N^2(\sigma_W^2+\alpha^2\sigma_X^2))}{\sigma_W^2(\sigma_W^2+\sigma_X^2+\sigma_N^2)(1-\alpha)^2\sigma_X^2}\right].$$

Theoretical results are plotted in Figures 4.8, 4.9, and 4.10, showing their dependence on the DWR, the WNR and α. Since $I(\mathbf{Y};\mathcal{U})$ depends on the transmission rate and this depends in turn on the expected WNR, the WNR has been fixed in order to plot the results. Under the light of these plots, several conclusions can be drawn:

- The information leakage increases with α, because a smaller self-noise power is introduced.
- Conversely, the information leakage decreases for growing DWRs, because the uncertainty about the watermarked signal given the chosen $\mathbf{U}$ sequence is increased. Furthermore, when the DWR is increased, the number of codewords in the codebook $\mathcal{U}$ is also increased, therefore increasing the a priori uncertainty.
- The larger the expected WNR, the smaller the mutual information, because the embedder can achieve a higher reliable rate, thus increasing the uncertainty of the attacker about the sent symbol, which makes more difficult his/her job.

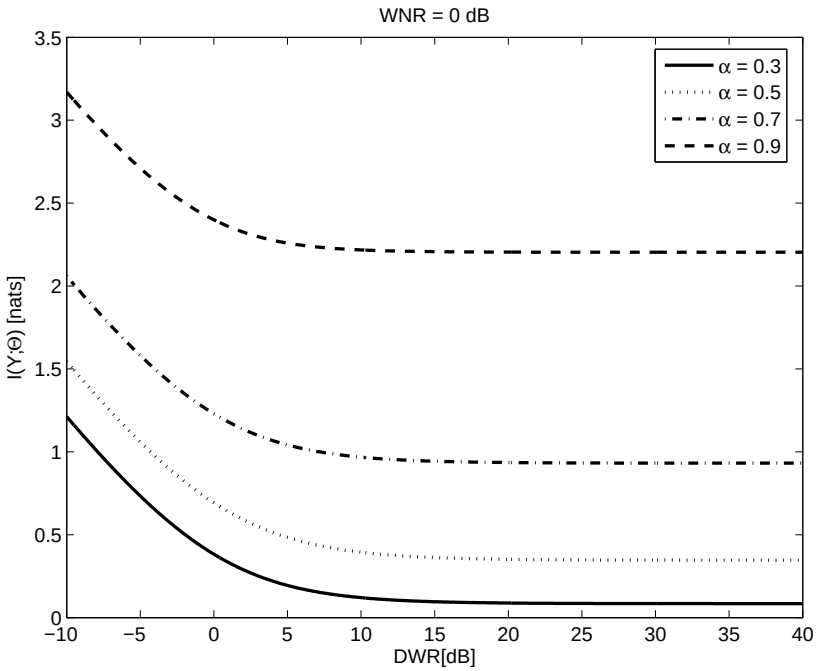


Figure 4.8: $I(\mathbf{Y};\mathcal{U})$ vs. DWR in Costa, for different values of α and WNR = 0 dB. $L_1 = 1$.

4.5.2.2. Multiple observations ($N_o \geq 1$)

In this case, the mutual information can be written as

$$\begin{aligned} I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U}) &= h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) \\ &- h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|B^1,\cdots,B^{N_o},\mathcal{U}) \\ &- I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};B^1,\cdots,B^{N_o}|\mathcal{U}), \end{aligned} \tag{4.21}$$

where $I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};B^1,\cdots,B^{N_o}|\mathcal{U}) = N_o I(\mathbf{Y};B|\mathcal{U})$ due to the independence of the observations. On the other hand, and taking into account the argument introduced in Section 4.5.1.2, we can see that

$$h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) \approx h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|B^1,\cdots,B^{N_o}), \tag{4.22}$$

since the probability of having two observations related to the same codeword $\mathbf{U}$ (this is the case where the entropy could be reduced with respect to the case of independent $\mathbf{Y}$'s) exponentially goes to 0; therefore, the fact of knowing the message each observation is related to does not to help in gaining any knowledge about Y^n.

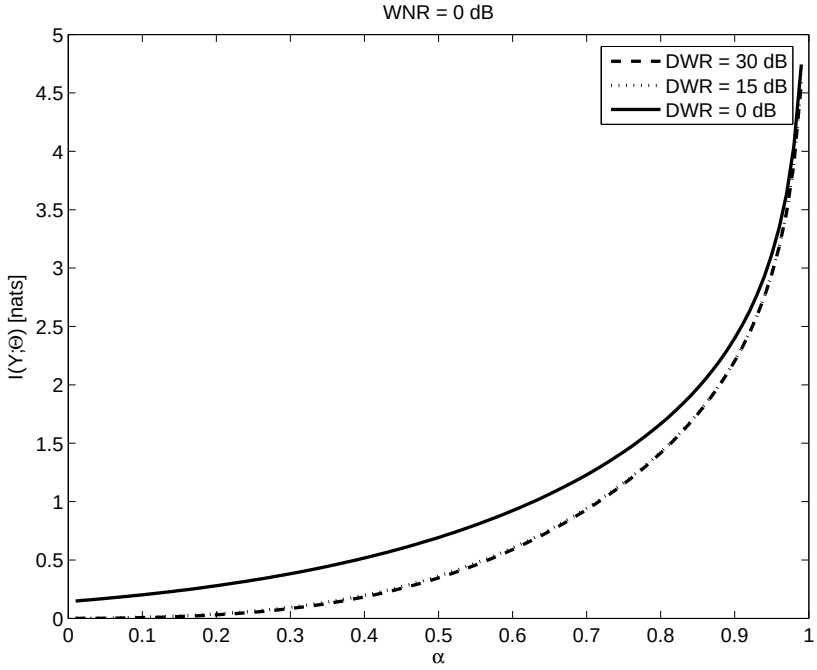


Figure 4.9: $I(\mathbf{Y};\mathcal{U})$ vs. α in Costa, for different values of DWR and WNR = 0 dB. $L_1 = 1$.

From (4.21) and (4.22), we can write

$$\begin{aligned} I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U}) &\approx h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|B^1,\cdots,B^{N_o}) \\ &- h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|B^1,\cdots,B^{N_o},\mathcal{U}) \\ &- N_o I(\mathbf{Y};B|\mathcal{U}) \\ &= I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U}|B^1,\cdots,B^{N_o}) - N_o I(\mathbf{Y};B|\mathcal{U}), \end{aligned}$$

where the leftmost term is the information leakage for the KMA case (see Section 4.5.1.2).

4.5.3. Estimated Original Attack

4.5.3.1. One available observation ($N_o = 1$)

In Appendix F.3 it is shown that if $\alpha > 0.2$, then

$$I(\mathbf{Y};\mathcal{U}|\hat{\mathbf{X}}) \approx \frac{L_1}{2}\log\left[\frac{(\sigma_W^2+\sigma_E^2)\left\{\sigma_W^2\sigma_E^2(1-\alpha)^2+\sigma_N^2(\sigma_W^2+\alpha^2\sigma_E^2)\right\}}{\sigma_W^2(\sigma_W^2+\sigma_E^2+\sigma_N^2)(1-\alpha)^2\sigma_E^2}\right], \quad (4.23)$$

so

$$h(\mathcal{U}|\mathbf{Y},\hat{\mathbf{X}}) \approx h(\mathcal{U}) - \frac{L_1}{2}\log\left[\frac{(\sigma_W^2+\sigma_E^2)\left\{\sigma_W^2\sigma_E^2(1-\alpha)^2+\sigma_N^2(\sigma_W^2+\alpha^2\sigma_E^2)\right\}}{\sigma_W^2(\sigma_W^2+\sigma_E^2+\sigma_N^2)(1-\alpha)^2\sigma_E^2}\right].$$

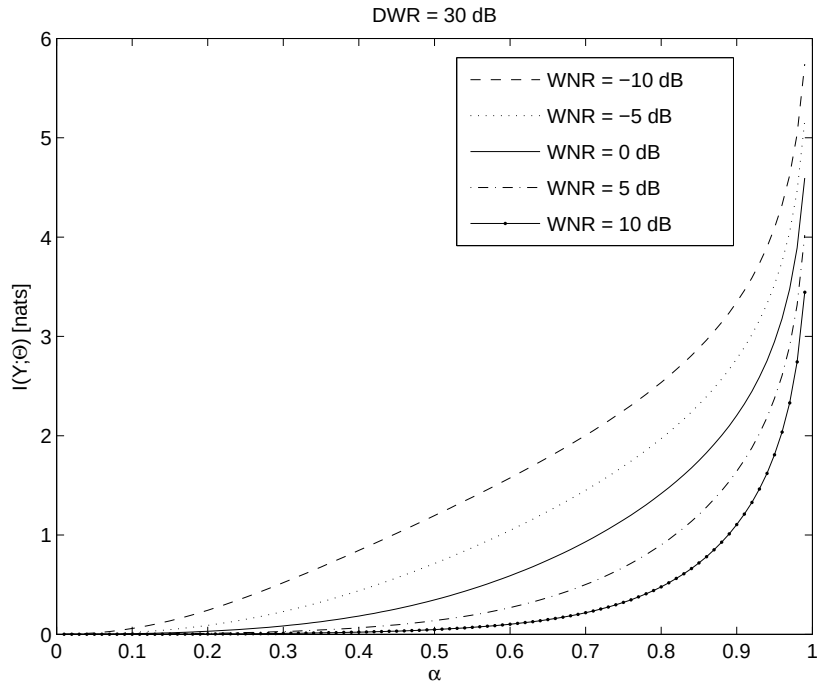


Figure 4.10: $I(\mathbf{Y};\mathcal{U})$ vs. α in Costa, for different values of WNR, setting DWR = 30 dB. $L_1 = 1$.

Therefore, when the attacker has perfect knowledge of the original host signal, $\sigma_E^2 = 0$ so $I(\mathbf{Y};\mathcal{U}|\hat{\mathbf{X}}) = \infty$. The minimum value of the mutual information corresponds to $\alpha = 0$.

It can be seen that (4.23) is equivalent to (4.20) but replacing σ_X^2 by σ_E^2. For that reason, Figures 4.8, 4.9 and 4.10 are still valid, but replacing the DWR by the EWR. In fact, when an estimate of the original host signal is available, the actual host signal can be thought of as being on a sphere centered at that estimate and with squared radius equal to the variance of the estimation error. Since such a shift should not modify the results, this problem must be equivalent to having the host on a sphere with squared radius equal to the variance of the estimation error, but centered at the origin (so σ_X^2 should be replaced by σ_E^2). Nevertheless, note that the codebook is not designed for this scenario, but for the original one (where the host signal is completely unknown), so the analysis performed in Appendix F.3 is still pertinent.

4.5.3.2. Multiple observations ($N_o \geq 1$)

In this case, the mutual information can be written as

$$\begin{aligned} I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U},\hat{\mathbf{X}}_1,\cdots,\hat{\mathbf{X}}_{N_o}) &= h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\hat{\mathbf{X}}_1,\cdots,\hat{\mathbf{X}}_{N_o}) \\ &- h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|B^1,\cdots,B^{N_o},\mathcal{U},\hat{\mathbf{X}}_1,\cdots,\hat{\mathbf{X}}_{N_o}) \\ &- I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};B^1,\cdots,B^{N_o}|\mathcal{U},\hat{\mathbf{X}}_1,\cdots,\hat{\mathbf{X}}_{N_o}), \end{aligned} \tag{4.24}$$

where

$$I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};B^1,\cdots,B^{N_o}|\mathcal{U},\hat{\mathbf{X}}_1,\cdots,\hat{\mathbf{X}}_{N_o}) \approx N_o I(\mathbf{Y};B|\mathcal{U},\hat{\mathbf{X}}),$$

due to the independence of the observations, and the righmost term was studied in Appendix F.3. Recalling again the argument introduced in Section 4.5.1.2, where now σ_X^2 must be replaced by σ_E^2, one can see that the probability of having two observations related to the same codeword can be upper bounded by $P_{u,\text{EOA}} \triangleq [1-(1-p)^{N_o}]N_o$, in such a way that when L_1 goes to infinity and $N_o \leq (1/p)^{L_1/8}$

$$\lim_{L_1\to\infty} P_{u,\text{EOA}} = p^{L_1/4},$$

with

$$p \triangleq \frac{1}{|\mathcal{U}_{\hat{\mathbf{X}}}|} = \left(\frac{\sigma_W^2\sigma_E^2(1-\alpha)^2+\sigma_N^2(\sigma_W^2+\alpha^2\sigma_E^2)}{(\sigma_W^2+\sigma_E^2+\sigma_N^2)(\sigma_W^2+\alpha^2\sigma_E^2)}\right)^{L_1/2},$$

where we have denoted by $|\mathcal{U}_{\hat{\mathbf{X}}}|$ the cardinality of the set of codewords of the codebook $\mathcal{U}$ which are in a hypersphere of radius $L_1\sigma_E^2$. Be aware that this hypershpere will be centered at different points $\hat{\mathbf{X}}^n$ for different observations, so the probability of having two observations related to the same codeword is even smaller than the estimated one. We would also like to remark that the bound on N_o is necessary to ensure the exponential decrease of $P_{u,\text{EOA}}$, and this decrease itself has a different argument p for this case and the KMA one studied in Section 4.5.3.2.

Taking these considerations into account, one can realize that when the number of observations is exponentially bounded, the mutual information for the EOA is that obtained for the WOA, but replacing σ_X^2 by σ_E^2.

4.5.4. Constant Message Attack

4.5.4.1. One available observation ($N_o = 1$)

Since in this case the attacker does not know which was the embedded symbol, this scenario is equivalent to the WOA; be aware that this equivalence does not

hold for the case of multiple observations ($N_o \geq 2$), because in the CMA scenario the attacker will take advantage of the additional information provided by the fact that all the observations are related to codewords belonging to the same bin, as it will be shown next.

4.5.4.2. Multiple observations ($N_o \geq 1$)

Following an approach similar to that of Section 4.5.1.2, the residual entropy when the attacker knows the index of the observed codeword can be written as

$$\begin{aligned} h(\mathcal{U}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\mathrm{CM},\mathcal{J}) &= h(\mathcal{U}|\mathrm{CM},\mathcal{J}) - I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U}|\mathrm{CM},\mathcal{J}) \\ &\leq h(\mathcal{U}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\mathrm{CM}), \end{aligned}$$

and given that the constant message and the index of the observed codeword are independent of the codebook, we have that $h(\mathcal{U}|\mathrm{CM},\mathcal{J}) = h(\mathcal{U})$. On the other hand,

$$\begin{aligned} I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathcal{U}|\mathrm{CM},\mathcal{J}) &= h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathrm{CM},\mathcal{J}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathcal{U},\mathrm{CM},\mathcal{J}) \\ &= \sum_{i=1}^{N_o} h(\mathbf{Y}^i|\mathrm{CM},\mathcal{J},\mathbf{Y}^1,\cdots,\mathbf{Y}^{i-1}) \\ &\quad - h(\mathbf{Y}^i|\mathcal{U},\mathrm{CM},\mathcal{J},\mathbf{Y}^1,\cdots,\mathbf{Y}^{i-1}), \end{aligned}$$

but the i-th observation is independent of the previous observations which are not related to the same codeword, so for a given constant message b and index $\mathcal{J}$ the previous equation yields

$$\sum_{i=1}^{|\mathcal{U}_b|}\sum_{k=1}^{L_{i,b}} h(\mathbf{Y}_{b,i,k}|\mathbf{Y}_{b,i}^{k-1}) - h(\mathbf{Y}_{b,i,k}|\mathcal{U},\mathrm{CM},\mathbf{Y}_{b,i}^{k-1},\mathbf{Y}_{b,i-1}^{L_{i-1,b}},\cdots,\mathbf{Y}_{b,1}^{L_{1,b}}), \quad (4.25)$$

where we have followed the notation introduced in Section 4.5.1.2. The leftmost term was already computed there. Assuming that α is chosen in such a way that perfect decoding is possible, the righmost one can be seen to be

$$h(\mathbf{Y}_{b,i,k}|\mathcal{U},\mathrm{CM},\mathbf{Y}_{b,i}^{k-1},\mathbf{Y}_{b,i-1}^{L_{i-1,b}},\cdots,\mathbf{Y}_{b,1}^{L_{1,b}}) = \begin{cases} \log(|\mathcal{U}|) + h(\mathbf{Y}|\mathbf{U}), & \text{if } i=1, k=1 \\ \log(|\mathcal{U}_b| - i + 1) + h(\mathbf{Y}|\mathbf{U}), & \text{if } i>1, k=1 \\ h(\mathbf{Y}|\mathbf{U}), & \text{if } k>1 \end{cases}.$$

Similarly to Section 4.5.1.2, one must average (4.25) over all possible realizations of B and $\mathcal{J}$, so, following the same reasoning, we can see that the probability of $L_{i,b} > 1$ for a fixed b and any $1 \leq i \leq |\mathcal{U}_b|$ can be upperbounded by

$$P\big(i \in \{1,\cdots,|\mathcal{U}_b|\} : L_{i,b} \geq 2\big) \leq [1-(1-p)^{N_o}]N_o \triangleq P_{u,\mathrm{CMA}},$$

where

$$p \triangleq \frac{1}{|\mathcal{U}_b|} = \left(\frac{\sigma_W^2}{\sigma_W^2 + \alpha^2\sigma_X^2}\right)^{L_1/2}.$$

Thus, recalling that if N_o is such that $N_o \leq (1/p)^{L_1/8}$, then $\lim_{L_1\to\infty} P_{u,\text{CMA}} = p^{L_1/4}$, so the probability of having 2 or more observations related to the same codeword decreases exponentially. Therefore, for large values of L_1, if $|\mathcal{U}_b| >> N_o$, which is reasonable since $|\mathcal{U}_b|$ increases also exponentially with L_1, then $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|\text{CM})$ can be accurately approximated as

$$\begin{aligned} I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|\text{CM}) &\approx I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|\text{CM}, \mathcal{J}) \\ &\approx N_o\big[h(\mathbf{Y}|\mathbf{U}) + h(\mathbf{U}) - h(\mathbf{U}|\mathbf{Y}) \\ &\quad - \log(|\mathcal{U}_b|) - h(\mathbf{Y}|\mathbf{U})\big] - \log(|\mathcal{U}|) + \log(|\mathcal{U}_b|) \\ &= I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o}) - I(\mathbf{U}; \mathbf{Z}) + I(\mathbf{U}; \mathbf{X}) \\ &= I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathcal{U}|B^1, \cdots, B^{N_o}) - I(\mathbf{Y}; B|\mathcal{U}), \end{aligned}$$

4.6. DC-DM security and comparison

In this section, we will briefly review the results obtained by Pérez-Freire et al. for DC-DM schemes (see [127] and [128]) following the same information theoretic approach that was used in this thesis to analyze Add-SS and Costa's scheme. In order to get the general picture, the results obtained for all three schemes will be compared, achieving interesting conclusions.

Before starting to enumerate the results about DC-DM security, we would like to remark that, contrarily to the random nature of Costa's codebook, DC-DM is based on a highly structured codebook. In fact, this is one of its advantages, since this structure dramatically reduces the complexity of embedding and decoding, allowing practical implementations, that are not possible at all for Costa's scheme, given that it requires an exhaustive search over the set of possible codewords. Nevertheless, this structure of DC-DM also reduces the number of parameters that uniquely characterizes the codebook, therefore reducing the a priori uncertainty about it.[4] But the differences from a security point of view are not just constrained to this reduction in the a priori uncertainty, but also affect the information leakage (i.e. mutual information), as it will be shown in the following by comparing the results obtained by Pérez-Freire for DC-DM with those exposed above in this work for Add-SS and Costa's scheme.

[4]The a priori uncertainty about the parameters of the system could be increased by augmenting the dimensionality of the space of those parameters. A way of doing so, is the rotation of the lattice proposed by Goteti and Moulin in [81] and [117].

In [127] and [128] Pérez-Freire analyzed the security of lattice-based DC-DM; in his approach, the security of the system relies only on the secret dither vector $\mathbf{D}$, which is just known by embedder and decoder. Taking into account some of the results of his analyses, the following comparison and conclusions can be driven for the different attacks:

- Known Message Attack (KMA):

  For one observation, i.e. $N_o = 1$, the mutual information for DC-DM, when $\sigma_X^2 >> \Delta$, i.e. under the flat-host assumption, is given by [128]

  $$I(\mathbf{Y};\mathbf{D}|\mathbf{B}) = -L_1 \log(1-\alpha).$$

  Unfortunately, a similar closed formula is not available in general when $N_o > 1$; nevertheless, for $\Lambda = \mathbb{Z}^{L_1}$ and $\alpha \geq 0.5$, Pérez-Freire showed (see [127]) that

  $$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o};\mathbf{D}|\mathbf{B}^1, \cdots, \mathbf{B}^{N_o}) = L_1 \left(-\log(1-\alpha) + \sum_{i=2}^{N_o} \frac{1}{i} \right).$$

  Be aware that in this case the mutual information is clearly non-linear and concave with the number of observations N_o; this is due to the fact that in DC-DM the attacker is always estimating *the same codeword*, meaning the same dither vector. On the other hand, in Costa's scheme each observation is related to a different codeword, so the previously learned information will not provide any information about the codeword to be estimated with the present observation; this implies the observed linear growth of the mutual information. A behavior somewhat similar to that of DC-DM can be also observed in Add-SS, where the same set of spreading vectors is estimated for all the observations, yielding a clearly concave non-linear (logarithmic) growth rate.

