Digital Watermarking Security

Ph.D. Thesis

Author: Luis Pérez Freire
Advisor: Fernando Pérez González

Escola Técnica Superior de Enxeñeiros de Telecomunicación

Universidade de Vigo

2008

Submitted for the degree of Doctor in Telecommunications Engineering

Abstract

Traditionally, digital watermarking and data hiding technologies have been evaluated according to robustness, capacity and imperceptibility measures. In this thesis we address a new performance measure based on security considerations. In our context, security has to do with the ability of the watermarking systems to properly conceal the information that must be kept secret, such as the secret keys and the embedded messages. The results of this thesis show that traditional watermarking techniques lack, in general, good security properties. This, in conjunction with the often simplistic key management strategies, puts at risk the security of watermarking systems in many practical situations.

Our security assessment is inspired in a cryptanalytic approach. First, we propose a formal framework for assessing security, and we choose an information-theoretic security measure based on the work by Shannon in the field of secrecy systems. This information-theoretic measure is used to quantify in a precise manner the amount of secret information that is made available to an attacker upon the observation of marked contents, and it is also the key component of several useful performance bounds. A second part of this thesis is concerned with more practical security considerations: we design estimators and algorithms to exploit the security weaknesses identified in the theoretical analysis, in order to assess security from a practical point of view and compare the obtained results with the theoretical limits predicted by our security measure. In some cases, surprisingly, the devised estimators achieve performance close to the theoretical limit. In general however, it is not possible to extract all the available information with manageable complexity, so several simplifications must be made. It is also shown how attackers can exploit the non publicly available information after performing a security attack.

Both analyses, from theoretical and practical points of view, are performed for the two main groups of watermarking methods: spread spectrum and quantization-based ones. Watermarking security brings new challenges to the field, so a good understanding of the problem is a must for the design of next-generation watermarking and data hiding systems.

Agradecimientos

Al fin, el momento por el que he trabajado durante tanto tiempo se hace realidad. Han sido unos años de trabajo duro, de cambios, de madurez, de aprender... y algún que otro momento difícil. Personalmente, creo que cuando echamos la vista atrás tendemos a desechar lo malo y recordar con más cariño aquellos momentos que nos marcaron. Por eso, lo que recuerdo de esta etapa son sobre todo cosas buenas, como las personas que mi periplo investigador me ha llevado a conocer. Hay personas sin la que, sencillamente, esta tesis no habría llegado a buen fin. Otra muchas me han facilitado, de una u otra manera, mi camino hasta aquí. Para todos ellas, aquí van unas palabras de agradecimiento. Perdonadme si me olvido de alguno/a.

En primer lugar, tengo que expresar mi más sincero agradecimiento al responsable directo de que esta tesis haya tenido lugar: Fernando, gracias por creer en mí desde el principio y por enseñarme tantas cosas (no sólo de la investigación). Ya llovió desde el día en que llamé a la puerta de tu despacho para dejar mi CV; nunca habría imaginado en aquel momento lo que suponía hacer una tesis. Gracias por haberme dejado formar parte del GPSC. Es un honor estar rodeado de compañeros tan profesionales (y sobre todo, buena gente) como Roberto, Carlos y Nuria, que siempre están dispuestos a echar una mano. Gracias a Carmen por ser siempre tan atenta y facilitarnos la vida en la medida de lo posible. Quiero recordar también a Carmen García Mateo, sin la cual seguramente hoy no estaría en el mundo de la investigación.

A todos mis compañeros de trabajo en la escuela: gracias, os considero amigos a todos y cada uno de vosotros. Hacéis que el trabajo no parezca tal. Gracias al WatermarkingWorld de Vigo: Pedro, nuestro hombre fuerte del Information Theory, un excelente compañero con el que compartí parte del trabajo de esta tesis en los inicios; Juan, el “hombre tranquilo”; Gabi, siempre dispuesto a crear polémica; David, el más reciente aunque muy prometedor fichaje. Mención especial a Dani, compañero de fatigas dentro y fuera de los muros de la escuela (al final estamos condenados a ser compañeros de por vida!). También al resto de compañeros del TSC-5: Fran, Abu, Gonzalo, Sysman... no cambiéis nunca.

I will never forget the time I spent in Urbana-Champaign and the hospitality of the people I got to know there: Pierre, many thanks for kindly hosting me in your group and for your advice; Maha, thanks for helping me so many times and for making my stay more pleasant (good luck with your PhD); Tie, thanks for your help and for the insightful discussions. My best wishes for you all. Hablando de Urbana-Champaign, no podía dejar de recordar a la gente de UIBERIA: Diego, Irune, Chino, Juan, Ginger, Susana, Claudio, Letania, Antonio, Albert, Iker... gracias por ayudarme mitigar mi "morriña". Mucha suerte a todos con vuestros proyectos, espero que volvamos a vernos.

En cuanto a mi familia, ellos son los que han hecho posible todo esto. A mis padres quiero decirles muchas gracias por su apoyo incondicional (a pesar de que a mamá no le guste demasiado la faceta viajera de mi trabajo), y a mi hermano que siempre puede contar conmigo (aunque eso ya lo sabe). Y no podían faltar unas palabras para Patri: te mereces un lugar de honor en esta lista, por aguantar pacientemente y con una sonrisa todas mis historias sobre este mundillo y apoyarme siempre en todo lo que hago. Creo que tú sabes mejor que nadie lo que han significado para mí estos años.

Quiero agradecer también a toda esa gente que me encontré por el mundo de la investigación adelante, que con sus palabras me animaron a seguir en todo momento. A veces es importante recibir una palmadita en el hombro.

Soy consciente de que ahora, al escribir estas palabras, se cierra una etapa y se abre otra. De la nueva espero muchas cosas, aunque sé que tendré que trabajar duro para conseguirlas.

Marín, 15 de Junio de 2008

Contents

List of Figures

List of Algorithms

List of Examples

Acronyms and abbreviations

add-SS	Additive Spread Spectrum
BSS	Blind Source Separation
CM	Constant Modulus criterion
CLT	Central Limit Theorem
CMA	Constant Message Attack
CW	Circular Watermarking
DC-DM	Distortion Compensation - Dither Modulation
DRM	Digital Rights Management
DWR	Document to Watermark Ratio (dB)
ICA	Independent Component Analysis
i.i.d.	independent and identically distributed
ISS	Improved Spread Spectrum
KMA	Known Message Attack
LMI	Linear Matrix Inequality
MSE	Mean Squared Error
MMSE	Minimum Mean Squared Error
PCA	Principal Component Analysis
pdf	probability density function
pmf	probability mass function
OBE	Optimal Bounding Ellipsoid
OVE	Optimal Volume Ellipsoid
PSNR	Peak Signal to Noise Ratio
QIM	Quantization Index Modulation
r.v.	random variable
SCS	Scalar Costa Scheme
SME	Set Membership Estimation
ST-DM	Spread-Transform Dither Modulation
WNR	Watermark to Noise Ratio (dB)
WOA	Watermarked Only Attack
γ-SS	Attenuated Spread Spectrum

Notation

$\mathbb{R}$	Set of reals
$\mathbb{Z}$	Set of integers
$\Delta\mathbb{Z}^n$	n-dimensional cubic lattice scaled by Δ
$\mathcal{B}(\mathbf{c}, r)$	n-dimensional closed hypersphere of radius r centered in $\mathbf{c}$
α	distortion compensation parameter for lattice data hiding schemes
λ	host-rejection parameter for ISS
D_w	average embedding distortion per dimension
$E\left[g(X)\right]$	expectation of the function $g(X)$, where X is a random variable
$\Gamma(z)$	complete Gamma function
$h(\mathbf{V})$	differential entropy of the continuous random variable V
$H(\mathbf{M})$	differential entropy of the discrete random variable M
$\mathbf{I}_n$	$n \times n$ identity matrix
$I(X;Y)$	mutual information between the r.v.s X and Y
$I(X;Y\vert Z)$	mutual information between X and Y conditioned on Z
v	scalar variable
V	scalar random variable
$f(a)$	pdf of the continuous random variable A
$\mathbf{v}$	n-dimensional column vector
$\mathbf{v}^T$	transpose of the vector $\mathbf{v}$
$\hat{\mathbf{v}}$	estimate of the vector $\mathbf{v}$
$\mathbf{v} \mod \Lambda$	"modulo operation" or "modulo reduction", defined as $\mathbf{v} - Q_\Lambda(\mathbf{v})$
$\tilde{\mathbf{v}}$	$\mathbf{v} \mod \Lambda$
$P(\Lambda)$	second order moment per dimension of the lattice Λ
$Q_\Lambda(\cdot)$	nearest neighbor lattice quantizer
$\mathcal{S}_{N_o}$	feasible region for the secret dither in the KMA scenario
$\mathcal{S}_{N_o}(\mathbf{m}^{(k)})$	feasible region for the secret dither in the WOA scenario
$\mathcal{V}(\Lambda)$	Voronoi region of Λ
$\mathcal{D}_p$	set of coset leaders of a nested lattice code
$\mathcal{M}$	message alphabet
R	Embedding rate, defined as $R \triangleq \log(\vert\mathcal{M}\vert)/n$
$\mathcal{M}^{N_o}$	"message space" or set of all possible messages in N_o channel uses

vol($\mathcal{X}$)	volume of the bounded set $\mathcal{X}$
$\lvert\mathcal{X}\rvert$	cardinality of the discrete (countable) set $\mathcal{X}$
p	size of the alphabet
M_k	message embedded in the kth host vector
$\mathbf{X}, \mathbf{x}$	host signal (r.v., deterministic)
$\mathbf{Y}, \mathbf{y}$	watermarked signal (r.v., deterministic)
$\mathbf{S}$	secret spreading vector for spread-spectrum-based methods
$\mathbf{T}$	secret dither signal for lattice data hiding methods
N_o	Number of observations available to the attacker
H_{N_o}	Harmonic number
σ_V^2	variance of a scalar r.v. V
$\mathbf{\Sigma_V}$	covariance matrix of an n-dimensional r.v. $\mathbf{V}$
$\lvert\mathbf{\Sigma_V}\rvert$	determinant of the covariance matrix $\mathbf{\Sigma_V}$
$\mathbf{\Theta}$	secret key
ξ	ratio between host variance and watermark variance: $\xi \triangleq \frac{\sigma_X^2}{D_w}$
$\psi(\cdot)$	transformation applied to the secret key $\mathbf{\Theta}$
$\phi_{\mathcal{J}}(\mathbf{v})$	indicator function ($\phi_{\mathcal{J}}(\mathbf{v}) = 1$ if $\mathbf{v} \in \mathcal{J}$, and 0 otherwise)
tr($\cdot$)	trace of a matrix
$\lvert\lvert\mathbf{v}\rvert\rvert$	Euclidean norm of the vector $\mathbf{v}$
$U(\mathcal{Z})$	Uniform distribution over $\mathcal{Z}$
$\mathcal{N}(\mathbf{v}, \mathbf{\Sigma})$	Gaussian distribution with covariance matrix $\mathbf{\Sigma}$ and mean $\mathbf{v}$
$\chi^2(n, \sigma_X)$	Chi-squared distribution with n degrees of freedom
$\chi'^2(n, \mathbf{v}, \sigma_X)$	non-central Chi-squared distribution with n degrees of freedom

Chapter 1

Introduction

1.1 A brief history

Digital watermarking can be defined as the process of embedding information in a certain digital signal that acts as a "host" or "cover" signal. The invention of digital watermarking can be traced back to 1954 [90], when Hembrooke filed a patent [130] describing a method for the identification of music signals through the embedding of inaudible codes, with the objective of proving ownership. However, it was not until the early/mid-nineties when the interest in digital watermarking technologies started to grow significantly, due to the advent of Internet and the wide spread of digital contents distribution. During the first years, digital watermarking was mainly intended as a tool for protecting the intellectual property rights over digital works [213],[209],[52],[95], as originally conceived in [130]. However, the range of potential applications soon became largely increased: copy protection [51],[144], fingerprinting (traitor tracing) for piracy deterrence [53],[163],[212], metadata annotation [47],[58], covert communications (steganography) [34],[56][223],[118], monitoring of broadcast transmissions [98],[177], authentication of digital items [153],[43],[109], etc., giving rise to the wider term "information hiding" technologies. A variety of algorithms were developed for embedding information in still images [40],[132], video [128],[210], audio [127],[151], text [33],[216], natural language [211],[219], digital circuits [154], and more.

As a result of this intense activity, digital watermarking shortly became a very active field of research with a promising future, as one can see in the number of workshops, conferences and journals devoted nowadays to watermarking research. However, it was an excess of self-confidence during these first years what probably led to several remarkable failures, such as the Secure Digital Music Initiative (SDMI) [21] and the well-known SDMI challenge [20]. Some voices raised strong criticisms against watermarking technologies [131], claiming that they would never succeed. Fortunately, the watermarking community reacted in consequence [171], clarifying that digital watermarking was still an open topic and far from being a mature technology.

For the moment, digital watermarking technologies have not successfully addressed some of the challenges they were originally devised to solve, mainly in the Digital Rights Management field [27]. However, they have found other market opportunities, and there exist indeed a number of companies successfully exploiting watermarking technologies in diverse scenarios, such as broadcast monitoring, audience metering, or audio and video watermarking for forensic applications. Digimarc [7], Teletrax [12], Philips [2], Thomson [13], Cinea [3], Verimatrix [16], Verance [15], and Aquamobile [1] are some examples of companies with business models based on digital watermarking technologies. Furthermore, a number of small companies are incorporating watermarking capabilities to their video surveillance solutions, such as GeoVision [9], MediaSec [10] and TRedess [14], for instance. Recently, the Digital Watermarking Alliance (DWA) [8] has been created by the sector's main companies for promoting the adoption of digital watermarking technologies. Recent reports [152] state that content identification technologies such as digital watermarking and fingerprinting will experiment a rapid grow in the next years, probably surpassing US $500 million worldwide by 2012. The main applications will be, according to this report, in Digital Rights Management (DRM) related applications and forensics. Due to the increasing processing capabilities of handheld devices, other emerging applications such as interactive advertising (linking printed content to the web) will also enter the picture.[1]

When talking about DRM related applications, we no longer mean copy control and conditional access solutions. Indeed, the failure of traditional DRM technologies has much to do with this restrictive approach that takes into account only the interests of the industry, neglecting those of consumers. Therefore, there is a common agreement in the need of rethinking the digital content distribution and DRM models in order to achieve a better balance of interests [160]. The new models will probably be more based on content monitoring and traceability approaches than on usage restriction [38],[50],[146]. As such, watermarking and data hiding will play an important role as enabling technologies (along with content-based identification) for these new business models.

Forensic applications, on the other hand, have two main facets:

1. Fingerprinting (traitor tracing) applications, where the objective is to embed information in the digital contents in order to 1) enable their traceability and 2) link them to particular individuals. This has immediate application in transactional business models where redistribution of the contents is not allowed, in general, such as pay-TV and online music selling. Fingercasting techniques (joint decryption and watermark embedding in set top boxes) seem promising in this

[1] Interestingly, these trends had been foreseen by Cox et al. some years ago [90].

regard [30]. The digital cinema industry, through the Digital Cinema Initiative specification [142], has also planned to benefit from this kind of forensic tools, requiring the embedding of fingerprints during the public exhibitions of movies.

2. Authentication-like applications, where the aim is to check the authenticity of a digital item taking advantage of a certain watermark embedded in the content beforehand. This has a clear application when digital contents want to be used as legal evidence in a court of law.

Forensic applications are probably the best examples so far of the usefulness of watermarking technologies. The two cases below are highly illustrative.

- **Case 1: The cinema industry vs Sprague and Caridi** [4]. This case was highly publicized. Carmine Caridi, a member of the Academy of Motion Picture Arts and Sciences, leaked to a friend of his (William Sprague) several promotional DVDs (screeners) for the Academy Awards during 2003, that the latter illegally commercialized on Internet. The screeners came from major picture studios, like Sony Pictures, Universal, Warner Brothers... All of them were protected by means of digital watermarks for identifying each of the individuals receiving a promotional copy. The FBI Forensic Services detected the embedded watermark in several illegal copies, and this automatically led to the primary source of the leak: Carmine Caridi. After arresting Caridi and Sprague, they were found to be guilty and sentenced to pay US $600,000 to the cinema industry for the lost revenues.

- **Case 2: Warner Music Group** [6]. It is usual that many songs delivered as promotional copies to radio stations and related media appear on the Internet and P2P networks before the corresponding music albums are available for sale. Several years ago, Warner Music Group (WMG) resorted to digital watermarking technology in the promotional copies of one of their artists' new album. These copies were distributed to radio and TV stations all over the world, and shortly after this, the first copies started to circulate on the Internet. Fortunately for WMG, the watermarks embedded in the promotional copies permitted to identify the sources of the leak and take the pertinent legal actions.

Forensics, despite being probably one of the most promising applications of watermarking technologies, is also highly demanding from the technological point of view. In applications where forensic marking is involved, money and legal issues usually come into play. Hence, it is of utmost importance to guarantee that the watermarking techniques being used are highly reliable, in such a way that innocent users will not be incriminated (this is currently an important problem in the case of fingerprinting for large populations, for example).

In the next section we will analyze more in detail some of the issues arising from the legal application of watermarking technologies.

1.2 Legal considerations

Legal aspects are rarely taken into account when considering watermarking technologies. We can find in the literature some references [201],[32], focused on DRM applications (not specifically addressing watermarking) and the circumvention of Technological Protection Measures (TPMs). These papers are mainly concerned with the legislative framework for DRM, especially the "Digital Millennium Copyright Act" (DMCA) [26], and its foreseen influence in the development of future content distribution models. On the other hand, [94] and [160] are more specific to watermarking techniques, yet focused on DRM and proof of ownership scenarios. The considerations presented below are more oriented to forensic applications.

1.2.1 Digital images as legal evidence

It is well known that visual evidence (still pictures, videos) is often crucial for determining the result of a trial. The most recent techniques in digital photography and video allow to obtain high-quality images that can be provided as valuable proofs, but there is an inherent reasonable doubt about the authenticity of such images. Of course, it is possible to tamper with images in analog format, but it is much easier to do it with digital images (nowadays, any average user has access to powerful editing tools). In fact, there exist documented cases [48] where it has been proved that digital retouching techniques were used to incriminate innocents. Currently, it is enough to suggest that a digital image could have been maliciously modified in order to reject it as a valid proof, if there is no reliable mechanism for proving its authenticity.

1.2.2 Forensic analysis of digital images

Due to the simplicity for tampering with digital images and videos, the work of the forensic scientist as a specialist in digital imagery becomes particularly relevant when using digital images as legal evidence. The forensic scientist establishes whether a certain image is authentic or not. The most important techniques for proving authenticity are hashing, encryption and watermarking [28]. All these measures must be applied beforehand (i.e. in a preventive manner). Being conscious of these legal needs, many companies have incorporated these technological measures as an added value in their video surveillance solutions ([9],[10],[11],[5], just to reference some examples).

In spite of this, the introduction of watermarking technology in the legal machinery is not so easy. Watermarks may be seen as irrefutable tools for authenticating digital images and videos, but we can identify at least the following shortcomings:

1. At the embedding side: from the very moment that the watermark is embedded in the original content, a reasonable doubt may arise: is the watermarked image indeed the same as the original image? Obviously, the watermark must have been embedded in an imperceptible manner, but the software manipulating the images (the embedder) will be susceptible of being analyzed and certified.
2. At the decoding side: a serious concern, already raised in [94], is the need for the application of certain transformations (inversion of a geometrical attack, for example) to the image under analysis in order to properly detect/decode the embedded watermark. Moreover, the answer of the watermark detector/decoder cannot be regarded as 100% reliable. There exist always nonnull false negative and false positive probabilities, and potential decoding errors.

The above issues are even more critical in the case of medical images, where the requirements are very restrictive (cf. [180] by the English College of Radiographers). Obviously, the validity of a scientific method as legal evidence is determined in last instance by its error rate. Let us consider, for instance, the case of paternity proofs. Depending on the used method, the false positive probability varies from 0.01% (with a virtually null false negative rate) up to 1%, being DNA-based methods the most effective. An error probability of 0.01% is considered, from the legal point of view, as *"beyond any reasonable doubt"*. As for watermarking techniques, the theoretical performance of many of them are around this figure, but the theoretical performance measures are not always representative of all real scenarios.

In general, judges and lawyers are reluctant to accept as legal evidence new methods and techniques for which there are no documented legal precedents (let us recall that the introduction of digital imagery in the courts is relatively new). To be more precise, according to US laws [48], a legal evidence obtained by scientific means is admissible in court only *"if the used techniques are commonly accepted by the scientific community"* (United States vs. Kilgus 1978).[2] In this regard it is interesting to note that, albeit watermarking technologies have experimented great advances during the last decade, in general they have not been tested at such extent. This being said, it seems that we still have a long way to go before watermarking technologies gain full legal acknowledgement. The example considered below may be enlightening in this sense.

[2] Actually, a more restrictive requirement was adopted in 1993, known as Daubert (Daubert vs. Merrell Dow Pharmaceuticals, Inc. 1993).

1.2.3 An illustrative example: digital signature

A digital signature is a data string which associates a message (in digital form) with some originating entity (the "signing entity") [164]. Besides, it guarantees the integrity of the message and its non-repudiation. The digital signature relies on cryptographic technology: hash functions (SHA, Secure Hashing Algorithm) and the well known public key cryptographic scheme RSA (Rivest-Shamir-Adelman), invented in 1977 [164]. However, the first law on digital signature was not enacted until 1995 in the Utah state (US), whereas the first European countries in adopting it were Germany and Italy in 1997 [25],[19], even before the first European directive on the subject [24]. In Spain, for instance, the law on digital signature was not enforced until late 2003 [18].

In general, the massive deployment of a new technology is possible after two fundamental steps:

1. Reach of a satisfactory degree of maturity for the considered technology. In the case of digital signature we have to main blocks: on one hand the hash functions, whose more sophisticated versions can guarantee negligible collision probabilities; on the other hand, the RSA scheme which has been proved to be secure for more than two decades now. These two characteristics have made possible to create secure digital signatures.

2. A standardization process. The definition, use and implementation of a technology must be reflected on internationally recognized standards in order to make possible its deployment while guaranteing a minimum quality level. In the case of digital signature, this process began in 1991 [22], dealing with aspects of RSA and secure systems for electronic transactions, among others.

By translating the above considerations to the watermarking field, it is evident that we are for the moment immerse in the first stage of the process of adoption of the technology, working for constructing watermarking systems with provable robustness and security levels. This is not surprising as watermarking technology is quite recent. However, despite its short life, watermarking research has evolved enormously. The next section gives an overview of the major advances that contributed to arrive at this situation, helping to place this thesis in its proper context.

1.3 Milestones in digital watermarking research

The first digital watermarking and data hiding techniques (devised for digital images in the early-middle 90s) were ad-hoc algorithms based on elementary image processing manipulations. During the last years, these technologies have matured considerably.

Watermarking research no longer pertains solely to the image processing field, but it has become a complex technology at the crossroads of many technical fields, such as digital communications, statistics, information theory, optimization (game theory), perceptual analysis, etc. All these fields have played a fundamental role in the advance of watermarking technology:

- The contribution of digital communications was in the modeling of watermarking as a classical communications problem, where certain information (the watermark) was to be placed in a channel (the host document) by means of a modulator (the embedder) and recovered by a receiver (the watermark detector). The model also allowed for the inclusion of attacks to the watermarking schemes, paving the way to a rigorous analysis of the problem [39].
- The digital communications approach, together with the introduction of the statistical modeling of the problem, allowed to perform the first rigorous performance analyses, providing probabilities of false positive/negative in detection applications [134], and Bit Error Rate (BER) in data hiding applications [195].
- Information theory was crucial in establishing the fundamental limits of the considered communications problem [66], [175],[176].
- Optimization and game theory made possible the modeling of the watermarking problem as a game between embedder and attacker. The derivation of optimal strategies [173] and worst-case attacks [172] for some cases has been possible.
- The perceptual analysis is key to finding the optimal domains of embedding and the application of perceptual masks that maximize the power of the watermark without impairing the perceptual quality of the marked assets [39, Chapter 5],[89],[221].

The first milestone in watermarking research is probably the invention of spread spectrum embedding by Cox et al. [86], inspired in the well known homonymous technique in the field of digital communications [42]. Spread spectrum techniques greatly increased the robustness of the existing watermarks, withstanding lossy compressions and many intentional removal attacks. However, spread spectrum (as well as the previous watermarking techniques) were still far from being optimal, since they suffered from the so-called "host interference", which is the name given to the serious degradation of performance due to the interference introduced by the host signal.

The true inflection point in watermarking research came from the realization of watermarking as a problem of communications with side information by Cox et al. [91]. This, together with the rediscovery of the surprising Costa's result [83] by Chen and Wornell [62], gave rise to the "side-informed" paradigm. The adaptation of Costa's result (which dates back to 1983) to the watermarking field basically says that, under certain conditions, the amount of information that can be embedded in a digital content and successfully recovered by the decoder is not affected by the host interference (even when the decoder does not have access to the original content) if the embedder properly exploits the availability of the host signal. The side-informed paradigm led to the development of new and revolutionary data hiding schemes based on structured codebooks [63],[108] (aka quantization-based methods) and Trellis-based embedding [168], which significantly outperform the former spread spectrum methods in a wide range of scenarios. Among the side-informed techniques, the most deeply studied are those based on quantization, and specially those using lattice quantizers [63],[108],[174]. The idea was extended to spread spectrum schemes as well, giving raise to hybrid methods usually known as Spread Transform [63] schemes.

Other remarkable inventions and discoveries in the field have been the following:

- The use of channel coding techniques [196],[76] and resynchronization schemes [182],[36], inspired by the digital communications field.
- The invention of asymmetric and public key watermarking schemes [120],[110].
- The introduction of zero-knowledge watermark detection [31] and other protocols for secure implementation of transactional watermarking [163].
- The invariability against certain geometrical attacks, either embedding the information in invariant domains [203] or using specific codes for information embedding [197].
- More recently, the proposal and study of universal decoders for coping with geometrical transformations [170], and results on optimal embedding and detection strategies under limited resources [165],[70].

Despite the tremendous progress done during the last years, watermarking schemes have been mostly tested against conventional attacks (noise addition, lossy compression, geometrical transformations, etc.) that do not fully exploit the attacker's knowledge and resources, in general. However, as recognized in [166, Chapter 6], it is realistic to assume that next-generation attackers will devise much more sophisticated and challenging tools for attacking the watermarking schemes, *"possibly obtaining information about the algorithmic steps of the information hiding schemes"*, and trying *"to find the secrets, already having a good enough knowledge about the algorithm itself"*. A well-known class of this kind of attacks in the watermarking community are the so-called oracle attacks [156],[74],[65], which exploit the availability of a watermark detector in order to extract information about the secret parameters of the watermarking algorithm. Recently, other related attacks based on cryptographic-like weaknesses have shown their potential [41],[121],[61], so they have started to be seriously considered. The core of the work presented in this thesis is about the analysis of the latter kind of attacks, which can be cast in the framework of "watermarking security" [61],[71]. The detailed objectives are explained in the next section.

1.4 Summary and thesis objectives

As mentioned above, the robustness and fundamental communication limits of the watermarking / data hiding schemes are well studied topics. This thesis is concerned with the study of another problem in the watermarking field, which has been rarely addressed in the literature until very recently: watermarking security. Contrary to classical robustness studies, security is focused on the capabilities of the watermarking schemes for concealing the secret information. By "secret information" we mean certain information or parameters of the watermarking schemes that are supposed to remain secret to any unauthorized party, such as the secret keys used in the embedding/decoding process.

The rigorous study of watermarking security is a necessary step towards the development of mature watermarking technologies in the future, capable of addressing the requirements mentioned in Section 1.2. Thus, one of the objectives of this thesis is to propose a formal model for assessing watermarking security. We are interested in assessing, according to this model, the security of the two main groups of watermarking methods introduced above: spread spectrum and quantization-based schemes. We will check whether these watermarking schemes properly conceal the secret information or not. If it is not the case, we will quantify how much of the secret information can be disclosed by an attacker. For those cases that have shown to be not secure, we will propose practical algorithms for disclosing the secret information, in a cryptanalytic fashion. The experimental results will be also used for supporting the conclusions obtained in the theoretical analysis.

These objectives are addressed in the remaining chapters.

- Chapter 2 motivates the interest in the study of watermarking security, introduces the fundamental concepts and definitions that will be used throughout this thesis, and presents the measure for quantifying security. This formal framework has been published in two conference papers [72],[71], and one journal paper [186].

- Chapter 3 presents a theoretical security analysis of the security of spread spectrum methods. The main results reveal fundamental limits and bounds on security, providing insight into other properties such as the impact of the embedding parameters and the tradeoff between robustness and security. Moreover, this work formalizes and generalizes the work by other authors. One conference paper has been published [187], and one journal paper has been submitted [191].

- Chapter 4 studies the security of spread spectrum methods from a practical point of view, proposing estimators for extracting the secret information. Previous methods proposed by other authors are theoretically analyzed for the first time. New estimators, mainly based on the blind equalization paradigm, are proposed, analyzed and tested in a variety of scenarios, comparing their performance to that of the previous ones. The results have been included in an internal report [183] and submitted in a journal paper [191].

- Chapter 5 analyzes, for the first time in the literature, the fundamental security limits of quantization-based data hiding methods. The implementation considered for these methods is by means of nested lattice codes, which is the most widely extended approach. Surprisingly, these methods will be shown to be highly secure under some circumstances. In the other cases, their security will be quantified. The obtained results have been published so far in four conference papers [185],[192],[193],[189], and two journal papers [194],[190].

- Chapter 6 addresses the problem of practical estimation of the secret information for data hiding schemes using nested lattice codes. The proposed methods are based on the set-theoretic estimation paradigm and in tree search algorithms. The obtained results have been published in four conference papers [185],[192],[193],[189], and two journal papers [194],[190].

- Finally, Chapter 7 summarizes the main conclusions that can be extracted from this thesis and presents a series of challenges and open topics that deserve further analysis in the future.

The present thesis is the first one completely devoted to the problem of watermarking security, although the very first one in addressing the problem was Comesaña [68]. Much of the work presented here is part of the contribution by the University of Vigo to the European Network of Excellence in Cryptology (ECRYPT) [105], where watermarking technologies are considered as a fundamental component of digital media security.

The approach followed in this thesis is not focused on particular watermarking applications, but it rather looks at the fundamental security properties of the information embedding methods. As such, no formal distinction between watermarking and information hiding terms will be made. One must bear in mind that, although the approach can be extended to a variety of scenarios, it cannot be directly applied to the steganography scenario, which falls out of the scope of the present thesis. As a final remark, we clarify that no security considerations at the protocol level are done; our analysis is completely focused on the “physical layer”.

Chapter 2

A formal framework for assessing watermarking security

In this chapter, the formal model for the study of watermarking security is introduced. First, Section 2.1 presents the general model for digital data hiding that will used throughout this thesis. Section 2.2 exposes the motivations for the study of watermarking security, and provides a a formal formulation of the problem. Section 2.3 poses the need for a fundamental security measure, and justifies the selection of an information-theoretic measure. The main properties of this measure are addressed in Section 2.4. Some relevant concepts related to the chosen information-theoretic security measure are defined on Section 2.5. Finally, Section 2.6 presents the different scenarios for security assessment that will be considered in this thesis.

2.1 General theoretical model of digital data hiding

The generic model for data hiding followed in this thesis is shown in Figure 2.1. Essentially, digital data hiding can be modeled as a digital communications problem. First, the digital content to be marked undergoes a certain transformation that outputs a collection of coefficients (DCT, DWT, FFT...), which are arranged in N column vectors of length n, denoted by $\{\mathbf{x}_k,\ i = 1, \ldots, N\}$. This set of feature vectors will be termed "host vectors" or "host signals". The message to be embedded in the host vectors may be channel coded, yielding the letters $\{m_k, k = 1, \ldots, N\}$, which belong to a p-ary alphabet $\mathcal{M} = \{0, 1, \ldots, p-1\}$. The embedder modifies each host vector $\mathbf{x}_k$, yielding a marked vector $\mathbf{y}_k$, that conveys information about the corresponding

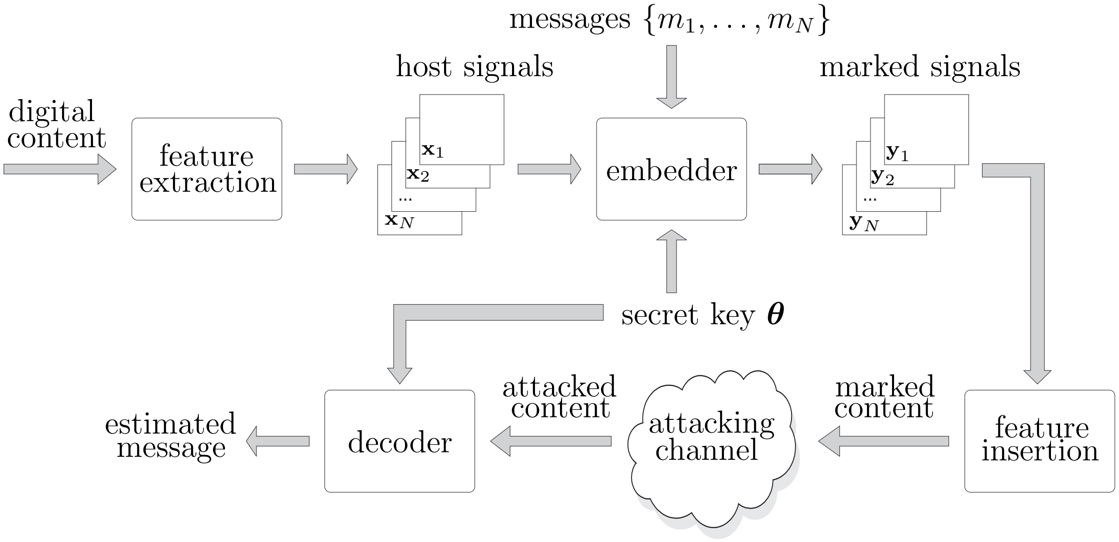


Figure 2.1: Communications model of digital data hiding

message m_k.[1] The resulting data hiding rate is

$$R = \frac{1}{n} \log_2(p)$$

bits per coefficient. In general, any embedding function can be written as

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{w}_k, \tag{2.1}$$

where the vector $\mathbf{w}_k$ is known as the "watermark" vector. In order to randomize the embedding function, the latter is made dependent on a secret key, which in Figure 2.1 is represented by the vector $\boldsymbol{\theta}$. The secret key is not directly used by the embedder; instead, it is previously transformed by means of a certain function $\psi(\cdot)$ (which can be a simple pseudorandom generator) that yields the secret parameters to be used in the embedding function.

Once the marked signal is generated, the inverse feature extraction operation is applied to the set of marked vectors $\{\mathbf{y}_k\}$ in order to construct the marked content. Thereafter, the content may be subject to attacks, which are usually modeled by means of a probabilistic channel. These attacks are sometimes mere signal processing operations applied to the marked assets that can occur during the normal lifecycle of the latter. In other cases, however, the attacks are performed by malicious users which try to get some benefit from the attack: removal of the embedded information, reading of

[1] In the embedding operation, the characteristics of the human perceptual system can be taken into account in order to minimize the perceptual impact without affecting the robustness of the watermark.

the hidden data, etc. After the possible attacks have taken place, the attacked content is presented to the decoder, which after performing the pertinent operations will produce an estimate $\{\hat{m}_k,\ k = 1, \ldots, N\}$ of the sequence of embedded messages.

In order to analyze the data hiding problem from a theoretical point of view, the signals involved in the theoretical analysis will be modeled as random variables, which will be denoted by capital letters. Instatiations of these random variables will be denoted by lowercase letters. The probability density function of a random variable X will be denoted by $f(x)$.

The embedding of information in a host signal invariably introduces a certain amount of distortion. Due to the imperceptibility requirements usually imposed, it is necessary to quantify the "embedding distortion" in a precise manner. We will define the average embedding distortion per dimension as

$$D_w \triangleq \frac{1}{n} E[||\mathbf{W}_i||^2],$$

where the operator $E[\cdot]$ denotes mathematical expectation, and $||\mathbf{x}||^2 \triangleq \sum_i x_i^2$ denotes the squared Euclidean norm of $\mathbf{x}$. In order to quantify the relative powers between the host and the watermark, we will introduce the "Document to Watermark Ratio" (DWR):

$$\text{DWR} \triangleq 10 \log_{10} \frac{\frac{1}{n} E[||\mathbf{X}_i||^2]}{D_w}. \tag{2.2}$$

Although the DWR is not necessarily the best choice for taking into account the perceptual aspects of the distortion in an accurate manner, it is nonetheless a very useful measure for assessing different embedding functions from a fair point of view, yet making the mathematical formulation of the problem affordable.

In addition to the theoretical model of digital data hiding just presented, we will add a series of assumptions aimed at simplifying the theoretical analysis in the subsequent chapters.

1. As for the host vectors $\{\mathbf{X}_i, i = 1, \ldots, l\}$, they will be assumed to be zero mean. Should this not be true, the mean of the host vectors must be subtracted so that the theoretical results shown in this thesis will hold. Furthermore, the different host vectors $\mathbf{X}_i$ are assumed to be mutually independent and identically distributed.

2. The messages $\{M_k, k = 1, \ldots, l\}$ embedded in different blocks are also assumed to be mutually independent and equiprobable in $\mathcal{M}$.

3. Although not strictly necessary, we assume that both the selection of the extracted coefficients and the partitioning in length-n blocks of the host is public, i.e. these operations do not depend on any secret parameter of the watermarking/data hiding scheme.

2.2 Motivation and statement of the problem

As mentioned in the previous section, the secret key $\boldsymbol{\theta}$ is an input to some mapping function $\psi(\cdot)$ that outputs the secret parameters of the embedding and decoding functions. The aim of such parameterization is twofold:

1. it is a way of protecting the digital contents from unauthorized embedding and/or decoding;
2. it makes the marked contents more robust to attacks.

The latter assertion is easy to see, for instance, in spread-spectrum-based methods: if the attacker ignores the secret subspace where the watermark lives, the best he can do is to perform his attack in a "random" direction of the space. However, if an accurate estimate of the spreading vector is available to the attacker, then he can put all the attacking power on the estimated subspace, so in that case the advantage brought about by spreading vanishes. Similar arguments would hold for any method that performs embedding in a secret subspace or in a secret transform domain [100]. Clearly, when evaluating attacks to data hiding systems it is important to consider the degree of knowledge about the secret key and about the data hiding scheme being attacked. Based on that amount of knowledge, the following classification of attacks to watermarking systems can be introduced.

1. **Blind attacks**. The attacker just tries to erase/modify the watermark without taking care of the secret key, even when the watermarking algorithm could be perfectly known. This is why these attacks are termed "blind". These are the kind of attacks traditionally considered in the watermarking literature concerned with robustness assessment, and they include addition of noise, compression/filtering attacks, geometric distortions, etc. However, as suggested above, if the attacker manages to gain some knowledge about the secret key, he could devise more harmful attacks. In this sense, "blind" attacks represent the most optimistic scenario for the watermarker.
2. **Attacks based on estimation of secret parameters**. When the attacker has knowledge about the data hiding scheme being attacked, he can try to obtain an estimate of $\psi(\boldsymbol{\theta})$ before attempting at attacking the system since, as discussed

above, this estimate can help him in succeeding in his task. Notice that we are talking about estimation of $\psi(\boldsymbol{\theta})$ instead of $\boldsymbol{\theta}$ itself; this is so because even when $\psi(\boldsymbol{\theta})$ and the mapping function $\psi(\cdot)$ are perfectly known, it may not be possible to recover the secret key $\boldsymbol{\theta}$, since $\psi(\cdot)$ is (or should be) designed so as not to be easily invertible. However, the knowledge of $\psi(\boldsymbol{\theta})$ is usually enough for the attacker's purposes, in general.

3. **Tampering attacks**. As mentioned above, the observation of the outputs of the embedder or the decoder only gives information about $\psi(\boldsymbol{\theta})$ but not about $\boldsymbol{\theta}$; however, the attacker can try other ways for obtaining such information. For instance, when the watermark embedder/decoder is part of an electronic device which is publicly available (such as a DVD player), the attacker can try to tamper with it in order to disclose the secret key. If the attacker manages to get perfect knowledge about $\boldsymbol{\theta}$, this implies a complete break of the watermarking system because he could perform the same actions as any authorized user.

It is clear that the first category of attacks just introduced (blind attacks) is concerned with the classical concept of robustness in watermarking and data hiding. The tampering attacks are also well known in the literature; a possible countermeasure at a hardware level against these kind of attacks is the use of tamper-proofing devices, as proposed in [35]. The second category of attacks has been started to being considered more recently and their importance should not be neglected. There are two basic scenarios where the attacker may attempt estimation of the secret parameters:

- Through the observation of the outputs of the embedder, i.e. the marked signals [41],[121]. Similarly to cryptographic scenarios, the watermarker is usually given his own secret key, that he will use repeatedly for marking images; hence, all the contents (or at least a large number) marked by the same user will contain information about the same secret key. Typically, a reliable computation of $\psi(\boldsymbol{\theta})$ will require a large number of signals watermarked with the same secret key, but once an estimate has been obtained it can be used for attacking more contents of the same user without additional effort.

- Through queries to the watermark decoder [74], [156]. In some cases the attacker may have access to a watermark decoder in the form of a black box. He can take advantage of the answers of the decoder to some chosen inputs in order to disclose information about the secret parameters of the decoding operation, such as the shape of the detection region. This kind of attacks, which are well known in the field of cryptography, receive usually the name of “oracle” or “sensitivity” attacks.

Once the attacker has estimated the secret parameters of the embedding and/or decoding function, he can exploit this knowledge in order to devise powerful attacks against the data hiding scheme which would not be possible for a "blind" attacker. The following are some examples:

1. Unauthorized embedding of messages: in copy protection scenarios, for instance, the attacker may remove the watermark inserted in a certain protected content and embed later a different message: e.g. he may change the status of a protected video from "Copy Never" to "Copy Once".

2. Unauthorized decoding of messages embedded in other contents marked with the same key.

3. Generation of forgeries: in some authentication scenarios, such as the one proposed by Eggers et al. [109], the approach is based on the embedding of a certain reference message, using a specific secret key. This means that both the secret message and the secret key remain constant for the contents marked by the same user, so an attacker with a good estimate of the latter could easily generate a forgery.

4. Complete watermark removal or "host recovery": many embedding functions are invertible if all the embedding parameters are known. This is not surprising, since the embedding process can be expressed in general as an additive function (cf. Equation 2.1). If the attacker manages to obtain a good estimate of the secret parameters of the embedding function, then he could easily estimate the actual watermark embedded in many cases. Should this be possible, the attacker would be able to recover the original host signal, completely removing the hidden data.

5. Modification of the marked signal in such a way that the decoded message is changed with minimum distortion. This could be possible, for instance, after parameterizing the decoding region with the help of an oracle attack.

The importance of properly concealing the secret parameters of the data hiding schemes becomes clear from these examples. These considerations motivate a view of the attacks to data hiding systems from a cryptographic (or better to say, "cryptanalytic") standpoint. The disclosure of the secret key (or the secret parameters) implies, roughly speaking, a "break" of the system. The term "break" is familiar in the field of cryptanalysis. Indeed, the similarities between the latter and the considered problem are numerous, and analogies between both concepts will be frequently made during this chapter. We formulate below more precisely the problem to be studied in this thesis.

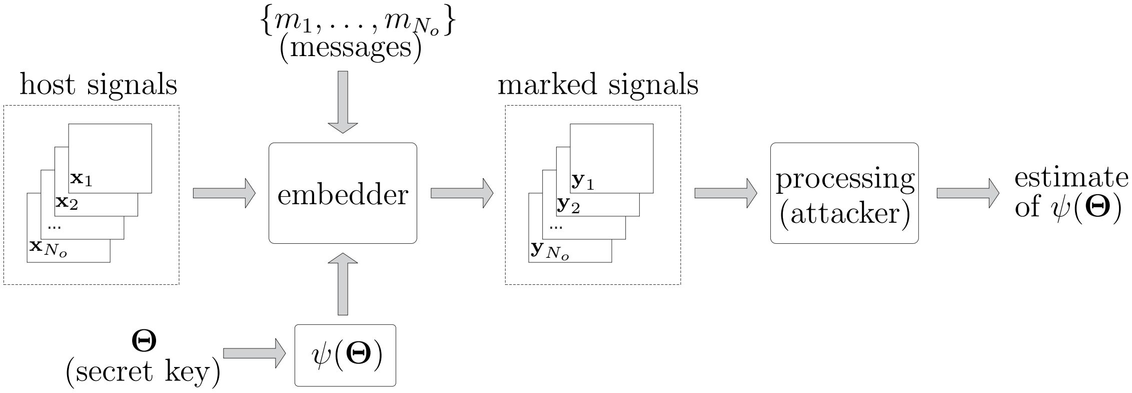


Figure 2.2: Considered model for security analysis.

2.2.1 Problem formulation

Although the concept of watermarking security is somewhat diffuse and it is still a matter of discussion, it is commonly accepted that if the secret parameters of the embedding/decoding function of a certain data hiding scheme can be estimated, then the scheme cannot be claimed to be secure. In the remainder of this thesis, we will work with the following definition in mind.

Definition 2.1. Attacks to security are those aimed at obtaining information from the secret parameters of the embedding and/or decoding functions.

From this definition, it follows that a data hiding scheme is secure if it properly conceals the secret parameters. Although this definition of security may appear too restrictive, it poses two clear advantages over previous definitions:

1. it establishes a clear frontier between robustness and security;
2. it allows to model precisely the problem of watermarking security.

The scenario considered in the remainder of this thesis is the depicted in Figure 2.2:

1. The secret key of a certain user, $\boldsymbol{\Theta}$, is reutilized N_o times for marking a set of host vectors $\{\mathbf{X}_k\}$. This is a reasonable assumption, since a user is very unlikely to change his secret key every time he watermarks a digital item; in fact, depending on the considered scenario, the value of N_o could be fairly large. Due to the assumptions on the statistical model made in Section 2.1, the marked signals $\mathbf{Y}_k$ are all conditionally independent given $\psi(\boldsymbol{\Theta})$, i.e.

$$f(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}|\psi(\boldsymbol{\Theta})) = \prod_{k=1}^{N_o} f(\mathbf{Y}_k|\psi(\boldsymbol{\Theta})). \tag{2.3}$$

2. The attacker manages to gather an ensemble of marked blocks $\{\mathbf{Y}_k,\ k=1,\ldots,N_o\}$, which may belong to different host signals, but all of them were marked with the same secret key $\boldsymbol{\Theta}$. Besides, some extra "side information" can be available to the attacker (see Section 2.6 below). The information at hand for the attacker will be termed "observations". From this ensemble of observations, he will try to extract an estimate of the secret parameters $\psi(\boldsymbol{\Theta})$ of the embedding function. We assume that the transformation $\psi(\cdot)$ is deterministic (in the sense that the same input always yields the same output). This is desirable from the point of view of the data hider, since a non-deterministic transformation would introduce an additional source of randomness for the decoder, impairing the communication. Finally, we further assume that the output of $\psi(\boldsymbol{\Theta})$ is independent both from the host signals and the embedded messages.[2]

3. As stated in Section 2.1, the procedures for extracting the features of the to-be-marked content and constructing the host vectors are publicly known, so they can be obviated in the security model. We will further assume that the remainder parameters of the embedding function are also publicly known. Hence, the only secret parameter is the secret key and its transformed version $\psi(\boldsymbol{\Theta})$. This assumption is compliant with Kerckhoffs' principle [150], well known in the field of cryptanalysis, which states that the security of a cryptosystem must be based solely on the secrecy of the secret key.

The problem of watermarking security under these assumptions will be addressed in this thesis from theoretical and practical points of view. The followed approach is essentially inspired in the field of cryptanalysis. We are interested in evaluating whether a certain data hiding scheme properly conceals the secret parameters. If it is not the case, we are also interested in quantifying how much information about the secret key leaks from the observations. For those schemes that have shown to not be secure, we will propose practical algorithms for extracting the information about $\psi(\boldsymbol{\Theta})$. Bear in mind that if the function $\psi(\cdot)$ is publicly available, the attacker may try to infer the original secret key with help of the estimate of $\psi(\boldsymbol{\Theta})$. However, such an approach completely pertains to the domain of cryptanalysis, and as such falls out of the scope of this thesis. The reader must be also aware that we will not deal with the security analysis of data hiding schemes against oracle attacks.

[2] Notice that in a more general formulation, $\psi(\boldsymbol{\Theta})$ could be made dependent of the host signal at some extent in order to add more diversity to the secret parameterization of the embedding function. Nevertheless, this strategy poses serious synchronization problems at the decoder side, and as such it is rarely used.

2.3 Tools for measuring security

From the definition of security given in Section 2.2.1, it follows that the security of a watermarking system is directly related to the difficulty in estimating the secret parameters of the embedding function from the observations at hand. Thus, a natural question is how can we quantify the hardness of such estimation problem. Let us first recall the classical criteria for evaluating the security of cryptosystems [206]:

- **Computational security**. This measure is concerned with the computational effort needed to break a given cryptosystem. If the best known algorithm for breaking the system demands a high number of computational resources, the system is said to be computationally secure. Clearly, this measure is not very useful as it is impossible to prove that a certain cryptosystem is secure (being secure against a certain attack does not imply being secure against a different class of attacks).

- **Provable security**. This measure consists in reducing the proof of security to another problem which is well understood and known to be difficult, such as the factorization of integers in prime numbers or many other combinatorial problems, which are known to be NP-complete. This definition of security is more useful than the former because it is possible to quantify the minimum effort needed to break the system. However, the drawback of this approach is that it is impossible to prove the security of a given cryptosystem in absolute terms, since the security proof is always relative to some other problem.

- **Unconditional security**. A cryptosystem is said to be "unconditionally secure" if it cannot be broken regardless the computational resources employed by the attacker.

The above criteria can be readily translated to the data hiding scenario:

- A data hiding system would be computationally secure if the computational complexity of the best known algorithm for extracting the secret parameters of the embedding function is unaffordable.

- A data hiding system would be said to be provably secure if it is proved that disclosing the randomization applied to the embedding function is a hard computational problem. The main drawback of this approach is that, unlike for classical cryptography (which deals with discrete variables), perfect disclosure of the secret parameters is not needed, but an estimate is usually enough for the attacker's purposes.

- A data hiding system is unconditionally secure if it is impossible to obtain any information about the secret parameters, regardless the computational effort by the attacker.

We will focus our approach on the third criterion. We are interested in a measuring tool capable of assessing whether a given embedding function is unconditionally secure or not. In the latter case, it would be useful if the measure could provide the amount of information that is leaked to the attacker, without resorting to computational considerations. Intuitively, a large information leakage would imply that the system is potentially less secure. In other words, we look for a measure that reveals the "fundamental" security weaknesses of a given embedding function.[3] This is why we resort to the seminal work by Shannon on the theory of secrecy systems [205], where he proposed the mutual information between the secret key and the cyphertexts for evaluating the security of cryptosystems. Assuming that both the cyphertexts and the secret key can be modeled as random variables, let $\{\mathbf{C}_k,\ k = 1, \ldots, N_o\}$ denote a set of cyphertexts generated with a secret key $\mathbf{\Theta}$. The mutual information between the cyphertexts and the secret key is defined as [84]

$$I(\mathbf{\Theta}; \mathbf{C}_1, \ldots, \mathbf{C}_{N_o}) = H(\mathbf{\Theta}) - H(\mathbf{\Theta}|\mathbf{C}_1, \ldots, \mathbf{C}_{N_o}), \tag{2.4}$$

where $H(\cdot)$ denotes the entropy function [84]. The first term in the right hand side of (2.4) is the so-called "a priori" entropy of the secret key, and it represents the total uncertainty about the key. The second term in the right hand side of (2.4) receives the name of "equivocation", and represents the remaining uncertainty or "residual entropy" about the key after the observation of cyphertexts. The mutual information quantifies in a precise manner the information leakage about the secret key provided by the observation of cyphertexts without imposing any restriction to the computational efforts of the attacker. Based on the equivocation, Shannon defines the "unicity distance" as the N_u for which $H(\mathbf{\Theta}|\mathbf{C}_1, \ldots, \mathbf{C}_{N_u}) = 0$. Therefore, the unicity distance is the number of observations needed for the attacker, in average, to reduce his uncertainty about the secret key to a unique candidate.

The adaptation of Shannon's measure to the watermarking security framework is straightforward, but one must be aware of one subtle difference: the paper by Shannon [205] deals with discrete random variables, whereas watermarking deals usually with continuous ones; this amounts to replacing the entropies in the computation of the mutual information by "differential" entropies [84]. Denoting the ith observation by $\mathbf{O}_i$, the mutual information in the data hiding scenario is defined as

$$I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_{N_o}) = h(\psi(\mathbf{\Theta})) - h(\psi(\mathbf{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N_o}). \tag{2.5}$$

[3] In this thesis we also evaluate the security from a practical point of view, but only with the purpose of confirming the identified security weaknesses, not with computational security measures in mind.

Similarly to (2.4), the first term in the right hand side of (2.5) is the "a priori" entropy of the secret parameters, and the other term is the equivocation given N_o observations. Note that we measure the mutual information between the observations and the secret parameters, which are a function of the secret key.

As suggested above, the information-theoretic model of watermarking security captures the worst case for the watermarker, because it quantifies the total amount of information about the secret parameters that is provided by each observation. An interesting question is the gap between theoretical and practical security, since the complexity of extracting all such information may be unaffordable, in general. In this sense, the mutual information $I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_{N_o})$ must be regarded to as the "achievable rate" in a hypothetical communications problem, where the information to be transmitted is $\psi(\mathbf{\Theta})$, and both the host signal and the embedded messages play the role of interfering channel. In other words, the information-theoretic analysis will provide a pessimistic bound on security, a bound which is achievable by means of an infinite computational power, in general. Of course, the watermarker will be interested in minimizing the achievable rate about the secret key while simultaneously maximizing $I(\mathbf{Y}; M|\psi(\mathbf{\Theta}))$, i.e., the achievable rate about the embedded message for a fair user.[4] For a given embedding function and embedding distortion, this can be posed as an optimization problem where the variable to be optimized is the statistical distribution of the secret key.[5] The parameters of the embedding function affect both the security and the robustness of the scheme; their influence and the relation between security and robustness can be made patent by the representation of $I(\mathbf{Y}; M|\psi(\mathbf{\Theta}))$ vs. $I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N)$, yielding a collection of "achievable regions" similar to those in classical broadcast channels [84].

The study of security from an information-theoretic point of view requires a statistical modeling of all the variables involved in the problem: the host signals, the secret key, and the embedded messages. The properties of the equivocation and the mutual information as security measures are discussed in Section 2.4.

2.3.1 How we arrived at this approach

The formulation of the watermarking security problem presented in this thesis and the choice of the information-theoretic measure are the results of the research on a continuously-evolving concept that was addressed by numerous authors during the last years.

During its infancy, digital watermarking research focused on the study of robustness

[4] $I(\mathbf{Y}; M|\psi(\mathbf{\Theta}))$ denotes the mutual information between the watermarked image $\mathbf{Y}$ and the embedded message M when the secret key $\mathbf{\Theta}$ is known.

[5] In this optimization problem, a constraint on the a priori entropy of the secret key must be imposed in order to avoid a trivial solution, such as a deterministic key.

against simple attacks such as noise addition, coarse quantization, low pass filtering, etc. Cox and Linnartz [88] were the first to establish a classification of the attacks to watermarking systems in "intentional" and "non-intentional" attacks, but without a clear notion of security yet. The pioneer in proposing a theoretical framework for security in general watermarking scenarios is probably Mittelholzer [169]. The work in [169], which is inspired on the work by Cachin [56] in the field of steganography, already introduces the concepts of equivocation and secrecy borrowed from Shannon's approach [205]. Mittelholzer considers the security problem in terms of secrecy of the embedded messages, but his model does take into account the possible information leakage about the secret key, and its reuse for marking several objects.

A first attempt at clarifying the meaning of security in watermarking is due to Kalker [145], who proposed definitions for robustness and security in the watermarking context; however, these definitions were too broad so as to be adopted in formal frameworks. As a result, the classification of attacks and the distinction between robustness and security continued to be matter of study in certain subsequent works such as [121] and [41]. The work by Furon et al. in [121] meant a great leap in the development of the theory of watermarking security due to the establishment of an analogy between watermarking security and cryptography, which had been already suggested by Hernández et al. in [133]:

- As in cryptography, watermarking systems make use of secret keys. According to Kerckhoffs' principle [150] in cryptography, if one wants to provide a quantitative measure of security, all the parameters of the system must be made public, and the security must rely only on the secret key.
- Attacks to security can be classified according to the amount of information made available to the attacker, similarly to the Diffie-Hellman classification of attacks to cryptosystems [101]. This classification decouples at a great extent the study of security from the specific watermarking application being considered.
- The "security level" of a watermarking system is said to be the effort required for disclosing the secret key.

In the work by Barni et al. [41], the restriction imposed by the Kerckhoffs' principle was relaxed with the aim of providing a wider view of the security problem. Essentially, watermarking is seen as a game with certain rules that determine what is the information about the watermarking scheme that is publicly available, giving rise to "fair" and "unfair" attacks depending on whether the attacker exploits only public information or, on the contrary, tries to access the secret information. Despite being quite general, the framework introduced in [41] is not easily applicable to real scenarios. Instead, the key ideas contained in [121] were the basis for the solid work by Cayre et al. [61], where the security of a real watermarking scheme was analyzed for the first time in a quantitative manner by resorting to the "Fisher information" [115],[84] and the statistical modeling of the watermarking problem.

Shannon's approach (i.e. using mutual information as security measure) is finally recovered for watermarking security in [71]. The main difference between [71] and [61] turns out to be the information leakage measure. The authors of [61] choose the Fisher Information instead of Shannon's mutual information by arguing that when the latter is applied to continuous random variables it makes no sense as a measure of uncertainty because it can take negative values. However, this is not a reason for discard it as an information measure, as will become apparent in Section 2.4. Furthermore, the Fisher Information presents some drawbacks, such as the computation of the Fisher information needs the existence and differentiability of the log-likelihood function of the observations given the key, precluding its application to the analysis of some practical methods (dither modulation data hiding [63], for instance); fortunately, these problems do not appear when using the mutual information measure. Nonetheless, the work in [61] constitutes the main reference for the present thesis.

Besides the information-theoretic models already introduced [169],[61],[71] there have been very few additional attempts at establishing theoretical measures of watermarking security: in fact, we are only aware of the so-called "computational" security model proposed in [148], directly inspired in the computational models commonly used in cryptography (enumerated at the beginning of this section). This model imposes a complexity constraint to the attacker in the sense that only polynomial time computations are allowed, and the security is related to the probability of successfully inferring (after an interaction between watermarker and attacker) which secret key out of two was used for watermarking a certain object. Nevertheless, as recognized in [148], the application of the computational model to existing watermarking schemes may be very difficult, and no results in this direction have been published so far. In contrast, the information-theoretic model presented in this thesis has been already applied for assessing the security of the two main classes of watermarking methods: spread-spectrum and quantization-based ones.

For further reference on this topic, the reader may want to check the recent survey by Furon [119]. Watermarking security under this viewpoint is also briefly addressed in [174, Section X].

2.4 Fundamental properties of information-theoretic measures

The properties of the security measure used in this thesis are given by those of the mutual information and differential entropy.

The differential entropy is basically a measure of the randomness of a random variable. The differential entropy of a random variable $\mathbf{V}$ with support domain $\mathcal{R}_V$ and pdf $f(\mathbf{v})$ is defined as [84]

$$h(\mathbf{V}) \triangleq -\int_{\mathcal{R}_V} f(\mathbf{v}) \cdot \log(f(\mathbf{v}))d\mathbf{v} = -E\left[\log(f(\mathbf{V}))\right]. \tag{2.6}$$

The base of the logarithm in the above definition can be arbitrarily chosen. This choice determines the units of the equivocation and of the mutual information. The most common choices are base 2 or base e. In this thesis we will work with natural logarithms (i.e. to the base e), so the equivocations and mutual informations will be always given in nats. Intuitively, the differential entropy of a continuous random variable can be interpreted as the logarithm of the volume of its typical set [84, Section 9.2]. Hence, small values of the entropy imply that the outcomes of the considered random variable are contained with high probability in a small region of space, and thus the uncertainty is small. A large entropy, on the contrary, means that the values of the random variable with significative probability are scattered over a large region. Bear in mind that, as opposed to entropy, the differential entropy can take negative values (meaning that the volume of the typical set is smaller than unity). In fact, the differential entropy of a deterministic random variable approaches $-\infty$.

In order to check the properties of the equivocation, we need to recall the concept of conditional entropy. Let $\mathbf{V}$, $\mathbf{A}$ two random variables with pdfs $f(\mathbf{v})$ and $f(\mathbf{a})$, respectively, and joint pdf $f(\mathbf{v}, \mathbf{a})$. Assume also that $\mathbf{V}$ and $\mathbf{A}$ have support domain $\mathcal{R}_V$ and $\mathcal{R}_A$, respectively. The entropy of $\mathbf{V}$ conditioned on $\mathbf{A}$ is defined as [84]

$$h(\mathbf{V}|\mathbf{A}) = \int_{\mathcal{R}_A} h(\mathbf{V}|\mathbf{A} = \mathbf{a}) f(\mathbf{a}) d\mathbf{a}, \tag{2.7}$$

where $h(\mathbf{V}|\mathbf{A} = \mathbf{a})$ is the entropy of $\mathbf{V}$ conditioned on a particular realization, $\mathbf{a}$, of the r.v. $\mathbf{A}$. It is a known result [84] that the conditioning of random variables reduces entropy. Formally, we say that

$$h(\mathbf{V}|\mathbf{A}) \leq h(\mathbf{V}). \tag{2.8}$$

The conditional entropy above is a measure of the remaining uncertainty about $\mathbf{V}$, in average, when $\mathbf{A}$ is observed. Intuitively, when considering several observations marked with the same secret key, the entropy of $\psi(\mathbf{\Theta})$ should become smaller with each observation. Recall that the equivocation has been defined as the a priori entropy minus the mutual information. The general behavior of the mutual information is formally stated in Lemma 2.2, which makes use of the result in Lemma 2.1. We first recall the definition of concavity for a real-valued function $f(x)$.

Definition 2.2 (Concavity). A function $f(x)$ defined in a real interval $[a, b]$ is said to be concave if and only if

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2), \; \forall\, x_1, x_2 \in [a,b], \text{ with } \lambda \in [0,1].$$

If strict inequality holds, then the function is said to be strictly concave.

Lemma 2.1 (Discrete concavity). A discrete function $g(n)$, with $n \in \mathbb{Z}$, is (strictly) concave if and only if $\Delta g(n)$ is (decreasing) non-increasing, with $\Delta g(n) \triangleq g(n+1) - g(n)$.

Proof: See Appendix A.1. ■

Lemma 2.2 (Concavity of the mutual information). The mutual information between $\psi(\mathbf{\Theta})$ and the observations is an increasing, concave function of the number of observations N_o.

Proof: See Appendix A.2. ■

Therefore, using the definition of equivocation and taking into account the result of this lemma, it is straightforward to see that the equivocation is a decreasing, convex function of the number of observations N_o.

Another property of the equivocation which makes it particularly useful for measuring security is that it defines a lower bound on the estimation error of the secret parameters. This is formally expressed in the following lemma.

Definition 2.3 (Variance per dimension of the estimation error). Let $\mathbf{\Sigma}_E$ denote the covariance matrix of the estimation error about the secret parameters. The variance per dimension of the estimation error is defined as

$$\sigma_E^2 \triangleq \frac{\operatorname{tr}(\mathbf{\Sigma}_E)}{n}. \tag{2.9}$$

Lemma 2.3 (Information-theoretic bound on the estimation error). The variance per dimension of the estimation error can be lower bounded as

$$\sigma_E^2 \geq \frac{1}{2\pi e} \exp\left(\frac{2}{n} h(\psi(\mathbf{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N_o})\right). \tag{2.10}$$

Proof: Let us define the estimation error as $\mathbf{e} \triangleq \psi(\boldsymbol{\theta}) - \hat{\boldsymbol{\psi}}$, where $\hat{\boldsymbol{\psi}} \triangleq \zeta(\mathbf{O}_1, \ldots, \mathbf{O}_{N_o})$ is the estimate of $\psi(\boldsymbol{\theta})$ obtained from the observations. If the covariance matrix of the estimation error is given by $\boldsymbol{\Sigma}_E$, then it is immediate to upper bound its entropy by [84, Th. 9.6.5]

$$h(\mathbf{E}) \leq \frac{1}{2}\log\left((2\pi e)^n |\boldsymbol{\Sigma}_E|\right). \tag{2.11}$$

Furthermore, note that

$$\begin{aligned} h(\mathbf{E}) &= h(\psi(\boldsymbol{\Theta}) - \hat{\boldsymbol{\psi}}) \geq h(\psi(\boldsymbol{\Theta}) - \hat{\boldsymbol{\psi}} | \mathbf{O}_1, \ldots, \mathbf{O}_{N_o}) \\ &= h(\psi(\boldsymbol{\Theta}) | \mathbf{O}_1, \ldots, \mathbf{O}_{N_o}), \end{aligned} \tag{2.12}$$

since $\hat{\boldsymbol{\psi}}$ is a function of the observations. Thus,

$$h(\psi(\boldsymbol{\Theta}) | \mathbf{O}_1, \ldots, \mathbf{O}_{N_o}) \leq \frac{1}{2}\log\left((2\pi e)^n |\boldsymbol{\Sigma}_E|\right) \leq \frac{n}{2}\log\left(2\pi e \frac{\mathrm{tr}(\boldsymbol{\Sigma}_E)}{n}\right), \tag{2.13}$$

where the second inequality follows from the fact that $|\boldsymbol{\Sigma}_E|^{\frac{1}{n}} \leq \frac{\mathrm{tr}(\boldsymbol{\Sigma}_E)}{n}$ [84, Th. 16.8.4]. Now, using Definition 2.3 and rearranging terms in (2.13), we arrive at the bound (2.10). ■

From (2.10) it can be observed that, in order to achieve an error-free estimate of the secret parameters, the equivocation must necessarily approach $-\infty$. The lower bound (2.10) is nothing but the entropy power of $\psi(\boldsymbol{\Theta})$ given N_o observations [84]. For this bound to be achievable, the estimation error must be Gaussian-distributed, although this does not necessarily occur in practice. Thus, the lower bound provided by the equivocation can be regarded as a fundamental bound that provides a pessimistic estimation of the security of a data hiding scheme. Notice also that the bound (2.10) is also valid as a bound on the mean-squared error (MSE) per dimension; in this case, the bound would be achievable if, besides having a Gaussian estimation error, the considered estimator of $\psi(\boldsymbol{\Theta})$ is unbiased.

It is interesting to note that this bound on the estimation error is similar to the well known Cramér-Rao lower bound [93], which relates the Fisher Information Matrix (FIM) to the minimum variance σ_E^2 achievable by an unbiased estimator:

$$\sigma_E^2 \geq \mathrm{tr}(\mathrm{FIM}(\psi(\boldsymbol{\Theta}))^{-1}), \tag{2.14}$$

and has been used in [61] for evaluating the security of additive spread spectrum watermarking.

2.5 More definitions based on information-theoretic measures

The mutual information and the equivocation are the basis for the definition of some fundamental concepts on the security of data hiding schemes:

Perfect secrecy: a watermarking system is said to achieve perfect secrecy (or unconditional security) whenever the observations do not provide any information about the secret parameters, that is,

$$I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N) = 0.$$

The meaning of the perfect secrecy condition, in statistical terms, is that the a posteriori probability of the observations given $\psi(\mathbf{\Theta})$ is the same as the a priori probability, i.e.

$$f(\mathbf{O}_1, \ldots, \mathbf{O}_{N_o} | \psi(\mathbf{\Theta})) = f(\mathbf{O}_1, \ldots, \mathbf{O}_{N_o}). \tag{2.15}$$

In practical terms, perfect secrecy means that all efforts by the attacker for disclosing the secret key will be useless, even if he could afford infinite computational power. Clearly, the construction of watermarking systems complying with this definition may be an extremely difficult task, or lead to unpractical systems (due to complexity requirements or length of the key, for instance).

$\varepsilon - N$ **security**: a watermarking system is said to be $\varepsilon - N$ secure if

$$I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N) \leq \varepsilon, \tag{2.16}$$

for a positive constant ε. Anyway, one must be careful with the definition of $\varepsilon - N$ security and perfect secrecy: maybe the information leakage is small (null), but this might be due to a small (null) a priori entropy of the secret key; to see this, consider the extreme case where the secret key is deterministic: in this situation, the information leakage is null, but in turn the system completely lacks security, since no secret parameterization takes place. This consideration gives rise to the notion of "security level", defined next, as a more convenient measure of security.

γ**-security level**: for those systems with $I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N) \neq 0$, the γ-security level is defined as the number of observations N_γ needed to make

$$h(\psi(\mathbf{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N_\gamma}) = h(\psi(\mathbf{\Theta})) - I(\psi(\mathbf{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N) \leq \gamma, \tag{2.17}$$

where the threshold γ (which can be negative) is established according to some criteria, as discussed below.

Unicity distance: it is defined as the number of observations N_u needed to yield a deterministic key, i.e., $h(\psi(\mathbf{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N_u}) = -\infty$.In the case of an a priori deterministic key, the unicity distance would be 0; however, it can approach ∞ in a general case, thus making useful the definition of the γ-security level. Furthermore, many attacks to the robustness can be performed without having perfect knowledge of $\psi(\mathbf{\Theta})$; instead, an accurate estimate may be enough for the attacker's purposes.

As mentioned above, it is not possible in general to construct perfectly secure watermarking systems; hence, the question is whether the achievable security levels are good enough for practical scenarios. The required security level will be determined by the specific application and the computational power of the attacker; in video watermarking, for instance, the large number of observations available [103] imposes severe restrictions in terms of security.

One of the main criticisms [148] to information-theoretic models for watermarking security is how can they be related to practical security levels, or equivalently, what should be the criteria for establishing the threshold γ in Eq. (2.17). From a practical point of view, the success of an attack based on the estimate of the secret mapping $\psi(\mathbf{\Theta})$ is closely related to the estimation error attainable by the attacker: the more reliable the estimate, the easier for the attacker to achieve his goals. Thus, it seems natural to fix the threshold γ in accordance with this estimation error. As discussed in Section 2.4, the equivocation gives a pessimistic bound on the estimation error attainable by an attacker. Hence, the definition of γ-security level can be readily used for establishing conservative security levels in the considered data hiding schemes. In this regard, it is interesting to note that the strong relation between information-theoretic and statistical measures has been recently reinforced by some works, where exact relations between mutual information and minimum mean-squared error are established for a variety of additive channels [124],[125],[181].

2.6 Evaluation scenarios

Under the Kerckhoffs' assumption, all the parameters and details of the data hiding scheme are known by the attacker, with the exception of the secret key. He must infer information about the secret key using the observations at hand. However, depending on the considered scenario, the amount and type of observations available to the attacker may vary. This gives rise to a Diffie-Hellman-like classification [101], which was applied to the watermarking field for the first time by Furon et al. [121], establishing the following scenarios:

- **Only watermarked content attack**, where the attacker has access only to marked signals, but not to embedder neither decoder. Thus, the inference about the secret parameters of the embedded must be made relying solely on the marked signals.

- **Watermarked content pair attack**, where the the observations are pairs of marked signals and their corresponding original versions. Here, the attacker has an extra source of information with respect to the previous scenario.

- **Chosen original content attack**, where the observations are pairs of marked and original signals chosen by the attacker. This situation may correspond, for example, to an scenario where the attacker has access to a watermark embedder as a black box.
- **Chosen watermarked content attack**, where the watermarker has access to a watermark detector/decoder as a black box. The attacker may feed the decoder with a chosen watermarked signal and observe the output in order to gain knowledge about the secret parameters of the decoder. This kind of attacks are also known as "oracle attacks" or "sensitivity attacks" [156],[147],[73], already mentioned in Section 2.2.

This type of classification decouples at a large extent the security analysis from the specific application. Furthermore, it allows to identify the components of security more clearly: the embedding method itself, the additional randomness brought about by the embedded messages, the availability of an embedder/detector as a black box, etc.

In this thesis we will address the problem of watermarking security in the following classes of scenarios, which are based on the classification proposed in [61].

Known Message Attack (KMA)

The attacker is assumed to have access to watermarked signals and to the messages embedded in each of those signals. Hence, in this case the security measure is given by

$$\begin{aligned} &I(\mathbf{Y}_1,\dots,\mathbf{Y}_{N_o},M_1,\dots,M_{N_o};\psi(\mathbf{\Theta})) \\ &\qquad = I(\mathbf{Y}_1,\dots,\mathbf{Y}_{N_o};\psi(\mathbf{\Theta})|M_1,\dots,M_{N_o}) \\ &\qquad = h(\psi(\mathbf{\Theta})) - h(\psi(\mathbf{\Theta})|\mathbf{Y}_1,\dots,\mathbf{Y}_{N_o},M_1,\dots,M_{N_o}), \end{aligned} \tag{2.18}$$

where the second and third equalities follow, respectively, from the definition of mutual information and by the assumption of independence between the messages and the secret key (cf. Section 2.2.1). This scenario constitutes the basis for the study of more involved scenarios and provides the main insight into the security problem, since the influence of the embedding parameters of the data hiding scheme can be more easily identified. Furthermore, it is also representative of some practical applications such as the following:

1. In the copy protection application, the embedded messages are usually known by any user. For instance, in copy protection for DVD video [51], the DVD player

may inform about the usage restrictions on certain DVD disks with messages like "Copy Never", "Copy Once", which were previously embedded in the digital contents. Moreover, due to severe resynchronization issues, watermark embedding strategies for video applications usually work with a unique secret key which is reused for marking the video sequences in a frame-by-frame basis. This, together with the fact of knowing the embedded messages, implies a serious threat to the security of copy protection applications.

2. Certain watermark detection schemes are adaptations of other schemes originally devised for data hiding applications. For instance, [157] and [184] propose the use of lattices for detection. The proposed methods are based on the well-known lattice DC-DM and ST-DM [63] schemes, respectively, and their adaptation to detection applications consists basically in limiting the possible transmitted messages to one representative. Hence, this framework would also fit in the KMA scenario as described in this thesis.

Watermarked Only Attack (WOA)

WOA models most of the data hiding scenarios of practical interest. The only information available to the attacker are the marked signals, without any knowledge of the embedded messages.

The security measure is given by

$$I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; \psi(\mathbf{\Theta})) = h(\psi(\mathbf{\Theta})) - h(\psi(\mathbf{\Theta})|\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}). \tag{2.19}$$

As can be expected, the information about $\psi(\mathbf{\Theta})$ provided to the attacker in the WOA scenario never exceeds that in the KMA scenario. We will pay special attention to quantifying how much information is lost due to the ignorance of the embedded messages. This loss will be represented by the "loss function", defined as

$$\delta(N_o) \triangleq I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}, M_1, \ldots, M_{N_o}; \psi(\mathbf{\Theta})) - I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; \psi(\mathbf{\Theta})). \tag{2.20}$$

This expression can be further simplified. By using the chain rule for mutual informations [84, Chapter 2], Eq. (2.19) can be rewritten as

$$\begin{aligned} & I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; \psi(\mathbf{\Theta})) \\ &\quad = \; I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; \psi(\mathbf{\Theta}), M_1, \ldots, M_{N_o}) - I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; M_1, \ldots, M_{N_o}|\psi(\mathbf{\Theta})) \\ &\quad = \; I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; \psi(\mathbf{\Theta})|M_1, \ldots, M_{N_o}) + I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; M_1, \ldots, M_{N_o}) \\ &\quad - \; I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; M_1, \ldots, M_{N_o}|\psi(\mathbf{\Theta})). \end{aligned} \tag{2.21}$$

The first term of Eq. (2.21) is the mutual information in the KMA scenario. The second and third terms of (2.21) represent the amount of information that can be learned by an attacker and by a user knowing the key, respectively, about the sequence of embedded messages. Using (2.21), the loss function can be rewritten as

$$\begin{aligned}\delta(N_o) &= I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o}|\psi(\mathbf{\Theta})) - I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o})\\ &= H(M_1,\ldots,M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o}) - H(M_1,\ldots,M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},\psi(\mathbf{\Theta})),\end{aligned} \tag{2.22}$$

where the second equality follows from the definition of mutual information. The loss function, which is always positive, quantifies the information about $\mathbf{T}$ that is lost due to the a priori ignorance of the embedded messages. As can be seen from (2.22), the loss can be formulated as "the information that $\psi(\mathbf{\Theta})$ provides about the embedded messages upon observation of marked signals".

As an illustrative example that fits in the WOA scenario we can mention fingerprinting applications, where a seller embeds identification codes of the buyers in the digital contents to be sold, with the purpose of enabling their traceability. In this case, each content contains different embedded information, but all the contents will have been watermarked with the secret key of the seller. Thus, a possible approach for the buyers in order to implement a collusion attack (although many other approaches are possible) is to estimate first the secret key and then to remove the watermark of each content. In this case, a security analysis could provide an approximate figure of the observations needed for successfully implementing such collusion attack, giving a certain advantage to the seller.

Constant Message Attack (CMA)

In some particular instances of the WOA scenario, the actual embedded message is unknown by the attacker but it can be known to be repeated in a certain set of observations (marked signals). Clearly, this scenario can be thought of as an intermediate case between KMA and WOA.

The CMA assumption turns out to hold in many practical applications (the cases 2 and 3 below could be possibly regarded as valid in certain KMA scenarios as well):

1. **Data hiding for audio and video**. Usually, when marking audio and video signals, the information is repeatedly embedded in consecutive blocks due to simplicity of the embedding, resynchronization and granularity issues. We can find an example in the requirements imposed by DCI [142] to forensic marking technologies for digital cinema. The DCI specification demands the periodic insertion in the audio and video contents of information with the place and date of exhibition. This information must be repeated every few minutes in order to identify the source of an illegal camcording even if the available fragment is small.

2. **Embedding of pilot signals**. In general, watermarking and data hiding schemes are vulnerable to desynchronization attacks, either in the form of geometrical attacks [218] or temporal shiftings [37]. In order to simplify the resynchronization task, some authors have proposed the insertion of a reference message as a pilot signal in the marked contents [218],[107],[108]. This implies that a large amount of marked signals can contain the same reference message embedded. If the exact location (i.e. the subset of coefficients) of the pilot signal is known, the attacker can try to estimate it in order to remove it and fool the watermarking system.

3. **Authentication**. Certain authentication schemes (e.g. [109]), are based on the embedding of a reference message (which may be derived from the secret key) in the contents to be protected. The authentication process consists in decoding the message embedded in the to-be-authenticated content and checking whether it matches the reference message or not. This implies that the contents marked by the same user contain the same reference message, similarly to the case of pilot signals mentioned above.

Chapter 3

Security of Spread spectrum Watermarking: Theory

In this chapter, the framework introduced in Chapter 2 is used for developing a complete security analysis of spread spectrum methods for data hiding. In spread spectrum methods, the secret key is mapped to a pseudorandom sequence, which is usually known as secret “carrier” or ”spreading vector’ (both terms will be used without distinction throughout this thesis). Roughly speaking, watermark embedding is performed by modulating the secret carrier with the message to be embedded and adding the result to the original host signal. This way, the watermark is hidden in a secret subspace, preventing its access or removal by unauthorized users. On the other hand, this makes that all the signals marked with the same secret key contain the same pseudorandom pattern, a fact that constitutes a potential security hole.

In this chapter, three well-known data hiding methods are considered: additive Spread Spectrum [87], attenuated Spread Spectrum [173], and Improved Spread Spectrum [161]. Spread spectrum methods continue to be widely used, as many embedding functions existing nowadays are based on spreading. Thus, the analysis presented in this chapter is expected to provide useful insights in the identification of security weaknesses of current spread spectrum schemes and the design of improved ones.

This chapter is organized as follows: Section 3.1 gives an overview of the previous related works on this topic and outlines the main contributions of this thesis. In Section 3.2, the problem to be studied is formalized, recalling the working assumptions and introducing the notation to be used in this chapter and the next one. Section 3.3 studies the security of the classical spread spectrum embedding function, whereas Section 3.4 addresses the security when host rejection is considered. In Section 3.5, we provide bounds for the estimation of the spreading vector based on the information-theoretic analysis of the previous sections. The conclusions are summarized in Section 3.6.

3.1 Related work and contributions of this thesis

Our main reference work is the paper by Cayre et al. in [61]. There, the security of spread spectrum was quantified for the first time using the Fisher information in the KMA and WOA scenarios (and also under the “Known Original Attack”). The main conclusions of the analysis presented in [61] are the following:[1]

- The information leakage is linear with the number of observations N_o, i.e. all the observations provide the same amount of information about the spreading vector. However, in this chapter we show that the information leakage, using the mutual information measure, is strictly concave in N_o. Though apparently contradictory, this difference is readily justified by the fact that our measure explicitly considers the a priori uncertainty (entropy) of the spreading vector. Further comments are given in Section 3.3.1.

- The difficulty of the estimation depends on the relative powers between host signal and watermark (i.e. the DWR), in such a way that more embedding distortion implies a larger information leakage. Similar conclusions are obtained in this chapter using the mutual information measure.

- Perfect estimation of the spreading sequence is only possible in the KMA scenario; for the WOA scenario, a sign ambiguity will remain independently of the number of observations. In this chapter we arrive at the same conclusion by following a different reasoning than that of [61].

Further contributions of the present thesis to the theoretical analysis of spread spectrum security are the following:

- The analysis of more general embedding functions, considering for the first time the tradeoff between robustness and security in this kind of methods from a theoretical point of view.

- The evaluation of the security in asymptotic conditions.

- The derivation of new bounds on the estimation performance for the attackers. In [61], the Cramér-Rao lower bound for the estimation of the spreading vector had been considered. In this chapter, we consider its homologous information-theoretic bound based on Lemma 2.3, and another bound which specifically addresses the difficulty of estimating the subspace spanned by the spreading vector.

[1] Bear in mind that all the assumptions relevant for the analysis are the same in [61] and in this thesis, so the comparison between [61] and our results is fair.

Bear in mind that the analysis in [61] considers the existence of several carriers in the same marked signal, while we focus on a single carrier. The case of several carriers will be briefly addressed in Section 4.7 in the next chapter. Information-theoretic measures for the information leakage with several carriers are provided in [71], but only for one observation.

In a recent paper [59], Cayre et al. further explore the security of spread spectrum methods from a theoretical point of view. They propose two new embedding functions named Natural Watermarking (NW) and Circular Watermarking (CW). NW is shown to achieve perfect secrecy in the WOA scenario using information-theoretic arguments, although a more intuitive explanation in terms of blind source separation (BSS) theory [140] is also given. However, the advantage of perfect secrecy of NW comes at the price of a significant degradation of the robustness with respect to the original spread-spectrum method. The CW method proposed in the same paper is a generalization of NW, improving its robustness, although not preserving the perfect secrecy property. In this thesis we do not address these methods from the theoretical point of view, but some practical considerations are given in Section 4.7.3 in the next chapter.

3.2 Formal problem statement

In this section, the problem of watermarking security for spread spectrum methods is formalized. Hereinafter, boldface letters denote column vectors, whereas italicized letters denote scalar variables. Random variables and their realizations are denoted by capital and lowercase letters, respectively.

The secret key is transformed into an n-dimensional "spreading vector" or secret "carrier", which will be denoted by $\mathbf{S} = \Phi(\mathbf{\Theta})$, and parameterizes the embedding function. For the family of methods considered in this chapter, the embedding function can be written as

$$\mathbf{Y}_i = \mathbf{X}_i + (-1)^{M_i}\mathbf{S} + \Psi(\mathbf{X}_i, \mathbf{S}) = \mathbf{X}_i + \mathbf{W}_i,\ i = 1, \ldots, N_o, \tag{3.1}$$

where $M_i \in \mathcal{M} = \{0, 1\}$ denotes the embedded message that modulates $\mathbf{S}$, and the function $\Psi : \mathbb{R}^{2n\times 1} \to \mathbb{R}^{n\times 1}$ is used for host-rejection purposes. The resulting embedding rate of the scheme is $R = \log(2)/n$.

According to the assumptions introduced in Section 2.1, the host signals $\{\mathbf{X}_i, i = 1, \ldots, N_o\}$ are assumed to be mutually independent. The messages $\{M_i, i = 1, \ldots, N_o\}$ embedded in different observations are also assumed to be mutually independent and equiprobable in $\mathcal{M} = \{0, 1\}$. Besides, we introduce two additional assumptions:

1. The host vectors $\mathbf{X}_i$ are i.i.d. Gaussian-distributed with zero mean: $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ denotes the identity matrix of size $n \times n$.

2. The spreading vector is assumed to be Gaussian-distributed and i.i.d: $\mathbf{S} \sim \mathcal{N}(0, \sigma_S^2 \mathbf{I}_n)$.

Notice that the Gaussian modeling of the host may be somewhat restrictive. However, this model allows for a tractable mathematical formulation, yet providing the main insights into the security of this methods. On the other hand, the i.i.d. assumption about the host vectors makes the computation of the DWR (Eq. (2.2)) very simple:

$$\mathrm{DWR} = 10 \log_{10} \frac{\frac{1}{n} E[||\mathbf{X}_i||^2]}{D_w} = 10 \log_{10} \xi,$$

where $\xi \triangleq \sigma_X^2 / D_w$, and D_w depends on the particular method considered.

The objective of the attacker is to obtain an estimate of $\mathbf{S}$ using the information contained in the sequence of observations $\{\mathbf{O}_i,\ i = 1, \ldots, N_o\}$, where $\mathbf{O}_i = [\mathbf{Y}_i^T, M_i]^T$ in the KMA case,[2] and $\mathbf{O}_i = \mathbf{Y}_i$ in the WOA case.

Other notational conventions widely used in this chapter are the following: $\Gamma(z)$ denotes the complete Gamma function. If $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_n)$, then $T \triangleq ||\mathbf{X}||^2$ follows a Chi-squared distribution with n degrees of freedom, which is denoted as $\chi^2(n, \sigma_X)$. If $\mathbf{X} \sim \mathcal{N}(\mathbf{v}, \sigma_X^2 \mathbf{I}_n)$, then $T' \triangleq ||\mathbf{X}||^2$ follows a noncentral Chi-squared, denoted by $\chi'^2(n, \mathbf{v}, \sigma_X)$. The probability density function (pdf) of a continuous random variable A is denoted by $f(a)$.

3.2.1 Spread-Spectrum-based embedding

We summarize in this section the embedding functions considered in this chapter along with their parameters.

Additive Spread Spectrum (add-SS)

Binary add-SS, as proposed by Cox et al. [87], is the most popular and widely studied watermarking method. No host rejection is performed, so the embedding function is simply given by

$$\mathbf{Y}_i = \mathbf{X}_i + (-1)^{M_i} \mathbf{S}, \text{ for } i = 1, \ldots, N_o. \tag{3.2}$$

The embedding distortion results simply in $D_w = \sigma_S^2$.

[2] The notation $[a, b]$ indicates concatenation of the row vectors a and b.

Attenuated Spread Spectrum (γ-SS)

The attenuated spread spectrum technique proposed in [173] consists in attenuating the host prior to embedding, in order to optimize the power transmission subject to an MSE distortion constraint (the embedding distortion). The embedding function is as follows:

$$\mathbf{Y}_i = (1-\gamma)\cdot \mathbf{X}_i + (-1)^{M_i}\cdot \mathbf{S}, \text{ for } i = 1,\dots,N_o, \tag{3.3}$$

where $0 \leq \gamma \leq 1$ is a host-rejection parameter. The embedding distortion of this scheme is given by $D_w = \gamma^2\sigma_X^2 + \sigma_S^2$. We will refer to this technique in the following as γ-SS. The parameter γ can be adjusted so as to optimize some performance measure, usually the error probability. Thus, the performance (in terms of robustness) of γ-SS is at least as good as that of add-SS, as the latter is just a particular case of γ-SS for $\gamma = 0$.

Improved Spread Spectrum (ISS)

ISS [161] is the result of introducing a host-interference-rejection mechanism in add-SS, fundamentally different from γ-SS in that ISS attenuates the host only in the direction of embedding, thus saving in embedding distortion and improving the performance of the latter in terms of robustness. We will consider the linear version of ISS, whose embedding function is as follows:

$$\mathbf{Y}_i = \mathbf{X}_i + (-1)^{M_i}\nu\mathbf{S} - \lambda\frac{\mathbf{X}_i^T\mathbf{S}}{||\mathbf{S}||^2}\mathbf{S}, \text{ for } i = 1,\dots,N_o, \tag{3.4}$$

where $0 \leq \lambda \leq 1$ is the host-rejection parameter, and ν is a parameter for fixing the embedding distortion. The embedding distortion in ISS can be computed as follows. For the ith observation and a particular $\mathbf{s}$ we have

$$E[||\mathbf{W}_i||^2|\mathbf{S}=\mathbf{s}] = E\left[\left|\left|\left((-1)^{M_i}\nu - \lambda\frac{\mathbf{X}_i^T\mathbf{s}}{||\mathbf{s}||^2}\right)\mathbf{s}\right|\right|^2\right] = \nu^2||\mathbf{s}||^2 + \lambda^2\sigma_X^2. \tag{3.5}$$

Finally, for a zero-mean Gaussian spreading vector,

$$D_w = \frac{1}{n}E[||\mathbf{W}_i||^2] = \nu^2\sigma_S^2 + \frac{\lambda^2}{n}\sigma_X^2, \tag{3.6}$$

which is the same result as for a spreading vector with constant norm equal to $n\sigma_S^2$ (the case originally considered in [161]).

3.3 Security analysis of Additive Spread Spectrum (add-SS)

3.3.1 KMA scenario

Under the KMA assumptions, the KMA scenario for add-SS can be seen as a simple additive Gaussian channel:

$$\mathbf{Y}_i = \mathbf{X}_i + (-1)^{m_i}\mathbf{S},$$

where $\mathbf{X}_i$ plays the role of interfering signal and $\mathbf{S}$ is the signal to be communicated. This is a classical communications scenario [84], where the mutual information is given by the well known expression

$$I(\mathbf{Y}_i;\mathbf{S}|M_i) = \frac{1}{2}\log\left(1+\frac{\sigma_S^2}{\sigma_X^2}\right). \tag{3.7}$$

Our problem is different in that we want to compute the information about $\mathbf{S}$ provided by N_o channel uses where $\mathbf{S}$ remains constant. Thus, it can be regarded as a sort of communication problem where the information is repetition-coded by means of several channel uses. This scenario was already considered in [71]. Here we provide a simple, alternative derivation for the sake of completeness. Moreover, the quantities defined below will be frequently recalled in the remaining of this chapter.

Let us denote by $\bar{\mathbf{S}}_{N_o}$ the random variable $\mathbf{S}$ conditioned on N_o observations. From the embedding function of add-SS, it follows that the components $\{\bar{S}_i,\ i=1,\ldots,n\}$, of $\bar{\mathbf{S}}_{N_o}$ are all mutually independent and Gaussian-distributed [199]. Given a particular realization $\{\mathbf{Y}_1=\mathbf{y}_1,\ldots,\mathbf{Y}_{N_o-1}=\mathbf{y}_{N_o}, M_1=m_1,\ldots,M_{N_o-1}=m_{N_o}\}$, we have $\bar{\mathbf{S}}_{N_o}\sim\mathcal{N}(\mathbf{v},\sigma^2_{\bar{S}_{N_o}}\mathbf{I}_n)$, with

$$v_i = \frac{\sigma_S^2}{N_o\sigma_S^2+\sigma_X^2}\boldsymbol{\mu}^T\mathbf{y}^{(i)},\ \ i=1,\ldots,n, \tag{3.8}$$

$$\sigma^2_{\bar{S}_{N_o}} = \frac{\sigma_X^2\sigma_S^2}{N_o\sigma_S^2+\sigma_X^2}, \tag{3.9}$$

where $\boldsymbol{\mu}\triangleq[(-1)^{m_1},\ldots,(-1)^{m_{N_o}}]^T$, and $\mathbf{y}^{(i)}\triangleq[y_{1,i},\ldots,y_{N_o,i}]^T$. Since $\bar{\mathbf{S}}_{N_o}$ is i.i.d. Gaussian, its entropy is given by $h(\bar{\mathbf{S}}_{N_o})=\frac{n}{2}\log(2\pi e\sigma^2_{\bar{S}_{N_o}})$, i.e. it does not depend on the particular realization of the observations. Hence, we can conclude that the equivocation per dimension is

$$\frac{1}{n}h(\mathbf{S}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},M_1,\ldots,M_{N_o})_{\text{add-SS}} = \frac{1}{2}\log\left(2\pi e\frac{\sigma_S^2}{1+N_o\cdot\xi^{-1}}\right). \tag{3.10}$$

Now, it is straightforward to see that the information leakage per dimension reads as

$$\frac{1}{n}I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};\mathbf{S}|M_1,\ldots,M_{N_o})_{\text{add-SS}} = \frac{1}{2}\log\left(1+N_o\cdot\xi^{-1}\right). \tag{3.11}$$

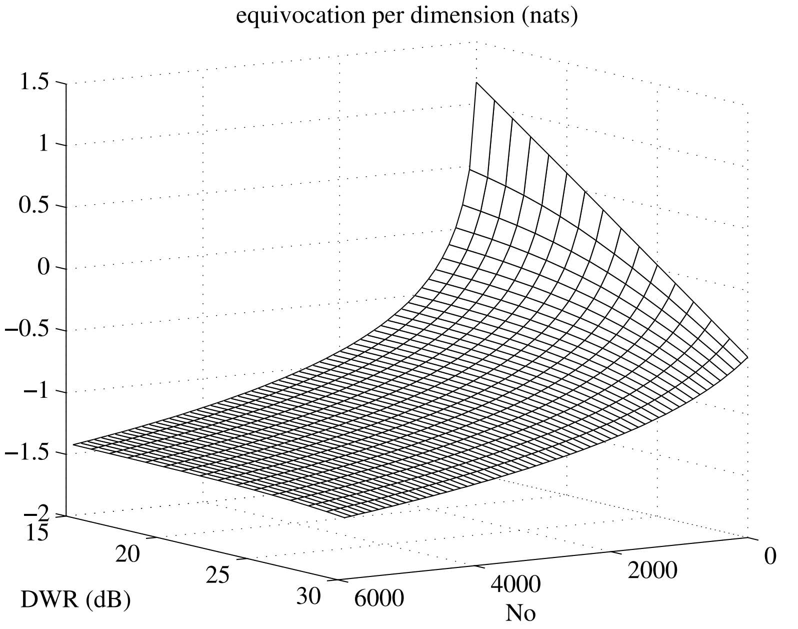


Figure 3.1: Equivocation per dimension for add-SS in the KMA scenario.

The information leakage (equivocation) is concave (convex) and strictly increasing (decreasing) with N_o, and its increasing (decreasing) rate is dependent on the DWR. Although (3.10) depends on the value of σ_S^2, as we can see in Figure 3.1, for large N_o we have

$$\frac{1}{n}h(\mathbf{S}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},M_1,\ldots,M_{N_o})_{\text{add-SS}} \approx \frac{1}{2}\log\left(2\pi e\sigma_X^2/N_o\right).$$

Finally, notice that both (3.10) and (3.11) are independent of n, meaning that the information leakage about each dimension is independent of the total number of dimensions. In other words, the difficulty of estimating each component of $\mathbf{S}$ does not depend on its total length.

Remark 3.1. According to [61], the information leakage about $\mathbf{S}$ should be linear in N_o, but (3.11) shows that it is instead logarithmic. Nevertheless, both results are not contradictory. The reason for obtaining a linear information leakage is that the analysis in [61] does not consider the random nature of $\mathbf{S}$. This randomness can be readily accounted for simply by introducing an additional term in the Fisher Information Matrix, yielding the so-called Total Information Matrix [214], as explained in [68] and [71].

3.3.2 WOA scenario

The WOA scenario can be seen as an additive Gaussian channel with an unknown scaling factor which, according to the binary transmission scheme given in Eq. (3.2) and the assumption of equiprobable, independent messages stated in Section 3.2, takes values ±1 equiprobably in each channel use. In this case, an exact expression for the information leakage cannot be obtained, so we derive upper and lower bounds. The information leakage is rewritten as in (2.21):

$$
\begin{aligned}
&I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};\mathbf{S})\\
&\quad= I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};\mathbf{S}),M_1,\ldots,M_{N_o}) - I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o}|\mathbf{S})\\
&\quad= I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};\mathbf{S}|M_1,\ldots,M_{N_o}) + I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o})\\
&\quad- I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o}|\mathbf{S}). \qquad (3.12)
\end{aligned}
$$

The first term of Eq. (3.12) has been already calculated in (3.11). The second and third terms of (3.12) represent the amount of information that can be learned by an attacker and by a fair user, respectively, about the sequence of embedded messages. These quantities are studied in Appendix B.1, resulting in the following upper and lower bounds to the information leakage per dimension:

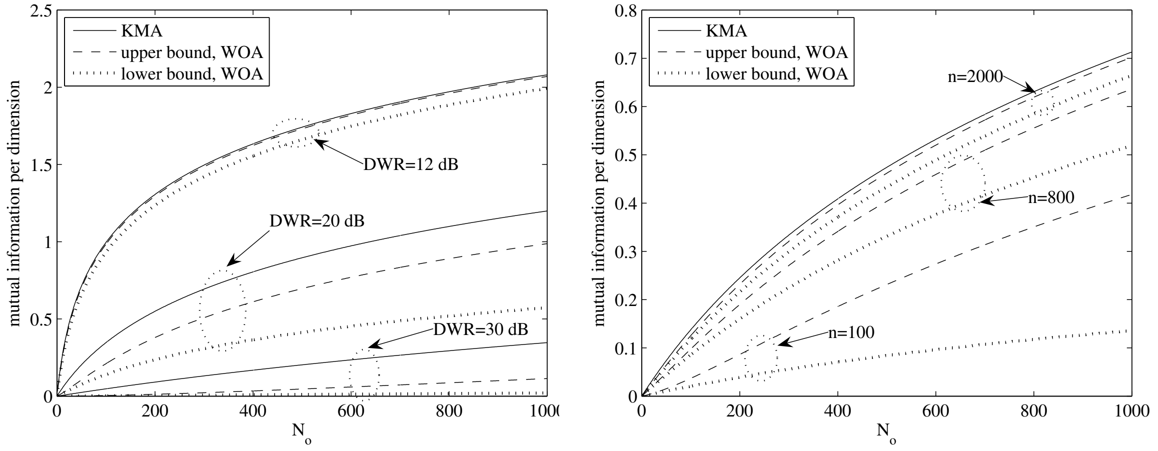
$$
\begin{aligned}
&\frac{1}{n}I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};\mathbf{S})_{\text{add-SS}}\\
&\quad\le \frac{1}{2}\log\left(1+N_o\cdot\xi^{-1}\right) + \frac{N_o}{n}I(\bar{\mathbf{X}}_{N_o}+(-1)^{M_{N_o}}\mathbf{V}_{N_o};M_{N_o}|\mathbf{V}_{N_o})\\
&\quad- \frac{N_o}{n}I(\mathbf{X}_1+(-1)^{M_1}\mathbf{S};M_1|\mathbf{S}),\ \text{for } N_o\ge 2, \qquad (3.13)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{n}I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};\mathbf{S})_{\text{add-SS}} &\ge \frac{1}{2}\log\left(1+N_o\cdot\xi^{-1}\right) + \frac{1}{n}\sum_{i=2}^{N_o} I(\bar{\mathbf{X}}_i+(-1)^{M_i}\mathbf{V}_i;M_i|\mathbf{V}_i)\\
&- \frac{N_o}{n}I(\mathbf{X}_1+(-1)^{M_1}\mathbf{S};M_1|\mathbf{S}),\ \text{for } N_o\ge 2. \qquad (3.14)
\end{aligned}
$$

In the expressions (3.13) and (3.14), $\bar{\mathbf{X}}_i \sim \mathcal{N}(\mathbf{0},(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)\mathbf{I}_n)$, with $\sigma_{\bar{S}_{i-1}}^2$ given by (3.9), and $\mathbf{V}_i \sim \mathcal{N}(\mathbf{0},\frac{(i-1)\sigma_S^4}{(i-1)\sigma_S^2+\sigma_X^2}\mathbf{I}_n)$. The second and third terms of (3.13) and (3.14) must be numerically computed by taking into account that

$$
\begin{aligned}
I(\mathbf{X}_1+(-1)^{M_1}\mathbf{S};M_1|\mathbf{S})_{\text{add-SS}} &= E\left[h((-1)^{M_1}||\mathbf{S}||^2+\mathbf{X}_1^T\mathbf{S}|\mathbf{S}=\mathbf{s})\right]\\
&- \frac{1}{2}E\left[\log\left(2\pi e\sigma_X^2\|\mathbf{S}\|^2\right)\right],
\end{aligned}
$$



(a) $n = 100$

(b) DWR = 25 dB

Figure 3.2: Comparison between the information leakage in KMA and WOA scenarios for add-SS. Figures 3.2(a) and 3.2(b) show the effect of varying the DWR and n, respectively.

where the expectation is taken over $\mathbf{S}$. Notice that the above upper and lower bounds differ only in their second term. Nevertheless, they cannot be given in closed-form, so numerical integration (on a scalar domain) is needed. A comparison between the information leakage (per dimension) in KMA and WOA scenarios is shown in Figure 3.2:

- Figure 3.2(a) shows that, when the parameter n is fixed, decreasing the DWR increases the information leakage in the KMA scenario (recall Eq. (3.11)), and simultaneously reduces the gap between KMA and WOA.

- Figure 3.2(b) shows the effect of varying the length of the spreading vector, n, when the DWR is fixed. In this case, we can see that the information leakage of the KMA scenario is approached as n is increased. Thus, the security level of the WOA scenario is strongly dependent on the value of n, contrarily to the KMA scenario.

The asymptotic behavior of the information leakage is formalized in the next theorem. The "loss function" defined in (2.22) now becomes

$$\begin{aligned}\delta(N_o)_{\text{add-SS}} &= I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o}|\mathbf{S}) - I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o})\\ &= H(M_1,\ldots,M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o}) - H(M_1,\ldots,M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},\mathbf{S}),\end{aligned} \tag{3.15}$$

Theorem 3.1 (Asymptotics of the loss function for add-SS). The loss function for add-SS in the WOA scenario can be upper bounded as

$$\delta(N_o)_{\text{add-SS}} \leq \log(2) + \sum_{i=2}^{N_o} H\left(\frac{\tau_i^{\frac{n}{2}}}{2}\right), \tag{3.16}$$

where

$$\tau_i = \frac{i\sigma_S^2\sigma_X^2 + \sigma_X^4}{(i-1)\sigma_S^4 + i\sigma_S^2\sigma_X^2 + \sigma_X^4},$$

and $H(\cdot)$ denotes the binary entropy function. The right hand side of (3.16) is decreasing with n and DWR^{-1}, and the following asymptotic properties hold:

1. For fixed n, $\lim\limits_{\text{DWR}\to-\infty} \delta(N_o)_{\text{add-SS}} \leq \log(2)$.

2. For fixed DWR, $\lim\limits_{n\to\infty} \delta(N_o)_{\text{add-SS}} \leq \log(2)$.

Proof: See Appendix B.2. ■

Theorem 3.1 basically states that the ignorance of the embedded messages does not affect the difficulty of estimating **S** if either the DWR or the embedding rate R (recall that $R = \log(2)/n$) are small enough. Although the first case is of virtually null relevance in practice, the second case is of major importance for practical applications. When high robustness is sought, the watermark is usually embedded at very low rates that allow to recover the message with low complexity. In such case a few observations suffice to obtain an estimate of **S** that in turn allows accurate recovery of the embedded message. Nevertheless, it is important to point out, as other researchers realized before [61] (by means of Blind Source Separation theoretic arguments), that the penalty to pay for not knowing the messages M_i comes in the form of an ambiguity in the sign of **S**, independently of n and of the DWR, that cannot be undone if no additional knowledge about the embedded messages is available. An alternative way for proving this ambiguity is to show that the a posteriori probability of the spreading vector in the WOA scenario is independent of its sign. One can easily check that

$$\begin{aligned} f(\mathbf{s}_0|\mathbf{Y}_i = \mathbf{y}_i) &= \frac{f(\mathbf{y}_i|\mathbf{S} = \mathbf{s}_0)f(\mathbf{s}_0)}{f(\mathbf{y}_i)} = \frac{f(\mathbf{y}_i|\mathbf{S} = -\mathbf{s}_0)f(-\mathbf{s}_0)}{f(\mathbf{y}_i)} \\ &= f(-\mathbf{s}_0|\mathbf{Y}_i = \mathbf{y}_i),\ \forall\ \mathbf{s}_0 \in \mathbb{R}^n, \end{aligned} \tag{3.17}$$

so the sign ambiguity becomes patent. Notice that (3.17) holds regardless of the statistical distribution of the host, but it is needed that **S** be circularly symmetric. In the information-theoretic analysis, the sign ambiguity is reflected by the factor log(2).

3.3.3 CMA scenario

The CMA scenario can be seen as a particular case of WOA where the unknown scaling factor in the Gaussian channel remains constant for N_o channel uses and takes the value ±1 equiprobably. For this reason, the asymptotic results of Theorem 3.1 are also applicable to CMA. However, in this case it is much easier to derive the asymptotic behavior of the loss function. For instance, taking into account that $M_1 = M_2, \ldots, M_{N_o} = M$, with unknown M, it is straightforward to upper bound the loss function as

$$\delta(N_o) \leq \log(2)$$

without any further consideration on the DWR. Moreover, as $n \rightarrow \infty$, this upper bound obviously goes to 0. Hence, the impact in security level of not knowing the embedded message in the CMA scenario is negligible. Nevertheless, it must be noted that the sign ambiguity mentioned in Section 3.3.2 is also irreducible in the CMA scenario if no information about the embedded message is available. Nonetheless, either in CMA or WOA scenarios, the sign ambiguity does not prevent from estimating the one-dimensional subspace spanned by $\mathbf{s}$, a feature that will be exploited by the practical estimators in Section 4.

3.4 Spread Spectrum with host rejection

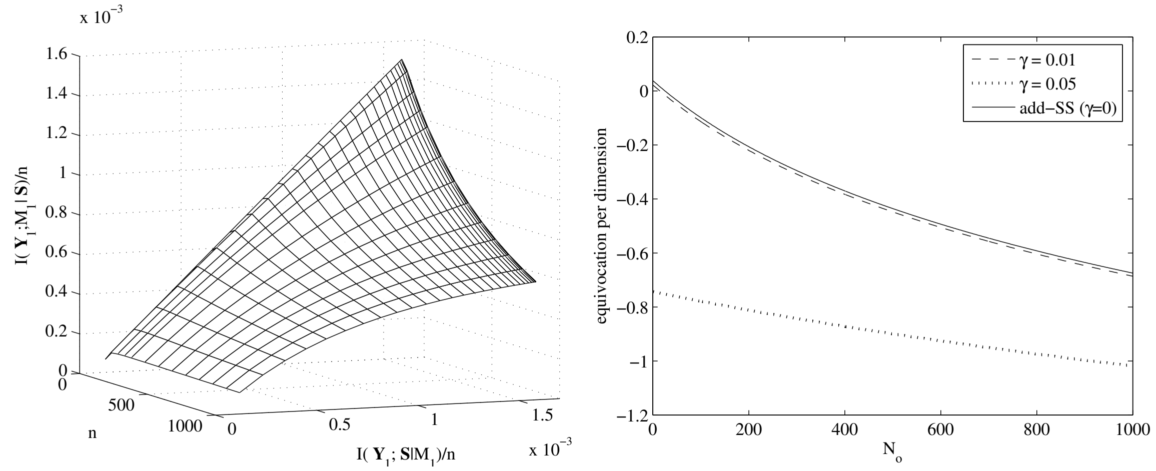
After studying the security properties of add-SS, the influence of the host rejection mechanisms in the security level is addressed in this section. Two particular methods are studied: "attenuated spread spectrum" [173] and "improved spread spectrum" [161].

3.4.1 Attenuated Spread Spectrum (γ-SS)

Although the robustness of the optimized γ-SS scheme introduced in Section 3.2.1 is better than that of add-SS, its security level can be shown to be always worse for the same values of DWR and n. In order to provide a fair comparison between both γ-SS and add-SS, we impose that $\sigma_S^2 = \sigma_X^2\,(\xi^{-1} - \gamma^2)$ so as to get $D_w = \sigma_S^2$ as in add-SS.[3] It is easy to see that the results obtained for add-SS can be straightforwardly adapted to γ-SS replacing σ_S^2 by $\sigma_X^2\,(\xi^{-1} - \gamma^2)$ and σ_X^2 by $(1-\gamma)^2\sigma_X^2$ in the corresponding expressions. The equivocation for the KMA scenario, for instance, results in

$$\frac{1}{n}h(\mathbf{S}|\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}, M_1, \ldots, M_{N_o})_{\gamma\text{-SS}} = \frac{1}{2}\log\left(2\pi e\frac{\sigma_X^2(\xi^{-1}-\gamma^2)(1-\gamma)^2}{(1-\gamma)^2 + N_o(\xi^{-1}-\gamma^2)}\right). \quad (3.18)$$

[3] Note that this condition restricts the value of γ to the interval $0 \leq \gamma \leq \xi^{-1}$, i.e. in practical scenarios (with low embedding distortion) γ takes values close to 0.



(a) (b)

Figure 3.3: γ-SS for KMA scenario and DWR = 25 dB. Tradeoff robustness-security as a result of varying γ in the interval $[0, \xi^{-1}]$, for $N_o = 1$ (a) and equivocation per dimension (b).

The expression above can be shown to be monotonically decreasing with γ. The results for add-SS WOA can be easily generalized as well to γ-SS. The most interesting consequence of introducing the parameter γ is the existence of a tradeoff between robustness and security. This tradeoff is illustrated in Figure 3.3(a), which shows the plot of the information leakage for $N_o = 1$ (in the KMA scenario) vs. the achievable rate for a fair user. The latter can be computed by numerical integration by taking into account that

$$\begin{aligned} I(\mathbf{X}_1 + (-1)^{M_1}\mathbf{S}; M_1|\mathbf{S})_{\gamma\text{-SS}} &= E\left[h((-1)^{M_1}||\mathbf{S}||^2 + (1-\gamma)\mathbf{X}_1^T\mathbf{S}|\mathbf{S})\right] \\ &- \frac{1}{2}E\left[\log\left(2\pi e(1-\gamma)^2\sigma_X^2\|\mathbf{S}\|^2\right)\right]. \end{aligned}$$

Figure 3.3(a) basically shows that increasing the achievable rate for fair users will provide more information for attackers interested in estimating $\mathbf{S}$. As can be seen, the tradeoff is also dependent of n, since this parameter affects the achievable rate. Figure 3.3(b) shows the equivocation per dimension in the KMA scenario for several values of the parameter γ, evidencing the degradation of the security level as the host rejection is increased.

3.4.2 Improved Spread Spectrum (ISS)

For a fair comparison between the ISS scheme and add-SS, the embedding distortion of the former is fixed to $D_w = \sigma_S^2$, which imposes

$$\nu = \left(1 - \frac{\lambda^2\xi}{n}\right)^{\frac{1}{2}}. \tag{3.19}$$

Since ν must be real, the maximum allowable value for λ is determined by

$$\lambda \leq \min\{1, \sqrt{n\xi^{-1}}\}.$$

Clearly, λ can be made arbitrarily close to 1 by increasing n, thus achieving complete rejection of the host interference. In general, the parameter λ is tuned so as to optimize the performance of ISS in terms of error probability. Clearly, ISS with $\lambda = 0$ is equivalent to add-SS as described in Section 3.3. We are concerned in this section with determining the effect on the security level of using $\lambda > 0$. The study will be carried out for the KMA scenario, where the information leakage is given by

$$\begin{aligned} I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; \mathbf{S}|M_1, \ldots, M_{N_o})_{\text{ISS}} &= h(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}|M_1, \ldots, M_{N_o}) \\ &- h(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}|\mathbf{S}, M_1, \ldots, M_{N_o}). \quad (3.20) \end{aligned}$$

The second term is easy to compute: given the messages and $\mathbf{S} = \mathbf{s}$, the observations are mutually independent, following a Gaussian distribution

$$\mathbf{Y}_i|\mathbf{S} = \mathbf{s}, M_i = m_i \sim \mathcal{N}((-1)^{m_i}\nu\mathbf{s}, \mathbf{\Sigma_S}), \quad (3.21)$$

with $\mathbf{\Sigma_S} = E\left[(\mathbf{Y}_i - (-1)^{m_i}\nu\mathbf{s})^T \cdot (\mathbf{Y}_i - (-1)^{m_i}\nu\mathbf{s})\right]$ the covariance matrix of the $\mathbf{Y}_i$ conditioned on the realization of $\mathbf{S}$ and M_i. The eigenvalue decomposition of $\mathbf{\Sigma_S}$ is given by $= \mathbf{U_S}\mathbf{\Lambda}\mathbf{U_S}^T$, where

$$\mathbf{\Lambda} = \left[\begin{array}{cc} (1-\lambda)^2\sigma_X^2 & 0 \\ 0 & \sigma_X^2\mathbf{I}_{n-1} \end{array} \right], \quad (3.22)$$

and $\mathbf{U_S}$ is a unitary matrix whose first column is collinear to $\mathbf{s}$. That is, the marked signal $\mathbf{Y}_i$ can be seen as a signal $(-1)^{M_i}\nu\mathbf{S}$ transmitted in a Gaussian channel with noise correlated with the signal (the statistics of the marked signal are illustrated in Figure 3.4 for $n = 2$). It turns out that in ISS not only the mean of the observations provides information about $\mathbf{S}$, but also the covariance matrix of the noise (here, the attenuated host). Using (3.21) and (3.22) we can write

$$\begin{aligned} h(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}|\mathbf{S}, M_1, \ldots, M_{N_o}) &= \sum_{i=1}^{N_o} h(\mathbf{Y}_i|M_i, \mathbf{S}) \\ &= \frac{N_o}{2}\log\left((2\pi e)^n \cdot (\sigma_X^2)^n \cdot (1-\lambda)^2\right). \quad (3.23) \end{aligned}$$

Since the first term of (3.20) is hard to compute analytically, we first formalize the general behavior of the information leakage in the following theorem for $N_o = 1$.

Theorem 3.2 (Convexity in λ of the information leakage for ISS). The information leakage in ISS is a convex and increasing function of the host-rejection parameter λ, and for $N_o = 1$ it is given by

$$\frac{1}{n}I(\mathbf{Y}_1; \mathbf{S}|M_1)_{\text{ISS}} = \frac{1}{2}\log\left(1 + \frac{\lambda(\lambda-2)}{n} + \nu^2\xi^{-1}\right) - \frac{1}{n}\log(1-\lambda). \quad (3.24)$$

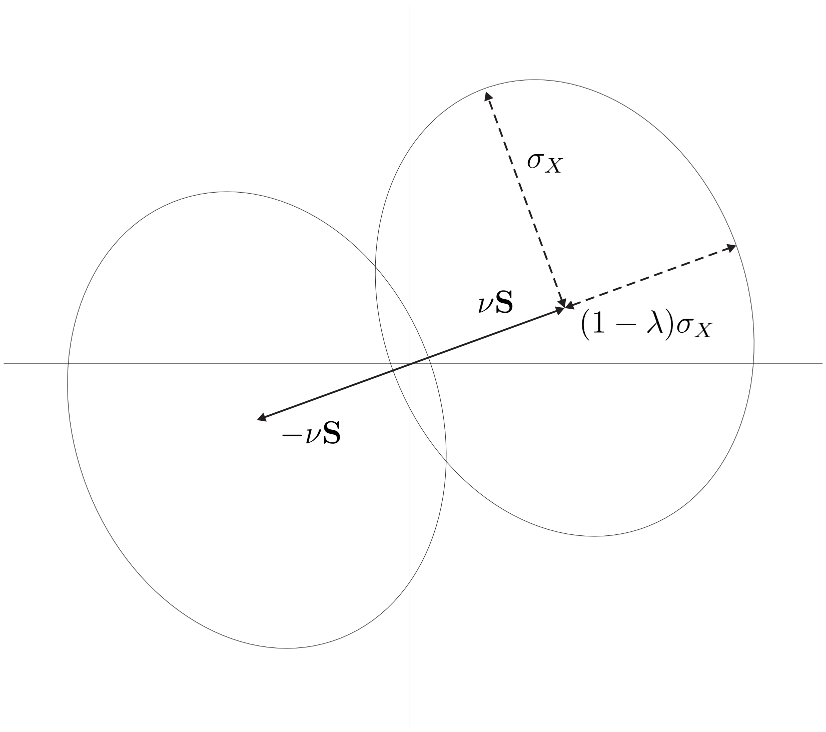


Figure 3.4: Statistics of the watermarked signal in ISS for $n = 2$.

Proof: In Appendix B.3, the exact value of the information leakage for $N_o = 1$ is shown to be given by (3.24). If we compute the first and second derivatives of the information leakage in terms of λ, we find out that the function is convex and increasing in the interval $\lambda \in [0, \min\{1, \sqrt{n\xi^{-1}}\}]$. ■

In ISS, the achievable rate depends on the value of λ. The optimum λ that maximizes this rate depends on the DWR and the power of the attacking noise (see [161] for further discussion). This behavior in conjunction with Theorem 3.2 shows that, similarly to the γ-SS scheme, the host-rejection mechanism of ISS induces a tradeoff between information leakage and achievable rate. This tradeoff is illustrated in Figure 3.5(a) by plotting $I(\mathbf{Y}_1; \mathbf{S}|M_1)/n$ vs $I(\mathbf{Y}_1; M_1|\mathbf{S})/n$ in terms of λ, for $\lambda \in [0, \min\{1, \sqrt{n\xi^{-1}}\}]$. It can be noticed the concavity of the curves, which present a global maximum of the achievable rate for a certain λ_{max} (dependent of n). Increasing λ beyond this value has the double (negative) effect of not increasing further the achievable rate but increasing the information leakage. Notice that for $\lambda = \sqrt{n\xi^{-1}} < 1$, from (3.19) we have $\nu = 0$, so $I(\mathbf{Y}_1; M_1|\mathbf{S}) = 0$. Even in this case, we can see in Figure 3.5(a) that the information leakage is not null, due to the dependence of the covariance matrix

$(\mathbf{\Sigma_S})$ on the spreading vector $\mathbf{S}$.[4]

For the case $N_o \geq 1$, an upper bound to the information leakage is derived in Appendix B.4, obtaining

$$\begin{aligned}&\tfrac{1}{n} I(\mathbf{Y}_1, \dots, \mathbf{Y}_{N_o}; \mathbf{S}|M_1, \dots, M_{N_o})_{\text{ISS}}\\ &\leq \frac{1}{2}\log\left(\left(1+\frac{\lambda(\lambda-2)}{n}\right)^{N_o}\left(1+\frac{N_o\nu^2\sigma_S^2}{\sigma_X^2\left(1+\frac{\lambda(\lambda-2)}{n}\right)}\right)\right) - \frac{N_o}{n}\log(1-\lambda),\ \text{for } N_o \geq 1.\end{aligned} \tag{3.25}$$

The bound (3.25) on the information leakage produces a lower bound on the equivocation, which is plotted in Figure 3.5(b) for different values of λ and compared to add-SS. We can see that the equivocation decreases as λ increases, in accordance with Theorem 3.2. This bound can be used to derive a conservative security level. Notice that (3.25) coincides with (3.24) for $N_o = 1$.

Remark 3.2. As expected, for $\lambda = 0$ (which implies $\nu = 1$) the bound (3.25) coincides with (3.11), the information leakage for add-SS. This is because the hypothesis of independence used in the bounding of Eq. (B.35) of Appendix B.4 is fulfilled for $\lambda = 0$.

Remark 3.3. For $n \to \infty$, (3.25) tends to (3.11), i.e.

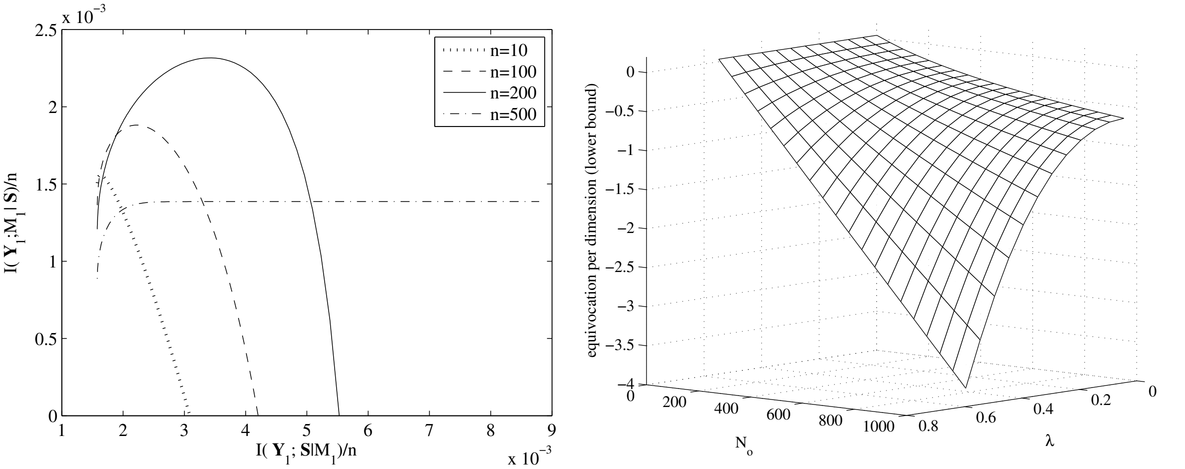
$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n} I(\mathbf{Y}_1, \dots, \mathbf{Y}_{N_o}; \mathbf{S}|M_1, \dots, M_{N_o})_{\text{ISS}}\\ &\qquad = \frac{1}{n} I(\mathbf{Y}_1, \dots, \mathbf{Y}_{N_o}; \mathbf{S}|M_1, \dots, M_{N_o})_{\text{add-SS}},\ \text{for } \lambda < 1,\end{aligned} \tag{3.26}$$

This means that asymptotically there is no penalty in security level for using host rejection in one dimension, constituting a major advantage over γ-SS. Remember that the latter performs host rejection in all dimensions, and as such it cannot benefit from increasing n for concealing the information about $\mathbf{S}$. Nevertheless, we want to remark that increasing n in ISS has the same effect as for add-SS stated in Theorem 3.1, namely, that the information leakage in the WOA scenario approaches that of the KMA scenario when $n \to \infty$. This can be easily proved for ISS following similar guidelines as those of Appendix B.2.

3.5 Bounds on the estimation error

In this section we provide fundamental performance bounds for practical estimators of the spreading vector. These bounds are based on the information-theoretic results

[4] Obviously, the shape of these curves will depend on the attacking noise (the regions of Figure 3.5(a) were obtained in the absence of noise).



(a) (b)

Figure 3.5: ISS for KMA scenario and DWR = 25 dB. Tradeoff robustness-security as a result of varying λ in the interval $[0, \min\{1, \sqrt{n\xi^{-1}}\}]$ (a), and lower bound on the equivocation per dimension for $n = 100$ (b).

derived in the previous sections. The aim is to translate the equivocation into other measures that result useful for the evaluation of the security from a practical point of view. The first bound is concerned with the mean-squared error between the spreading vector ($\mathbf{s}$) and its estimate ($\hat{\mathbf{s}}$), and the second one with the normalized correlation between $\mathbf{s}$ and $\hat{\mathbf{s}}$. The achievability of each bound is also discussed.

3.5.1 Bound on the variance of the estimation error

Let us define the estimation error as $\mathbf{e} \triangleq \mathbf{s} - \hat{\mathbf{s}}$ and its variance per dimension as $\sigma_E^2 \triangleq \frac{\mathrm{tr}(\mathbf{\Sigma}_E)}{n}$, where $\mathbf{\Sigma}_E$ is the covariance matrix of $\mathbf{e}$, and $\mathrm{tr}(\mathbf{\Sigma}_E)$ denotes the trace of $\mathbf{\Sigma}_E$. As shown in Lemma 2.3, we have the following lower bound:

$$\sigma_E^2 \geq \frac{1}{2\pi e} \exp\left(\frac{2}{n} h(\mathbf{S}|\mathbf{O}_1, \ldots, \mathbf{O}_{N_o})\right). \tag{3.27}$$

Hence, the equivocation can be regarded as the exponent of the estimation error lower bound. Inserting (3.10) into (3.27), for add-SS in the KMA scenario we have

$$\sigma_{E_{\mathrm{KMA}}}^2 \geq \frac{\sigma_S^2}{1 + N_o \cdot \xi^{-1}} \approx \frac{\sigma_X^2}{N_o}, \tag{3.28}$$

which is achievable when the estimation error is Gaussian-distributed, and the approximation holds for large N_o. This bound coincides with the Cramér-Rao lower bound, which was calculated in [71] using the Total Information Matrix [214], and with the variance of the MMSE estimator [199]. This is not surprising, as the MMSE estimator is unbiased and its estimation error is Gaussian-distributed, thus fulfilling the condition for achieving the bound.

A similar bound for the WOA scenario could be obtained by inserting the corresponding equivocation into (3.27). Taking into account Theorem 3.1 we find out that

$$\lim_{n\to\infty} \sigma^2_{E_{\text{WOA}}} \geq \frac{\sigma_S^2}{1 + N_o \cdot \xi^{-1}},$$

exactly as for KMA. However, we must bear in mind that, contrarily to KMA, the latter bound is obviously not achievable due to the sign ambiguity in the estimate of $\mathbf{S}$ (recall Sect. 3.3.2). Hence, the best estimate possible (for $N_o \to \infty$) is $\hat{\mathbf{S}} = \pm\mathbf{S}$ with probability 1/2 each, which leads to an error variance $2\sigma_S^2$, the minimum achievable without knowledge of the embedded messages.

3.5.2 Bound on the normalized correlation

In spread spectrum methods using binary antipodal constellations, exact knowledge of $\mathbf{s}$ is not necessary for performing correct decoding. Usually, decoding is implemented by means of a cross-correlation operation, estimating the message embedded in $\mathbf{y}_i$ as $\hat{m}_i = \text{sign}\{\mathbf{y}_i^T \cdot \mathbf{s}\}$. Also, in watermark detection applications based on spread spectrum, the detector decides on the presence of the watermark upon the angle between $\mathbf{y}_i$ and $\mathbf{s}$. In other words, the norm of $\mathbf{s}$ is important for the embedding operation (e.g. for controlling the embedding distortion), but not for detection/decoding. This implies that the attacker is mainly interested in disclosing the direction of $\mathbf{s}$, which spans the subspace where the watermark is contained. Thus, it is useful to quantify the difficulty in estimating the direction of $\mathbf{s}$. The natural performance measure is the normalized correlation, defined as

$$\rho \triangleq \frac{\hat{\mathbf{s}}^T \mathbf{s}}{||\hat{\mathbf{s}}|| \cdot ||\mathbf{s}||} = \cos(\omega) \in [-1, 1], \tag{3.29}$$

where ω denotes the angle between $\mathbf{s}$ and $\hat{\mathbf{s}}$. The closer to 1 is the value of ρ, the more accurate is the estimate of $\mathbf{s}$. The relation between ρ and the Cramér-Rao lower bound on the estimation error of $\mathbf{S}$ has been pointed out in [61, Sect. V], although not in deep detail. Here we pursue a bound on ρ using the equivocation.

Notice that the vector $\mathbf{s} \in \mathbb{R}^n$ can be expressed by means of its norm and an n-dimensional unit vector collinear to $\mathbf{s}$, i.e. we consider the transformation $\mathbf{s} \to (q, \mathbf{r})$, with $q = ||\mathbf{s}||$ and $\mathbf{r}$ a unit vector in the direction of $\mathbf{s}$. Thus, we have a coordinate change $\mathbb{R}^n \to \mathbb{R}^+ \times \mathbb{R}^n$.

Lemma 3.1 (Bound on the normalized correlation). For an unbiased estimator, the mean value of the normalized correlation can be bounded from above as

$$E\left[\cos(\Omega)\right] \leq 1 - \frac{n}{4\pi e}\exp\left(\frac{2}{n}h(\mathbf{R}|\mathbf{O}_1,\ldots,\mathbf{O}_{N_o})\right), \tag{3.30}$$

where $h(\mathbf{R}|\mathbf{O}_1,\ldots,\mathbf{O}_{N_o})$ represents the equivocation about $\mathbf{R}$, given N_o observations.

Proof: Let us define the estimation error as $\mathbf{d} \triangleq \mathbf{r} - \hat{\mathbf{r}}$. From Eq. (3.27), for an unbiased estimator we have

$$\frac{E\left[||\mathbf{D}||^2\right]}{n} = \frac{\mathrm{tr}(\mathbf{\Sigma}_D)}{n} \geq \frac{1}{2\pi e}\exp\left(\frac{2}{n}h(\mathbf{R}|\mathbf{O}_1,\ldots,\mathbf{O}_{N_o})\right). \tag{3.31}$$

By the cosine theorem, we have that $||\mathbf{d}||^2 = 2(1-\cos(\omega))$, with ω the angle between $\mathbf{r}$ and $\hat{\mathbf{r}}$. Combining this with (3.31), we arrive at the bound (3.30). ■

Lemma 3.1 relates the normalized correlation with the equivocation about $\mathbf{R}$. The a priori equivocation, $h(\mathbf{R})$, achieves its maximum when $\mathbf{R}$ is uniformly distributed over the surface of the unit-radius hypersphere. Note that this is the case when $\mathbf{S}$ is i.i.d. Gaussian, as we are assuming in this chapter. We are concerned now with the equivocation about $\mathbf{R}$. First, note that using the coordinate change $\mathbf{s} \to (q, \mathbf{r})$ introduced above, the differential entropy of $\mathbf{S}$ can be rewritten as

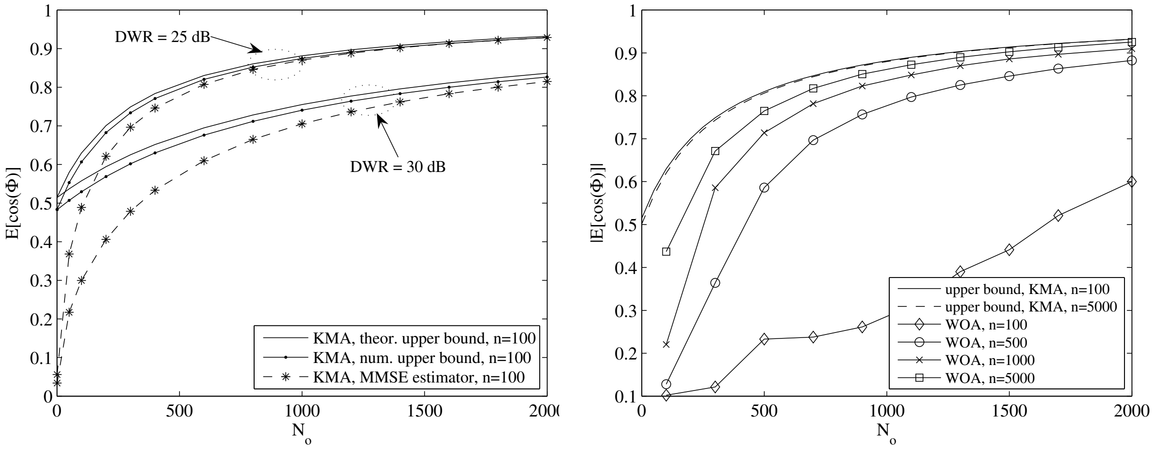
$$h(\mathbf{S}) = h(Q,\mathbf{R}) + E[\log(J)], \tag{3.32}$$

where J denotes the Jacobian of the coordinate change and the expectation is taken over $\mathbf{S}$. This change of coordinates can be seen as a QR factorization [122], $\mathbf{s} = q \cdot \mathbf{r}$, with $q \in \mathbb{R}$, $\mathbf{r} \in \mathbb{R}^n$, and $||\mathbf{r}|| = 1$. The Jacobian of this QR factorization is given by $J = Q^{n-1}$ [99]. Hence, using this result and (3.32), we can lower bound the equivocation on the embedding direction as

$$\begin{aligned} h(\mathbf{R}|\mathbf{O}_1,\ldots,\mathbf{O}_{N_o}) \;\geq\; & h(\mathbf{S}|\mathbf{O}_1,\ldots,\mathbf{O}_{N_o}) - h(Q|\mathbf{O}_1,\ldots,\mathbf{O}_{N_o}) \\ - \;& (n-1)E\left[E[\log(Q)|\mathbf{O}_1 = \mathbf{o}_1,\ldots,\mathbf{O}_{N_o} = \mathbf{o}_{N_o}]\right], \end{aligned} \tag{3.33}$$

where the inner expectation is taken over Q, and the outer expectation is over the observations. Equality in (3.33) is achieved when the norm and direction of $\mathbf{S}$ are mutually independent. Using (3.33), we will specialize the bound of Lemma 3.1 to add-SS in the KMA scenario.

Lemma 3.2 (Bound on the equivocation about the normalized correlation). For add-SS in the KMA scenario, the equivocation about $\mathbf{R}$ can be bounded from below as



(a) (b)

Figure 3.6: Upper bound to the normalized correlation for add-SS. Comparison with practical estimators for KMA with $n = 100$ (a) and WOA with DWR=25 dB (b).

$$
\begin{aligned}
&h(\mathbf{R}|\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}, M_1, \ldots, M_{N_o})\\
&\quad\geq \frac{n-1}{2}\log\left(\frac{2\pi e}{n\sigma_S^2}\right) + \frac{n}{2}\log\left(\frac{\sigma_S^2}{1+N_o\xi^{-1}}\right)\\
&\quad- \frac{1}{2}\log\left(n\sigma_S^2 - \frac{2\sigma_S^2}{1+N_o\xi^{-1}}\left(\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\right)^2 {}_1F_1\left(-\frac{1}{2};\frac{n}{2};-\frac{nN_o}{2}\xi^{-1}\right)^2\right), \quad (3.34)
\end{aligned}
$$

where ${}_1F_1$ denotes the confluent hypergeometric function of the first kind [29], and $\Gamma(\cdot)$ is the complete Gamma function.

Proof: The first term in the right hand side of (3.33) is given by (3.10). The remaining terms, related to the norm of the spreading vector, are upper bounded in Appendix B.5. The combination of these results yields (3.34). ■

We have represented in Figure 3.6(a) the upper bound on the normalized correlation for add-SS resulting from the insertion of (3.34) in (3.30). This theoretical bound is compared to the result using numerical integration for computing the norm-related terms, supporting the tightness of the bounds derived in Appendix B.5. For checking the tightness of the bound on $E[\cos(\Omega)]$, we have computed numerically (by Monte Carlo) the average normalized correlation resulting from applying the MMSE estimator of $\mathbf{S}$. As can be seen, the bound is loose for small N_o, but it becomes tight as N_o is increased. The reason is that the distributions of $\mathbf{R}$ and Q conditioned on the observations become approximately independent when N_o is increased and because the boundings based on Jensen's inequality are also asymptotically tight when the variance of the considered random variable approaches 0.

As seen in Section 3.3.2, when the embedding rate is small enough, the information that the WOA scenario provides about **S** approaches that of KMA except for the sign ambiguity. This ambiguity also arises when evaluating a practical estimator, but we can get rid of it simply by taking the absolute value of ρ as performance measure. Notice that, in this case, we would be evaluating the accuracy in the estimation of the subspace spanned by $\mathbf{s}$. This performance evaluation is illustrated in Figure 3.6(b), showing that the accuracy of the subspace estimation in the WOA case tends (as expected) to that of the KMA as n is increased. The curves for WOA were obtained empirically with the "PCA estimator", that will be studied in Section 4.

Remark 3.4. Notice that the theoretical bound in Figure 3.6(b) remains approximately invariant with n. Moreover, it can be checked that it is approximately independent of the specific values of σ_S or σ_X, depending only on ξ. Thus, it looks more appealing than the bound derived in Section 3.5.1 for evaluating the security level.

3.6 Conclusions

The security of spread-spectrum-based data hiding methods has been investigated from a theoretical point of view. Among the theoretical results obtained in this chapter, we would like to remark the following:

1. The use of Shannon's mutual information permits to show that the growth of the information gained by the attacker about the secret carrier is far from being linear, as had been stated in [61].

2. The use of low embedding rates (i.e. very large n) has a harmful impact in the security level of WOA scenarios. In limiting cases of zero-rate watermarking, quite extended for their robustness against blind attacks, the penalty for ignoring the embedded messages becomes negligible, representing a serious threat to the security of the system.

3. A tradeoff between security and robustness has been shown to exist in the spread spectrum methods that perform host rejection. For the schemes studied in this chapter, which cover a wide range of the spread spectrum schemes considered in the literature, host rejection can significantly decrease the security level of plain spread spectrum (add-SS). Nevertheless, different host rejection strategies can yield very different results: whereas the penalty for the ISS scheme vanishes as n is increased, the security level of γ-SS cannot be improved by increasing n.

Chapter 4

Security of Spread Spectrum Watermarking: Practical algorithms

After the theoretical analysis carried out in Chapter 3, this chapter is devoted to the evaluation of the security of spread spectrum methods from a more practical point of view. We will focus on the ISS embedding function, since the attacks devised for it are applicable to add-SS and γ-SS as well. The purpose of this chapter is twofold: on one hand, we analyze the approaches previously proposed for tackling the spreading vector estimation problem, highlighting their limitations; on the other hand, new estimators for the WOA scenario are proposed and analyzed. The estimation of the spreading vector is basically defined as an optimization problem, and as such it can be decomposed on two subproblems:

1. The definition of a suitable cost function, and its analysis.
2. The choice of an appropriate optimization algorithm.

The analysis of the corresponding cost functions is performed under the assumption of an infinite-length sample (i.e. $N_o \rightarrow \infty$) for showing the asymptotically achievable performance of an estimator based on such cost function. Although the present analysis can be extended to more general host distributions, we have focused on i.i.d. Gaussian hosts in order to keep the derivations more compact and easy to interpret.

This chapter is structured as follows: Section 4.1 summarizes the previous approaches for the estimation of spreading vectors in spread spectrum modulations. In Section 4.2, the estimation problem is formulated, and the performance measures introduced. Section 4.3 analyzes the estimators proposed for the WOA scenario in [61]. New estimators for both the KMA and WOA scenarios are proposed in sections 4.4 and 4.5, respectively. The choice of numerical methods for optimizing the cost functions defined in these sections is discussed in Section 4.6. In Section 4.7, the performance of the proposed estimators is experimentally evaluated with real images. Finally, their application to more general problems is discussed in Section 4.7.3.