  Figure 4.11 shows the information leakage of Add-SS, Costa's scheme and DC-DM when only one observation and unidimensional hosts are considered (i.e., $N_o = 1$ and $L_1 = 1$). In Figure 4.12, these results are plotted again, but using linear ordinate axis; this allows better comparison between the results of Costa's scheme and DC-DM. It is specially remarkable the large resemblance between the theoretical results for Costa and the results numerically obtained by Pérez-Freire for DC-DM. Furthermore, it is interesting to note the similarity of both Costa and DC-DM with Add-SS at the range of small DWRs and small values of the distortion compensation parameter α. As it was already discussed above, one can see the dependence of the mutual information with α (the largest α, the largest the mutual information), and with the DWR (the largest the DWR, the smallest the mutual information). Finally, Figure 4.13 shows the residual entropy of the three methods when $N_o = 1$ and $L_1 = 1$. Probably the most interesting

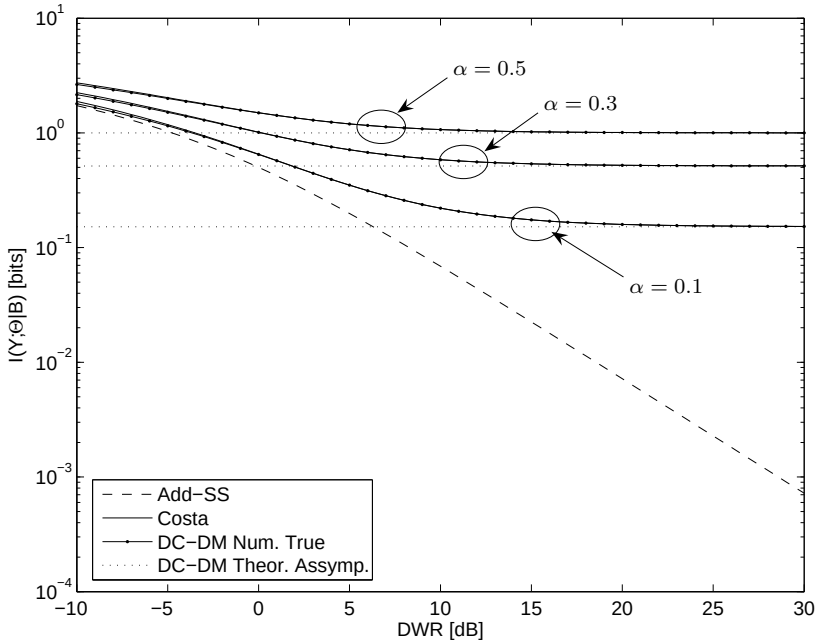


Figure 4.11: Comparison of information leakage for Add-SS, Costa's scheme and DC-DM, with $L_1 = 1$ and $N_o = 1$, for the KMA case.

result of this plot, is the quick increase of the residual entropy with DWR for Costa. The explanation is that the a priori uncertainty is also increased with the DWR; the number of codewords needed by Costa's scheme is increased with the DWR, and the larger the number of codewords, the larger the a priori uncertainty about the codebook will be. Furthermore, if $N_o \geq 1$ were considered, we could observed how the number of codewords needed for Costa (and therefore the a priori uncertainty) really explodes with the number of observations, specially for large DWRs.

- Watermarked Only Attack (WOA):

  For DC-DM with $\Lambda = \mathbb{Z}^{L_1}$, $P = 2$, and $\alpha \geq 0.5$ the mutual information when one observation is available can be written as [127]

  $$I(\mathbf{Y}; \mathbf{D}) = -L_1 \log(2(1-\alpha)),$$

  and when more observations are at hand, if $\alpha \geq 0.75$, then

  $$\begin{aligned} I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{D}) &= L_1 \left(-\log(1-\alpha) - \log(2) + \sum_{i=2}^{N_o} \frac{1}{i} \right) \\ &= I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{D} | \mathbf{B}^1, \cdots, \mathbf{B}^{N_o}) - L_1 \log(2). \end{aligned}$$

  Be aware that when $\alpha = 0.5$, then $I(\mathbf{Y}; \mathbf{D}) = 0$, meaning that no knowledge about the dither can be learnt from the observation of watermarked contents.

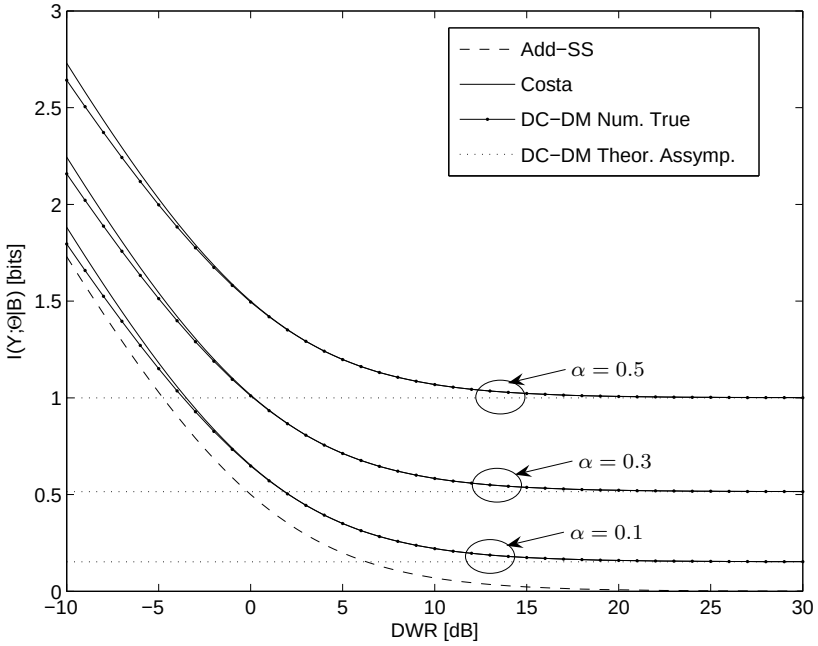


Figure 4.12: Comparison of information leakage for Costa's scheme and DC-DM, with $L_1 = 1$ and $N_o = 1$, for the KMA case.

Furthermore, the differences with Costa's scheme results are also clear. On one hand, the result obtained for Costa differs with respect to that obtained in the KMA scenario in the total rate that could be transmitted using N_o usages of the channel. On the other hand, the difference between the mutual information for the KMA and WOA scenarios for binary DC-DM is just the achievable rate in *one* usage of the channel, i.e. L_1 bits. The explanation to this phenomenon is straightforward. When an attacker gets the first observation in DC-DM, he/she has uncertainty about the value of the bit it is related to. Due to this uncertainty, there are two possible values of $D_i, 1 \leq i \leq L_1$ for each dimension; therefore, there are 2^{L_1} possible choices, i.e. L_1 information bits. This uncertainty is not increased with the subsequent observations, since they must be coherent with the first one. On the other hand, when an attacker gets the first observation in Costa's scheme, he/she has an equivalent uncertainty about the value of the message such that observation is related to. The difference is that this uncertainty is increased when more observations are available; even if the attacker were told the message the first observation is related to, he/she would have the same uncertainty about the message to which the second observation is related. This would not be obviously the case for binary DC-DM, with $\Lambda = \mathbb{Z}^{L_1}$ and $\alpha \geq 0.75$, where such information would uniquely determine the message the subsequent observations are related to. This explains why the difference between the KMA and WOA cases is linear with N_o for Costa, but constant for DC-DM.

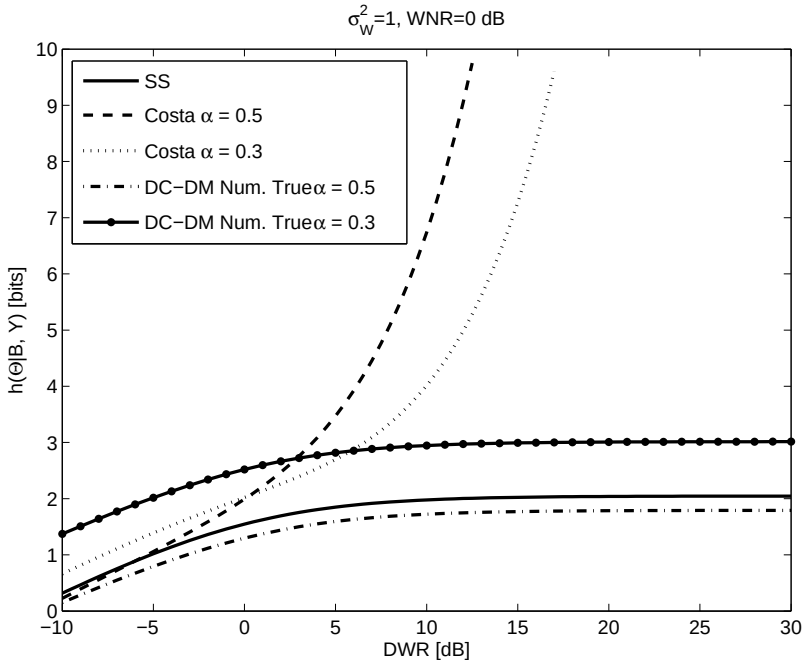


Figure 4.13: Comparison of residual entropy for different data hiding schemes, with $L_1 = 1$ and $N_o = 1$, for the KMA case.

- Estimated Original Attack (EOA):

  The Estimated Original Attack does not seem to make sense for DC-DM data hiding methods. Even when some approximations were computed in [63], they are just valid when the estimation error is uniformly distributed in a small fraction of the quantization step. Given that this assumption will be rarely verified in practical scenarios, since it implies a really accurate estimate of the host, the corresponding comparison will not be made here.

- Constant Message Attack (CMA):

  As it was previously discussed, CMA with $N_o = 1$ is completely equivalent to WOA with $N_o = 1$, so we will focus on the case of $N_o \geq 1$. When $N_o \geq 1$, the mutual information for DC-DM can be bounded as [128]

$$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{D}|\text{CM}) \geq I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{D}|\mathbf{B}^1, \cdots, \mathbf{B}^{N_o}) - \log(P^{L_b}),$$

  where the rightmost term is the mutual information obtained for the KMA case, minus the logarithm of the messages that can be transmitted in one usage of the channel, i.e. P^{L_b}. Thus, the relation between this result and that obtained for Costa is evident.

4.7. Conclusions

In this chapter we have focused on watermarking security. First of all, a brief historical introduction showed the evolution of this concept on watermarking research community. Taking into account this discussion, a new watermarking security definition was proposed, and related to this, an information theoretic measure was used for the first time to analyze the security of Add-SS and Costa's scheme (based on random codebooks). Both schemes' security is compared with that of DC-DM obtained by Pérez-Freire. At the sight of these results, we can say that the host interference can be positive from a security point of view; effectively, the host makes more difficult the estimation of the system parameters (spreading sequence, codebook or dither vector, depending on the algorithm), in such a way that the information leakage for a given DWR is minimized for Add-SS; for the side-informed methods (both Costa and DC-DM) one can observe that the smaller the distortion compensation parameter α (and therefore the greater the host intereference due to the self-noise), the smaller the information leakage. Furthermore, for all the three methods, the information leakage is reduced when the DWR is increased. It is also noticeable that the information leakage is reduced when the rate of the watermarking system is increased; this has an easy intuitive idea: the higher the rate, the more uncertainty the attacker will have about the sent symbol, making more difficult the estimate of the system parameters.

Nevertheless, one should also consider that the watermarking security is not just a function of the information leakage, but the a priori uncertainty about the system parameters has also to be taken into account; in this sense, the attacker can take advantage of the highly structured nature of practical watermarking codes, as those used by Add-SS or lattice-based DC-DM, since in that case the code is usually uniquely characterized by a reduced (in fact just linear with the number of dimensions) number of parameters , in such a way that the search space, and therefore the associated a priori uncertainty, is also reduced. This is not the case of Costa's scheme, where the codewords, whose number is increased exponentially with the number of dimensions, are independently randomly generated, achieving a huge a priori uncertainty. Unfortunately, it is well-known that Costa's scheme is not practical, since it requires exhaustive search over this huge set of possible codewords.

Chapter 5

Dirty Paper Codes: when channel-coding meets source-coding

Structured codes are known to be necessary in practical implementations of capacity-approaching "dirty paper schemes". In this chapter we study the performance of a recently proposed dirty paper technique, by Erez and ten Brink which is firstly applied to data-hiding, and compare it with other existing approaches. Specifically, we compare this technique with conventional side-informed schemes previously used in data-hiding based on repetition and turbo coding. We show that a significant improvement can be achieved using Erez and ten Brink's proposal. We also discuss the adaptation of these codes to data hiding, mainly related with perceptual questions.

5.1. Introduction

In the last years the usefulness of approaching watermarking as a communication problem with side information known at the encoder but not at the decoder has been proven. This model with i.i.d. Gaussian random variables was shown by Costa [50] to achieve the same capacity as if the side information were also made available to the decoder. Nevertheless, the main problem with Costa's construction is that it relies on random codes, which require an exhaustive search strategy for selecting the codeword to be used, something that is largely impractical. Due to the importance of Costa's result, not only to watermarking, but also to many other applications in communications, a large number of papers dealing with the possibility of approaching the same result using structured codes have been written [32, 70].

Erez and Zamir have recently shown [70] that Costa's result can be achieved with nested lattices. In fact, they have proven a stronger result in which a modulo-lattice transformation of the received signal is considered at the decoder; this result obviously links with the lattice-based DC-DM using *lattice decoding* already introduced in this thesis in Section 2.5. As is explained there, lattice decoding allows a huge reduction in complexity, as well as the possibility of achieving capacity without explicitly knowing the pdf of the host signal. Nevertheless, the question of code construction is not completely solved: Erez and Zamir's result applies to lattices verifying quite strict conditions which require that the fundamental regions approach hyperspheres asymptotically as the number of dimensions is increased. Unfortunately, those conditions fall short of being met by the simplest (and mostly used) lattices, such as the cubic ones. Therefore, practical solutions demand the use of strategies whose complexity does not rely exclusively on these simple lattices.

The usually followed solution is to encode the information bits with a near-Shannon-limit channel code and then take the output bits to index the sub-lattice used to quantize the host signal (i.e., DC-DM with channel coding). Due to the redundancy introduced by the channel code, this lattice can be a very simple one, even allowing for scalar quantization. The good results obtained with this kind of schemes can be explained from the fact that the channel code concatenated with a simple lattice is equivalent to a better (and also more involved) lattice.

Summarizing, most of the practical schemes that use structured (in the sense of lattice-based) codes to approach Costa's result are composed of a good channel code concatenated with a quite simple lattice. The encoding and decoding with the channel code is usually relatively easy, and the same applies when a simple lattice is chosen, in such a way that the resulting dirty paper coding schemes fall quite close to Shannon's limit, while keeping a reasonable computational cost.

Nevertheless, as it was already introduced in Section 3.6.3, if a truly capacity approaching system is to be designed, the gap to capacity due to the shaping gain must be reduced. In order to measure this gap, in Figure 5.1 we plot the achievable rate of a system based on scalar lattices (for both uniform and binary input distributions) versus the ratio between the energy per bit and the power spectral density of the noise. The energy per bit will be denoted by E_b, and it is computed as $E_b = \sigma_W^2/R$, with R rate of the system. On the other hand, the power spectral density of the noise is usually denoted by $N_0/2$, and it is computed as $N_0 = 2\sigma_N^2$. Therefore, the aforementioned ratio E_b/N_0 can be expressed as

$$\frac{E_b}{N_0} = \frac{\sigma_W^2}{2R\sigma_N^2} = \mathrm{WNR}\frac{1}{2R}. \tag{5.1}$$

The resulting plot is typically used in the literature to measure the gap to capacity, since it normalizes the SNR by the achieved rate, so it is a good measure of the efficency of the system. Retaking Figure 5.1, we can see again, as shown in

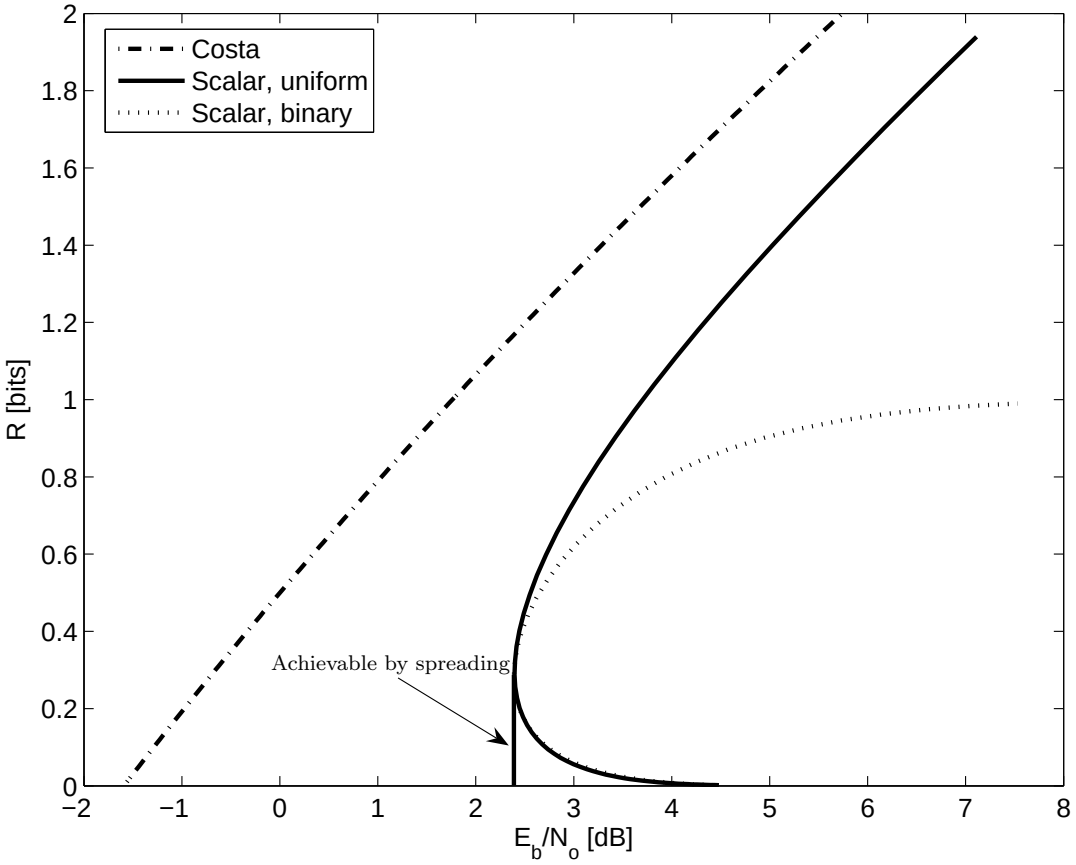


Figure 5.1: Achievable rate vs. E_b/N_0 for scalar dirty paper codes, with binary and uniform input. Costa's result is also plotted for comparison purposes.

Section 3.6.3, that the achievable rate for the case of binary input is upper-bounded by 1 bit per use of the channel, whereas the gap to capacity for the case of uniformly distributed input and scalar quantizers goes to 1.53 dB for large rates. Nevertheless, the main problem of a scheme based on scalar quantizers lies in the range of small rates, which in fact are the most used in data-hiding applications. There, the gap to capacity is unbounded; this explains why it would be more advantageous to use a spreading sequence in order to follow a STDM-like strategy. Those strategies are represented in Figure 5.1 by a vertical line from the point corresponding to the base channel code, meaning that all the rates lower than that of the used code are achievable for the same E_b/N_0; this is based on the fact that spreading does not modify the E_b/N_0, since the WNR needed to achieve the same performance as the no-spreading case is reduced by the same amount as the spreading rate. Taking this into account, if scalar quantizers without shaping are to be used, a smart strategy seems to be using a code of rate about 1/3, and then spreading in order to improve robustness. Furthermore, binary signalling yields almost the same results as a uniformly distributed input. This kind of

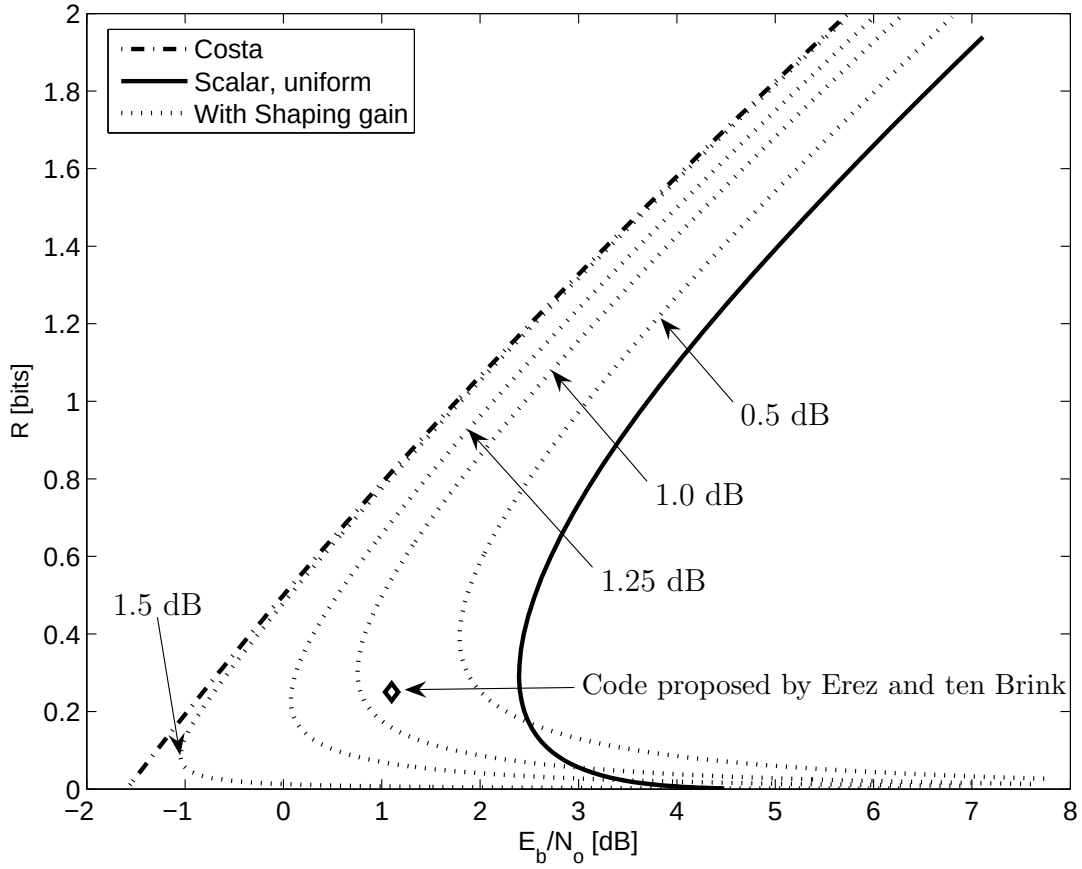


Figure 5.2: Achievable rate vs. E_b/N_0 for scalar dirty paper codes and uniform input, compared with the lower bounds obtained for different shaping gain values. Concretely, the values of shaping gain considered were 0.5 dB, 1.0 dB, 1.25 dB and 1.5 dB. Furthermore, we have plotted the point corresponding to the code proposed in [69], i.e. $(1.1, 0.25)$, and that will the used in Section 5.4.

strategy is sometimes called *time-sharing* in the literature [69].