4.1 Related work and contributions of this thesis

Attacks to the security of spread-spectrum methods are aimed at estimating the pseudorandom spreading vector which is derived from the secret key. For the methods studied in the previous chapter, the correspondence between spreading vector and watermark is one to one (except for a scaling factor due to the sign of the embedded message or to the host rejection). A consequence of this correspondence between watermark and spreading vector is that most attacks previously proposed for watermark estimation are indeed attacks to security, such as the "Wiener filtering" attack [208] and the statistical averaging attack [88] (which typically needs a large number of marked signals to succeed). Related approaches using denoising techniques besides averaging are discussed in [103]. Another attempt at performing watermark estimation is due to Mihçak et al. in [167], where the authors estimate the watermark based on the fact that the components of the watermark vector take discrete values ($\pm\Delta$), paying special attention to the case where these values are repeated in blocks of a certain length. Under a Gaussian host assumption, the maximum a posteriori (MAP) estimate of the watermark is computed. The final aim of estimation attacks is to provide the information necessary to perform a "remodulation" attack [217] in order to remove the watermark.

The problem of watermark estimation in general scenarios (continuous-valued watermarks, decoding applications) remained unaddressed for some time. A maximum likelihood watermark estimator (assuming Gaussian-distributed host signals) is proposed in [61] for the add-SS scheme in the KMA scenario, whereas BSS techniques, namely Principal Component Analysis (PCA) and Independent Component Analysis (ICA) [140, 141] are used in more involved scenarios. The rationale behind PCA and ICA-based estimation is that the energy of the watermark is concentrated in one particular subspace; moreover, the latter takes advantage of non-Gaussianity of the message distribution and the independence between the embedded messages and the host images. An extension of this approach (focused on the WOA scenario) is considered in [60] using ICA jointly with the Expectation-Maximization (EM) algorithm [97] in order to reduce the computational complexity of the attack when the dimensionality of the spreading vector is very large. It is also pertinent to mention a simultaneous work, [102], in which the subspace generated by the secret key is estimated with PCA in order to remove the watermark. A previous work which used ICA to estimate the watermark signal, although without taking into account security considerations, is given in [104].

As for the main contributions of this chapter on this topic, they can be summarized as follows:

- The consideration, for the first time, of host rejection in the analysis of the practical estimators.
- A theoretical analysis of the estimators proposed in [61]. This analysis clarifies, for the first time, in which situations (depending on the embedding parameters) the attacker may expect a successful estimation when employing such estimators.
- The proposal and analysis of new estimators, and their application to practical scenarios. The main objective is to devise new estimators that work in scenarios where the approaches studied before fail, constituting this way a wider battery of methods for performing practical security tests.

4.2 Problem formulation

Hereinafter, for the sake of clarity, the spreading vector used by the watermarker and which the attacker wants to estimate will be denoted by $\mathbf{s}_0$. We will assume, for simplicity, that $\mathbf{s}_0$ is of unit norm. Hence, the embedding function that we consider is

$$\mathbf{y}_i = \mathbf{x}_i + (-1)^{m_i}\nu\mathbf{s}_0 - \lambda(\mathbf{x}_i^T\mathbf{s}_0)\mathbf{s}_0, \tag{4.1}$$

which is equivalent to (3.4) if we use $\nu = (n\sigma_S^2 - \lambda^2\sigma_X^2)^{\frac{1}{2}}$.

We are interested in obtaining an estimate of the spreading vector $\mathbf{s}_0$. Two different scenarios will be considered:

1. The “Known Message Attack scenario” (KMA), where the message embedded in each observation is known by the attacker. From the results of Chapter 3, we know that in this scenario it is possible to achieve a perfect estimate of $\mathbf{s}_0$. The performance of the estimator will be measured in terms of the normalized correlation:
$$\rho = \frac{\hat{\mathbf{s}}_0^T\mathbf{s}}{||\hat{\mathbf{s}}_0||\cdot||\mathbf{s}||} \in [-1, 1], \tag{4.2}$$
i.e. we look for the unit vector $\hat{\mathbf{s}}_0$ that minimizes the angle with $\mathbf{s}_0$.
2. The ’Watermarked Only Attack’ (WOA), where the embedded messages are not known. In this scenario it is possible to estimate $\mathbf{s}_0$ up to a sign ambiguity. In other words, it is only possible to estimate the subspace spanned by $\mathbf{s}_0$. Thus, the estimator’s performance will be measured by the absolute value of the normalized correlation, $|\rho|$.

In order to illustrate the performance of the analyzed methods, we will plot the cost functions versus ρ (or $|\rho|$) and the DWR. This yields a surface which, according to the conventional optimization terminology, will be referred to as the "cost surface".

This estimation problem strongly resembles the problem of linear equalization in Digital Communications [143], where the objective is to find the linear filter of a given length that best matches the inverse impulse response of the channel. Usually, this filter is obtained by optimizing of a suitable cost function that exploits the statistics of the channel and the underlying modulation scheme. Roughly speaking, the spreading vector of our problem plays the role of linear filter. In the KMA scenario, the estimator can exploit the knowledge of the embedded messages, i.e. the latter play the role of training sequences or pilot signals. The WOA scenario is clearly more related to the blind equalization paradigm.

4.3 Analysis of previous approaches for the WOA scenario

We analyze the estimation setup proposed in [61], based on Independent Component Analysis (ICA) and Principal Component Analysis (PCA). ICA and PCA, which are well known statistical tools for performing blind source separation (BSS) [140], were applied for the first time in [61] to the watermarking security problem for estimating spreading vectors in the WOA scenario.

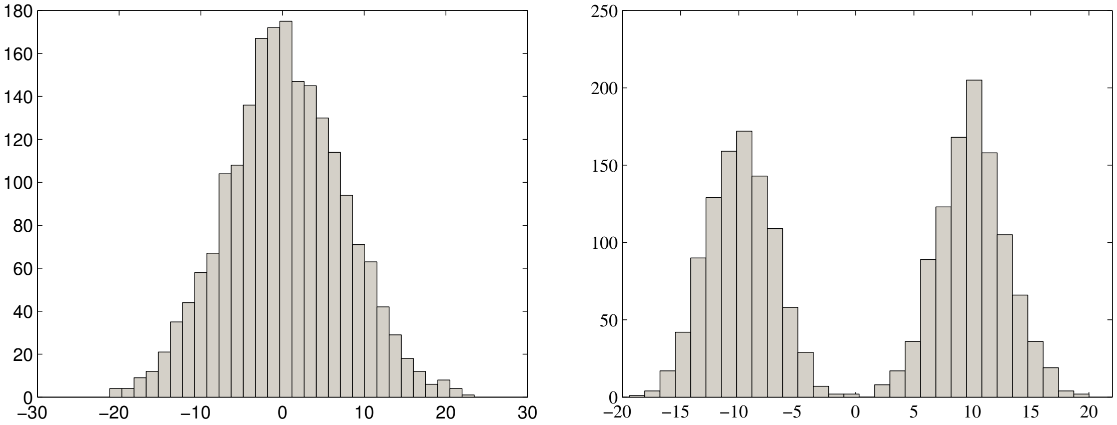
4.3.1 Independent Component Analysis (ICA)

ICA is a well known statistical tool for performing blind source separation. The idea behind ICA methods is to optimize a cost function that represents the mutual independence between the separated sources. The intuitive meaning of the negentropy resembles that of the fourth order cumulant, as both can be interpreted as measures of distance to the Gaussian distribution. Given a continuous random variable R with variance σ_R^2 and probability density function $f(r)$, its negentropy is defined as

$$\zeta(R) \triangleq h(R) - h(U), \tag{4.3}$$

where $h(\cdot)$ denotes differential entropy and $U \sim \mathcal{N}(0, \sigma_R^2)$. For implementing a practical ICA estimator based on negentropy maximization, the latter must be estimated from the observed data without knowing its probability density function, in general. The first negentropy estimators that were proposed were based on approximations of the differential entropy by means of high order cumulants [77]. In [138], it is shown that better approximations to the negentropy are of the form

$$\hat{\zeta}(R) = E\left[g(R)\right] - E\left[g(U)\right], \tag{4.4}$$



(a) (b)

Figure 4.1: Histogram of the projection $\mathbf{y}_i^T\mathbf{s}$ for $n = 300$, $\lambda = 0.6$, and DWR $= 21$ dB. The projection in (a) is for a randomly chosen vector $\mathbf{s}$, and the projection in (b) is for $\mathbf{s} = \mathbf{s}_0$.

where again $U \sim \mathcal{N}(0, \sigma_R^2)$ as in (4.3), and $g(\cdot)$ is the so-called "contrast function", which in practice can be almost any non-linear smooth function.

ICA is not restricted to the BSS paradigm, but it is often used as a tool for extracting "interesting" components of high-dimensional data, which is precisely the target pursued here. Indeed, under this point of view, ICA must be seen as a way of performing Projection Pursuit [137] rather than BSS. Figure 4.1 provides the main intuition on why the ICA approach should be valid for our problem. Figure 4.1(a) shows the histogram of the projection $\mathbf{y}_i^T\mathbf{s}$ when the vector $\mathbf{s}$ is randomly chosen (hence, approximately orthogonal to $\mathbf{s}_0$), and Figure 4.1(b) shows the same histogram when $\mathbf{s} = \mathbf{s}_0$. In the first case we see that the histogram of the projection strongly resembles a zero-mean Gaussian. However, in the second case, the histogram strongly diverges from the Gaussian (actually, it corresponds to a two-sided Gaussian pdf). Thus, deviation from the Gaussian distribution seems to be a good cost function for performing estimation of the secret spreading vector.

Among the variety of ICA tools existing in the literature, we will focus on the approach used in [61] and [44], where the ICA cost function for our problem is defined as [139]

$$J_{\text{ICA}}(\mathbf{s}) = \left(E\left[g(\mathbf{Y}^T\mathbf{s})\right] - E\left[g(U)\right]\right)^2, \tag{4.5}$$

where $U \sim \mathcal{N}(0, \text{var}(\mathbf{Y}^T\mathbf{s}))$, and $g(\cdot)$ is the so-called "contrast function". The ICA estimator results in

$$\hat{\mathbf{s}}_0 = \arg\max_{\mathbf{s}} J_{\text{ICA}}(\mathbf{s}). \tag{4.6}$$

Intuitively, (4.6) looks for the unit-norm vector $\mathbf{s}$ that maximizes the divergence between the distribution of $g(\mathbf{Y}^T\mathbf{s})$ and a Gaussian distribution with the same variance. The choice of the appropriate contrast function in (4.5) is largely application-dependent. Nevertheless, in informed applications where the statistical distribution of the independent components is known, the optimal choice is clearly $g(z) = \log(f(z))$, where $f(z)$ is the pdf of the independent component to be estimated, since it gives directly the entropy of Z. The optimality of this choice is also supported by [138] from the point of view of minimum asymptotic variance of the estimation error. For the problem we are considering, with an i.i.d. Gaussian host, the pdf of the component to be estimated is

$$\begin{aligned}
f(z) &= \frac{K}{2}\left(\exp\left(-\frac{(z-\nu)^2}{2\sigma_Z^2}\right) + \exp\left(-\frac{(z+\nu)^2}{2\sigma_Z^2}\right)\right) \\
&= \frac{K}{2}\exp\left(\frac{-z^2-\nu^2}{2\sigma_Z^2}\right)\left(\exp\left(\frac{z\nu}{\sigma_Z^2}\right) + \exp\left(-\frac{z\nu}{\sigma_Z^2}\right)\right) \\
&= K'\cdot\exp\left(-\frac{z^2}{2\sigma_Z^2}\right)\cosh\left(\frac{z\nu}{\sigma_Z^2}\right). \qquad (4.7)
\end{aligned}$$

where K' is a constant, and $\sigma_Z^2 = (1-\lambda)^2\sigma_X^2$. Hence, the optimal ICA cost function for our problem results in

$$\begin{aligned}
J_{\text{ICA}}(\mathbf{s}) &= \left(E\left[\log\cosh\left(\frac{\nu}{\sigma_Z^2}\mathbf{Y}^T\mathbf{s}\right)\right] - \frac{1}{2\sigma_Z^2}E\left[(\mathbf{Y}^T\mathbf{s})^2\right]\right. \\
&- \left. E\left[\log\cosh\left(\frac{\nu}{\sigma_Z^2}U\right)\right] + \frac{1}{2\sigma_Z^2}E\left[U^2\right]\right)^2 \\
&= \left(E\left[\log\cosh\left(\frac{\nu}{\sigma_Z^2}\mathbf{Y}^T\mathbf{s}\right)\right] - E\left[\log\cosh\left(\frac{\nu}{\sigma_Z^2}U\right)\right]\right)^2, \qquad (4.8)
\end{aligned}$$

where the second equality follows because the variance of U equals, by definition, the variance of $\mathbf{Y}^T\mathbf{s}$ and both are zero-mean (recall that $U \sim \mathcal{N}(0, \text{var}(\mathbf{Y}^T\mathbf{s}))$. It is interesting to note that log cosh is the contrast function recommended in [139] for general purposes, although the reasoning for arriving there is essentially different. Using the change of variable $a = \nu/\sigma_Z^2$, Eq. (4.8) can be rewritten as

$$J_{\text{ICA}}^{blind}(\mathbf{s}, a) = \left(E\left[\log\cosh(a\cdot\mathbf{Y}^T\mathbf{s})\right] - E\left[\log\cosh(a\cdot U)\right]\right)^2, \qquad (4.9)$$

where the parameter a can be fixed by the user. The notation "*blind*" is adopted for emphasizing the fact that, as in [61] and [44], the cost function does not depend on any parameter to be estimated from the observed data. It is interesting to note that the choice of the log cosh contrast function coincides with the recommendation in [139] for general purpose ICA, although the reasoning for arriving there is essentially different. The evaluation of (4.9) requires numerical integration. We have

$$\begin{aligned}\mathbf{Y}^T\mathbf{s} &= \mathbf{X}^T\mathbf{s} + \nu(-1)^M\rho - \lambda(\mathbf{X}^T\mathbf{s}_0)\rho \\ &= \mathbf{X}^T\left(\mathbf{s} - \lambda\rho\mathbf{s}_0\right) + \nu(-1)^M\rho = \mathbf{X}^T\mathbf{t} + \nu(-1)^M\rho. \end{aligned} \tag{4.10}$$

Assuming an i.i.d. Gaussian host,

$$\mathbf{Y}^T\mathbf{s}|M = m \sim \mathcal{N}(\nu(-1)^m\rho, \sigma_X^2||\mathbf{t}||^2), \tag{4.11}$$

where

$$||\mathbf{t}||^2 = ||\mathbf{s} - \lambda\rho\mathbf{s}_0||^2 = 1 + \rho^2(\lambda^2 - 2\lambda). \tag{4.12}$$

Therefore, the term $\mathbf{Y}^T\mathbf{s}$ in (4.9) is a binary Gaussian mixture,

$$\mathbf{Y}^T\mathbf{s} \sim \frac{1}{2}\left(\mathcal{N}(\nu\rho, \sigma_X^2||\mathbf{t}||^2) + \mathcal{N}(-\nu\rho, \sigma_X^2||\mathbf{t}||^2)\right), \tag{4.13}$$

with $||\mathbf{t}||^2$ given by (4.12), which is coherent with the situation shown in Figure 4.1. Notice that the cost function depends solely on the value of ρ, not on the particular realization of $\mathbf{s}_0$.

If we use the approximation $\log\cosh(x) \approx |x|$, a tight closed form approximation of (4.9) becomes

$$J_{\mathrm{ICA}}^{blind}(\mathbf{s}, a) \approx \left(\sqrt{\frac{2}{\pi}}\sigma_1 \exp\left(\frac{-\beta^2}{2\sigma_1^2}\right) + \beta\cdot\operatorname{erf}\left(\frac{\beta}{\sqrt{2}\sigma_1}\right) - \sqrt{\frac{2}{\pi}}\sigma_2\right)^2, \tag{4.14}$$

where

$$\begin{aligned}\sigma_1 &= a\sigma_X(1 + \rho^2(\lambda^2 - 2\lambda))^{\frac{1}{2}}, \\ \sigma_2 &= a(\sigma_X^2 + \nu^2\rho^2 + \lambda^2\rho^2\sigma_X^2 - 2\lambda\rho^2\sigma_X^2)^{\frac{1}{2}}, \\ \beta &= a\nu\rho.\end{aligned}$$

Figure 4.2 shows the cost function versus $|\rho|$ and the DWR for $a = 1$ and $\lambda = 0.5$. For each DWR, the plots have been normalized by the maximum value of the cost function for ease of comparison. In practice, this normalized cost function has shown to be virtually insensitive to the chosen value of a. Although $J_{\mathrm{ICA}}^{blind}(\mathbf{s}, a)$ is convex and its maximum is clearly located at $|\rho| = 1$ as desired, this cost function is not well suited to practical applications where the DWR is moderately high, because of its remarkable flatness for small ρ. Due to this flatness, the initial vector must be very close to $\mathbf{s}_0$ for assuring convergence when the cost function is to be optimized iteratively. Indeed, in real experiments the ICA estimator has shown to get stuck most of the times at $\rho \approx 0$.

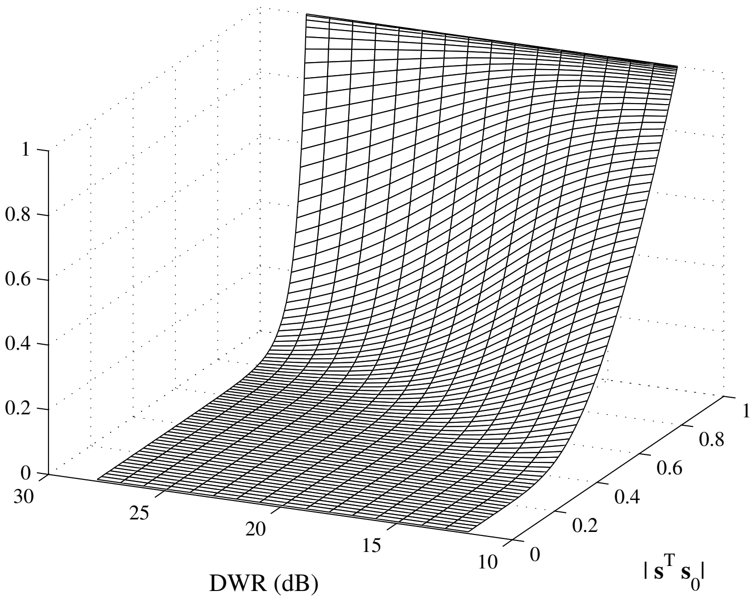


Figure 4.2: Cost surface of blind ICA, for $n = 200$, $\lambda = 0.5$, and $a = 1$.

4.3.2 ICA with pre-whitening

We must notice that the definition of the ICA estimator (4.6) does not consider any kind of preprocessing of the observations. However, in most ICA implementations, the observations are whitened before applying the ICA estimator (this is also the case of [139]). There are several reasons for applying this whitening, being the most important the fact that this operation reduces the estimation of the whole mixing matrix in the BSS problem to the estimation of a rotation matrix, and that it brings the problem to more "controlled" conditions, since the statistics of the problem are unknown, in general. In our particular problem of carrier estimation there is no justified reason for whitening the observations. Thus, we are interested in checking its impact in the performance of the ICA estimator. The new cost function is

$$J_{\text{ICA}}^{white}(\mathbf{s}, a) = \left(E\left[\log\cosh(a \cdot \mathbf{Y}_w^T \mathbf{s})\right] - E\left[\log\cosh(a \cdot U)\right]\right)^2, \tag{4.15}$$

where $U \sim \mathcal{N}(0, 1)$, and $\mathbf{Y}_w^T$ represents the whitened observations. We consider whitening through linear Principal Component Analysis (PCA), so $\mathbf{Y}_w \triangleq \mathbf{PY}$, where $\mathbf{P}$ denotes the "whitening matrix". Let us denote by $\mathbf{Q}$ the covariance matrix of the observations. For $N_o \to \infty$, an eigenvalue decomposition yields $\mathbf{Q} = \mathbf{VDV}^T$, with

$$\mathbf{V} = [\mathbf{s}_0, \mathbf{V}_{\mathbf{s}_0}] \in \mathbb{R}^{n \times n}, \qquad \mathbf{D} = \begin{bmatrix} \nu^2 + (1-\lambda)^2 \sigma_X^2 & 0 \\ 0 & \sigma_X^2 \cdot \mathbf{I}_{n-1} \end{bmatrix}, \tag{4.16}$$

where $\mathbf{V}_{\mathbf{s}_0} \in \mathbb{R}^{n\times n-1}$ is a unitary matrix whose columns span the orthogonal complement of the subspace spanned by $\mathbf{s}_0$. Thus, for $N_o \to \infty$, the whitening matrix can be written as $\mathbf{P} = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^T$. The statistics of $\mathbf{Y}_w^T\mathbf{s}$ can be computed by taking into account that

$$\mathbf{Y}_w^T\mathbf{s} = \mathbf{X}^T(\mathbf{P}\mathbf{s} - \lambda\mathbf{s}_0(\mathbf{s}_0^T\mathbf{P}\mathbf{s})) + \nu(-1)^M\mathbf{s}_0^T\mathbf{P}\mathbf{s} = \mathbf{X}^T\mathbf{q} + \nu(-1)^M\mathbf{s}_0^T\mathbf{P}\mathbf{s}. \tag{4.17}$$

By realizing that $\mathbf{s}_0^T\mathbf{P}\mathbf{s} = \rho(\nu^2 + (1-\lambda)^2\sigma_X^2)^{-\frac{1}{2}}$, we have

$$\mathbf{Y}_w^T\mathbf{s}|M = m \sim \mathcal{N}((-1)^m\nu\rho(\nu^2 + (1-\lambda)^2\sigma_X^2)^{-\frac{1}{2}}, \sigma_X^2||\mathbf{q}||^2), \tag{4.18}$$

where

$$||\mathbf{q}||^2 = ||\mathbf{P}\mathbf{s} - \lambda\mathbf{s}_0(\mathbf{s}_0^T\mathbf{P}\mathbf{s})||^2 = \mathbf{s}^T\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T\mathbf{s} + \rho^2(\nu^2 + (1-\lambda)^2\sigma_X^2)^{-1}(\lambda^2 - 2\lambda). \tag{4.19}$$

The computation of (4.15), taking into account (4.18) and (4.19), requires again numerical integration. Figure 4.3 compares $J_{\text{ICA}}^{white}(\mathbf{s}, a)$ with $J_{\text{ICA}}^{blind}(\mathbf{s}, a)$.[1] As can be seen, $J_{\text{ICA}}^{white}(\mathbf{s}, a)$ is worse than $J_{\text{ICA}}^{blind}(\mathbf{s}, a)$ (for that particular DWR) in practical terms, because of its remarkable flatness: convergence is only guaranteed for initialization vectors close to $\mathbf{s}_0$. However, it is interesting to note that $J_{\text{ICA}}^{white}(\mathbf{s}, a)$, after being normalized, is nearly invariant to the embedding parameters and to the free parameter a.

4.3.3 Principal Component Analysis (PCA)

As can be guessed from Figure 4.2(a), $J_{\text{ICA}}^{blind}(\mathbf{s}, a)$ becomes a suitable cost function when the DWR is small. Hence, if some preprocessing for reducing the effective DWR can be made, then $J_{\text{ICA}}^{blind}(\mathbf{s}, a)$ can be used for our estimation purposes. Being probably aware of this fact, the authors in [61] propose a dimensionality reduction through Principal Component Analysis (PCA). This is based on the following rationale: for large spreading sequences, the variance in the direction of the watermark dominates over the remaining directions, so for carrier estimation it suffices to keep for the ICA only the components of the observations in the direction of the eigenvectors with the largest associated eigenvalues. This reasoning holds in many practical scenarios, especially in watermark detection applications, where n uses to be in the order of thousands. However, when considering data hiding applications, the situation may be different.

When a single carrier is being used, the estimator of $\mathbf{s}_0$ by PCA is simply given by [61]

$$\hat{\mathbf{s}}_0 = \mathbf{V}[\arg\max_i D_{i,i}], \tag{4.20}$$

[1] The other plot in Figure 4.3 corresponds to the “informed ICA” estimator, which will be defined in Section 4.5.1.

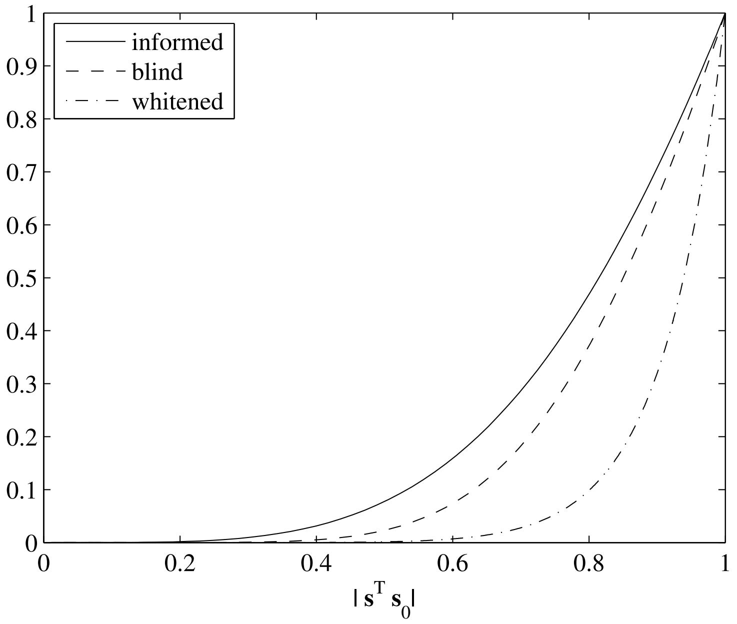


Figure 4.3: Comparison of the ICA cost functions with $a = 1$, for ISS with $n = 200$, $\lambda = 0.5$ and DWR = 16 dB.

where $\mathbf{V}[k]$ denotes the kth column of the matrix $\mathbf{V}$, and $D_{i,i}$ is the ith element in the diagonal of the matrix $\mathbf{D}$, both defined in (4.16). The performance of this estimator, which will be referred to as the "PCA estimator", was already plotted for add-SS (i.e. with $\lambda = 0$) in Figure 3.6(b), where we can see that it works remarkably well for not so large values of n. In order to perform a correct estimate, the variance in the direction of $\mathbf{s}_0$ must be larger than in the remaining directions. For the i.i.d. Gaussian host, this is equivalent to

$$\nu^2 + (1-\lambda)^2\sigma_X^2 > \sigma_X^2 \Leftrightarrow \text{DWR} < 10\log_{10}\left(\frac{n}{2\lambda}\right) \tag{4.21}$$

as can be seen from (4.16). Note that if the pdf of the host signal is not circular (i.e. if the covariance matrix is not diagonal), then the condition for ensuring correct estimation is more restrictive, as one must take into account the directions with the highest variance. In any case, the watermarker could easily fool the PCA estimator by properly tuning the embedding parameters n and λ so as to reduce the variance in the embedding direction (possibly loosing some robustness).

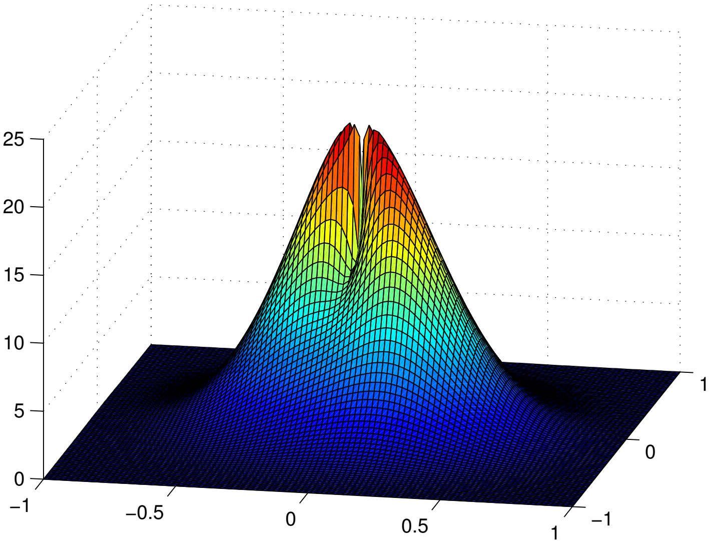


Figure 4.4: Probability density function of the carrier conditioned on one observation for ISS-KMA with $n = 2$ and $\lambda = 0.9$.

4.4 New estimator for the KMA scenario

This scenario was considered in [61, Section V.A] for $\lambda = 0$ under the i.i.d. Gaussian host assumption, deriving the maximum likelihood (ML) estimator of $\mathbf{s}_0$ for that case. Unfortunately, when $\lambda > 0$, no closed-form expression exists for the ML or MMSE estimators, even under the Gaussian host assumption (Figure 4.4 shows an example of the conditional pdf of $\mathbf{s}_0$ for $\lambda = 0.9$ and $n = 2$). For overcoming this problem we will consider the pdf of the observations when projected onto the subspace defined by $\mathbf{s}_0$. Let us define the variables $z_i \triangleq \mathbf{y}_i^T \mathbf{s}_0,\ i = 1, \ldots, N_o$. Since the $\mathbf{y}_i$ are conditionally independent when the m_i are known, we have

$$f((z_1, \ldots, z_{N_o})^T | \mathbf{S} = \mathbf{s}_0) = K_1 \cdot \exp\left(-K_2 \cdot \sum_{i=1}^{N_o} (z_i - \mu_i)^2 \right), \tag{4.22}$$

where K_1, K_2 are constants, and $\mu_i \triangleq \nu(-1)^{m_i} \|\mathbf{s}_0\|^2$, for $i = 1, \ldots, N_o$. It is easy to show that the ML estimator of $\mathbf{s}_0$ under the model (4.22) is the solution to the linear system $\mathbf{A}\hat{\mathbf{s}}_0 = \mathbf{M}$, which is expressed as

$$\begin{bmatrix} \sum_{i=1}^{N_o} y_{i,1}^2 & \sum_{i=1}^{N_o} y_{i,2}\cdot y_{i,1} & \cdots & \sum_{i=1}^{N_o} y_{i,n}\cdot y_{i,1} \\ \sum_{i=1}^{N_o} y_{i,1}\cdot y_{i,2} & \sum_{i=1}^{N_o} y_{i,2}^2 & \cdots & \sum_{i=1}^{N_o} y_{i,n}\cdot y_{i,1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N_o} y_{i,1}\cdot y_{i,n} & \sum_{i=1}^{N_o} y_{i,2}\cdot y_{1,n} & \cdots & \sum_{i=1}^{N_o} y_{i,n}^2 \end{bmatrix} \times \begin{bmatrix} \hat{s}_{0,1} \\ \hat{s}_{0,2} \\ \vdots \\ \hat{s}_{0,n} \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{i=1}^{N_o} \mu_i \cdot y_{i,1} \\ \sum_{i=1}^{N_o} \mu_i \cdot y_{i,2} \\ \vdots \\ \sum_{i=1}^{N_o} \mu_i \cdot y_{i,n} \end{bmatrix}, \qquad (4.23)$$

where $y_{i,j}$ is the jth component of the ith observation. This estimator is easy to compute by matrix inversion or by numerical methods. However, as a penalty for using the statistics of the host in the projected domain, we need at least n observations in order to get a full-rank matrix $\mathbf{A}$. Note that the hypothesized values of ν and $||\mathbf{s}_0||$ do not affect the performance of the estimator, since they would only produce a scaling of the estimate. This estimator can be thought of as a regressor that tries to estimate the supporting hyperplane of the scaled observations $(-1)^{m_i}\mathbf{y}_i$, which is given by the equation $\mathbf{v}^T\mathbf{s}_0 = \nu||\mathbf{s}_0||^2$. For $\lambda = 1$, all the observations $(-1)^{m_i}\mathbf{y}_i$ fall exactly on this hyperplane. Hence, it is expected that the performance of this estimator is improved as λ approaches 1 (moreover, it is obvious that in this case $n-1$ observations would suffice for obtaining a perfect estimate of $\mathbf{s}_0$).

Notice that the validity of this estimator is not necessarily restricted to Gaussian hosts. It will work for a variety of host distributions, as long as the Central Limit Theorem (CLT) holds. The reason is that, for moderately large n, Eq. (4.22) is approximately valid by virtue of the CLT.

4.5 New estimators for the WOA scenario

4.5.1 Informed ICA

We have empirically observed that the drawbacks mentioned in Sect. 4.3.1 for the blind ICA estimator can be partially overcome if the variance of the random variable U in $J_{\text{ICA}}^{blind}(\mathbf{s}, a)$ is fixed to a proper constant value, instead of varying it according to the variance of $\mathbf{Y}^T\mathbf{s}$. We term the new cost function $J_{\text{ICA}}^{inf}(\mathbf{s}, a)$, where "*inf*" stands for "informed". After some experiments, it was found that fixing that value to σ_X^2 yields good results. In order to keep the estimator blind, we compute an estimate of σ_X from

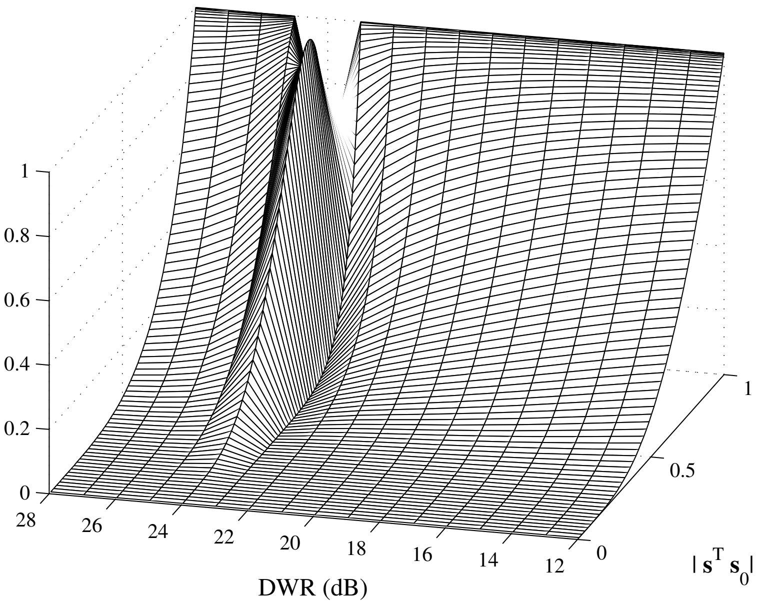


Figure 4.5: Cost surface of informed ICA, for $n = 200$, $\lambda = 0.5$ and $a = 1$.

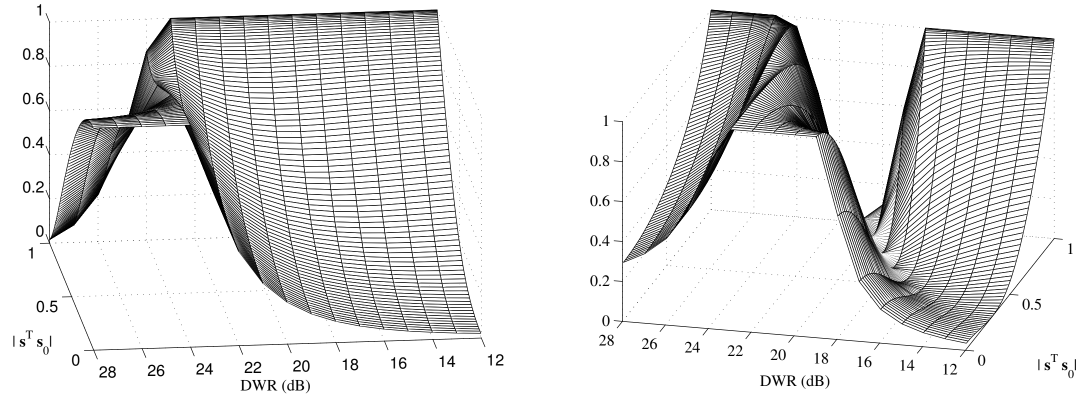
the observations. Our estimate of σ_X is

$$\hat{\sigma}_X = \left(\frac{1}{n}\mathrm{tr}\left(\mathbf{Q}\right)\right)^{\frac{1}{2}} = \left(\frac{1}{n}\sum_{i=1}^{n} D_{i,i}\right)^{\frac{1}{2}}, \tag{4.24}$$

where $\mathbf{Q}$ is the covariance matrix of the observations and $D_{i,i}$ are the diagonal elements of the matrix of eigenvalues defined in (4.16). The expression of $J_{\mathrm{ICA}}^{inf}(\mathbf{s}, a)$ is still given by (4.5), with the only difference that $U \sim \mathcal{N}(0, \hat{\sigma}_X^2)$.

Figure 4.5 shows the cost surface of $J_{\mathrm{ICA}}^{inf}(\mathbf{s}, a)$ (obtained by means of numerical integration) under the same conditions as $J_{\mathrm{ICA}}^{blind}(\mathbf{s}, a)$ in Figure 4.2. As can be seen, $J_{\mathrm{ICA}}^{inf}(\mathbf{s}, a)$ is convex for all DWRs except for a small range, and no problems of flatness for small ρ appear, so an iterative optimization algorithm can easily reach the maximum. Although $J_{\mathrm{ICA}}^{inf}(\mathbf{s}, a)$ has shown to be quite sensitive to value of the variance we fix for U (two examples of the resulting cost surface for underestimation and overestimation of σ_X are shown in Figure 4.6), it performs reasonably well with real images, as will be seen in Section 4.7). Similarly to $J_{\mathrm{ICA}}^{blind}(\mathbf{s}, a)$, the value of a has no noticeable influence on the normalized cost function.

Informed ICA is also compared with blind ICA and blind ICA with pre-whitening in Figure 4.3, showing that the first is better behaved.



(a) $\hat{\sigma}_X = 0.5\sigma_X$

(b) $\hat{\sigma}_X = 1.5\sigma_X$

Figure 4.6: Cost surface $J_{\text{ICA}}^{inf}(\mathbf{s}, a)$ with $a = 1$ for ISS with $n = 200$, $\lambda = 0.5$ and different estimates of the host variance.

4.5.2 Constant Modulus (CM) criterion

According to (4.13), when $\mathbf{Y}$ is correlated with a vector $\mathbf{s}$ orthogonal to $\mathbf{s}_0$ (i.e. with $\rho = 0$), the resulting random variable is zero-mean Gaussian with variance σ_X^2. However, when $\mathbf{Y}$ is correlated with $\mathbf{s}_0$, we have

$$\mathbf{Y}^T\mathbf{s}_0 \sim \frac{1}{2}\left(\mathcal{N}(\nu, (1-\lambda)^2\sigma_X^2) + \mathcal{N}(-\nu, (1-\lambda)^2\sigma_X^2)\right).$$

Hence, if we take the modulus of $\mathbf{Y}^T\mathbf{s}$, it should lie (in average) closer to ν as $\mathbf{s}$ approximates $\mathbf{s}_0$. Notice, for instance, that in the particular case of $\lambda = 1$ the observations $\mathbf{y}_i$ fall exactly on two parallel hyperplanes given by $\mathbf{v}^T\mathbf{s}_0 = \pm\nu||\mathbf{s}_0||^2$. The main idea behind the CM method is to define a cost function that penalizes the deviations of $|\mathbf{Y}^T\mathbf{s}|$ from ν. A possible cost function is then

$$J_{\text{CM}}(\mathbf{s}) \;=\; E\left[\left((\mathbf{Y}^T\mathbf{s})^2 - \nu^2\right)^2\right] = E\left[(\mathbf{Y}^T\mathbf{s})^4\right] - 2\nu^2 \cdot E\left[(\mathbf{Y}^T\mathbf{s})^2\right] + \nu^4. \quad (4.25)$$

By definition, our CM estimator is essentially equivalent to the methods for blind equalization based on the constant modulus criterion which are well known in the literature of Digital Communications [143] and have been extensively studied there. Nevertheless, the problem setup is different in both cases (e.g. host rejection is not considered in Communications, and the observations are not necessarily processed in a block-by-block basis), so a new analysis of the CM cost function for our problem is justified.

The cost function of the CM estimator is analyzed in Appendix B.6 assuming an i.i.d. Gaussian host. The cost function results in a fourth order degree polynomial in the normalized correlation,

$$\begin{aligned} J_{\mathrm{CM}}(\mathbf{s}) &= J_{\mathrm{CM}}(\rho) \\ &= \rho^4\left(\nu^4 - 12\nu^2\lambda\sigma_X^2 + 6\nu^2\lambda^2\sigma_X^2 + 12\lambda^2\sigma_X^4 - 12\lambda^3\sigma_X^4 + 3\lambda^4\sigma_X^4\right) \\ &\quad -2\rho^2\left(\nu^2 - 3\sigma_X^2\right)\left(\nu^2 - 2\lambda\sigma_X^2 + \lambda^2\sigma_X^2\right) + 3\sigma_X^4 - 2\nu^2\sigma_X^2 + \nu^4. \end{aligned} \tag{4.26}$$

In the particular case of $\lambda = 0$ (add-SS), Eq. (4.31) admits a clearer expression:

$$J_{\mathrm{CM}}^{blind}(\mathbf{s}, \hat{\nu})\,|_{\lambda=0} \;=\; \rho^4\nu^4 - 2\rho^2\nu^2(\nu^2 - 3\sigma_X^2) + 3\sigma_X^4 - 2\nu^2\sigma_X^2 + \nu^4. \tag{4.27}$$

Now let us denote by ρ_{min} the value of ρ for which the global minimum of (4.27) is achieved. From the attacker's point of view, it is desirable that $|\rho_{min}|$ is as close to 1 as possible. For $\nu \leq \sqrt{3}\sigma_X$, the minimum of (4.27) is always achieved for $\rho_{min} = 0$. In turn, for $\nu > \sqrt{3}\sigma_X$, if the attacker wants to achieve $|\rho_{min}| \geq \tau \in [0, 1]$, then the condition

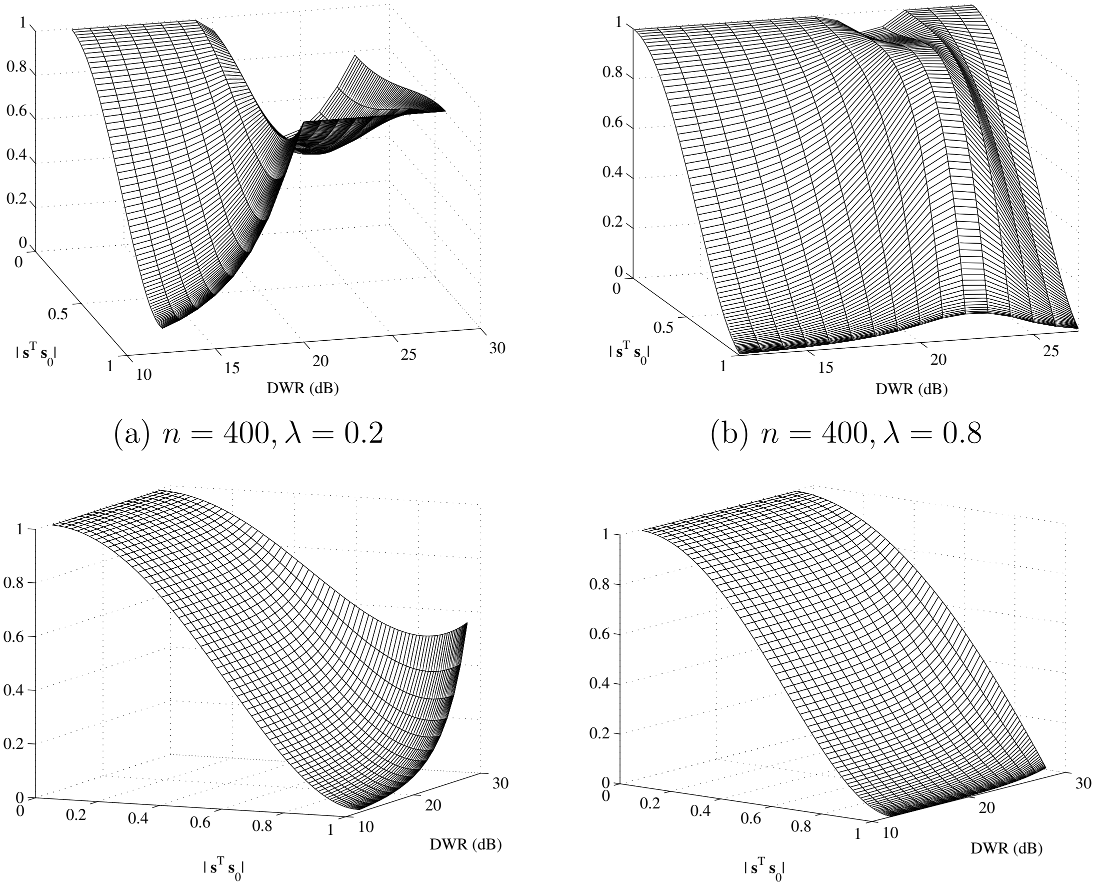
$$\mathrm{DWR} < 10\log_{10}\left(\frac{n(1-\tau^2)}{3}\right) \tag{4.28}$$

must hold. This means that for achieving $|\rho_{min}| \geq 0.9$ the DWR must be below 1 dB and 18 dB for $n = 20$ and $n = 1000$, respectively. Bear in mind that, when optimizing iteratively the CM cost function, the condition derived above is necessary but not sufficient, in general, for arriving at the optimum that guarantees $|\rho_{min}| > \tau$, since (4.27) is not necessarily monotonically decreasing for $|\rho| \in [0, 1]$.

For $\lambda > 0$ it is hard to infer from (4.26) the conditions for the successful estimate of $\mathbf{s}_0$, so we will plot the cost surface for different embedding parameters. Figure 4.7 shows the cost surface (4.26) for different values of n. The value of (4.26) has been normalized for each DWR by the value of its maximum. As can be seen, the cost surface is not guaranteed to be neither monotonic nor convex for all DWRs. Depending on the values of n and λ, there exists a range of DWRs where the global minimum of the cost surface may be far from $|\rho| = 1$; in the worst cases, the minimum is achieved for $\rho = 0$). Hence, the results of an iterative optimization algorithm will be strongly dependent on the initialization point. Notice that the shape of the cost surface near $\rho = 0$ is of utmost importance, since random initializations of $\mathbf{s}$ will yield $\mathbf{s}^T\mathbf{s}_0 \approx 0$ with high probability, so an iterative algorithm will get stuck at $\rho \approx 0$ if the cost surface achieves its minimum at that point. Nevertheless, Figure 4.7 shows that the cost surface becomes more amenable to be optimized as n and λ are increased (keeping the DWR constant).

Blind CM estimator

In general, the attacker has no a priori information about the embedding parameter ν, so he needs an estimate before applying the CM method. We assume now that the



(a) $n = 400, \lambda = 0.2$

(b) $n = 400, \lambda = 0.8$

(c) $n = 4000, \lambda = 0.2$

(d) $n = 4000, \lambda = 0.8$

Figure 4.7: Cost surface of the CM method for ISS with different embedding parameters.

attacker manages to obtain an estimate $\hat{\nu} = \nu + \tilde{\nu}$ from the observations at hand. The new CM cost function, now termed "blind", is

$$\begin{aligned} J_{\mathrm{CM}}^{blind}(\mathbf{s}, \hat{\nu}) &= E\left[(\mathbf{Y}^T\mathbf{s})^4\right] - 2\hat{\nu}^2 \cdot E\left[(\mathbf{Y}^T\mathbf{s})^2\right] + \hat{\nu}^4 \\ &= J_{\mathrm{CM}}(\mathbf{s}) - 2E\left[(\mathbf{Y}^T\mathbf{s})^2\right]\left(\tilde{\nu}^2 + 2\nu\tilde{\nu}\right) + \hat{\nu}^4 - \nu^4, \end{aligned} \tag{4.29}$$

and the blind CM estimator is defined as

$$\hat{\mathbf{s}}_0 = \arg\min_{\mathbf{S}} J_{\mathrm{CM}}^{blind}(\mathbf{s}, \hat{\nu}). \tag{4.30}$$

By combining the Eq. (4.26) above and (B.72) obtained in Appendix B.6, after rearranging terms (4.29) results in

$$\begin{aligned} J_{\mathrm{CM}}^{blind}(\mathbf{s},\hat{\nu}) &= J_{\mathrm{CM}}^{blind}(\rho,\hat{\nu}) \\ &= \rho^4\left(\nu^4-12\nu^2\lambda\sigma_X^2+6\nu^2\lambda^2\sigma_X^2+12\lambda^2\sigma_X^4-12\lambda^3\sigma_X^4+3\lambda^4\sigma_X^4\right) \\ &\quad -2\rho^2\left(\nu^2-3\sigma_X^2+2\nu\tilde{\nu}+\tilde{\nu}^2\right)\left(\nu^2-2\lambda\sigma_X^2+\lambda^2\sigma_X^2\right) \\ &\quad +3\sigma_X^4-2\nu^2\sigma_X^2+\nu^4+4\nu^3\tilde{\nu}+6\nu^2\tilde{\nu}^2+4\nu\tilde{\nu}^3+\tilde{\nu}^4-4\nu\tilde{\nu}\sigma_X^2-2\tilde{\nu}^2\sigma_X^2. \end{aligned} \tag{4.31}$$

The only difference between (4.31) and (4.26) is a constant additive factor and the term $2\nu\tilde{\nu}+\tilde{\nu}^2$ in the multiplier of ρ^2. Although underestimating the true value of ν may have a harmful effect in the cost function, using $\hat{\nu} > \nu$ can even be beneficial for the attacker. Figure 4.8 shows examples of the cost surface with $\hat{\nu} = 1.5\nu$ and $\hat{\nu} = 0.5\nu$. As can be seen in Figure 4.8(b), the local minima that are apparent in Figure 4.8(a) have disappeared; the penalty to pay is the general flattening of the cost surface that may affect the convergence of iterative optimization algorithms. For a truly "blind CM" estimator we propose to extract $\hat{\nu}$ from $\mathbf{Q}$, the covariance matrix of the observations, as

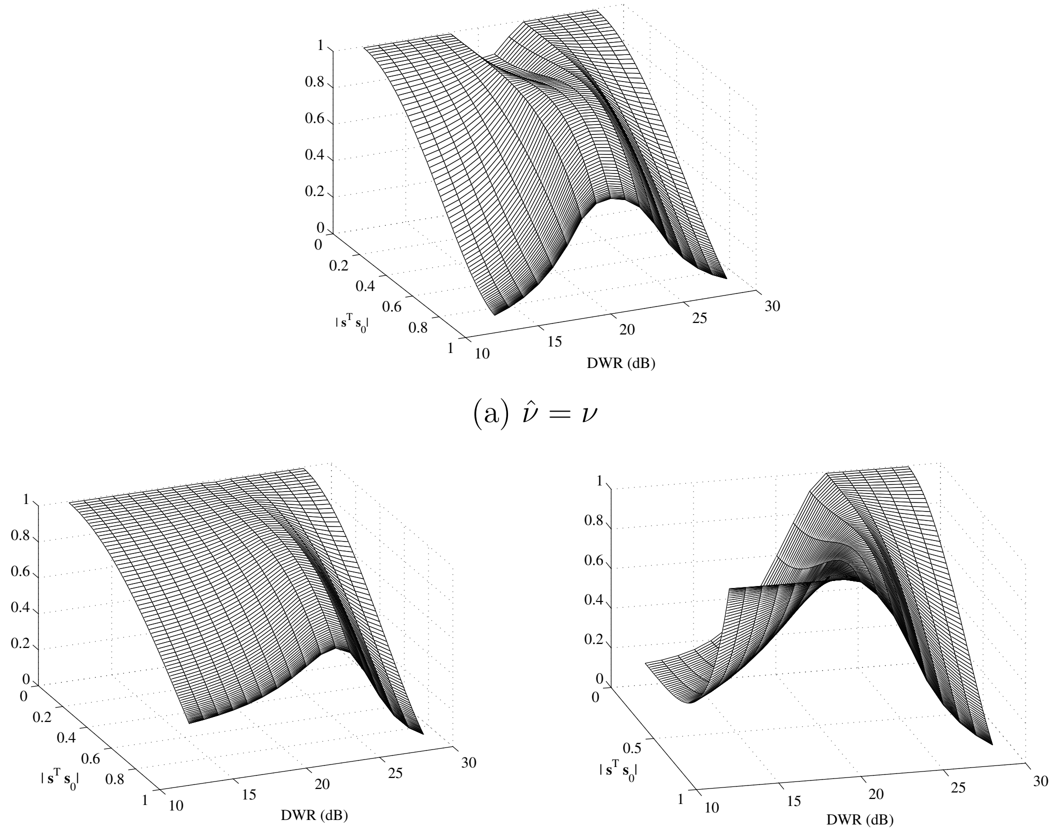
$$\hat{\nu}_{blind} \triangleq \max_i \left\{ (D_{i,i})^{\frac{1}{2}} \right\}, \tag{4.32}$$

where $D_{i,i}$ are the diagonal elements of $\mathbf{D}$, defined in (4.16). The resulting cost surface of the blind CM estimator is shown in Figure 4.9 for $\lambda = 0.5$. Notice that for $n = 200$ there exists a small range of DWRs (around 20 dB) with a local minimum close to $|\rho| = 0$. For those DWRs, an iterative optimization of the cost function would get stuck at $|\rho| \approx 0$ with high probability. However, for the remaining DWRs, the shape of the cost function is very appealing. Increasing the parameter n (Figure 4.9(b)) has the effect of increasing the range of DWRs where the function is unimodal (i.e. with no local minima) and achieves its minimum for $|\rho| = 1$. As can be seen, the range of DWRs with undesired local minima is moved towards higher DWRs. Figure 4.10 depicts the variation of J_{CM}^{blind} with λ, showing that increasing λ clearly benefits the attacker using this method.

4.5.3 Average Maximum Likelihood (AML)

Given a sequence of observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_o}\}$, a candidate vector $\mathbf{s}$, and the sequence of messages $\{m_1, \ldots, m_{N_o}\}$ embedded in each observation, the likelihood of the cross-correlation between the observations and $\mathbf{s}$ is given by

$$f(\mathbf{z}|\mathbf{s},\mathbf{m}) = K \cdot \prod_{i=1}^{N_o} \exp\left(-\frac{(z_i - \nu(-1)^{m_i})^2}{2\sigma_Z^2} \right), \tag{4.33}$$



(a) $\hat{\nu} = \nu$

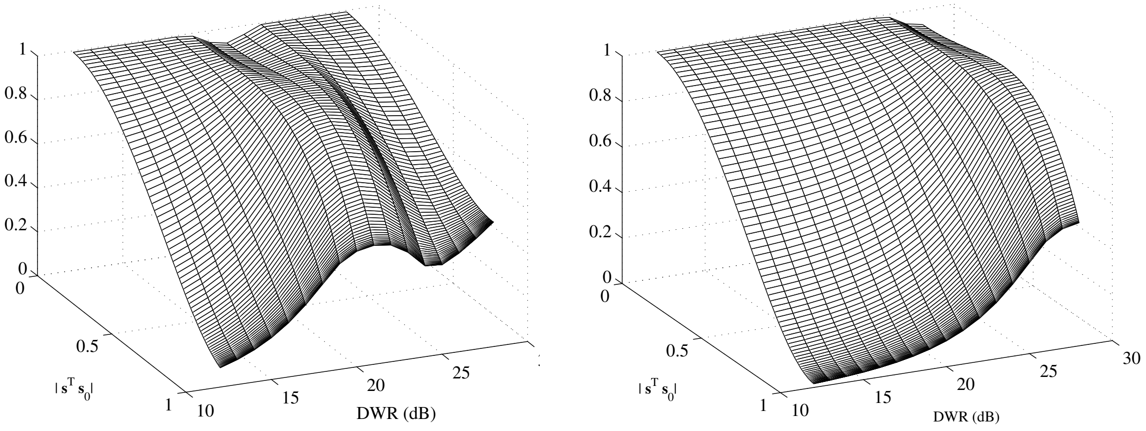
(b) $\hat{\nu} = 1.5\nu$

(c) $\hat{\nu} = 0.5\nu$

Figure 4.8: Cost surface of the blind CM method for ISS with $n = 200$ and $\lambda = 0.5$, for different estimates of ν.

where $\mathbf{z} = \{z_1, \ldots, z_{N_o}\}$, with $z_i \triangleq \mathbf{y}_i^T \mathbf{s}$, K is a constant, and $\sigma_Z^2 = (1-\lambda)^2\sigma_X^2$. Since we do not know the actual sequence of embedded messages, we take the average of the likelihood function:

$$
\begin{aligned}
f(\mathbf{z}|\mathbf{s}) &= K \cdot \prod_{i=1}^{N_o} E\left[\exp\left(-\frac{(z_i - \nu(-1)^{M_i})^2}{2\sigma_Z^2}\right)\right] \\
&= K \cdot \prod_{i=1}^{N_o} \left(\frac{1}{2}\exp\left(-\frac{(z_i-\nu)^2}{2\sigma_Z^2}\right) + \frac{1}{2}\exp\left(-\frac{(z_i+\nu)^2}{2\sigma_Z^2}\right)\right) \\
&= K \cdot \prod_{i=1}^{N_o} \exp\left(\frac{-z_i^2 - \nu^2}{2\sigma_Z^2}\right)\left(\frac{1}{2}\exp\left(\frac{z_i\nu}{\sigma_Z^2}\right) + \frac{1}{2}\exp\left(-\frac{z_i\nu}{\sigma_Z^2}\right)\right) \\
&= K' \cdot \prod_{i=1}^{N_o} \exp\left(\frac{-z_i^2}{2\sigma_Z^2}\right)\cosh\left(\frac{z_i\nu}{\sigma_Z^2}\right). \qquad (4.34)
\end{aligned}
$$



(a) $n = 200$

(b) $n = 1000$

Figure 4.9: Cost surface of the blind CM method for ISS with $\lambda = 0.5$, using the estimate (4.32) for parameter ν.

The average log-likelihood results in

$$\log(f(\mathbf{z}|\mathbf{s})) = \log(K') + \sum_{i=1}^{N_o} \log\cosh\left(\frac{z_i \nu}{\sigma_Z^2}\right) - \sum_{i=1}^{N_o} \frac{z_i^2}{2\sigma_Z^2}. \tag{4.35}$$

If we consider $\frac{1}{N_o}\log(f(\mathbf{z}|\mathbf{s}))$ for $N_o \to \infty$, then we can write the AML cost function as

$$J_{\mathrm{AML}}(\mathbf{s}) = E\left[\log\cosh\left(\frac{\nu \mathbf{Y}^T\mathbf{s}}{\sigma_Z^2}\right)\right] - \frac{1}{2\sigma_Z^2} E\left[(\mathbf{Y}^T\mathbf{s})^2\right]. \tag{4.36}$$

Thus, the AML estimator is given by

$$\hat{\mathbf{s}}_0 = \arg\max_{\mathbf{s}} J_{\mathrm{AML}}(\mathbf{s}). \tag{4.37}$$

For a i.i.d. Gaussian host, the second term of (4.36) has been already calculated in Appendix B.6. The first term has to be calculated by numerical integration. The pdf of $\mathbf{Y}^T\mathbf{s}$ has been obtained in (4.13). Now, if we define the r.v. $A_m \triangleq \frac{\nu}{\sigma_Z^2}\mathbf{Y}^T$, then

$$\frac{\nu}{\sigma_Z^2}\mathbf{Y}^T\mathbf{s}|M = m \sim \mathcal{N}\left(\frac{\nu^2(-1)^m\rho}{\sigma_Z^2}, \frac{\sigma_X^2\nu^2||\mathbf{t}||^2}{\sigma_Z^4}\right), \tag{4.38}$$

the value of the first term in (4.36) can be readily computed by means of numerical integration. Figure 4.11 shows the resulting cost surface. As can be seen, the cost surface is monotonically increasing in $|\rho|$ for all DWRs. However, as the DWR is increased the cost surface gets flatter, which constitutes a problem in practical setups. In general, a practical AML estimator working with high DWRs needs a large number of observations in order to perform well.

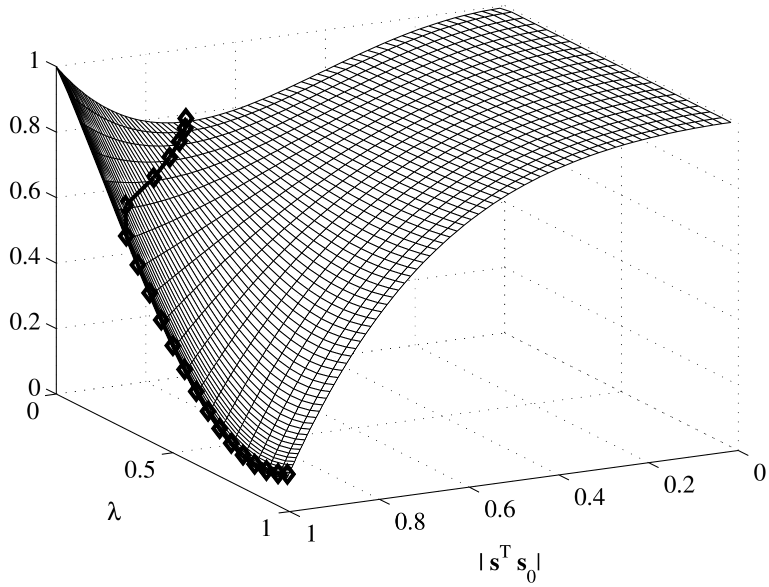


Figure 4.10: Cost surface of the "blind CM" method for ISS in terms of λ, for $n = 500$ and DWR = 21 dB. The markers on the black solid line depict the minimum of the cost surface for each value of λ.

Deviation from the true parameters: "blind AML"

The application of (4.36) requires in practice the knowledge of ν and σ_Z. We assume that the attacker has estimates of the form $\hat{\nu} = \nu + \tilde{\nu}$ and $\hat{\sigma}_Z = \sigma_Z + \tilde{\sigma}_Z$. The cost function is now given by

$$J_{\text{AML}}(\mathbf{s}, \hat{\nu}, \hat{\sigma}_Z) = E\left[\log\cosh\left(\frac{\hat{\nu}\mathbf{Y}^T\mathbf{s}}{\hat{\sigma}_Z^2}\right)\right] - \frac{1}{2\hat{\sigma}_Z^2}E\left[(\mathbf{Y}^T\mathbf{s})^2\right]. \tag{4.39}$$

In practice, the effect of varying σ_Z on the cost function has shown to be negligible. Serious problems may arise when ν is underestimated, but the convexity properties of the cost function seem to remain the same when $\hat{\nu} > \nu$. This suggests that in practical problems we can resort to the estimate of ν given in (4.32) for implementing a blind AML estimator.

4.5.4 Final remarks

Although the analysis carried out here for the ICA, PCA, CM and AML cost functions was made considering Gaussian hosts, the conclusions can be extended to more general host distributions under some mild assumptions. All the cost functions are based on functionals of the type $\mathbf{Y}^T\mathbf{s}$. Hence, if the components of the observations are approximately independent, the resulting random variable can be well approximated

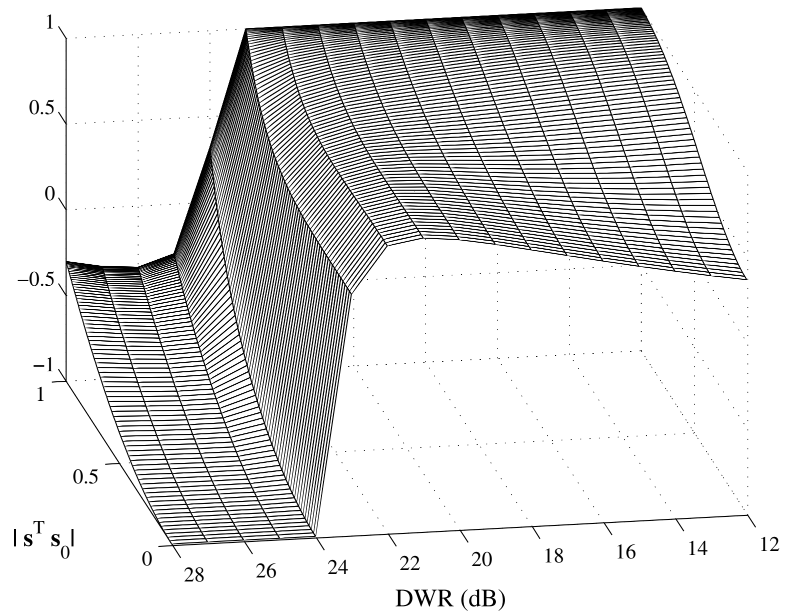


Figure 4.11: Cost surface of the AML method for ISS with $n = 200$ and $\lambda = 0.5$.

for a wide variety of host distributions by a Gaussian whenever n is sufficiently large, by virtue of the CLT. This is true, for instance, when the embedding occurs in the DCT or DWT domains, where the coefficients are distributed according to a Generalized Gaussian. This latter case is explicitly analyzed in Appendix B.7 for the CM cost function, supporting the above arguments.

The final remark comes from the links between the theoretical security analysis and the behavior of the estimators:

1. From Section 3.3.2, it is known that in the WOA scenario large spreading sequences provide more information to the attacker than short ones. Thus, it should be easier for the attacker to perform estimation of $\mathbf{s}_0$ as n is increased. This is the case for the PCA estimator, as discussed in Section 4.3.3, and also for the blind CM estimator, under the light of Figure 4.9.

2. From Section 3.4.2, it is known that increasing λ, i.e. increasing the host rejection, provides more information about $\mathbf{s}_0$. This additional information is effectively translated into an advantage for the blind CM estimator, which is reflected in its cost function (cf. Figure 4.10). However, the PCA estimator works better for $\lambda = 0$, according to (4.21).

4.6 Practical implementation of the WOA estimators

This section is concerned with the application of the previously proposed estimators to practical problems. The first difficulty that arises in practice is that the estimators considered do not admit closed form solutions, so one has to resort to iterative optimization algorithms. Another difficulty introduced by practical setups is the availability of a finite number of samples (observations), so the expectations have to be approximated by some estimate. If the number of observations is small, then the approximate cost function may differ significantly from that derived for asymptotic conditions. In our implementations we make use of the usual sample mean estimator, which simplifies the computation of the gradients needed by the optimization algorithms. Examples for the blind CM method and for the AML method are shown in Example 4.1 and Example 4.2, respectively.

Example 4.1 Cost function of blind CM for a finite number of observations

The cost function for blind CM for a finite set of observations is:

$$J_{\mathrm{CM}}^{blind}(\mathbf{s}) = \frac{1}{N_o}\sum_{i=1}^{N_o}\left((\mathbf{y}_i^T\mathbf{s})^2 - \hat{\nu}^2\right)^2. \tag{4.40}$$

The partial derivatives of (4.40) are

$$\frac{\partial J_{\mathrm{CM}}^{blind}(\mathbf{s})}{\partial s_i} = \frac{4}{N_o}\sum_{k=1}^{N_o}\left((\mathbf{y}_k^T\mathbf{s})^2 - \hat{\nu}^2\right)^2\mathbf{y}_k^T\mathbf{s},\ i = 1,\ldots,n. \tag{4.41}$$

Let us define the vectors $\mathbf{z}_i \in \mathbb{R}^{N_o\times 1}$, $\mathbf{b} \in \mathbb{R}^{N_o\times 1}$ as

$$\mathbf{z}_i \triangleq [y_{1,i},\ldots,y_{N_o,i}]^T,\ i = 1,\ldots,n \tag{4.42}$$

$$\mathbf{b} \triangleq \left[\left((\mathbf{y}_1^T\mathbf{s})^2 - \hat{\nu}^2\right)^2\mathbf{y}_1^T\mathbf{s},\ldots,\left((\mathbf{y}_{N_o}^T\mathbf{s})^2 - \hat{\nu}^2\right)^2\mathbf{y}_{N_o}^T\mathbf{s}\right]^T. \tag{4.43}$$

The gradient of blind CM can be expressed as

$$\nabla J_{\mathrm{CM}}^{blind}(\mathbf{s}) = \frac{4}{N_o}\left[\mathbf{b}^T[\mathbf{z}_i,\ldots,\mathbf{z}_n]\right]^T = \frac{4}{N_o}\mathbf{Y}^T\mathbf{b}, \tag{4.44}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times N_o}$ is the matrix of observations.

In order to optimize the chosen cost function, one can resort to the wide variety of algorithms available in the literature. The most widely used method in Digital Communications is the stochastic gradient descent [129] that allows to construct lightweight iterative methods. In the problem of watermarking security, the attackers are not especially concerned about computational complexity or speed of convergence issues, so

Example 4.2 Cost function of AML for a finite set of observations

The cost function of AML for a finite set of observations is

$$J_{\mathrm{AML}}^{blind}(\mathbf{s}) = \frac{1}{N_o}\sum_{k=1}^{N_o} \log\cosh\left(\frac{\nu}{\sigma_Z^2}\mathbf{y}_k^T\mathbf{s}\right) - \frac{1}{2N_o\sigma_Z^2}\sum_{k=1}^{N_o}(\mathbf{y}_k^T\mathbf{s})^2. \tag{4.45}$$

The partial derivatives of (4.45) are

$$\frac{\partial J_{\mathrm{AML}}^{blind}(\mathbf{s})}{\partial s_i} = \frac{\nu}{N_o\sigma_Z^2}\sum_{k=1}^{N_o} \tanh\left(\frac{\nu}{\sigma_Z^2}\mathbf{y}_k^T\mathbf{s}\right) y_{k,i},\ i = 1, \ldots, n. \tag{4.46}$$

Let us define the vector $\mathbf{c} \in \mathbb{R}^{N_o\times 1}$ as

$$\mathbf{c} = \left[\tanh\left(\frac{\nu}{\sigma_Z^2}\mathbf{y}_1^T\mathbf{s}\right), \ldots, \tanh\left(\frac{\nu}{\sigma_Z^2}\mathbf{y}_{N_o}^T\mathbf{s}\right)\right]^T. \tag{4.47}$$

The gradient of AML can be expressed as

$$\begin{aligned}\nabla J_{\mathrm{AML}}^{blind}(\mathbf{s}) &= \frac{\nu}{N_o\sigma_Z^2}\left[\mathbf{c}^T[\mathbf{z}_1, \ldots, \mathbf{z}_n]\right]^T - \frac{1}{N_o\sigma_Z^2}\left[\mathbf{b}^T[\mathbf{z}_1, \ldots, \mathbf{z}_n]\right]^T \\ &= \frac{\mathbf{Y}}{N_o\sigma_Z^2}\left(\nu\mathbf{c} - \mathbf{b}\right). \end{aligned} \tag{4.48}$$

we choose a different family of optimization methods that, though not very popular in the Communications field, permits us to extend easily the proposed approach to other related watermarking scenarios, as we will see in Section 4.7.3.

Recall that our setup is the following: we are interested in estimating a vector $\mathbf{s}_0$, which is supposed to be of unit norm, i.e. $\mathbf{s}^T\mathbf{s} = 1$. The usual approach is to apply one of the classical, off-the-shelf optimization methods [179] imposing the unit-norm restriction. However, those methods have been designed for performing optimization on Euclidean spaces, i.e. on rectangular coordinates. The correct geometrical setting for this problem is on the "Stiefel manifold" [106]. The usual definition of a manifold is as a Haussdorf topological space that looks locally Euclidean [106, 54]. Specifically, the Stiefel manifold is defined as

Definition 4.1. The Stiefel manifold $\mathcal{S}(m, n)$ is the set of all unitary matrices $\mathbf{U} \in \mathbb{R}^{n\times m}$, and has dimensionality $mn - \frac{1}{2}m(m+1)$.

For $m = 1$, which is the case of interest for us, $\mathcal{S}(m, n)$ becomes the surface of the unit sphere in $\mathbb{R}^n$. The advantage of optimizing directly on the manifold is that one naturally gets rid of the unit-norm constraint in our optimization problem. Thus, our estimate $\hat{\mathbf{s}}_0$ can be expressed as a point in $\mathcal{S}(1,n)$.

Now, recall that in the WOA scenario it is not possible to estimate the sign of the vector $\mathbf{s}_0$. In fact, the vectors $\mathbf{s}_0$ and $-\mathbf{s}_0$ yield the same value for all the cost functions defined in Section 4.5. Hence, the relevant geometrical object for estimation in the WOA scenario is the Grassman manifold.

Definition 4.2. The Grassman manifold $\mathcal{G}(m,n)$ is the set of all m-dimensional hyperplanes through the origin of $\mathbb{R}^n$, and has dimensionality $m(n-m)$. If for $\mathbf{U}_1, \mathbf{U}_2 \in \mathcal{S}(m,n)$ we define the equivalence relation

$$\mathbf{U}_1 \sim \mathbf{U}_2 \text{ if } \operatorname{span}(\mathbf{U}_1) = \operatorname{span}(\mathbf{U}_2),$$

the Grassman manifold results as the quotient space of $\mathcal{S}(m,n)$ under this equivalence relation, and points in the Grassman manifold can be interpreted as m-dimensional subspaces in $\mathbb{R}^n$.

In our case ($m=1$), $\mathcal{G}(m,n)$ is simply the set of one-dimensional spaces that can be spanned by a n-dimensional vector, so our estimate can be expressed as one point in $\mathcal{G}(1,n)$. In general, optimization on the Grassman manifold can be applied to problems where, besides the unitary constraint, the cost function $J(\mathbf{S})$ to be optimized fulfills the condition $J(\mathbf{S}) = J(\mathbf{SH})$, for $\mathbf{S} \in \mathbb{R}^{n\times m}$, and $\mathbf{H} \in \mathbb{R}^{m\times m}$ a unitary matrix.[2]

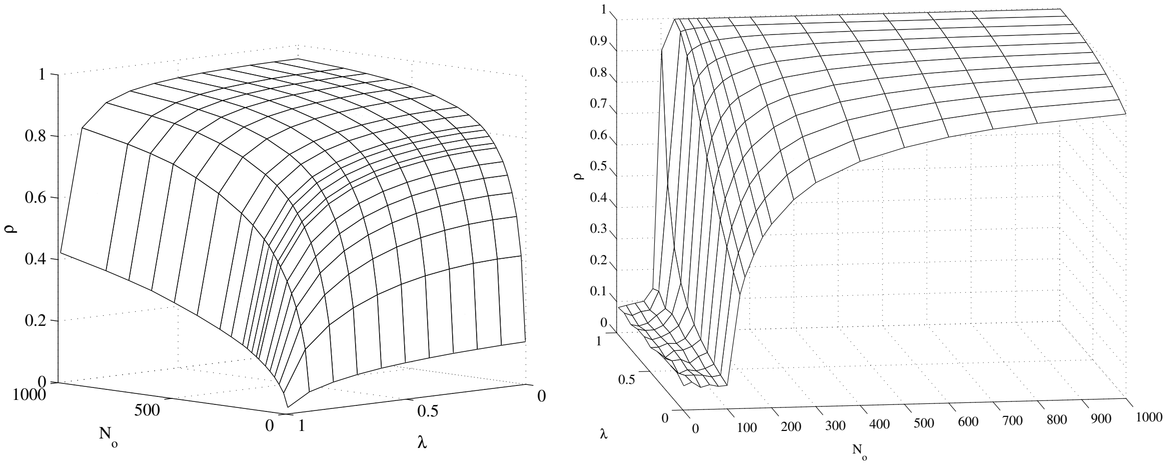
In general, optimization problems with unitary constraints tend to occur in Signal Processing applications involving subspaces. Grassman and Stiefel manifolds have a nice geometrical structure that allows to adapt classical iterative algorithms, such as Newton and conjugate gradient methods, to these curved spaces [106, 162]. For the practical implementations used in this thesis we chose a conjugate gradient method over the Grassman manifold [106]. The algorithm remains essentially equivalent to a classical conjugate gradient, and the main difference lies on how the gradients and updates are computed, since we must move through geodesics instead of straight lines. A complete description of the algorithm can be found in [106, Sect. 3.3].

4.7 Experimental results

4.7.1 Results for the KMA estimator

We compare in terms of performance the estimator proposed in Section 4.4 for ISS with the estimator proposed in [61], originally devised for add-SS (i.e. ISS with

[2] Notice that in the WOA scenario we could still carry out the optimization over the Stiefel manifold, but in this case we would treat $\mathbf{s}$ and $-\mathbf{s}$ as different points.



(a)

(b)

Figure 4.12: Estimation for ISS in the KMA scenario, with DWR = 25 dB, $n = 100$, and Gaussian host. Results with the ML estimator devised in [61, Section V] for add-SS (a), and results with the estimator (4.23) proposed in Section 4.4 (b).

$\lambda = 0$). We remark here that the latter is the ML estimator of $\mathbf{s}_0$ when the host is i.i.d. Gaussian and $\lambda = 0$. Figure 4.12 shows the estimation results for different values of λ and Gaussian host, for both estimators. As can be seen, the add-SS estimator (Figure 4.12(a)) achieves its best performance when $\lambda = 0$. On the contrary, the performance of the ISS estimator (Figure 4.12(b)) is improved as λ increases.[3] Actually, in the limiting case when $\lambda = 1$, n observations would suffice to disclose the secret subspace, since in that case the observations $\mathbf{y}_i \cdot (-1)^{m_i}$ would lie in the same hyperplane, as discussed in Section 4.4.

4.7.2 Results for the WOA estimators

The statistical analysis carried out in Section 4.3.1 shows that blind ICA approaches fail when directly applied to our estimation problem, and this has been shown to be indeed the case in practice. This is why we focus here on the "blind CM" and "informed ICA" (using the estimates (4.32) and (4.24, respectively) estimators. We are also interested in comparing these approaches with the one proposed in [61], which is based on the PCA estimator discussed in 4.3.3.

Figure 4.13(a) shows the comparison between the PCA and blind CM estimators for i.i.d. Gaussian host and different values of λ. As can be seen, the PCA estimator achieves its best performance for $\lambda = 0$, as explained in the previous section, and it

[3] Recall that, according to Theorem 3.2, the closer to 1 is the value of λ, the more information about $\mathbf{s}_0$ leaks.

is below the theoretical upper bound (derived in Section 3.5.2) in all cases. As for the blind CM estimator, it can be seen that it clearly benefits from increasing λ. In general, PCA presents better performance than blind CM for small N_o, but this is not necessarily true as N_o is increased. Figure 4.13(b)-(c)-(d) show results obtained for real images: "man", "couple", and "stream & bridge", available for download in [23]. All of them have been marked in the DCT domain, using a subset of coefficients that is assumed to be known by the attacker, and which is the input to the estimators (thus, depending on the size of the image and the value of n, the maximum allowed N_o is different in each case). This scenario is less favorable for PCA than with the i.i.d. Gaussian host, because the embedding direction is not guaranteed to have the highest variance. As can be seen, the blind CM estimator offers in all cases better performance than the informed ICA. It is also interesting to see that the PCA estimator gets trapped in the directions of maximum variance that, in general, are far from the direction of the spreading vector. Furthermore, it can be observed that for PCA the estimation accuracy can decrease (see Figure 4.13(b),4.13(d)) when the number of observations is increased. This is due to the fact that the host vectors are not i.i.d., and the directions of maximum variance depend on the considered region of the image.

4.7.3 Extension to other scenarios

Estimation of several carriers (multibit data hiding)

We consider here the case where several bits of information are embedded in the same host signal by means of several carriers, a setup that could correspond to scenarios of multiuser information embedding, for instance. The embedding function for a multibit ISS system with embedding rate $R = \frac{1}{n}\log(N_b)$ can be written as

$$\mathbf{Y}_i = \mathbf{X}_i + \nu \sum_{k=1}^{N_b} (-1)^{M_i} \mathbf{s}_k - \lambda \sum_{k=1}^{N_b} \frac{\mathbf{X}_i^T \mathbf{s}_k}{||\mathbf{s}_k||^2} \mathbf{s}_k, \text{ for } i = 1, \ldots, N_o, \tag{4.49}$$

where for simplicity we have considered that the parameters ν and λ are the same for all carriers. Hence, the secrecy of the system relies on the matrix of secret carriers $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{N_b}] \in \mathbb{R}^{n \times N_b}$.

For this kind of multibit setups, the authors of [61] have already proposed an estimator based on PCA and ICA. However, the application of PCA to this scenario presents the drawbacks noted in Sect. 4.3.3 for single carrier schemes. On the contrary, the methods for optimization on manifolds introduced above can be effectively adapted to this scenario. If the carriers are known to be (almost) perfectly orthogonal,[4] then

[4] This will be the case in most scenarios as long as the length (n) of the carriers is large enough, even if they were not generated fulfilling the perfect orthogonality condition.

the matrix of carriers $\mathbf{S}$ is (approximately) unitary. For performing estimation of the carriers in the multibit scenario we resort to a "deflation" approach which, although clearly suboptimal,[5] yields good performance at reasonably low complexity. The deflation approach is described in Algorithm 4.1, and basically consists in estimating the carriers one by one, properly orthogonalizing the observations by means of the Gram-Schmidt procedure.

Algorithm 4.1 Deflation approach for estimating the secret spreading vectors

```
1. Initialize the matrix of observations $\mathbf{O}^{(1)} = [\mathbf{y}_1, \dots, \mathbf{y}_{N_o}]$.
2. For $i = 1, \dots, N_b$
    (a) Compute an estimate $\hat{\mathbf{s}}_i$ by applying the estimator of Section 4.6 on $\mathbf{O}^{(i)}$.
    (b) Orthogonalize the observations by means of the Gram-Schmidt procedure:
            $\mathbf{O}^{(i+1)} = \mathbf{O}^{(i)} - \hat{\mathbf{s}}_i((\mathbf{O}^{(i)})^T \hat{\mathbf{s}}_i).$    (4.50)
3. The final estimate is $\hat{\mathbf{S}} = [\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_{N_b}]$.
```

Regardless of the followed estimation procedure, there is an ambiguity in the order of the estimated carriers which is inherent to the multibit scenario and cannot be undone, as noted in [61]. Figure 4.14(a) shows the estimation results when "man" is marked with 3 different carriers, after removing the ambiguity in the order of the estimated carriers. As can be seen, the blind CM estimator outperforms the PCA-ICA estimator in this scenario.

Circular modulations

Some authors have proposed several variations of the spread spectrum embedding function with the aim of improving its security. In [44], the circular watermarking method was proposed as a slight variation of the multibit ISS. In order to make more difficult the estimation of the carriers, an additional random sequence is used for getting the marked signal sphered.[6] According to [44], the embedding function of the circular

[5] A presumably better approach would be to perform optimization directly in the Stiefel manifold for estimating the whole set of carriers at once, although this path will not be pursued in this thesis. Of course, a suitable cost function must be defined for such problem.

[6] Actually, this is equivalent to use a non-deterministic transformation $\psi(\boldsymbol{\theta})$ every time a signal is marked.

version of ISS reads as

$$\mathbf{Y}_i = \mathbf{X}_i + \nu \sum_{k=1}^{N_b} (-1)^{M_i} \mathbf{s}_k d_k - \lambda \sum_{k=1}^{N_b} \frac{\mathbf{X}_i^T \mathbf{s}_k}{||\mathbf{s}_k||^2} \mathbf{s}_k, \text{ for } i = 1, \ldots, N_o, \tag{4.51}$$

where $\mathbf{d} = [d_1, \ldots, d_{N_b}]$ is a vector uniformly distributed in the unit sphere. Although this randomization effectively conceals the secret carriers, it has the drawback of impairing the communication between embedder and decoder, since the latter ignores the randomization sequence $\mathbf{d}$, which is changed for each marked vector. Furthermore, this modulation does not conceal properly the embedding subspace, so an attacker can take advantage of this fact for removing the watermark with low distortion, if he manages to disclose this subspace. For this purpose one can think of a generalized CM cost function, that in fact can be interpreted as a Constant Norm criterion (CN) [123], for exploiting the structure of the embedding subspace:

$$J_{CNA}(\mathbf{S}, \nu) = E\left[\left(||\mathbf{Y}^T \mathbf{S}||^2 - \nu^2 N_b\right)^2\right]. \tag{4.52}$$

For measuring the performance of the CN estimator we resort to the "chordal distance" [78], which is the natural measure of distance between two subspaces spanned by unitary matrices $\mathbf{P}$ and $\mathbf{Q}$. The chordal distance is usually defined in terms of the "principal angles" $\theta_1, \ldots, \theta_{N_b} \in [0, \pi/2]$ between $\mathbf{P}$ and $\mathbf{Q}$, which in turn are defined as

$$\cos\theta_i = \max_{\mathbf{p}} \max_{\mathbf{q}} \mathbf{p}^T \mathbf{q}, \tag{4.53}$$

where $\mathbf{p}$ and $\mathbf{q}$ are the columns of $\mathbf{P}$ and $\mathbf{Q}$, respectively. This way, the chordal distance can be computed as

$$d_c(\mathbf{P}, \mathbf{Q}) = \left(\sum_{i=1}^{N_b} \sin^2\theta_i\right)^{\frac{1}{2}}. \tag{4.54}$$

A more compact definition of the chordal distance is by means of the "projection matrices" $\mathbf{P}\mathbf{P}^T$ and $\mathbf{Q}\mathbf{Q}^T$ as follows:

$$d_c(\mathbf{P}, \mathbf{Q}) = \frac{1}{\sqrt{2}} ||\mathbf{P}\mathbf{P}^T - \mathbf{Q}\mathbf{Q}^T||_F, \tag{4.55}$$

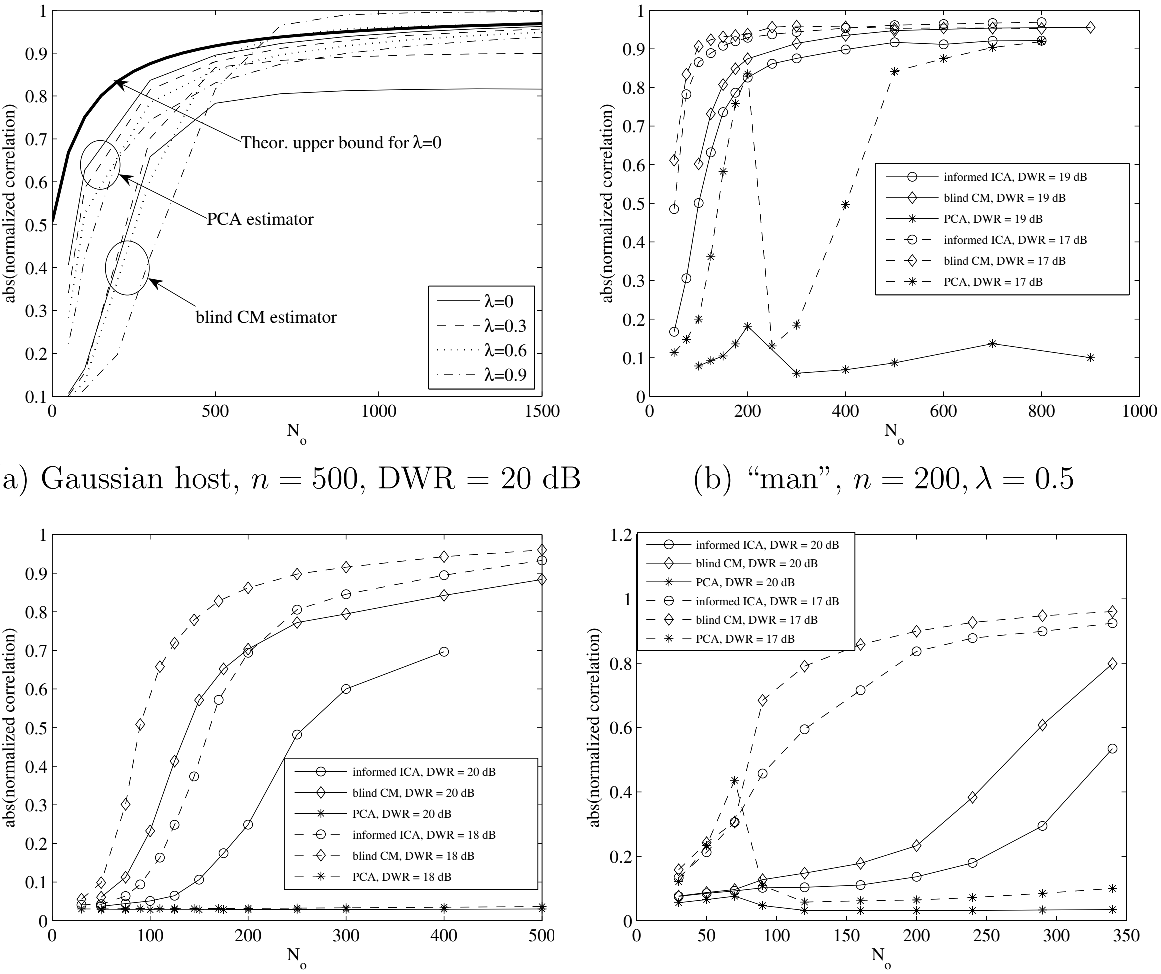
where $||\cdot||_F$ denotes the Frobenius norm for matrices. The chordal distance achieves its maximum, $\sqrt{N_b}$, when the two subspaces are perfectly orthogonal, and it equals 0 when both matrices generate the same subspace. Notice that for optimizing (4.52) we can resort to the Grassman manifold, since the norm is invariant to rotations. Figure 4.14(b) shows the performance comparison between the CN and PCA estimators applied to "man" with $N_b = 2$, $n = 200$ and $\lambda = 0.7$. As we can see, CN performs significantly better than PCA.

4.8 Conclusions

We have analyzed the previously proposed ICA and PCA estimators [61]. Although these estimators have shown their good performance in practical scenarios (see e.g. [61],[44]), we have highlighted their limitations, showing that they can be fooled by appropriately choosing the embedding parameters. As for the new estimators proposed in this chapter (informed ICA and blind CM), the statistical analysis performed in Section 4.5 showed that they also present some drawbacks when facing certain combinations of embedding parameters. However, they have proved to work in a number of scenarios where the previous approaches have failed, as seen in Section 4.7. The combination of new and previous methods provides a wider battery of estimators for performing practical security tests that work for most practical situations.

These results show that it is dangerous to draw conclusions about the security of a particular data hiding scheme using the results obtained with a particular estimator. If this estimator is far from being optimal, then the security level of the method under study could be largely overestimated. In this sense, the role of the information-theoretic security analysis is necessary for providing the fundamental limits of watermarking security.

Finally, we want to note that with the chosen optimization algorithm, computational problems may appear if the size of the spreading vector (n) is very large. In this case, one should look for other optimization methods more suitable for “large scale” problems.



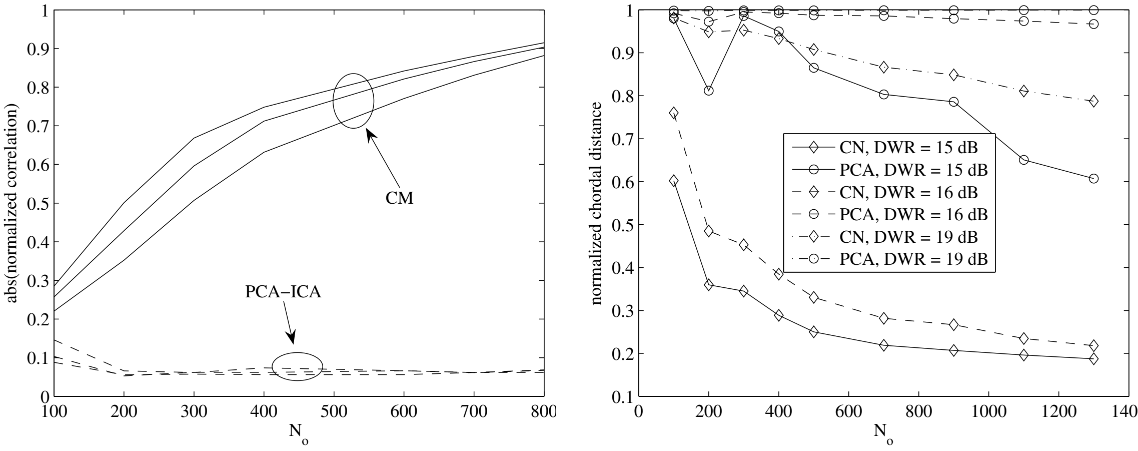
(a) Gaussian host, $n = 500$, DWR = 20 dB

(b) "man", $n = 200, \lambda = 0.5$

(c) "couple", $n = 150, \lambda = 0.6$

(d) "stream & bridge", $n = 250, \lambda = 0.6$

Figure 4.13: Performance comparison of blind CM, informed ICA and PCA for Gaussian hosts and real images. Results averaged over 100 realizations of the spreading vector in each case.



(a) (b)

Figure 4.14: Performance of blind CM and blind CN in multicarrier setups. In (a) we have the performance of "blind CM" for "man" marked with multibit ISS, $N_b = 3$, $n = 200$, $\lambda = 0.8$ and DWR=17dBs. In (b) we have the performance of the "blind CN" and PCA estimators for "man" marked with "circular ISS", $N_b = 2$, $n = 200$, and $\lambda = 0.7$. Results averaged over 100 realizations of the spreading vector.

Chapter 5

Security of Lattice-Based Data Hiding: Theory

The security of spread spectrum methods has been addressed in chapters 3 and 4. In this chapter we study from a theoretical point of view the security of another important group of data hiding schemes: quantization-based methods using structured codebooks, proposed by Chen and Wornell [63] with the name of "Quantization Index Modulation" (QIM). The specific implementation of QIM considered in this thesis is by means of nested lattice codes [174], which encompasses most of the proposed QIM formulations so far, and it will be referred to as "lattice data hiding". In the literature it is also usual to find the nomenclature DC-DM, where DC-DM stands for "Distortion Compensation - Dither Modulation" [63]. Other related family of quantization-based methods is Spread Transform - Dither Modulation (ST-DM) [63],[108], which combines certain characteristics of spread spectrum and QIM methods, although it is not specifically addressed in this thesis.

Lattice data hiding methods have been widely studied during the last years. These methods are the most significant representatives of the so-called "side-information" paradigm for data hiding, which was mainly inspired by Costa's result [83]. Much of the interest in lattice data hiding schemes comes from the fact that they have been shown to provide a practical, manageable construction for approaching the capacity limit predicted by Costa [112]. Despite this interest, however, the security offered by this kind of methods has not been properly addressed so far; the existing results are more focused on practical approaches [61, Section V-D],[45],[46].

This chapter is organized as follows: Section 5.1 introduces the theoretical model for the study of lattice data hiding methods and their security. Section 5.2 studies the security in the KMA and CMA scenarios, characterizing the influence of the embedding parameters. The security in the WOA scenario is addressed in Section 5.3, putting special emphasis in the comparison between KMA and WOA, and in the possibility of achieving perfect secrecy. In Section 5.4, the security of lattice data hiding schemes is linked to that of Costa's set-up [83]. Finally, in Section 5.5 the main conclusions are summarized and some remarks are given.

The main notational conventions followed throughout this chapter are the following: random variables and their occurrences are denoted by capital and lowercase letters, respectively; boldface letters denote column vectors, whereas scalar variables are represented in non-boldface characters. Calligraphic letters are reserved for sets. The volume of a bounded set $\mathcal{X} \in \mathbb{R}^n$ is denoted by $\text{vol}(\mathcal{X})$, whereas the cardinality of a discrete set $\mathcal{C}$ with a countable number of elements is denoted by $|\mathcal{C}|$. The indicator function, denoted by $\psi_{\mathcal{B}}(\cdot)$ and defined as

$$\phi_{\mathcal{B}}(\mathbf{z}) = \begin{cases} 1, & \mathbf{z} \in \mathcal{B} \\ 0, & \text{otherwise}, \end{cases} \tag{5.1}$$

will be widely used throughout the text. The expectation of a function $\varphi(X)$ over X is denoted by $E[\varphi(X)]$. The p-ary alphabet that represents the messages to be transmitted is defined as $\mathcal{M} \triangleq \{0, 1, \ldots, p-1\}$. The "messsage space", defined as $\mathcal{M}^{N_o} \triangleq \mathcal{M} \times \ldots \times \mathcal{M}$, represents the p^{N_o} possible message sequences that can be formed with such an alphabet in N_o channel uses. The sequences in $\mathcal{M}^{N_o}$ can be arranged using any arbitrary ordering (their value in base p, for instance), not relevant for our analysis. The notation $\mathbf{m}^{(k)} = [m_1^k, \ldots, m_{N_o}^k]$ will be used for indexing the kth sequence in $\mathcal{M}^{N_o}$.

5.1 Theoretical model of lattice-based data hiding

This section introduces the mathematical model for lattice data hiding and the lattice constructions used in the paper. As this introduction is not intended to be exhaustive, the interested reader is referred to [82] for a comprehensive description of lattices, and to [174] for a more complete description of lattice codes for data hiding.

5.1.1 Lattices and nested lattice codes

We introduce in this section a series of concepts about lattices and nested lattice codes necessary for the security analysis.

Algebraically, a lattice Λ is defined as a discrete subgroup of $\mathbb{R}^n$ endowed with the natural addition operation. A lattice in n-dimensional space can be generated by any integer combination of a set of n linearly independent basis vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, which form the "generating matrix" of Λ, defined as

$$\mathbf{M} \triangleq [\mathbf{v}_1, \ldots, \mathbf{v}_n]. \tag{5.2}$$

Hence, any point $\mathbf{v} \in \Lambda$ can be written as

$$\mathbf{u} = \mathbf{M}^T\mathbf{z}, \text{ for } \mathbf{z} \in \mathbb{Z}^n. \tag{5.3}$$

Every lattice has an associated nearest neighbor lattice quantizer which maps any vector $\mathbf{x} \in \mathbb{R}^n$ to the nearest lattice point of Λ, and is defined as

$$Q_\Lambda(\mathbf{x}) = \arg\min_{\mathbf{u}\in\Lambda}\{||\mathbf{x} - \mathbf{u}||\}, \tag{5.4}$$

where $||\cdot||$ denotes the Euclidean norm. Ties in (5.4) can be broken arbitrarily but systematically. The fundamental Voronoi region of Λ, a concept that will be frequently recalled in this paper, is defined as

$$\mathcal{V}(\Lambda) \triangleq \{\mathbf{x} \in \mathbb{R}^n : Q_\Lambda(\mathbf{x}) = \mathbf{0}\}, \tag{5.5}$$

and corresponds to the n-dimensional polytope wherein all points are quantized to $\mathbf{0}$. The union of all the lattice points along with the properly translated Voronoi regions tessellates $\mathbb{R}^n$, i.e.

$$\mathbb{R}^n = \Lambda + \mathcal{V}(\Lambda). \tag{5.6}$$

This, together with (5.4), implies that any point $\mathbf{x} \in \mathbb{R}^n$ can be uniquely written as

$$\mathbf{x} = Q_\Lambda(\mathbf{x}) + \mathbf{e}, \tag{5.7}$$

where $\mathbf{e} \in \mathcal{V}(\Lambda)$. The volume of a lattice Λ is defined as the volume of its Voronoi region, and it is computed as

$$\mathrm{vol}(\mathcal{V}(\Lambda)) = \int_{\mathcal{V}(\Lambda)} d\mathbf{x} = \det \mathbf{M}, \tag{5.8}$$

where $\mathbf{M}$ is the generating matrix defined in (5.2). We will also define the modulo-Λ reduction of a vector $\mathbf{x} \in \mathbb{R}^n$ as

$$\mathbf{x} \mod \Lambda \triangleq \mathbf{x} - Q_\Lambda(\mathbf{x}),$$

and the modulo-Λ reduced vector will be denoted by $\tilde{\mathbf{x}}$. Notice that $\tilde{\mathbf{x}} \in \mathcal{V}(\Lambda)$, and it can be regarded as the quantization error resulting from the quantization operation.

Other important parameters of lattices that will be used in this chapter are the following:

- Second order moment per dimension:

$$P(\Lambda) \triangleq \frac{1}{n}\frac{\int_{\mathcal{V}(\Lambda)} ||\mathbf{x}||^2 d\mathbf{x}}{\mathrm{vol}(\mathcal{V}(\Lambda))}, \tag{5.9}$$

which measures the MSE distortion incurred when using Λ as a quantizer.

- Normalized second order moment:

$$G(\Lambda) \triangleq \frac{P(\Lambda)}{\mathrm{vol}(\mathcal{V}(\Lambda))^{\frac{2}{n}}}, \tag{5.10}$$

  which measures the goodness of Λ for quantization in MSE terms.

- The "covering radius" of the lattice Λ:

$$r_c(\Lambda) \triangleq \min\{r : \mathcal{V}(\Lambda) \subseteq \mathcal{B}(\mathbf{0}, r)\}. \tag{5.11}$$

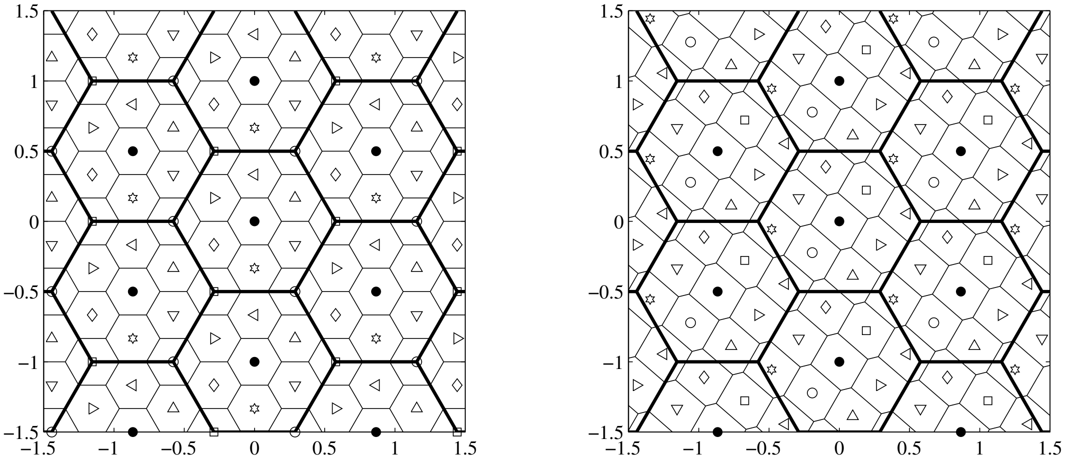
The application of lattices to data hiding is based on the concept of "lattice partitioning." Given a certain lattice Λ, we can define a sublattice Λ' which is a subset of the points in Λ (i.e. $\Lambda' \subset \Lambda$) and is itself a lattice. The set $\{\mathbf{u} + \Lambda'\}$, with $\mathbf{u} \in \Lambda$, is known as a coset of Λ'. Due to the periodic structure of lattices, there exist infinite $\mathbf{u}$ that yield the same coset. The vector $\mathbf{u}$ with the smallest Euclidean norm is termed a "coset leader." From the definitions of lattice and Voronoi region, it follows that the coset leaders always belong to $\mathcal{V}(\Lambda')$. The set of all cosets of Λ' with respect to Λ is called the "partition" of Λ induced by Λ', and it carries a group structure with the natural addition operation. It can be proved that the number of different cosets is given by the so-called "nesting ratio" $\frac{\mathrm{vol}(\mathcal{V}(\Lambda'))}{\mathrm{vol}(\mathcal{V}(\Lambda))} = p$. The union of the p cosets yields the lattice Λ, i.e. $\bigcup_{k=0}^{p-1} \mathbf{d}_k + \Lambda' = \Lambda$, where $\mathbf{d}_k$ denotes the coset leaders.

A nested lattice code is defined by two parameters: a shaping (coarse) lattice Λ and a fine lattice Λ_f such that $\Lambda \subset \Lambda_f$. The pair (Λ, Λ_f) defines a partition which in turn yields a set of p coset leaders or codewords $\mathcal{C}_p = \{\mathbf{d}_k,\ k \in \mathcal{M}\}$. Each letter $k \in \mathcal{M}$ is mapped to one coset leader $\mathbf{d}_k \in \mathcal{C}_p$, and thus to the kth coset of Λ. Although the mapping between elements of $\mathcal{M}$ and $\mathcal{C}$ can be arbitrary, we will assume that the letter 0 corresponds to $\mathbf{d}_0 = \mathbf{0}$. Nested lattice codes can be constructed in a number of ways. Although most results in this chapter are quite general, we focus in some cases on two particular constructions, given their importance: the self-similar lattice construction [116] and the so-called "Construction A" [82],[111].

In the self-similar construction, the nested code is obtained as follows:[1]

1. Define a positive integer $p^{\frac{1}{n}} \in \mathbb{N}$, where n is the dimensionality of Λ.
2. Compute the fine lattice as $\Lambda_f \triangleq p^{-\frac{1}{n}}\Lambda$. It follows that the lattice Λ is a sublattice of Λ_f, resulting in a nesting ratio $\frac{\mathrm{vol}(\mathcal{V}(\Lambda))}{\mathrm{vol}(\mathcal{V}(\Lambda_f))} = p$, and an embedding rate $R = \log(p)/n$.

[1] More general self-similar partitions consider also rotations of Λ [79], but we will restrict our attention to those obtained through scaling.



(a)

(b)

Figure 5.1: Nested lattice codes of rate $R = \log(9)/2$ with hexagonal shaping lattice, obtained by means of self-similar construction (a) and with Construction A (b). The Voronoi regions of Λ_f and Λ are represented by thin and thick lines, respectively.

3. Obtain the set of coset leaders $\mathcal{C}_p$ as $\Lambda_f \cap \mathcal{V}(\Lambda)$.

In Construction A, the nested lattice code is completely specified by a "generating vector" and the lattice Λ. It is summarized as follows:

1. Define a positive integer p and a generating vector $\mathbf{g} \in \mathbb{Z}_p^n$, where $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$. Compute the codebook $\mathcal{Q} \triangleq \{\mathbf{c} \in \mathbb{Z}_p^n : \mathbf{c} = k \cdot \mathbf{g} \mod p,\ k \in \mathcal{M}\}$, which is contained in the hypercube $[0, p)^n$. Then, construct the lattice $\Lambda' = p^{-1}\mathcal{Q} + \mathbb{Z}^n$.

2. Define the generating matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ (where each column is a basis vector) of the coarse (shaping) lattice Λ. Apply the linear transformation $\Lambda_f = \mathbf{G}\Lambda'$. It immediately follows that Λ is a sublattice of Λ_f and the nesting ratio is $\frac{\mathrm{vol}(\mathcal{V}(\Lambda))}{\mathrm{vol}(\mathcal{V}(\Lambda_f))} = p$, resulting in a coding rate $R = \log(p)/n$.

3. The set of coset leaders $\mathcal{C}_p$ is given by $\Lambda_f \cap \mathcal{V}(\Lambda)$, or equivalently, $p^{-1}\mathbf{G}\mathcal{Q} \mod \Lambda$. Notice that the mapping between the letters of the alphabet and the coset leaders follows directly from the construction procedure.

The advantage of Construction A over the self-similar construction is that it allows to build codes of arbitrary rate (without the restriction $p^{\frac{1}{n}} \in \mathbb{N}$). Moreover, if p is chosen as a prime number and Λ is a good lattice for MSE quantization, then good (asymptotic) properties of the code are ensured [111].

Examples of 2-dimensional nested lattice codes are shown in Figure 5.1, using a hexagonal shaping lattice and $p = 9$. For Construction A, the generating vector $\mathbf{g} = [1, 2]^T$ was chosen. In both cases, the lattice points belonging to the same coset are represented by the same symbol. Notice that (as in Figure 5.1(a)) a coset leader may fall exactly in the frontier between several quantization cells.

5.1.2 Embedding and decoding

The mathematical model for lattice data hiding is shown in Figure 5.2. First, the host signal is partitioned into non-overlapping blocks $\mathbf{X}_k$ of length n. The message to be embedded may undergo channel coding, yielding the symbols $M_k \in \mathcal{M}$. In our setup, the messages embedded in different blocks are assumed to be equiprobable in $\mathcal{M}$ and mutually independent, unless otherwise stated. $\mathbf{\Theta}$ represents the secret key, shared between embedder and decoder. The parameter $\mathbf{T} = \psi(\mathbf{\Theta})$ is a n-dimensional vector, termed "secret dither," which is used to randomize the embedding and decoding functions. This vector plays the role of secret key. In the lattice DC-DM scheme, each letter M_k is embedded in one block $\mathbf{X}_k$ by means of a randomized lattice quantizer as shown in Figure 5.2. First, using $\mathbf{T}$, M_k and Λ (the shaping lattice), the coset $\mathcal{U}_{M_k,\mathbf{T}} = \Lambda + \mathbf{d}_{M_k} + \mathbf{T}$ is obtained, where $\mathbf{d}_{M_k}$ is the coset leader associated to M_k. Thereafter, the block $\mathbf{X}_k$ is quantized to the nearest point in $\mathcal{U}_{M_k,\mathbf{T}}$ and the resulting quantization error is computed. Finally, this quantization error is scaled by the "distortion compensation parameter" $\alpha \in [0, 1]$, and added back to $\mathbf{X}_k$ in order to obtain the marked block $\mathbf{Y}_k$. Mathematically, this is expressed as

$$\mathbf{Y}_k = \mathbf{X}_k + \alpha(Q_{\mathcal{U}_{M_k,\mathbf{T}}}(\mathbf{X}_k) - \mathbf{X}_k), \tag{5.12}$$

where $Q_{\mathcal{U}_{M_k,\mathbf{T}}}(\mathbf{x})$ is a nearest-neighbor quantizer whose centroids are distributed according to $\mathcal{U}_{M_k,\mathbf{T}}$. Taking into account that $Q_{\Lambda+\mathbf{p}}(\mathbf{x}) = Q_\Lambda(\mathbf{x} - \mathbf{p}) + \mathbf{p}$, the embedding function (5.12) is usually implemented in practice by means of a "dithered" lattice quantizer as follows:

$$\mathbf{Y}_k = \mathbf{X}_k + \alpha(Q_\Lambda(\mathbf{X}_k - \mathbf{d}_{M_k} - \mathbf{T}) - \mathbf{X}_k + \mathbf{d}_{M_k} + \mathbf{T}). \tag{5.13}$$

Notice that Eq. (5.13) is equivalent to (5.12) and to the embedder depicted in Figure 5.2. The distortion caused by the embedding process can be computed by resorting to the usual assumption (a.k.a. "flat-host assumption") that the variance of the components of $\mathbf{X}_k$ is sufficiently large, in such a way that the host distribution is uniform inside each quantization cell. Thus, from (5.13), the embedding distortion per dimension in a mean-squared-error sense results in

$$D_w = \frac{1}{n} E\left[||\mathbf{X}_k - \mathbf{Y}_k||^2\right] = \alpha^2 P(\Lambda), \tag{5.14}$$

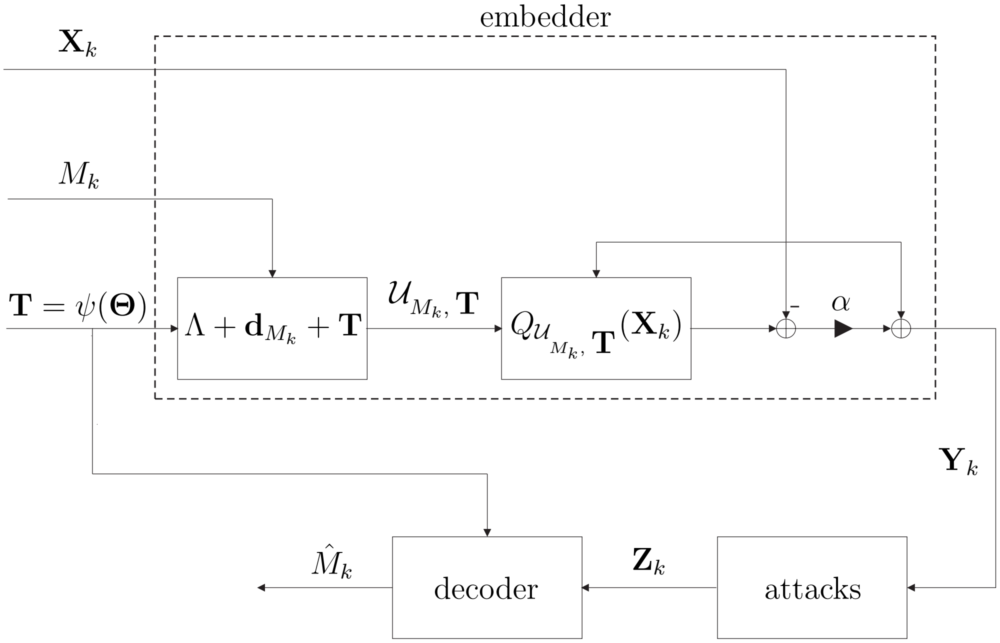

embedder
$\mathbf{X}_k$
M_k
$\mathbf{T} = \psi(\mathbf{\Theta})$
$\Lambda + \mathbf{d}_{M_k} + \mathbf{T}$
$\mathcal{U}_{M_k}, \mathbf{T}$
$Q_{\mathcal{U}_{M_k}, \mathbf{T}}(\mathbf{X}_k)$
-
α
$\mathbf{Y}_k$
$\hat{M}_k$
decoder
$\mathbf{Z}_k$
attacks

Figure 5.2: Block diagram showing the lattice data hiding model. The parameter $\mathbf{T}$ is the secret dither.

where $P(\Lambda)$ denotes the second-order moment per dimension of the Voronoi region of Λ, defined in Eq. (5.9).[2]

The most popular decoders are those termed "lattice decoders", where the embedded message is estimated by choosing the coset which is closest to the received (attacked) vector $\mathbf{Z}_k = \vartheta(\mathbf{Y}_k)$, where $\vartheta(\cdot)$ represents the transformation applied by the attacker to $\mathbf{Y}_k$. Hence, the lattice decoder can be mathematically formulated as

$$\hat{M}_k = \min_{m \in \mathcal{M}} \left\{ ||Q_\Lambda(\mathbf{Z}_k - \mathbf{d}_m - \mathbf{T}) - \mathbf{Z}_k + \mathbf{d}_m + \mathbf{T}|| \right\}, \tag{5.15}$$

where $|| \cdot ||$ denotes the Euclidean norm. Bear in mind that the decoder needs the correct realization of $\mathbf{T}$ for successful performance. Traditionally, the attacker designs the function $\vartheta(\cdot)$ without taking into account the secret dither $\mathbf{T}$. In this thesis, however, the attacker focuses his strategy on $\mathbf{T}$, as explained below.

5.1.3 Problem formulation

It is assumed that the attacker manages to gather an ensemble of marked blocks $\{\mathbf{Y}_k,\ k = 1, \ldots, N_o\}$, which may belong to different host signals, but all of them were

[2] Thanks to the flat-host assumption, the analysis of lattice data hiding methods is independent of the actual host statistics.

marked with the same secret key, and hence with the same secret dither vector $\mathbf{T}$. Under the assumptions of Chapter 2, the attacker knows the embedding parameters being used, i.e. Λ, $\mathcal{C}_p$, and α, whereas he ignores the host blocks $\mathbf{X}_k$, the embedded symbols M_k, and $\mathbf{T}$. The objective of the attacker is to estimate this secret dither $\mathbf{T}$. The first step performed by the attacker is the modulo reduction of the marked blocks as $\tilde{\mathbf{Y}}_k \triangleq \mathbf{Y}_k \mod \Lambda$. Under the flat-host assumption introduced above (Sect. 5.1.2), such modulo reduction does not imply any loss of information for the attacker, as discussed in [188]. Note that (5.13) can be rewritten as

$$\mathbf{Y}_k = \mathbf{d}_{M_k} + \mathbf{T} + \mathbf{Q}_\Lambda(\mathbf{U}_k) + (1-\alpha)\mathbf{N}_k, \tag{5.16}$$

where $\mathbf{U}_k \triangleq \mathbf{X}_k - \mathbf{d}_{M_k} - \mathbf{T}$ and $\mathbf{N}_k \triangleq \mathbf{U}_k - Q_\Lambda(\mathbf{U}_k)$. Thus, the modulo-$\Lambda$ reduced observations seen by the attacker are given by

$$\tilde{\mathbf{Y}}_k = (\mathbf{d}_{M_k} + \mathbf{T} + (1-\alpha)\mathbf{N}_k) \mod \Lambda. \tag{5.17}$$

From (5.17), it becomes clear that the secret dither $\mathbf{T}$ is concealed by the transmitted message M_k and by the host-interference $\mathbf{N}_k$. The parameter α controls the amount of host-interference or "self-noise", thus affecting the robustness of the lattice data hiding scheme. However, from a security standpoint, one can take benefit of the randomness introduced by the self-noise for achieving good secrecy and complicating the attacker's task, as will be seen in Section 5.3.1. As for the secret dither $\mathbf{T}$, its statistical distribution will be fixed later according to the result of Lemma 5.2, in order to maximize the equivocation.

The statistical distribution of the observations seen by the attacker can be easily identified by recalling the flat-host assumption, which makes $\mathbf{N}_k$ uniformly distributed in $\mathcal{V}(\Lambda)$. Hence, it follows from (5.17) that

$$\begin{aligned}\varphi(\mathbf{x}) \triangleq f(\tilde{\mathbf{y}}_k|m_k=0,\mathbf{t}=\mathbf{0}) &= (\mathrm{vol}(\mathcal{Z}(\Lambda)))^{-1}\cdot\phi_{\mathcal{Z}(\Lambda)}(\mathbf{x})\\ &= \begin{cases}(\mathrm{vol}(\mathcal{Z}(\Lambda)))^{-1}, & \mathbf{x}\in\mathcal{Z}(\Lambda)\\ 0, & \text{otherwise,}\end{cases}\end{aligned} \tag{5.18}$$

with $\mathcal{Z}(\Lambda) \triangleq (1-\alpha)\mathcal{V}(\Lambda)$. Using again the flat-host assumption,

$$f(\tilde{\mathbf{y}}_k|m_k,\mathbf{t}) = \varphi(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k} - \mathbf{t} \mod \Lambda). \tag{5.19}$$

Hence, $f(\tilde{\mathbf{y}}_k|\mathbf{t}) = \frac{1}{p}\sum_{m_k=0}^{p-1} f(\tilde{\mathbf{y}}_k|m_k,\mathbf{t})$, and $f(\tilde{\mathbf{y}}_k) = \int f(\tilde{\mathbf{y}}_k|\mathbf{t})f(\mathbf{t})d\mathbf{t}$.

5.2 Theoretical analysis of the KMA

When a sequence of marked signals $\{\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o}\}$ and their associated messages $\{M_1,\ldots,M_{N_o}\}$ are observed, the information leakage about $\mathbf{T}$ is calculated as

$$I(\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o};\mathbf{T}|M_1,\ldots,M_{N_o}) = h(\mathbf{T}) - h(\mathbf{T}|\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o},M_1,\ldots,M_{N_o}), \tag{5.20}$$

where we have made use of the mutual independence between $\mathbf{T}$ and the embedded messages, also assumed to be mutually independent. Recall that the second term in the right hand side of (5.20) represents the equivocation about $\mathbf{T}$, which measures the remaining ignorance about the secret dither. Thus, the appropriate distribution for $\mathbf{T}$ should be chosen in order to maximize the equivocation. To this end, let us first introduce the next definition:

Definition 5.1. The "feasible region" of the secret dither is defined as the support of its conditional pdf after N_o observations.

The next property will be widely used throughout the text.

Property 5.1. The feasible region is bounded by

$$\mathcal{S}_{N_o} \triangleq \bigcap_{i=1}^{N_o} \mathcal{D}_i,$$

where

$$\mathcal{D}_i \triangleq (\tilde{\mathbf{y}}_i - \mathbf{d}_{m_i} - \mathcal{Z}(\Lambda)) \mod \Lambda,\ i = 1, \ldots, N_o. \tag{5.21}$$

Proof: Application of Bayes' rule yields

$$\begin{aligned} f(\mathbf{t}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o}) &= \frac{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o}|\mathbf{t}) \cdot f(\mathbf{t})}{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o})} \\ &= \frac{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}|m_1, \ldots, m_{N_o}, \mathbf{t}) \cdot f(\mathbf{t})}{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}|m_1, \ldots, m_{N_o})}, \end{aligned} \tag{5.22}$$

where $\tilde{\mathbf{y}}_i \in (\mathbf{d}_m + \mathbf{t} + \mathcal{Z}(\Lambda)) \mod \Lambda,\ i = 1, \ldots, N_o$. Notice that each random variable $\tilde{\mathbf{Y}}_i$ is a function of the triple $(\mathbf{X}_i, M_i, \mathbf{T})$, and the host samples $\mathbf{X}_i$ in our model are mutually independent. This means that the observations $\{\tilde{\mathbf{Y}}_i\}$ are conditionally independent given the dither; hence, Eq. (5.22) can be rewritten as

$$f(\mathbf{t}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o})$$

$$= \frac{f(\mathbf{t}) \cdot \prod_{i=1}^{N_o} f(\tilde{\mathbf{y}}_i|m_i, \mathbf{t})}{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}|m_1, \ldots, m_{N_o})} \tag{5.23}$$

$$= \frac{f(\mathbf{t}) \cdot \prod_{i=1}^{N_o} f((\tilde{\mathbf{y}}_i - \mathbf{d}_{m_i} - \mathbf{t}) \mod \Lambda | M_i = 0, \mathbf{T} = \mathbf{0})}{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}|m_1, \ldots, m_{N_o})}, \tag{5.24}$$

where (5.24) follows from the flat-host assumption. By recalling Eq. (5.19), it is clear that each term in the numerator of (5.24) is nonzero iff $(\tilde{\mathbf{y}}_i - \mathbf{d}_{m_i} - \mathbf{t}) \mod \Lambda \in \mathcal{Z}(\Lambda)$, or equivalently, iff $\mathbf{t} \in \mathcal{D}_i$, with $\mathcal{D}_i$ given by (5.21). Hence, it is clear that the feasible region of $\mathbf{t}$ is contained in $\bigcap_{i=1}^{N_o} \mathcal{D}_i$, independently of the distribution of $\mathbf{T}$. ■

Consider now the following definition.

Definition 5.2. A set $\mathcal{S}$ is said to be modulo-Λ convex if there exists $\mathbf{r}$ such that $\mathcal{S} - Q_{\Lambda+\mathbf{r}}(\mathcal{S})$ is convex.

The notion of modulo-Λ convexity is key to our analysis, due to the next property.

Property 5.2. For $\alpha \geq 0.5$, the feasible region $\mathcal{S}_{N_o}$ is always a modulo-Λ convex set.

Proof: Let us define

$$\tilde{\mathbf{V}}_i \triangleq (\tilde{\mathbf{Y}}_i - \mathbf{d}_{M_i} - \mathbf{T}) \mod \Lambda \tag{5.25}$$

and $\mathcal{D}'_i \triangleq \tilde{\mathbf{V}}_i - \mathcal{Z}(\Lambda)$, for $i = 1, \ldots, N_o$. By recalling (5.17), it is clear that $\tilde{\mathbf{V}}_i \sim U(\mathcal{Z}(\Lambda))$. If $\alpha \geq 0.5$, then $\tilde{\mathbf{V}}_i + \mathbf{r} \in \mathcal{V}(\Lambda)$, $\forall\ \mathbf{r} \in \mathcal{Z}(\Lambda)$. Hence, $\mathcal{D}'_i \subset \mathcal{V}(\Lambda)$, and obviously $\bigcap_i \mathcal{D}'_i \subset \mathcal{V}(\Lambda)$. Since $\mathcal{D}'_i$, $i = 1, \ldots, N_o$, are convex sets, their intersection is also a convex set. Taking into account that $\mathcal{D}_i = (\mathbf{T} + \mathcal{D}'_i) \mod \Lambda$, the property follows. ■

The use of $\alpha < 0.5$ may lead to non-convex feasible regions, as illustrated in Figure 5.3(b), where the feasible region for the dither is composed of three different modulo-Λ sets. However, as can be seen in the proof of Property 5.2, under the assumption of $\alpha \geq 0.5$ it is possible to find a shifted version of the problem such that the feasible region is always modulo-Λ convex, according to Definition 5.2. This property permits us to drop out the modulo operation from the expressions of the feasible regions. Bear in mind that the entropy is invariant to translations, so this simplification does not change the results. In the remainder of this chapter we will work under the assumption of $\alpha \geq 0.5$.

The next result shows that perfect estimation of the secret dither in the KMA scenario is always possible.

Lemma 5.1 (Asymptotic convergence of the feasible region). If the secret dither takes the value $\mathbf{t}$, then $\mathcal{S}_{N_o}$ converges almost surely to $\mathbf{t}$ as $N_o \to \infty$.

Proof: See Appendix C.1. ■

Although estimation of the secret dither is possible, different choices of the statistical distribution of $\mathbf{T}$ will yield very different security levels. The maximization of the security level is the subject of the next result.

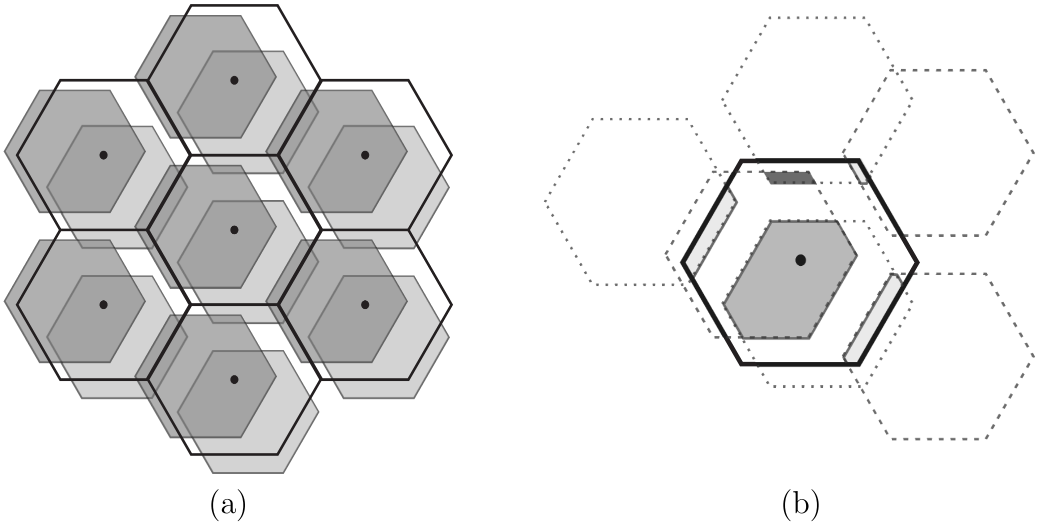


Figure 5.3: Illustration of a non-connected feasible region for two observations using small α. In (a), the solid lines are the Voronoi regions of Λ, and the feasible regions for the centroids defined by each observation are the shaded ones. The figure depicted in (b) is the modulo-Λ reduction of the intersection between the shaded regions in (a), showing three resulting modulo-Λ convex regions (illustrated with different shadings).

Lemma 5.2 (Maximization of the equivocation). The equivocation for any $N_o \geq 1$ is maximized for $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, yielding a conditional pdf uniformly distributed in $\mathcal{S}_{N_o}$, that is

$$f(\mathbf{t}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o}) = \begin{cases} (\mathrm{vol}\,(\mathcal{S}_{N_o}))^{-1}, & \mathbf{t} \in \mathcal{S}_{N_o} \\ 0 & \text{otherwise.} \end{cases} \tag{5.26}$$

Proof: By the definition of residual entropy, we have

$$h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}, M_1, \ldots, M_{N_o}) = E[h(\mathbf{T}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o})], \tag{5.27}$$

where the expectation is taken over the joint pdf $f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o})$. Since the feasible region of the dither is bounded by $\mathcal{S}_{N_o}$, its entropy will be maximized when the dither is uniformly distributed in $\mathcal{S}_{N_o}$, i.e,

$$\begin{aligned} h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}, M_1, \ldots, M_{N_o}) &= -E[\log(\mathbf{T}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o})] \\ &\leq E[\log(\mathrm{vol}(\mathcal{S}_{N_o}))]. \end{aligned} \tag{5.28}$$

Since the denominator of (5.24) does not depend on $\mathbf{t}$, then the choice $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$ suffices for achieving such distribution, and hence equality in (5.28). ∎

The optimal distribution resulting from Lemma 5.2 also brings additional desirable properties: it provides statistical independence between the self-noise and the host signal [204], and most important, it does not prevent from achieving capacity in the Gaussian channel in the asymptotic set-up ($n \to \infty$) [112]. Hence, the choice of $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$ is "good" from the robustness and security points of view, and this will be the chosen distribution in the remaining of this chapter unless otherwise stated. Hence, by combining Property 5.1 and Lemma 5.2, the residual entropy results in

$$h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}, M_1, \ldots, M_{N_o}) = E[\log(\text{vol}(\mathcal{S}_{N_o}))], \tag{5.29}$$

where the expectation is taken over the joint pdf of the observations. In case of one observation ($N_o = 1$) we have

$$h(\mathbf{T}|\tilde{\mathbf{Y}}_1, M_1) = \log(\text{vol}(\mathcal{Z}(\Lambda))) = \log((1-\alpha)^n \text{vol}(\mathcal{V}(\Lambda))), \tag{5.30}$$

and the information leakage is given by

$$I(\tilde{\mathbf{Y}}_1; \mathbf{T}|M_1) = h(\mathbf{T}) - h(\mathbf{T}|\tilde{\mathbf{Y}}_1, M_1) = -n\log(1-\alpha) \tag{5.31}$$

for all $\alpha \in [0,1]$, independently of the specific lattice chosen for embedding. This result clearly shows a trade-off between security and achievable rate: theoretical analyses [108], [112] show that, in AWGN channels, the value of α must approach 1 for maximizing the achievable rate in the high-SNR region. However, from (5.30) we have that $\lim_{\alpha\to 1} h(\mathbf{T}|\tilde{\mathbf{Y}}_1, M_1) = -\infty$, meaning that one observation suffices to get a perfect estimate of the secret dither. The intuitive interpretation is easy: for $\alpha \approx 1$, one observation is enough for pinpointing the exact location of one centroid of the lattice; given the structure imposed to the codebook, this observation provides all the information about $\mathbf{T}$.

If $N_o > 1$, the security level of the lattice data hiding scheme is very much dependent of the chosen lattice when $\alpha < 1$, as will be seen in the subsequent sections. Now we will try to shed some light on two fundamental questions:

1. Given n, what is the best lattice (if any) in terms of security?
2. Does an increase of n improve the security level?

In order to provide a fair comparison between different lattices, they are scaled so as to present the same embedding distortion. For computing the residual entropy, the expectation in (5.29) must be taken over $f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m_1, \ldots, m_{N_o})$, but the conditional pdf of $\mathbf{T}$, given by (5.26), does not depend on the specific sequence of messages embedded, as long as the latter is known; this implies that, for the expectations, the message sequence can be assumed to be deterministic. Since it is not always possible to obtain closed-form expressions for the information leakage (even for low-dimensional lattices), we must resort in general to Monte Carlo integration and bounding techniques.

5.2.1 Equivocation for the cubic lattice

For the scaled cubic lattice[3] $\Delta\mathbb{Z}^n = (x_1, \dots, x_n),\ x_i \in \Delta\mathbb{Z}$ it is possible to obtain a closed-form expression for the residual entropy. From Eq. (5.29), the residual entropy is given by the expectation of the log-volume of the feasible region for the dither. Since the latter for the cubic lattice is always a hyperrectangle, using Property 5.2 we can write

$$E[\log(\text{vol}(\mathcal{S}_{N_o}))] = \sum_{k=1}^{n} E[\log(W_k)] = n \cdot E[\log(W)], \tag{5.32}$$

where W_k is the random variable that measures the length of the feasible interval in the k-th dimension, and the last equality follows because the quantization step is the same for all dimensions. The random variable W is given by

$$W = \text{vol}(\bigcap_{i=1}^{N_o}(\tilde{V}_i - \mathcal{I})), \tag{5.33}$$

with $\tilde{V}_i$ a random variable uniformly distributed in

$$\mathcal{I} \triangleq [-(1-\alpha)\Delta/2, (1-\alpha)\Delta/2).$$

Hence, the problem is reduced to a scalar subproblem consisting in computing $E[\log(W)]$, i.e., the residual entropy in one dimension. The final expression of the residual entropy per dimension is expressed in the following theorem, which is proven in Appendix C.2.

Theorem 5.1 (Equivocation for the cubic lattices). The equivocation per dimension for the cubic lattice is given by

$$\begin{aligned}
&\tfrac{1}{n}h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}, M_1, \dots, M_{N_o}) \\
&\qquad = \log((1-\alpha)\Delta) - H_{N_o} + 1 \\
&\qquad = \log(\sqrt{12}D_w) - H_{N_o} + 1 + \log\left(\frac{1-\alpha}{\alpha}\right),\ \text{for } \alpha \geq 0.5,
\end{aligned} \tag{5.34}$$

where

$$H_{N_o} \triangleq \sum_{i=1}^{N_o} \frac{1}{i}$$

is the N_oth harmonic number [17], D_w is the embedding distortion according to (5.14), and we have taken into account that $D_w = \alpha P(\Lambda) = \alpha\Delta^2/12$.

[3] We consider the same quantization step in each dimension, although the results can be straightforwardly extended to a more general case.

5.2.2 Numerical computation of the equivocation

When the analytical evaluation of (5.29) becomes intractable we resort to Monte Carlo integration. The fact that the feasible region is reduced with each new observation makes necessary an additional task of computing a tight region of integration so as to preserve the accuracy of the Monte Carlo method (as will be seen in step 3 of the algorithm outlined below). In order to give a comparison between different standard lattices, we consider the root lattices and their duals (the best known lattice quantizers for $n \leq 8$), namely A_2 (hexagonal lattice), D_3, $D_4 \cong D_4^*$, D_5, E_7, $E_8 \cong E_8^*$. For their definition and properties, see [82], [81]. All these lattices are scaled so as to present the same embedding distortion per dimension as the cubic lattice $\Delta\mathbb{Z}^n$ with $\Delta = 1$, that is, $D_w = 1/12$. The scaling factor Δ to be applied to any lattice can be obtained as

$$\Delta = \left(\frac{G(\Lambda)^{-1} \cdot \mathrm{vol}(\mathcal{V}(\Lambda))^{-\frac{2}{n}}}{12} \right)^{\frac{1}{2}}, \tag{5.35}$$

where $G(\Lambda)$ denotes the normalized second order moment per dimension of Λ. The procedure followed for the Monte Carlo simulations is briefly outlined in Algorithm 5.1.

After performing the corresponding numerical optimizations, the results of Monte Carlo integration indicate that the lattice Λ_n^* that maximizes the residual entropy for each n is that with the best mean-squared quantization properties. This can be formally expressed as

$$\begin{aligned} \Lambda_n^* = \quad & \arg \min_{\Lambda \in \mathcal{L}_n} G(\Lambda) \\ & \text{subject to } P(\Lambda) = \text{constant} \end{aligned} \tag{5.36}$$

where $\mathcal{L}_n$ is the set of root lattices of dimensionality $n \leq 8$. Notice that Λ_n^* maximizes $\mathrm{vol}(\mathcal{V}(\Lambda))$ for given n and $P(\Lambda)$, and consequently Λ_n^* has the highest a priori entropy in $\mathcal{L}_n$, due to the uniformity of $\mathbf{T}$. For illustration purposes, Figure 5.4 gives a comparison between the residual entropy per dimension using the cubic lattice and that using some of the root lattices. Although we do not claim that the above result holds for the whole set of lattices with arbitrary n, at least it suggests that the security level of a lattice data hiding scheme can be improved by increasing n and choosing the lattice Λ with the lowest $G(\Lambda)$. This leads us to conjecture that a hypothetical spherically-shaped Voronoi region will provide an upper bound to the residual entropy, since the sphere is the region of $\mathbb{R}^n$ with the smallest normalized second order moment. This is indeed so for the set of lattices considered in our experiments: as an example, the result obtained with the 8-dimensional sphere (also obtained through Monte Carlo) is plotted in Figure 5.4. Unfortunately, the space can not be tessellated with spherical regions (except for $n = 1$), so it is not possible to construct "spherical" lattice quantizers; nevertheless, as it was shown in [222], as n increases there exist lattices whose normalized second order

Algorithm 5.1 Numerical computation of the equivocation for a generic lattice

1. We assume without loss of generality that $\mathbf{t} = \mathbf{0}$. Hence, a sequence of N_o observed vectors uniformly distributed in $(1-\alpha)\Delta\mathcal{V}(\Lambda)$, with Δ such that $P(\Lambda) = 1/12$, is generated.
2. $\mathcal{V}(\Lambda)$ is outer bounded by a hypercube whose edge length is twice the covering radius [82] of Λ. This gives an outer bound to $\mathcal{D}_i$ (Eq. (5.21)), which is used to compute an outer approximation $\mathcal{S}_{N_o}^u$ of the feasible region.
3. The feasible region resulting from the previous step (which is a hyperrectangle) is shrunk along each dimension so as to tightly bound the true feasible region $\mathcal{S}_{N_o}$. This is accomplished by means of a bisection algorithm which looks for the tightest limits of the outer bounding hyperrectangle in each dimension. The need for this step is justified by the fact that, for large N_o, the ratio $\mathrm{vol}(\mathcal{S}_{N_o}^u)/\mathrm{vol}(\mathcal{S}_{N_o})$ becomes too large, affecting the accuracy of Monte Carlo integration.
4. A large number of points uniformly distributed in the hyperrectangle of the previous step is generated. For each of these points, it is checked whether it belongs to $\bigcap_{i=1}^{N_o} \mathcal{D}_i$; if so, the considered point belongs to $\mathcal{S}_{N_o}$. Finally, the log-volume of $\mathcal{S}_{N_o}$ is computed by Monte Carlo integration, and the residual entropy is obtained by averaging the log-volume over a large number of realizations. In steps 3) and 4), fast quantizing algorithms [80] are used.

moment tend to that of a sphere.[4] The security of lattice DC-DM using this type of lattices is studied in the next section.

5.2.3 Bounds and asymptotics on the equivocation for "good" lattices

Throughout this section, we will make use of two assumptions:

1. $\alpha \geq 0.5$;
2. we are using Λ_n^*, the optimal (in an MSE sense) n-dimensional lattice quantizer.

As discussed in the proof of Property 5.2, Assumption 1 makes the modulo operation transparent for the computation of the entropy, since this is invariant to translations. Making use of the chain rule for mutual informations [84] we can write

[4] Moreover, this is a necessary condition for the lattices in order to achieve the channel capacity in the lattice DC-DM scheme [112].