Finally, following an approach similar to that in [69], where the achievable rate is lower-bounded by a function of the shaping gain as

$$I(\mathbf{B}; (\mathbf{Z} \mod \Lambda)) \geq \frac{1}{2}\left[\log_2(1+\mathrm{WNR}) - \log_2\left(\frac{2\pi e}{12}\cdot 10^{-g_s(\Lambda)/10}\right)\right], \quad (5.2)$$

we have plotted in Figure 5.2 these bounds to the achievable rate against the E_b/N_0 for different values of the shaping gain. Furthermore, we have plotted the point corresponding to the code proposed in [69], which will be used in Section 5.4, and that achieves error-free decoding for rate 1/4, and $E_b/N_0 = 1.1$ dB (WNR $= -1.9$ dB), using a 4-QAM constellation combined with trellis-shaping for reducing the gap to capacity. Summarizing, if a dirty paper code really ap-

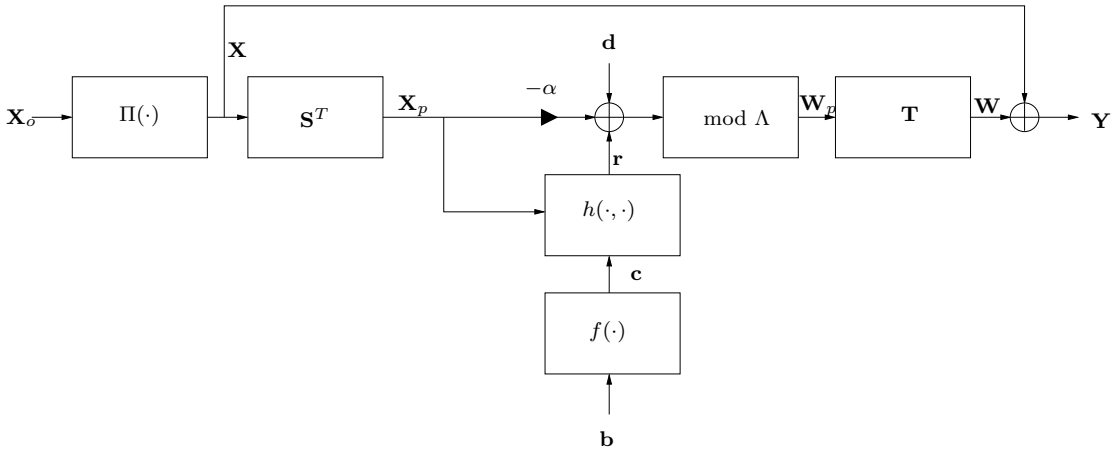


Figure 5.3: General structure of a dirty-paper encoder.

proaching capacity is to be designed, not just channel coding but also source coding must be employed.

Firstly, some modifications to the notation used throughout this thesis will be introduced. Afterwards, classical approaches which only use channel coding will be reviewed. Finally, the method proposed by Erez and ten Brink [68, 69] that is based on the combination of source coding and channel coding will be introduced, and its gain over the previous works explained. A fundamental part of our work consists in the experimental results; they will show the goodness of the studied method by comparing its performance on a real framework with those side-informed methods just using channel coding.

5.2. Notation and Unified Framework

The general diagram of the dirty-paper coding schemes is plotted in Figure 5.3. As it was explained in Section 2.1, we will assume that the host signal is modeled by a zero-mean random vector $\mathbf{X}^o = (X_1^o, \cdots, X_{L_1}^o)^T$, and prior to embedding we apply a key-dependent pseudorandom permutation $\Pi(\cdot)$ to $\mathbf{X}^o$. The permuted host, denoted by $\mathbf{X} \triangleq \Pi(\mathbf{X}^o)$, could be projected onto a L_3-dimensional space ($L_3 \leq L_1$), see Section 2.7; this yields $\mathbf{X}_p = \mathbf{S}^T \cdot \mathbf{X}$, where $\mathbf{S}$ is a $L_1 \times L_3$ matrix. Be aware that the concatenation of the permutation matrix and $\mathbf{S}^T$ can be seen as a simple projection matrix; nevertheless, we have preferred to separately define both the permutation and the projection matrix $\mathbf{S}$ in order to establish a special structure on $\mathbf{S}$ (diagonal or block-diagonal) without losing generality.

Since channel coding is considered, the length L_b infomation message $\mathbf{b}$ could go through a channel encoder $f(\cdot)$, so $\mathbf{c} = f(\mathbf{b})$ is the length L_c channel-coded

message. For the sake of simplicity, we will assume that both $\mathbf{b}$ and $\mathbf{c}$ are binary vectors ($P = 2$).

The source coding part will be performed by a vector quantizer, whose code will be denoted by $h(\cdot, \cdot)$; this will transform the length L_c channel-coded binary message $\mathbf{c}$ into a length L_3 vector $\mathbf{r}$, with elements in the alphabet $\mathcal{W}$. The vector $\mathbf{r}$ will depend on both $\mathbf{c}$ and $\mathbf{X}_p$. Therefore, $h(\cdot, \cdot)$ will only make sense when the vector quantizer is really used; for example, in Section 5.3, where Cartesian products of scalar vectors are used, its output $\mathbf{r}$ will be just a mapping from $\mathbf{c}$.

Let

$$\Lambda \triangleq |\mathcal{W}|\mathbb{Z}^{L_3}, \tag{5.3}$$

then, given $\mathbf{r}$, a shifted-lattice quantizer, $\mathbf{Q}_{\mathbf{r}}(\cdot)$, based on a minimum Euclidean distance criterion is defined as

$$\mathbf{Q}_{\mathbf{r}}(\mathbf{a}) = \mathbf{Q}_{\Lambda}\left(\mathbf{a} - \mathbf{v}(\mathbf{r})\right) + \mathbf{v}(\mathbf{r}), \quad \text{for any } \mathbf{a} \in \mathbb{R}^{L_3} \tag{5.4}$$

where $\mathbf{Q}_{\Lambda}(\cdot)$ is the minimum Euclidean distance quantizer induced by the lattice Λ, and $\mathbf{v}(\mathbf{r}) = \mathbf{r} + \mathbf{d}$. Vector $\mathbf{d}$ is a realization of a key-dependent pseudorandom dither vector $\mathbf{D}$, which is uniformly distributed over the Voronoi region of Λ, so in the j-th component $D_j \sim U(-|\mathcal{W}|/2, |\mathcal{W}|/2], 1 \leq j \leq L_3$.

The watermark in the projected domain $\mathbf{W}_p$ becomes

$$\mathbf{W}_p \triangleq \mathbf{Q}_{\mathbf{r}}(\alpha \mathbf{X}_p) - \alpha \mathbf{X}_p, \tag{5.5}$$

which is nothing but the quantization error resulting when quantizing $\alpha\mathbf{X}_p$ with the quantizer $\mathbf{Q}_{\mathbf{r}}(\cdot)$ corresponding to the message $\mathbf{r}$. Considering the structure of the lattice defined in (5.3), it is clear that the quantization in (5.5) can be implemented in a sample-by-sample basis. The distortion-compensation parameter α, $0 < \alpha \leq 1$, is an optimizable variable akin to the one in Costa's paper. We would like to recall that the watermarking scheme resulting from (5.5) is exactly that introduced in Section 2.7, where the power of the watermark is kept constant independently of α by inflating the lattice with a factor $1/\alpha$ (see Appendix A for the proof).

The inverse projection will be given by the $L_1 \times L_3$-matrix $\mathbf{T}$, so

$$\mathbf{W} = \mathbf{T}\mathbf{W}_p, \tag{5.6}$$

where $\mathbf{T}$ could be any matrix verifying

$$\mathbf{S}^T \cdot \mathbf{T} = \mathbf{I}_{L_3 \times L_3}, \tag{5.7}$$

although the matrix simultaneously verifying the last equality and minimizing the norm of W is the pseudoinverse $\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}$. Considering the last formula, (5.6)

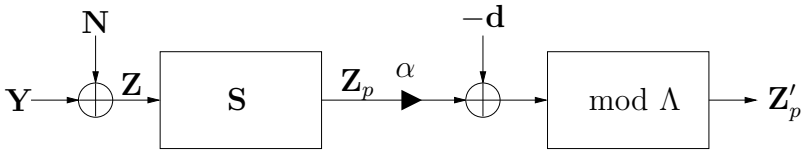


Figure 5.4: Scheme of the channel and predecoder.

and (5.3), one realizes that the joint design of $\mathbf{S}$ and $\mathbf{T}$ needs to take into account the human perceptual feature, in order not to produce perceptually noticeably watermarks. In this sense, we will see that both of them will depend on the perceptual mask $\boldsymbol{\gamma}$, introduced in Section 2.1.

On the other hand, the received signal $\mathbf{Z}$ will be projected using $\mathbf{S}$ to obtain $\mathbf{Z}_p = \mathbf{S}^T \cdot \mathbf{Z}$, the projected signal which will be used in the decoding process. In a similar way, $\mathbf{N}_p = \mathbf{S}^T \cdot \mathbf{N}$ denotes the projected noise. Both the channel and predecoder are plotted in Figure 5.4.

5.3. Classical approaches

Once the general framework for decoding has been introduced, we will show how classical approaches fit in this framework. These methods are typically based on the use of scalar quantizers instead of a vector one, and the differences among them are given by their specific values of $\mathbf{S}$, $\mathbf{T}$, $f(\cdot)$ and $h(\cdot,\cdot)$.

5.3.1. Repetition coding with no projection

In this case, the following identities apply:

$$\begin{aligned}
&\mathbf{S} = \mathrm{diag}(1/\gamma_1, \cdots, 1/\gamma_{L_1}),\\
&\mathbf{T} = \mathrm{diag}(\gamma_1, \cdots, \gamma_{L_1}),\\
&c_j = b_i, \quad (i-1)L_1/L_b < j \leq iL_1/L_b, \text{ and } 1 \leq i \leq L_b,\\
&r_j = c_j, \quad 1 \leq j \leq L_1,
\end{aligned} \tag{5.8}$$

in such a way that any bit b_j is repeated L_1/L_b times[1], so $L_c = L_1 = L_3$. Note also that given (5.8), $\mathbf{r}$ will not depend on $\mathbf{X}_p$ but only on $\mathbf{c}$, since a scalar quantizer is being used.

Summarizing, the initial values of $\mathbf{X}$ are normalized by the corresponding value of the perceptual mask $\boldsymbol{\gamma}$ in order to take into account the perceptual constraints in the embedding, and the input bits are repeated L_1/L_b times.

[1]We will assume that L_1/L_b is an integer.

5.3.2. Repetition coding with projection

For this method, we have:

$$
\begin{aligned}
s_{ij} &= \begin{cases} \frac{L_b}{L_1}\frac{q_j}{\gamma_j}, & \text{if } (i-1)L_1/L_b < j \leq iL_1/L_b, \text{ and } 1 \leq i \leq L_b \\ 0, & \text{otherwise} \end{cases}, \\
t_{ij} &= \begin{cases} q_i\gamma_i, & \text{if } (j-1)L_1/L_b < i \leq jL_1/L_b, \text{ and } 1 \leq j \leq L_b \\ 0, & \text{otherwise} \end{cases}, \\
c_j &= b_j, \quad 1 \leq j \leq L_1, \\
r_j &= c_j, \quad 1 \leq j \leq L_1,
\end{aligned} \tag{5.9}
$$

with $q_j \in \{-1, +1\}$ a pseudorandomly generated spreading sequence, known to both encoder and decoder. Note that $\mathbf{q}$ could follow any other zero-mean unit-variance distribution (e.g., a Gaussian). The definition of $\mathbf{T}$ is also based on perceptual constraints. The fact of not having a vector quantizer, but a scalar one is again reflected in (5.9).

Be aware that both of these methods could be seen as extreme cases of a general one, where the repetition rate L_1/L_b is achieved by a first step which projects from L_1 dimensions to L_3 and then a repetition channel code which transforms the L_b bits into L_3. Nevertheless, the optimal value for L_3 when all the samples are i.i.d. is $L_3 = L_b$, i.e. no repetition coding, but only projection, as was shown in [132].

5.3.3. Channel coding with no projection

In this case, we have:

$$
\begin{aligned}
\mathbf{S} &= \text{diag}(1/\gamma_1, \cdots, 1/\gamma_{L_1}), \\
\mathbf{T} &= \text{diag}(\gamma_1, \cdots, \gamma_{L_1}), \\
\mathbf{c} &= f(\mathbf{b}), \\
r_j &= c_j, \quad 1 \leq j \leq L_1.
\end{aligned} \tag{5.10}
$$

As it can be clearly seen, repetition coding without projection is just a particular case of the previous methods. Nevertheless, it is interesting to address it separately due to its practical importance. In practical situations $f(\cdot)$ could be any kind of channel code: turbo [23, 88], serially concatenated [22], block [108], LDPC [79], etc.

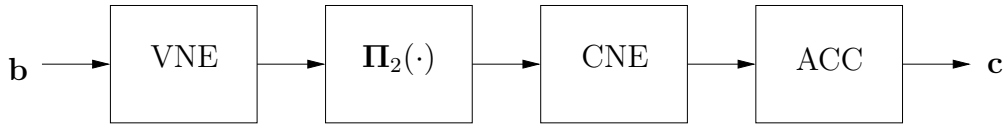


Figure 5.5: Structure of $f(\cdot)$ for Erez and Ten Brink's scheme.

5.3.4. Channel coding with projection

Finally, a last alternative could be:

$$
\begin{aligned}
s_{ij} &= \begin{cases} \frac{L_c}{L_1}\frac{q_j}{\gamma_j}, & \text{if } (i-1)L_1/L_c < j \leq iL_1/L_c, \text{ and } 1 \leq i \leq L_c \\ 0, & \text{otherwise} \end{cases}, \\
t_{ij} &= \begin{cases} q_i\gamma_i, & \text{if } (j-1)L_1/L_c < i \leq jL_1/L_c, \text{ and } 1 \leq j \leq L_c \\ 0, & \text{otherwise} \end{cases}, \\
\mathbf{c} &= f(\mathbf{b}), \\
r_j &= c_j, \quad 1 \leq j \leq L_1,
\end{aligned} \tag{5.11}
$$

where the same comments made in Section 5.3.3 are still valid.

5.4. Erez and ten Brink's approach

Erez and ten Brink's scheme [68, 69] can be regarded to as one of the foremost existing dirty paper codes, which to the best of our knowledge has not been applied yet in data hiding scenarios.

It consists of a check-biregular, repeat-irregular nonsystematic repeat-accumulate code concatenated with a vector quantizer. In other words, the encoder is composed of a variable node encoder (VNE), which is nothing but a variable-rate repetition encoder, whose output is permuted using $\mathbf{\Pi}_2(\cdot)$ to become the input of a check node encoder (CNE), which is a single parity check encoder. The variable node encoder has 64.36% of the nodes of degree 3, 31.24% of degree 10 and 4.4% of degree 76. 80% of the check nodes have degree 1 and 20% degree 3. The concatenation of both of them, yields a total rate 1/6. The bits in the output of this check node encoder go through a recursive accumulator (ACC). All the variable node encoder, the permuter, the check node encoder and the recursive accumulator can be seen as a channel code $f(\cdot)$ and its output $\mathbf{c}$ constitutes the input of a vector quantizer, that will be explained in Section 5.4.1. The structure of $f(\cdot)$ for this scheme is plotted in Figure 5.5.

This quantizer finds that centroid of a lattice (which depends on the input bits) which minimizes the distortion between the side information $\mathbf{X}$ and the output

signal $\mathbf{Y}$. This distortion measure can be changed depending on the requirements of our system, although for Erez and ten Brink's paper the Euclidean distance between both signals is employed. The search of this centroid implies using a Viterbi algorithm, so the embedding process is computationally much more expensive than for turbo-codes. In the data-hiding problem, a typical choice for the distortion measure could be a perceptual measure, which will obviously depend on the nature of the host signal. For example, when $\mathbf{X}$ is the 8×8 block-wise DCT of an image, the perceptual measure by Watson could be used [160]. Other alternative could be a weighted Euclidean distance, which normalizes the distortion in each dimension by the perceptual mask $\boldsymbol{\gamma}$. In our implementation, we have followed the last strategy for the sake of simplicity.

Another problem to be solved is how to increase the redundancy for a fixed structure (which implies a fixed rate) of the channel code and vector quantizer. The solution we have adopted is based on projecting the initial vector $\mathbf{X}$ onto a lower-dimensional space (using $\mathbf{S}$). In this way the SNR per dimension will be increased in average by L_1/L_3.

As a consequence of the previous discussion, we can write

$$s_{ij} = \begin{cases} \frac{L_3}{L_1}\frac{q_j}{\gamma_j}, & \text{if } (i-1)L_1/L_1 < j \leq iL_1/L_3, \text{ and } 1 \leq i \leq L_3 \\ 0, & \text{otherwise} \end{cases},$$

$$t_{ij} = \begin{cases} q_i\gamma_i, & \text{if } (j-1)L_1/L_3 < i \leq jL_1/L_3, \text{ and } 1 \leq j \leq L_3 \\ 0, & \text{otherwise} \end{cases},$$

The decoding is carried out by the iterative decoding of three blocks: vector quantizer and accumulator (VQ + ACC), check node decoder (CND), and variable node decoder (VND). In exchange for this increase in complexity, significant performance gains can be achieved, as it is shown in Section 5.5. Figure 5.6 shows the structure of the decoder.

5.4.1. Vector Quantizer

The vector quantizer proposed by Erez and ten Brink [68, 69] (see Figure 5.7) groups the bits into triplets. One bit per triplet is duplicated and combined with the output of a non-systematic convolutional code with feedforward polynomials 07_8 and 05_8, whose input are the *virtual bits*. These *virtual bits* are not information bits but a tool to shape the quantization region of the vector quantizer; they can be arbitrarily flipped and give a degree of freedom to modify the watermark in such a way that a distortion measure between the original host signal and the watermarked one is minimized. The presence of these *virtual bits* is what accounts for the difference between a scalar quantizer and a vector one. The optimal sequence of virtual bits, i.e. that minimizing the target distortion measure, is computed

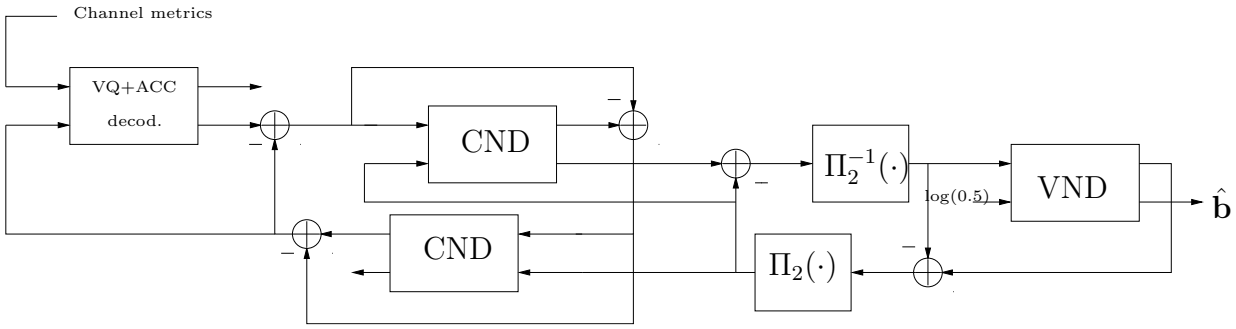


Figure 5.6: Structure of Erez and ten Brink's decoder. The lower input and output to the decoding blocks are the *a priori* and *a posteriori* log-probabilities of the input of the corresponding encoding block. The upper ones are the same probabilities, but corresponding to the output of the encoding block.

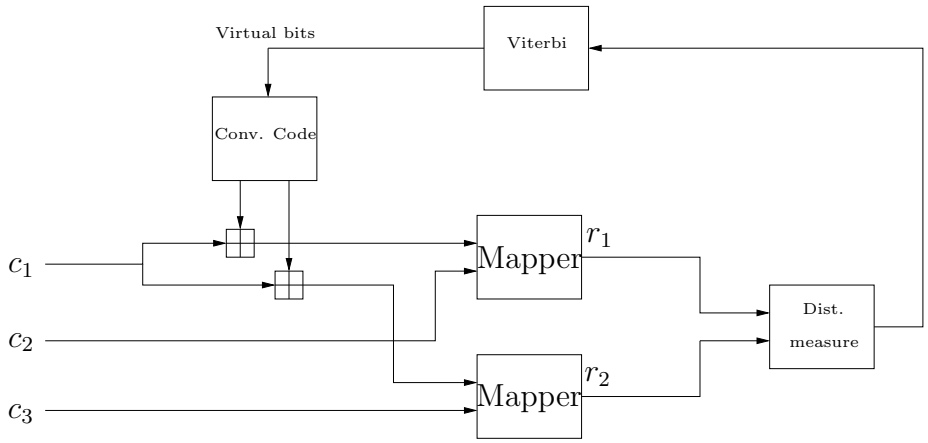


Figure 5.7: Structure of Erez and ten Brink's vector quantizer.

using a Viterbi algorithm, and the resulting output is combined with the information coded bits, yielding 4 bits which are used to index two 4-PAM symbols (or, equivalently, a 16-QAM symbol) with alphabet $\mathcal{W} = \{-3/2, -1/2, 1/2, 3/2\}$, obtaining $\mathbf{r}$, wich is used in (2.21) to get $\mathbf{W}_p$. Moreover, $\mathbf{r}$ is taken into account to measure the distortion, which is used by a Viterbi algorithm to determine the optimal virtual bits sequence. Bearing this structure in mind, the total rate of the scheme is 1/4.