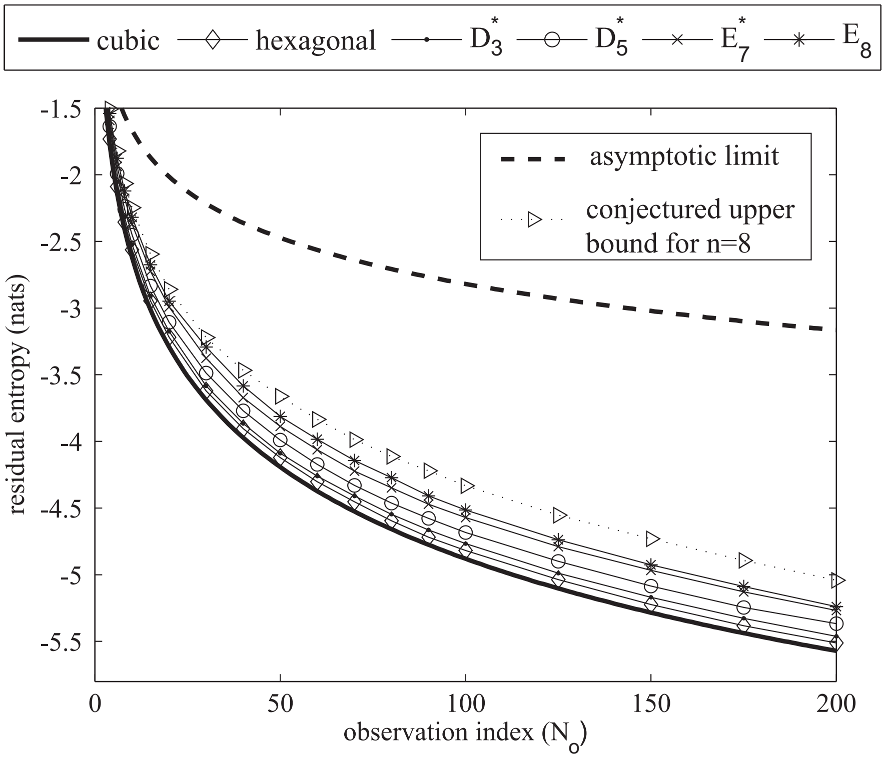


Figure 5.4: Equivocation per dimension for different lattices. All plots for the root lattices (but for the cubic one, which is theoretical) were obtained through Monte Carlo integration. The asymptotic limit corresponds to Eq. (5.41). The embedding distortion in all cases is $D_w = \alpha^2/12$, with $\alpha = 0.5$.

$$
\begin{aligned}
&I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}|M_1, \dots, M_{N_o}) \\
&\qquad = I(\tilde{\mathbf{Y}}_1; \mathbf{T}|M_1) + I(\tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}|\tilde{\mathbf{Y}}_1, M_1, \dots, M_{N_o}) \\
&\qquad = I(\tilde{\mathbf{Y}}_1; \mathbf{T}|M_1) + I(\tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}'|M_2, \dots, M_{N_o}), \qquad (5.37)
\end{aligned}
$$

where $\mathbf{T}' \sim U((1-\alpha)\mathcal{V}(\Lambda_n^*))$ is the dither conditioned on the first observation (as it follows from Property 5.1 and Lemma 5.2). Thus, each new observation conditioned on $\tilde{\mathbf{Y}}_1$ and M_1 can be written as[5]

$$
\tilde{\mathbf{Y}}_i = \mathbf{Z}_i + \mathbf{T}' + \mathbf{d}_{m_i},\ i = 2, \dots, N_o, \qquad (5.38)
$$

where $\mathbf{Z}_i \triangleq (1-\alpha)(\mathbf{X}_i - Q_{\Lambda_n^*}(\mathbf{X}_i))$ is the self-noise term, with the same statistical distribution as $\mathbf{T}'$, and hence with second moment per dimension $(1-\alpha)^2 P(\Lambda_n^*)$. From Eq. (5.37), it can be seen that the following equality holds:

$$
h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}, M_1, \dots, M_{N_o}) = h(\mathbf{T}'|\tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}, M_2, \dots, M_{N_o}),\ \text{for } N_o \geq 2, \qquad (5.39)
$$

[5] As discussed before, the residual entropy in the KMA scenario does not depend on the specific message sequence as long as this is known, so we consider $\mathbf{d}_{m_i} = \mathbf{0}\,\forall\ i = 1, \dots, N_o$, without loss of generality for the remaining of this section and in the corresponding appendices.

so we can use the second term of (5.37) for obtaining a lower bound on the equivocation per dimension, as shown in Appendix C.3:

$$\tfrac{1}{n}h(\mathbf{T}'|\tilde{\mathbf{Y}}_2,\ldots,\tilde{\mathbf{Y}}_{N_o},M_2,\ldots,M_{N_o}) \geq$$

$$\frac{N_o}{2}\log\left(\frac{P(\Lambda_n^*)}{G(\Lambda_n^*)}\right) - \frac{(N_o-1)}{2}\log\left(2\pi e P(\Lambda_n^*)\right) - \frac{1}{2}\log(N_o) + \log(1-\alpha). \qquad (5.40)$$

This lower bound is loose for small n, but the next result demonstrates its asymptotic tightness.

Theorem 5.2 (Asymptotic equivocation for good lattices). In the limit when $n \to \infty$, using the optimum lattice quantizer Λ_n^*, the equivocation per dimension in lattice DC-DM is given by

$$\lim_{n\to\infty}\frac{1}{n}h(\mathbf{T}'|\tilde{\mathbf{Y}}_2,\ldots,\tilde{\mathbf{Y}}_{N_o},M_2,\ldots,M_{N_o})$$

$$= \frac{1}{2}\log(2\pi e D_w) - \frac{1}{2}\log(N_o) + \log\left(\frac{1-\alpha}{\alpha}\right), \text{ for } N_o \geq 2, \qquad (5.41)$$

where D_w is the embedding distortion per dimension (5.14).

Proof: See Appendix C.4. ■

Notice that when $n \to \infty$, (5.40) coincides with (5.41), because $G(\Lambda_n^*) \to 1/2\pi e$. The first term in (5.41) accounts for the relation between the embedding distortion and the a priori entropy of the secret dither. The second term tells us how the equivocation decreases with N_o, and the third term shows the dependence with the distortion compensation parameter α, which basically introduces a constant shift in the equivocation curve (recall that for $\alpha = 1$, the residual entropy is $-\infty$ for $N_o \geq 1$). The asymptotic value of the equivocation is plotted in Figure 5.4 for reference, showing the gap with the root lattices studied before. The above theorem is the formal statement of a more intuitive result: the Voronoi region of Λ_n^* tends to a sphere, and in turn the uniform distribution in $\mathcal{V}(\Lambda_n^*)$ tends asymptotically to a Gaussian distribution (in the normalized entropy sense) [222]; hence, roughly speaking, each modulo-Λ reduced observation (Eq. (5.38)) becomes closer to a Gaussian distribution with variance D_w/α^2, whose mean is given by the secret dither (also with the same statistical distribution). This interpretation brings more insight in the comparison of the theoretical security between lattice DC-DM and additive spread spectrum methods, whose embedding function is given by Eq. (3.2). Notice that the resemblance between this embedding function and (5.38) implies similar security properties for both methods. Considering

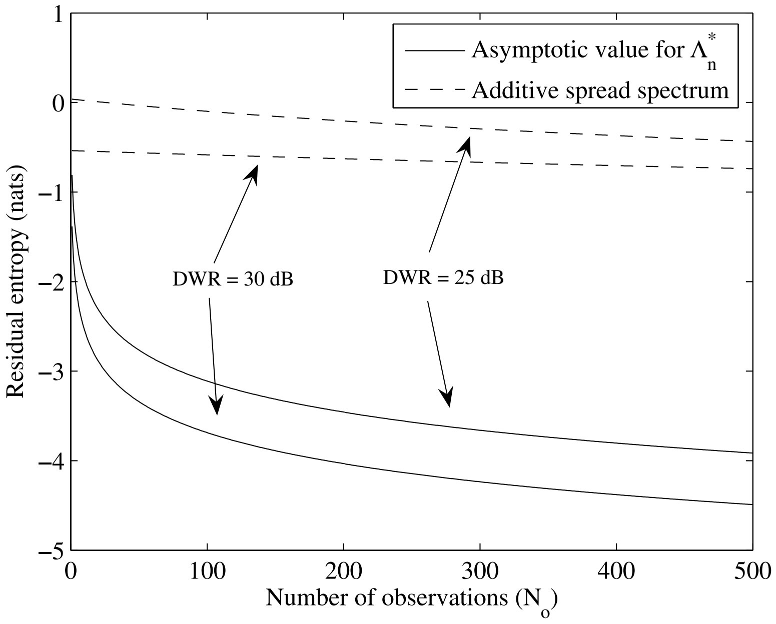


Figure 5.5: Comparison, in terms of equivocation per dimension, between lattice DC-DM and additive spread spectrum. $\alpha = 0.7$ for DC-DM.

that $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \cdot \mathbf{I}_n)$ and $\mathbf{S} \sim \mathcal{N}(\mathbf{0}, \sigma_S^2 \cdot \mathbf{I}_n)$, it was shown in Section 3.3.1 that

$$\frac{1}{n} h(\mathbf{S}|\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}, M_1, \ldots, M_{N_o}) = \frac{1}{2}\log(2\pi e D_w) - \frac{1}{2}\log\left(1 + N_o \frac{D_w}{\sigma_X^2}\right), \quad (5.42)$$

where now $D_w = \sigma_S^2$. It can be readily seen that the decrease in the equivocation for additive spread spectrum is determined by the ratio σ_S^2/σ_X^2, which is usually very small due to imperceptibility constraints. Instead, for the lattice data hiding scheme after the modulo-Λ reduction, the power of both the watermark and the host interference are the same, i.e., $(1-\alpha)^2 P(\Lambda_n^*)$; this explains the term $\frac{1}{2}\log(N_o)$ in (5.41) and the rapid decrease of the equivocation, compared to that of (5.42).

Figure 5.5 shows a comparison between lattice DC-DM and additive spread spectrum for different values of embedding distortion, parameterized by the DWR.

5.2.4 Bounds on the estimation error

We resort here to the lower bound on the estimation error provided by Lemma 2.3. Let σ_E^2 denote the variance per dimension of the estimation error defined in Section 2.4. Substituting Eq. (5.41) into (2.10), we arrive at the following bound for $n \to \infty$ and the optimal lattice quantizer:

$$\sigma_E^2 \geq \frac{(1-\alpha)^2 P(\Lambda_n^*)}{N_o}, \tag{5.43}$$

The above bound is attained using the simple averaging estimator, but taking into account that the observations must be properly shifted in order to avoid problems with the modulo-Λ reduction; thus, if we define

$$\tilde{\mathbf{v}}_i = (\tilde{\mathbf{y}}_i - \mathbf{d}_{m_i} - \tilde{\mathbf{y}}_1 + \mathbf{d}_{m_1}) \bmod \Lambda,\ i = 1, \ldots, N_o, \tag{5.44}$$

then the optimal dither estimator for $\Lambda_n^*, n \to \infty$, is given by

$$\hat{\mathbf{t}}_{av} = \left(\tilde{\mathbf{y}}_1 - \mathbf{d}_{m_1} + \frac{1}{N_o} \sum_{i=1}^{N_o} \tilde{\mathbf{v}}_i \right) \bmod \Lambda. \tag{5.45}$$

The achievability of (5.43) follows from the fact that, for Λ_n^*, the self-noise and the secret dither follow asymptotically a Gaussian distribution as $n \to \infty$. Thus, this result about the estimation error can be compared to the estimation error for the cubic lattice; since we are interested in computing the behavior for large N_o, we make use of the approximation

$$H_{N_o} \approx \log(N_o) + \gamma,$$

which is asymptotically tight for large N_o, with $H_{N_o} \triangleq \sum_{i=1}^{N_o} \frac{1}{i}$ the harmonic number and γ the Euler-Mascheroni constant [17], defined as $\gamma \triangleq \lim_{N_o \to \infty} H_{N_o} - \log(N_o)$. In this case we have, using (5.34)

$$\begin{aligned} \sigma_E^2 &\geq \frac{1}{2\pi e} \exp(2\left(\log((1-\alpha)\Delta) - H_{N_o} + 1\right)) \\ &\approx \frac{1}{2\pi e} \exp(2\left(\log((1-\alpha)\Delta) - \log(N_o) + 1 - \gamma\right)) \\ &= \frac{1}{2\pi e^{2\gamma - 1}} \cdot \frac{(1-\alpha)^2 \Delta^2}{N_o^2}. \end{aligned} \tag{5.46}$$

Thus, the variance per dimension approximately decreases with the inverse of the squared number of observations.

In order to illustrate the tightness of this bound, it will be compared with the exact error variance of the optimal dither estimator. For the cubic lattice, dither estimation may be carried out independently for each component without loss of optimality. It is a well known result that the optimal dither estimator in a mean-squared error sense is given by the mean value of the dither conditioned on the N_o observations: in our case, the ith component of the dither is uniformly distributed in an interval $[x_1, x_2]$;

hence, the optimal estimate is $\hat{t} = (x_1 + x_2)/2$, and the variance per dimension of the estimation error is

$$\sigma_E^2 = E\left[(T - \hat{t})^2\right] = \mathrm{var}(T) = \frac{1}{12} \cdot E[W^2], \tag{5.47}$$

where $w = |x_2 - x_1|$ is the width of the feasible interval, and the expectation is taken over the joint pdf of the observations. Actually, this expectation may be computed by replacing $\log(w)$ by w^2 in Eq. (C.3) of Appendix C.2, resulting in

$$\sigma_E^2 = \frac{1}{2} \cdot \frac{(1-\alpha)^2 \Delta^2}{2 + 3N_o + N_o^2}, \tag{5.48}$$

which for large N_o is dominated by the term N_o^2, differing from the right hand side of (5.46) only in a constant multiplying factor. Note that due to the approximation of H_{N_o} used in (5.46), the latter is a lower bound only for $N_o \geq 2$; nevertheless, making use of the exact expression for H_{N_o}, the right hand side of (5.46) can be shown to be always lower than (5.48).

5.2.5 Constant Message Attack (CMA)

The discussion about the KMA case given above can be easily extended in order to consider the CMA scenario. The easiest way of addressing this scenario is to regard it as a collection of several KMA problems. When the message embedded is unknown but unchanged for the whole sequence of observations, the conditional pdf of the dither after N_o observations can be expressed as

$$f(\mathbf{t}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m) = \frac{1}{p} \sum_{m=0}^{p-1} f(\mathbf{t}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, m, \ldots, m), \tag{5.49}$$

where m stands for the constant message, and p denotes the size of the alphabet. This means that the feasible region $\mathcal{S}_{N_o}^{CMA}$ for the dither in the CMA case is simply the union of the feasible regions of p KMA problems. Formally,

$$\mathcal{S}_{N_o}^{CMA} = \bigcup_{m=0}^{|\mathcal{M}|-1} (\mathcal{S}_{N_o} + \mathbf{d}_m), \tag{5.50}$$

with $\mathcal{S}_{N_o}$ defined in Property 5.1. Using the result of Lemma 5.1, it follows that the the different regions that constitute $\mathcal{S}_{N_o}^{CMA}$ become disjoint for sufficiently large N_o. In such case, the residual entropy is again maximized if $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$ is chosen, but it is not necessarily the optimal distribution for all N_o. Due to (5.49), the residual entropy can be upper bounded as

$$h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}, \mathrm{CMA}) \leq h(\mathbf{T}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}, M_1, \ldots, M_{N_o}) + \log(|\mathcal{M}|), \tag{5.51}$$

resulting in a lower bound to the information leakage. Equality in (5.51) is achieved when the regions $\mathcal{S}_{N_o} + \mathbf{d}_m$ are disjoint, which means that, as N_o increases, the bound will be asymptotically tight. However, if the value of α is above a certain threshold (which depends on the lattice partition) such regions are always disjoint, and the bound is reached for all N_o; this is the case, for instance, when $\alpha > \alpha_T = 1 - \frac{1}{p}$, for self-similar partitions [174], [112].

Although the gap between KMA and CMA in terms of equivocation is small, the CMA scenario introduces an ambiguity that makes impossible the perfect estimate of the secret dither. This ambiguity will be formally stated in Lemma 5.4 and further discussed in Section 5.3.2.

5.3 Theoretical security analysis of WOA

The amount of information that leaks from the observations is quantified by means of the mutual information $I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T})$. Recall that, according to (2.21), this mutual information can be rewritten in a more illustrative manner as

$$\begin{aligned} I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}) & \\ = \; & I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T} | M_1, \ldots, M_{N_o}) + I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o}) \\ - \; & I(\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}; M_1, \ldots, M_{N_o} | \mathbf{T}). \end{aligned} \tag{5.52}$$

The first term in the right hand side of (5.52) is the information leakage for KMA, which was studied in Sect 5.2. The third term in the right hand side of (5.52) represents the achievable rate for a fair user, i.e. knowing the secret dither $\mathbf{T}$, whereas the second term is the rate achievable by unfair users (which is not null, in general) that do not know $\mathbf{T}$. In this section we consider, first, the possibility of achieving perfect secrecy about $\mathbf{T}$. Thereafter, we study the asymptotic behavior of the information leakage about $\mathbf{T}$ in other situations where perfect secrecy is not achieved, and finally we analyze a practical lattice data hiding scheme.

The concept of feasible region, defined in Section 5.2, will be often recalled for proving several results in the WOA scenario. However, this time it is necessary to generalize the definition of feasible region to account for any possible embedded message (since the true embedded message is a priori unknown). Therefore, in this section we will take into account the next definition.

Definition 5.3. The feasible region for a sequence of observations $\{\tilde{\mathbf{y}}_k, k = 1, \ldots, N_o\}$

and a message sequence $\mathbf{m}^{(k)},\ k=1,\ldots,p^{N_o}$, is defined as

$$\mathcal{S}_{N_o}(\mathbf{m}^{(k)}) \triangleq \bigcap_{j=1}^{N_o} \mathcal{D}_j(m_j^k), \tag{5.53}$$

$$\mathcal{D}_j(m_j^k) \triangleq (\tilde{\mathbf{y}}_j - \mathbf{d}_{m_j^k} - \mathcal{Z}(\Lambda)) \mod \Lambda. \tag{5.54}$$

5.3.1 Theoretical and practical perfect secrecy

If the lattice data hiding scheme fulfills the condition

$$\begin{aligned} I(\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o};\mathbf{T}) &= h(\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o}) - h(\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o}|\mathbf{T}) \\ &= h(\mathbf{T}) - h(\mathbf{T}|\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o}) = 0,\ \forall\, N_o, \end{aligned} \tag{5.55}$$

then it is said to provide "perfect secrecy", meaning that no information about $\mathbf{T}$ can be obtained from the observations, no matter the computational effort employed by the attacker. Under the assumption that $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, it is elementary to prove that $\tilde{\mathbf{Y}}_k$ is uniformly distributed over $\mathcal{V}(\Lambda)$. Hence, the closer is the distribution of $\tilde{\mathbf{Y}}_k$ conditioned on $\mathbf{T}$ to the uniform over $\mathcal{V}(\Lambda)$, the more secure is the scheme. Once the shaping lattice is fixed, the two free parameters for tuning such distribution are the set of coset leaders $\mathcal{C}_p$ and the parameter α. The possibility of achieving perfect secrecy is considered below in Lemma 5.3 and Proposition 5.1. We recall that the messages embedded in different observations are assumed to be independent.

Lemma 5.3 (Asymptotic perfect secrecy). Consider a sequence of nested lattice codes $(\Lambda, \mathcal{C}_p^*)$ such that $r_c(\Lambda_f) \to 0$ as $p \to \infty$. If $\alpha < 1$, this sequence of lattice codes asymptotically achieves perfect secrecy as $p \to \infty$, and the statistical distribution of $\tilde{\mathbf{Y}}_k$ conditioned on $\mathbf{T}$ uniformly converges to the uniform over $\mathcal{V}(\Lambda)$, $\forall\, k = 1,\ldots,N_o$.

Proof: See Appendix C.5.[6] ■

For the lattice codes considered in Lemma 5.3, the information leakage is reduced as the size of the alphabet is increased, since the distribution of $\tilde{\mathbf{Y}}_k$ is uniformly convergent. Bear in mind that if the condition $r_c(\Lambda_f) \to 0$ does not hold (i.e. if the coset leaders do not properly "cover" the whole Voronoi region of Λ), the resulting code does not necessarily offer good secrecy, even for high embedding rates (see sections 5.3.3 and 5.3.4, which deal with a lattice repetition code). It must be observed that an infinite

[6] The condition $\alpha < 1$ imposed in Lemma 5.3 comes from the necessity of having a continuous, Riemann integrable pdf for our proof to be valid. The case $\alpha = 1$, for which $f(\tilde{\mathbf{y}}_1|\mathbf{t})$ becomes a probability mass function, cannot be dealt with using the arguments of Appendix C.5.

alphabet size is not affordable in practice. However, this choice of alphabet is used in [112] for showing that a nested lattice code asymptotically achieves the capacity of the modulo-lattice Gaussian channel. Thus, Lemma 5.3 shows, in conjunction with [112], that simultaneous maximization of robustness and security is theoretically (asymptotically) possible. Finally, we would like to remark that the result of asymptotic perfect secrecy holds for any distribution of the secret dither $\mathbf{T}$, not necessarily the uniform over $\mathcal{V}(\Lambda)$ (cf. Appendix C.5). In the next proposition, a nested code achieving perfect secrecy and realizable in practice is proposed.

Proposition 5.1 (Achievable perfect secrecy). Any self-similar lattice code of rate $R = \log(p)/n$ and distortion compensation parameter $\alpha = 1 - p^{-\frac{1}{n}}$ achieves perfect secrecy. In that case $\tilde{\mathbf{Y}}_k$ conditioned on $\mathbf{T}$ is uniform over $\mathcal{V}(\Lambda)$, $\forall\, k = 1, \ldots, N_o$.

Proof: Similarly to the proof of Lemma 5.3, the proof of perfect secrecy can be reduced to showing that the resulting scheme fulfills the condition $I(\tilde{\mathbf{Y}}_1; \mathbf{T}) = h(\tilde{\mathbf{Y}}_1) - h(\tilde{\mathbf{Y}}_1|\mathbf{T}) = 0$.

For $\alpha = 1 - p^{-\frac{1}{n}}$, we have $\mathcal{Z}(\Lambda) = p^{-\frac{1}{n}}\mathcal{V}(\Lambda) = \mathcal{V}(\Lambda_f)$, so the pdf defined in (5.18) is given by

$$\varphi(\mathbf{x}) = p \cdot (\mathrm{vol}(\mathcal{V}(\Lambda)))^{-1} \cdot \phi_{\mathcal{V}(\Lambda_f)}(\mathbf{x}).$$

Hence, under the assumption of equiprobable symbols we can write

$$\begin{aligned} f(\tilde{\mathbf{y}}_1|\mathbf{T}=\mathbf{t}) &= \frac{1}{p}\sum_{i=0}^{p-1} \varphi(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_i \mod \Lambda) \\ &= (\mathrm{vol}(\mathcal{V}(\Lambda)))^{-1} \sum_{i=0}^{p-1} \phi_{\mathcal{V}(\Lambda_f)}(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_i \mod \Lambda). \end{aligned} \tag{5.56}$$

Taking into account that self-similar partitions fulfill the next "covering property":

$$\bigcup_{i=0}^{p-1} (\mathcal{V}(\Lambda_f) - \mathbf{t} - \mathbf{d}_i) \mod \Lambda = \mathcal{V}(\Lambda), \quad \bigcap_{i=0}^{p-1} (\mathcal{V}(\Lambda_f) - \mathbf{t} - \mathbf{d}_i) \mod \Lambda = \emptyset, \tag{5.57}$$

it follows that for every $\tilde{\mathbf{y}}_1 \in \mathcal{V}(\Lambda)$ there exists exactly one $\mathbf{d}_i$ such that $(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_i) \mod \Lambda \in \mathcal{V}(\Lambda_f)$. Hence, (5.56) becomes

$$f(\tilde{\mathbf{y}}_1|\mathbf{T}=\mathbf{t}) = \frac{1}{\mathrm{vol}(\mathcal{V}(\Lambda))}\ \forall\, \tilde{\mathbf{y}}_1 \in \mathcal{V}(\Lambda), \tag{5.58}$$

and $h(\tilde{\mathbf{Y}}_1|\mathbf{T}=\mathbf{t}) = \log(\mathrm{vol}(\mathcal{V}(\Lambda)))$. Since the entropy of a continuous random variable with bounded support is upper bounded by the log-volume of its support set, we can write

$$h(\tilde{\mathbf{Y}}_1|\mathbf{T}=\mathbf{t}) \leq h(\tilde{\mathbf{Y}}_1) \leq \log(\mathrm{vol}(\mathcal{V}(\Lambda))).$$

Thus, we have $h(\tilde{\mathbf{Y}}_1) = h(\tilde{\mathbf{Y}}_1|\mathbf{T}) = \log(\mathrm{vol}(\mathcal{V}(\Lambda)))$, resulting in a null information leakage, and $\tilde{\mathbf{Y}}_1$ is necessarily uniform over $\mathcal{V}(\Lambda)$ regardless the distribution of $\mathbf{T}$. ■

Some important remarks to this result are given below.

Remark 5.1. The proof of perfect secrecy in Proposition 5.1 relies on the lattice code itself rather than on the statistical distribution of $\mathbf{T}$. Actually, the result holds for any distribution of the secret dither $\mathbf{T}$. However, it must be taken into account that the value of α that provides perfect secrecy can be conflicting with other requirements (e.g. error probability), so it is important to properly choose the distribution of $\mathbf{T}$ for maximizing the security when perfect secrecy cannot be attained. This distribution has been shown in [194] to be the uniform over $\mathcal{V}(\Lambda)$, which yields $\tilde{\mathbf{Y}}_k$ also uniform over $\mathcal{V}(\Lambda)$.

Remark 5.2. It is possible to show that for $\alpha_k = 1 - kp^{-\frac{1}{n}}$, $k = 1, \ldots, p^{\frac{1}{n}} - 1$, the condition of perfect secrecy still holds. However, for $k > 1$ there are overlaps between adjacent symbols that produce nonzero error probability even in the absence of noise. This makes necessary the use of channel coding (i.e. error correcting codes) to recover the embedded message reliably, thus breaking the hypothesis of independence between messages embedded in different blocks. As a result, perfect secrecy about $\mathbf{T}$ cannot be assured. To see this, consider a simple example with 2 observations $\{\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2\}$, where the embedding parameters fulfill the conditions for perfect secrecy stated in Proposition 5.1. The mutual information between observations and secret dither is written as

$$I(\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2|\mathbf{T}) = I(\tilde{\mathbf{Y}}_1; \mathbf{T}) + I(\tilde{\mathbf{Y}}_2; \mathbf{T}|\tilde{\mathbf{Y}}_1), \tag{5.59}$$

whereas the rightmost term of (5.59) can be written as $I(\tilde{\mathbf{Y}}_2; \mathbf{T}|\tilde{\mathbf{Y}}_1) = h(\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1) - h(\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1, \mathbf{T})$. Given only $\tilde{\mathbf{Y}}_1$, no information about $\mathbf{T}$ is leaked, so $h(\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1) = h(\tilde{\mathbf{Y}}_2)$. Given $\mathbf{T}$, we have that $\tilde{\mathbf{Y}}_1$ and $\tilde{\mathbf{Y}}_2$ are independent because M_1, M_2 are independent. Hence, $h(\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1, \mathbf{T}) = h(\tilde{\mathbf{Y}}_2|\mathbf{T})$, resulting $I(\tilde{\mathbf{Y}}_2; \mathbf{T}|\tilde{\mathbf{Y}}_1) = I(\tilde{\mathbf{Y}}_2; \mathbf{T}) = 0$. Now, consider the case where M_1, M_2 are not independent. Given only $\tilde{\mathbf{Y}}_1$, no information about $\mathbf{T}$ nor M_1 is obtained, so we have $h(\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1) = h(\tilde{\mathbf{Y}}_2)$ again. However, given $\tilde{\mathbf{Y}}_1$ and $\mathbf{T}$, information about M_1 is leaked. Since M_2 is dependent on M_1, we have $h(\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1, \mathbf{T}) \leq h(\tilde{\mathbf{Y}}_2|\mathbf{T})$, resulting $I(\tilde{\mathbf{Y}}_2; \mathbf{T}|\tilde{\mathbf{Y}}_1) \geq 0$.

Remark 5.3. The proof of the proposition resorts to the flat-host assumption to show null information leakage. This means that, in practice, small information leakages may exist due to the finite variance of the host signal, which causes the host distribution to not be strictly uniform in each quantization cell. However, this information leakage seems to be hardly exploitable in practical attacks.

Remark 5.4. Perfect secrecy about $\mathbf{T}$ does not necessarily mean perfect secrecy about the embedded messages. Using (5.52), the rate for the attacker under the condition of perfect secrecy is given by

$$
\begin{aligned}
&I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; M_1, \dots, M_{N_o}) \\
&\qquad = I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; M_1, \dots, M_{N_o}|\mathbf{T}) - I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}|M_1, \dots, M_{N_o}). \qquad (5.60)
\end{aligned}
$$

This "unfair" rate is studied in Section 5.3.3 for the repetition coding lattice scheme.

5.3.2 Asymptotic analysis and comparison with KMA

When perfect secrecy is not attained, a closed-form expression for the information leakage cannot be given, in general. We are interested here in studying the general behavior of the information leakage for large N_o and comparing it with the KMA scenario.

The next lemma shows that without the knowledge of $\mathbf{T}$ there exists an irreducible ambiguity in the estimation of the embedded message sequence.

Lemma 5.4 (Equivalence classes in the message space). Given N_o observations, consider the a priori message space $\mathcal{M}^{N_o}$ where all the message sequences are assumed to be a priori equiprobable. For $\mathbf{m}^{(1)}, \mathbf{m}^{(2)} \in \mathcal{M}^{N_o}$, let us define the equivalence relation

$$
\begin{aligned}
&\mathbf{m}^{(1)} \sim \mathbf{m}^{(2)} \text{ if } [\mathbf{d}_{m_1^2}, \dots, \mathbf{d}_{m_{N_o}^2}] \\
&\qquad = [(\mathbf{d}_{m_1^1} + \mathbf{d}_j) \mod \Lambda, \dots, (\mathbf{d}_{m_{N_o}^1} + \mathbf{d}_j) \mod \Lambda], \text{ for some } j \in \mathcal{M}. \qquad (5.61)
\end{aligned}
$$

Each equivalence class is composed of p elements, and the sequences belonging to the same equivalence class have all the same a posteriori probability.

Proof: By the additive structure of Λ_f, the operation $\mathbf{d}_l = (\mathbf{d}_k + \mathbf{d}_j) \mod \Lambda$, with $l, k, j \in \mathcal{M}$, defines a bijective mapping $\mathbf{d}_k \to \mathbf{d}_l$. Then, for a message sequence $\mathbf{m}^{(1)} = [m_1^1, \dots, m_{N_o}^1]$, the operation

$$
[(\mathbf{d}_{m_1^1} + \mathbf{d}_j) \mod \Lambda, \dots, (\mathbf{d}_{m_{N_o}^1} + \mathbf{d}_j) \mod \Lambda]
$$

yields p different sequences of length N_o when j is varied from 0 to $p-1$. Thus, each equivalence class as defined in (5.61) is composed of p elements. The a posteriori probability of a message sequence is derived in Appendix C.6. For a sequence $\mathbf{m}^{(1)} \in \mathcal{M}^{N_o}$, combining equations (C.56) and (C.60) of Appendix C.6 we arrive at

$$
\begin{aligned}
\Pr(\mathbf{m}^{(1)}|\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{N_o}) &= \Pr(\mathbf{d}_{m_1^1}, \dots, \mathbf{d}_{m_{N_o}^1}|\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{N_o}) \\
&= \frac{\mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(1)}))}{(\mathrm{vol}(\mathcal{Z}(\Lambda)))^{N_o} \cdot \mathrm{vol}(\mathcal{V}(\Lambda))} \cdot \frac{\Pr(\mathbf{m}^{(1)})}{f(\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{N_o})}. \qquad (5.62)
\end{aligned}
$$

Now, let us define $\hat{\mathcal{S}}_{N_o}(\mathbf{m}^{(1)}) = \bigcap_{k=1}^{N_o}(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^1} - \mathbf{u} - \mathcal{Z}(\Lambda)) \mod \Lambda$, with $\mathbf{u}$ an arbitrary vector constant for all k. Clearly,

$$\hat{\mathcal{S}}_{N_o}(\mathbf{m}^{(1)}) = (\mathcal{S}_{N_o}(\mathbf{m}^{(1)}) - \mathbf{u}) \mod \Lambda,$$

so

$$\mathrm{vol}(\hat{\mathcal{S}}_{N_o}(\mathbf{m}^{(1)})) = \mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(1)})).$$

Hence, if $\mathbf{u} = \mathbf{d}_j$, then

$$[(\mathbf{d}_{m_1^1} + \mathbf{u}) \mod \Lambda, \ldots, (\mathbf{d}_{m_{N_o}^1} + \mathbf{u}) \mod \Lambda]$$

$$= [(\mathbf{d}_{m_1^1} + \mathbf{d}_j) \mod \Lambda, \ldots, (\mathbf{d}_{m_{N_o}^1} + \mathbf{d}_j) \mod \Lambda] = [\mathbf{d}_{m_1^2}, \ldots, \mathbf{d}_{m_{N_o}^2}],$$

and it follows that $\mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(1)})) = \mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(2)}))$. Inserting this result in (5.62) and recalling that the message sequences are assumed to be a priori equiprobable, the lemma follows. ■

By virtue of Lemma 5.4, if the messages embedded in different blocks are mutually independent, then the attacker can aspire (at most) at reducing the uncertainty about the embedded message sequence to a set of p equiprobable sequences. The following theorem states how this ambiguity affects the information leakage about $\mathbf{T}$ for large N_o.

Let us denote by $\mathcal{L}_k$ the elements of $\mathcal{M}$ with nonnull probability, given $\tilde{\mathbf{y}}_k$ and $\mathbf{t}$. Note that for any nested lattice code there exists a value α_0 such that for $\alpha > \alpha_0$ we can assure $\mathcal{Z}(\Lambda) \subset \mathcal{V}(\Lambda_f)$, so $|\mathcal{L}_k| = 1$, i.e. error-free decoding is guaranteed in the absence of noise. Obviously, $\alpha_0 > 0.5$ in any case, although it will depend on the lattice code, in general.

Theorem 5.3 (Asymptotics of the loss function for lattice data hiding). For any nested lattice code, if α is chosen such that $\mathcal{Z}(\Lambda) \subset \mathcal{V}(\Lambda_f)$, then

$$\lim_{N_o \to \infty} \frac{1}{n}\delta(N_o) = \frac{1}{n}\log(p) = R,$$

where $\delta(N_o)$ is the loss function defined in (2.20).

Proof: See Appendix C.7. ■

The result of Theorem 5.3 has two main implications:

1. For low embedding rates (i.e. low R) the information per dimension that leaks about $\mathbf{T}$ is approximately the same as if the attacker knew the embedded messages. This case is highly relevant in practice, since in practical scenarios the watermarker usually resorts to low embedding rates that allow to recover the embedded message without the use of complex channel coding schemes.

2. As shown in Appendix C.7, when $N_o \to \infty$, the feasible regions associated to the only message sequences with nonnull probability converge to p different vectors of the form $(\mathbf{t} - \mathbf{d}_j) \mod \Lambda$, $j \in \mathcal{M}$, which are equiprobable. Thus, unambiguous estimation of the secret dither vector is not possible in the WOA scenario. It is interesting to note that this ambiguity also holds in the CMA scenario, which was addressed in Section 5.2.5.

The values of α for which Theorem 5.3 holds guarantee that no decoding errors occur in the absence of noise. However, in some cases it is advantageous to choose smaller values of α, e.g. when a certain degree of attacking noise is expected [112],[108]. In such cases, $|\mathcal{L}_k|$ may be larger than 1, giving rise to several sequences with nonnull probability, complicating the analysis of the loss function. However, for lattice codes obtained through Construction A, the equivalence classes of Lemma 5.4 present a simple structure. The equivalence relation (5.61) can be expressed as

$$\mathbf{m}^{(1)} \sim \mathbf{m}^{(2)} \text{ if } \mathbf{m}^{(2)} = (\mathbf{m}^{(1)} + j \cdot \mathbf{1}) \mod p, \text{ for some } j \in \mathcal{M}, \tag{5.63}$$

where $\mathbf{1}$ denotes the vector with its components equal to 1, and the modulo operation is applied componentwise. The proof follows from Lemma 5.4 simply by observing that $(\mathbf{d}_k + \mathbf{d}_j) \bmod \Lambda = \mathbf{d}_{(k+j) \bmod p}$ for Construction A. Using this result, a lower bound for the loss function is derived in Lemma 5.5 below.

Lemma 5.5 (Bound on the loss function for Construction A). If α is such that $|\mathcal{L}_k| \leq \left\lceil \frac{p}{2} \right\rceil$, $\forall\ k = 1, \ldots, N_o$, then, for any nested lattice code obtained through Construction A, the asymptotic loss per dimension between the information leakage in KMA and WOA, given by $\lim_{N_o \to \infty} \frac{1}{n} g(N_o))$, is bounded from below by R.

Proof: See Appendix C.8. ■

The achievability of this lower bound is discussed in Section 5.3.3 for the repetition coding lattice scheme.

Remark 5.5. If the messages conveyed by different observations were not independent, the residual uncertainty expressed in Theorem 5.3 and Lemma 5.5 would be further reduced. This is the case, for instance, when a channel code is applied, since it could provide the attacker with information about the a priori probabilities of each message sequence. This consideration is important when high robustness against noise is sought or when α is small, since the use of channel codes is mandatory in these cases for lowering the decoding error probability.

5.3.3 Theoretical results for cubic shaping lattices with repetition coding

Lattice data hiding with repetition coding using scalar quantizers, also known as Scalar Costa Scheme (SCS) with repetition coding [108], is one of the most popular schemes for lattice data hiding. Redundant embedding of the information is performed by repeatedly embedding the same message in n different host samples, using a pair of scalar lattices $\Lambda = \Delta\mathbb{Z}$, $\Lambda_f = \Delta\mathbb{Z}/p$, which yield $\mathcal{V}(\Lambda) = [-\Delta/2, \Delta/2)$, and $d_k = (\Delta k/p) \mod \Delta,\ k = 0, \dots, p-1$. For a n-dimensional host vector $\mathbf{X}_k$, the embedding function (5.13) is particularized to

$$Y_{k,i} = X_{k,i} + \alpha(Q_\Lambda(X_{k,i} - d_{M_k} - T_i) - X_{k,i} + d_{M_k} + T_i),\ i = 1, \dots, n, \tag{5.64}$$

where the subindex i indicates the ith component of the n-dimensional vector. This simple coding scheme, which results in a code of rate $R = \log(p)/n$, is equivalent to a code obtained through Construction A with $\mathbf{g} = [1, \dots, 1]^T$ and $\Lambda = \Delta\mathbb{Z}^n$. Bear in mind that, due to the redundant embedding of the message, the repetition scheme provides the attacker with more information about the embedded message than a scheme where n independent scalar embeddings are performed in parallel. Thus, one can intuitively realize that the repetition scheme analyzed here is less secure than the simple SCS. Figure 5.6 shows two examples of lattice repetition codes for $n = 2$. Although the robustness of this data hiding code has been analyzed in depth in [75], here we are interested in analyzing its security properties.

In order to obtain the information leakage, we rewrite the third term of Eq. (5.52) as

$$\begin{aligned} I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; M_1, \dots, M_{N_o} | \mathbf{T}) & \\ &= H(M_1, \dots, M_{N_o}) - H(M_1, \dots, M_{N_o} | \tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}, \mathbf{T}) \\ &= N_o \cdot (\log(p) - H(M_1 | \tilde{\mathbf{Y}}_1, \mathbf{T})). \end{aligned} \tag{5.65}$$

In turn, the second term in the right hand side of (5.52) can be expressed as

$$I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; M_1, \dots, M_{N_o}) = N_o \cdot \log(p) - H(M_1, \dots, M_{N_o} | \tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}).$$

By combining the two above equations with the information leakage for cubic lattices in the KMA scenario (cf. Sect. 5.2.1), the information leakage per dimension reads as

$$\begin{aligned} \frac{1}{n} I(\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}) &= \frac{N_o}{n} H(M_1 | \tilde{\mathbf{Y}}_1, \mathbf{T}) - \frac{1}{n} H(M_1, \dots, M_{N_o} | \tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_{N_o}) \\ &+ \sum_{i=2}^{N_o} \frac{1}{i} - \log(1 - \alpha),\quad N_o \geq 2,\ \alpha \geq 0.5. \end{aligned} \tag{5.66}$$

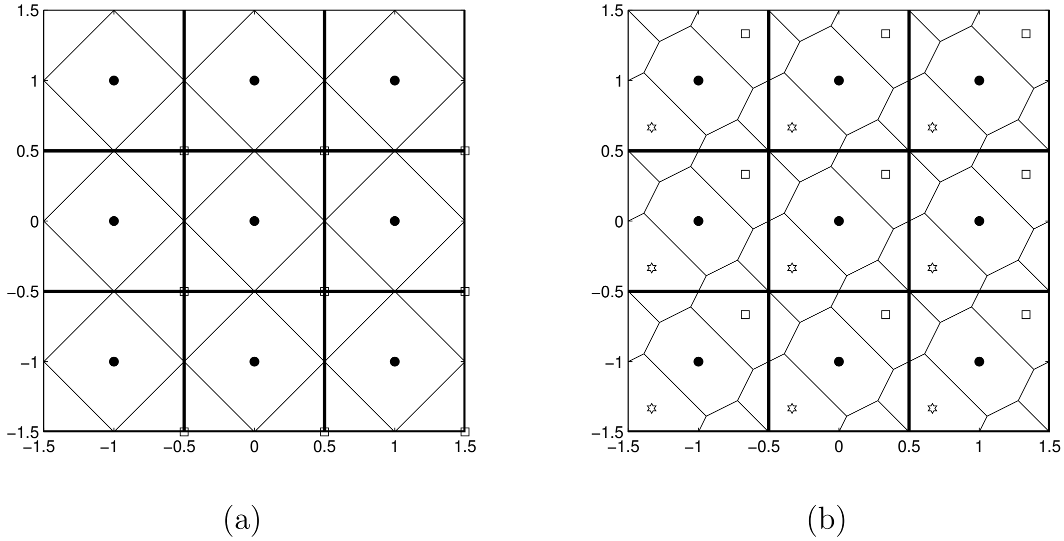


Figure 5.6: Illustration of repetition codes with $n = 2$ and $\Delta = 1$. Figure (a) represents a code with $p = 2$, where dots and squares represent the cosets for $m = 0$ and $m = 1$, respectively, and the fine lattice Λ_f is the well-known "checkerboard lattice" [82]. Figure (b) represents a code with $p = 3$. Notice that the coset leaders fall in the main diagonal of the Voronoi region.

Eq. (5.66) does not admit a closed-form expression, although it is possible to accurately obtain the entropies of interest in a numerical manner.

Computation of the achievable rate for fair users

The first term in the right hand side of (5.66) represents the uncertainty about the embedded message when the secret dither is known. Under the flat-host assumption, we can write

$$H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T}) = H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T} = \mathbf{0}) = E\left[H(M_1|\tilde{\mathbf{Y}}_1 = \tilde{\mathbf{y}}, \mathbf{T} = \mathbf{0})\right], \qquad (5.67)$$

where the expectation is taken over $\tilde{\mathbf{Y}}_1$. The a posteriori probability of a certain message m is given by

$$\begin{aligned}
&\Pr(m|\tilde{\mathbf{y}}, \mathbf{t} = \mathbf{0}) \\
&= \frac{f(\tilde{\mathbf{y}}|m, \mathbf{t} = \mathbf{0}) \cdot \Pr(m)}{f(\tilde{\mathbf{y}}|\mathbf{t} = \mathbf{0})} = \frac{\Pr(m)}{f(\tilde{\mathbf{y}}|\mathbf{t} = \mathbf{0})} \prod_{i=1}^{n} f(\tilde{y}_i|m, t_i = 0) \\
&= \frac{\Pr(m)}{f(\tilde{\mathbf{y}}|\mathbf{t} = \mathbf{0})} \prod_{i=1}^{n} \varphi((\tilde{y}_i - d_m) \mod \Lambda), \text{ for } \tilde{y}_i \in \bigcup_{k=0}^{p-1} (d_k + \mathcal{Z}(\Lambda) \mod \Lambda), \qquad (5.68)
\end{aligned}$$

where $\tilde{y}_i,\ i = 1, \ldots, n$, are the components of $\tilde{\mathbf{y}}$. From (5.68), we can see that the feasible messages (i.e. with non-null probability) are those whose coset leader is contained in the interval given by $\bigcap_{i=1}^{n}(\tilde{y}_i - d_m - \mathcal{Z}(\Lambda) \mod \Lambda)$, and that the feasible messages are equiprobable. Given the symmetry of the pdf of $\tilde{y}_i | t_i = 0$, we can write

$$H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T} = \mathbf{0}) = E\left[\log\left(\sum_{k=0}^{p-1} \phi_{\mathcal{H}}((\Delta \cdot k/p) \mod \Lambda)\right)\right], \text{ for } \alpha \geq 0.5, \quad (5.69)$$

where $\phi_{\mathcal{H}}(\cdot)$ is the indicator function defined in Eq. (5.1), and

$$\mathcal{H} \triangleq \left[\max_{i=1,\ldots,n}\{Z_i\} - (1-\alpha)\Delta/2,\ \min_{i=1,\ldots,n}\{Z_i\} + (1-\alpha)\Delta/2\right),$$

with $Z_i \sim U((1-\alpha)[-\Delta/2, \Delta/2))$.[7] Hence, the expectation in (5.69) is taken over Z_i. This expectation is obtained numerically by Monte Carlo integration.

Computation of the achievable rate for unfair users

The second term in the right hand side of (5.66) is given by

$$\frac{1}{n} E\left[H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{Y}}_1 = \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o} = \tilde{\mathbf{y}}_{N_o})\right], \quad (5.70)$$

where the expectation is taken over the observations. The a posteriori probability distribution of the message sequences can be obtained by combining the equations (C.56) and (C.60) of Appendix C.6. For a message sequence $\mathbf{m}^{(k)} = [m_1^k, \ldots, m_{N_o}^k]$,

$$\Pr(m_1^k, \ldots, m_{N_o}^k|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o})$$

$$= \Pr(m_1^k, \ldots, m_{N_o}^k) \cdot \frac{f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}|m_1^k, \ldots, m_{N_o}^k)}{\sum_{i=1}^{p^{N_o}} f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, \mathbf{m}^{(i)})} = \frac{\text{vol}(\mathcal{S}_{N_o}(m_1^k, \ldots, m_{N_o}^k))}{\sum_{i=1}^{p^{N_o}} \text{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(i)}))}. \quad (5.71)$$

For $\alpha \geq 0.5$, the feasible regions involved in the calculation of (5.71) are always modulo-Λ convex hypercubes [194], and as such they can be easily computed componentwise. The entropy (5.70) is obtained by Monte Carlo, computing the probability of the all the message sequences in a large set of realizations of $\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o}$. Notice that, although the cardinality of the message space grows exponentially with N_o, only the message sequences with non-null probability need to be taken into account, making the problem computationally feasible.

[7] Hence, in the case of repetition coding the problem of computing $H(M_1|\tilde{\mathbf{Y}}_1, \mathbf{T})$ can be seen as the dual of the problem of computing $h(T|\tilde{Y}_1, \ldots, \tilde{Y}_{N_o}, M_1, \ldots, M_{N_o})$ for a scalar lattice, which was addressed in Section 5.2.1.

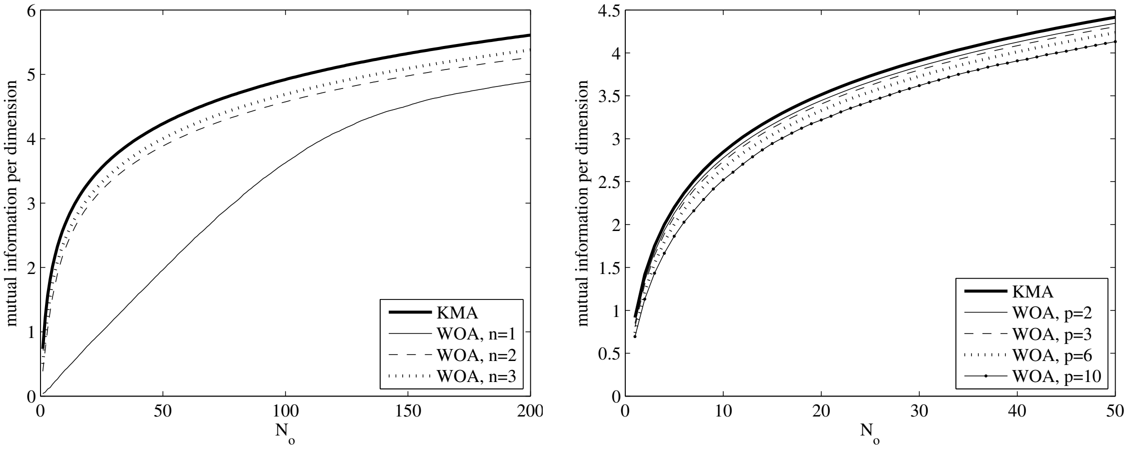
5.3.4 Numerical results

The results shown in this section support some of the conclusions drawn in sections 5.3.1 and 5.3.2. Figure 5.7 illustrates the information leakage about $\mathbf{T}$ for the repetition code and provides a comparison with the results obtained for the KMA scenario. Figure 5.7(a) shows the negative impact on the security level of increasing the dimensionality whilst keeping constant the size of the alphabet. Figure 5.7(b) shows the security improvement brought about by the increase of the alphabet size, although this improvement is not very significant. Finally, notice in both Figure 5.7(a) and Figure 5.7(b) that the loss between the information leakage for KMA and WOA tends asymptotically to a constant. For the case of Figure 5.7(a), Theorem 5.3 holds, so the asymptotic value of the loss is actually $\log(p)/n$. In Figure 5.7(b), Theorem 5.3 holds only for $p = 2$, although for larger values of p (for which Lemma 5.5 holds) the loss can still be seen to be approximately $\log(2)/n$. The loss is more deeply considered in the next paragraph.

Figure 5.8 shows the loss function defined in (2.22) for two instances of the lattice repetition code with different parameters. The code considered in Figure 5.8(a) is for $n = 1$ and $p = 2$, which is equivalent to binary SCS [108], the simplest lattice code. The marked signal observed by the attacker is (recall Eq. (5.17))

$$\tilde{Y}_k = (d_{M_k} + T + (1-\alpha)N_k) \mod \Lambda,$$

where N_k is uniform over the interval $[-\Delta/2, \Delta/2)$, with variance $(1-\alpha)^2\Delta^2/12$. For this code, the equivocation for a fair user is always 0 in a noiseless scenario whenever $\alpha \geq 0.5$ [108]. Thus, in this case the loss function is equivalent to the equivocation about the embedded messages for the attacker, i.e. $H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{N_o})$. It can be seen that the loss tends in all cases to $\log(2)$ as N_o is increased, as stated in Theorem 5.3, except in the case $\alpha = 0.5$, where the loss increases indefinitely. The explanation for this behavior is simple: according to the scheme proposed in Proposition 5.1, the combination $n = 1$, $p = 2$ and $\alpha = 0.5$ provides perfect secrecy about $\mathbf{T}$, so the secret dither cannot be disclosed, and consequently the embedded message sequence cannot be reliably estimated. However, as mentioned in Remark 4 of Sect. 5.3.1, perfect secrecy about $\mathbf{T}$ does not imply perfect secrecy about the message. In fact, if perfect secrecy about the message were achieved, then the loss should increase linearly in N_o, but we can check in Figure 5.8(a) that it is far from being the case. Finally, Figure 5.8(b) shows the loss for a repetition code with $p = 8$, $n = 10$. Bear in mind that this code provides error-free decoding only if $\alpha \geq 1 - 1/8$. It can be seen that for values of α sufficiently large, the bound given in Lemma 5.5 ($\log(8)/10 \approx 0.21$ nats) is tight. However, for smaller values of α, the loss becomes slightly larger.



(a) $\alpha = 0.52$, $p = 2$

(b) $\alpha = 0.6$, $n = 10$

Figure 5.7: Information leakage per dimension for DC-DM with repetition coding. Impact of the repetition rate (a), and impact of the alphabet size (b).

5.4 Comparison: lattice data hiding vs. Costa

For the lattice data hiding scheme analyzed in Section 5.1, the entropy of the codebook is rather limited due to the codeboook structure and the chosen form of randomization, negatively affecting security. Lattice data hiding schemes are deeply connected with the theoretical construction developed by Costa [83]. However, the codebook in the latter is totally different, since it is random by definition. The main purpose of this brief comparison is to quantify how much can be gained in terms of security by using a codebook with these characteristics. The theoretical security analysis for Costa's scheme is due to Comesaña [68] so its details will not be included here.

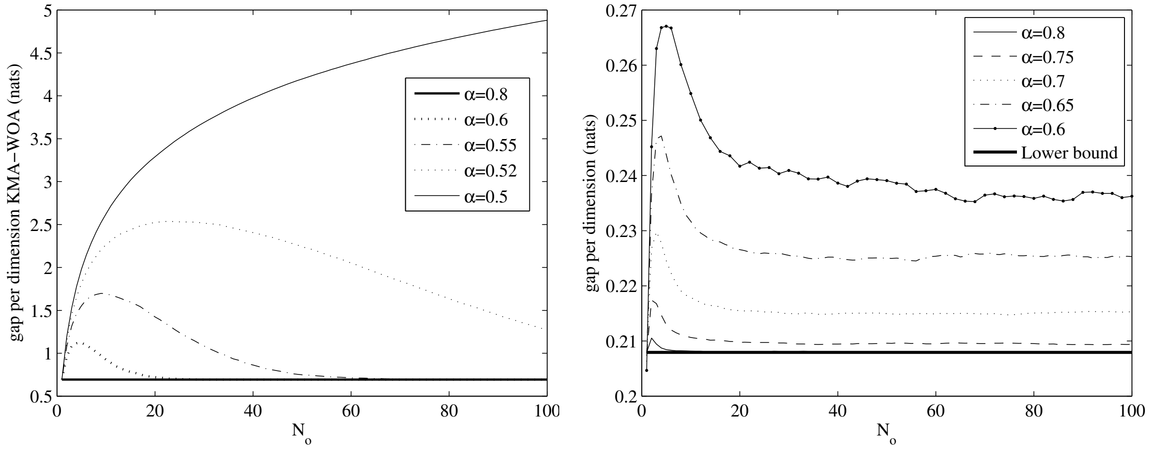
In the following, $\mathcal{U}$ denotes the codebook of Costa's construction. This codebook is randomly generated, so it plays the role of secret parameter of the embedding function to be estimated by the attacker. In Costa's scheme, for the KMA case and $N_o = 1$, it can be shown that

$$\frac{h(\mathcal{U}|\mathbf{Y}, M)}{n} = \frac{h(\mathcal{U})}{n} - \frac{1}{2}\log\left(\frac{D_w + \sigma_X^2}{(1-\alpha)^2\sigma_X^2}\right), \tag{5.72}$$

where σ_X^2 and D_w stand for host and watermark power, respectively, and $h(\mathcal{U})$ denotes the differential entropy of the codebook, given by

$$h(\mathcal{U}) = \frac{n}{2}|\mathcal{U}|\log\left[2\pi e(D_w + \alpha^2\sigma_X^2)\right].$$

Eq. (5.72) depends on the ratio $\lambda \triangleq \sigma_X^2/D_w$ which quantifies the embedding distortion, whereas $|\mathcal{U}|$ depends both on λ and on the ration D_w/σ_N^2, where σ_N^2 is the channel noise.



(a) $p = 2$, $n = 1$

(b) $p = 8$, $n = 10$

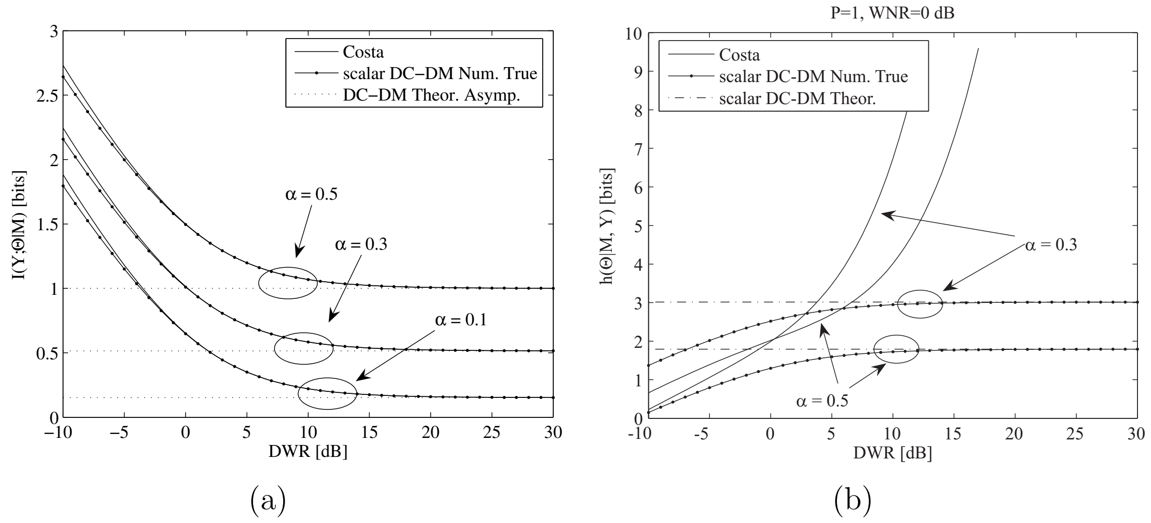
Figure 5.8: Loss per dimension between KMA and WOA for DC-DM with repetition coding.

Interestingly, if we make $\lambda \to \infty$ (which corresponds to a low embedding distortion regime), the information leakage for Costa tends to $-n\log(1-\alpha)$, exactly as for the lattice scheme (see Eq. (5.31)). Actually, the information leakage in lattice data hiding also depends on λ, and in fact it is possible to compute this dependency numerically, by means of numerical integration. In Figure 5.9(a), the information leakage for Costa and scalar DC-DM (i.e., SCS) is shown. It is remarkable the striking similarity in the behavior of both schemes. Furthermore, it can be seen that the asymptotic analysis (DWR→ ∞) performed in this chapter for lattice data hiding is in good agreement with the numerical results for the range of embedding distortions of practical interest.

Nevertheless, when the comparison between Costa and the lattice scheme is made in terms of residual entropy, the similarities disappear (see Figure 5.9(b)): whereas for the lattice scheme the entropy of the codebook is bounded by $\log(\text{vol}(\mathcal{V}(\Lambda)))$, the residual entropy in Costa's scheme is unbounded when $\lambda \to \infty$. The last fact is a consequence of the codebook construction in Costa, where all codewords are mutually independent and its number increases with λ. This constitutes the main advantage, in terms of security, of the random codebook scheme over the lattice scheme that relies solely on dithering. For lattice data hiding, the number of codewords follow a similar dependence with λ, but every codeword just depends on Λ, the corresponding coset representative, and the secret dither.

On the other hand, for the WOA case, and assuming that the watermarker is transmitting information at the maximum reliable rate allowed by the channel, we have for Costa's scheme (with $N_o = 1$)

$$\frac{I(\mathbf{Y}_1;\mathcal{U})}{n} = \frac{I(\mathbf{Y}_1;\mathcal{U}|M)}{n} - \frac{I(\mathbf{Y}_1;M|\mathcal{U})}{n}. \tag{5.73}$$



(a) (b)

Figure 5.9: Comparison of the security provided by Costa's scheme and lattice DC-DM, in terms of mutual information (a), and residual entropy (b) per dimension. WNR $\triangleq \log_{10}(D_w/\sigma_N^2)$.

This result is clearly related to those given in (5.51) and Theorem 5.3 for the lattice data hiding scheme. Here, we can see that the uncertainty about the codebook increases exactly in the same quantity as the reliable transmission rate.

5.5 Conclusions

The main conclusion to be drawn from this chapter is that lattice data hiding schemes relying only on secret dithering are vulnerable to security attacks both in the KMA and CMA scenarios, and also in the WOA scenario if the embedding parameters are not properly chosen.

Other important conclusions are summarized below:

1. The security is largely dependent on the distortion compensation parameter α. Values of α close to 1 make the scheme extremely vulnerable to security attacks. However, in the WOA scenario it is possible to make the scheme highly secure by choosing the appropriate value of α and an appropriate lattice code, as stated in Proposition 5.1. This implies the existence of a trade-off between security and achievable rate in information transmission.

2. It has been shown in Section 5.2.2 that the embedding lattice plays an important role in the security of the lattice data hiding scheme. The security level can be increased by increasing the dimensionality of the embedding lattice and choosing

that lattice with the best mean-squared error quantization properties, although the gain for small n is rather limited. The best security level achievable for lattice data hiding is conjectured to be given by those lattices whose Voronoi cells are the closest (in the normalized second order moment sense) to hyperspheres.[8]

3. When the embedding distortion is sufficiently small (as it is the case in scenarios of practical interest) the information leakage is virtually independent of the DWR, contrarily to spread-spectrum methods.

4. Although it has been shown that it is theoretically possible to achieve perfect secrecy, the security level of many practical scenarios (i.e., simple shaping lattices, low embedding rates, etc.) can be fairly low. Asymptotic values for the equivocation and the variance of the estimation error have been obtained, explaining the fundamental gap between the security of lattice data hiding schemes and spread spectrum methods (without host rejection). The security level of the lattice data hiding schemes has been found to be fairly lower than for spread spectrum methods. The main reason for this gap lies in the host-rejecting nature of the lattice data hiding scheme.

5. One obvious strategy for minimizing security risks is to reuse the secret key as few times as possible, but this may introduce serious synchronization problems. Another strategy is to look for more secure forms of randomization[9] or choosing the embedding parameters that maximize the security. In general, the information leakage about the secret dither can be reduced by increasing the embedding rate or decreasing α, but this solution demands for more powerful error correcting codes (ECC) if one wants to guarantee reliable transmission. A possible drawback, as noted in this chapter, is that the use of ECCs may introduce statistical dependence between different observations that could be exploited by an attacker, particularly for simple ECCs. The complexity of exploiting the information leakage provided by ECCs deserves further study in the future.

6. The comparison given in Section 5.4 shows that the security weaknesses of lattice data hiding are not inherent to side-informed schemes, but they are due to the fact that the randomness of the codeboook in the lattice methods analyzed here relies solely on secret dithering.

[8] The reader interested in a detailed discussion about lattices is referred to the classical text by Conway and Sloane [82].

[9] Note that virtually all implementations of lattice data hiding proposed use the dithering randomization.

Chapter 6

Security of Lattice-Based Data Hiding: Practical algorithms

After the theoretical analysis carried out in Chapter 5, this chapter shows how the information about the secret dither provided by the observations can be extracted and used in practical scenarios, proposing a reversibility attack based on the estimated dither. An estimation algorithm for the KMA and CMA scenarios is proposed. This algorithm is the core of another estimation algorithm proposed for the WOA case. These estimation algorithms work with any arbitrary nested lattice code, and are applicable to high embedding rate scenarios.

The previous works about practical security evaluations of quantization-based methods are focused on ST-DM methods [63],[108]. Contrarily to spread-spectrum, for this kind of methods the watermark depends both on the secret key and the host signal; thus, a simple watermark estimation does not necessarily provide information about the secret key. However, in ST-DM methods, the aim of the attacker is also to disclose a secret subspace, which is precisely where quantization takes place. The estimation tools that have been proposed for attacking ST-DM schemes [61],[46] are PCA and ICA, as for spread spectrum systems. Particularly, the good performance of ICA-based estimators was shown in [46], where a large ensemble of natural images marked with the same secret key are taken as input to the ICA algorithm, which outputs an estimate of the spreading vector. This estimate is used in a subsequent stage for attacking the robustness of the ST-DM scheme, and the results are compared to other attacks that do not exploit the estimate of the spreading vector at hand. More recently, a practical security analysis of trellis-based watermarking methods has been presented in [45]. The attack to the security of these methods is based on the fact that trellis-based embedding creates clusters that identify the codewords (pseudorandom patterns) which are derived from the secret key. An estimator based on the K-means algorithm [159] is proposed, showing good performance in many practical cases. Finally, the estimated codewords are used for changing the embedded message with low distortion.

The estimation algorithms considered in this chapter are devised for the lattice data hiding model described in Chapter 5, so ST-DM and Trellis-based schemes are out of scope. The chapter is structured as follows: Section 6.1 presents the estimation algorithm for the KMA and CMA scenarios. The estimator for the WOA scenario is addressed in Section 6.2. Sections 6.3 and 6.4 show the performance evaluation of the proposed estimators for KMA and WOA, respectively. In the latter section, a "reversibility attack" or "host recovery" (cf. Section 2.2) based on the dither estimate is proposed and implemented. In Section 6.5, the extension of the estimation framework proposed in this chapter to more general scenarios is discussed. The conclusions are summarized in Section 6.6. The reader must be aware that the final purpose of this chapter is not to propose optimal estimation algorithms for all cases, but rather to show that the security weaknesses noticed in Chapter 5 are exploitable in practice with reasonable complexity.

6.1 Dither estimation in the KMA scenario

The theoretical analysis carried out in the previous sections, besides quantifying the information leakage about the secret dither, gives important hints about how to perform dither estimation. Indeed, the information-theoretic formulation given in Section 5.1 is closely related to the theory of "set-membership estimation" (SME), aka "set-theoretic estimation" [96], [67], which is widely known in the field of Automatic Control and in certain Signal Processing areas, such as image recovery.[1] In the set-membership formulation of a problem with solution space Ξ, the ith observation $\mathbf{o}_i$ is associated to a certain subset $\mathcal{F}_i \in \Xi$ that contains all estimates which are "consistent" with that observation. Formally, we say that $\mathbf{z} \in \Xi$ is consistent with the observation $\mathbf{o}_i$ if $\phi_{\mathcal{F}_i}(\mathbf{z}) = 1$, where $\phi_{\mathcal{F}_i}(\mathbf{z})$ denotes the indicator function, and N_o is the number of available observations. The subset $\mathcal{F}$ of estimates which are consistent with all the available information is the so-called "feasible solution set" and it is given by $\mathcal{F} = \bigcap_{i=1}^{N_o} \mathcal{F}_i$; finally, a set-membership estimate consists in choosing any point $\mathbf{z} \in \mathcal{F}$.

In the dither estimation problem, the solution space of interest is $\mathbb{R}^n$. We will deal for now only with the KMA scenario, deferring until Section 6.1.3 the (minor) modifications needed to cope with the CMA case. Thus, the indicator function of interest for KMA is given by

$$\phi_{\mathcal{D}_i}(\mathbf{z}) = \begin{cases} 1, & \mathbf{z} \in \mathcal{D}_i \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

[1] Interestingly, the set-membership framework has been previously applied to watermark embedding in speech signals [126].

so $\mathcal{F}_i = \mathcal{D}_i$ and $\mathcal{F} = \mathcal{S}_{N_o}$, with $\mathcal{D}_i$ and $\mathcal{S}_{No}$ defined in Property 5.1. Moreover, if $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, which is the worst case for the attacker, the set-membership estimator becomes the maximum likelihood dither estimator. Although intuitively simple, such estimator may not be practical, since exact computation of the solution sets may be computationally prohibitive, because of the increasing number of vertices in $\mathcal{S}_{N_o}$ for $N_o > 1$. Nevertheless, the attacker may not be interested in obtaining the exact $\mathcal{S}_{N_o}$, but instead be satisfied with an accurate approximation of the feasible solution set. Algorithms that are suitable for performing such approximation are discussed in this section.

According to Property 5.2, the assumption $\alpha \geq 0.5$ allows us to consider the feasible region as a modulo-Λ convex set. Furthermore, if we shift all observations by $-\tilde{\mathbf{y}}_1+\mathbf{d}_{m_1}$, then the modulo operation is transparent, so the feasible regions for each observation (Eq. (5.21)) can be now simplified to[2]

$$\mathcal{D}_i = \tilde{\mathbf{v}}_i + (1-\alpha)\mathcal{V}(\Lambda),\ i = 1, \ldots, N_o, \tag{6.2}$$

with $\tilde{\mathbf{v}}_i$ defined in (5.44), rendering the problem convex, since the feasible solution sets (which are in fact polytopes) result from the intersection of convex sets. Some guidelines about how to modify the algorithms in order to work with $\alpha < 0.5$ will be given in Section 6.5.

The Voronoi region of any lattice can be described in a variety of ways; for our purposes the most appropriate description is by means of the bounding hyperplanes corresponding to its facets. In the following we assume that, for a Voronoi cell $\mathcal{V}(\Lambda)$ with n_f facets, we know:

1. a vector $\boldsymbol{\phi}_k$ which is outward-pointing normal to the k-th facet;
2. a point $\mathbf{z}_{0,k}$ on the k-th facet.

Taking into account each of the modified observations $\tilde{\mathbf{v}}_i$, we have

$$\mathcal{D}_i = \{\mathbf{z} \in \mathbb{R}^n : \boldsymbol{\phi}_k^T(\mathbf{z} - \mathbf{z}_{0,k}) \leq \boldsymbol{\phi}_k^T\tilde{\mathbf{v}}_i,\ k = 1, \ldots, n_f;\ i = 1, \ldots, N_o\}. \tag{6.3}$$

6.1.1 Inner polytope algorithm

The set of modified observations $\{\tilde{\mathbf{v}}_i\}$ together with Eq. (6.3) define an ensemble of linear inequalities, which in turn describe a polytope in n-dimensional space [92]. Hence, the feasible solution set can be expressed as

$$\mathcal{S}_{N_o} = \left\{\mathbf{z} \in \mathbb{R}^n : \boldsymbol{\phi}_k^T\mathbf{z} \leq \boldsymbol{\phi}_k^T\tilde{\mathbf{v}}_i + \boldsymbol{\phi}_k^T\mathbf{z}_{0,k},\ k = 1, \ldots, n_f;\ i = 1, \ldots, N_o\right\}. \tag{6.4}$$

[2] Obviously, the offset $-\tilde{\mathbf{y}}_1 - \mathbf{d}_{m_1}$ must be removed from the final estimate.

We are interested in computing an approximation of the feasible region. For such an approximation to be valid, it must outer bound $\mathcal{S}_{N_o}$ (as tightly as possible), since we do not want to discard any point in $\mathcal{S}_{N_o}$ a priori, and it is also desirable that the approximate region is easy to describe. Then, a reasonable choice is to search for the ellipsoid of minimum volume that contains $\mathcal{S}_{N_o}$ (formally known as the *Löwner-John* ellipsoid of $\mathcal{S}_{N_o}$[55]). Unfortunately, the problem of finding the ellipsoid of interest is ill-posed (indeed, it has been shown to be an NP-complete problem) [178], but on the other hand, the problem of finding the maximum volume ellipsoid contained in the polytope defined by a set of linear inequalities is well-posed. Moreover, if we scale such ellipsoid by a factor of n around its center (n is the dimensionality of the lattice), then the resulting ellipsoid is guaranteed to bound $\mathcal{S}_{N_o}$ [55]. An ellipsoid $\mathcal{E}(\boldsymbol{\theta}, \mathbf{P})$ in Euclidean space is defined by its center $\boldsymbol{\theta}$ and a symmetric positive definite matrix $\mathbf{P}$ such that

$$\mathcal{E}(\boldsymbol{\theta}, \mathbf{P}) = \left\{\mathbf{z} \in \mathbb{R}^n : |(\mathbf{z}-\boldsymbol{\theta})^T \mathbf{P}^{-1}(\mathbf{z}-\boldsymbol{\theta})| \leq 1\right\} = \left\{\mathbf{P}^{1/2}\mathbf{r} + \boldsymbol{\theta} : ||\mathbf{r}|| \leq 1\right\}. \quad (6.5)$$

The computation of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{P}}$ for the maximum volume ellipsoid contained in $\mathcal{S}_{N_o}$ can be written as a convex minimization problem with second order cone constraints [55]:

$$\begin{aligned} (\hat{\boldsymbol{\theta}}, \hat{\mathbf{P}}) = \arg\min_{\boldsymbol{\theta}, \mathbf{P}} \quad & \log\det(\mathbf{P}^{-1/2}) \\ \text{subject to} \quad & ||\mathbf{P}^{1/2}\boldsymbol{\phi}_k|| \leq \boldsymbol{\phi}_k^T \tilde{\mathbf{v}}_i + \boldsymbol{\phi}_k^T \mathbf{z}_{0,i} - \boldsymbol{\phi}_k^T \boldsymbol{\theta}, \\ & \forall\, k = 1, \ldots, n_f;\ i = 1, \ldots, N_o. \end{aligned} \quad (6.6)$$

This problem can be recast as a “semidefinite problem” [215] where a linear function is minimized subject to Linear Matrix Inequality (LMI) constraints; this kind of optimization problems can be efficiently solved by means of interior-point methods [178]. As will be checked in Section 6.4, this approach yields tight approximations to $\mathcal{S}_{N_o}$, but it presents an obvious drawback: the potential complexity of the minimization problem arising from the huge number of constraints imposed by large n and N_o. The scheme presented in the next section reduces the complexity by means of an iterative approach.

6.1.2 Optimal Volume Ellipsoid (OVE) [64]

This is a classical SME algorithm that was originally devised for estimation in noisy AR models:

$$y_k = \sum_{j=1}^{n} \theta_j y_{k-j} + u_k = \boldsymbol{\theta}^T \boldsymbol{\phi}_k + u_k,$$

where $\boldsymbol{\phi}_k = (y_{k-1}, \ldots, y_{k-n})^T$ are the n past observations, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ is the vector of parameters to be estimated, and u_k is the noise term, whose absolute value is assumed to be bounded by γ_k. For the k-th observation, the feasible solution set $\mathcal{F}_k$ is given by all points in $\mathbb{R}^n$ that are "consistent" with the observation, i.e.

$$\mathcal{F}_k = \{\mathbf{z} \in \mathbb{R}^n : |y_k - \mathbf{z}^T \boldsymbol{\phi}_k| \leq \gamma_k\}. \tag{6.7}$$

Equation (6.7) defines a region of $\mathbb{R}^n$ delimited by two parallel hyperplanes:

$$H_{k,1} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T \boldsymbol{\phi}_k = y_k - \gamma_k\}, \;\; H_{k,2} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T \boldsymbol{\phi}_k = y_k + \gamma_k\},$$

which encloses the true parameter vector $\boldsymbol{\theta}$. The series of solution sets is then constructed iteratively as $\mathcal{S}_k = \bigcap_{i=1}^{k} \mathcal{F}_i,\; k = 1, \ldots, N_o$. In order to avoid the costly computation of the exact $\{\mathcal{S}_k\}$, the solution sets are approximately described by means of bounding ellipsoids.

This algorithm can be straightforwardly applied to our problem by slightly modifying the description of the feasible region given in (6.3): in our case, we need to parameterize $\mathcal{D}_i$ as the intersection of a finite number of parallel hyperplanes. Assuming that the Voronoi cell of the considered lattice is composed of n_f pairwise parallel facets (see Figure 6.1(a)),[3] the feasible solution set for the i-th observation can be specified by a matrix $\boldsymbol{\Phi}_{n \times n_f/2}$, and a vector $\boldsymbol{\gamma}_{n_f/2 \times 1}$ such that $\mathcal{D}_i = \bigcap_{j=1}^{n_f/2} \mathcal{F}_{i,j}$, where

$$\mathcal{F}_{i,j} = \{\mathbf{z} \in \mathbb{R}^n : |\tilde{\mathbf{v}}_i^T \boldsymbol{\phi}_j - \mathbf{z}^T \boldsymbol{\phi}_j| \leq \gamma_j\}, \tag{6.8}$$

being $\boldsymbol{\phi}_j$ the j-th column of $\boldsymbol{\Phi}$, and $\gamma_j \triangleq \boldsymbol{\phi}_j^T \mathbf{z}_{0,k}$ is the j-th element of $\boldsymbol{\gamma}$. Hence, the series of solution sets is given by

$$\mathcal{S}_k = \bigcap_{i=1}^{k} \mathcal{D}_i = \bigcap_{i=1}^{k} \bigcap_{j=1}^{n_f/2} \mathcal{F}_{i,j}, \;\; k = 1, \ldots, N_o. \tag{6.9}$$

The computation of the $(k+1)$-th solution set amounts to obtaining an ellipsoid $\mathcal{E}(\hat{\boldsymbol{\theta}}_{k+1}, \hat{\mathbf{P}}_{k+1}) \supseteq \mathcal{E}(\hat{\boldsymbol{\theta}}_k, \hat{\mathbf{P}}_k) \cap \mathcal{D}_k$. Such ellipsoid is iteratively computed following Algorithm 6.1. This way, in Step 2 of Algorithm 6.1 we are intersecting iteratively one ellipsoid with one set $\mathcal{F}_{k,i}$, as is depicted in Figure 6.1(b). Clearly, we are interested in finding the ellipsoid with minimum volume that contains such intersection, i.e.

$$\begin{aligned} (\mathbf{c}_{i+1}^*, \mathbf{B}_{i+1}^*) = \arg\min_{\mathbf{c},\mathbf{B}} \quad & \mathrm{vol}(\mathcal{E}(\mathbf{c}, \mathbf{B})) \\ \text{subject to} \quad & \mathcal{E}(\mathbf{c}_i, \mathbf{B}_i) \cap \mathcal{F}_{k,i+1} \subseteq \mathcal{E}(\mathbf{c}, \mathbf{B}). \end{aligned} \tag{6.10}$$

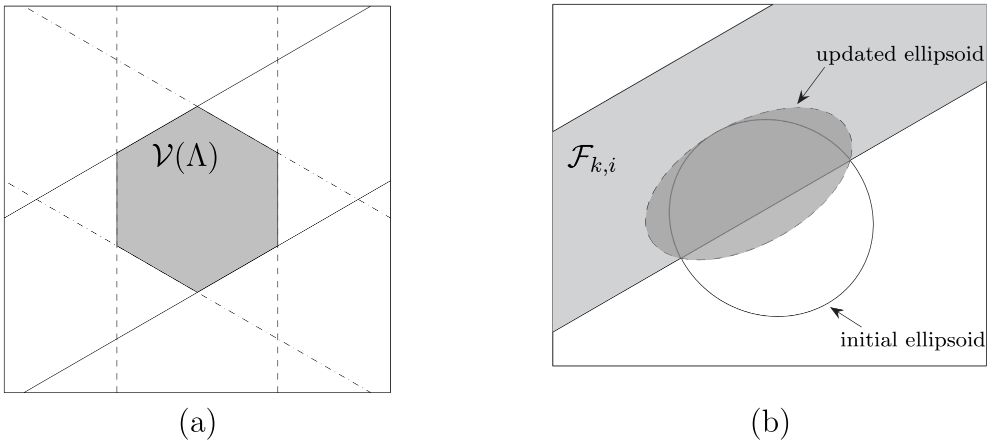
which is precisely the minimization problem addressed in the OVE algorithm [64], whose analytic solution reads as

$$\mathbf{c}_{i+1}^* = \mathbf{c}_i + \frac{\tau_i \mathbf{B}_i \boldsymbol{\phi}_i}{\left(\boldsymbol{\phi}_i^T \mathbf{B}_i \boldsymbol{\phi}_i\right)^{1/2}}, \qquad \mathbf{B}_{i+1}^* = \delta_i \left(\mathbf{B}_i - \sigma_i \frac{\mathbf{B}_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T \mathbf{B}_i}{\boldsymbol{\phi}_i^T \mathbf{B}_i \boldsymbol{\phi}_i}\right), \tag{6.11}$$

[3] Should this not be true, the problem can still be recast in a similar manner by adding some additional hyperplanes.

Algorithm 6.1 Iterative computation of the bounding ellipsoid

```
1. Initialization: $\mathcal{E}(\mathbf{c}_0, \mathbf{B}_0) = \mathcal{E}(\hat{\boldsymbol{\theta}}_k, \hat{\mathbf{P}}_k)$
2. Compute $\mathcal{E}(\mathbf{c}_{i+1}, \mathbf{B}_{i+1}) \supseteq \mathcal{E}(\mathbf{c}_i, \mathbf{B}_i) \cap \mathcal{F}_{k,i+1},\ i = 0, \ldots, n_f/2 - 1$
3. Finally, make $\mathcal{E}(\hat{\boldsymbol{\theta}}_{k+1}, \hat{\mathbf{P}}_{k+1}) = \mathcal{E}(\mathbf{c}_{n_f/2}, \mathbf{B}_{n_f/2})$
```



(a) (b)

Figure 6.1: (a) Voronoi region of the hexagonal lattice delimited by three pairs of parallel hyperplanes. (b) Intersection between an ellipsoid and a pair of hyperplanes.

where τ_i, σ_i, δ_i are variables that depend on the observation $\tilde{\mathbf{v}}_k$, the current ellipsoid $\mathcal{E}(\mathbf{c}_i, \mathbf{B}_i)$ and $\mathcal{F}_{k,i+1}$ (details about their calculation can be found in [64]), and finally $\boldsymbol{\phi}_i$ is the i-th column of matrix $\mathbf{\Phi}$. Figure 6.2 illustrates the series of bounding ellipsoids obtained in 2-dimensional problem with a hexagonal shaping lattice.

The algorithm just described is obviously optimal in one dimension, since the ellipsoids are simply real intervals. Another interesting feature of this approach, and common to many other iterative SME algorithms, is that further refinements on the solution set are possible by recirculating the observed data, i.e., by feeding to the system the same set of observations repeatedly (as if they were stored in a circular buffer). This is possible because the resulting bounding ellipsoid in the ith iteration depends on both the $(i-1)$th bounding ellipsoid and the ith observation. This important feature provides performance similar to that of the above "inner polytope" algorithm, as will be checked in Section 6.4.

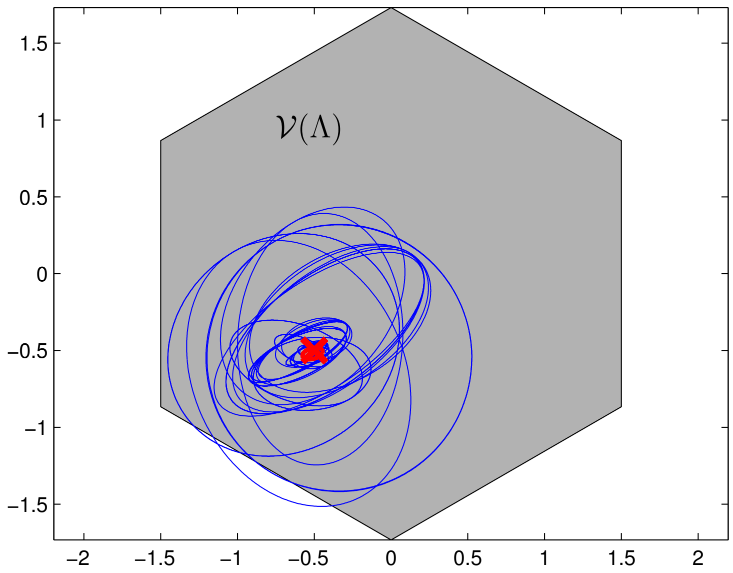


Figure 6.2: Series of bounding ellipsoids, with the true dither vector indicated by a cross.

6.1.3 Dither estimation in the CMA scenario

The CMA scenario implies minor changes to the estimation algorithms proposed above for the KMA case. An estimator can be implemented following the steps detailed in Algorithm 6.2.

6.2 A practical dither estimator for the WOA scenario

A practical dither estimator for the WOA scenario is proposed in this section. The core of the estimation procedure is the estimator devised in Sect. 6.1. Throughout this section, the secret dither will be again assumed to be uniformly distributed in $\mathcal{V}(\Lambda)$.

6.2.1 Joint Bayesian and set-membership estimation for the WOA scenario

The ML estimate of the secret dither in the WOA scenario can be expressed as

$$\begin{aligned}\hat{\mathbf{t}}_{ML} &= \arg\max_{\mathbf{t}\in\mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o}|\mathbf{T}=\mathbf{t}) \\ &= \arg\max_{\mathbf{t}\in\mathcal{V}(\Lambda)} \sum_{k=1}^{p^{N_o}} f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o},\mathbf{m}^{(k)}|\mathbf{T}=\mathbf{t}). \qquad (6.12)\end{aligned}$$

Algorithm 6.2 Dither estimation in the CMA scenario

1. Assume that the sequence of observations is marked with message $m \in \mathcal{M}$.
2. Perform estimation as in the KMA scenario.
3. Once $\hat{\mathcal{S}}_{N_o}$ has been obtained, compute the approximate feasible region $\hat{\mathcal{S}}_{N_o}^{CMA}$ as in Eq. (5.50).
4. Provided that $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, two possible cases may arise after performing Step 3:
    - The resulting feasible regions $(\hat{\mathcal{S}}_{N_o} + \mathbf{d}_m)$ overlap; then, according to Eq. (5.49), the probability of finding the dither in their intersection is higher than in the remaining regions.
    - The regions do not overlap; then, the dither is equally likely in any of the feasible regions.

Estimation based on Eq. (6.12) is impractical due to the number of summation terms, which is exponentially increasing with the number of observations. In order to keep an affordable complexity for the estimation algorithm we resort to the usual "Viterbi approximation", where the summation in (6.12) is approximated by the value of the maximum term. This way, approximate ML estimation of the dither amounts to estimating the most probable message sequence, and then performing dither estimation as in Sect. 6.1 using the estimate of the embedded message. Mathematically, it can be written as

$$\hat{\mathbf{t}} = \arg \max_{\mathbf{t} \in \mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, \mathbf{m}^{(\hat{k})} | \mathbf{t}), \text{ with } \hat{k} = \arg \max_{k=1,\ldots,p^{N_o}} f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} | \mathbf{m}^{(k)}). \tag{6.13}$$

At this point we recall that the a priori path space $\mathcal{M}^{N_o}$ is divided in equivalence classes (with p elements each) defined by the relation (5.61). Since the paths belonging to the same equivalence class have the same a posteriori probability (according to Lemma 5.4), the path $\mathbf{m}^{(\hat{k})}$ in (6.13) is not unique. In order to get rid of this ambiguity, we will reduce the search space to one representative per equivalence class. At the same time, this strategy reduces the cardinality of the search space by a factor p. Notice that this complexity reduction does not imply any loss in performance, since the whole set of feasible paths can be recovered from the set of equivalence classes.

It is important to clarify that the two-stage estimator just defined is suboptimal, in general, since its performance is subject to the correct estimate of the embedded message. However, if α fulfills the condition imposed in Theorem 5.3, then no loss of optimality is incurred for large N_o. The reason is that for sufficiently large N_o there exist only p equiprobable feasible message sequences (which belong to the same equivalence class), as explained in the proof of the theorem. Since we are restricting the search to one representative per equivalence class, it is clear that the summation (6.12) will equal the value of the maximum term.

Hereinafter, we will use the term "path" for denoting each message sequence $\mathbf{m}^{(k)} = [m_1^k, \ldots, m_{N_o}^k]$, $k = 1, \ldots, p^{N_o}$. From Appendix C.6, we know that the a posteriori probability of the observations given $\mathbf{m}^{(k)}$ is given by

$$f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}|\mathbf{m}^{(k)}) = \frac{\mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(k)}))}{(\mathrm{vol}(\mathcal{Z}(\Lambda)))^{N_o} \cdot \mathrm{vol}(\mathcal{V}(\Lambda))}. \tag{6.14}$$

The only term of (6.14) that depends on the hypothesized path is $\mathcal{S}_{N_o}(\mathbf{m}^{(k)})$. This implies that, in practical terms, the most probable paths are those with the largest feasible region. Hence, we can define the "score" of the path $\mathbf{m}^{(k)}$ as

$$\lambda(\mathbf{m}^{(k)}) \triangleq \mathrm{vol}(\mathcal{S}_{N_o}(\mathbf{m}^{(k)})). \tag{6.15}$$

which can be used to compare the probabilities of different paths as long as they have the same length. It follows that, given N_o observations, the ML estimate of the most probable path is simply given by $\mathbf{m}^{(\hat{k})}$, where $\hat{k} = \arg\max_k \lambda(\mathbf{m}^{(k)})$. The search for the most probable path can be carried out by means of a tree search where each branch of the tree represents a hypothesized path with an associated score. For saving computational resources, the outer bound introduced in Section 6.1.1 and a "beam search" strategy [136, Chapt. 12] will be applied during the tree search. The steps of the proposed dither estimation algorithm are summarized in Algorithm 6.3 (page 144). The input data are the observations $\{\tilde{\mathbf{y}}_i,\ i = 1 \ldots, N_o\}$ and the parameters of the nested lattice code.

Remark 6.1. Through the variation of the "beam factor" β, defined in Algorithm 6.3, one can control the tradeoff between computational complexity and accuracy. If $\beta = 1$, only the most probable path is retained in each iteration of Algorithm 6.3. Hence, complexity reaches its minimum for $\beta = 1$, but the probability of missing the correct path may be very high. As β is increased, the number of surviving paths per iteration increases, in general. In the case $\beta \to \infty$, all the paths are retained in each iteration, making the complexity unaffordable, in general, but reducing to 0 the probability of missing the correct path. The impact of varying β is shown in Sect. 6.4.1. Notice that β does not limit the absolute number of paths to be considered in each iteration. This is why the parameter K_{max} is also introduced in Step 2.c of Algorithm 6.3, for limiting the complexity in absolute terms.

Remark 6.2. It is possible that Step 2.c of Algorithm 6.3 results in $K_i = 0$. If this is the case, then it means that the true path has been discarded at some previous iteration of the algorithm due to too restrictive beam factors β and/or K_{max}. If this happens, Algorithm 6.3 must be restarted after increasing the values of the beam factors.

Remark 6.3. The outer bounding of the feasible regions may impact negatively the estimator performance, due to the introduction of spurious paths and variation of the scores.

Remark 6.4. When the shaping lattice $\mathcal{V}(\Lambda)$ is cubic, the feasible regions are hyperrectangles. In such case, they can be easily computed componentwise, since they are simply defined by n real segments. Thus, there is no need to apply the inner polytope algorithm when the shaping lattice is cubic.

The computation of all the outer bounding ellipsoids in Step 2.b of Algorithm 6.3 is the most time-consuming task of the estimation algorithm. Clearly, this step can be sped up if we can use a fast algorithm to discard the "unfeasible" paths. Formally, a certain path $\mathbf{m}^{(k)},\ k = 1, \dots, p^{N_o}$, is said to be "unfeasible" or "inconsistent" with the observations if the associated feasible region $\mathcal{S}_{N_o}(\mathbf{m}^{(k)})$ is an empty set; otherwise, the path is said to be "feasible" or "consistent". That is, the unfeasible paths are those that yield a null score (i.e. null a posteriori probability), so it is not worth keeping them for the next iteration. Thus, Step 2.b of Algorithm 6.3 is broken down in two steps: Step 2.b.i, that checks the feasibility of the candidate paths, and Step 2.b.ii, which is the same as the original 2.b of Algorithm 6.3, but computing only the feasible region of the feasible paths.

For the $i-1$ first observations, we have the pairs $\{\mathbf{m}^{(k)}, \mathcal{E}_{i-1}(\mathbf{m}^{(k)}),\ k = 1, \dots, K_{i-1}\}$. In order to check the feasibility of a certain candidate path $\mathbf{m}^{(k,l)},\ k = 1, \dots, K_{i-1},\ l = 0, \dots, p-1$, we need to check whether the intersection $\mathcal{E}_{i-1}(\mathbf{m}^{(k)}) \bigcap \mathcal{D}_i(l)$ is empty or not. To this end, we have used an algorithm based on the OVE algorithm proposed in [64], which is described in Algorithm 6.4 in page 145.

6.3 Experimental results for the KMA scenario

This section provides a comparison of the practical performance for the different estimators proposed in Section 6.1, considering only the KMA scenario. The optimization problems involving LMIs were solved using the optimization packages YALMIP [158] and SeDuMi [207] for Matlab®, and the set of observations $\{\tilde{\mathbf{y}}_i\}$ was generated according to the distribution given in (5.19). As for the theoretical part, we will consider here some of the so-called "root lattices" and their duals, introduced in Section 5.2.2. The Voronoi regions of these lattices are described in [81], from which we derived all the parameters needed for implementing our attack. For illustration purposes, some examples are given in the next subsection.

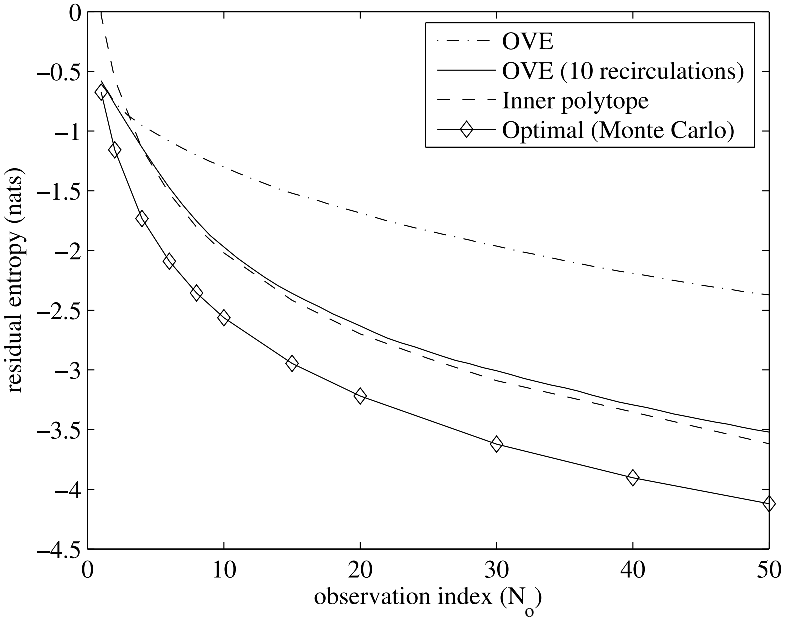


Figure 6.3: Performance comparison for the hexagonal lattice (KMA, $\alpha = 0.5$).

6.3.1 Parameters of the estimation algorithms

We will consider here some of the so-called "root lattices" and their duals, introduced in Section 5.2.2. The first step is to obtain the equations of the hyperplanes that define their Voronoi regions, which are described in [81]. This is simple for the root lattices, since their facets are the hyperplanes bisecting the vectors that join the origin to the lattice points of minimal norm, i.e. the nearest neighbors of $\mathbf{0}$. This procedure is illustrated in Example 6.1.

The Voronoi region of the dual of the root lattices cannot be obtained so easily. For instance, lattices D_n^* can be written as the union of two cosets of a cubic lattice, and as such their Voronoi region is given by the intersection between a hypercube and a generalized octahedron, both in n-dimensional space. Example 6.2 considers the particular case of the lattice D_4^* for illustration purposes.

6.3.2 Results

We provide two different measures of performance of the proposed estimators:

1. The first one is based on the volume of the estimated feasible regions. The volume

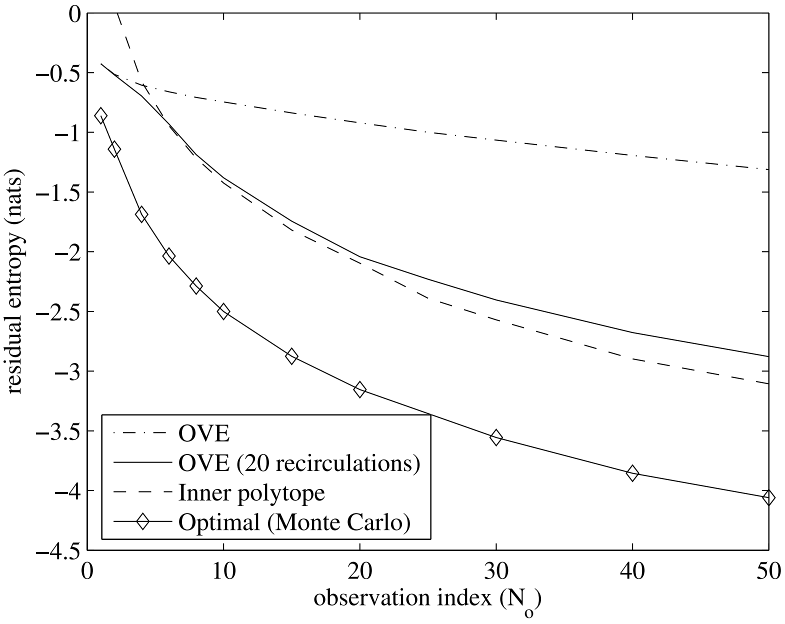


Figure 6.4: Performance comparison for the lattice D_4^* (KMA, $\alpha = 0.5$).

of the kth ellipsoid reads as

$$\mathrm{vol}(\mathcal{E}(\hat{\boldsymbol{\theta}}_k, \hat{\mathbf{P}}_k)) = (\det \hat{\mathbf{P}}_k)^{1/2} \cdot V_n(1), \tag{6.22}$$

where $V_n(1)$ stands for the volume of the n-dimensional sphere of unit radius. When $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, all points in the interior of the estimated feasible region $\hat{\mathcal{S}}_{N_o}$ have the same probability of being the true dither vector $\mathbf{t}_0$, so it is immediate to estimate the residual entropy of the dither as $\log(\mathrm{vol}(\hat{\mathcal{S}}_{N_o}))$. The average value of this "empirical" residual entropy is computed over a large number of realizations. The performance of each method is quantified by the gap between this measure and the theoretical result of Section 5.2.2.

2. The second measure of performance is the MSE per dimension, i.e. $\frac{1}{n}||\mathbf{t} - \hat{\mathbf{t}}||^2$, where $\hat{\mathbf{t}}$ has been taken as the center of the resulting ellipsoid. Note that, as long as this center is close to the center of masses of $\mathcal{S}_{N_o}$, the resulting estimator will be close to the MMSE estimator (i.e., the conditional mean estimator). Again, the plots represent this squared error averaged over a large number of observations.

In the experiments, the embedding distortion was fixed to $D_w = \alpha^2/12$, with $\alpha = 0.5$. Figures 6.3, 6.4 and 6.5 show the performance (in terms of the equivocation) of the different estimators when the embedding lattices are the hexagonal, D_4^* and E_8 [82], respectively. Although the inner polytope algorithm provides the best performance, it

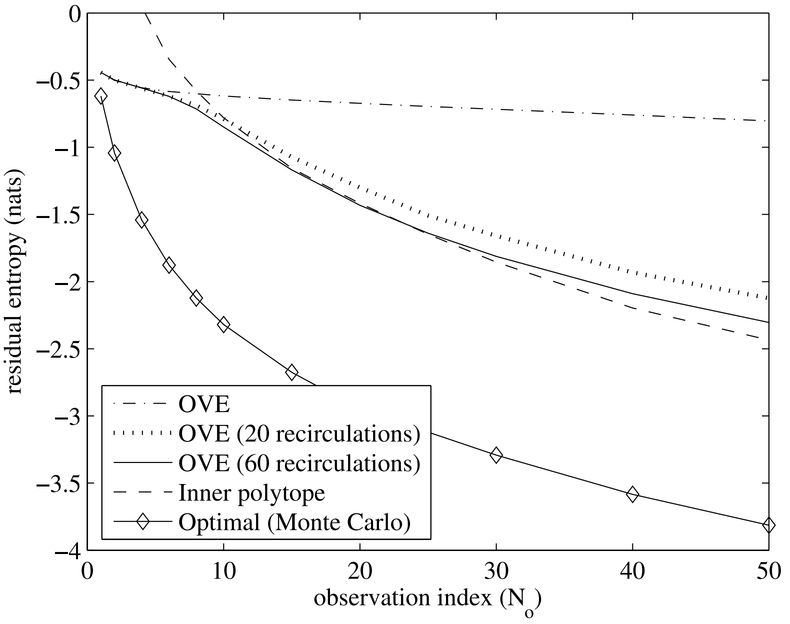
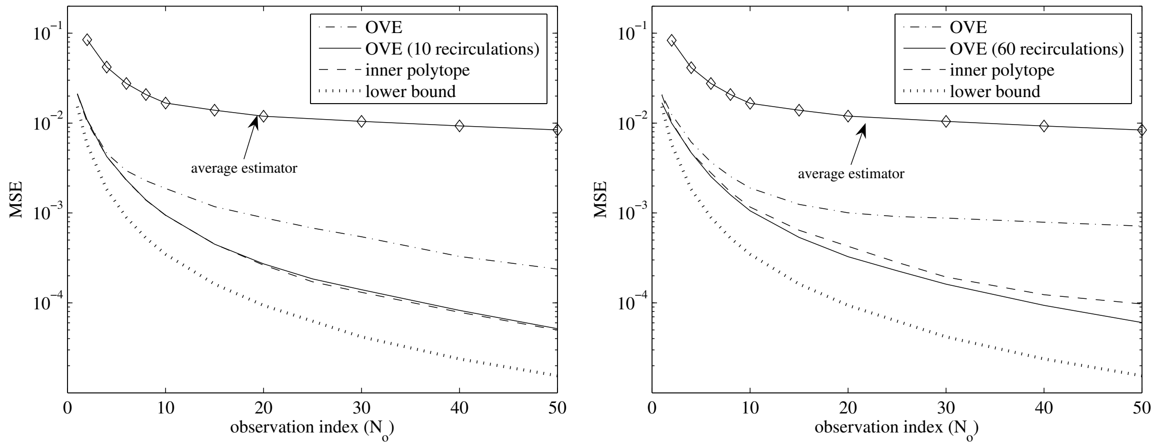


Figure 6.5: Performance comparison for the Gosset lattice (E_8) (KMA, $\alpha = 0.5$).

can be observed that the property of recirculation allows to compensate for the loss of optimality of the OVE algorithm. The performance gain is remarkable for the first recirculations, but marginal above a certain number, as can be seen in Figure 6.5. Notice also that the number of recirculations must be increased with n in order to match the performance of the inner polytope algorithm. Figure 6.4 shows the results obtained with the lattice D_4^*. Finally, the plots in Figure 6.6 show the empirical mean squared error per dimension obtained with each method. The lower bound given by Eq. (2.10) is plotted for comparison, showing the good performance of both methods. Interestingly, the OVE algorithm seems to perform better than the inner polytope in terms of mean squared error. The performance of the averaging estimator is also plotted for reference; such estimator is optimal for $n \to \infty$ and Λ_n^*, as discussed in Section 5.2.4, but for small n it is clearly far from being so.

6.3.3 Complexity issues

One can find in the literature of set-membership estimation approaches that offer better performance than the ellipsoidal approximations, by computing the exact solution sets [67],[220]. Nevertheless, they may be very computationally demanding in large-scale problems. Instead, the algorithms considered in this chapter have proved to be efficient in giving approximate solutions for several hundreds of observations. For the optimization problem in (6.6), it has been shown that the number of itera-



(a) (b)

Figure 6.6: Mean squared error per dimension of the dither estimate, for the hexagonal lattice (a) and Gosset lattice E_8 (b), for KMA and $\alpha = 0.5$.

tions needed to solve the problem (by means of interior-point methods) does not grow faster than a polynomial of the problem size [215].[5] Most of the computational cost of each iteration lies in the least-squares problem (of the same size as the original problem) that must be solved, whose number of iterations is again polynomial with the problem size. However, in practice it is possible to exploit the problem structure (sparsity, for instance) so as to reduce complexity: in our case, for example, there is a potentially large number of redundant constraints that can be removed for alleviating the computational burden. For high-dimensional lattices it is also possible to simplify the problem description (albeit resulting in looser estimates) by approximating the considered Voronoi region by another simpler polytope that bounds $\mathcal{V}(\Lambda)$.

For the OVE algorithm, the number of arithmetic operations (scalar sums and products) carried out in each iteration is $O(n^2)$. Also, in the OVE algorithm we perform exactly $N_o \cdot \frac{n_f}{2} \cdot n_r$ iterations, where N_o is the number of observations, n_f is the number of facets of the Voronoi cell (equivalently, the number of linear inequalities specifying the problem), and n_r is the number of recirculations of the data. The term n_f will largely depend on the considered lattice, in general, and n_r will be determined by the required accuracy, giving a degree of freedom to the attacker. Finally, it is interesting to note that OVE-like algorithms automatically get rid of redundant constraints, using only those pairs of hyperplanes that produce an update on the solution set.

[5] The size of an optimization problem is commonly understood as the dimensionality of a vector whose components are the coefficients of the analytical expressions for the constraints and the objective variables.

6.4 Experimental results for the WOA scenario

This section presents the results of applying Algorithm 6.3 over some practical lattice data hiding schemes in the WOA scenario. The experiments have been carried out under the following assumptions: the host signals follow an i.i.d. Gaussian distribution with zero mean and variance $\sigma_X^2 = 10$, the DWR is 30 dB in all cases, and the embedded messages are equiprobable and independent. In order to assess the performance of the dither estimator without ambiguities (due to the result of Lemma 5.4), it is assumed that the message conveyed by the first observation corresponds to the symbol 0. The parameter K_{max} of Algorithm 6.3 has been set to 250 in all cases.

6.4.1 Tradeoff complexity-accuracy

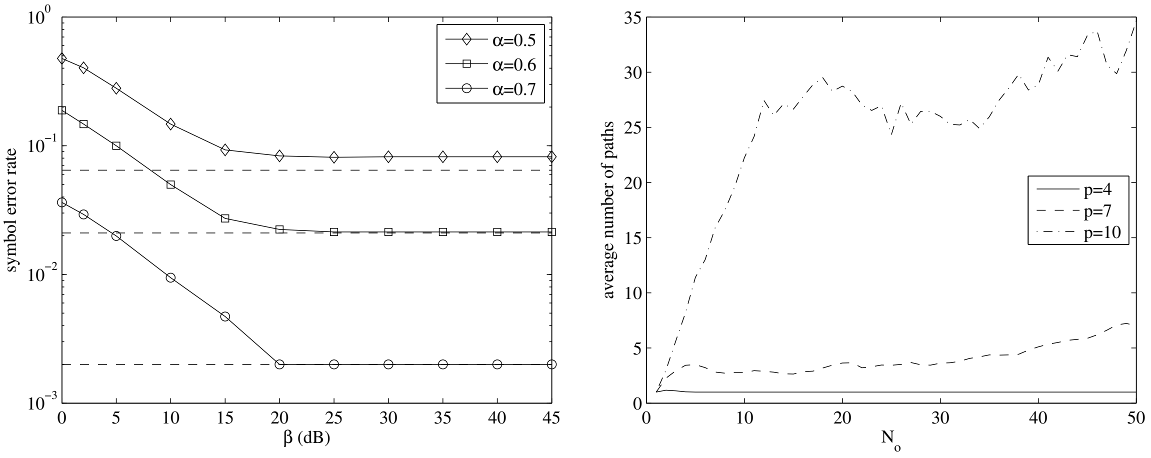
One interesting performance measure is the resulting probability of decoding error when the decoder uses the dither estimate, instead of the true dither vector. If this measure is represented in terms of the "beam factor" (β) of Algorithm 6.3, then the tradeoff between complexity and accuracy becomes patent. This tradeoff is illustrated in Figure 6.7(a) for a cubic shaping lattice and repetition coding (see Section 5.3.3) with $n = 10$ and $p = 6$. Using dither estimates obtained with $N_o = 100$ observations, the numerically computed symbol error rate (SER) is shown in Figure 6.7(a) for different values of α. For reference, Figure 6.7(a) also shows in dashed lines the SER obtained by a fair decoder, i.e. knowing the true dither signal. As can be seen, the SER is always decreased as β is increased, achieving the same decoding performance as the fair decoder in the cases of $\alpha = 0.6$ and $\alpha = 0.7$. However, for $\alpha = 0.5$, the SER of the unfair decoder is slightly larger. The reason is that in this case the probability of deciding a wrong dither is not negligible even for high β.

The average number of surviving paths in the tree search is plotted in Figure 6.7(b) for illustrating the complexity of the search procedure. In this regard, it can be seen that even in a difficult case as $p = 10$ with $\alpha = 0.6$, the tree search can still be performed with low complexity.

6.4.2 Estimation error

Now we measure the performance of the estimator in terms of the mean squared error (MSE) per dimension between the dither estimate and the actual dither vector. Three different shaping lattices have been considered. In all cases, a beam factor $\beta = 45$ dB has been used.

Figure 6.8 shows the results obtained for a scheme using a cubic shaping lattice in 10 dimensions and repetition coding with $\alpha = 0.6$. It can be seen that for $p = 4$ it is still possible to attain the same accuracy as in the KMA scenario, whereas for $p = 7$



(a)

(b)

Figure 6.7: Estimation results for a cubic lattice with repetition coding, with $n = 10$ and DWR $= 30$ dB. Figure (a) shows the symbol error rate for a dither estimated with $N_o = 100$ observations versus the beam factor β in dB, for $p = 6$. Figure (b) shows the average number of surviving paths in the tree search (for $\beta = 45$) with $\alpha = 0.6$ and different embedding rates.

and $p = 10$ a significant degradation of the MSE is observed. This degradation is a consequence of the fact that, as p is increased, the probability of correctly retrieving the embedded path decreases when α is kept constant (even knowing the true value of the dither).[6] In the experiments, the probability of choosing an incorrect path has been found to be around 0.05 and 0.1 for $p = 7$ and $p = 10$, respectively.

Figure 6.9(a) shows the results obtained for a hexagonal shaping lattice and $\alpha = 0.7$. Notice that, although α is higher than in the former case, the maximum embedding rate considered now is substantially larger ($\frac{1}{2}\log_2(9)$ bits vs. $\frac{1}{10}\log_2(10)$ bits). Similarly as above, we can see that increasing p degrades the MSE: for $p = 4$ it is still possible to achieve the same accuracy as in the KMA scenario, but for $p = 4$ the MSE is increased, and for $p = 9$ the MSE is not reduced with N_o. Finally, Figure 6.9(b) shows the results obtained for the E_8 shaping lattice [82], the best lattice quantizer in 8 dimensions. It can be seen that in this case, even with large alphabets (e.g. $p = 12$), the estimator achieves its optimal performance. For obtaining the results of Figure 6.9, the lattice codes have been obtained by Construction A, in the estimation procedure we have resorted to the inner polytope algorithm in order to compute the approximate feasible regions.

[6] Recall that, due to the Viterbi approximation, the performance of the estimator is degraded when the embedded message cannot be correctly decoded, as explained in 6.2.1.

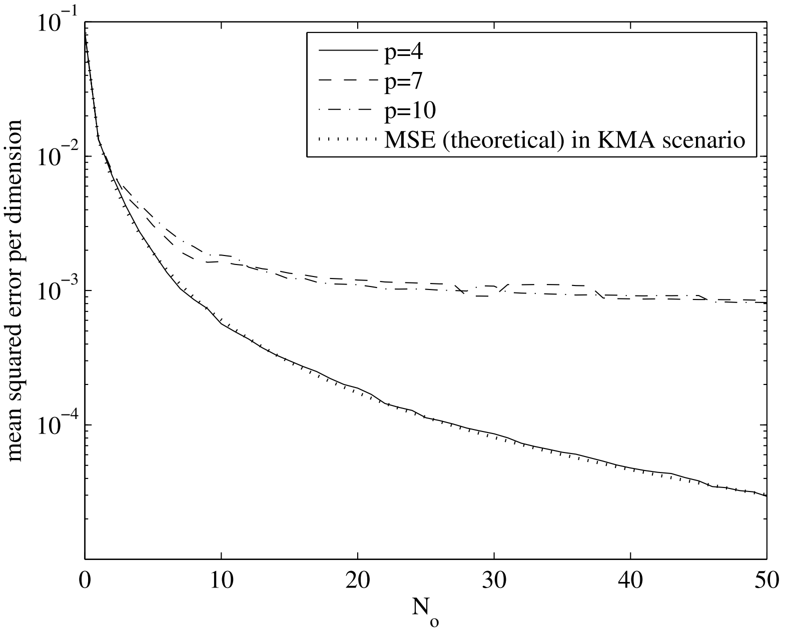


Figure 6.8: MSE per dimension in the dither estimation obtained for different embedding rates and $\alpha = 0.6$.

6.4.3 Reversibility attack

An accurate dither estimate (subjected to an unknown modulo-Λ shift, as the one obtained here) allows to implement a number of harmful attacks. As an illustrative example, we present here a reversibility attack, consisting in producing an estimate of the original host signal. In the context of lattice-data hiding methods, our attack is based on the fact that the embedding function is reversible whenever $\alpha < 1$ and we know both $\mathbf{T}$ and the embedded message. Using our dither and path estimates $\hat{\mathbf{t}}$ and $\mathbf{m}^{(\hat{k})}$, the host vector estimate corresponding to the ith marked block is computed as[7]

$$\hat{\mathbf{x}}_i = \mathbf{y}_i - \frac{\alpha}{1-\alpha}(Q_\Lambda(\mathbf{y}_i - \mathbf{d}_{m_i^{\hat{k}}} - \hat{\mathbf{t}}) - \mathbf{x}_i + \mathbf{d}_{m_i^{\hat{k}}} + \hat{\mathbf{t}}). \tag{6.23}$$

It is interesting to notice that the ambiguity in the estimated message does not affect negatively the host estimation whenever the estimated path $\mathbf{m}^{(\hat{k})}$ belongs to the equivalence class (by (5.61)) of the actual embedded path. The reason is that the dither estimate associated to any path in $[\mathbf{m}]$ yields the same fine lattice Λ_f, and thus it is valid for performing a successful reversibility attack. This can be readily seen if we realize that the dither estimates associated to the equivalence class of $\mathbf{m}^{(\hat{k})}$, given by (6.16), differ only in the term $\mathbf{d}_k$, and that $\Lambda_f = \Lambda_f - \mathbf{d}_k$, for $k \in \mathcal{M}$.

[7] The same reversibility function had been proposed in [108] in the context of scalar lattices.

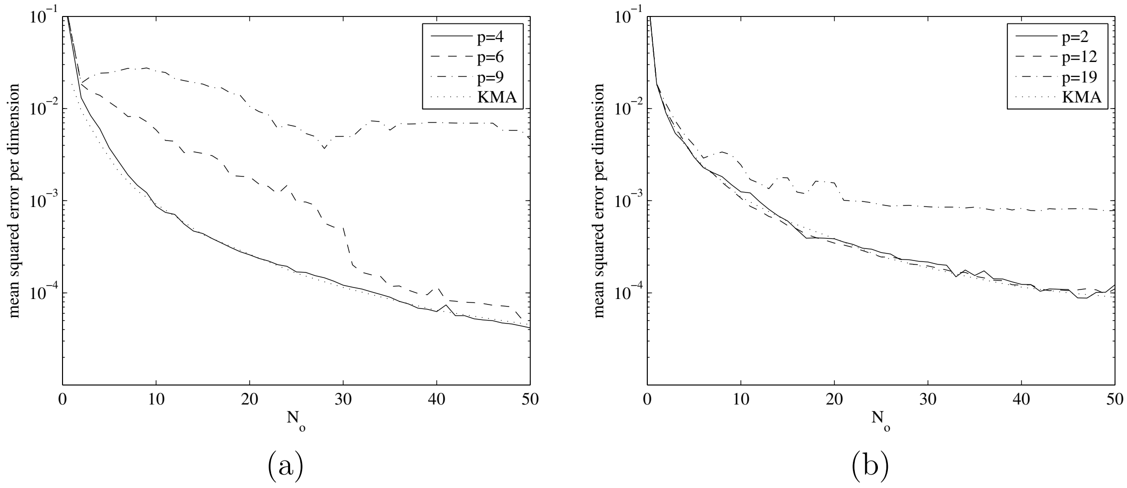


Figure 6.9: MSE per dimension for $\alpha = 0.7$ and different embedding rates. Results for $n = 2$ and $n = 8$ using hexagonal (a) and E_8 (b) shaping lattices, respectively. DWR = 30 dB in both cases.

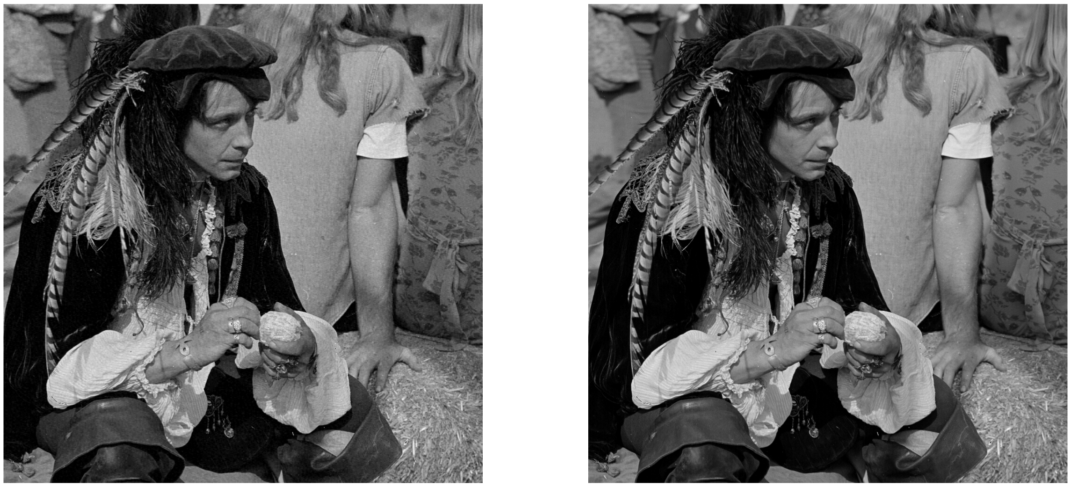
Figure 6.10 shows the result of implementing the proposed reversibility attack on a real marked image. The parameters of the watermarking algorithm are $\Lambda = E_8$, $\alpha = 0.7$, $p = 10$, and the coset leaders were obtained through Construction A. The watermark is embedded in the low frequency coefficients of 8×8 non-overlapping DCT blocks, resulting in a PSNR after embedding of 38.2 dB. The resulting host estimate, using only the message and dither estimates from the first 50 DCT blocks, is shown in Figure 6.10(b) and presents a PSNR of approximately 56 dB. If the estimated host is quantized back to integers, then PSNR $\rightarrow \infty$, meaning that the host has been estimated perfectly.

6.5 Application to other scenarios

In this section we discuss the application of the proposed estimation algorithms to other related but more involved scenarios.

1. $\alpha < 0.5$: Our analysis and algorithms were restricted to the case $\alpha \geq 0.5$. For the case $\alpha < 0.5$ the theoretical analysis gets more intricate, since the feasible region $\mathcal{S}_{N_o}$ may be composed of multiple modulo-Λ convex sets (cf. Figure 5.3). The difficulty of the estimation is also greatly increased, since it would be necessary to apply several estimators in parallel, one for each possible convex set. In such case, other set-membership approaches suited to non-convex solution sets may perform better [67].

2. Spread Transform - Dither Modulation (ST-DM) [63]: lattice data hiding schemes



(a) (b)

Figure 6.10: Illustration of a reversibility attack based on dither estimate according to Eq. (6.23). Image marked using $\Lambda = E_8$, $\alpha = 0.7$, $p = 10$ and PSNR = 38.2 dB (a), and estimate of the original image with PSNR = 55.9 dB (b).

may be applied in conjunction with spread transform in low-rate data hiding applications. In that kind of schemes, lattice quantization takes place in a secret projected domain, parameterized by certain projection matrix, and secret dithering can still be used in the projected domain for improving the security of the scheme. The ignorance of the projection matrix invalidates the direct application of the estimation algorithms proposed here; however, recent works [61], [46] have shown that Independent Component Analysis (ICA) may be used for estimating the projection matrix. Thus, if ICA is successful, dither estimators may be applied in a second step.

3. Permutations: the security of a lattice DC-DM scheme may be improved by applying secret permutations to the host vectors. This introduces an additional degree of uncertainty that invalidates the direct application of the estimators proposed in this chapter. However, if the same permutation is used in multiple marked blocks, it is still possible to exploit the information leakage, as shown in the next example: assume that the host is partitioned in l length-n vectors $\mathbf{x}_i$, $i = 1, \dots, l$, and these vectors are arranged in an $n \times l$ matrix $\mathbf{X}$. Given a secret permutation matrix $\mathbf{P}$, the columns of the new matrix $\mathbf{X}' = \mathbf{PX}$ are marked using the lattice data hiding scheme described in Chapter 5, yielding a marked matrix $\mathbf{Y}'$. Later on, the inverse permutation is applied to $\mathbf{Y}'$, obtaining $\mathbf{Y}$, and its rows are the observations that are made available to the attacker. Depending

on the symmetry properties of the embedding lattice, two possible cases arise:

(a) The lattice is symmetric to permutations of its components. This happens, for instance, to the cubic and "checkerboard" (aka "quincunx") lattices in 2 dimensions [174], [82]. If this is the case, then the attacker can run the dither estimation algorithm disregarding the actual permutation, obtaining an estimate of the permuted dither. It is easy to see that this permuted estimate allows the same attacks as those discussed in Section 2.2, as long as the permutation and the secret dither are the same in the attacked contents.

(b) The lattice is not symmetric to permutations. The main consequence is that the feasible regions for the dither are different under each permutation, and this can be exploited to detect inconsistent arrangements in the components of the observations, i.e., those arrangements that produce an empty feasible region cannot be correct. Some experiments performed with the OVE algorithm and the hexagonal lattice have shown that, using 10 recirculations, an average of 32 observations are needed to successfully detect inconsistent arrangements of the components. Using the inner polytope algorithm it is also possible to check for inconsistencies: one just needs to run the "feasibility test" to check whether all constraints in the optimization problem can be simultaneously satisfied or not. If not, the considered arrangement is inconsistent.

6.6 Conclusions

1. In Section 6.4 we have shown the strong link between the information-theoretic and set-membership estimation frameworks, applying the latter for the first time to attacks in the data hiding scenario. The estimator devised for the KMA scenario can be naturally extended to the WOA scenario with the help of a Viterbi-like estimator for the most likely embedded message.

2. The security weaknesses of the data hiding schemes studied in Chapter 5 have been shown to be exploitable in practice with affordable complexity, yielding accurate dither estimates and allowing to obtain host estimates with high fidelity.

3. Set-membership estimators are the optimal estimators for the dither estimation problem when $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$. However, in order to perform an attack with manageable complexity, it is necessary to do some approximations and simplifications that cause a loss of optimality (approximations with ellipsoids, for example, or with the Viterbi algorithm). Only in simple cases, such as embedding with cubic lattices, it is possible to achieve optimal estimator performance with low complexity. In other cases (as for the root lattices considered in Section 6.3), the

gap with the optimal dither estimator is still very close to the theoretical limit. However, as the complexity of the shaping lattice is increased, it seems that the gap with the optimal estimator will also be greatly increased if one wants to keep the complexity of the estimator within reasonable terms.

Algorithm 6.3 Secret dither estimation in the WOA scenario

1. Initialization: $\mathbf{m}^{(1)} = 0$, $\mathcal{D}_1(\mathbf{m}^{(1)}) = (1-\alpha)\mathcal{V}(\Lambda)$, $K_1 = 1$, with K_1 denoting the number of feasible paths for the first observation (1 in our case).

2. For $i = 2, \ldots, N_o$

   (a) Let $\{\mathbf{m}^{(k)},\ k = 1, \ldots, K_{i-1}\}$ be the set of feasible paths for the $i-1$ first observations. Construct a set of candidate paths as $\{\mathbf{m}^{(k,l)} \triangleq [\mathbf{m}^{(k)},\ l],\ k = 1, \ldots, K_{i-1},\ l = 0, \ldots, p-1\}$.

   (b) Compute the ellipsoids $\mathcal{E}_i(\mathbf{m}^{(k,l)}) \supseteq \mathcal{S}_i(\mathbf{m}^{(k,l)})$ using $\tilde{\mathbf{v}}_r(m_r^{(k,l)}) = (\tilde{\mathbf{y}}_r - \mathbf{d}_{m_r^{(k,l)}} - \tilde{\mathbf{y}}_1) \mod \Lambda,\ r = 1, \ldots, i$, where $m_r^{(k,l)}$ denotes the rth element of $\mathbf{m}^{(k,l)}$.

   (c) Compute the score $\lambda(\mathbf{m}^{(k,l)})$ of each path as $\text{vol}(\mathcal{E}_i(\mathbf{m}^{(k,l)}))$. Arrange the paths in order of descending score as $\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(p \cdot K_{i-1})}$. Compute now $K_i \triangleq \max_q q{+}1$, subject to the constraint $\lambda(\mathbf{m}^{(1)})/\lambda(\mathbf{m}^{(q)}) < \beta$, $q = 1, \ldots, p \cdot K_{i-1} \leq K_{max}$. The parameters $\beta > 0$, $K_{max} \in \mathbb{N}^+$ are termed "beam factors". The set of $K_i < K_{max}$ "surviving paths" for the next iteration is $\{\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(K_i - 1)}\}$.

3. Let $\mathbf{m}^{(1)}$ be the path with the highest score resulting from Step 2 of Algorithm 6.3, and let $\hat{\mathbf{t}}_1$, defined as the center of $\mathcal{E}_{N_o}(\mathbf{m}^{(1)})$, the dither estimate associated to $\mathbf{m}^{(1)}$ (recall that this choice minimizes the mean squared error of the estimate [194]). The p paths belonging to the equivalence class $[\mathbf{m}^{(1)}]$ can be computed according to (5.61), and the p corresponding dither estimates are given by

$$\hat{\mathbf{t}}_k = (\hat{\mathbf{t}}_1 - \mathbf{d}_k + \tilde{\mathbf{y}}_1) \mod \Lambda,\ k \in \mathcal{M}. \tag{6.16}$$

   Note that $\tilde{\mathbf{y}}_1$ is added in order to cancel the offset introduced in Step 2b of Algorithm 6.3.

Algorithm 6.4 Checking of unfeasible paths

According to [194, Sect. IV], assume the feasible region for the ith observation can be specified by a matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times n_f/2}$ and a vector $\boldsymbol{\gamma} \in \mathbb{R}^{n_f/2\times 1}$ such that $\mathcal{D}_i(m_i^{(k,l)}) = \bigcap_{j=1}^{n_f/2} \mathcal{F}_{i,j}$, where

$$\mathcal{F}_{i,j} = \{\mathbf{z} \in \mathbb{R}^n : |\tilde{\mathbf{v}}_i(m_i^{(k,l)})^T \boldsymbol{\phi}_j - \mathbf{z}^T \boldsymbol{\phi}_j| \leq \gamma_j\}, \tag{6.17}$$

being $\boldsymbol{\phi}_j$ the jth column of $\mathbf{\Phi}$, $\gamma_j \triangleq \boldsymbol{\phi}_j^T \mathbf{z}_{0,j}$ is the jth element of $\boldsymbol{\gamma}$, and $\mathbf{z}_{0,j}$ is a vector in the jth facet of $\mathcal{Z}(\Lambda)$. For $k = 1, \ldots, K_i$, and $l = 0, \ldots, p-1$,

1. Compute

$$\eta_j = \frac{\tilde{\mathbf{v}}_i(m_i^{(k,l)}) + \gamma_j - \boldsymbol{\phi}_j^T \mathbf{c}_{i-1}}{\sqrt{\boldsymbol{\phi}_j^T \mathbf{P}_{i-1} \boldsymbol{\phi}_j}}, \qquad \zeta_j = \frac{\gamma_j}{\sqrt{\boldsymbol{\phi}_j^T \mathbf{P}_{i-1} \boldsymbol{\phi}_j}}, \quad j = 1, \ldots, n_f/2, \tag{6.18}$$

   where $\mathbf{P}_{i-1}$ and $\mathbf{c}_{i-1}$ are the positive-definite matrix and center defining the ellipsoid $\mathcal{E}_{i-1}(\mathbf{m}^{(k)})$.

2. If $\eta_j \notin [-1, 1 + 2\zeta_j]$, for some $j = 1, \ldots, n_f/2$, then the hypothesized path $\mathbf{m}^{(k,l)}$ is unfeasible. Otherwise, the path is declared as feasible.[4]

Example 6.1 Parameters for the hexagonal lattice

Consider the standard hexagonal lattice whose generator matrix is [82]:

$$\mathbf{M} = \begin{bmatrix} \Delta & 0 \\ \Delta/2 & \sqrt{3}\Delta/2 \end{bmatrix}, \tag{6.19}$$

where Δ is the scaling factor for adjusting the embedding distortion. This lattice has 6 vectors of minimal norm that define a regular polytope bounded by 6 pairwise parallel hyperplanes (see Figure 6.1(a)). Let $\boldsymbol{\phi}_k$ be the vector corresponding to the k-th facet, the feasible region can be defined as

$$\mathcal{S}_{N_o} = \left\{ \mathbf{z} \in \mathbb{R}^n : \boldsymbol{\phi}_k^T \mathbf{z} \leq \boldsymbol{\phi}_k^T \tilde{\mathbf{v}}_i + \frac{1}{2} ||\boldsymbol{\phi}_k||^2 \right\},\ k = 1, \ldots, n_f;\ i = 1, \ldots, N_o, \tag{6.20}$$

which is a particular case of Eq. (6.4) with $\mathbf{z}_{0,k} = \frac{1}{2}\boldsymbol{\phi}_k$. For applying the SME algorithm, we need matrix $\boldsymbol{\Phi}$ and vector $\boldsymbol{\gamma}$ (see Eq. (6.8)). It is straightforward to see that, for the considered generator matrix,

$$\boldsymbol{\Phi} = \Delta \cdot (1-\alpha) \cdot \begin{bmatrix} 0 & \sqrt{3}/2 & \sqrt{3}/2 \\ 1 & 1/2 & -1/2 \end{bmatrix}, \quad \boldsymbol{\gamma} = \frac{1}{2} \cdot (\Delta \cdot (1-\alpha))^2 \cdot (1,1,1)^T. \tag{6.21}$$

Example 6.2 Parameters for the lattice D_4^*

For D_4^*, the faces of the hypercube are given by the hyperplanes bisecting all permutations of the vector $(\pm 1, 0, 0, 0)$. On the other hand, the faces of the generalized octahedron are given by the hyperplanes bisector to $\frac{1}{2}(\pm 1, \pm 1, \pm 1, \pm 1)$. Thus, the resulting Voronoi region has 24 pairwise parallel facets.

Chapter 7

Conclusions

In this thesis we have tried to provide an exhaustive analysis of watermarking security. The problem has been cast in an information-theoretic framework whose usefulness has been proved by its application to practical watermarking schemes in three different scenarios: KMA, CMA, and WOA. Our work shows the need for complementary security analyses, both from theoretical and practical points of view:

- The theoretical analysis must use precise measures in order to reveal the fundamental security weaknesses and establish bounds on security. Our measure based on Shannon's mutual information and equivocation, introduced in Chapter 2, has proved to be very useful for revealing the security properties of a given embedding method (cf. chapters 3 and 5) and deriving fundamental bounds on the estimation of the secret parameters.
- As the theoretical analysis does not show how to exploit the information leakage about the secret keys, a practical analysis is necessary for demonstrating that the security weaknesses highlighted in the theoretical part indeed pose a threat. We have done so in chapters 4 and 6.
- On the other hand, the security must not be solely evaluated from a practical point of view, because it may lead to overestimating the actual security of the watermarking scheme under study.

We have studied the two main groups of watermarking methods: spread spectrum and quantization-based ones. In some cases, similar conclusions can be drawn for both methods:

1. A tradeoff between robustness and security has been identified. In general, the choice of the embedding parameters that maximize the robustness against conventional attacks leads to reveal more information about the secret key. In this

regard, one must bear in mind that the analyzed embedding methods had not been originally designed from a security standpoint.

2. The gap of the information leakage between KMA and WOA scenarios (and CMA) has been shown to be negligible in many cases if the embedding rate is sufficiently small. As practical scenarios usually demand small embedding rates, a consequence is that the security level of the studied methods may be, in practice, fairly low.

3. Theoretically, some fundamental ambiguities have been shown to exist in the secret key estimation for both methods in the WOA scenario. These ambiguities prevent from perfect estimation of the secret key, no matter the number of observations available. Nevertheless, they do not prevent from devising harmful attacks, as already discussed in the corresponding chapters.

Regarding the security of lattice data hiding techniques, randomized via secret dithering, we have obtained some interesting findings:

1. Lattice data hiding schemes possess some desirable security properties: by adjusting the embedding parameters (the nested code and the distortion compensation parameters), they can achieve perfect secrecy in the WOA scenario, meaning that no information about the secret dither is leaked from the observation of watermarked signals. The price to pay comes in the form of a possible penalty in robustness.

2. In most practical cases, the security level of lattice data hiding schemes is low, since a small number of observations is enough for obtaining an accurate dither estimate. This is mainly due to their host-rejection properties and highly structured codebook.

3. The security level of the lattice data hiding scheme can be increased by increasing the dimensionality of the nested lattice code and properly choosing the shaping lattice. The optimal lattices seem to be those with the smallest normalized second order moment, which is also desirable from the robustness point of view. The main drawback of using these high-dimensional lattices is in the inherent embedding/decoding complexity (especially, the latter one), which grows nonlinearly with the number of dimensions. This illustrates the existence of a tradeoff between complexity and security.

As for spread spectrum modulations, whose security relies on the secrecy of the spreading vector, we can make the following remarks:

1. The methods analyzed in this thesis do not provide perfect secrecy in any instance. Nevertheless, due the high host interference, the security level of the spread spectrum embedding function is, in general, larger than that of lattice data hiding (see discussion in Section 5.2.3). We know it is possible to design perfectly secure spread spectrum modulations, but at the cost of impairing robustness (cf. the Natural Watermarking technique in [59]). This shows, once again, a tradeoff between robustness and security.

2. While the ICA and PCA approaches proposed by other authors are very useful for attacking spread spectrum systems from the security point of view, we have shown that such techniques are not always the best choice from the attacker's point of view. Their success depends very much on the operating conditions (the DWR, the length of the spreading vector, and the host rejection parameter) and on the statistical distribution of the host signals. We have presented a more general methodology for the practical security evaluation of spread spectrum schemes, based on 1) the definition and analysis of suitable cost functions, and 2) the optimization of such cost functions.

From the results shown in this thesis, one can infer that if the condition of perfect secrecy is not fulfilled, watermarking systems are in general "easy" to break, especially if robustness (low embedding rates) is a requirement. As mentioned in [85], secret keys in watermarking are "weaker" than in cryptography, meaning that the attacker does not need to perfectly disclose the key, but a rough estimate is sometimes enough for his purposes. This weakness is, in part, due to the continuous nature of the secret parameters to be estimated, and also to the relatively small size of the secret parameter space:[1]

- In lattice data hiding schemes, the codebook is highly structured, which is highly desirable for keeping a low embedding and decoding complexity. The drawback is that in such case there is no room for high entropy: the maximum entropy of the codebook in the nested lattice scheme is limited to the entropy of a uniform r.v. over $\mathcal{V}(\Lambda)$, whose volume is limited due to the small distortion constraint.

- In the case of spread spectrum methods, the attacker can restrict his search to a vector in the surface of an n-dimensional hypersphere. The number of degrees of freedom of such vector is $n-1$. Hence, the complexity of the problem increases linearly with the dimensionality of the embedding function.

[1] Note that when dealing with continuous parameters, it makes more sense to talk about the volume of the secret parameter space rather than its cardinality.

It is likely that, in the future, the development of watermarking research keeps on being driven by robustness requirements, as it has been up to now. However, under the light of the results presented in this thesis we believe it is wise to have in mind security considerations whenever the watermarking systems have to face hostile environments.

7.1 Future research lines

The work carried out in this thesis leaves a number of open questions which are worth addressing in the future. In the first place, we can enumerate a number of problems closely related to the security analysis:

1. In this thesis we have studied the security both from information-theoretic and practical points of view. The information-theoretic measure only accounts for attackers with unlimited computational resources. On the other hand, the practical security analysis only provides the security level obtained with the proposed estimation algorithms. In some cases, we have compared the theoretical bounds (which are pessimistic, in general) with the performance of the practical estimators, but we do not know yet how to compute the gap between theoretical and practical security levels as a function of the allowed computational complexity. This claims for the introduction of complexity theory in the field of watermarking security, which would lead to the obtention of provable security levels (in analogy with the provable security of cryptographic systems).

2. We have proved the existence of a fundamental tradeoff between security and robustness for spread spectrum and lattice data hiding methods. This tradeoff appears to be inherent to any watermarking scheme that performs host rejection or demands high robustness, but a proper unifying framework would be necessary in order to study this problem with a higher perspective.

3. We have studied the security of the most popular spread spectrum methods. Other spread spectrum techniques recently proposed, with the aim of improving the security of these methods, have been already analyzed [59]. However, the analysis of other spread spectrum schemes [165], [70] which are likely to receive a lot of attention in the future is still pending.

4. A theoretical study of the security of hybrid methods such as ST-DM [63],[108] and Trellis-based methods [168], which were not covered in this thesis, is advisable in order to get more insight into the security properties of the different side-informed embedding methods.

5. The effect of channel coding techniques in the security level must be addressed. When no channel coding across different host blocks takes place, the message

sequences can be considered to be a priori equiprobable. Otherwise, the attacker could exploit the a priori probability of each message sequence, simplifying the estimation of the secret parameters and thus reducing the security level.

6. In parallel to the design of new watermarking schemes with improved security properties, the design of countermeasures against security attacks must be also taken into account. Key management solutions are the most immediate countermeasures: since the security holes come from the availability of many observations marked with the same secret key, an obvious countermeasure is to limit the number of times a certain secret key can be reused. Several solutions based on host-dependent key generation functions have been proposed [135],[117],[155]. However, re-synchronization at the decoder side becomes a major issue when using host-dependent keys.

7. It is possible to study new forms of randomization for the existing watermarking schemes. For instance, the security of lattice data hiding schemes would be certainly improved by means of non-structured codebooks ([113],[69]) which, albeit strongly based on the nested lattice construction, introduce additional degrees of freedom that increase the entropy of the codebook and probably render the estimation more difficult. These codes could be seen as an intermediate step between Costa's construction (cf. Section 5.4) and the pure nested lattice construction. By varying the degree of additional randomness in the lattice code, one could balance the tradeoff between embedding/decoding complexity and security. Another possibility for lattice data hiding schemes is to apply a secret rotation to the embedding lattice [172]. The presumable advantage of this approach over the former is that it still keeps the structure of the codebook, which is highly desirable from the implementation point of view, whilst increasing its entropy (an $n \times n$ rotation matrix introduces $n(n-1)/2$ degrees of freedom). Further possibilities include the use of permutations (applicable to any watermarking scheme) or other combinatorial approaches, which could significantly increase the size of the secret parameter space.

The framework for watermarking security presented in this thesis must be regarded as a first step towards a more complete analysis. It is necessary to extend this framework to more general scenarios:

1. Extension to steganography scenarios: traditionally, a steganographic scheme has been considered secure if it is impossible for an eavesdropper to distinguish “innocent” cover messages from stego-messages. Security in steganography under this viewpoint has been addressed by a number of authors, like Zollner et al. [223], Petitcolas et al. [198], or Cachin [57] from an information-theoretic perspective,

and by Katzenbeisser et al. [149] resorting to more practical, computational considerations. However, an aspect which is usually absent from these formulations is the fact that the secret key used in the communication process is likely to be repeated in the stego-messages generated by the same user. This gives raise to the following considerations:

(a) An active eavesdropper, after detecting a series of stego-messages, can try to estimate the secret key in order to impair the communication process or to recover the hidden message.

(b) The repetition of the secret key in several stego-messages introduces memory (in the form of redundancy) in the communication channel, a fact that could be exploited for improving the detectability performance of the eavesdropper.

2. Extension to oracle attacks: currently, the security of a watermarking system where decoders/detectors are publicly available is roughly quantified by the difficulty of describing the detection/decoding region. The security measures are too empirical, mainly based on the difficulty of breaking the system with a particular algorithm. In this cases, no fundamental measure exists for comparing in fair terms the security of different watermarking systems.