This vector quantizing resembles the method proposed by Miller et al.[115], since both of them try to find a watermark which minimizes a distortion measure taking into account all the components of the watermark. Nevertheless, the differences are evident: Miller et al.'s method is based on an heuristic trellis coding, while the search of the optimal watermark is more systematic in Erez and ten Brink's scheme.

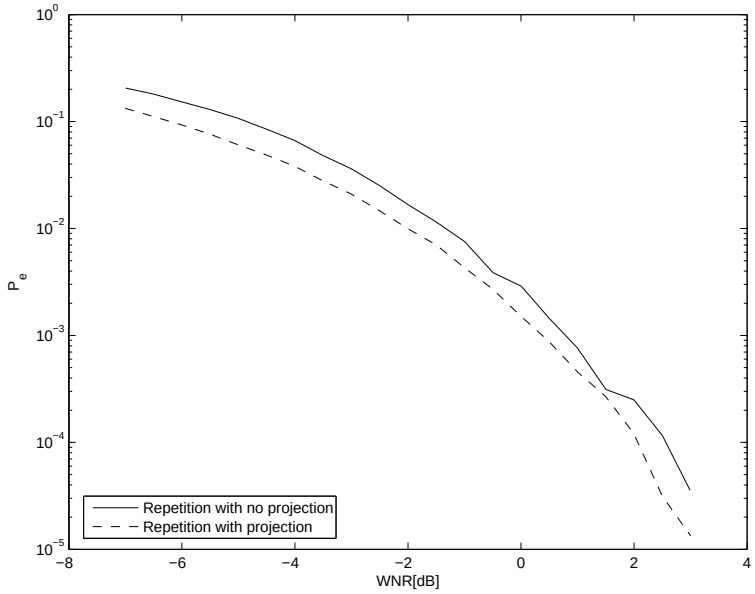


Figure 5.8: Comparison of repetition coding with and with no projection. $L_1/L_b = 20$.

5.5. Experimental results

For the experimental part of this section we have watermarked *Lena* 256×256 in the mid-frequencies of the 8×8-DCT domain [93], using a perceptual mask based on Watson's distortion [160]. In all experiments each information bit was hidden in 20 coefficients, giving a total payload of $1,122$ bits. The channel-noise was chosen to be Gaussian with the same power in all coefficients (i.i.d.). In order to address a real scenario, a value of α was set for each experiment and held constant for the entire range of WNR's.

First of all, we have compared the repetition coding schemes, both with and without projecting. The values of α were 0.5 and 0.9 respectively. This difference is due to the different SNR per dimension in each scheme, since the optimal α in the first case is computed by taking into account the SNR in the projected domain, which is increased by $10\log_{10}$ of the projection factor. In Figure 5.8 the improvement due to projecting is shown. Both schemes were decoded using Maximum Likelihood (ML) lattice decoding [70].

In order to compare dirty paper schemes which use repetition coding with those using channel coding, we have chosen a serially concatenated code proposed by Benedetto et al. [22] with outer code $G_o(D) = [1 + D, 1 + D + D^3]$ and inner $G_i(D) = [1, (1 + D + D^3)/(1 + D)]$, giving a total rate 1/4, which is used with projection of rate 1/5. In Figure 5.9 the turbo-cliff of this code for dirty paper coding when all the components are i.i.d. is shown. Figure 5.9 also shows the turbo-cliff of Erez and ten Brink's scheme for the same scenario. In the paper

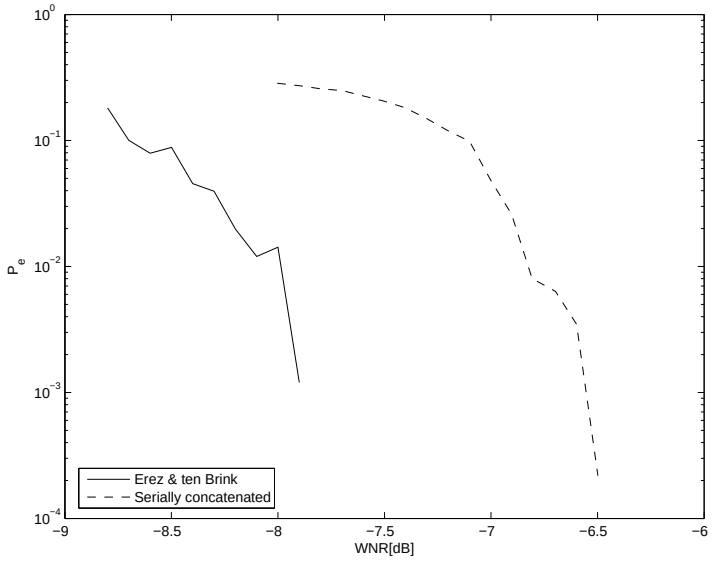


Figure 5.9: Comparison between a serially-concatenated code concatenated ($\alpha = 0.5$) with a scalar quantizer and Erez and ten Brink's scheme ($\alpha = 0.42$) when the noise components are i.i.d.. $L_b/L_c = 1/4$ and $L_c/L_1 = 1/5$ with projection.

by Erez and ten Brink[68] the turbo-cliff was at WNR $= -1.9$ dB (1.93 dB from capacity limit) or, equivalently, $E_b/N_0 = 1.1$ dB, so taking into account the increase in the WNR due to the projection, one would expect such turbo-cliff to show up at -8.9 dB (2.55 dB from capacity limit). Nevertheless, Figure 5.9 (where we use $\alpha = 0.42$) shows it around -7.8 dB. In fact, we can decompose the gap to the capacity limit (3.64 dB) into a gap due to the method itself (1.92 dB), another part due to projecting instead of using a more sophisticated code (0.63 dB), and finally the part corresponding to the use of a limited-size permuter (1.09 dB). In any case, the gain achieved by using Erez and ten Brink's scheme compared with the serially concatenated codes is around 1.3 dB, see Figure 5.9.

Figure 5.10 shows the results when noise samples are Gaussian and independent but not identically distributed. The gain by using Erez and ten Brink's scheme is still around 1.5 dB, but both plots are now shifted almost 2 dB to the right, so the turbo-cliffs are found now at -5.8 dB and -4.5 dB. Finally, it is interesting to remark that the gain due to projecting when the serially concatenated codes are used, is almost negligible, as can be seen in Figure 5.10.

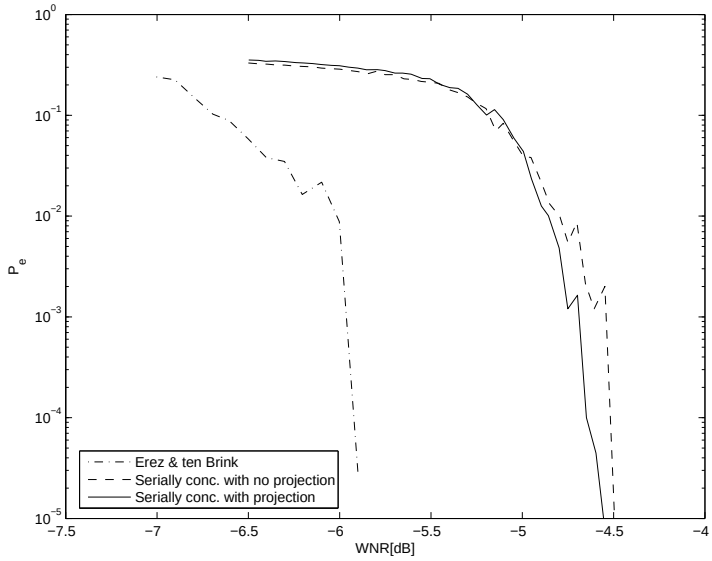


Figure 5.10: Comparison between a serially-concatenated code concatenated with projection ($\alpha = 0.6$) and with no projection ($\alpha = 0.3$) with a scalar quantizer, and Erez and ten Brink's scheme ($\alpha = 0.415$) when the noise components after normalizing by the perceptual mask are independent but not identically distributed. $L_b/L_c = 1/4$ and $L_c/L_1 = 1/5$.

5.6. Subsequent works on the combination of source-coding and channel coding

After the publication of the paper that constitutes the main part of this chapter, i.e. [47], other works have appeared dealing with the problem of combining source coding and channel coding for approaching the capacity of the dirty-paper problem. First, in [146] Sun et al. show that the gap-to-capacity in [69] can be reduced about 0.5 dB; the technique proposed is based on choosing *"a strong source coding first and then focusing on designing near-capacity channel codes"*. As the authors of [146] show, the source coding proposed in [69] is far from being optimal, achieving a shaping gain of 1.22 dB (and a distance of 1.32 dB to the SNR yielding a capacity of 0.25 bits) for the proposed 64-state vector quantizer, and 1.28 dB (1.15 dB distance to the SNR yielding a capacity of 0.25) for the 256-state one. Su et al. conjecture that this is due to the use of a 16-QAM constellation and the introduction of systematic doping. Taking this into account, their proposal uses a 16-PAM constellation and a vector quantizer with a high number of states (the implemented case has 1024 states), achieving a shaping gain of 1.38 dB, which is translated into a distance of 0.83 dB to the SNR needed to achieve a capacity of 0.25 bits.

The main ideas of this paper are used by Yang et al. in [163] for designing

an image data-hiding scheme. In that paper the authors follow a strategy similar to that introduced in [47] (and exposed previously here) for adapting the dirty-paper coding scheme to the data-hiding problem. In fact, they use the same mid-frequencies as in [47], and normalize each dimension by the corresponding value of Watson's mask, following also a similar projecting strategy for simulating time-sharing. Nevertheless, due to the specific design of the dirty-paper codes for this problem, where the authors also took into account the error floor phenomenon due to the small codeword lengths, the authors obtained a data-hiding scheme with performance similar to that based on Erez and ten Brink's structure and described in this chapter, but with reduced redundancy. In fact, while our adaptation of the method by Erez and ten Brink has the ratios $L_b/L_c = 1/4$, and $L_c/L_1 = 1/5$, that introduced in [163] is characterized by $L_b/L_C = 1/5$ and $L_c/L_1 = 1/3$, increasing the possible payload by 33%.

5.7. Conclusions

In this chapter we have proposed a framework that encompasses many side-informed methods with coding for data-hiding, and reviewed state-of-the-art methods, specifying two possible ways to increase the operating SNR: repetition coding with and without projection. Moreover, we have introduced for the first time in watermarking a capacity-approaching dirty-paper scheme by Erez and ten Brink. The gap to capacity of this scheme is measured for Gaussian i.i.d. noise, showing the different causes of this loss. Experimental results comparing the performance of that scheme with serially concatenated codes and repetition, with and without projection, have been also introduced for non-i.i.d. noise, showing again a similar improvement when the new scheme is used.

Chapter 6

Application to a Video Surveillance Authentication System

Apart from the theoretical results presented so far, in this thesis we have also paid special attention to the practical implementation of watermarking applications. In this sense, we have designed an algorithm for the authentication of images in a video surveillance application; the main features are summarized in this chapter. First of all, the video surveillance system we took under consideration is described. Taking its peculiarities into account, the requirements of our system are explained. Afterwards, we introduce the proposed solutions, and finalize this chapter describing the main problems found.

6.1. Framework

The considered video surveillance system (which is plotted in Figure 6.1) is constituted by a set of cameras placed at different locations and a central server, where the images are stored. This server could be located at the client's place or at the facilities of a service provider. Since the cameras usually output images with a high redundancy, in order to reduce the space needed to store the videos the images can undergo a lossy compression process before storing them. The objective of storing the videos is that if a theft were peperpetrated, they could provide information about the thieves' identities. Nevertheless, a smarter burglar could try to fool this system; for example, he/she could replace the live images by previously recorded ones, or even just replace the particular piece that compromises him/her. Therefore, this makes evident the need of additional methods which guarantee the integrity of the stored images; that is obviously the problem to be solved by digital watermarking.

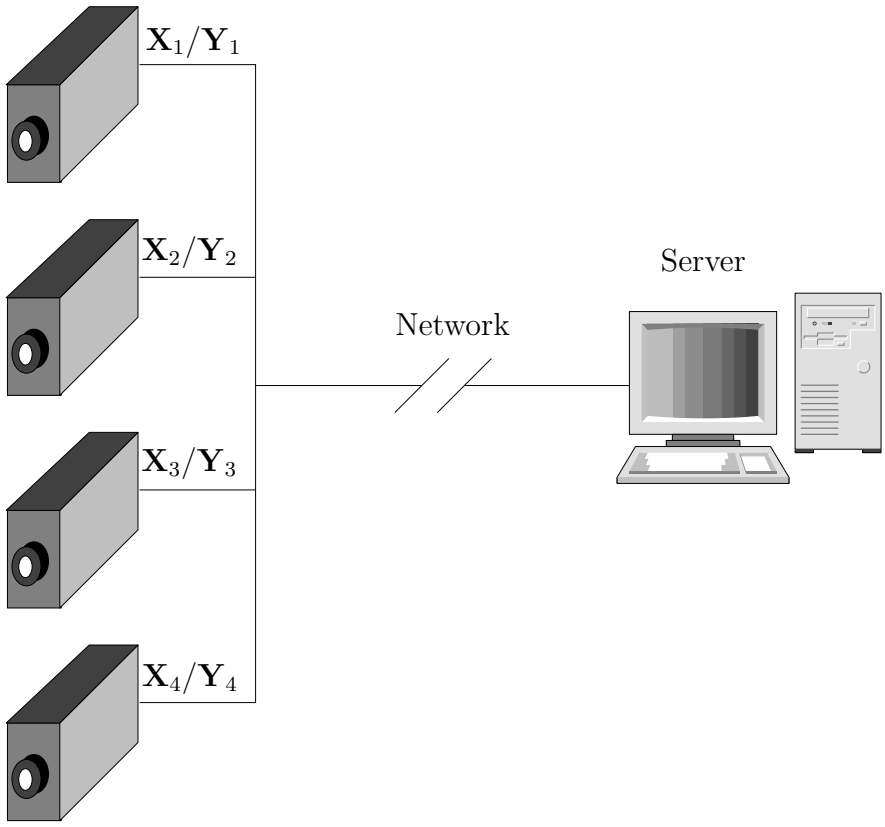


Figure 6.1: Video surveillance system model. If the images are watermarked in the camera, its output is $\mathbf{Y}$; in other case, $\mathbf{X}$, i.e. the original host signal.

Digital watermarking could be used to detect modifications in the original image; in fact, we will show in this chapter that we can even use it to detect which part of the image was modified, or to detect the replacement of the original video sequence by a previous one.

Concerning the place where the embedding is performed, two basic choices are available: at the camera itself or at the server. The first choice is possible due to the existence of *smart* cameras, i.e. cameras which have an operative system, allowing low-computational demanding programs to be run. Furthermore, this choice seems to be the most secure one; as long as the original host signal (without watermark) is output, the burglar could try to forge it in the midway to the server. Therefore, if the embedding were to be performed at the server, a secure communication channel should be established between the camera and the server (based on cryptographic tools). Just another advantage of using *smart* cameras for embedding is that then the server could check for the presence of the watermark in the received signal; if the watermark were not detected, this would mean that the image had been modified in the communication channel, and the server could make the alarm. In this last case, the trade-off between probability of false alarm and missed detection should be carefully evaluated: the client wants to be

warned when an intruder goes in his/her business, but probably he/she would not like to be woken up in the middle of the night because a transcoding artifact removed the watermark.

In any case, in both cases one should pay special attention to the security and reliability of the channel between the camera and the server; it is usually said that the security of a system is given by the security of its weakest link, and this channel seems to be one of the weakest parts of our system. In the remainder of this chapter we will just focus on the watermarking application, making some technical assumptions whose solution lies outside the scope of this work (as the camera-server channel, storage needs and security, legal issues, etc.).

6.2. Requirements

In this section, we will enumerate the requirements for our system, and in the next one we will describe the proposed solutions to meet such requirements. We will require our system to verify the following:

- As it was explained in Section 1.1, authentication systems may be based on the so-called fragile watermarks; this means that the detector should warn of modifications in the watermarked signal. Nevertheless, due to the nature of our application, we are not just interested in the detector to tell whether the image was modified or not, but also in signaling the modified regions. In that way, we could determine which part of the image the attacker wanted to remove/change, and take further actions. This is obviously a detection (binary hypothesis) watermarking problem.

- Another important point is the temporal ordering; the watermarked image must depend on the instant it is produced, in such a way that a video sequence can not be replaced by a previous one.

- We would also like the watermark to convey some data; in this way we could hide in the image itself information about the camera that took it, other data that the client may consider relevant, or even enable future extensions. It is straightforward to see that this constitutes a decoding (multiple hypothesis) data hiding problem.

- The detector should be robust against some simple synchronization problems. Even when this requirement is apparently opposed to the fragile nature of the watermark, sometimes certain rows or columns of the watermarked image could be accidentally removed, and we would be still interested in detecting if there was a further malevolent modification.

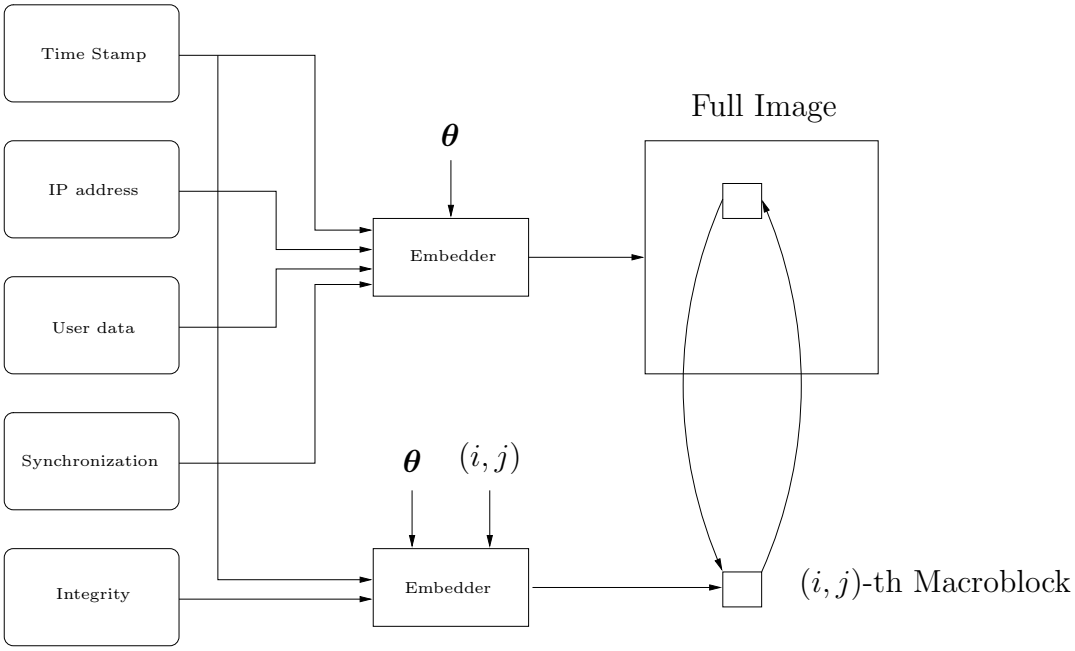


Figure 6.2: Scheme of the proposed solution.

- The system should be robust to unintentional attacks, such as transcoding; specifically, we have imposed our system to be robust to the conversion from MJPEG to MPEG formats.

6.3. Proposed Solution

As it was shown in the previous section, the proposed system will require both decoding and detection features, and in addition the synchronization of the blocks in the image must be also possible. Taking this into account, we have divided the watermark in three parts: synchronization, data hiding and integrity (the authenticacion itself). In the following, we describe how we met the requirements introduced in the last section, and which of those three parts is related to each requirement. Fine-grain details concerning parameters, coefficients, etc., will be reported elsewhere.

- The problem of temporal ordering can be solved by introducing a time stamp in each frame, and making the watermark depend on that stamp. In this way, the decoder can easily verify if the received frame is in the correct sequencial order, or if it is a repeated or an old one; this can be implemented by a simple counter. The part of the watermark performing the authentication can be made dependent on the time stamp by modifying some of its parameters (e.g., projecting sequences, dither vector). This entails a risk,

since, as it was discussed, the system should be robust, especially against unintentional attacks, such as transcoding, and if this time stamp is not correctly decoded, the system parameters that the detector will use will be incorrect, yielding invalid results. Therefore, we have decided to introduce a high redundancy in order to protect that data; the embedding mechanism is explained next.

The time stamp is coded as a 35-bit length vector; this enables one to distinguish every frame for about 40 years, assuming that the frame rate is 25 frames/second. Be aware that in real video surveillance applications the rate will be typically lower, so this is a conservative figure. In order to prevent decoding errors, those 35 bits go through a channel code with rate 7/8, obtaining a vector with 40 encoded bits. Each of those bits is pseudorandomly assigned to a large set of coefficients of the 8×8-block DCT transformed host all over the image; the vector with those coefficients is projected to a lower dimensional domain, where it is quantized using repetition coding and the Cartesian product of uniform scalar quantizers. The projecting sequences should be the same for all frames, or, at most, there could exist a reduced number of possible sets of projecting sequences. In any case, they depend on the secret key, which is only shared by embedder and decoder.

The deprojected version of a fraction α of the quantization error is added back to the original host signal. Note that this scheme is a particular case of generalized version of STDM described in [132], and that was described in Section 2.7 in this work.

- The additional data hiding corresponding to information about the camera (we proposed to hide its IP address, that represents a 32 bits payload), and the user data (in the current implementation 28 bits are allowed) are hidden using a strategy similar to that used for hiding the time stamp. Nevertheless, the projecting rate is not so small as in that case, since an error here is not as crucial as in the time stamp; in fact, no additional channel coding is used.

- In order to solve the synchronization problem, the use of a synchronization pattern repeated block-wise is proposed. The pattern is fixed independently of time. By doing so, the detector can tune the shift or cropping the image has undergone without having to estimate the temporal stamp.