3. Extension to global security: this thesis only covers security in the "physical layer". As we have seen, the disclosure of the secret parameters gives access to the raw embedded bits (possibly with some ambiguities). Although this can be enough for impairing the communication very efficiently, it may not be sufficient for reading the actual embedded message if, for instance, a cryptographic layer is placed upon the watermarking channel (the physical layer) [85]. Thus, a global security analysis of a watermarking system should account for attacks at the protocol level and even side-channel attacks, which may be very important in practical systems. As in any cryptographic system, the security level is always given by that of the weakest link in the chain.

Bibliography

[1] http://www.aquamobile.es/.

[2] http://www.business-sites.philips.com/contentidentification/home.

[3] http://www.cinea.com.

[4] http://www.cnn.com/2004/showbiz/01/23/oscar.arrest/index.html.

[5] http://www.commtech2000.co.uk.

[6] http://www.digimarc.com/docs/dmrc_content_id.pdf .

[7] http://www.digimarc.com/.

[8] http://www.digitalwatermarkingalliance.org/.

[9] http://www.geovision.com.tw.

[10] http://www.mediasec.com/.

[11] http://www.protectron.com.

[12] http://www.teletrax.tv.

[13] http://www.thomson.net/globalenglish/solutions/content-tracking-security.

[14] http://www.tredess.com/en/index.html.

[15] http://www.verance.com.

[16] http://www.verymatrix.com.

[17] Harmonic number. http://mathworld.wolfram.com/HarmonicNumber.html.

[18] Ley 59/2003, de 19 de diciembre, de firma electrónica. BOE, 20 de diciembre de 2003.

[19] Regolamento recante criteri e modalità per la formazione, l'archiviazione e la trasmissione di documenti con strumenti informatici e telematici. 10 novembre 1997, n. 513.

[20] SDMI challenge FAQ. http://www.cs.princeton.edu/sip/sdmi/faq.html.

[21] Secure Digital Music Initiative (SDMI).
http://en.wikipedia.org/wiki/ Secure_Digital_Music_Initiative .

[22] The Public-Key Criptography Standards. Available at http://www.rsasecurity.com/rsalabs/node.asp?id=2124.

[23] The USC-SIPI Image Database. Available at http://sipi.usc.edu/database/.

[24] Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a Community framework for electronic signatures, 19 January 2000. Official Journal of the European Communities.

[25] German Digital Signature Law (SigG), 19 June 1997. Available at http://www.kuner.com/data/sig/digsig4.htm.

[26] Digital Millenium Copyright Act, 1998. Available at http://www.copyright.gov/legislation/dmca.pdf.

[27] IEEE Signal Processing Magazine, 21(2), March 2004. Special issue on Digital Rights: Management, Protection, Standardization.

[28] Digital imaging procedure v1.0, March 2002. Available at
http://www.homeoffice.gov.uk/docs/digimpro.pdf.

[29] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

[30] André Adelsbach, Ulrich Huber, and Ahmad-Reza Sadeghi. Fingercasting - joint fingerprinting and decryption of broadcast messages. *Transactions on Data Hiding and Multimedia Security II*, 4499:1–34, 2007.

[31] André Adelsbach and Ahmad-Reza Sadeghi. Zero-knowledge watermark detection and proof of ownership. In *4th International Workshop on Information Hiding, IH'01*, volume 2137 of *Lecture Notes in Computer Science*, pages 273–288. Springer, 2001.

[32] Rajen Akalu and Deepa Kundur. Technological protection measures in the courts. *IEEE Signal Processing Magazine*, 21(2):109–117, March 2004. Special Issue on Digital Rights Management.

[33] 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 Wah Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents VI*, volume 5306, pages 685–695. SPIE, January 2004.

[34] Ross J. Anderson. Stretching the limits of steganography. In *1st International Workshop on Information Hiding, IH'96*, pages 39–48, Cambridge, UK, May 1996. Springer-Verlag.

[35] Ross J. Anderson and Markus G. Kuhn. Low cost attacks on tamper resistant devices. In *International Workshop on Security Protocols*, volume 1361 of *Lecture Notes in Computer Science*, pages 125–136, Paris, France, April 1997. Springer Verlag.

[36] Félix Balado, Kevin M. Whelan, Guénolé 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.

[37] Mauro Barni. Effectiveness of exhaustive search and template matching against watermark desynchronization. *IEEE Signal Processing Letters*, 12(2):158–161, February 2005.

[38] Mauro Barni and Franco Bartolini. Data hiding for fighting piracy, March 2004. Special issue on Digital Rights: Management, Protection, Standardization.

[39] Mauro Barni and Franco Bartolini. *Watermarking Systems Engineering*. Signal Processing and Communications. Marcel Dekker, 2004.

[40] Mauro Barni, Franco Bartolini, Vitto Cappellini, and Alessandro Piva. Robust watermarking of still images for copyright protection. In *13th International Conference on Digital Signal Processing, DSP'97*, volume 2, pages 499–502, Santorini, Greece, 2-4 July 1997.

[41] Mauro Barni, Franco Bartolini, and Teddy Furon. A general framework for robust watermarking security. *Signal Processing*, 83(10):2069–2084, October 2003. Special issue on Security of Data Hiding Technologies, invited paper.

[42] John R. Barry, Edward A. Lee, and David G. Messerschmitt. *Digital Communication*. Kluwer Academic Press, third edition, 2004.

[43] Franco Bartolini, Anastasios Tefas, Mauro Barni, and Ioannis Pitas. Image authentication techniques for surveillance applications. *Proceedings of the IEEE*, 89(10):1403–1418, October 2001.

[44] Patrick Bas and François Cayre. Achieving subspace or key security for WOA using natural or circular watermarking. In *ACM Multimedia and Security Workshop, MMSEC'06*, Geneva, Switzerland, September 2006.

[45] Patrick Bas and Gwenaël J. Doërr. Practical security analysis of dirty paper trellis watermarking. In *9th International Workshop on Information Hiding, IH'07*, volume 4567 of *Lecture Notes in Computer Science*, pages 174–188, Saint Malo, France, 11-13 June 2008. Springer-Verlag.

[46] Patrick Bas and Jarmo Hurri. Security of DM quantization watermarking schemes: a practical study for digital images. In Mauro Barni, Ingemar Cox, Ton Kalker, and Hyoung Joong Kim, editors, *4th International Workshop on Digital Watermarking, IWDW'05*, volume 3710, pages 186–200, Siena, Italy, September 2005. Springer-Verlag.

[47] Walter Bender, Daniel Gruhl, Norishige Morimoto, and Aiguo Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3-4):313–336, 1996.

[48] Eric C. Berg. Legal ramifications of digital imaging in law enforcement. *Forensic Science Communications*, 2(4), October 2000. Available at `http://www.fbi.gov/hq/lab/fsc/backissu/oct2000/berg.htm`.

[49] Patrick P. Bergmans. A simple converse for broadcast channels with additive white Gaussian noise. *IEEE Transactions on Information Theory*, 20(2):279–281, March 1974.

[50] Peter Biddle, Paul England, Marcus Peinado, and Bryan Willman. The darknet and the future of content protection. In Springer Berlin / Heidelberg, editor, *Digital Rights Management - Technological, Economic, Legal and Political Aspects*, volume 2770 of *Lecture Notes in Computer Science*, pages 344–365, 2003.

[51] 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. Special Issue on Identification and Protection of Multimedia Information.

[52] F. M. Boland, Joseph J. K. O'Ruanaidh, and C. Dautzenberg. Watermarking digital images for copyright protection. In *5th IEE International Conference on Image Processing and its Applications*, pages 326–330, Edimburg, July 1995.

[53] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, September 1998.

[54] William M. Boothby. *An introduction to differentiable manifolds and Riemannian geometry.* Pure and Applied Mathematics. Academic Press, New York, NY, 1975.

[55] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. SIAM Studies in Applied Mathematics. Cambridge University Press, Cambridge, UK, 2004.

[56] Christian Cachin. An information-theoretic model for steganography. In David Aucsmith, editor, *2nd International Workshop on Information Hiding, IH'98*, volume 1525 of *Lecture Notes in Computer Science*, pages 306–318, Portland, OR, USA, April 1998. Springer Verlag.

[57] Christian Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, 2004.

[58] Roberto Caldelli, Franco Bartolini, and Vito Cappellini. Metadata hiding tightly binding information to content. In *International Conference on Dublin Core and Metadata Applications, DCMI'02*, page 199, Florence, Italy, 13-17 October 2002. Dublin Core Metadata Initiative.

[59] François Cayre and Patrick Bas. Kerckhoffs-based embedding security classes for WOA data hiding. *IEEE Transactions on Information Forensics and Security*, 3(1):1–15, March 2008.

[60] François Cayre, Caroline Fontaine, and Teddy Furon. Watermarking attack: security of WSS techniques. In Ingemar J. Cox, Ton Kalker, and Heung-Kyu Lee, editors, *3rd International Workshop on Digital Watermarking, IWDW'04*, volume 3304, pages 171–183, Seoul, Korea, October 30 - November 1 2004. Springer.

[61] François Cayre, Caroline Fontaine, and Teddy Furon. Watermarking security: theory and practice. *IEEE Transactions on Signal Processing*, 53(10):3976–3987, October 2005.

[62] Brian Chen and Gregory W. Wornell. Dither modulation: a new approach to digital watermarking and information embedding. In Edward J. Delp III and Ping Wah Wong, editors, *Security and Watermarking of Multimedia Contents*, volume 3657, pages 342–353, San José, CA, USA, 25-27 January 1999. SPIE.

[63] 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.

[64] Man-Fung Cheung, Stephen Yurkovich, and Kevin M. Passino. An optimal volume ellipsoid algorithm for parameter set estimation. *IEEE Transactions on Automatic Control*, 38(8):1292–1296, August 1993.

[65] Maha El Choubassi and Pierre Moulin. Noniterative algorithms for sensitivity analysis attacks. *IEEE Transactions on Information Forensics and Security*, 2(3):113–126, June 2007.

[66] Aaron S. Cohen and Amos Lapidoth. The Gaussian watermarking game. *IEEE Transactions on Information Theory*, 48(6):1639–1667, June 2002.

[67] Patrick L. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, February 1993.

[68] Pedro Comesaña. *Side-informed data hiding: robustness and security analysis.* PhD thesis, University of Vigo, Vigo, Spain, 2006.

[69] Pedro Comesaña, Félix Balado, and Fernando Pérez-González. A novel interpretation of content authentication. In Edward J. Delp III and Ping W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, San Jose, California, USA, January 2007. SPIE.

[70] Pedro Comesaña, Mauro Barni, and Neri Merhav. Asymptotically optimum embedding strategy for one-bit watermarking under Gaussian attacks. In Edward J. Delp III, Ping W. Wong, Jana Dittmann, and Nasir Memon, editors, *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, volume 6819, San Jose, California, USA, January 2008. SPIE.

[71] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. Fundamentals of data hiding security and their application to spread-spectrum analysis. In *7th International Workshop on Information Hiding, IH'05*, volume 3727 of *Lecture Notes in Computer Science*, pages 146–160, Barcelona, Spain, June 2005. Springer Verlag.

[72] 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.

[73] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. The return of the sensitivity attack. In Mauro Barni, Ingemar J. Cox, Ton Kalker, and Hyoung Joong Kim, editors, *4th International Workshop on Digital Watermarking, IWDW'05*, volume 3710 of *Lecture Notes in Computer Science*, pages 260–274, Siena, Italy, September 2005. Springer.

[74] Pedro Comesaña, Luis Pérez-Freire, and Fernando Pérez-González. Blind Newton Sensitivity Attack. *IEE Proceedings on Information Security*, 153(3):115–125, September 2006.

[75] Pedro Comesaña, Fernando Pérez-González, and Félix Balado. On distortion-compensated dither modulation data-hiding with repetition coding. *IEEE Transactions on Signal Processing*, 54(2):585–600, February 2006.

[76] 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 Wah Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 298–307, San Jose, California, USA, January 2005. SPIE.

[77] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[78] John H. Conway, Ronald H. Hardin, and Neil J. A. Sloane. Packing lines, planes, etc.: packings in grassmanian spaces. *Experimental Mathematics*, 5(2):139–159, 1996.

[79] John H. Conway, E. M. Rains, and Neil J. A. Sloane. On the existence of similar sublattices. *Canadian Journal of Mathematics*, 51:1300–1306, 1999.

[80] John H. Conway and Neil J. A. Sloane. Fast quantizing and decoding algorithms for lattice quantizers and codes. *IEEE Transactions on Information Theory*, 28(2):227–232, March 1982.

[81] John H. Conway and Neil J. A. Sloane. Voronoi regions of lattices, second moments of polytopes, and quantization. *IEEE Transactions on Information Theory*, 28(2):211–226, March 1982.

[82] John H. Conway and Neil J. A. Sloane. *Sphere packings, lattices and groups*, volume 290 of *Comprehensive Studies in Mathematics.* Springer-Verlag, New York, 3rd edition, 1999.

[83] Max H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[84] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley series in Telecommunications, 1991.

[85] Ingemar J. Cox, Gwenaëll Doërr, and Teddy Furon. Watermarking is not cryptography. In Yun Qing Shi and Byeungwoo Yeon, editors, *5th International Workshop on Digital Watermarking, IWDW'06*, volume 4283 of *Lecture Notes in Computer Science*, pages 1–15, Jeju Island, Korea, November 2006. Springer Berlin / Heidelberg.

[86] Ingemar J. Cox, Joe Killian, Tom Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997.

[87] Ingemar J. Cox, John Killian, Tom Leighton, and Talal Shamoon. Secure spread spectrum watermarking for images, audio and video. *IEEE International Conference on Image Processing, ICIP97*, 3:243–246, December 1996.

[88] 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.

[89] Ingemar J. Cox and Matthew L. Miller. A review of watermarking and the importance of perceptual modeling. In *Human Vision and Electronic Imaging II*, volume 3016, pages 92–99. SPIE, June 1997.

[90] Ingemar J. Cox and Matthew L. Miller. The first 50 years of electronic watermarking. *EURASIP Journal on Applied Signal Processing*, 2:126–132, February 2002.

[91] Ingemar J. Cox, Matthew L. Miller, and Andrew L. McKellips. Watermarking as communications with side information. *Proceedings of the IEEE*, 87(7):1127–1141, July 1999.

[92] H. S. M. Coxeter. *Regular polytopes*. Dover, New York, third edition, 1973.

[93] Harald Cramér. *Mathematical methods of statistics*. Landmarks on Mathematics. Princeton University Press, 1999. Reprint.

[94] Scott Craver, Boon-Lock Yeo, and Minerva Yeung. Technical trials and legal tribulations. *Communications of the ACM*, 41(7):44–54, July 1998.

[95] Jean-François Delaigle. *Protection of intelectual property of images by perceptual watermarking*. PhD thesis, Université Catholique de Louvain, 2000.

[96] John R. Deller. Set membership identification in digital signal processing. *IEEE ASSP Magazine*, 6(4):4–20, October 1989.

[97] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[98] Geert Depovere, Ton Kalker, Jaap Haitsma, Maurice Maes, Lieven de Strycker, Pascale Termont, Jan Vandewege, Andreas Langell, Claes Alm, Per Norman, Gerry O'Reilly, Bob Howes, Henk Vaanholt, Rein Hintzen, Pat Donnelly, and Andy Hudson. The VIVA project: digital watermarking for broadcast monitoring. In *IEEE International Conference on Image Processing, ICIP'99*, volume 2, pages 202–205, 24-28 October 1999.

[99] José A. Díaz-García and Graciela González-Farías. Singular random matrix decompositions: Jacobians. *Journal of Multivariate Analysis*, 93(2):296–312, 2005.

[100] Werner Dietl, Peter Meerwald, and Andreas Uhl. Protection of wavelet-based watermarking systems using filter parametrization. *Elsevier Signal Processing*, 83(10):2095–2116, 2003.

[101] Whitfield Diffie and Martin Hellman. New directions in cryptography. *IEEE Transactions on Information Theory*, 22(6):644–684, November 1976.

[102] Gwenaël Doërr and Jean-Luc Dugelay. Danger of low-dimensional watermarking subspaces. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'04*, volume 3, pages 93–96, Montreal, Canada, May 2004.

[103] Gwenaël Doërr and Jean-Luc Dugelay. Security pitfalls of frame-by-frame approaches to video watermarking. *IEEE Transactions on Signal Procesing, Supplement on Secure Media*, 52(10):2955–2964, October 2004.

[104] Jiang Du, Chong-Hoon Lee, Heun-Kio Lee, and Youngho Suh. Watermark attack based on blind estimation without priors. In *1st International Workshop on Digital Watermarking, IWDW'02*, volume 2613 of *Lecture Notes in Computer Science*, Seoul, Korea, 2002. Springer.

[105] ECRYPT. European Network of Excellence in Cryptology, 2004-2008. `http://www.ecrypt.eu.org`.

[106] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[107] Joachim J. Eggers, Robert Bäuml, and Bernd Girod. Estimation of amplitude modifications before SCS watermark detection. In *Security and Watermarking of Multimedia Contents IV*, volume 4675, pages 387–398, San Jose, CA, USA, January 2002. SPIE.

[108] Joachim J. Eggers, Robert Bäuml, Roman Tzschoppe, and Bernd Girod. Scalar Costa Scheme for information embedding. *IEEE Transactions on Signal Processing*, 51(4):1003–1019, April 2003. Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery.

[109] Joachim J. Eggers and Bernd Girod. Blind watermarking applied to image authentication. In *IEEE International Conference on Audio, Speech and Signal Processing, ICASSP'01*, volume 3, pages 1977–1980, Salt-Lake City, USA, May 2001.

[110] Joachim J. Eggers, Jonathan K. Su, and Bernd Girod. Public key watermarking by eigenvectors of linear transforms. In *European Signal Processing Conference, EUSIPCO'00*, Tampere, Finland, 5-8 September 2000.

[111] Uri Erez, Simon Litsyn, and Ram Zamir. Lattices which are good for (almost) everything. *IEEE Transactions on Information Theory*, 51(10):3401–3416, October 2005.

[112] 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.

[113] Chuhong Fei, Deepa Kundur, and Raymond H. Kwong. Analysis and design of secure watermark-based authentication systems. *IEEE Transactions on Information Forensics and Security*, 1(1):43–55, March 2006.

[114] William Feller. *An Introduction to Probability Theory and Its Applications.* John Wiley and Sons, New York, 3rd edition, 1968.

[115] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368, 1922.

[116] G. David Forney. Multidimensional constellations - Part II: Voronoi constellations. *IEEE Journal of Selected Areas in Communications*, 7(6):941–958, August 1989.

[117] Jessica Fridrich and Miroslav Goljan. Robust hash functions for digital watermarking. In *Proceedings of the International Conference on Information Technology: Coding and Computing*, pages 173–178, Las Vegas, Nevada, USA, March 2000.

[118] Jessica Fridrich, Miroslav Goljan, and Rui Du. Detecting LSB steganography in color, and gray-scale images. *IEEE Multimedia*, 8(4):22–28, October-December 2001.

[119] Teddy Furon. A survey of watermarking security. In M. Barni, editor, *International Workshop on Digital Watermarking, IWDW'05*, volume 3710 of *Lecture Notes on Computer Science*, pages 201–215, Siena, Italy, sep 2005. Springer-Verlag.

[120] Teddy Furon and Pierre Duhamel. An asymmetric watermarking method. *IEEE Transactions on Signal Processing*, 51(4):981–995, April 2003. Special Issue on Signal Processing for Data Hiding in Digital Media & Secure Content Delivery.

[121] Teddy Furon et al. Security Analysis. *European Project IST-1999-10987 CERTIMARK, Deliverable D.5.5*, 2002.

[122] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in Mathematical Sciences. Johns Hopkins University Press, 3rd edition, 1996.

[123] Alban Goupil and Jacques Palicot. New algorithms for blind equalization: the Constant Norm Algorithm family. *IEEE Transactions on Signal Processing*, 55(4):1436–1444, April 2007.

[124] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Additive non-Gaussian noise channels: Mutual information and conditional mean estimation. In *IEEE International Symposium on Information Theory, ISIT'05*, pages 719–723, Adelaide, Australia, 4-9 September 2005.

[125] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, April 2005.

[126] Aparna R. Gurijala and John R. Deller. Speech watermarking with objective fidelity and robustness criteria. In *Conference record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1908–1912, Pacific Grove, CA, USA, November 2003.

[127] Jaap Haitsma, Michiel van der Veen, Ton Kalker, and Fons Bruekers. Audio watermarking for monitoring and copy protection. In *ACM workshops on Multimedia, MULTIMEDIA '00*, pages 119–122, New York, NY, USA, 2000. ACM.

[128] Frank Hartung and Bernd Girod. Watermarking of uncompressed and compressed video. *Signal Processing*, 66(3):283–302, May 1998.

[129] Simon Haykin. *Adaptive filter theory*. Prentice-Hall, Englewood Cliffs, NJ, USA, 3rd edition, 1996.

[130] Emil F. Hembrooke. Identification of sound and like signals, 1961. United States Patent, 3004104.

[131] Cormac Herley. Why watermarking is nonsense. *IEEE Signal Processing Magazine*, 19(5):10–11, Sep 2002.

[132] Juan R. 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.

[133] Juan R. Hernández and Fernando Pérez-González. Throwing more light on image watermarks. In D. Aucsmith, editor, *2nd International Workshop on Information Hiding, IH'98*, volume 1525 of *Lecture Notes in Computer Science*, pages 191–207, Portland, OR, USA, April 1998. Springer-Verlag.

[134] 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 in Communications*, 16(4):510–524, May 1998.

[135] Matthew Holliman, Nasir Memon, and Minerva M. Yeung. On the need for image dependent keys for watermarking. In *IEEE Conference on Content Security and Data Hiding in Digital Media*, Newark, NJ, USA, May 1999.

[136] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[137] Peter J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, June 1985.

[138] Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems 10*, pages 273–279, Denver, Colorado, USA, 1998.

[139] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, May 1999.

[140] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Adaptive and learning systems for signal processing, communications and control. John Wiley & Sons, 2001.

[141] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[142] Digital Cinema Initiatives. DCI specification v1.2. Available at `www.dcimovies.com`, March 2008.

[143] Richard C. Jonhson, Philip Schniter, Thomas J. Endres, James D. Behm, Donald R. Brown, and Raúl A. Casas. Blind equalization using the constant modulus criterion: a review. *Proceedings of the IEEE*, 86(10):1927–1950, October 1998.

[144] Ton Kalker. System issues in digital image and video watermarking for copy protection. In *IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 562–567, Florence, Italy, 7-11 June 1999.

[145] Ton Kalker. Considerations on watermarking security. In *IEEE International Workshop on Multimedia Signal Processing, MMSP'01*, pages 201–206, Cannes, France, October 2001.

[146] Ton Kalker, Dick H. J. Epema, Pieter H. Hartel, R.(Inald) L. Lagendijk, and Maarten van Steen. Music2share - copyright-compliant music sharing in P2P systems. *Proceedings of the IEEE*, 92(6):961–970, June 2004.

[147] Ton Kalker, Jean-Paul M.G. Linnartz, and Marten van Dijk. Watermark estimation through detector analysis. In *IEEE International Conference on Image Processing, ICIP'98*, pages 425–429, Chicago, IL, USA, October 1998.

[148] Stefan Katzenbeisser. Computational security models for digital watermarks. In *Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS'05*, Montreux, Switzerland, April 2005.

[149] Stefan Katzenbeisser and Fabien A. P. Petitcolas. Defining security in steganographic systems. In Edward J. Delp III and Ping W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents IV*, volume 4675, pages 50–56, San Jose, California, USA, January 2002. SPIE.

[150] Auguste Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 9:5–38, January 1883.

[151] Darko Kirovski and Henrique S. Malvar. Spread spectrum watermarking of audio signals. *IEEE Transactions on Signal Processing*, 51(4):1020–1033, April 2003.

[152] Mark Kirstein. Beyond traditional DRM: Moving to digital watermarking & fingerprinting in media monetization, January 2008. Available at `http://www.multimediaintelligence.com`.

[153] Deepa Kundur and Dimitrios Hatzinakos. Digital watermark for telltale tamperproofing and authentication. *Proceedings of the IEEE*, 87(7):1167–1180, July 1999.

[154] John Lach, William H. Mangione-Smith, and Miodrag Potkonjak. Enhanced intellectual property protection for digital circuits on programmable hardware. In *3rd International Workshop on Information Hiding, IH'99*, Dresden, Germany, October 1999. Springer-Verlag.

[155] Eugene T. Lin and Edward J. Delp. Temporal synchronization in video watermarking. *IEEE Transactions on Signal Processing*, 52(10):3007–3022, October 2004.

[156] Jean-Paul M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In D. Aucsmith, editor, *2nd International Workshop on Information Hiding, IH'98*, volume 1525 of *Lecture Notes in Computer Science*, pages 258–272, Portland, OR, USA, April 1998. Springer Verlag.

[157] Tie Liu and Pierre Moulin. Error exponents for one-bit watermarking. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'03*, volume 3, pages 65–68, 6-8 April 2003.

[158] Johan Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004. Toolbox available at `http://control.ee.ethz.ch/~joloef/yalmip.php`.

[159] James B. MacQueen. Some methods for classification and analysis of multivariate observations. In L.M. LeCam and J. Neyman, editors, *Proc. of the 5th Berkeley Symposium on Mathematics Statistics and Probability*, 1967.

[160] Thierry Maillard and Teddy Furon. Towards digital rights and exemptions management systems. *Computer law and security report*, 20(4):281–287, July 2004.

[161] Henrique S. Malvar and Dinei A. F. Florêncio. Improved Spread Spectrum: a new modulation technique for robust watermarking. *IEEE Transactions on Signal Processing*, 51(4):898–905, April 2003.

[162] Jonathan H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 5(3):635–650, March 2002.

[163] Nasir Memon and Ping Wah Wong. A buyer-seller watermarking protocol. *IEEE Transactions on Image Processing*, 10(4):643–649, April 2001.

[164] Alfred J. Menezes, Scott A. Vanstone, and Paul C. Van Oorschot. *Handbook of Applied Cryptography.* CRC Press, Inc., Boca Raton, FL, USA, 1996.

[165] Neri Merhav and Erez Sabbag. Optimal watermark embedding and detection strategies under limited detection resources. *IEEE Transactions on Information Theory*, 54(1):255–274, January 2008.

[166] M. Kivanç Mihçak. *Information hiding codes and their applications to images and audio.* PhD thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, 2002.

[167] M. Kivanç Mihçak, Ramarathnam Venkatesan, and Mustafa Kesal. Cryptanalysis of discrete-sequence spread spectrum watermarks. In Fabien A. P. Petitcolas, editor, *5th International Workshop on Information Hiding, IH'02*, pages 226–246, Noordwijkerhout, The Netherlands, October 2002. Springer-Verlag.

[168] Matthew L. Miller, Gwenaël Doërr, and Ingemar J. Cox. Applying informed coding and embedding to design a robust, high capacity watermark. *IEEE Transactions on Image Procesing*, 13(6):792–807, June 2004.

[169] Thomas Mittelholzer. An information-theoretic approach to steganography and watermarking. In A. Pfitzmann, editor, *3rd International Workshop on Information Hiding, IH'99*, volume 1768 of *Lecture Notes in Computer Science*, pages 1–17, Dresden, Germany, September 1999. Springer Verlag.

[170] Pierre Moulin. Universal decoding of watermarks under geometric attacks. In *IEEE International Symposium on Information Theory*, pages 2353–2357, July 2006.

[171] Pierre Moulin. Comments on "Why watermarking is nonsense". *Signal Processing Magazine, IEEE*, 20(6):57–59, Nov. 2003.

[172] Pierre Moulin and Anil K. Goteti. Block QIM watermarking games. *IEEE Transactions on Information Forensics and Security*, 3(1):293–310, September 2006.

[173] Pierre Moulin and Aleksandar Ivanović. The zero-rate spread-spectrum watermarking game. *IEEE Transactions on Signal Processing*, 51(4):1098–1117, April 2003.

[174] Pierre Moulin and Ralf Koetter. Data hiding codes. *Proceedings of IEEE*, 93(12):2083–2126, December 2005.

[175] Pierre Moulin and M. Kivanç Mihçak. A framework for evaluating the data-hiding capacity of image sources. *IEEE Transactions on Image Processing*, 11(9):1029–1042, September 2002.

[176] Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, March 2003.

[177] Taiga Nakamura, Ryuki Tachibana, and Seiji Kobayashi. Automatic music monitoring and boundary detection for broadcast using audio watermarking. In Edward J. Delp III and Ping Wah Wong, editors, *Security and Watermarking of Multimedia Contents IV*, volume 4675, pages 170–180, San Jose, CA, USA, January 2002. SPIE.

[178] Yuri Nesterov and Arkadi Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics.* Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[179] Jorge Nocedal and Stephen J. Wright. *Numerical optimization.* Springer, 1999.

[180] The College of Radiographers. Guidance for the provision of forensic radiography services. Available at `http://www.sor.org/public/document-library/sor_guidance_provision_forensic_radiography.pdf`.

[181] Daniel P. Palomar and Sergio Verdú. Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, 52(1):141–154, January 2006.

[182] Shelby Pereira. Robust template matching for affine resistant image watermarks. *IEEE Transactions on Image Processing*, 9(6):1123–1129, June 2000.

[183] Luis Pérez-Freire. Practical estimators of the secret spreading vector for Improved Spread Spectrum modulations. Statistical analysis and results. Technical report, Signal Theory and Communications Department, University of Vigo, November 2007. Available at `http://www.gts.tsc.uvigo.es/gpsc/cgi-bin/bibsearch3.cgi`.

[184] 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, *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 721–733, San Jose, California, USA, January 2005. SPIE.

[185] Luis Pérez-Freire, Pedro Comesaña, and Fernando Pérez-González. Information-theoretic analysis of security in side-informed data hiding. In *7th International Workshop on Information Hiding, IH'05*, volume 3727 of *Lecture Notes in Computer Science*, pages 131–145, Barcelona, Spain, June 2005. Springer Verlag.

[186] Luis Pérez-Freire, Pedro Comesaña, Juan Ramón Troncoso-Pastoriza, and Fernando Pérez-González. Watermarking security: a survey. *Transactions on Data Hiding and Multimedia Security I*, 4300:41–72, 2006.

[187] Luis Pérez-Freire, Pierre Moulin, and Fernando Pérez-González. Security of spread-spectrum-based data hiding. In Edward J. Delp III and Ping W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, San Jose, California, USA, January 2007. SPIE.

[188] Luis Pérez-Freire and Fernando Pérez-González. Spread-spectrum vs. quantization-based data hiding: misconceptions and implications. In Edward J. Delp III and Ping W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 341–352, San Jose, California, USA, January 2005. SPIE.

[189] Luis Pérez-Freire and Fernando Pérez-González. Exploiting security holes in lattice data hiding. In *9th International Workshop on Information Hiding, IH'07*, Lecture Notes in Computer Science, Saint Malo, France, June 11-13 2007. Springer Verlag.

[190] Luis Pérez-Freire and Fernando Pérez-González. Security of lattice-based data hiding against the watermarked only attack. *IEEE Transactions on Information Forensics and Security*, 2008. Accepted for publication.

[191] Luis Pérez-Freire and Fernando Pérez-González. Spread spectrum watermarking security. *IEEE Transactions on Information Forensics and Security*, 2008. Submitted.

[192] Luis Pérez-Freire, Fernando Pérez-González, and Pedro Comesaña. Secret dither estimation in lattice-quantization data hiding: a set-membership approach. In Edward J. Delp III and Ping W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents VIII*, San Jose, California, USA, January 2006. SPIE.

[193] Luis Pérez-Freire, Fernando Pérez-González, and Teddy Furon. On achievable security levels for lattice data hiding in the Known Message Attack scenario. In *ACM Multimedia and Security Workshop*, pages 68–79, Geneva, Switzerland, September 2006.

[194] 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*, 1(4):421–439, December 2006.

[195] 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 on Signal Processing for Data Hiding in Digital Media & Secure Content Delivery.

[196] Fernando Pérez-González, Juan R. Hernández, and Félix Balado. Approaching the capacity limit in image watermarking: A perspective on coding techniques for data hiding applications. *Signal Processing*, 81(6):1215–1238, June 2001. Special Section on Information Theoretic Aspects of Digital Watermarking.

[197] Fernando Pérez-González, Carlos Mosquera, Mauro Barni, and Andrea Abrardo. Rational Dither Modulation: a high-rate data-hiding method robust to gain attacks. *IEEE Transactions on Signal Processing*, 53(10):3960–3975, October 2005. Third supplement on secure media.

[198] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding-a survey. *Proceedings of the IEEE*, 87(7):1062–1078, 1999.

[199] H. Vincent Poor. *An introduction to signal detection and estimation.* Springer, New York, second edition, 1998.

[200] John G. Proakis. *Digital Communications.* McGraw-Hill, New York, 4th edition, 2001.

[201] R. Anthony Reese. Extreme lawsuits. *IEEE Spectrum*, 40(5):23–25, May 2003.

[202] Halsey L. Royden. *Real analysis.* Prentice Hall, 3rd edition, 1988.

[203] Joseph J. K. Ó Ruanaidh and Thierry Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, May 1998.

[204] Leonard Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, 12:162–165, December 1964.

[205] Claude E. Shannon. Communication theory of secrecy systems. *Bell system technical journal*, 28:656–715, October 1949.

[206] Douglas R. Stinson. *Cryptography: Theory and Practice.* Chapman and Hall/CRC, Boca Raton, FL, USA, 2nd edition, 2002.

[207] Jos F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11-12(1-4):625–653, 1999. Version 1.1 of the toolbox available at `http://sedumi.mcmaster.ca/`.

[208] Jonathan K. Su and Bernd Girod. Power-spectrum condition for energy-efficient watermarking. *IEEE Transactions on Multimedia*, 4(4):551–560, 2002.

[209] Mitchell D. Swanson, Bin Zhu, and Ahmed H. Tewfik. Robust data hiding for images. In *IEEE Digital Signal Processing Workshop*, pages 37–40, Loen, Norway, September 1996.

[210] Mitchell D. Swanson, Bin Zhu, and Ahmed H. Tewfik. Multiresolution scene-based video watermarking using perceptual models. *IEEE Journal on Selected Areas in Communications*, 16(4):540–550, May 1998.

[211] Mercan Topkara, Cuneyt Taskiran, and Edward J. Delp. Natural language watermarking. In Edward J. Delp and Ping Wah Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 441–452, San Jose, CA, USA, January 2005.

[212] Wade Trappe, Min Wu, Z. Jane Wang, and K. J. Ray Liu. Anti-collusion fingerprinting for multimedia. *IEEE Transactions on Signal Processing*, 51(4):1069.1087, April 2003.

[213] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne. A digital watermark. In *IEEE International Conference on Image Processing, ICIP'94*, volume 2, pages 86–89, Austin, Texas, USA, 1994.

[214] Harry L. van Trees. *Detection, Estimation, and Modulation Theory.* John Wiley and Sons, 1968.

[215] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, March 1996.

[216] Renato Villán, Sviatoslav Voloshynovskiy, Oleksiy Koval, Jose Emilio Vila-Forcén, Emre Topak, Frédéric Deguillaume, Yuri Rytsar, and Thierry Pun. Text data-hiding for digital and printed documents: Theoretical and practical considerations. In *Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, San Jose, CA, USA, 15-19 January 2006. SPIE.

[217] Sviatoslav Voloshynovskiy, Shelby Pereira, V. Iquise, and Thierry Pun. Attack modeling: Towards a second generation benchmark. *Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking*, 81(6):1177–1214, June 2001.

[218] Svyatoslav Voloshynovskiy, Frédéric Deguillaume, and Thierry Pun. Multibit digital watermarking robust against local nonlinear geometrical distortions. In *IEEE International Conference on Image Processing, ICIP'01*, volume 3, pages 999–1002, Thessaloniki, Greece, October 2001.

[219] Olga Vybornova and Benoit Macq. Natural language watermarking and robust hashing based on presuppositional analysis. In *IEEE International Conference on Information Reuse and Integration, IRI'07*, pages 177–182, Las Vegas, Nevada, USA, August 2007.

[220] Éric Walter and Hélène Piet-Lahanier. Exact recursive polyhedral description of the feasible parameter set for bounded-error models. *IEEE Transactions on Automatic Control*, 34(8):911–915, August 1989.

[221] Andrew B. Watson. Dct quantization matrices visually optimized for individual images. In Jan P. Allebach and Bernice E. Rogowitz, editors, *Human Vision, Visual Processing, and Digital Display IV*, volume 1913, pages 202–216, San Jose, CA, USA, February 1992. SPIE.

[222] Ram Zamir and Meir Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, July 1996.

[223] 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, *2nd International Workshop on Information Hiding, IH'98*, volume 1525 of *Lecture Notes in Computer Science*, pages 344–354, Portland, OR, USA, April 1998. Springer-Verlag.

Appendix A

A.1 Proof of Lemma 2.1

The proof follows directly from the definition of concavity, by taking into account that $g(\cdot)$ can take only integer arguments. The function $g(n)$ is concave if and only if

$$g(\lambda n_1 + (1-\lambda)n_2) \geq \lambda g(n_1) + (1-\lambda)g(n_2), \tag{A.1}$$

for all $n_1, n_2 \in \mathbb{Z}$ and $\lambda \in [0,1]$ such that $\lambda n_1 + (1-\lambda)n_2 \in \mathbb{Z}$. Let us assume, without loss of generality, that $n_1 < n_2 = n_1 + k$. Hence, the valid values for $\lambda n_1 + (1-\lambda)n_2$ are given by

$$\lambda_i n_1 + (1-\lambda_i)n_2 = n_1 + i, \text{ with } i = 0, \ldots, k \ , \tag{A.2}$$

where $\lambda_i = \frac{k-i}{k}$. Substituting into Eq. (A.1) and operating, we get the following equivalent condition for concavity:

$$g(n_1 + i) \geq \frac{k-i}{k} \cdot g(n_1) + \frac{i}{k} \cdot g(n_1 + k), \tag{A.3}$$

which in turn can be rewritten as

$$\sum_{j=0}^{i-1} \Delta g(n_1 + j) \geq \frac{i}{k} \sum_{j=0}^{k-1} \Delta g(n_1 + j). \tag{A.4}$$

Eq. (A.4) must hold for all $k \geq i$; in particular, for $k = 2$ we have

$$\Delta g(n_1) \geq \frac{1}{2}(\Delta g(n_1) + \Delta g(n_1 + 1)), \tag{A.5}$$

which implies that

$$\Delta g(n) \geq \Delta g(n+1) \ \forall \ n. \tag{A.6}$$

For proving sufficiency, assume now that $\Delta g(n) \geq \Delta g(n+1)$ for all n. If this condition holds, then it is easy to see that the partial averages satisfy the following inequality:

$$\frac{1}{i} \sum_{j=0}^{i-1} \Delta g(n_1 + j) \geq \frac{1}{k} \sum_{j=0}^{k-1} \Delta g(n_1 + j), \tag{A.7}$$

which is tantamount to (A.4).

A.2 Proof of Lemma 2.2

To see that the mutual information is always increasing, consider the function

$$\Delta I(N) = I(\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}; \psi(\mathbf{\Theta})) - I(\mathbf{O}_1, \ldots, \mathbf{O}_N; \psi(\mathbf{\Theta})), \tag{A.8}$$

which is nothing but the average information about $\mathbf{t}$ that is gained with the $(N+1)$th observation. Such function is easily seen to be always non-negative:

$$\Delta I(N) = h(\psi(\mathbf{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_N) - h(\psi(\mathbf{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}) \geq 0, \tag{A.9}$$

where (A.9) follows from the fact that conditioning reduces entropy [84].

By Lemma 2.1, the mutual information will be (strictly) concave iff $\Delta I(N)$ is (decreasing) non-increasing. By the definition of mutual information, we have

$$I(\mathbf{O}_1, \ldots, \mathbf{O}_N; \psi(\mathbf{\Theta})) = h(\mathbf{O}_1, \ldots, \mathbf{O}_N) - h(\mathbf{O}_1, \ldots, \mathbf{O}_N|\psi(\mathbf{\Theta})). \tag{A.10}$$

Using (A.10) we can write

$$\begin{aligned} \Delta I(N+1) &= I(\mathbf{O}_1, \ldots, \mathbf{O}_{N+2}; \psi(\mathbf{\Theta})) - I(\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}; \psi(\mathbf{\Theta})) \\ &= h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+2}) - h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+2}|\psi(\mathbf{\Theta})) \\ &- h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}) + h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}|\psi(\mathbf{\Theta})). \end{aligned} \tag{A.11}$$

By the chain rule for entropies [84, Sect 9.6], it follows that

$$h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+2}) = h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}) + h(\mathbf{O}_{N+2}|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}). \tag{A.12}$$

Likewise, it is straightforward to see that

$$h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+2}|\psi(\mathbf{\Theta})) = h(\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}|\psi(\mathbf{\Theta})) + h(\mathbf{O}_{N+2}|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}, \psi(\mathbf{\Theta})). \tag{A.13}$$

Hence, by combining (A.12) and (A.13), $\Delta I(N+1)$ can be rewritten as

$$\begin{aligned} \Delta I(N+1) &= h(\mathbf{O}_{N+2}|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}) - h(\mathbf{O}_{N+2}|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}, \psi(\mathbf{\Theta})) \\ &= h(\mathbf{O}_{N+2}|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}) - h(\mathbf{O}_{N+2}|\psi(\mathbf{\Theta})), \end{aligned} \tag{A.14}$$

where the last equality follows from the conditional independence of the observations given $\psi(\mathbf{\Theta})$. Using similar reasonings, it can be seen that

$$\Delta I(N) = h(\mathbf{O}_{N+1}|\mathbf{O}_1, \ldots, \mathbf{O}_N) - h(\mathbf{O}_{N+1}|\psi(\mathbf{\Theta})). \tag{A.15}$$

Since the $\mathbf{O}_i$ are i.i.d, $h(\mathbf{O}_{N+2}|\psi(\mathbf{\Theta})) = h(\mathbf{O}_{N+1}|\psi(\mathbf{\Theta}))$, so we finally have that

$$\Delta I(N+1) - \Delta I(N) = h(\mathbf{O}_{N+2}|\mathbf{O}_1, \ldots, \mathbf{O}_{N+1}) - h(\mathbf{O}_{N+1}|\mathbf{O}_1, \ldots, \mathbf{O}_N) \leq 0. \tag{A.16}$$

This concludes the proof of the lemma.

Appendix B

B.1 Bounds on the information leakage in the WOA scenario for add-SS

For obtaining the bounds to the information leakage, we will resort to the expression given by (3.12). Since its first term has been already computed in (3.11), we will focus on the remaining terms. For a fair user with perfect knowledge of $\mathbf{S}$, the observations are all mutually independent. Hence, the third term can be rewritten as

$$I(\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}; M_1, \ldots, M_{N_o}|\mathbf{S})_{\text{add-SS}} = N_o \cdot I(\mathbf{Y}_1; M_1|\mathbf{S}). \tag{B.1}$$

For computing (B.1) we take advantage of the fact that the statistic $\mathbf{Y}_i^T\mathbf{S}$ is a sufficient statistic for decoding M_i when $\mathbf{S}$ is known by the decoder [199]. Thus, we have

$$\begin{aligned} I(\mathbf{Y}_1; M_1|\mathbf{S}) &= I(\mathbf{Y}_1^T\mathbf{S}; M_1|\mathbf{S}) = h(\mathbf{Y}_1^T\mathbf{S}|\mathbf{S}) - h(\mathbf{Y}_1^T\mathbf{S}|M_1, \mathbf{S}) \\ &= h(\mathbf{Y}_1^T\mathbf{S}|\mathbf{S}) - \frac{1}{2}\left(\log(2\pi e\sigma_X^2) + E[\log(\|\mathbf{S}\|^2)]\right), \end{aligned} \tag{B.2}$$

where we have used that $\mathbf{Y}_1^T\mathbf{S}|M_1 = m_1, \mathbf{S} = \mathbf{s} \sim \mathcal{N}(\|\mathbf{s}\|^2(-1)^{m_1}, \|\mathbf{s}\|^2\sigma_X^2)$. For the term $h(\mathbf{Y}^T\mathbf{S}|\mathbf{S})$ we must take into account that

$$\mathbf{Y}_1^T\mathbf{S}|\mathbf{S} = \mathbf{s} \sim \frac{1}{2}\left(\mathcal{N}(\|\mathbf{s}\|^2, \|\mathbf{s}\|^2\sigma_X^2) + \mathcal{N}(-\|\mathbf{s}\|^2, \|\mathbf{s}\|^2\sigma_X^2)\right),$$

and that $h(\mathbf{Y}_1^T\mathbf{S}|\mathbf{S}) = E\left[h(\mathbf{Y}_1^T\mathbf{S}|\mathbf{S} = \mathbf{s})\right]$, where the expectation is taken over $\mathbf{S}$. The computation of the expectations in (B.2) requires numerical integration, taking into account that $||\mathbf{S}||^2 \sim \chi^2(n, \sigma_S)$. The second term of (3.12) is the only one that remains to be addressed. Notice that

$$\begin{aligned}
&I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o})_{\text{add-SS}}\\
&\quad= N_o\cdot H(M_1)-H(M_1,\ldots,M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o})\\
&\quad= N_o\cdot H(M_1)-\sum_{i=1}^{N_o}H(M_i|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},M_1,\ldots,M_{i-1})\\
&\quad\leq N_o\cdot H(M_1)-N_o\cdot H(M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},M_1,\ldots,M_{N_o-1}) \qquad \text{(B.3)}\\
&\quad= N_o\cdot I(\mathbf{Y}_{N_o};M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o-1},M_1,\ldots,M_{N_o-1}), \qquad \text{(B.4)}
\end{aligned}$$

where the inequality (B.3) follows from the fact that conditioning reduces entropy [84]. Conditioned on a particular realization of the observations, we have that $\mathbf{Y}_{N_o}=\mathbf{X}_{N_o}+(-1)^{M_{N_o}}\cdot\bar{\mathbf{S}}_{N_o-1}$, with $\bar{\mathbf{S}}_{N_o-1}$ a random variable that follows the distribution of $\mathbf{S}$ conditioned on the N_o-1 past observations,[1] i.e. $\bar{\mathbf{S}}_{N_o}\sim\mathcal{N}(\mathbf{v},\sigma^2_{\bar{S}_{N_o-1}}\mathbf{I}_n)$, with $\mathbf{v}$ and $\sigma^2_{\bar{S}_{N_o-1}}$ given by (3.8) and (3.9), respectively. Hence, considering a random variable $\mathbf{N}\sim\mathcal{N}(\mathbf{0},\sigma^2_{\bar{S}_{N_o-1}}\mathbf{I}_n)$, and noticing that $(-1)^{M_{N_o}}\mathbf{N}$ is identically distributed to $\mathbf{N}$, then we can write

$$\mathbf{Y}_{N_o}=\mathbf{X}_{N_o}+(-1)^{M_{N_o}}(\mathbf{N}+\mathbf{v})=\bar{\mathbf{X}}_{N_o}+(-1)^{M_{N_o}}\mathbf{v}, \qquad \text{(B.5)}$$

where $\bar{\mathbf{X}}_{N_o}\triangleq\mathbf{X}_{N_o}+\mathbf{N}\sim\mathcal{N}(\mathbf{0},(\sigma^2_X+\sigma^2_{\bar{S}_{N_o-1}})\mathbf{I}_n)$. Clearly, this implies that

$$I(\mathbf{Y}_{N_o};M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o},M_1,\ldots,M_{N_o-1})_{\text{add-SS}}=I(\mathbf{Y}_{N_o};M_{N_o}|\mathbf{V}_{N_o}), \qquad \text{(B.6)}$$

where the components of $\mathbf{V}_{N_o}$ are the realizations of (3.8). Since $\boldsymbol{\mu}^T\mathbf{y}^{(i)}$ is zero-mean Gaussian with variance $(N_o-1)\cdot(\sigma^2_X+(N_o-1)\sigma^2_S)$, then $\mathbf{V}_{N_o}\sim\mathcal{N}\left(\mathbf{0},\frac{(N_o-1)\sigma^4_S}{(N_o-1)\sigma^2_S+\sigma^2_X}\mathbf{I}_n\right)$. Hence, (B.6) becomes

$$\begin{aligned}
&I(\mathbf{Y}_{N_o};M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o-1},M_1,\ldots,M_{N_o-1})_{\text{add-SS}}\\
&\quad= I(\mathbf{Y}^T_{N_o}\mathbf{V}_{N_o};M_{N_o}|\mathbf{V}_{N_o})\\
&\quad= h(\mathbf{Y}^T_{N_o}\mathbf{V}_{N_o}|\mathbf{V}_{N_o})-h(\mathbf{Y}^T_{N_o}\mathbf{V}_{N_o}|M_{N_o},\mathbf{V}_{N_o})\\
&\quad= h(\mathbf{Y}^T_{N_o}\mathbf{V}|\mathbf{V}_{N_o})-E\left[\frac{1}{2}\log\left(2\pi e(\sigma^2_X+\sigma^2_{\bar{S}_{N_o}})\|\mathbf{V}_{N_o}\|^2\right)\right], \qquad \text{(B.7)}
\end{aligned}$$

where we have used again the fact that $\mathbf{Y}^T_{N_o}\mathbf{V}_{N_o}$ is a sufficient statistic for decoding (see [199]). Notice that the first term in the right hand side of (B.2) and (B.7) must be computed by means of numerical integration. Finally, combining (3.12) with (3.11), (B.1), (B.2), (B.4) and (B.7) we arrive at (3.13), the final expression of the upper bound.

[1] This is due to the unconditional independence between the hosts and messages of different observations.

A lower bound on the mutual information can be obtained by taking into account that

$$\begin{aligned} I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o})_{\text{add-SS}} & \\ &= \sum_{i=1}^{N_o} I(\mathbf{Y}_i;M_1,\ldots,M_{N_o}|\mathbf{Y}_1,\ldots,\mathbf{Y}_{i-1}) \\ &= \sum_{i=1}^{N_o}\sum_{j=1}^{N_o} I(\mathbf{Y}_i;M_j|\mathbf{Y}_1,\ldots,\mathbf{Y}_{i-1},M_1,\ldots,M_{j-1}) \\ &\geq \sum_{i=2}^{N_o} I(\mathbf{Y}_i;M_i|\mathbf{Y}_1,\ldots,\mathbf{Y}_{i-1},M_1,\ldots,M_{i-1}). \qquad \text{(B.8)} \end{aligned}$$

Note that each term of (B.8) was already computed in (B.7). The final expression of the lower bound is given in (3.14).

B.2 Proof of Theorem 3.1

An straightforward upper bound to the loss function is given by

$$\delta(N_o)_{\text{add-SS}} \leq N_o\log(2) - I(\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_o};M_1,\ldots,M_{N_o}). \qquad \text{(B.9)}$$

Using the lower bound (B.8) for the second term in the right hand side of (B.9) and recalling that the loss function is always non-negative, the latter can be bounded as

$$N_o\log(2) - \sum_{i=2}^{N_o} I(\mathbf{Y}_i;M_i|\mathbf{V}_i) \geq \delta(N_o)_{\text{add-SS}} \geq 0,\ \text{for } N_o \geq 2, \qquad \text{(B.10)}$$

where $\mathbf{Y}_i = \bar{\mathbf{X}}_i + (-1)^{M_i}\mathbf{V}_i$, $\mathbf{V}_i \sim \mathcal{N}(\mathbf{0}, \frac{(i-1)\sigma_S^4}{(i-1)\sigma_S^2+\sigma_X^2}\mathbf{I}_n)$, and $\bar{\mathbf{X}}_i \sim \mathcal{N}(\mathbf{0},(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)\mathbf{I}_n)$ with $\sigma_{\bar{S}_{i-1}}^2$ given by (3.9). The first term in the right hand side of (B.10) can be rewritten as $I(\mathbf{Y}_i;M_i|\mathbf{V}_i) = H(M_i) - H(M_i|\mathbf{Y}_i,\mathbf{V}_i)$. By inserting this expression into (B.10), we obtain

$$\log(2) + \sum_{i=2}^{N_o} H(M_i|\mathbf{Y}_i,\mathbf{V}_i) \geq \delta(N_o)_{\text{add-SS}} \geq 0, \qquad \text{(B.11)}$$

Now we focus on the second term in the right hand side of (B.11). Using Fano's inequality [84], the conditional entropy of M_i can be bounded from above as

$$H(M_i|\mathbf{Y}_i,\mathbf{V}_i) \leq H(P_e|\mathbf{V}_i) + P_e\cdot\log(|\mathcal{M}|-1), \qquad \text{(B.12)}$$

where $P_e|\mathbf{V}_i$ denotes the probability of decoding error conditioned on $\mathbf{V}_i$. In our case ($|\mathcal{M}|$=2), ineq. (B.12) is reduced to

$$H(M_i|\mathbf{Y}_i,\mathbf{V}_i) \leq H(P_e|\mathbf{V}_i), \tag{B.13}$$

where $H(P_e)$ is the binary entropy function [84], i.e. $H(P_e) = P_e \log(P_e) + (1-P_e)\log(1-P_e)$. Hence, (B.10) is rewritten as

$$\log(2) + \sum_{i=2}^{N_o} H(P_e|\mathbf{V}_i) \geq \delta(N_o)_{\text{add-SS}} \geq 0, \tag{B.14}$$

For computing the error probability we define the scalar random variable $Z_i \triangleq \mathbf{Y}_i^T\mathbf{V}_i$, which is a sufficient statistic for the decoding of M_i. The statistic of Z_i conditioned on $\mathbf{V}_i = \mathbf{v}_i$ is

$$\begin{aligned} Z_i|\mathbf{V}_i = \mathbf{v}_i \sim \frac{1}{2}\Big(&\mathcal{N}\left(-||\mathbf{v}_i||^2, (\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)||\mathbf{v}_i||^2\right) \\ &+ \mathcal{N}\left(||\mathbf{v}_i||^2, (\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)||\mathbf{v}_i||^2\right)\Big). \end{aligned} \tag{B.15}$$

Eq. (B.13) holds for any decoder, and in particular for a decoder based on sign-decision. For the latter, $P_e|\mathbf{v}_i = \Pr\{Z_i > 0|M_i = 1, \mathbf{V}_i = \mathbf{v}_i\} = \Pr\{Z_i \leq 0|M_i = -1, \mathbf{V}_i = \mathbf{v}_i\}$. Obviously, from (B.15), we have

$$\Pr\{Z_i > 0|M_i = 1, \mathbf{V}_i = \mathbf{v}_i\} = Q\left(\frac{t_i^{\frac{1}{2}}}{(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)^{\frac{1}{2}}}\right), \tag{B.16}$$

where $Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt$ denotes the Gaussian Q-function, and $T_i \triangleq ||\mathbf{V}_i||^2 \sim \chi^2(n,\sigma_{V_i})$, with $\sigma_{V_i}^2 = \frac{(i-1)\sigma_S^4}{(i-1)\sigma_S^2+\sigma_X^2}$. Using the well known Chernoff bound [200], we can upper bound the error probability (B.16) as

$$\Pr\{Z_i > 0|M_i = 1, \mathbf{V}_i = \mathbf{v}_i\} \leq \frac{1}{2}\exp\left(\frac{-t_i}{2(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)}\right). \tag{B.17}$$

Since $H(P_e)$ is increasing in $P_e \in [0, 0.5]$, (B.17) can be used to upper bound $H(P_e|\mathbf{V}_i = \mathbf{v}_i)$. Hence,

$$\begin{aligned} H(P_e|\mathbf{V}_i) &\leq E\left[H\left(\frac{1}{2}\exp\left(\frac{-T_i}{2(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)}\right)\right)\right] \\ &\leq H\left(\frac{1}{2}E\left[\exp\left(\frac{-T_i}{2(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)}\right)\right]\right), \end{aligned} \tag{B.18}$$

where the second inequality follows from Jensen's inequality [84]. The expectation above can be computed by taking into account the pdf of the Chi-square distribution [200]:

$$E\left[\exp\left(\frac{-T_i}{2(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)}\right)\right]$$

$$= \left(\sigma_{V_i}^n \cdot 2^{\frac{n}{2}-1} \cdot \Gamma(n/2)\right)^{-1} \int_0^\infty \exp\left(-t\frac{\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2+\sigma_{V_i}^2}{2\sigma_{V_i}^2(\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2)}\right) \cdot t^{\frac{n}{2}-1}dt. \tag{B.19}$$

Identifying $k = \frac{\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2+\sigma_{V_i}^2}{\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2}$ and using the variable change $r = tk$, the integral in (B.19) can be written as

$$\int_0^\infty \exp(-tk) \cdot t^{\frac{n}{2}-1}dt = k^{-\frac{n}{2}} \int_0^\infty \exp(-r)r^{\frac{n}{2}-1}dr = k^{-\frac{n}{2}} \cdot \Gamma(n/2). \tag{B.20}$$

Substituting successively into (B.19) and (B.18) we arrive at

$$H(P_e|\mathbf{V}_i) \leq H\left(\frac{1}{2}\left(\frac{\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2}{\sigma_X^2+\sigma_{\bar{S}_{i-1}}^2+\sigma_{V_i}^2}\right)^{\frac{n}{2}}\right) \triangleq H\left(\frac{\tau_i^{\frac{n}{2}}}{2}\right). \tag{B.21}$$

By combining (B.21) with (B.14), (3.16) is obtained. Now, the asymptotic results of Theorem 3.1 follow easily:

1. For proving the first result, we rewrite the term τ_i in (B.21), after some algebraic manipulations, as

$$\tau_i = \frac{i\sigma_S^2\sigma_X^2+\sigma_X^4}{(i-1)\sigma_S^4+i\sigma_S^2\sigma_X^2+\sigma_X^4} = \frac{1}{1+\frac{(i-1)\sigma_S^4}{i\sigma_S^2\sigma_X^2+\sigma_X^4}}. \tag{B.22}$$

   In the limit, we have $\lim_{\sigma_S\to\infty} H\left(\frac{1}{2}\tau_i^{\frac{n}{2}}\right) = \lim_{\sigma_X\to 0} H\left(\frac{1}{2}\tau_i^{\frac{n}{2}}\right) = 0$ for $i \geq 2$. Hence,

$$\lim_{\text{DWR}\to-\infty} \delta(N_o)_{\text{add-SS}} \leq \log(2).$$

   This proves the first asymptotic result of Theorem 3.1.

2. The term τ_i in (B.21) is strictly smaller than 1; hence, (B.21) decreases with n, in such a way that $\lim_{n\to\infty} H\left(\frac{1}{2}\tau_i^{\frac{n}{2}}\right) = 0$. Thus,

$$\lim_{n\to\infty} \delta(N_o)_{\text{add-SS}} \leq \log(2).$$

   Since $\delta(N_o)_{\text{add-SS}} \geq 0$, according to (B.14), we have proved the second asymptotic result of Theorem 3.1.

B.3 Proof of Eq. (3.24) in Theorem 3.2

For computing the information leakage for ISS, we first need to prove the following lemma.

Lemma B.1 (Gaussianity of the marked signal in ISS). For the ISS embedding function, $\mathbf{Y}_i$ conditioned on the embedded message M_i is i.i.d. Gaussian.

Proof: For $N_o = 1$, $\mathbf{Y}_i$ is the sum of an i.i.d. Gaussian ($\mathbf{S}$) and another Gaussian whose covariance matrix depends on $\mathbf{S}$. Noticing that the sum of Gaussian random variables is Gaussian, we have that $\mathbf{Y}_i$ is Gaussian. If a Gaussian random variable is circularly symmetric, then it is i.i.d. Thus, it remains to be proved that the pdf of $\mathbf{Y}$ is circularly symmetric, i.e. that its pdf is invariant under rotations, or equivalently, $f(\mathbf{y}|M=0) = f(\mathbf{Hy}|M=0)$, for $\mathbf{H} \in \mathbb{R}^{n\times n}$ any unitary matrix.

In the following we drop the subindex i from the notation for simplicity, and we assume that the embedded message is $m = 0$ without loss of generality. The pdf of $\mathbf{y}$ conditioned on $\mathbf{s}$ is

$$f(\mathbf{y}|\mathbf{S}=\mathbf{s}, M=0) = \frac{1}{(2\pi)^{\frac{1}{2}}|\mathbf{\Sigma_S}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\nu\mathbf{s})^T\mathbf{\Sigma_S}^{-1}(\mathbf{y}-\nu\mathbf{s})\right), \qquad \text{(B.23)}$$

and the pdf $f(\mathbf{Hy}|M=0)$ is calculated as

$$\begin{aligned} f(\mathbf{Hy}|M=0) &= \int_{\mathbb{R}^n} f(\mathbf{Hy}|\mathbf{S}=\mathbf{s}, M=0)\cdot f(\mathbf{s})d\mathbf{s} \\ &= K\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left((\mathbf{Hy}-\nu\mathbf{s})^T\mathbf{\Sigma_S}^{-1}(\mathbf{Hy}-\nu\mathbf{s}) + \frac{||\mathbf{s}||^2}{2\sigma_S^2}\right)\right)d\mathbf{s}, \text{(B.24)} \end{aligned}$$

where K is a constant (notice, from Eq. (3.22), that $|\mathbf{\Sigma_S}|$ does not depend on $\mathbf{s}$). The first term of the exponent in (B.24) can be rewritten as

$$\begin{aligned} (\mathbf{Hy}-\nu\mathbf{s})^T\mathbf{\Sigma_S}^{-1}(\mathbf{Hy}-\nu\mathbf{s}) &= \mathbf{y}^T\mathbf{H}^T\mathbf{\Sigma_S}^{-1}\mathbf{Hy} - 2\nu\mathbf{s}^T\mathbf{\Sigma_S}^{-1}\mathbf{Hy} + \nu^2\mathbf{s}^T\mathbf{\Sigma_S}^{-1}\mathbf{s} \\ &= \mathbf{y}^T\mathbf{H}^T\mathbf{\Sigma_S}^{-1}\mathbf{Hy} - 2\nu\mathbf{s}^T\mathbf{HH}^T\mathbf{\Sigma_S}^{-1}\mathbf{Hy} \\ &+ \nu^2\mathbf{s}^T\mathbf{HH}^T\mathbf{\Sigma_S}^{-1}\mathbf{HH}^T\mathbf{s} \\ &= (\mathbf{y}-\nu\mathbf{H}^T\mathbf{s})^T(\mathbf{H}^T\mathbf{\Sigma_S}\mathbf{H})^{-1}(\mathbf{y}-\nu\mathbf{H}^T\mathbf{s}), \qquad \text{(B.25)} \end{aligned}$$

where the second equality follows because $\mathbf{HH}^T = \mathbf{I}_n$. By realizing that $||\mathbf{s}||^2 = ||\mathbf{H}^T\mathbf{s}||^2$ (or equivalently, that $f(\mathbf{s})$ is circularly symmetric), we can write

$$\begin{aligned}
&f(\mathbf{H}\mathbf{y}|M=0)\\
&\quad= \int_{\mathbb{R}^n} f(\mathbf{y}|\mathbf{H}^T\mathbf{S}=\mathbf{H}^T\mathbf{s}, M=0)\cdot f(\mathbf{H}^T\mathbf{s})d\mathbf{s}\\
&\quad= K\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left((\mathbf{y}-\nu\mathbf{H}^T\mathbf{s})^T(\mathbf{H}^T\mathbf{\Sigma_S}\mathbf{H})^{-1}(\mathbf{y}-\nu\mathbf{H}^T\mathbf{s})+\frac{||\mathbf{H}^T\mathbf{s}||^2}{2\sigma_S^2}\right)\right)d\mathbf{s}\\
&\quad= K\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left((\mathbf{y}-\nu\mathbf{s}')^T\mathbf{\Sigma}_{\mathbf{S}'}^{-1}(\mathbf{y}-\nu\mathbf{s}')+\frac{||\mathbf{s}'||^2}{2\sigma_S^2}\right)\right)d\mathbf{s}'\\
&\quad= f(\mathbf{y}|M=0), \qquad (B.26)
\end{aligned}$$

where for the third equality we have used the change of variable $\mathbf{s}'=\mathbf{H}^T\mathbf{s}$ (notice that the Jacobian of this change of variable is 1). This concludes the proof of the lemma. ■

By Lemma B.1, the entropy of $\mathbf{Y}_j$ conditioned on M_j is

$$h(\mathbf{Y}_j|M_j)=\frac{1}{2}\left(\frac{n}{2}\log(2\pi e\sigma_{Y_0}^2)+\frac{n}{2}\log(2\pi e\sigma_{Y_1}^2)\right), \qquad (B.27)$$

with $\sigma_{Y_0}^2$ and $\sigma_{Y_1}^2$ the variance of the components of $\mathbf{Y}_j$ conditioned on $M_j=0$ and $M_j=1$, respectively. Let $Y_{j,i}$ be the ith component of the jth observation,

$$Y_{j,i}=X_{j,i}+(-1)^{M_j}\nu S_i-\lambda\frac{\mathbf{X}_j^T\mathbf{S}}{\|\mathbf{S}\|^2}S_i,\ \text{for } i=1,\ldots,n. \qquad (B.28)$$

Since the components of $\mathbf{X}_j$ are zero-mean, it is easy to see that $E[Y_{j,i}|M_j=m_j]=0$. Thus, the variance of $Y_{j,i}$ conditioned on M_j is given by

$$\begin{aligned}
E\left[Y_{j,i}^2|M_j=m_j\right] &= \sigma_X^2+\nu^2\sigma_S^2-2\lambda\sigma_X^2\cdot E\left[\left(\frac{S_i}{\|\mathbf{S}\|}\right)^2\right]+\lambda^2\cdot E\left[\frac{(\mathbf{X}_j^T\mathbf{S})^2S_i^2}{\|\mathbf{S}\|^4}\right]\\
&= \sigma_X^2+\nu^2\sigma_S^2-2\lambda\sigma_X^2\cdot E\left[\left(\frac{S_i}{\|\mathbf{S}\|}\right)^2\right]\\
&+ \lambda^2\sigma_X^2\left(E\left[\left(\frac{S_i}{\|\mathbf{S}\|}\right)^4\right]\right.\\
&+ \left.E\left[\sum_{l=1,l\neq i}^{n}\left(\frac{S_l}{\|\mathbf{S}\|}\right)^2\left(\frac{S_i}{\|\mathbf{S}\|}\right)^2\right]\right). \qquad (B.29)
\end{aligned}$$

We need to compute the second and fourth order statistics of $S_i/\|\mathbf{S}\|$. For a vector $\mathbf{S}$ isotropically distributed (i.e., with its probability density function invariant under rotations), the random variable defined as $\mathbf{S}'\triangleq\frac{\mathbf{S}}{\|\mathbf{S}\|}\cdot r$ is uniformly distributed on the surface of the n-dimensional sphere of radius r. Notice that the Gaussian vector with i.i.d. components, which is the case of interest for us, is indeed isotropically distributed. The marginal probability density function of one component S_i' can be computed by integrating the probability density function of $\mathbf{S}'$ over the surface of the $(n-1)$-dimensional sphere with radius $r\sqrt{1-(s_i'/r)^2}$, yielding[2]

$$f(s_i') = \frac{\Gamma\left(\frac{n}{2}\right)}{r\cdot\sqrt{\pi}\cdot\Gamma\left(\frac{n-1}{2}\right)}\left(1-\left(\frac{s_i}{r}\right)^2\right)^{\frac{n-3}{2}}, \ \forall\, i=1,\ldots,n;\ s_i'\in[-r,r], \tag{B.30}$$

where $\Gamma(\cdot)$ denotes the complete Gamma function. Due to the symmetry of the pdf, it is easy to see that $E[S_i']=0$, and the moment generating function $\int_{-r}^{r}(s_i')^p f(s_i')ds_i'$ yields $E[(S_i')^2]=r^2/n$ and $E[(S_i')^4]=3r^4/(2n+n^2)$. For computing the other statistic involved in (B.29), the joint pdf $f(s_i',s_j'),\ i\neq j$ must be calculated. Conditioned on $S_i'=s_i'$, the remaining components of $\mathbf{S}'$ are uniformly distributed over the surface of a $(n-1)$-dimensional sphere of radius $r_i = r\sqrt{1-(s_i'/r)^2}$. Hence, the conditional marginal pdf of S_j' is given by

$$\begin{aligned} &f(s_j'|S_i'=s_i') \\ &\quad= \frac{\Gamma\left(\frac{n-1}{2}\right)}{r_i\cdot\sqrt{\pi}\cdot\Gamma\left(\frac{n-2}{2}\right)}\left(1-\left(\frac{s_j'}{r_i}\right)^2\right)^{\frac{n-4}{2}}, \ \forall\, j=1,\ldots,n;\ j\neq i;\ s_j'\in[-r_i,r_i]. \end{aligned} \tag{B.31}$$

After some algebraic simplifications, we arrive at the following expression for the joint pdf:

$$f(s_i',s_j') = \frac{n-2}{2\pi}\cdot\frac{r^2\left(1-\left(\frac{s_i'}{r}\right)^2\right)^{\frac{n}{2}}\left(1-\left(\frac{s_j'}{r_i}\right)^2\right)^{\frac{n}{2}}}{(s_i')^2+(s_j')^2-r^2}. \tag{B.32}$$

Now we can calculate

$$E[(S_i')^k\cdot(S_j')^k] = \int_{-r}^{r}\int_{-r_i}^{r_i}(s_i')^k\cdot(s_j')^k f(s_i',s_j')ds_j'ds_i', \tag{B.33}$$

finding out that S_i' and S_j' are uncorrelated, but $E[(S_i')^2\cdot(S_j')^2]=r^4/(2n+n^2)$. Substituting in (B.29) with $r=1$, which is the case of interest for us, we obtain

$$\begin{aligned} E\left[Y_{j,i}^2|M_j=m_j\right] &= \sigma_X^2+\nu^2\sigma_S^2-2\lambda\frac{\sigma_X^2}{n}+\lambda^2\sigma_X^2\left(\frac{3}{2n+n^2}+\frac{n-1}{2n+n^2}\right) \\ &= \sigma_X^2+\nu^2\sigma_S^2+\sigma_X^2\frac{\lambda(\lambda-2)}{n}. \end{aligned} \tag{B.34}$$

[2] A similar calculus is performed in [42, Chapter 14.3.1] for obtaining the marginal pdf of the uniform inside the n-dimensional sphere.