  One can realize that the fact of repeating the pattern implies a security flaw. Nevertheless, *collusion*-like attacks are not a problem in this case, since even when the attacker could estimate the pattern, he/she is not interested in removing it.

- Finally, integrity is implemented by a watermarking detection system similar to that in [126]. First, the 8×8-block DCT image is arranged in

macroblocks. These macroblocks are constituted by neighbor 8×8 blocks, and the detection of the watermark will be performed over each of them. In this way, the detector will decide in what of the macroblocks the watermark is present (i.e., the watermarked signal is assumed to have not been strongly modified in those macroblocks), and in which of them the watermark is absent (i.e., the watermarked signal has been modified so much that the watermark has been removed). In the embedding and detection processes, the coefficients of a macroblock devoted to checking integrity are projected to a smaller dimensional domain. In this case, the projecting sequence will depend on the time stamp and the secret key, ensuring that the attacker will not have access to the watermarking channel. Furthermore, the spreading sequences will change from frame to frame, disabling an attack oriented to estimate the spreading sequences by observing different watermarked frames.

The embedding is also based on quantization in the projected domain, and adding back a fraction α of the quantization error, similarly to the data hiding part. Nevertheless, the detection stage is clearly different of the decoding performed for recovering the time stamp or the additional data. In this case, the detector is just interested in knowing if the watermark is present or not. In order to do so, we establish a region around the centroids of the quantizing lattice, and, if the projected received signal is inside one of those regions, the detector will state that the watermark is present (or absent if the received signal is out of the region). The size of those regions is strongly related to the probability of false alarm and missed detection of the scheme. Depending on the level of accuracy desired by the client, different thresholds can be established to play with these probabilities.

6.4. Main problems found and conclusions

In the design and later implementation of this scheme [48], several problems have been found. First of all, the dependence of the quantization noise due to transcoding and the watermarked signal invalidates most of the results in the literature for the quantization based methods. As a direct consequence of this dependence, we had to assume that the embedder knows the target image format, so the amplitude of the projecting and deprojecting sequences can depend on the quantization step of that image format; by doing so, the embedder can envisage the signal received by the detector, and therefore, the result of the detection. This assumption would not be necessary if the watermark were very much larger that the transcoding quantization step; unfortunately, this is not usually the case, since a figure-of-merit of the system will be its storage capacity, and if the size of the host signal is wanted to be reduced, large transcoding quantization steps must be used.

Another parameter we played with was the number of DCT coefficients of each 8×8-block we used to embed the watermark. We have observed that the larger this number is, the larger the perceptual distortion will be. This is especially harmful on the high frequencies, where very large quantization steps are used, and the amplitude distortion due to the embedding must be also proportional to those steps. On the other hand, the performance of the system will be increased with the number of coefficients devoted to convey the watermark, where the performance can be measured as the robustness to transcoding, or the probality of false alarm for a given probability of missed detection (ROC).

Other parameter that must be taken into account is the watermark power at a given coefficient. As in the previous case, the larger the power of the watermark is, the better the perfomance will be, but also the larger the distortion that is introduced. Therefore, the typical trade-off distortion vs. performance is also observed in this case; the choice of the operating point of the system will typically depend on the client's choice.

6.5. Results

In this section we show the behavior of our system when modifications are introduced. In Figure 6.3 some original frames are plotted, and their watermarked versions were depicted in Figure 6.4. The watermarked video was modified in order to remove the man coming into the room, change the date in the upper part of the frame, and remove the dark square in the background; the resulting frames, corresponding to the images in Figures 6.3 and 6.4, are depicted in Figure 6.5. Finally, in Figure 6.6 we have plotted the result of performing the detection of the modified signal, depicting in white the macroblocks that the algorithm detected to have been altered. It can be observed that the three modifications are detected when the sequence number is correctly decoded; when this is not the case, i.e. when the frame is so strongly modified that the sequence number can not be properly decoded, a *"sequence number not in order"* error is raised, and given that the spreading sequence for the integrity part depends on that sequence number, all the frame is found to be modified.

Figure 6.3: Original frames.

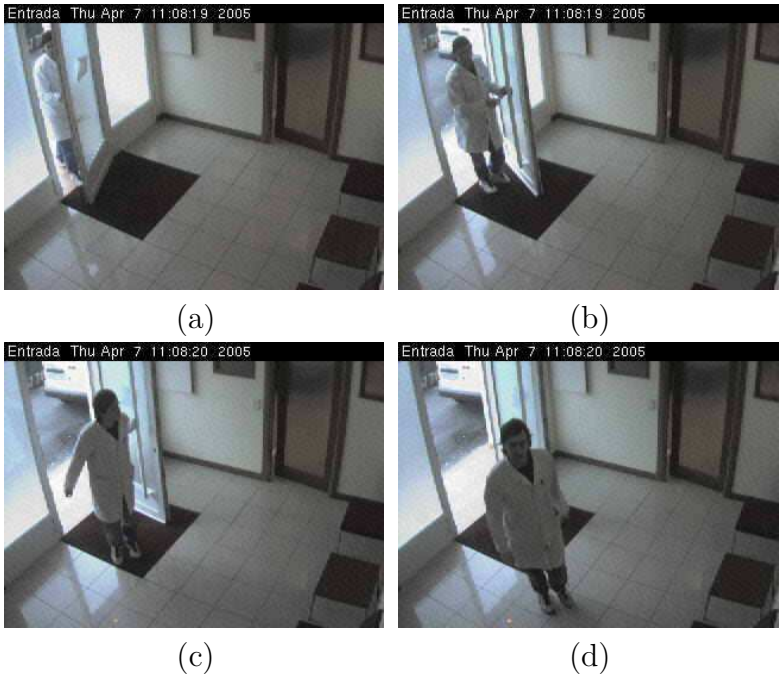


Figure 6.4: Watermarked frames.

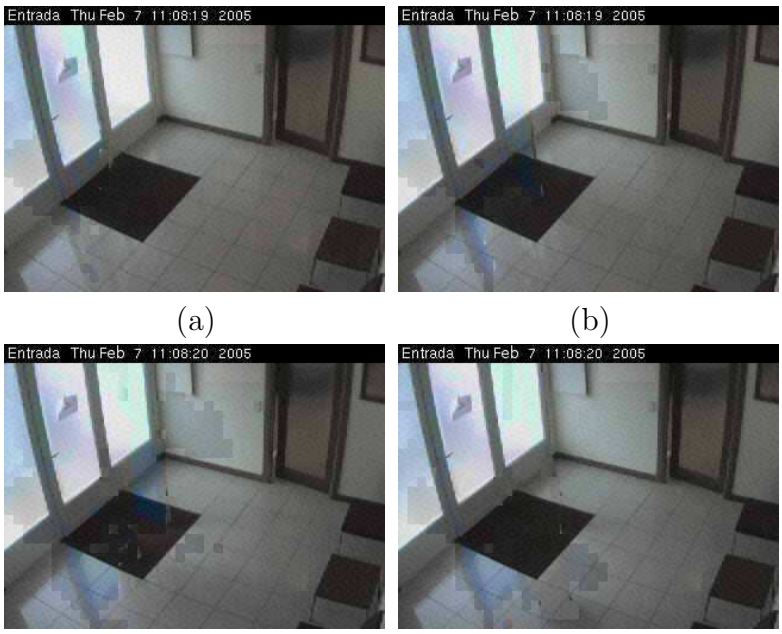


Figure 6.5: Modified frames.

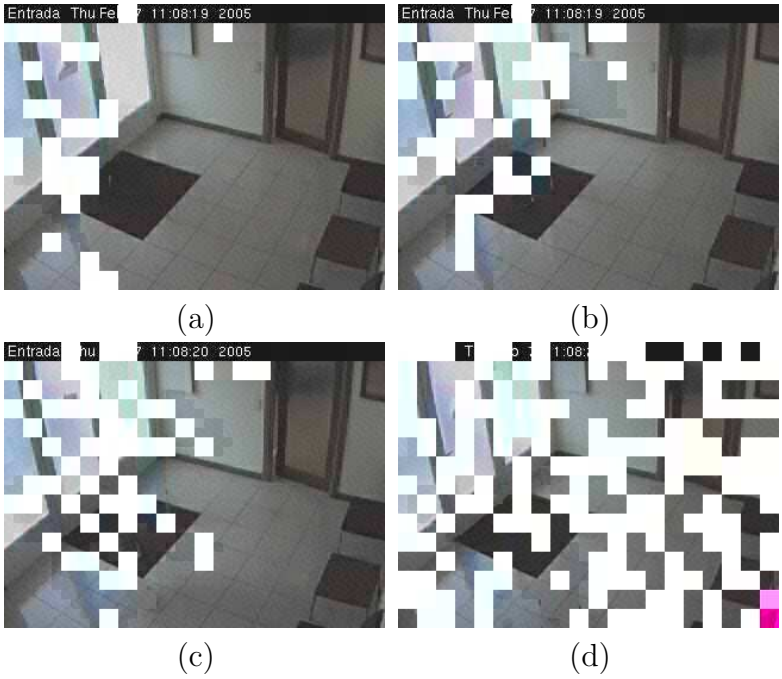


Figure 6.6: Frames result of the integrity check. Those macroblocks that are detected to have been modified are plotted in white.

Chapter 7

Conclusions

In this thesis, we have focused on the analysis of two of the main characteristics of data hiding systems: robustness and security. Moreover, even though our main attention was paid to side-informed methods, we have also introduced and analyzed other schemes, such as spread-spectrum based, when the comparison was judged to be valuable.

Among the main contributions of this work, we will now emphasize those that we consider to be more representative and useful:

- Computation of the exact probability of decoding error for DC-DM based on uniform scalar quantizers and repetition coding under additive noise attacks: even for that scheme, that could be considered to be the simplest version of DC-DM with channel coding, there was a lack of literature dealing with its exact performance. In fact, the previous works usually provided upper bounds to that probability of error. In contrast, in this thesis we have obtained for the first time the exact probability of decoding error for DC-DM based on uniform scalar quantizers and repetition coding. This result, which is based on the modulo-reduced version of the total noise, i.e. self-noise plus channel noise, is completed with the proposal of some bounds and approximations to that probability of error, whose computational cost is significantly reduced compared to the exact computation. Considering these approximations, some improved decoding weights are proposed, noticeably improving the performance of the studied scheme.

- Computation of the exact probability of decoding error for DC-DM based on uniform scalar quantizers and repetition coding under coarse quantization attacks: we have statistically analyzed the probability of error of that version of DC-DM by considering the pdf of the watermarked signal, and the way coarse quantization modifies that pdf.

- The performance analysis of SSTDM under cropping attacks constitutes also a novel approach: previously in the literature it was assumed that SSTDM was more robust that DC-DM based on uniform scalar quantizers and repetition coding; nevertheless, this was shown not to be the case for this kind of attacks. Furthermore, the performance degradation of SSTDM has been compared with that of Add-SS, and a new method has been proposed trying to encompass the advantages of SSTDM against additive noise attacks and those of DC-DM with uniform scalar quantizers and repetition coding against the cropping attack.

- Proposal of a sensitivity attack which has been shown to be suitable for attacking most of state-of-the-art data hiding methods. The proposed method, termed Blind Newton Sensitivity Attack (BNSA), was used both for removing the watermark from watermarked contents and for creating forgeries (i.e., falsely watermarked contents). The main advantages of BNSA compared with previous proposals in the literature lie on the fact of just needing the binary output of the detector, since it does not require any knowledge about the detection function. The effectivity of BNSA was shown with several experiments.

- Trying to address smarter attacks, we have also studied Add-SS, DC-DM with uniform scalar quantizers and repetition coding, and SSTDM from a game-theoretic approach. Nevertheless, due to the complexity of the resulting expressions, closed formulas were only obtained in some particular cases.

- Computation of the worst case (from an information theoretic point of view) additive attack for scalar DC-DM. These results show how far is the optimal attack in that scenario from being Gaussian, as it is usually the case in most of communications frameworks.

- Concerning security, definitions of both robustness and security (two concepts usually mistaken in the literature) were proposed. Later, information theoretic measures were used for quantifying the security of Add-SS and Costa's scheme, comparing the results obtained following this novel approach with existing ones. As the main conclusions of this analysis, we can enumerate:

    1. The host interference can be profitable from a security point of view: the host interference makes more difficult the communication between embedder and decoder, but it also complicates the estimate of the secret key by the attacker.
    2. For both Costa and DC-DM, the smaller the distortion compensation parameter α, the smaller the information leakage: this is a direct consequence of the increase in the self-noise.
    3. The information leakage is reduced when the DWR is increased.

4. The information leakage is reduced when the rate of the watermarking system is increased: a higher rate increases the uncertainty about the sent symbol, making more difficult the estimate of the secret parameters.
5. For a given information leakage, the security of the system could be increased by increasing the *a priori* uncertainty about the parameters to be estimated. This last point can be achieved by increasing the number of parameters to be estimated, whose upper bound is given by Costa's scheme, where all the codewords are independently randomly generated.
6. The structure of the codebook helps to obtain feasible embedding and decoding algorithms, but it also makes easier the attacker's job: in both Add-SS and DC-DM the attacker has to estimate just a codeword for each sent message, since the remaining ones can be written as a function of that codeword. Nevertheless, when there is a complete lack of structure, as it happens for Costa's scheme, the information learnt from one observation is applicable just to the codeword related to that observation, saying nothing about any other codeword.

Summarizing, several trade-off's can be established when security is analyzed: security vs. host interference (with interesting links to performance), security vs. self-noise, security vs. structure of the codebook, etc.

- Another topic which has been paid special attention in this thesis is the need of combining channel coding and source coding in the design of any data hiding system truly approaching capacity. In fact, a dirty paper coding scheme aproaching capacity designed by Erez and ten Brink was adapted to data hiding, comparing the obtained results with those of the typical strategies followed so far, where just channel coding was considered. Although the results obtained for the adaptation of Erez and ten Brinks's scheme clearly outperform those methods just based on serially-concatenated codes, the gap to capacity of the former is still quite large.

- Finally, another contribution of this thesis is the design of a video surveillance authentication system. With the proposed scheme the detector is not just able to determine if a given frame was modified or not, but it can also define what parts of the image were tampered with.

7.1. Future Research Lines

In this section we will enumerate some topics that stay as possible future research lines:

- Information theoretic analysis of sensitivity attacks: in order to obtain the final performance of sensitivity attacks, it seems to be reasonable to try to analyze those attacks from an information-theoretic point of view.

- Computation of the Worst Case Additive Attack for lattice based quantization methods more general than that described in this work: some research groups are already currently working on this topic, replacing the unidimensional lattice by a bidimensional one [117]. In any case, closed-form results seem to be difficult to obtain in this field; moreover, increasing the number of dimensions also seems to be hard, since the number of parameters to be optimized exponentially increases with such dimensionality.

- Relation between attacks to security and attacks to robustness: As it was previously introduced in this work, attacks to security could be performed as a previous step to attacks to robustness. Therefore, it makes sense to try to define some measure that quantifies the goodness of a security attack from that point of view, as well as the design of attacks to robustness which take advantage of knowing an estimate of the secret parameters.

- Security analysis of other methods in the literature, for example the Rational Dither Modulation [134], QIM-based methods with rotated lattices [81, 117], or methods based on quantization on a projected domain (like STDM) [32].

- Concerning the joint use of channel coding and source coding, one could think of replacing the distortion compensation parameter α by its optimal value, i.e., the Wiener filter. This would improve the system performance in the case of Gaussian, independent but not identically distributed noise. Another challenge is the design of similar methods to those introduced in this thesis, by specifying different repetition and checking rates for the VNE and CNE respectively, in order to use them in scenarios with lower SNRs, instead of increasing the operating SNR through projection.

- Detailed analysis and proposal of new data hiding methods for emerging applications, as authentication and fingerprinting: in the last years these two applications have been paid an increasing attention by the academy, and specially by the industry. A consequence of this interest is the appearance of some commercial applications dealing with these problems. Nevertheless, we think that large improvements can still be obtained.

Appendix A

Comparison of two lattice schemes

In this section we will compare the embedding and decoding scheme used in this work, with a similar one used in the literature by Erez and colleagues (see for example [166, 70, 69]). Following the notation introduced in Section 2.5, we can write the watermarked signal as

$$\begin{aligned}\mathbf{Y} &= (1-\alpha)\mathbf{X} + \alpha\Big[Q_\Lambda\big(\mathbf{x}-\mathbf{v}(b)\big)+\mathbf{v}(b)\Big] = \mathbf{X} + \alpha\Big[Q_\Lambda\big(\mathbf{X}-\mathbf{v}(b)\big)-\mathbf{X}+\mathbf{v}(b)\Big] \\ &= \mathbf{X} + \alpha\Big[\big(-\mathbf{X}+\mathbf{v}(b)\big) \mod \Lambda\Big];\end{aligned}$$

if the modulo-Λ reduced host signal is uniformly distributed over $\mathcal{V}(\Lambda)$, then the power of the watermark is

$$D_w = \frac{\alpha^2}{L_2}\frac{\int_{\mathcal{V}}||x||^2dx}{\mathrm{Vol}(\mathcal{V})}, \tag{A.1}$$

where $\mathrm{Vol}(\mathcal{V})$ denotes the volume of $\mathcal{V}$. In decoding, the modulo reduced received signal is given by

$$\begin{aligned}\mathbf{Z}_{\mathrm{mod}} \triangleq \mathbf{Z} \mod \Lambda &= \Big\{\mathbf{X}+\mathbf{N}+\alpha\Big[\big(-\mathbf{X}+\mathbf{v}(b)\big) \mod \Lambda\Big]\Big\} \mod \Lambda \\ &= \Big\{\mathbf{N} \mod \Lambda + (1-\alpha)\Big[\big(\mathbf{X}-\mathbf{v}(b)\big) \mod \Lambda\Big] + \mathbf{v}(b)\Big\} \mod \Lambda.\end{aligned}$$

The corresponding block diagram is plotted in Figure A.1.

On the other hand, the scheme described in [166, 70, 69] computes the watermark as

$$\mathbf{W} = [-\alpha\mathbf{X} + \mathbf{v}(b)] \mod \Lambda,$$

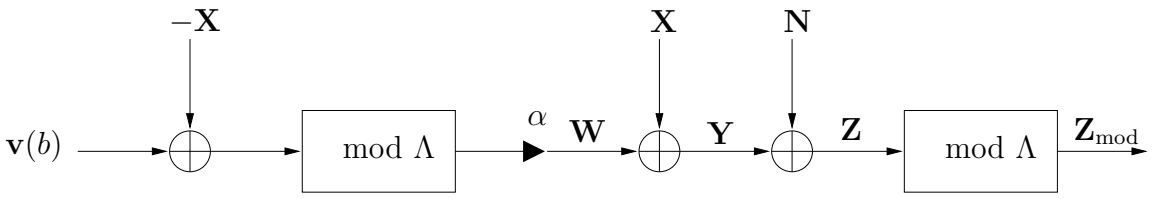


Figure A.1: Lattice Scheme used in this work.

so assuming $(\alpha\mathbf{X}) \mod \Lambda$ to be uniformly distributed over $\mathcal{V}(\Lambda)$, the power of the watermark is now

$$D_w = \frac{1}{L_2}\frac{\int_{\mathcal{V}} ||x||^2 dx}{\mathrm{Vol}(\mathcal{V})}, \tag{A.2}$$

that is, α^2 times smaller than that obtained in (A.1). Finally, the modulo reduced received signal reads as,

$$\begin{aligned}
\mathbf{Z}_{\mathrm{mod}} = \mathbf{Z} \mod \Lambda &= \Big[\alpha\big(\mathbf{W}+\mathbf{X}+\mathbf{N}\big)\Big] \mod \Lambda \\
&= \Big\{(\alpha\mathbf{N}) \mod \Lambda + \alpha\Big(\big[-\alpha\mathbf{X}+\mathbf{v}(b)\big] \mod \Lambda\Big) \\
&\quad + \Big[\alpha\mathbf{X}-\mathbf{v}(b)\Big] \mod \Lambda + \mathbf{v}(b)\Big\} \mod \Lambda \\
&= \Big\{(\alpha\mathbf{N}) \mod \Lambda + (1-\alpha)\Big(\big[\alpha\mathbf{X}-\mathbf{v}(b)\big] \mod \Lambda\Big) \\
&\quad + \mathbf{v}(b)\Big\} \mod \Lambda,
\end{aligned}$$

and its block diagram is shown in Figure A.2.

Considering these analyses, if we want to make a fair comparison between the two schemes, i.e., with both of them yielding the same watermark power, the lattice and shifting vectors $\mathbf{v}(b)$ of the first scenario should be inflated by $1/\alpha$; by doing this, and taking into account that

$$\alpha[\mathbf{X} \mod (\Lambda/\alpha)] = (\alpha\mathbf{X}) \mod \Lambda, \tag{A.3}$$

since

$$\alpha[\mathbf{X} - Q_{\Lambda/\alpha}(\mathbf{X})] = \alpha\mathbf{X} - Q_{\Lambda}(\alpha\mathbf{X}), \tag{A.4}$$

and $\alpha Q_{\Lambda/\alpha}(\mathbf{X}) = Q_{\Lambda}(\alpha\mathbf{X})$, it is straightforward to see the equivalent performance of both approaches.

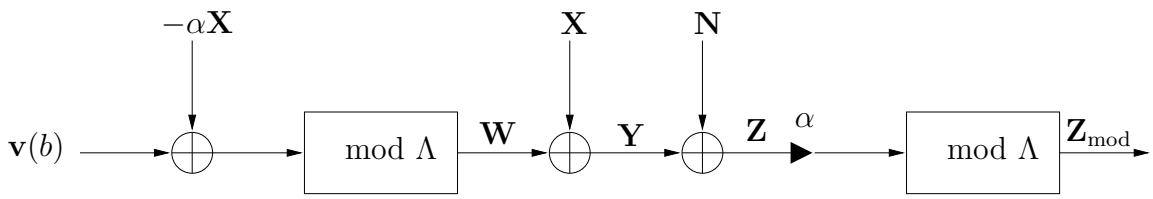


Figure A.2: Lattice Scheme used in [166, 70, 69].

Appendix B

Characteristic Function for the Beaulieu Approach under Gaussian Distortion

We derive next the characteristic function $F_{U_i^+}(u_i^+)$ required for computing P_e in front of Gaussian noise following Beaulieu's method. Let $\sigma_{G_i} \triangleq \frac{\sigma_{N_i}}{\Delta_i}$ be the standard deviation of the Gaussian attack after the normalization by Δ_i. Taking into account (2.29), (2.30) and (3.3), the pdf of U_i^+ can be written as

$$f_{U_i^+}(u_i^+) = \begin{cases} \sum_{k=-\infty}^{\infty} \frac{1}{\mu_i} \left[\mathcal{Q}\left(\frac{u_i^+ - (1-\alpha) - 2k}{\sigma_{G_i}} \right) \right. \\ \quad \left. - \mathcal{Q}\left(\frac{u_i^+ + (1-\alpha) - 2k}{\sigma_{G_i}} \right) \right], \quad \text{if } 0 \leq u_i^+ \leq 1, \\ 0, \quad \text{otherwise,} \end{cases} ,$$

with $\mu_i \triangleq \frac{1-\alpha}{\sigma_{G_i}}$. For the sake of simplicity we define $M_i \triangleq \frac{U_i^+}{\sigma_{G_i}}$, whose characteristic function is

$$\begin{aligned}
F_{M_i}(\omega) = & \int_0^{\delta_i} e^{j\omega m_i} \sum_{k=-\infty}^{\infty} \frac{1}{\mu_i} \left[\mathcal{Q}\left(m_i - \mu_i - 2k\delta_i\right) \right. \\
& - \left. \mathcal{Q}\left(m_i + \mu_i - 2k\delta_i\right) \right] dm_i = \\
& = \frac{j}{2\mu_i\omega} \left\{ - \operatorname{erf}\left(\frac{-\mu_i - 2k\delta_i}{\sqrt{2}} \right) \right. \\
& + \operatorname{erf}\left(\frac{\mu_i - 2k\delta_i}{\sqrt{2}} \right) \\
& + e^{j\delta_i\omega} \left[\operatorname{erf}\left(\frac{\delta_i - 2k\delta_i - \mu_i}{\sqrt{2}} \right) \right. \\
& - \left. \operatorname{erf}\left(\frac{\delta_i - 2k\delta_i + \mu_i}{\sqrt{2}} \right) \right] \\
& + e^{-\omega(\frac{\omega}{2} - j(\mu_i + 2k\delta_i))} \left[- \operatorname{erf}\left(\frac{\mu_i + 2k\delta_i + j\omega}{\sqrt{2}} \right) \right. \\
& + \left. \operatorname{erf}\left(\frac{-\delta_i + \mu_i + 2k\delta_i + j\omega}{\sqrt{2}} \right) \right] \\
& + e^{-\omega(\frac{\omega}{2} - j(-\mu_i + 2k\delta_i))} \left[- \operatorname{erf}\left(\frac{-\mu_i + 2k\delta_i + j\omega}{\sqrt{2}} \right) \right. \\
& + \left. \left. \operatorname{erf}\left(\frac{-\delta_i - \mu_i + 2k\delta_i + j\omega}{\sqrt{2}} \right) \right] \right\}
\end{aligned} \tag{B.1}$$

with $\delta_i \triangleq \frac{1}{\sigma_{G_i}}$. It is straightforward to see that $F_{U_i^+}(\omega) = F_{M_i}(\omega \cdot \sigma_{G_i})$. The $\operatorname{erf}(\cdot)$ function is defined as

$$\begin{aligned}
\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-t^2/2} dt = \frac{2z}{\sqrt{\pi}} M\left(\frac{1}{2}, \frac{3}{2}, -z^2\right), \\
\text{with } z \in \mathbb{C},
\end{aligned} \tag{B.2}$$

with $M(\cdot,\cdot,\cdot)$ the Kummer confluent hypergeometric function of the first kind. The evaluation of (B.1) presents numerical problems due to the evaluation of (B.2), which is computed as

$$\begin{aligned}
\operatorname{erf}(x + jy) \approx & \operatorname{erf}(x) + \frac{\mathrm{e}^{-x^2}}{2\pi x}[(1 - \cos(2xy) + j\sin(2xy)] \\
& + \frac{2}{\pi}\mathrm{e}^{-x^2} \sum_{n=1}^{\infty} \frac{\mathrm{e}^{-n^2/4}}{n^2 + 4x^2}[f_n(x,y) + j g_n(x,y)],
\end{aligned}$$

where

$$\begin{aligned}
f_n(x,y) &= 2x - 2x\cosh(ny)\cos(2xy) + n\sinh(ny)\sin(2xy), \\
g_n(x,y) &= 2x\cosh(ny)\sin(2xy) + n\sinh(ny)\cos(2xy).
\end{aligned}$$

Appendix C

BNSA explanation

In this Appendix we show that (3.42) is equivalent to

$$\arg \min_{\mathbf{s} \in \mathbb{R}^{L_1}} d^{\star}_{\mathbf{y}}(h_{\mathbf{y}}(\mathbf{s})), \tag{C.1}$$

with $d^{\star}_{\mathbf{y}}(\mathbf{t})$ the restriction of $d_{\mathbf{y}}(\mathbf{t})$ to those $\mathbf{t} \in \partial\mathcal{B}$, i.e.,

$$\begin{aligned} d^{\star}_{\mathbf{y}}(\mathbf{t}) : \partial\mathcal{B} &\rightarrow \mathbb{R}^{+} \\ \mathbf{t} &\rightarrow d_{\mathbf{y}}(\mathbf{t}), \end{aligned}$$

and $h_{\mathbf{y}}(\mathbf{s})$ is a surjection from $\mathbb{R}^{L_1}$ to $\partial\mathcal{B}$,[1] i.e., $h_{\mathbf{y}}(\mathbf{s}) : \mathbb{R}^{L_1} \rightarrow \partial\mathcal{B}$, such that $h_{\mathbf{y}}(\mathbb{R}^{L_1}) = \partial\mathcal{B}$, verifying that $h_{\mathbf{y}}(\mathbf{s}) = \mathbf{s}$ for all $\mathbf{s} \in \partial\mathcal{B}$; we will also assume that $h_{\mathbf{y}}(\mathbf{s}) \in C^2$, i.e., its second derivative exists and is continuous, in a neighborhood of $\mathbf{s}$ (this last point is related to the differentiability of $g \circ f$). Note that $h_{\mathbf{y}}(\mathbf{s})$ just maps the vector $\mathbf{s}$ to a point on $\partial\mathcal{B}$; following this approach the constraint in (3.42) is straightforwardly verified and we no longer have to care about it. In this way, if $\mathbf{t}_1^*$ is a solution to (3.42), it will verify $g \circ f(\mathbf{y} + \mathbf{t}_1^*) = \eta$, so $\mathbf{t}_1^* \in \partial\mathcal{B}$ and we can define the set of vectors $\mathcal{S}_1 \triangleq \{\mathbf{s}_1^* \in \mathbb{R}^{L_1} : h_{\mathbf{y}}(\mathbf{s}_1^*) = \mathbf{t}_1^*\}$. Taking into account that $h_{\mathbf{y}}$ is a surjection there will be at least one such vector $\mathbf{s}_1^* \in \mathcal{S}_1$, so that $d^{\star}_{\mathbf{y}}(h_{\mathbf{y}}(\mathbf{s}_1^*)) = d_{\mathbf{y}}(\mathbf{t}_1^*)$, and $\mathbf{s}_1^*$ is a solution to (C.1). On the other hand, if $\mathbf{s}_2^*$ is a solution to (C.1), we can define $\mathbf{t}_2^* = h_{\mathbf{y}}(\mathbf{s}_2^*)$, which minimizes $d^{\star}_{\mathbf{y}}(\mathbf{t})$ over $\partial\mathcal{B}$, so $\mathbf{t}_2^*$ also minimizes $d_{\mathbf{y}}(\mathbf{t})$ for all $\mathbf{t} \in \partial\mathcal{B}$, and is a solution to (3.42).

Therefore, a vector $\mathbf{s}$ is a solution to (C.1) if and only if $h_{\mathbf{y}}(\mathbf{s})$ is a solution to (3.42), in such a way that we can restrict our problem to look for a function $h_{\mathbf{y}}$ and an algorithm which finds a solution to (C.1).

[1] This means that for all $\mathbf{b} \in \partial\mathcal{B}$, there is an $\mathbf{a} \in \mathbb{R}^{L_1}$ such that $h_{\mathbf{y}}(\mathbf{a}) = \mathbf{b}$.

Appendix D

Calculation of mutual information for spread spectrum

D.1. Known Message Attack (KMA) for a single observation

For a single observation ($N_o = 1$) and $L_b = 1$, we have

$$I(\mathbf{Y}; \mathbf{S}_1|\mathbf{B}) = \sum_{i=1}^{L_1} \sum_{j=1}^{L_1} I(Y_i; S_{1,j}|\mathbf{B}, Y_{i-1}, \ldots, Y_1, S_{1,j-1}, \ldots, S_{1,1}) \tag{D.1}$$

$$= \sum_{i=1}^{L_1} I(Y_i; S_{1,i}|\mathbf{B}, Y_{i-1}, \ldots, Y_1) \tag{D.2}$$

$$= \sum_{i=1}^{L_1} I(Y_i; S_{1,i}|\mathbf{B}) \tag{D.3}$$

$$= L_1 I(Y_i; S_{1,i}|\mathbf{B}), \tag{D.4}$$

where (D.2) follows from the fact that Y_i and $S_{1,j}$ are independent for all $i \neq j$; (D.3) follows from the independence between the components of $\mathbf{Y}$ given the message, and (D.4) follows from the fact that $\mathbf{Y}$ and $\mathbf{S}_1$ are i.i.d. processes. The analytical expression for (D.4) is easy to calculate:

$$I(Y_i; S_{1,i}|\mathbf{B}) = I(Y_i; S_{1,i}|\mathbf{B} = \mathbf{0}) = h(Y_i|\mathbf{B} = \mathbf{0}) - h(Y_i|\mathbf{B} = \mathbf{0}, S_{1,i}),$$

where $h(Y_i|\mathbf{B} = \mathbf{0})$ will obviously depend on the distribution of $S_{1,i}$. Assuming $\mathbf{S}_1$ to be Gaussian, i.e. $\mathbf{S}_1 \sim \mathcal{N}(0, \sigma_S^2 \mathbf{I}_{L_1})$, we can write

$$I(Y_i; S_{1,i}|\mathbf{B}) = h(\mathcal{N}(0, \sigma_X^2 + \sigma_S^2)) - h(\mathcal{N}(0, \sigma_X^2)) = \frac{1}{2} \log \left(1 + \frac{\sigma_S^2}{\sigma_X^2}\right).$$

Next, the case of multiple carriers is analyzed. When $L_b > 1$, we can write

$$\begin{aligned} I(\mathbf{Y}; \mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{L_b}|\mathbf{B}) &= L_1 I(Y_i; S_{1,i}, S_{2,i}, \ldots, S_{L_b,i}|\mathbf{B}) \\ &= L_1 \left\{h(Y_i|\mathbf{B}) - h(Y_i|S_{1,i}, \ldots, S_{L_b,i}, \mathbf{B})\right\} \\ &= L_1 \left\{h\left(X_i + \textstyle\sum_{j=1}^{L_b} (L_b)^{-1/2} S_{j,i}\right) - h(X_i)\right\} \\ &= L_1 \left\{h(\mathcal{N}(0, \sigma_X^2 + \sigma_S^2)) - h(\mathcal{N}(0, \sigma_X^2))\right\} \\ &= \frac{L_1}{2} \log\left(1 + \frac{\sigma_S^2}{\sigma_X^2}\right). \end{aligned} \tag{D.5}$$

D.2. Known Message Attack (KMA) for multiple observations

When $L_1 = 1$, there are several available observations ($N_o > 1$) watermarked with the same secret key and there is one embedded bit for each observation ($L_b = 1$) which we will assume without loss of generality to be the same for all the observations,[1] it can be seen that the covariance matrix of $(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o})$, denoted by $R_{\mathbf{Y}}$, becomes

$$R_{\mathbf{Y}} = \begin{pmatrix} \sigma_X^2 + \sigma_S^2 & \sigma_S^2 & \cdots & \sigma_S^2 \\ \sigma_S^2 & \sigma_X^2 + \sigma_S^2 & \cdots & \sigma_S^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_S^2 & \sigma_S^2 & \cdots & \sigma_X^2 + \sigma_S^2 \end{pmatrix},$$

so its entropy is [51]

$$h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}) = \frac{1}{2} \log\left((2\pi e)^{N_o} |R_{\mathbf{Y}}|\right) = \frac{1}{2} \log\left((2\pi e)^{N_o} \left[\frac{N_o \sigma_S^2}{\sigma_X^2} + 1\right] \sigma_X^{2N_o}\right),$$

and we can write $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{S}_1|\mathbf{B}^1, \cdots, \mathbf{B}^{N_o}) = \frac{1}{2} \log\left(1 + \frac{N_o \sigma_S^2}{\sigma_X^2}\right)$.

[1] If the embedded bits are different, it will be enough to multiply the observations by -1 when the embedded bit is 0.

Appendix E

Fisher Information Matrix for SS-KMA

In this section we will compute the Fisher Information Matrix of the estimate of the constant multiple parameter $\boldsymbol{\theta}$ taking into account the observations $\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}$. Let us consider $\mathbf{Y}^j = \mathbf{X}^j + \boldsymbol{\theta}$, with $\mathbf{X}^j \sim \mathcal{N}(0, \sigma_X^2 \mathbf{I}_{L_1})$, and the $\mathbf{X}^j$'s to be mutually independent for $1 \leq j \leq N_o$ [1]. Following the definition of Fisher Information Matrix [151], we can write

$$\text{FIM}_{ii}(\boldsymbol{\theta}) = \int f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) \left(\frac{\partial}{\partial \theta_i} \log f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) \right)^2 d\mathbf{y}^1 \cdots d\mathbf{y}^{N_o},$$

where $f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) = \prod_{k=1}^{L_1} \prod_{j=1}^{N_o} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-(y_k^j - \theta_k)^2}{2\sigma_X^2}}$, in such a way that

$$\frac{\partial}{\partial \theta_i} \log f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) = \sum_{j=1}^{N_o} \frac{y_i^j - \theta_i}{\sigma_X^2} = \frac{\sum_{j=1}^{N_o} x_i^j}{\sigma_X^2},$$

and, finally, after a change of variable,

$$\text{FIM}_{ii}(\boldsymbol{\theta}) = \int \left(\frac{\sum_{j=1}^{N_o} x_i^j}{\sigma_X^2} \right)^2 \prod_{j=1}^{N_o} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-(x_i^j)^2}{2\sigma_X^2}} dx_i^1 \cdots dx_i^{N_o} = \frac{N_o}{\sigma_X^2}, \ 1 \leq i \leq L_1.$$

[1] Be aware that this is the case described in Section 4.4.1 for $L_b = 1$, after multiplying the j-th observation by $-(-1)^{B_1^j}$. In that case, the parameter to be estimated is $\mathbf{S}_1$.

On the other hand,

$$\begin{aligned}
\mathrm{FIM}_{ik}(\boldsymbol{\theta}) &= \int f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) \left(\frac{\partial}{\partial \theta_i} \log f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) \right) \\
&\quad \left(\frac{\partial}{\partial \theta_k} \log f(\mathbf{y}^1, \cdots, \mathbf{y}^{N_o}|\boldsymbol{\theta}) \right) d\mathbf{y}^1 \cdots d\mathbf{y}^{N_o} \\
&= \left(\int \frac{\sum_{j=1}^{N_o} x_i^j}{\sigma_X^2} \prod_{j=1}^{N_o} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-(x_i^j)^2}{2\sigma_X^2}} dx_i^1 \cdots dx_i^{N_o} \right) \\
&\quad \cdot \left(\int \frac{\sum_{l=1}^{N_o} x_k^l}{\sigma_X^2} \prod_{l=1}^{N_o} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-(x_k^l)^2}{2\sigma_X^2}} dx_k^1 \cdots dx_k^{N_o} \right) = 0, \text{ for all } i \neq k,
\end{aligned}$$

so, we can conclude $\mathrm{FIM}(\boldsymbol{\theta}) = \frac{N_o}{\sigma_X^2} \mathbf{I}_{L_1}$.

Appendix F

Mutual information for a single observation in Costa's scheme

F.1. Known Message Attack (KMA)

The mutual information between the received signal and the codebook when the sent message is known by the attacker can be written as

$$I(\mathbf{Y};\mathcal{U}|B) = h(\mathbf{Y}|B) - h(\mathbf{Y}|\mathcal{U},B) = h(\mathbf{Y}) - I(\mathbf{Y};B) - h(\mathbf{Y}|\mathcal{U}_B). \quad \text{(F.1)}$$

Studying the second term, $I(\mathbf{Y};B) = h(\mathbf{Y}) - h(\mathbf{Y}|B)$, it can be seen to be 0 whenever $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}|B}(\mathbf{y}|B=b)$ for all the possible values of b. Taking into account that $\mathbf{Y} = f_1(\mathcal{U}, B, \mathbf{X})$, this will be true in several cases. For example, if $\mathcal{U}_B$ is a lattice shifted by a random variable uniform over its Voronoi region (as in [70]), since the value of that random variable is not known by the attacker, the former equality is verified and $I(\mathbf{Y};B) = 0$. This will be also the case when $\mathcal{U}_B$ is a random codebook [50]; the attacker could know exactly all the $\mathbf{u}$'s in $\mathcal{U}$, but if he/she does not know the value of B corresponding to each of them, so the best he/she can do is to apply his/her a priori knowledge about $P(B=b)$, implying $I(\mathbf{Y};B) = 0$ again; this is the scenario studied here. Nevertheless, in the general case $0 \leq I(\mathbf{Y};B) \leq I(\mathbf{Y};B|\mathcal{U})$.

To compute $h(\mathbf{Y}|\mathcal{U}_B)$ we will focus on the implementations using random codebooks. In those schemes every $\mathbf{u}$ in $\mathcal{U}_B$ has the same probability of being chosen. In order to facilitate the analysis, we will see $\mathbf{y}$ as the combination of a scaled version of $\mathbf{u}$ and a component orthogonal to $\mathbf{u}$, $\mathbf{y} = c\mathbf{u} + \mathbf{u}^{\perp}$; recalling that $\mathbf{u} = \mathbf{w} + \alpha\mathbf{x}$, we can write $\mathbf{u}^{\perp} = \mathbf{x} + \mathbf{w} - c\mathbf{w} - c\alpha\mathbf{x}$. Therefore, the value of c can be computed taking into account that $\sigma_X^2 + \sigma_W^2 = c^2(\sigma_W^2 + \alpha^2\sigma_X^2) + \sigma_X^2(1 - c\alpha)^2 + \sigma_W^2(1-c)^2$; after some trivial algebraic operations, one obtains

$$c = \frac{\sigma_W^2 + \alpha\sigma_X^2}{\sigma_W^2 + \alpha^2\sigma_X^2}.$$

Since all the variables are Gaussian, if L_1 is large enough the samples of $\mathbf{y}$ will be very close to a sphere with radius $\sqrt{L_1 \mathrm{Var}\{\mathbf{U}^\perp\}}$ centered at some $c\mathbf{u}_o$; these spheres will be disjoint if[1] $\frac{\mathrm{Var}\{\mathbf{U}^\perp\}}{c^2} < \sigma_W^2$, which is true for any DWR if $\alpha > 0.2$. If this is the case, then we can write $h(\mathbf{Y}|\mathcal{U}_B) = h(\mathbf{Y}|\mathbf{U}) + \log(|\mathcal{U}_B|)$. Concerning $\log(|\mathcal{U}_B|)$, it is easy to see that

$$|\mathcal{U}_B| \approx e^{I(\mathbf{U};\mathbf{X})} = \left(\frac{\sigma_W^2 + \alpha^2 \sigma_X^2}{\sigma_W^2} \right)^{L_1/2}. \tag{F.2}$$

On the other hand,

$$h(\mathbf{Y}|\mathbf{U}) = h(\mathbf{U}^\perp) = \frac{L_1}{2} \log \left[2\pi e \frac{(1-\alpha)^2 \sigma_W^2 \sigma_X^2}{\sigma_W^2 + \alpha^2 \sigma_X^2} \right], \tag{F.3}$$

so,

$$\begin{aligned} h(\mathbf{Y}|\mathcal{U}_B) &= \frac{L_1}{2} \log \left[2\pi e \frac{(1-\alpha)^2 \sigma_W^2 \sigma_X^2}{\sigma_W^2 + \alpha^2 \sigma_X^2} \right] + \frac{L_1}{2} \log \left[\frac{\sigma_W^2 + \alpha^2 \sigma_X^2}{\sigma_W^2} \right] \\ &= \frac{L_1}{2} \log \left[2\pi e (1-\alpha)^2 \sigma_X^2 \right]. \end{aligned} \tag{F.4}$$

Note that this value is just an upper bound when the spheres described above are not disjoint.

Finally, the information leakage is given by

$$\begin{aligned} I(\mathbf{Y};\mathcal{U}|B) &= \frac{L_1}{2} \log \left[2\pi e (\sigma_W^2 + \sigma_X^2) \right] - \frac{L_1}{2} \log \left[2\pi e (1-\alpha)^2 \sigma_X^2 \right] \\ &= \frac{L_1}{2} \log \left[\frac{\sigma_W^2 + \sigma_X^2}{(1-\alpha)^2 \sigma_X^2} \right]. \end{aligned} \tag{F.5}$$

F.2. Watermarked Only Attack (WOA)

In this case, the mutual information between the observations and the codebook is

$$\begin{aligned} I(\mathbf{Y};\mathcal{U}) &= h(\mathbf{Y}) - h(\mathbf{Y}|\mathcal{U}) = h(\mathbf{Y}) - I(\mathbf{Y};B|\mathcal{U}) - h(\mathbf{Y}|\mathcal{U},B) \\ &= h(\mathbf{Y}) - I(\mathbf{Y};B|\mathcal{U}) - h(\mathbf{Y}|\mathcal{U}_B) = I(\mathbf{Y};\mathcal{U}|B) - I(\mathbf{Y};B|\mathcal{U}). \end{aligned}$$

The only term that has not been analyzed yet is $I(\mathbf{Y};B|\mathcal{U})$, which is the reliable rate that can be reached when the codebook is known. Note that the fact of not knowing the transmitted message produces a decrease in $I(\mathbf{Y};\mathcal{U})$ equal to

[1] It can be shown that this is a sufficient, but not necessary, condition.

the transmission rate $I(\mathbf{Y}; B|\mathcal{U})$, since the increase in the uncertainty of the sent symbol complicates the attacker's work. In [50] it is shown that

$$I(\mathbf{Y}; B|\mathcal{U}) = \frac{L_1}{2} \log \left[\frac{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_N^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_N^2(\sigma_W^2 + \alpha^2\sigma_X^2)} \right]. \tag{F.6}$$

So in this case, assuming again $\alpha > 0.2$, we can write

$$\begin{aligned} I(\mathbf{Y};\mathcal{U}) &= \frac{L_1}{2} \log \left[2\pi e(\sigma_W^2 + \sigma_X^2)\right] - \frac{L_1}{2} \log \left[\frac{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_N^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_N^2(\sigma_W^2 + \alpha^2\sigma_X^2)} \right] \\ &- \frac{L_1}{2} \log \left[2\pi e(1-\alpha)^2\sigma_X^2\right] \\ &= \frac{L_1}{2} \log \left[\frac{(\sigma_W^2 + \sigma_X^2)\left\{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_N^2(\sigma_W^2 + \alpha^2\sigma_X^2)\right\}}{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_N^2)(1-\alpha)^2\sigma_X^2} \right]. \end{aligned} \tag{F.7}$$

F.3. Estimated Original Attack (EOA)

In this Appendix we compute

$$I(\mathbf{Y};\mathcal{U}|\hat{\mathbf{X}}) = h(\mathbf{Y}|\hat{\mathbf{X}}) - h(\mathbf{Y}|\mathcal{U}, \hat{\mathbf{X}}), \tag{F.8}$$

where $\hat{\mathbf{X}} \triangleq \mathbf{X} + \tilde{\mathbf{X}}$ is an estimate of $\mathbf{X}$ and $\tilde{\mathbf{X}}$ is the estimation error; $\tilde{\mathbf{X}}$ is assumed to be i.i.d. Gaussian with power σ_E^2 and independent of $\mathbf{X}$. In this way, we can write

$$h(\mathbf{Y}|\hat{\mathbf{X}}) = h(\mathbf{X} + \mathbf{W}|\mathbf{X} + \tilde{\mathbf{X}}) < h(\mathbf{W} - \tilde{\mathbf{X}}). \tag{F.9}$$

In fact, if $\sigma_X^2 >> \sigma_E^2$, $\tilde{\mathbf{X}}$ will be almost independent of $\mathbf{X} + \tilde{\mathbf{X}}$ and

$$h(\mathbf{Y}|\hat{\mathbf{X}}) \approx h(\mathbf{W} - \tilde{\mathbf{X}}) = \frac{L_1}{2} \log \left[2\pi e(\sigma_W^2 + \sigma_E^2)\right]. \tag{F.10}$$

For the rightmost term in (F.8), we can write

$$h(\mathbf{Y}|\mathcal{U}, \hat{\mathbf{X}}) = I(\mathbf{Y}; B|\mathcal{U}, \hat{\mathbf{X}}) + h(\mathbf{Y}|\mathcal{U}, B, \hat{\mathbf{X}}). \tag{F.11}$$

Adapting the achievable rate from [50] we have

$$I(\mathbf{Y}; B|\mathcal{U}, \hat{\mathbf{X}}) = I(\mathbf{U}; \mathbf{Z}|\hat{\mathbf{X}}) - I(\mathbf{U}; \mathbf{X}|\hat{\mathbf{X}}), \tag{F.12}$$

where $I(\mathbf{U}; \mathbf{Z}|\hat{\mathbf{X}}) = h(\mathbf{Z}|\hat{\mathbf{X}}) - h(\mathbf{Z}|\hat{\mathbf{X}}, \mathbf{U})$, with

$$h(\mathbf{Z}|\hat{\mathbf{X}}) = h(\mathbf{X} + \mathbf{W} + \mathbf{N}|\hat{\mathbf{X}}) \approx \frac{L_1}{2} \log(2\pi e(\sigma_E^2 + \sigma_W^2 + \sigma_N^2)), \tag{F.13}$$

where it has been assumed $\sigma_X^2 >> \sigma_E^2$. On the other hand, $\mathbf{Z}$ conditioned on $\mathbf{U}$ and $\hat{\mathbf{X}}$ will be a Gaussian variable, so the computation of its entropy is done by simply determining its variance. Therefore, we can write

$$\mathbf{Y}_{\hat{\mathbf{X}}} = c_{\hat{\mathbf{X}}}\mathbf{U}_{\hat{\mathbf{X}}} + \mathbf{U}_{\hat{\mathbf{X}}}^{\perp} \tag{F.14}$$

where the notation implies that $\hat{\mathbf{X}}$ is given. Since $\mathbf{U}_{\hat{\mathbf{X}}} = \mathbf{W} + \alpha \mathbf{X}_{\hat{\mathbf{X}}}$, $\mathrm{Var}(\mathbf{U}_{\hat{\mathbf{X}}}) = \mathrm{Var}(\mathbf{U}|\hat{\mathbf{X}}) = \mathrm{Var}(\mathbf{W} + \alpha(\mathbf{X} + \tilde{\mathbf{X}}) - \alpha\tilde{\mathbf{X}}|\mathbf{X} + \tilde{\mathbf{X}}) \approx \sigma_W^2 + \alpha^2\sigma_E^2$, where we have assumed that $\sigma_X^2 >> \sigma_E^2$; in the same way, $\mathbf{U}^{\perp}_{\hat{\mathbf{X}}} = \mathbf{X}_{\hat{\mathbf{X}}}(1 - c_{\hat{\mathbf{X}}}\alpha) + \mathbf{W}(1 - c_{\hat{\mathbf{X}}})$, so $\mathrm{Var}(\mathbf{U}^{\perp}_{\hat{\mathbf{X}}}) \approx \sigma_E^2(1 - c_{\hat{\mathbf{X}}}\alpha)^2 + \sigma_W^2(1 - c_{\hat{\mathbf{X}}})^2$. Therefore, $c_{\hat{\mathbf{X}}}$ must verify

$$\sigma_W^2 + \sigma_E^2 = c_{\hat{\mathbf{X}}}^2(\sigma_W^2 + \alpha^2\sigma_E^2) + \sigma_E^2(1 - c_{\hat{\mathbf{X}}}\alpha)^2 + \sigma_W^2(1 - c_{\hat{\mathbf{X}}})^2, \tag{F.15}$$

so

$$c_{\hat{\mathbf{X}}}^2 = \frac{\sigma_W^2 + \alpha\sigma_E^2}{\sigma_W^2 + \alpha^2\sigma_E^2}. \tag{F.16}$$

Taking this into account,

$$\mathrm{Var}(\mathbf{Z}|\hat{\mathbf{X}}, \mathbf{U}) = \mathrm{Var}(\mathbf{Y}_{\hat{\mathbf{X}}}|\mathbf{U}_{\hat{\mathbf{X}}}) + \mathrm{Var}(\mathbf{N}) = \mathrm{Var}(\mathbf{U}^{\perp}_{\hat{\mathbf{X}}}) + \mathrm{Var}(\mathbf{N}) \approx \frac{\sigma_W^2\sigma_E^2(1-\alpha)^2}{\sigma_W^2 + \alpha^2\sigma_E^2} + \sigma_N^2,$$

so we can write

$$h(\mathbf{Z}|\hat{\mathbf{X}}, \mathbf{U}) \approx \frac{L_1}{2}\log\left(2\pi e\left[\frac{\sigma_W^2\sigma_E^2(1-\alpha)^2}{\sigma_W^2 + \alpha^2\sigma_E^2} + \sigma_N^2\right]\right). \tag{F.17}$$

Going back to (F.12), we should compute

$$\begin{aligned} I(\mathbf{U}; \mathbf{X}|\hat{\mathbf{X}}) &= h(\mathbf{U}|\hat{\mathbf{X}}) - h(\mathbf{U}|\mathbf{X}, \hat{\mathbf{X}}) = h(\mathbf{W} + \alpha\mathbf{X}|\hat{\mathbf{X}}) - h(\mathbf{W} + \alpha\mathbf{X}|\mathbf{X}, \hat{\mathbf{X}}) \\ &\approx \frac{L_1}{2}\log\left(\frac{\sigma_W^2 + \alpha^2\sigma_E^2}{\sigma_W^2}\right) \end{aligned} \tag{F.18}$$

Finally, the last needed term is $h(\mathbf{Y}|\mathcal{U}, B, \hat{\mathbf{X}})$. Under the same assumption made in Appendix F.1 ($\alpha > 0.2$), we obtain it as

$$h(\mathbf{Y}|\mathcal{U}, B, \hat{\mathbf{X}}) = h(\mathbf{Y}|\mathcal{U}_B, \hat{\mathbf{X}}) = h(\mathbf{Y}|\mathbf{U}, \hat{\mathbf{X}}) + \log(|\mathcal{U}_{B\hat{\mathbf{X}}}|) \tag{F.19}$$

where $|\mathcal{U}_{B\hat{\mathbf{X}}}|$ is the number of centroids associated with symbol B needed to verify the watermark power restriction when $\hat{\mathbf{X}}$ is given, and $h(\mathbf{Y}|\mathbf{U}, \hat{\mathbf{X}})$ coincides with $h(\mathbf{Z}|\mathbf{U}, \hat{\mathbf{X}})$ (see (F.17)) when $\sigma_N^2 = 0$. It can be shown that

$$\log(|\mathcal{U}_{B\hat{\mathbf{X}}}|) = I(\mathbf{U}; \mathbf{X}|\hat{\mathbf{X}}) \tag{F.20}$$

which has already already been derived in (F.18). Summarizing, $I(\mathbf{Y}; \mathcal{U}|\hat{\mathbf{X}})$ will be

$$\begin{aligned} I(\mathbf{Y}; \mathcal{U}|\hat{\mathbf{X}}) &\approx \frac{L_1}{2}\log\left[2\pi e(\sigma_W^2 + \sigma_E^2)\right] - \frac{L_1}{2}\log\left[\frac{\sigma_W^2(\sigma_W^2 + \sigma_E^2 + \sigma_N^2)}{\sigma_W^2\sigma_E^2(1-\alpha)^2 + \sigma_N^2(\sigma_W^2 + \alpha^2\sigma_E^2)}\right] \\ &\quad - \frac{L_1}{2}\log\left[2\pi e(1-\alpha)^2\sigma_E^2\right] \\ &= \frac{L_1}{2}\log\left[\frac{(\sigma_W^2 + \sigma_E^2)\left\{\sigma_W^2\sigma_E^2(1-\alpha)^2 + \sigma_N^2(\sigma_W^2 + \alpha^2\sigma_E^2)\right\}}{\sigma_W^2(\sigma_W^2 + \sigma_E^2 + \sigma_N^2)(1-\alpha)^2\sigma_E^2}\right], \end{aligned} \tag{F.21}$$

which is (F.7), but replacing σ_X^2 by σ_E^2. This is explained because the uncertainty about the host signal, which makes difficult the attack, is reduced, being $\tilde{\mathbf{X}}$ (with power σ_E^2) the only unknown component (recall that in the WOA case, it was $\mathbf{X}$, with power σ_X^2). Following this idea, WOA could be also seen as a particular case of EOA where the power of the estimation error is just σ_X^2. Notice in any case that in several equations we have assumed $\sigma_X^2 >> \sigma_E^2$ to ensure the independence between $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$. Examining the final result and the equivalent one for WOA, this condition does not seem to be critical, perhaps because of the cancellation of this dependence between different terms.

Bibliography

[1] http://www.digimarc.com.

[2] http://www.nextamp.com.

[3] ITU-T Recommendation T.81. JPEG Standard.

[4] Digital Imaging and Communications in Medicine (DICOM), 2004. National Electrical Manufacturers Association.

[5] André Adelsbach and Ahmad-Reza Sadeghi. Zero-knowledge watermark detection and proof of ownership. In Ira S. Moskowitz, editor, *Information Hiding International Workshop*, volume 2137 of *Lecture Notes in Computer Science*, pages 273–288, Pittsburgh, PA, USA, April 2001. Springer.

[6] A. J. Ahumada Jr. and H. A. Peterson. Luminance-model-based DCT quantization for color image compression. In *Proceedings of SPIE*, volume 1666, pages 365–374, 1992. Human Vision, Visual Processing and Digital Display III.

[7] Adnan M. Alattar and Osama M. Alattar. Watermarking electronic text documents containing justified paragraphs and irregular line spacing. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5306 of *Security, Steganography, and Watermarking of Multimedia Contents VI*, pages 685–695, San Jose, CA, USA, January 2004. SPIE.

[8] Patrice Rondao Alface and Benoit Macq. Feature-based watermarking of 3d objects: toward robustness against remeshing and desynchronization. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 400–409, San Jose, CA, USA, January 2005. SPIE.

[9] Patrice Rondao Alface, Benoit Macq, François Cayre, Francis Schmitt, and Henri Maitre. Lapped spectral decomposition for 3D triangle mesh compression. In *IEEE International Conference on Image Processing*, volume 1, pages 781–784, Barcelona, Spain, September 2003.

[10] Ross J. Anderson and Fabien A. P. Petitcolas. On the limits of steganography. *IEEE Journal on Selected Areas in Communications*, 16(4):474–481, May 1998.

[11] Félix Balado. *Digital image data hiding using side information*. PhD thesis, Universidade de Vigo, 2003.

[12] Félix Balado, Kevin M. Whelan, Guénolé C. M. Silvestre, and Neil J. Hurley. Joint iterative decoding and estimation for side-informed data hiding. *IEEE Transactions on Signal Processing*, 53(10):4006–4019, October 2005.

[13] Mauro Barni. What is the future for watermarking? (Part I). *IEEE Signal Processing Magazine*, 20(5):55–60, September 2003.

[14] Mauro Barni. What is the future for watermarking? (Part II). *IEEE Signal Processing Magazine*, 20(6):53–57, November 2003.

[15] Mauro Barni. Effectiveness of exhaustive search and template matching against watermark desynchronization. *IEEE Signal Processing Letters*, 12(2):158–161, February 2005.

[16] Mauro Barni and Franco Bartolini. *Watermarking Systems Engineering*. Marcel Dekker, 2004.

[17] Mauro Barni, Franco Bartolini, Alessia de Rosa, and Alessandro Piva. A new decoder for the optimum recovery of nonadditive watermarks. *IEEE Transactions on Image Processing*, 10(5):755–766, May 2001.

[18] Mauro Barni, Franco Bartolini, and Teddy Furon. A general framework for robust watermarking security. *Signal Processing*, 83(10):2069–2084, February 2003.

[19] Mauro Barni, Franco Bartolini, and Alessia De Rosa. Advantages and drawbacks of multiplicative spread spectrum watermarking. In Edward J. Delp III and Ping W. Wong, editors, *Proceeding of SPIE*, volume 5020 of *Security and Watermarking of Multimedia Contents V*, pages 290–299, Santa Clara, CA, USA, January 2003. SPIE.

[20] Franco Bartolini, Mauro Barni, and Alessandro Piva. Performance analysis of ST-DM watermarking in presence of nonadditive attacks. *IEEE Transactions on Signal Processing*, 52(10):2965–2974, October 2004.

[21] Norman C. Beaulieu. An infinite series for the computation of the complementary probability distribution function of a sum of independent random variables and its application to the sum of Rayleigh random variables. *IEEE Transactions on Communications*, 38(9):1463–1474, September 1990.

[22] Sergio Benedetto, Dariush Divsalar, Guido Montorsi, and Fabrizio Pollara. Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding. *IEEE Transactions on Information Theory*, 44(3):909–926, May 1998.

[23] Claude Berrou, Alain Glavieux, and Punya Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. In *IEEE International Conference on Communications*, pages 1064–1070, May 1993.

[24] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

[25] Jeffrey A. Bloom, Ingemar J. Cox, Ton Kalker, Jean-Paul M. G. Linnartz, Matthew L. Miller, and C. Brendan S. Traw. Copy protection for DVD video. *Proceedings of the IEEE*, 87(7):1267–1276, July 1999.

[26] Andrew P. Bradley. A wavelet visible difference predictor. *IEEE Trans. on Image Processing*, 8(5):717–730, May 1999.

[27] Christian Cachin. An information-theoretic model for steganography. In David Aucsmith, editor, *Information Hiding International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 306–318, Portland, OR, USA, April 1998. Springer.

[28] Roberto Caldelli, Alessandro Piva, Mauro Barni, and Andrea Carboni. Effectiveness of ST-DM watermarking against intra-video collusion. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 158–170, Siena, Italy, September 2005. Springer.

[29] Patrizio Campisi, Marco Carli, Gaetano Giunta, and Alessandro Neri. Blind quality assessment system for multimedia communications using tracing watermarking. *IEEE Transactions on Signal Processing*, 51(4):996–1002, April 2003.

[30] François Cayre, Caroline Fontaine, and Teddy Furon. Watermarking security: theory and practice. *IEEE Transactions on Signal Processing*, 53(10):3976–3987, October 2005.

[31] Brian Chen and Gregory W. Wornell. Achievable performance of digital watermarking systems. In *IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 13–18, June 1999.

[32] Brian Chen and Gregory W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47(4):1423–1443, May 2001.

[33] Jim Chou, Kannan Ramchandran, and Antonio Ortega. Next generation techniques for robust and imperceptible audio data hiding. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1349–1352, May 2001.

[34] Aaron S. Cohen and Amos Lapidoth. The Gaussian watermarking game. *IEEE Transactions on Information Theory*, 48(6):1639–1667, June 2002.

[35] Christian Collberg and Clark Thomborson. Software watermarking: Models and dynamic embeddings. In *Principles of Programming Languages*, pages 311–324, San Antonio, TX, USA, January 1999.

[36] Christian Collberg, Clark Thomborson, and Douglas Low. A taxonomy of obfuscating transformations. Technical report, University of Auckland, 1997.

[37] Pedro Comesaña. Data hiding techniques with side information. Master's thesis, University of Vigo, September 2002. In Spanish.

[38] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. Fundamentals of data hiding security and their application to Spread-Spectrum analysis. In Mauro Barni, Jordi Herrera Joancomartí, Stefan Katzenbeisser, and Fernando Pérez-González, editors, *Information Hiding International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 146–160, Barcelona, Spain, June 2005. Springer.

[39] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. An information-theoretic framework for assessing security in practical watermarking and data hiding scenarios. In *6th International Workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Switzerland, April 2005.

[40] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. The return of the sensitivity attack. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 260–274, Siena, Italy, September 2005. Springer.

[41] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. The blind Newton sensitivity attack. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 6072 of *Security, Steganography and Watermarking of Multimedia contents VIII*, San Jose, CA, USA, January 2006. SPIE.

[42] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. Blind Newton sensitivity attack. *IEE Proceedings on Information Security*, 2006. Accepted for publication.

[43] Pedro Comesaña and Fernando Pérez-González. The impact of the cropping attack on scalar STDM data hiding. *IEEE Signal Processing Letters*, 13, June 2006.

[44] Pedro Comesaña, Fernando Pérez-González, and Félix Balado. Optimal data-hiding strategies for games with BER payoffs. In Ton Kalker, Yong M. Ro, and Ingemar J. Cox, editors, *International Workshop on Digital Watermarking*, volume 2939 of *Lecture Notes in Computer Science*, pages 101–116, Seoul, Korea, October 2003. Springer.

[45] Pedro Comesaña, Fernando Pérez-González, and Félix Balado. Optimal strategies for spread-spectrum and quantized-projection image data hiding games with BER payoffs. In *IEEE International Conference on Image Processing*, volume 1, pages 479–482, Barcelona, Spain, September 2003.

[46] Pedro Comesaña, Fernando Pérez-González, and Félix Balado. On distortion-compesated dither modulation data-hiding with repetition coding. *IEEE Transactions on Signal Processing*, 54(2):585–600, February 2006.

[47] Pedro Comesaña, Fernando Pérez-González, and Frans M. J. Willems. Applying Erez and ten Brink's dirty paper codes to data-hiding. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography and Watermarking of Multimedia contents VII*, pages 298–307, San Jose, CA, USA, January 2005. SPIE.

[48] Gabriel Domínguez Conde. Image authenticacion for video surveillance applications. Master's thesis, University of Vigo, April 2006. In Spanish.

[49] John Horton Conway and Neil J. A. Sloane. *Sphere Packings, Lattices and Groups*, volume 290 of *Comprehensive Studies in Mathematics*. Springer, 3rd edition, 1999.

[50] Max H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[51] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley series in Telecommunications, 1991.

[52] Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997.

[53] Ingemar J. Cox and Jean-Paul M. G. Linnartz. Public watermarks and resistance to tampering. In *IEEE International Conference on Image Processing*, volume 3, pages 3–6, Santa Barbara, California, USA, October 1997.

[54] Ingemar J. Cox and Jean-Paul M. G. Linnartz. Some general methods for tampering with watermarks. *IEEE Journal on Selected Areas in Communications*, 16(4):587–593, May 1998.

[55] Ingemar J. Cox and Matthew L. Miller. The first 50 years of electronic watermarking. *Journal of Applied Signal Processing*, pages 126–132, 2002.

[56] Ingemar J. Cox, Matthew L. Miller, and Jeffrey A. Bloom. *Digital Watermarking*. Morgan Kaufmann Publishers, 2002.

[57] Ingemar J. Cox, Matthew L. Miller, and Andrew L. McKellips. Watermarking as communication with side information. *Proceedings of the IEEE*, 87(7):1127–1141, July 1999.

[58] Harald Cramér. *Mathematical methods of statistics*. Landmarks on Mathematics. Princeton University Press, 1999. Reprint.

[59] Jean-François Delaigle. *Protection of Intellectual Property of Images by Perceptual Watermarking*. PhD thesis, Université catholique de Louvain, 2000.

[60] Whitfield Diffie and Martin E. Hellman. New directions in cryptography. *IEEE Transactions on Information Theory*, 22(6):644–684, November 1976.

[61] Gwenaël Doërr and Jean-Luc Dugelay. Countermeasures for collusion attacks exploiting host signal redundancy. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 216–230, Siena, Italy, September 2005. Springer.

[62] Gwenaël Doërr and Jean Luc Dugelay. Danger of low-dimensional watermarking subspaces. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 93–96, Montreal, Canada, May 2004.

[63] ECRYPT D.WVL.1. First summary report on fundamentals. Technical report, ECRYPT-Network of Excellence in Cryptology, 2004.

[64] Joachim Eggers and Bernd Girod. *Informed Watermarking*. Kluwer Academic Publishers, 2002.

[65] Joachim J. Eggers, Robert Bäuml, Roman Tzschoppe, and Bern Girod. Scalar Costa Scheme for information embedding. *IEEE Transactions on Signal Processing*, 51(4):1003–1019, April 2003.

[66] Joachim J. Eggers and Bernd Girod. Blind watermarking applied to image authentication. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1977–1980, May 2001.

[67] Maha El Choubassi and Pierre Moulin. New sensitivity analysis attack. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography and Watermarking of Multimedia contents VII*, pages 734–745, San Jose, CA, USA, January 2005. SPIE.

[68] Uri Erez and Stephan ten Brink. Approaching the dirty paper limit for cancelling known interference. In *41st Annual Allerton Conference on Communications, Control, and Computing*, October 2003.

[69] Uri Erez and Stephan ten Brink. A close-to-capacity dirty paper coding scheme. *IEEE Transactions on Information Theory*, 51(10):3417–3432, October 2005.

[70] Uri Erez and Ram Zamir. Achieving $\frac{1}{2}\log(1+\mathrm{SNR})$ on the AWGN channel with lattice encoding and decoding. *IEEE Transactions on Information Theory*, 50(10):2293–2314, October 2004.

[71] Ronald Aylmer Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368, 1922.

[72] G. David Forney, Mitchell D. Trott, and Sae-Young Chung. Sphere-bound-achieving coset codes and multilevel coset codes. *IEEE Transactions on Information Theory*, 46(3):820–850, May 2000.

[73] G. David Forney Jr. On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener. In *41st Annual Allerton Conference on Communications, Control, and Computing*, October 2003.

[74] Jessica Fridrich, Miroslav Goljan, and Rui Du. Lossless data embedding for all image formats. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 4675 of *Security and Watermarking of Multimedia Contents IV*, pages 572–583, San Jose, CA, USA, January 2002. SPIE.

[75] Teddy Furon and Pierre Duhamel. An asymmetric watermarking method. *IEEE Transactions on Signal Processing*, 51(4):981–995, April 2003.

[76] Teddy Furon et al. Security Analysis. *European Project IST-1999-10987 CERTIMARK, Deliverable D.5.5*, 2002.

[77] Teddy Furon, Benoit Macq, Neil Hurley, and Guénolé Silvestre. JANIS: Just Another N-order side-Informed watermarking Scheme. In *IEEE International Conference on Image Processing*, volume 2, pages 153–156, Rochester, NY, USA, September 2002.

[78] Teddy Furon, Ilaria Venturini, and Pierre Duhamel. An unified approach of asymmetric watermarking schemes. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 4314 of *Security and Watermarking of Multimedia Contents III*, pages 269–279, San Jose, CA, USA, January 2001. SPIE.

[79] Robert G. Gallager. *Low-Density Parity-Check Codes*. PhD thesis, M.I.T., 1963.

[80] Bernd Girod. *What's Wrong with Mean-Squared Error?*, chapter 15, pages 207–220. MIT Press, 1993. in Digital images and human vision.

[81] Anil K. Goteti and Pierre Moulin. QIM watermarking games. In *IEEE International Conference on Image Processing*, volume 2, pages 717–720, Singapore, October 2004.

[82] Anil K. Goteti and Pierre Moulin. Two private, perceptual data-hiding games. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 373–376, Montreal, Canada, May 2004.

[83] Hongmei Gou and Min Wu. Data hiding in curves with application to fingerprinting maps. *IEEE Transactions on Signal Processing*, 53(10):3988–4005, October 2005.

[84] Christian Grothoff, Krista Grothoff, Ludmila Alkhutova, Ryan Stutsman, and Mikhail Atallah. Translation-Based Steganography. In Mauro Barni, Jordi Herrera Joancomartí, Stefan Katzenbeisser, and Fernando Pérez-González, editors, *Information Hiding International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 219–233, Barcelona, Spain, June 2005. Springer.

[85] Gaëtan Le Guelvouit and Stéphane Pateux. Wide spread spectrum watermarking with side information and interference cancellation. In Edward J. Delp III and Ping W. Wong, editors, *Proceeding of SPIE*, volume 5020 of *Security and Watermarking of Multimedia Contents V*, pages 278–289, Santa Clara, CA, USA, January 2003. SPIE.

[86] Gaëtan Le Guelvouit, Stéphane Pateux, and Christine Guillemot. Information-theoretic resolution of perceptual wss watermarking of non i.i.d. gaussian signals. In *European Signal Processing Conference*, volume 1, pages 454–457, Toulouse, France, September 2002.

[87] Gaëtan Le Guelvouit, Stéphane Pateux, and Christine Guillemot. Perceptual watermarking of non i.i.d. signals based on wide spread spectrum using side information. In *IEEE International Conference on Image Processing*, volume 3, pages 477–480, Rochester, NY, USA, September 2002.

[88] Joachim Hagenauer, Elke Offer, and Lutz Papke. Iterative decoding of binary block and convolutional codes. *IEEE Transactions on Information Theory*, 42(2):429–445, March 1996.

[89] Frank Hartung and Bernd Girod. Fast public-key watermarking of compressed video. In *IEEE International Conference on Image Processing*, volume 1, pages 528–531, Santa Barbara, CA, USA, October 1997.

[90] Frank Hartung and Martin Kutter. Multimedia watermarking techniques. *Proceedings of the IEEE*, 87(7):1079–1107, July 1999.

[91] Juan R. Hernández and Fernando Pérez-González. Throwing more light on image watermarks. In David Aucsmith, editor, *Information Hiding International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 191–207, Portland, OR, USA, April 1998. Springer.

[92] Juan R. Hernández, Fernando Pérez-González, José M. Rodríguez, and Gustavo Nieto. Performance analysis of a 2D-multipulse amplitude modulation scheme for data hiding and watermarking of still images. *IEEE Journal on Selected Areas on Communications*, 16(4):510–524, May 1998.

[93] Juan Ramón Hernández, Martín Amado, and Fernando Pérez-González. DCT-Domain Watermarking Techniques for Still Images: Detector Performance Analysis and a New Structure. *IEEE Transactions on Image Processing*, 9(1):55–68, January 2000.

[94] José Herskovits. Feasible direction interior-point technique for nonlinear optimization. *Journal of optimization theory and applications*, 1998.

[95] David Kahn. *The Comprehensive History of Secret Communication from Ancient Times to the Internet*. Scribner, 1996.

[96] Ton Kalker. Considerations on watermarking security. In *IEEE International Workshop on Multimedia Signal Processing*, pages 201–206, Cannes, France, October 2001.

[97] Ton Kalker, Jean-Paul Linnartz, and Marten van Dijk. Watermark estimation through detector analysis. In *IEEE International Conference on Image Processing*, volume 1, pages 425–429, Chicago, IL, USA, October 1998.

[98] Ton Kalker and Frans M. J. Willems. Capacity bounds and constructions for reversible data-hiding. In *14th International Conference on Digital Signal Processing*, volume 1, pages 71–76, 2002.

[99] Auguste Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 9:5–38, January 1883.

[100] Joe Killian, F. Thomson Leighton, Lesley R. Matheson, Talal G. Shamoon, and Robert E. Tarjan. Resistance of watermarked documents to collusional attacks. Technical report, NEC Research Institute, Princeton, NJ, 1997.

[101] Min-Su Kim, Sébastien Valette, Ho-Youl Jung, and Rémy Prost. Watermarking of 3d irregular meshes based on wavelet multiresolution analysis. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 244–259, Siena, Italy, September 2005. Springer.

[102] Darko Kirovski, Marcus Peinado, and Fabien A. P. Petitcolas. Digital rights management for digital cinema. In *Security in Imaging: Theory and Applications, International Symposium on Optical Science and Technology*, July 2001.

[103] Darko Kirovski and Fabien A. Petitcolas. Blind pattern matching attack on watermarking systems. *IEEE Transactions on Signal Processing*, 51(4):1045–1053, April 2003.

[104] Negar Kiyavash and Pierre Moulin. Regular simplex fingerprints and their optimality properties. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 97–109, Siena, Italy, September 2005. Springer.

[105] Alan J. Larkin, Félix Balado, Neil J. Hurley, and Guenolé C.M. Silvestre. Dither modulation watermarking of dynamic memory traces. In Mauro Barni, Jordi Herrera Joancomartí, Stefan Katzenbeisser, and Fernando Pérez-González, editors, *Information Hiding International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 372–386, Barcelona, Spain, June 2005. Springer.

[106] Avi Levy and Neri Merhav. An image watermarking scheme based on information theoretic principles. Technical report, HP Labs, 2001. Available at *http://www.hpl.hp.com/techreports/2001/HPL-2001-13.html*.

[107] Jae S. Lim. *Two-dimensional signal and image processing*. Prentice Hall, 1990.

[108] Shu Lin and Daniel Costello. *Error Control Coding: Fundamentals and Applications*. Prentice-Hall, 2004.

[109] Jean-Paul M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In David Aucsmith, editor, *Information Hiding International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 258–272, Portland, OR, USA, April 1998. Springer.

[110] Tie Liu and Pierre Moulin. Error exponents for one-bit watermarking. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 65–68, April 2003.

[111] Mohamed F. Mansour and Ahmed H. Tewfik. LMS-based attack on watermark public detectors. In *IEEE International Conference on Image Processing*, volume 3, pages 649–652, Rochester, NY, USA, September 2002.

[112] Andrew L. McKellips and Sergio Verdú. Worst case additive noise of binary-input channels and zero-threshold detection under constraints of power and divergence. *IEEE Transactions on Information Theory*, 43(4):1256–1264, July 1997.

[113] Andrew L. McKellips and Sergio Verdú. Maximin performance of binary-input channels with uncertain noise distributions. *IEEE Transactions on Information Theory*, 44(3):947–972, May 1998.

[114] Neri Merhav. An information-theoretic view of watermark embedding-detection and geometric attacks. Barcelona, Spain, June 2005. WaCha.

[115] Matt L. Miller, Gwenaël J. Doërr, and Ingemar J. Cox. Applying informed coding and embedding to design a robust high-capacity watermark. *IEEE Transactions on Image Processing*, 13(6):792–807, June 2004.

[116] Thomas Mitthelholzer. An information-theoretic approach to steganography and watermarking. In Andreas Pfitzmann, editor, *Information Hiding International Workshop*, volume 1768 of *Lecture Notes in Computer Science*, pages 1–16, Dresden, Germany, September 1999. Springer.

[117] Pierre Moulin and Anil K. Goteti. Minmax strategies for QIM watermarking subject to attacks with memory. In *IEEE International Conference on Image Processing*, volume 1, pages 985–988, Genoa, Italy, September 2005.

[118] Pierre Moulin, Anil K. Goteti, and Ralf Koetter. Optimal sparse-QIM codes for zero-rate blind watermarking. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 73–76, Montreal, Canada, May 2004.

[119] Pierre Moulin and Aleksandar Ivanovic. The zero-rate spread-spectrum watermarking game. *IEEE Transactions on Signal Processing*, 51(4):1098–1117, April 2003.

[120] Pierre Moulin and Ralph Koetter. Data-hiding codes. *Proceedings of the IEEE*, 93(12):2083–2126, December 2005.

[121] Pierre Moulin and M. Kivanç Mihçak. The parallel-Gaussian watermarking game. *IEEE Transactions on Information Theory*, 50(2):272–289, February 2004.

[122] Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, March 2003.

[123] Steven J. Murdoch and Stephen Lewis. Embedding covert channels into TCP/IP. In Mauro Barni, Jordi Herrera Joancomartí, Stefan Katzenbeisser, and Fernando Pérez-González, editors, *Information Hiding International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 247–261, Barcelona, Spain, June 2005. Springer.

[124] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 1999.

[125] Stéphane Pateux and Gaëtan Le Guelvouit. Practical watermarking scheme based on wide spread spectrum and game theory. *Signal Processing: Image Communication*, 18(4):283–296, April 2003.

[126] Luis Pérez-Freire, Pedro Comesaña, and Fernando Pérez-González. Detection in quantization-based watermarking: Performance and security issues. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 721–733, San Jose, CA, USA, January 2005. SPIE.

[127] Luis Pérez-Freire, Pedro Comesaña, and Fernando Pérez-González. Information-theoretic analysis of security in side-informed data hiding. In Mauro Barni, Jordi Herrera Joancomartí, Stefan Katzenbeisser, and Fernando Pérez-González, editors, *Information Hiding International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 131–145, Barcelona, Spain, June 2005. Springer.

[128] Luis Pérez-Freire, Fernando Pérez-González, Teddy Furon, and Pedro Comesaña. Security of lattice-based data hiding against the known message attack. *IEEE Transactions on Information Forensics and Security*, 2006. Accepted for publication.

[129] Luis Pérez-Freire, Fernando Pérez-González, and Sviatoslav Voloshinovskiy. An accurate analysis of scalar quantization-based data-hiding. *IEEE Transactions on Information Forensics and Security*, 1(1):80–86, March 2006.

[130] Fernando Pérez-González. The importance of aliasing in structured quantization index modulation data hiding. In Ton Kalker, Yong M. Ro, and Ingemar J. Cox, editors, *International Workshop on Digital Watermarking*, volume 2939 of *Lecture Notes in Computer Science*, pages 1–17, Seoul, Korea, October 2003. Springer.

[131] Fernando Pérez-González and Félix Balado. Nothing but a kiss: A novel and accurate approach to assessing the performance of multidimensional

distortion-compensated dither modulation. In Fabien A. P. Petitcolas, editor, *Information Hiding International Workshop*, volume 2578 of *Lecture Notes in Computer Science*, pages 87–105, Noorwijkerhout, The Netherlands, October 2002. Springer.

[132] Fernando Pérez-González, Félix Balado, and Juan R. Hernández. Performance analysis of existing and new methods for data hiding with known-host information in additive channels. *IEEE Transactions on Signal Processing*, 51(4):960–980, April 2003. Special Issue "Signal Processing for Data Hiding in Digital Media & Secure Content Delivery".

[133] Fernando Pérez-González, Pedro Comesaña, and Félix Balado. Dither-modulation data hiding with distortion-compensation: Exact performance analysis and an improved detector for JPEG attacks. In *IEEE International Conference on Image Processing*, volume 1, pages 503–506, Barcelona, Spain, September 2003.

[134] Fernando Pérez-González, Carlos Mosquera, Mauro Barni, and Andrea Abrardo. Rational dither modulation: a high-rate data-hiding method invariant to gain attack. *IEEE Transactions on Signal Processing*, 53(10):3960–3975, October 2005.

[135] Fabien A. P. Petitcolas. Watermarking schemes evaluation. *IEEE Signal Processsing Magazine*, 17(5):58–64, September 2000.

[136] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding-a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999.

[137] Alessandro Piva, Franco Bartolini, Iuve Coppini, Alessia De Rosa, and Elena Tamurini. Analysis of data hiding technologies for medical images. In *Procedings of SPIE*, volume 5020 of *Security and Watermarking of Multimedia Contents V*, pages 379–390, Santa Clara, CA, USA, January 2003. SPIE.

[138] Mahalingam Ramkumar and Ali N. Akansu. Capacity estimates for data hiding in compressed images. *IEEE Transactions on Image Processing*, 10(8):1252–1263, August 2001.

[139] Mahalingam Ramkumar and Ali N. Akansu. Signaling methods for multimedia steganography. *IEEE Transactions on Signal Processing*, 52(4):1100–1111, April 2004.

[140] Claude E. Shannon. Communication theory of secrecy systems. *Bell system technical journal*, 28:656–715, October 1949.

[141] Shlomo Shamai (Shitz) and Sergio Verdú. Worst-case power-constrained noise for binary-input channels. *IEEE Transactions on Information Theory*, 38(5):1494–1511, September 1992.

[142] Anelia Somekh-Baruch and Neri Merhav. On the error exponent and capacity games of private watermarking systems. *IEEE Transactions on Information Theory*, 49(3):537–562, March 2003.

[143] Anelia Somekh-Baruch and Neri Merhav. On the capacity game of public watermarking systems. *IEEE Transactions on Information Theory*, 50(3):511–524, March 2004.

[144] Jonathan K. Su, Joachim J. Eggers, and Bernd Girod. Analysis of digital watermarks subjected to optimum linear filtering and additive noise. *Signal Processing*, (81):1141–1175, 2001.

[145] Jonathan K. Su and Bernd Girod. Power-spectrum condition for energy-efficient watermarking. *IEEE Transactions on Multimedia*, 4(4):551–560, December 2002.

[146] Yong Sun, Angelos D. Liveris, Vladimir Stankovic, and Zixiang Xiong. Near-capacity dirty-paper code designs based on TCQ and IRA codes. In *IEEE International Symposium on Information Theory*, pages 184–188, Adelaide, Australia, September 2005.

[147] Mitchell D. Swanson, Mei Kobayashi, and Ahmed H. Tewfik. Multimedia data-embedding and watermarking technologies. *Proceedings of the IEEE*, 86(6):1064–1087, June 1998.

[148] Emre Topak, Sviatoslav Voloshynovskiy, Oleksiy Koval, and Thierry Pun. On security of geometrically-robust data-hiding. In *6th International Workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Switzerland, April 2005.

[149] Mercan Topkara, Cuneyt M. Taskiran, and Edward J. Delp III. Natural language watermarking. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 441–452, San Jose, CA, USA, January 2005. SPIE.

[150] Roman Tzschoppe, Robert Bäuml, Robert Fischer, Johannes Huber, and André Kaup. Additive non-Gaussian noise attacks on the Scalar Costa Scheme (SCS). In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 114–123, San Jose, CA, USA, January 2005. SPIE.

[151] Harry L. van Trees. *Detection, Estimation, and Modulation Theory.* John Wiley and Sons, 1968.

[152] Miguel Ernesto Vázquez-Méndez. *Análisis y control óptimo de problemas relacionados con la dispersión de contaminantes*. PhD thesis, Universidade de Santiago de Compostela, 1999.

[153] Ramarathnam Venkatesan, Vijay V. Vazirani, and Saurabh Sinha. A graph theoretic approach to software watermarking. In Ira S. Moskowitz, editor, *Information Hiding International Workshop*, volume 2137 of *Lecture Notes in Computer Science*, pages 157–168, Pittsburgh, PA, USA, April 2001. Springer.

[154] Ilaria Venturini. Oracle attacks and covert channels. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 171–185, Siena, Italy, September 2005. Springer.

[155] José Emilio Vila-Forcén, Sviatoslav Voloshynovskiy, Oleksiy Koval, Fernando Pérez-González, and Thierry Pun. Practical data-hiding: Additive attacks peformance analysis. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, pages 244–259, Siena, Italy, September 2005. Springer.

[156] José Emilio Vila-Forcén, Sviatoslav Voloshynovskiy, Oleksiy Koval, Fernando Pérez-González, and Thierry Pun. Quantization-based methods: Additive attacks performance analysis. *IEEE Transactions on Signal Processing*, 2005. Submitted.

[157] José Emilio Vila-Forcén, Sviatoslav Voloshynovskiy, Oleksiy Koval, Fernando Pérez-González, and Thierry Pun. Worst case additive attack against quantization-based data-hiding methods. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 5681 of *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 136–146, San Jose, CA, USA, January 2005. SPIE.

[158] José Emilio Vila-Forcén, Sviatoslav Voloshynovskiy, Oleksiy Koval, Thierry Pun, and Fernando Pérez-González. Worst Case Additive Attack against Quantization-Based Watermarking Techniques. In *IEEE Workshop on Multimedia Signal Processing*, pages 135–138, Siena, Italy, September-October 2004.

[159] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, April 2004.

[160] A. B. Watson. DCT quantization matrices visually optimized for individual images. In *Proceedings of SPIE*, volume 1913-14, pages 202–216, 1993. Human Vision, Visual Processing and Digital Display IV.

[161] Wavila. D.WVL.1 First summary report on fundamentals. Technical report, ECRYPT. European Network of Excellence in Cryptology, 2005.

[162] Stefan Winkler, Elisa Drelie Gelasca, and Touradj Ebrahimi. Toward perceptual metrics for video watermark evaluation. In Andrew G. Tescher, editor, *Proceedings of SPIE*, volume 5203 of *Applications of Digital Image Processing XXVI*, pages 371–378, San Diego, CA, USA, August 2003. SPIE.

[163] Yang Yang, Yong Sun, Vladimir Stankovic, and Zixiang Xiong. Image data hiding based on capacity-approaching dirty-paper coding. In Edward J. Delp III and Ping W. Wong, editors, *Proceedings of SPIE*, volume 6072 of *Security, Steganography and Watermarking of Multimedia contents VIII*, San Jose, CA, USA, January 2006. SPIE.

[164] Wei Yu, Arak Sutivong, David Julian, Thomas M. Cover, and Mung Chiang. Writing on colored paper. Available at *http://www.comm.toronto.edu/ weiyu/publications.html*.

[165] Ram Zamir and Meir Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, July 1996.

[166] Ram Zamir, Shlomo Shamai, and Uri Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Transactions on Information Theory*, 48(6):1250–1276, June 2002.

[167] Jan Zöllner, Hannes Federrath, Herbert Klimant, Andreas Pfitzmann, Rudi Piotraschke, Andreas Westfeld, Guntram Wicke, and Gritta Wolf. Modeling the security of steganographic systems. In David Aucsmith, editor, *Information Hiding International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 344–354, Portland, OR, USA, April 1998. Springer.