Since (B.34) is independent of the actual value of M_j, it follows that (B.27) is given by

$$h(\mathbf{Y}_j|M_j) = \frac{n}{2}\log\left(2\pi e\left(\sigma_X^2 + \nu^2\sigma_S^2 + \sigma_X^2\frac{\lambda(\lambda-2)}{n}\right)\right).$$

Finally, combining this result with (3.23) for $N_o = 1$ and rearranging terms, we arrive at (3.24).

B.4 Upper bound to the entropy of the observations in the KMA scenario for ISS

An upper bound to the entropy of the observations can be derived as follows:

$$\begin{aligned} h(\mathbf{Y}_1,\dots,\mathbf{Y}_{N_o}|M_1,\dots,M_{N_o}) &\leq \sum_{i=1}^{n} h(Y_{1,i},\dots,Y_{N_o,i}|M_1,\dots,M_{N_o}) && \text{(B.35)}\\ &\leq \sum_{i=1}^{n}\frac{1}{2}E\left[\log\left((2\pi e)^{N_o}|\mathbf{\Sigma}_{\mathbf{Y}_i}|\right)\right], && \text{(B.36)} \end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{Y}_i}$ denotes the covariance matrix of $Y_{1,i},\dots,Y_{N_o,i}|M_1 = m_1,\dots,M_{N_o} = m_{N_o}$, and the expectation is taken over all possible realizations of the messages sequences. It can be seen that, for a particular realization , the off-diagonal terms of the covariance matrix are given by

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{Y}_i}(j,k) &= E\left[Y_{j,i}\cdot Y_{k,i}|M_1 = m_1,\dots,M_{N_o} = m_{N_o}\right] = (-1)^{m_j+m_k}\nu^2\cdot E\left[S_i^2\right]\\ &= (-1)^{m_j+m_k}\nu^2\sigma_S^2,\ j\neq k. && \text{(B.37)} \end{aligned}$$

The diagonal terms have been already calculated in (B.34). As can be seen, the covariance matrix $\mathbf{\Sigma}_{\mathbf{Y}_i}$ is of the form

$$\mathbf{\Sigma}_{\mathbf{Y}_i} = \begin{bmatrix} P+C & (-1)^{m_1+m_2}C & \cdots & (-1)^{m_1+m_{N_o}}C\\ (-1)^{m_2+m_1}C & P+C & \cdots & (-1)^{m_2+m_{N_o}}C\\ \vdots & \vdots & \ddots & \vdots\\ (-1)^{m_{N_o}+m_1}C & (-1)^{m_{N_o}+m_2}C & \cdots & P+C \end{bmatrix}, \qquad \text{(B.38)}$$

with $C = \nu^2\sigma_S^2$, and $P = \sigma_X^2(1+\frac{\lambda(\lambda-2)}{n})$. Taking into account that multiplying a row or a column of a matrix by a scalar multiplies the determinant by that scalar, it is not difficult to show that the determinant of the above matrix is equal to $P^{N_o}\left(1+\frac{N_oC}{P}\right)$, independently of the actual values of m_i. We can insert this result in Eq. (B.36), arriving at

$$h(\mathbf{Y}_1,\dots,\mathbf{Y}_{N_o}|M_1,\dots,M_{N_o})$$
$$\leq \frac{n}{2}\log\left((2\pi e)^{N_o}(\sigma_X^2)^{N_o}\left(1+\frac{\lambda(\lambda-2)}{n}\right)^{N_o}\left(1+\frac{N_o\nu^2\sigma_S^2}{\sigma_X^2\left(1+\frac{\lambda(\lambda-2)}{n}\right)}\right)\right). \quad \text{(B.39)}$$

This result is finally combined with (3.23) for obtaining (3.25).

B.5 Upper bounds to the conditional entropy and the log-expectation of the norm

In the KMA scenario, the conditional pdf of the spreading vector follows a nonzero-mean Gaussian $\mathcal{N}(\mathbf{v},\sigma^2_{\bar{S}_{N_o}})$, where $\mathbf{v}$ and $\sigma^2_{\bar{S}_{N_o}}$ are given by (3.8) and (3.9), respectively. The bounds derived in this appendix are based on the fact that the norm of a nonzero-mean Gaussian follows a noncentral Chi-square distribution. For the sake of notational clarity, let us define the random variable $T\sim\chi'^2(n,\mathbf{v},\sigma_{\bar{S}_{N_o}})$.

We will first derive the bound for the log-expectation of the norm. We can write

$$E\left[\log\left(T^{\frac{1}{2}}\right)\right]=\frac{1}{2}E\left[\log(T)\right]\leq\frac{1}{2}\log\left(E[T]\right)=\frac{1}{2}\log\left(n\sigma^2_{\bar{S}_{N_o}}+||\mathbf{v}||^2\right), \quad \text{(B.40)}$$

where the upper bound follows from Jensen's inequality [84]. Since $\mathbf{V}\sim\mathcal{N}\left(\mathbf{0},\frac{N_o\sigma_S^4}{N_o\sigma_S^2+\sigma_X^2}\mathbf{I}_n\right)$, the third term of (3.33) can be bounded from above as follows:

$$\begin{aligned}(n-1)&E\left[E[\log(Q)|\mathbf{O}_1=\mathbf{o}_1,\dots,\mathbf{O}_{N_o}=\mathbf{o}_{N_o}]\right]\\ &\leq \frac{(n-1)}{2}E\left[\log(n\sigma^2_{\bar{S}_{N_o}}+||\mathbf{V}||^2)\right]\\ &\leq \frac{(n-1)}{2}\log\left(n\sigma^2_{\bar{S}_{N_o}}+n\frac{N_o\sigma_S^4}{N_o\sigma_S^2+\sigma_X^2}\right)=\frac{1}{2}\log(n\sigma_S^2), \quad \text{(B.41)}\end{aligned}$$

where we have applied again Jensen's inequality. An upper bound to the second term of (3.33) is now derived. First, note that $h(T^{\frac{1}{2}})\leq\frac{1}{2}\log(2\pi e\cdot\mathrm{var}(T^{\frac{1}{2}}))$. For the variance, we have [200, Chapter 2]

$$\begin{aligned}\mathrm{var}(T^{\frac{1}{2}}) &= E[T]-E^2[T^{\frac{1}{2}}]\\ &= n\sigma^2_{\bar{S}_{N_o}}+||\mathbf{v}||^2\\ &- \left(\sqrt{2}\sigma_{\bar{S}_{N_o}}\exp\left(-\frac{||\mathbf{v}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right)\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\;{}_1F_1\left(\frac{n+1}{2};\frac{n}{2};\frac{||\mathbf{v}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right)\right)^2, \text{(B.42)}\end{aligned}$$

where ${}_1F_1(\alpha;\beta;x)$ denotes the confluent hypergeometric function of the first kind [29]. Hence, the second term of (3.33) can be upper bounded as

$$
\begin{aligned}
&h(Q|\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}, M_1, \ldots, M_{N_o}) \\
&\quad \leq \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log\left(n\sigma^2_{\bar{S}_{N_o}} + n\frac{N_o\sigma_S^4}{N_o\sigma_S^2 + \sigma_X^2} \right. \\
&\quad - 2\sigma^2_{\bar{S}_{N_o}} \left(\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\right)^2 \times \\
&\quad \left. E\left[\left(\exp\left(-\frac{||\mathbf{V}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right) {}_1F_1\left(\frac{n+1}{2}; \frac{n}{2}; \frac{||\mathbf{V}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right)\right)^2\right]\right), \qquad \text{(B.43)}
\end{aligned}
$$

Using the Kummer transformation [29, Chapter 13], the expectation in (B.43) can be written as

$$
E\left[\left(\exp\left(-\frac{||\mathbf{V}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right) {}_1F_1\left(\frac{n+1}{2}; \frac{n}{2}; \frac{||\mathbf{V}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right)\right)^2\right] = E\left[{}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -\frac{||\mathbf{V}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right)^2\right]. \tag{B.44}
$$

Now, consider the integral representation of the hypergeometric function [29]

$$
{}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -z\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(-\frac{1}{2}\right)} \int_0^1 e^{-zt} t^{-\frac{3}{2}} (1-t)^{\frac{n-1}{2}} dt. \tag{B.45}
$$

We want to prove that the function ${}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -z\right)^2$ is convex in z. Using Leibniz's rule [29, Chapter 3], we take the second derivative of the squared integral in (B.45), obtaining

$$
\begin{aligned}
&\tfrac{\partial^2}{\partial z^2}\left(\int_0^1 e^{-zt} t^{-\frac{3}{2}} (1-t)^{\frac{n-1}{2}} dt\right)^2 \\
&\quad = 2\left[\left(\int_0^1 e^{-zt} t^{-\frac{1}{2}} (1-t)^{\frac{n-1}{2}} dt\right)^2\right. \\
&\quad + \left.\left(\int_0^1 e^{-zt} t^{-\frac{3}{2}} (1-t)^{\frac{n-1}{2}} dt\right)\left(\int_0^1 e^{-zt} t^{\frac{1}{2}} (1-t)^{\frac{n-1}{2}} dt\right)\right]. \qquad \text{(B.46)}
\end{aligned}
$$

Eq. (B.46) is easily seen to be positive for all z, since all the integrands are positive. Thus, ${}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -z\right)^2$ is convex in z. This implies that we can lower bound (B.44) using Jensen's inequality as

$$
\begin{aligned}
E\left[{}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -\frac{||\mathbf{V}||^2}{2\sigma^2_{\bar{S}_{N_o}}}\right)^2\right] &\geq {}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -\frac{n\frac{N_o\sigma_S^4}{N_o\sigma_S^2+\sigma_X^2}}{2\sigma^2_{\bar{S}_{N_o}}}\right)^2 \\
&= {}_1F_1\left(-\frac{1}{2}; \frac{n}{2}; -\frac{nN_o\sigma_S^2}{2\sigma_X^2}\right)^2. \qquad \text{(B.47)}
\end{aligned}
$$

Combining (B.43), (B.44), and (B.47), we obtain

$$
\begin{aligned}
&h(Q|\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}, M_1, \ldots, M_{N_o}) \\
&\quad \leq \frac{1}{2}\log(2\pi e) \\
&\quad + \frac{1}{2}\log\left(n\sigma_S^2 - \frac{2\sigma_S^2}{1+N_o\xi^{-1}}\left(\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\right)^2 {}_1F_1\left(-\frac{1}{2};\frac{n}{2};-\frac{nN_o}{2}\xi^{-1}\right)^2\right). \quad \text{(B.48)}
\end{aligned}
$$

Finally, combining (3.10), (B.41), (B.48), and simplifying terms, the lower bound (3.34) follows.

B.6 Analysis of the blind CM cost function

Taking into account that both $\mathbf{s}$ and $\mathbf{s}_0$ are of unit norm, the first term of (4.25) can be computed as

$$
\begin{aligned}
E\left[(\mathbf{Y}^T\mathbf{s})^4\right] &= E\left[\left(\mathbf{X}^T\mathbf{s} + \nu(-1)^M\rho - \lambda(\mathbf{X}^T\mathbf{s}_0)\rho\right)^4\right] && \text{(B.49)}\\
&= E\left[(\mathbf{X}^T\mathbf{s})^4\right] + 6\nu^2\rho^2E\left[(\mathbf{X}^T\mathbf{s})^2\right] + \nu^4\rho^4 && \text{(B.50)}\\
&\quad +6\nu^2\lambda^2\rho^4E\left[(\mathbf{X}^T\mathbf{s}_0)^2\right] + \lambda^4\rho^4E\left[(\mathbf{X}^T\mathbf{s}_0)^4\right] && \text{(B.51)}\\
&\quad -4\lambda\rho E\left[(\mathbf{X}^T\mathbf{s})^3(\mathbf{X}^T\mathbf{s}_0)\right] && \text{(B.52)}\\
&\quad -12\lambda\nu^2\rho^3E\left[(\mathbf{X}^T\mathbf{s})(\mathbf{X}^T\mathbf{s}_0)\right] && \text{(B.53)}\\
&\quad +6\lambda^2\rho^2E\left[(\mathbf{X}^T\mathbf{s})^2(\mathbf{X}^T\mathbf{s}_0)^2\right] && \text{(B.54)}\\
&\quad -4\lambda^3\rho^3E\left[(\mathbf{X}^T\mathbf{s})(\mathbf{X}^T\mathbf{s}_0)^3\right]. && \text{(B.55)}
\end{aligned}
$$

By recalling that $\mathbf{X}$ is i.i.d. Gaussian, it follows that $E\left[(\mathbf{X}^T\mathbf{s})^4\right] = E\left[(\mathbf{X}^T\mathbf{s}_0)^4\right] = 3\sigma_X^4$, and $E\left[(\mathbf{Y}^T\mathbf{s})^2\right] = E\left[(\mathbf{Y}^T\mathbf{s}_0)^2\right] = \sigma_X^2$. Similarly, for the expectation in (B.53) we have

$$
E\left[(\mathbf{X}^T\mathbf{s})(\mathbf{X}^T\mathbf{s}_0)\right] = \mathbf{s}^TE\left[\mathbf{X}\mathbf{X}^T\right]\mathbf{s}_0 = \sigma_X^2\rho. \quad \text{(B.56)}
$$

Hence, computation of the terms (B.50), (B.51) and (B.53) is straightforward. The remaining terms require more involvement:

- For the expectation in the term (B.52), we first expand the terms of the correlation as follows

  $$
  E\left[(\mathbf{X}^T\mathbf{s})^3(\mathbf{X}^T\mathbf{s}_0)\right] = \sum_i\sum_j\sum_k\sum_m E\left[X_iX_jX_kX_ms_is_js_ks_{0,m}\right]. \quad \text{(B.57)}
  $$

  We consider now the non-null terms of (B.57):

$(i = j = k = m)$

$$\sum_i E\left[X_i^4\right] s_i^3 s_{0,i} = 3\sigma_X^4 \sum_i s_i^3 s_{0,i} \tag{B.58}$$

$(i = j) \neq (k = m)$

$$\sum_i \sum_{k \neq i} E\left[X_i^2 X_k^2\right] s_i^2 s_k s_{0,k} = \sigma_X^4 \sum_i \sum_{k \neq i} s_i^2 s_k s_{0,k} \tag{B.59}$$

$(i = k) \neq (j = m)$

$$\sum_i \sum_{j \neq i} E\left[x_i^2 x_j^2\right] s_i^2 s_j s_{0,j} = \sigma_X^4 \sum_i \sum_{j \neq i} s_i^2 s_j s_{0,j} \tag{B.60}$$

$(i = m) \neq (j = k)$

$$\sigma_X^4 \sum_i \sum_{j \neq i} s_j^2 s_i s_{0,i} \tag{B.61}$$

Combining the above expressions, rearranging terms and using the fact that

$$\sum_{i=1}^n s_i^2 \left(\sum_{k \neq i} s_k s_{0,k} \right) = \rho \sum_{i=1}^n s_i^2 - \sum_{i=1}^n s_i^3 s_{0,i},$$

Eq. (B.57) can be rewritten as

$$\begin{aligned} E\left[(\mathbf{X}^T\mathbf{s})^3(\mathbf{X}^T\mathbf{s}_0)\right] &= 3\sigma_X^4 \sum_{i=1}^n s_i^3 s_{0,i} + 3\sigma_X^4 \left(\rho - \sum_{i=1}^n s_i^3 s_{0,i} \right) \\ &= 3\rho\sigma_X^4. \end{aligned} \tag{B.62}$$

- The computation of the statistic in (B.55) is analogous to that of (B.52). It can be shown that

$$E\left[(\mathbf{X}^T\mathbf{s})^3(\mathbf{X}^T\mathbf{s}_0)\right] = 3\rho\sigma_X^4. \tag{B.63}$$

- For computing the statistic in (B.54), we expand again the correlations

$$E\left[(\mathbf{X}^T\mathbf{s})^2(\mathbf{X}^T\mathbf{s}_0)^2\right] = \sum_i \sum_j \sum_k \sum_m E\left[X_i X_j X_k X_m s_i s_j s_{0,k} s_{0,m}\right]. \tag{B.64}$$

The non-null terms of (B.57) are:

$(i = j = k = m)$

$$\sum_i E\left[X_i^4\right] s_i^2 s_{0,i}^2 = 3\sigma_X^4 \sum_i s_i^2 s_{0,i}^2 \tag{B.65}$$

$(i = j) \neq (k = m)$

$$\sigma_X^4 \sum_i \sum_{k \neq i} s_i^2 s_{0,k}^2 \tag{B.66}$$

$(i = k) \neq (j = m)$

$$\sigma_X^4 \sum_i \sum_{j \neq i} s_i s_j s_{0,i} s_{0,j} \tag{B.67}$$

$(i = m) \neq (j = k)$

$$\sigma_X^4 \sum_i \sum_{j \neq i} s_i s_j s_{0,i} s_{0,j} \tag{B.68}$$

Now, taking into account that

$$\begin{aligned} \sum_{i=1}^{n} s_i^2 \left(\sum_{k \neq i} s_{0,k}^2 \right) &= \sum_{i=1}^{n} s_i^2 \sum_{k=1}^{n} s_{0,i}^2 - \sum_{i=1}^{n} s_i^2 s_{0,i}^2 \\ &= ||\mathbf{s}||^2 ||\mathbf{s}_0||^2 - \sum_{i=1}^{n} s_i^2 s_{0,i}^2 &\text{(B.69)} \\ \sum_{i=1}^{n} s_i s_{0,i} \left(\sum_{j \neq i} s_j s_{0,j} \right) &= \rho^2 - \sum_{i=1}^{n} s_i^2 s_{0,i}^2, &\text{(B.70)} \end{aligned}$$

and combining the above terms, we can rewrite (B.64) as

$$\begin{aligned} E\left[(\mathbf{X}^T\mathbf{s})^2(\mathbf{X}^T\mathbf{s}_0)^2\right] &= 3\sigma_X^4 \sum_{i=1}^{n} s_i^2 s_{0,i}^2 + \sigma_X^4 \left(1 + 2\rho^2 - 3 \sum_{i=1}^{n} s_i^2 s_{0,i}^2 \right) \\ &= \sigma_X^4 (1 + 2\rho^2). &\text{(B.71)} \end{aligned}$$

By combining all the terms computed above, we can arrive at the final expression for the fourth order moment of the correlation between the observations and the estimate, i.e Eq. (B.49). The calculation of the second order moment is easier:

$$\begin{aligned} E\left[(\mathbf{Y}^T\mathbf{s})^2\right] &= E\left[\left(\mathbf{X}^T\mathbf{s} + \nu(-1)^M \rho - \lambda(\mathbf{X}^T\mathbf{s}_0)\rho\right)^2\right] \\ &= E\left[(\mathbf{X}^T\mathbf{s})^2\right] + \nu^2\rho^2 + \lambda^2\rho^2 E\left[(\mathbf{X}^T\mathbf{s}_0)^2\right] \\ &- 2\lambda\rho E\left[(\mathbf{X}^T s)(\mathbf{X}^T\mathbf{s}_0)\right] \\ &= \sigma_X^2 + \nu^2\rho^2 + \lambda^2\rho^2\sigma_X^2 - 2\lambda\rho^2\sigma_X^2. &\text{(B.72)} \end{aligned}$$

Finally, by combining all the statistics obtained above, we can arrive at the expression of the cost function (4.25)

$$\begin{aligned} J_{\mathrm{CM}}(\mathbf{s}) &= J_{\mathrm{CM}}(\rho) = \rho^4 \left(\nu^4 - 12\nu^2\lambda\sigma_X^2 + 6\nu^2\lambda^2\sigma_X^2 + 12\lambda^2\sigma_X^4 - 12\lambda^3\sigma_X^4 + 3\lambda^4\sigma_X^4\right) \\ &-2\rho^2\left(\nu^2 - 3\sigma_X^2\right)\left(\nu^2 - 2\lambda\sigma_X^2 + \lambda^2\sigma_X^2\right) + 3\sigma_X^4 - 2\nu^2\sigma_X^2 + \nu^4. &\text{(B.73)} \end{aligned}$$

B.7 Cost function of the blind CM method for Generalized Gaussian hosts

We assume that the components of the host signal follow a zero-mean Generalized Gaussian distribution with shape parameter c and variance σ_X^2, i.e. $X_i \sim GG(c,\sigma_X^2)\ , i = 1,\ldots,n$. Following similar calculations to those in Appendix B.6, the CM cost function for a GG host can be shown to be given by

$$\begin{aligned}
J_{\mathrm{CM}}(\mathbf{s},\mathbf{s}_0,c) &= \rho^4\Bigg(\nu^4 - 12\nu^2\lambda\sigma_X^2 + 6\nu^2\lambda^2\sigma_X^2 + 12\lambda^2\sigma_X^4 - 12\lambda^3\sigma_X^4 + 3\lambda^4\sigma_X^4 \\
&\quad +\lambda^4\sigma_X^4\kappa(c)\sum_{i=1}^n s_{0,i}^4\Bigg) - 4\rho^3\lambda^3\sigma_X^4\kappa(c)\sum_{i=1}^n s_i s_{0,i}^3 \\
&\quad -2\rho^2\left(\left(\nu^2-3\sigma_X^2\right)\left(\nu^2-2\lambda\sigma_X^2+\lambda^2\sigma_X^2\right) - 3\lambda^2\sigma_X^4\kappa(c)\sum_{i=1}^n s_i^2 s_{0,i}^2\right) \\
&\quad -4\rho\lambda\sigma_X^4\kappa(c)\sum_{i=1}^n s_i^3 s_{0,i} + 3\sigma_X^4 - 2\nu^2\sigma_X^2 + \nu^4 + \sigma_X^4\kappa(c)\sum_{i=1}^n s_i^4, \qquad \text{(B.74)}
\end{aligned}$$

where $\kappa(c) \triangleq \frac{\Gamma(1/c)\Gamma(5/c)}{\Gamma^2(3/c)} - 3$ is the kurtosis excess (γ_2 in the notation of [29]) for a GG with parameter c. Notice that for $\kappa(c) = 0$ and $c = 2$ (i.e. for a Gaussian host), the expression above is reduced to (B.73). Intuitively, the behavior of CM for any GG host should approach the behavior for a Gaussian host as n grows, because the sum of GGs tends asymptotically to a Gaussian distribution. As one can see, Eq. (B.74) depends on the actual realizations of $\mathbf{s}$ and $\mathbf{s}_0$, not only on their cross-correlation ρ, as happened with (B.73). However, taking the expectation of $J_{\mathrm{CM}}(\mathbf{s},\mathbf{s}_0,c)$ over $\mathbf{s}_0$ and taking into account that, for unit norm $\mathbf{s}$, $E[S_i^4] = 3/(2n+n^2)$, $E[S_i^2] = 1/n$ and $E[S_i^k] = 0$ for odd k (cf. Appendix B.3),

$$\begin{aligned}
E\left[J_{\mathrm{CM}}(\mathbf{s},\mathbf{S}_0,c)\right] &= \rho^4\Bigg(\nu^4 - 12\nu^2\lambda\sigma_X^2 + 6\nu^2\lambda^2\sigma_X^2 + 12\lambda^2\sigma_X^4 - 12\lambda^3\sigma_X^4 + 3\lambda^4\sigma_X^4 \\
&\quad +\lambda^4\sigma_X^4\kappa(c)\frac{n}{2n+n^2}\Bigg) \\
&\quad -2\rho^2\left(\left(\nu^2-3\sigma_X^2\right)\left(\nu^2-2\lambda\sigma_X^2+\lambda^2\sigma_X^2\right) - 3\lambda^2\sigma_X^4\kappa(c)\frac{1}{n}\right) \\
&\quad +3\sigma_X^4 - 2\nu^2\sigma_X^2 + \nu^4 + \sigma_X^4\kappa(c)\sum_{i=1}^n s_i^4. \qquad \text{(B.75)}
\end{aligned}$$

If we take the limit of the above expression when $n \to \infty$, we arrive at

$$\lim_{n\to\infty} E\left[J_{\mathrm{CM}}(\mathbf{s},\mathbf{S}_0,c)\right] = J_{\mathrm{CM}}(\rho) + \sigma_X^4\kappa(c)\sum_{i=1}^n s_i^4, \qquad \text{(B.76)}$$

where $J_{\mathrm{CM}}(\rho)$ is given by (B.73). Also for large n, $\sum_{i=1}^{n} s_i^4$ will be close to zero unless $\mathbf{s}$ is very close to any of the vectors of the canonical basis of $\mathbb{R}^n$. Hence, when the number of dimensions is very large, the cost function of CM is, in average, nearly independent of the shape parameter c and CM behaves approximately as for a Gaussian host.

Appendix C

C.1 Proof of Lemma 5.1

Assume that the secret dither, which is fixed for all k, takes the value $\mathbf{t}$. The random variable defined as $\tilde{\mathbf{V}}_k \triangleq (\tilde{\mathbf{Y}}_k - \mathbf{d}_{M_k}) \mod \Lambda$ is uniformly distributed over $(\mathbf{t} + \mathcal{Z}(\Lambda)) \mod \Lambda$. For $\alpha \geq 0.5$, the feasible region $\mathcal{S}_{N_o}$ is a modulo-Λ convex set. For any lattice Λ, $\mathcal{Z}(\Lambda)$ can be upper bounded by $\mathcal{B}(\mathbf{0}, (1-\alpha)r_c(\Lambda))$, where $r_c(\Lambda)$ is the covering radius of Λ defined in (5.11), and $\mathcal{B}(\mathbf{c}, r)$ denotes the n-dimensional closed hypersphere of radius r centered in $\mathbf{c}$. Therefore,

$$\mathcal{S} \subseteq \bigcap_{k=1}^{N_o} \mathcal{B}(\tilde{\mathbf{V}}_k, (1-\alpha)r_c(\Lambda)) \mod \Lambda.$$

The intersection between two hyperspheres of radius r becomes arbitrarily small as the distance between the centers of both spheres approaches $2r$. In the limit, their intersection is the unique point equidistant to the two centers. This means that the feasible region $\mathcal{S}_{N_o}$ converges to $\mathbf{t}$ if the maximum distance between the modulo-Λ reduced observations $\tilde{\mathbf{V}}_k$ approaches $2(1-\alpha)r_c(\Lambda)$. For this condition to hold, at least one observation $\tilde{\mathbf{V}}_i$ must be arbitrarily close to a certain vertex $\mathbf{a}$ of $(\mathbf{t}+\mathcal{Z}(\Lambda)) \mod \Lambda$, and at least another observation $\tilde{\mathbf{V}}_j$, with $j \neq i$, must be arbitrarily close to another vertex $\mathbf{b}$ at distance $2(1-\alpha)r_c(\Lambda)$. Intuitively, the probability of finding such a pair of observations goes to 1 almost surely as $N_o \to \infty$.

Let us define the random variable $D_{N_o} \triangleq \max\{\delta_{N_o}(\mathbf{a}), \delta_{N_o}(\mathbf{b})\}$, with $\delta_{N_o}(\mathbf{a}) = \min_k |\mathbf{a} - \tilde{\mathbf{V}}_k|$ and $\delta_{N_o}(\mathbf{b}) = \min_k |\mathbf{b} - \tilde{\mathbf{V}}_k|$. Formally, we want to show that $D_{N_o} \xrightarrow{\text{a.s.}} 0$, which is equivalent to show that for all $\epsilon > 0$

$$\Pr\left(\lim_{N_o \to \infty} |D_{N_o}| > \epsilon\right) = 0. \tag{C.1}$$

In order to prove almost sure convergence, it is sufficient to prove that the sum $\sum_{N_o=1}^{\infty} \Pr\left(|D_{N_o}| > \epsilon\right)$ is finite. In such case, almost sure convergence follows by virtue

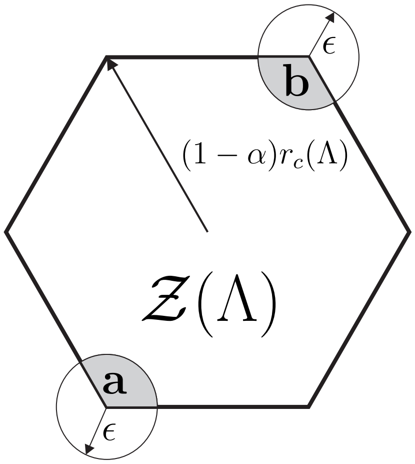


Figure C.1: Geometrical interpretation of the variables involved in the proof of Lemma 5.1. The value $(1-\alpha)r_c(\Lambda)$ is the covering radius of $\mathcal{Z}(\Lambda)$. The regions of $\mathcal{Z}(\Lambda)$ within distance ϵ of the vertices of interest, a and b, are shaded.

of the Borel-Cantelli lemma [114]. The probability of interest can be obtained as

$$\begin{aligned}
\Pr\left(|D_{N_o}| > \epsilon\right) &= \Pr\left(D_{N_o} > \epsilon\right) = \Pr\left(\delta_{N_o}(\mathbf{a}) > \epsilon \bigcup \delta_{N_o}(\mathbf{b}) > \epsilon\right) \\
&\leq \Pr\left(\delta_{N_o}(\mathbf{a}) > \epsilon\right) + \Pr\left(\delta_{N_o}(\mathbf{b}) > \epsilon\right) \\
&= \Pr\left(\bigcap_{k=1}^{N_o} |\mathbf{a} - \tilde{\mathbf{V}}_k| > \epsilon\right) + \Pr\left(\bigcap_{k=1}^{N_o} |\mathbf{b} - \tilde{\mathbf{V}}_k| > \epsilon\right) \\
&= \Pr\left(|\mathbf{a} - \tilde{\mathbf{V}}_k| > \epsilon\right)^{N_o} + \Pr\left(|\mathbf{b} - \tilde{\mathbf{V}}_k| > \epsilon\right)^{N_o} \qquad \text{(C.2)}
\end{aligned}$$

The probabilities in (C.2) can be easily computed by taking into account the geometrical interpretation provided by Figure C.1, using an hexagonal lattice as example. For the first probability in the right hand side of (C.2) we have

$$\begin{aligned}
\Pr\left(|\mathbf{a} - \tilde{\mathbf{V}}_k| > \epsilon\right) &= \frac{\operatorname{vol}(\mathcal{Z}(\Lambda)) - \operatorname{vol}(\mathcal{B}(\mathbf{a},\epsilon) \cap \mathcal{Z}(\Lambda))}{\operatorname{vol}(\mathcal{Z}(\Lambda))} \\
&= 1 - \frac{\operatorname{vol}(\mathcal{B}(\mathbf{a},\epsilon) \cap \mathcal{Z}(\Lambda))}{\operatorname{vol}(\mathcal{Z}(\Lambda))} = 1 - \rho_{\mathbf{a}}(\epsilon),
\end{aligned}$$

with $0 < \rho_{\mathbf{a}}(\epsilon) \leq 1$. Following a similar calculation for the other term involved in (C.2), we can write

$$\Pr\left(|D_{N_o}| > \epsilon\right) \leq (1 - \rho_{\mathbf{a}}(\epsilon))^{N_o} + (1 - \rho_{\mathbf{b}}(\epsilon))^{N_o},$$

and consequently

$$\sum_{N_o=1}^{\infty} \Pr(|D_{N_o}| > \epsilon) \leq 1/\rho_{\mathbf{a}}(\epsilon) + 1/\rho_{\mathbf{b}}(\epsilon) < \infty.$$

Since the above sum is finite, the result (C.1) follows as a consequence of the Borel-Cantelli lemma, thus proving Lemma C.1.

C.2 Proof of Theorem 5.1

Here we compute the mean value of (5.33). It can be seen that

$$W = 2\mu + \min\{\tilde{V}_1, \ldots, \tilde{V}_{N_o}\} - \max\{\tilde{V}_1, \ldots, \tilde{V}_{N_o}\}.$$

Hence, $W = 2\mu + X$, where X is the random variable defined as

$$X \triangleq \min\{\tilde{V}_1, \ldots, \tilde{V}_{N_o}\} - \max\{\tilde{V}_1, \ldots, \tilde{V}_{N_o}\},$$

where $x \in (-2\mu, 0]$, so the pdf of W is $f_W(w) = f_X(w - 2\mu)$. This allows us to rewrite the problem as

$$E[\log(W)] = \int_0^{2\mu} \log(w) \cdot f_W(w) dw. \tag{C.3}$$

First, let us see how the pdf of X can be computed. For having $X = x$, it should be $\min\{\ldots\} = t$ and $\max\{\ldots\} = t - x$; this is so when $\tilde{v}_i = t$, $\tilde{v}_j = t - x$, and $t \leq \tilde{v}_k \leq t - x$, for $k = \{1, \ldots, N_o\} \setminus \{i, j\}$, but taking into account that there are infinite values of t that yield $X = x$. Hence, the pdf of X reads as

$$f_X(x) = N_o(N_o - 1) \int_{-\mu}^{\mu + x} f_{\tilde{V}_i}(t) \cdot f_{\tilde{V}_i}(t - x) \cdot (\mathrm{Prob}\{t < \tilde{V}_i < t - x\})^{N_o - 2} dt, \tag{C.4}$$

where the factor $N_o(N_o - 1)$ comes from the number of different orderings of the minimum and the maximum in vector $(\tilde{v}_1, \ldots, \tilde{v}_{N_o})$; since all observations are i.i.d., we can simply multiply the integral by this factor. When $\tilde{V}_i \sim U(-\mu, \mu)$, computation of (C.4) in this case is straightforward and yields

$$f_X(x) = N_o(N_o - 1) \cdot \frac{(-x)^{N_o - 2}}{((1 - \alpha)\Delta)^{N_o}} \cdot [(1 - \alpha)\Delta + x], \tag{C.5}$$

for $\mu = (1 - \alpha)\Delta/2$. Thus, we must solve the following integral:

$$E[\log(W)] = \frac{N_o(N_o - 1)}{2\mu} \int_0^{2\mu} (2\mu - w)^{N_o - 2} w \log(w) dw, \tag{C.6}$$

where $\mu = (1-\alpha)\Delta/2$. Integration by parts must be recursively applied to (C.6). We define

$$u \triangleq (2\mu - w)^{N_o-2} \quad \Rightarrow \quad du = -(N_o-2)(2\mu - w)^{N_o-3}dw, \tag{C.7}$$

$$dv \triangleq w\log(w)dw \quad \Rightarrow \quad v = \frac{w^2}{2}\left(\log(w) - \frac{1}{2}\right), \tag{C.8}$$

hence,

$$\begin{aligned}\int_0^{2\mu}(2\mu - w)^{N_o-2}w\log(w)dw &= (2\mu - w)^{N_o-2}\cdot\frac{w^2}{2}\left(\log(w)-\frac{1}{2}\right)\Bigg|_0^{2\mu} \\ &+ (N_o-2)\int_0^{2\mu}(2\mu - w)^{N_o-3}\cdot\frac{w^2}{2}\left(\log(w)-\frac{1}{2}\right)dw.\end{aligned} \tag{C.9}$$

Integration by parts can be successively applied to the resulting integrals by using variables u and dv analogous to those of (C.7) and (C.8), arriving at a summation with $N_o - 1$ terms. For the k-th term, $k = 1, \ldots, N_o - 1$, we must compute an integral of the form

$$\int w^{k+1}\left(\log(w) - b_{k-1}\right)dw = \frac{w^{k+2}}{k+2}\left(\log(w) - b_k\right), \tag{C.10}$$

where

$$b_0 = \frac{1}{2}; \;\; b_k = \frac{1 + b_{k-1} + (k+1)\cdot b_{k-1}}{k+2}, \; k = 1, \ldots, N_o-2. \tag{C.11}$$

The expression with all the terms in the summation reads as

$$\begin{aligned}\int_0^{2\mu}(2\mu - w)^{N_o-2}w\log(w)dw &= (2\mu - w)^{N_o-2}\cdot\frac{w^2}{2}\left(\log(w)-\frac{1}{2}\right)\Bigg|_0^{2\mu} \\ &+ \sum_{k=1}^{N_o-2}(2\mu - w)^{N_o-2-k}w^{k+2}\left(\log(w) - b_k\right)\cdot c_k\Bigg|_0^{2\mu},\end{aligned} \tag{C.12}$$

where b_k are given by (C.11), and

$$c_0 = \frac{1}{2}; \;\; c_k = c_{k-1}\cdot\frac{N_o - 1 - k}{k+2}, \; k = 1, \ldots, N_o-2. \tag{C.13}$$

It can be seen that all the terms in (C.12) are null but the last one; hence, the value of the integral results in

$$\int_0^{2\mu}(2\mu - w)^{N_o-2}w\log(w)dw = (2\mu)^{N_o}\cdot\left(\log(2\mu) - b_{N_o-2}\right)\cdot c_{N_o-2}. \tag{C.14}$$

Notice that

$$c_{N_o-2} = \frac{(N_o-2)!}{N_o!} = \frac{1}{N_o(N_o-1)}, \tag{C.15}$$

and

$$b_k = \frac{1 + b_{k-1} + (k+1)\cdot b_{k-1}}{k+2} = b_{k-1} + \frac{1}{k+2},\ k = 1, \ldots, N_o-2, \tag{C.16}$$

thus $b_{N_o-2} = -1 + \sum_{i=1}^{N_o} \frac{1}{i} = H_{N_o} - 1$, where H_{N_o} is known as the N_o-th "harmonic number". Hence, the value of (C.14) can be succinctly expressed as

$$\int_0^{2\mu} (2\mu - w)^{N_o-2} w \log(w) dw \;=\; (\log(2\mu) - H_{N_o} + 1) \cdot \frac{(2\mu)^{N_o}}{N_o(N_o-1)}. \tag{C.17}$$

Finally, by plugging (C.17) into (C.6) with $\mu = (1-\alpha)\Delta/2$, we arrive at

$$E[\log(W)] = \log{(1-\alpha)\Delta} - H_{N_o} + 1. \tag{C.18}$$

C.3 Lower bound on the equivocation

By the definition of mutual information, we have

$$\begin{aligned} I(\tilde{\mathbf{Y}}_2, \ldots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}' | M_2, \ldots, M_{N_o}) &= h(\tilde{\mathbf{Y}}_2, \ldots, \tilde{\mathbf{Y}}_{N_o} | M_2, \ldots, M_{N_o}) \\ &- \sum_{i=2}^{N_o} h(\tilde{\mathbf{Y}}_i | \mathbf{T}', M_i). \end{aligned} \tag{C.19}$$

The first term of (C.19) can be bounded from above as [84]

$$\begin{aligned} h(\tilde{\mathbf{Y}}_2, \ldots, \tilde{\mathbf{Y}}_{N_o} | M_2, \ldots, M_{N_o}) &= h(\mathbf{Z}_2 + \mathbf{T}', \ldots, \mathbf{Z}_{N_o} + \mathbf{T}') \\ &\leq \sum_{i=1}^{n} h(Z_{i,2} + T_i', \ldots, Z_{i,N_o} + T_i'), \end{aligned} \tag{C.20}$$

where $Z_{i,j}$ is the i-th component of $\mathbf{Z}_j$, and T_i' denotes the i-th component of $\mathbf{T}'$. Since the host signals $\mathbf{X}_j$ and the secret dither $\mathbf{T}$ are mutually independent, it follows that $Z_{i,j}$ and T_i' are independent. Hence, we can write

$$\mathbf{R} \triangleq \mathrm{Cov}(Z_{i,2} + T_i', \ldots, Z_{i,N_o} + T_i') = \mathbf{R}_{Z_i} + \mathbf{R}_{T_i'}, \tag{C.21}$$

where $\mathbf{R}_{Z_i} \triangleq \mathrm{Cov}(Z_{i,2}, \ldots, Z_{i,N_o})$, and $\mathbf{R}_{T_i'} \triangleq \mathrm{Cov}(T_i', \ldots, T_i')$. Furthermore, it follows from Assumption 2 that the self-noise is white [222] with variance per dimension $(1-\alpha)^2 P(\Lambda_n^*)$. Hence, by considering that $Z_{i,j}$ are mutually independent for all j, we have

$$\mathbf{R}_{Z_i} = (1-\alpha)^2 P(\Lambda_n^*) \cdot \mathbf{I}_{N_o-1}, \qquad \mathbf{R}_{T_i'} = (1-\alpha)^2 P(\Lambda_n^*) \cdot \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1. \end{pmatrix}, \tag{C.22}$$

for $i = 1, \dots, n$. This allows us to bound Eq. (C.20) as [84, Th. 9.6.5]:

$$\begin{aligned} h(\tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o} | M_2, \dots, M_{N_o}) &\leq \frac{n}{2} \log \left((2\pi e)^{N_o - 1} |\mathbf{R}| \right) \\ &= \frac{n}{2} \log \left((2\pi e (1-\alpha)^2 P(\Lambda_n^*))^{N_o-1} \cdot N_o \right). \quad \text{(C.23)} \end{aligned}$$

The equivocation or residual entropy is

$$h(\mathbf{T}' | \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}, M_2, \dots, M_{N_o}) = h(\mathbf{T}') - I(\tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}; \mathbf{T}' | M_2, \dots, M_{N_o}), \text{(C.24)}$$

hence, using (C.19) and (C.23), Eq. (C.24) can be lower bounded as

$$\begin{aligned} h(\mathbf{T}' | \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}, M_2, \dots, M_{N_o}) &\geq h(\mathbf{T}') + \sum_{i=2}^{N_o} h(\tilde{\mathbf{Y}}_i | \mathbf{T}', M_i) \\ &- \frac{n}{2} \log \left((2\pi e (1-\alpha)^2 P(\Lambda_n^*))^{N_o-1} \cdot N_o \right). \quad \text{(C.25)} \end{aligned}$$

Taking into account that $h(\tilde{\mathbf{Y}}_i | \mathbf{T}', M_i) = h(\mathbf{T}') = h(\mathbf{T}) + n \log(1-\alpha)$, and rearranging terms, we finally arrive at the following lower bound to the equivocation per dimension:

$$\begin{aligned} &\tfrac{1}{n} h(\mathbf{T}' | \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}, M_2, \dots, M_{N_o}) \\ &\qquad \geq N_o \frac{h(\mathbf{T})}{n} - \frac{1}{2} \log \left((2\pi e P(\Lambda_n^*))^{N_o-1} \cdot N_o \right) + \log(1-\alpha), \qquad \text{(C.26)} \end{aligned}$$

and after substituting $\frac{1}{n} h(\mathbf{T}) = \frac{1}{n} \log(\mathrm{vol}(\mathcal{V}(\Lambda_n^*))) = \frac{1}{2} \log \left(\frac{P(\Lambda_n^*)}{G(\Lambda_n^*)} \right)$, we obtain Eq. (5.40).

C.4 Proof of Theorem 5.2

In order to arrive at Eq. (5.41), we start from the expression

$$\begin{aligned} \frac{1}{n} h(\mathbf{T}' | \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o}, M_2, \dots, M_{N_o}) &= N_o \frac{h(\mathbf{T})}{n} + N_o \log(1-\alpha) \\ &- \frac{1}{n} h(\tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{N_o} | M_2, \dots, M_{N_o}), \quad \text{(C.27)} \end{aligned}$$

which can be straightforwardly obtained by following the reasoning in Appendix C.3. First, we note that for the sequence of optimum lattice quantizers Λ_n^* we have [222]

$$\lim_{n \to \infty} \frac{h(\mathbf{T})}{n} = \frac{1}{2} \log(2\pi e P(\Lambda_n^*)). \qquad \text{(C.28)}$$

On the other hand, we want to prove that the following relation holds:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}h(\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o-1}|M_1,\ldots,M_{N_o-1}) &= \lim_{n\to\infty}\frac{1}{n}h(\mathbf{Z}_1+\mathbf{T}',\ldots,\mathbf{Z}_{N_o-1}+\mathbf{T}') \\ &= \frac{1}{2}\log\left((2\pi e(1-\alpha)^2\cdot P(\Lambda_n^*))^{N_o-1}\cdot N_o\right),\end{aligned} \tag{C.29}$$

with $\mathbf{Z}_i$, $\mathbf{T}'$ independent and uniformly distributed in $(1-\alpha)\mathcal{V}(\Lambda_n^*)$, being $\mathcal{V}(\Lambda_n^*)$ the Voronoi cell of Λ_n^* with second moment per dimension $P(\Lambda_n^*)$. Notice that we have rearranged the observation indices from 1 to N_o-1, for the sake of clarity. We will prove this result by making use of two lemmas.

Lemma C.1. Let $\mathbf{Z}$, $\mathbf{T}'$ be two independent random variables uniformly distributed in $(1-\alpha)\mathcal{V}(\Lambda_n^*)$. We have that

$$\lim_{n\to\infty}\frac{h(\mathbf{Z}+\mathbf{T}')}{n} = \frac{1}{2}\log\left(2\pi e(1-\alpha)^2 P(\Lambda_n^*)\cdot 2\right). \tag{C.30}$$

Proof: The entropy power inequality [84] states that

$$e^{\frac{2}{n}h(\mathbf{Z}+\mathbf{T}')} \geq e^{\frac{2}{n}h(\mathbf{Z})} + e^{\frac{2}{n}h(\mathbf{T}')}. \tag{C.31}$$

Furthermore, we know that [222]

$$\lim_{n\to\infty}\frac{h(\mathbf{Z})}{n} = \lim_{n\to\infty}\frac{h(\mathbf{T}')}{n} = \frac{1}{2}\log\left(2\pi e(1-\alpha)^2 P(\Lambda_n^*)\right), \tag{C.32}$$

so we can write

$$\begin{aligned}\lim_{n\to\infty} e^{\frac{2}{n}h(\mathbf{Z})} + e^{\frac{2}{n}h(\mathbf{T}')} &= 2\cdot e^{\log\left(2\pi e(1-\alpha)^2 P(\Lambda_n^*)\right)} \\ &= 2\pi e(1-\alpha)^2 P(\Lambda_n^*)\cdot 2 = e^{\frac{2}{n}h(\mathbf{U})},\end{aligned} \tag{C.33}$$

with $\mathbf{U}\sim\mathcal{N}(\mathbf{0},2(1-\alpha)^2P(\Lambda_n^*)\cdot\mathbf{I}_n)$. Thus, from Eq. (C.31) we have that

$$\lim_{n\to\infty}\frac{1}{n}h(\mathbf{Z}+\mathbf{T}') \geq \frac{h(\mathbf{U})}{n} = \frac{1}{2}\log(2\pi e(1-\alpha)^2P(\Lambda_n^*)\cdot 2), \tag{C.34}$$

and we know from Eq. (C.23) that

$$\frac{h(\mathbf{Z}+\mathbf{T}')}{n} \leq \frac{1}{2}\log\left(2\pi e(1-\alpha)^2P(\Lambda_n^*)\cdot 2\right) \tag{C.35}$$

for all n. Hence, by combining (C.34) and (C.35) the lemma follows. ■

Lemma C.2. For $\mathbf{Z}_i$, $\mathbf{T}'$ uniformly distributed in $(1-\alpha)\mathcal{V}(\Lambda_n^*)$, the following result holds

$\lim_{n\to\infty} \frac{1}{n} h(\mathbf{Z}_m + \mathbf{T}' | \mathbf{Z}_{m-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}') =$

$$\frac{1}{2}\log\left(2\pi e (1-\alpha)^2 P(\Lambda_n^*) \cdot \frac{m+1}{m}\right), \text{ for } m \geq 1. \tag{C.36}$$

Proof: We will prove the result by induction. Since it was proven for $m=1$ in Lemma C.1, we will show now that it is true for $m=i$, assuming that it holds for $m \leq i-1$. Making use of the entropy power inequality and the convexity of $\log(e^x + c)$ in x [49], we can write

$$\frac{2}{n} h(\mathbf{Z}_i + \mathbf{T}' | \mathbf{Z}_{i-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}) \;\geq\; \log\left(e^{\frac{2}{n}h(\mathbf{Z}_i)} + e^{\frac{2}{n}h(\mathbf{T}' | \mathbf{Z}_{i-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}')}\right). \tag{C.37}$$

By using the chain rule for entropies, it can be shown that the following equivocation can be written as

$$h(\mathbf{T}' | \mathbf{Z}_{i-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}') \;=\; i \cdot h(\mathbf{Z}_i) - \sum_{j=1}^{i-1} h(\mathbf{Z}_j + \mathbf{T}' | \mathbf{Z}_{j-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}'), \tag{C.38}$$

and making use of the inductive hypothesis we have that

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} h(\mathbf{T}' | \mathbf{Z}_{i-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}') \;&=\; \frac{i}{2}\log\left(2\pi e (1-\alpha)^2 P(\Lambda_n^*)\right) \\ &-\; \frac{1}{2}\sum_{j=1}^{i-1} \log\left(2\pi e (1-\alpha)^2 P(\Lambda_n^*) \cdot \frac{j+1}{j}\right) \\ &=\; \frac{1}{2}\log\left(2\pi e \cdot \frac{(1-\alpha)^2 P(\Lambda_n^*)}{i}\right). \end{aligned} \tag{C.39}$$

Thus, if we take limits in (C.37) we arrive at the following bound:

$$\lim_{n\to\infty} \frac{1}{n} h(\mathbf{Z}_i + \mathbf{T}' | \mathbf{Z}_{i-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}') \geq \frac{1}{2}\log\left(2\pi e (1-\alpha)^2 P(\Lambda_n^*) \cdot \frac{i+1}{i}\right). \tag{C.40}$$

Note that from the bounding given in (C.23) and the inductive hypothesis it follows that

$$\lim_{n\to\infty} \frac{1}{n} h(\mathbf{Z}_i + \mathbf{T}' | \mathbf{Z}_{i-1} + \mathbf{T}', \ldots, \mathbf{Z}_1 + \mathbf{T}') \leq \frac{1}{2}\log\left(2\pi e (1-\alpha)^2 P(\Lambda_n^*) \cdot \frac{i+1}{i}\right). \tag{C.41}$$

Hence, by combining (C.40) and (C.41), the lemma follows. ■

Now, using the chain rule for differential entropies we can write

$$\frac{1}{n}h(\mathbf{Z}_1+\mathbf{T}',\ldots,\mathbf{Z}_{N_o-1}+\mathbf{T}')=\frac{1}{n}\sum_{i=1}^{N_o-1}h(\mathbf{Z}_i+\mathbf{T}'|\mathbf{Z}_{i-1}+\mathbf{T}',\ldots,\mathbf{Z}_1+\mathbf{T}'), \tag{C.42}$$

and taking the limit when $n\to\infty$, by virtue of Lemma C.2, we arrive at the result given in (C.29). Finally, by combining (C.27), (C.28) and (C.29) we can conclude that

$$\lim_{n\to\infty}\frac{1}{n}h(\mathbf{T}'|\tilde{\mathbf{Y}}_2,\ldots,\tilde{\mathbf{Y}}_{N_o},M_2,\ldots,M_{N_o})=\frac{1}{2}\log(2\pi eP(\Lambda_n^*))-\frac{1}{2}\log(N_o)+\log(1-\alpha),$$

which is the desired result. If we identify now $P(\Lambda_n^*)=D_w/\alpha^2$, then Theorem 5.2 follows.

C.5 Proof of Lemma 5.3

First, one must realize that if the assumption of independence between embedded messages holds, then null information leakage for one observation implies perfect secrecy for all N_o. Thus, the proof of perfect secrecy can be reduced to show that the condition $I(\tilde{\mathbf{Y}}_1;\mathbf{T})=h(\tilde{\mathbf{Y}}_1)-h(\tilde{\mathbf{Y}}_1|\mathbf{T})=0$ is fulfilled.

With a little abuse of notation, in this appendix we will denote the pdf of $\tilde{\mathbf{y}}_1$ conditioned on $\mathbf{t}$ by $f_p(\tilde{\mathbf{y}}_1|\mathbf{t})$, for making clear the dependence with p, the alphabet size. We will first consider the term $h(\tilde{\mathbf{Y}}_1|\mathbf{T})$. Due to the flat-host assumption, we have $h(\tilde{\mathbf{Y}}_1|\mathbf{T})=h(\tilde{\mathbf{Y}}_1|\mathbf{T}=\mathbf{t})$ for any $\mathbf{t}$. Hence, we need to calculate

$$\lim_{p\to\infty}h(\tilde{\mathbf{Y}}_1|\mathbf{T}=\mathbf{t})=-\lim_{p\to\infty}\int_{\mathcal{V}(\Lambda)}f_p(\tilde{\mathbf{y}}_1|\mathbf{t})\log(f_p(\tilde{\mathbf{y}}_1|\mathbf{t}))d\tilde{\mathbf{y}}_1. \tag{C.43}$$

Computation of the integral in (C.43) is unaffordable. However, by virtue of the bounded convergence theorem [202, Chapt. 4], integral sign and limit can be interchanged if the integrand converges pointwise and

$$|f_p(\tilde{\mathbf{y}}_1|\mathbf{t})\log(f_p(\tilde{\mathbf{y}}_1|\mathbf{t}))|\leq g(\tilde{\mathbf{y}}_1)\ \forall\, p, \tag{C.44}$$

where $g(\tilde{\mathbf{y}}_1)$ is any function such that $\int_{\mathcal{V}(\Lambda)}|g(\tilde{\mathbf{y}}_1)|d\tilde{\mathbf{y}}_1<\infty$. In our case it suffices to choose a constant function $g(\tilde{\mathbf{y}}_1)=\mathrm{vol}(\mathcal{Z}(\Lambda))^{-1}\log(\mathrm{vol}(\mathcal{Z}(\Lambda))^{-1})\ \forall\ \tilde{\mathbf{y}}_1\in\mathcal{V}(\Lambda)$. The integral of $|g(\tilde{\mathbf{y}}_1)|$ for this choice is finite, since $\alpha<1$ by assumption. Thus, (C.43) can be computed as

$$\lim_{p\to\infty}h(\tilde{\mathbf{Y}}_1|\mathbf{T}=\mathbf{t})\ =\ -\int_{\mathcal{V}(\Lambda)}\lim_{p\to\infty}f_p(\tilde{\mathbf{y}}_1|\mathbf{t})\log(f_p(\tilde{\mathbf{y}}_1|\mathbf{t}))d\tilde{\mathbf{y}}_1. \tag{C.45}$$

We turn, for the moment, our attention to the computation of the limit of the conditioned pdf. The pdf of $\tilde{\mathbf{y}}_1$ conditioned on $\mathbf{t}$ is given by

$$\begin{aligned} f_p(\tilde{\mathbf{y}}_1|\mathbf{t}) &= \frac{1}{p}\sum_{k=0}^{p-1}\varphi(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_k \mod \Lambda) \\ &= \frac{1}{\mathrm{vol}(\mathcal{V}(\Lambda))}\sum_{k=0}^{p-1}\varphi(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_k \mod \Lambda)\cdot \mathrm{vol}(\mathcal{V}(\Lambda_f)), \end{aligned} \tag{C.46}$$

where the second equality follows from the definition of nesting ratio (cf. Sect. 5.1.1). By recalling the definition of the function $\varphi(\mathbf{x})$ in (5.18), Eq. (C.46) can be further rewritten as

$$f_p(\tilde{\mathbf{y}}_1|\mathbf{t}) = \frac{\mathrm{vol}(\mathcal{Z}(\Lambda))^{-1}}{\mathrm{vol}(\mathcal{V}(\Lambda))}\sum_{k=0}^{p-1}\phi_{\mathcal{Z}(\Lambda)}(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_k \mod \Lambda)\cdot \mathrm{vol}(\mathcal{V}(\Lambda_f)), \tag{C.47}$$

where $\mathcal{Z}(\Lambda) = (1-\alpha)\mathcal{V}(\Lambda)$. Let us analyze the sum in (C.47). Each term is given by

$$\phi_{\mathcal{Z}(\Lambda)}(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_k \mod \Lambda) = \begin{cases} 1, & \text{for } \mathbf{d}_k \in (\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathcal{Z}(\Lambda)) \mod \Lambda \\ 0, & \text{elsewhere} \end{cases} \tag{C.48}$$

Hence, the sum $\sum_{k=0}^{p-1}\phi_{\mathcal{Z}(\Lambda)}(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_k \mod \Lambda)$ is equivalent to counting the points $\mathbf{d}_k \in \mathcal{D}_p$ that fall inside the region $(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathcal{Z}(\Lambda)) \mod \Lambda$. Let us denote by $\mathcal{F}$ the subset of points of $\mathcal{D}_p$ for which the indicator function (C.48) is nonnull. The set $\mathcal{F}$ induces a partition (according to Λ_f) of the region $(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathcal{Z}(\Lambda) \mod \Lambda$. Since each term in the sum is multiplied by $\mathrm{vol}(\mathcal{V}(\Lambda_f))$, the result is an approximation of the volume of $-\mathcal{Z}(\Lambda)$ (which is equal to $\mathcal{Z}(\Lambda)$ up to a set of measure zero), i.e.

$$\sum_{k=0}^{p-1}\phi_{\mathcal{Z}(\Lambda)}(\tilde{\mathbf{y}}_1 - \mathbf{t} - \mathbf{d}_k \mod \Lambda)\cdot \mathrm{vol}(\mathcal{V}(\Lambda_f)) = \mathrm{vol}(\mathcal{Z}(\Lambda)) + \varepsilon(p), \tag{C.49}$$

where $\varepsilon(p)$ represents the discrepancy between $\mathrm{vol}(\mathcal{Z}(\Lambda))$ and the value of the sum, which comes from the points of $\mathcal{F}$ close to the boundary of $\mathcal{Z}(\Lambda)$. Recall that, by definition of covering radius (cf. Eq. (5.11)),

$$\mathcal{V}(\Lambda_f) \subset \mathcal{B}(\mathbf{0}, r_c(\Lambda_f)), \tag{C.50}$$

where $\mathcal{B}(\mathbf{0}, r_c(\Lambda_f))$ is the n-dimensional closed ball of radius $r_c(\Lambda_f)$. Hence, if $r_c(\Lambda_f) \to 0$ as $p \to \infty$, the term $\varepsilon(p)$ goes to 0 as well, and we have the definition of n-dimensional Riemann integral:[1]

[1] Notice that the volume of $\mathcal{Z}(\Lambda)$ can be computed by means of a Riemann integral, because for $\alpha < 1$, $\mathcal{Z}(\Lambda)$ is compact, and it is the linear (and invertible) image of a n-dimensional hypercube.

$$\lim_{p\to\infty}\sum_{k=0}^{p-1}\phi_{\mathcal{Z}(\Lambda)}(\tilde{\mathbf{y}}_1-\mathbf{t}-\mathbf{d}_k \mod \Lambda)\cdot \mathrm{vol}(\mathcal{V}(\Lambda_f))$$
$$=\int_{\mathcal{Z}(\Lambda)} d\mathbf{d}=\mathrm{vol}(\mathcal{Z}(\Lambda)),\ \forall\,\tilde{\mathbf{y}}_1\in\mathcal{V}(\Lambda). \tag{C.51}$$

Finally, by substituting (C.51) into (C.47) we arrive at

$$\lim_{p\to\infty} f_p(\tilde{\mathbf{y}}_1|\mathbf{t})=\frac{1}{\mathrm{vol}(\mathcal{V}(\Lambda))},\ \forall\,\tilde{\mathbf{y}}_1\in\mathcal{V}(\Lambda). \tag{C.52}$$

Furthermore, from the properties of the Riemann sum, it follows that $\forall\ \epsilon>0$ there exists p_0 such that, for $p\geq p_0$,

$$\left|f_p(\tilde{\mathbf{y}}_1|\mathbf{t})-\mathrm{vol}(\mathcal{V}(\Lambda))^{-1})\right|<\epsilon\ \forall\,\tilde{\mathbf{y}}_1\in\mathcal{V}(\Lambda). \tag{C.53}$$

Thus, we have proved that $f_p(\tilde{\mathbf{y}}_1|\mathbf{t})$ is uniformly convergent [202] with limit $\mathrm{vol}(\mathcal{V}(\Lambda))^{-1}$, $\forall\,\tilde{\mathbf{y}}_1\in\mathcal{V}(\Lambda)$.

In order to compute the limit in (C.45), we observe that the function $x\log(x)$ is continuous in $[0,Q]$, with finite Q, so it is uniformly continuous on that interval. Therefore, (C.45) can be computed as

$$\begin{aligned}\lim_{p\to\infty} h(\tilde{\mathbf{Y}}_1|\mathbf{T}=\mathbf{t}) &= -\int_{\mathcal{V}(\Lambda)}\mathrm{vol}(\mathcal{V}(\Lambda))^{-1}\log(\mathrm{vol}(\mathcal{V}(\Lambda))^{-1}d\tilde{\mathbf{y}}_1\\ &= \log(\mathrm{vol}(\mathcal{V}(\Lambda))).\end{aligned} \tag{C.54}$$

As for the other term involved in the mutual information, $h(\tilde{\mathbf{Y}}_1)$, we know that the entropy of a continuous random variable with bounded support can be upper bounded by the log-volume of its support set. Thus, we can write

$$h(\tilde{\mathbf{Y}}_1|\mathbf{T})\leq h(\tilde{\mathbf{Y}}_1)\leq\log(\mathrm{vol}(\mathcal{V}(\Lambda))). \tag{C.55}$$

Since $\lim_{p\to\infty}h(\tilde{\mathbf{Y}}_1|\mathbf{T})=\log(\mathrm{vol}(\mathcal{V}(\Lambda)))$ it is immediate, by (C.54) and (C.55), that $\lim_{p\to\infty}h(\tilde{\mathbf{Y}}_1)=\log(\mathrm{vol}(\mathcal{V}(\Lambda)))$, fulfilling the condition of perfect secrecy regardless the distribution of $\mathbf{T}$, and Lemma 5.3 follows.

C.6 A posteriori probability of the message sequences

In order to compute the a posteriori probability of a message sequence $\mathbf{m}^{(i)}=(m_1^i,\ldots,m_{N_o}^i)$ (hereinafter, a "path"), this probability is first rewritten using Bayes' rule:

$$\Pr(m_1^i,\ldots,m_{N_o}^i|\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o})=\frac{f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o}|m_1^i,\ldots,m_{N_o}^i)\cdot\Pr(m_1^i,\ldots,m_{N_o}^i)}{f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o})}. \tag{C.56}$$

We will focus on the probability a posteriori of the observations, which can be written as

$$\begin{aligned}
&f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o}|m_1^i,\ldots,m_{N_o}^i)\\
&\qquad = \int_{\mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o}|m_1^i,\ldots,m_{N_o}^i,\mathbf{t})\cdot f(\mathbf{t})d\mathbf{t}\\
&\qquad = \int_{\mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_{N_o}|m_{N_o}^i,\mathbf{t})\cdot f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o-1}|m_1^i,\ldots,m_{N_o-1}^i,\mathbf{t})\cdot f(\mathbf{t})d\mathbf{t}, \quad \text{(C.57)}
\end{aligned}$$

where the second equality follows from the mutual independence between the observations when the secret dither $\mathbf{t}$ is known. Eq. (C.57) can be rewritten as

$$\begin{aligned}
&f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o}|m_1^i,\ldots,m_{N_o}^i)\\
&\qquad = f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o-1}|m_1^i,\ldots,m_{N_o-1}^i)\times\\
&\qquad\quad \int_{\mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_{N_o}|m_{N_o}^i,\mathbf{t})\cdot f(\mathbf{t}|\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o-1},m_1^i,\ldots,m_{N_o-1}^i)d\mathbf{t}. \quad \text{(C.58)}
\end{aligned}$$

If the same procedure is applied recursively to the leftmost term in the right hand side of (C.58), we arrive at the following factorization for the a posteriori probability:

$$\begin{aligned}
&f(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{N_o}|m_1^i,\ldots,m_{N_o}^i)\\
&\qquad = \prod_{k=1}^{N_o} f(\tilde{\mathbf{y}}_k|m_1^i,\ldots,m_k^i,\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{k-1})\\
&\qquad = \prod_{k=1}^{N_o} \int_{\mathcal{V}(\Lambda)} f(\tilde{\mathbf{y}}_k|m_k^i,\mathbf{t})\cdot f(\mathbf{t}|\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{k-1},m_1^i,\ldots,m_{k-1}^i)d\mathbf{t}. \quad \text{(C.59)}
\end{aligned}$$

In order to compute each factor of (C.59), we recall the flat-host assumption, which implies that $f(\tilde{\mathbf{y}}_k|m_k^i,\mathbf{t}) = \varphi(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i} - \mathbf{t} \mod \Lambda)$. Thus, each factor of (C.59) can be seen as a circular convolution over $\mathcal{V}(\Lambda)$. Furthermore, under the assumption that $\mathbf{T} \sim U(\mathcal{V}(\Lambda))$, we have that the conditional pdf of the dither is given by Eq. (5.26). By combining (5.18) and (5.26), it can be seen that the integrand of the kth factor in Eq. (C.59) is

$$\begin{cases}
\dfrac{\mathrm{vol}(\mathcal{S}_{k-1}(\mathbf{m}^{(i)}))^{-1}}{\mathrm{vol}(\mathcal{Z}(\Lambda))}, & \text{for } \mathbf{t}\in\mathcal{S}_{k-1}(\mathbf{m}^{(i)}) \text{ such that } (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i} - \mathbf{t}) \mod \Lambda \in \mathcal{Z}(\Lambda)\\
0, & \text{otherwise.}
\end{cases}$$

The condition on $\mathbf{t}$ in the equation above is equivalent to $\mathbf{t} \in S_{k-1}(\mathbf{m}^{(i)})$ such that $\mathbf{t} \in (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i} - \mathcal{Z}(\Lambda)) \mod \Lambda$, so each factor in (C.59) is proportional to the volume of

$\mathcal{S}_k(\mathbf{m}^{(i)}) = \mathcal{S}_{k-1}(\mathbf{m}^{(i)}) \cap \mathcal{D}_k(m_k^i)$. Finally, Eq. (C.59) can be succinctly expressed as

$$f(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} | m_1^i, \ldots, m_{N_o}^i)$$

$$= \begin{cases} \prod_{k=1}^{N_o} \frac{\mathrm{vol}(\mathcal{S}_k(m_1^i, \ldots, m_k^i))}{\mathrm{vol}(\mathcal{Z}(\Lambda)) \cdot \mathrm{vol}(\mathcal{S}_{k-1}(m_1^i, \ldots, m_{k-1}^i))}, & \text{for } \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} \in \mathcal{S}_{N_o}(\mathbf{m}^{(i)}) \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{\mathrm{vol}(\mathcal{S}_{N_o}(m_1^i, \ldots, m_{N_o}^i))}{(\mathrm{vol}(\mathcal{Z}(\Lambda)))^{N_o} \cdot \mathrm{vol}(\mathcal{V}(\Lambda))}, & \text{for } \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o} \in \mathcal{S}_{N_o}(\mathbf{m}^{(i)}) \\ 0, & \text{otherwise} \end{cases} \quad \text{(C.60)}$$

C.7 Proof of Theorem 5.3

As shown in Eq. (C.56) and (C.60) from App. C.6, the a posteriori probability of a message sequence is nonnull only if its feasible region is not an empty set. Using the definition of feasible region (cf. (5.53) and (5.54)), we can state that a certain message sequence $\mathbf{m}^{(i)}$ has nonnull a posteriori probability if

$$\bigcap_{k=1}^{N_o} (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i} - \mathcal{Z}(\Lambda)) \mod \Lambda \neq \emptyset. \quad \text{(C.61)}$$

We say that the sequences fulfilling the above condition are "feasible" given the set of observations. The proof of the theorem is based on the concept of feasibility.

First, let us denote by $\mathbf{m}^{(1)}$ the message sequence embedded in the observations. We will arrange, without loss of generality, the message space $\mathcal{M}^{N_o}$ in two disjoint subsets: $\{\mathbf{m}^{(i)},\ i = 1, \ldots, p\}$, which will represent the sequences in the equivalence class of $\mathbf{m}^{(1)}$, and $\{\mathbf{m}^{(i)},\ i = p+1, \ldots, p^{N_o}\}$, which will represent the remaining sequences in $\mathcal{M}^{N_o}$. For the sake of clarity, the proof will be conducted in 4 steps.

Step 1)

In this step we will show that all the sequences in the equivalent class of $\mathbf{m}^{(1)}$ are feasible.

Notice that from Eq. (5.17), taking into account that $\mathbf{N}_k$ is uniformly distributed over $\mathcal{V}(\Lambda)$, we can write

$$(\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i}) \mod \Lambda \in (\mathcal{Z}(\Lambda) + \mathbf{t} + \mathbf{d}_{m_k^1} - \mathbf{d}_{m_k^i}) \mod \Lambda, \text{ for all } k = 1, \ldots, N_o,$$

or equivalently,

$$(\mathbf{t} + \mathbf{d}_{m_k^1} - \mathbf{d}_{m_k^i}) \mod \Lambda \in (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i} - \mathcal{Z}(\Lambda)) \mod \Lambda, \text{ for all } k = 1, \ldots, N_o. \quad \text{(C.62)}$$

By Lemma 5.4, for $i = 1, \dots, p$, and for all k, we have $\mathbf{d}_{m_k^i} = (\mathbf{d}_{m_k^1} + \mathbf{d}_l) \mod \Lambda$, for some fixed $l \in \mathcal{M}$. Hence, $(\mathbf{t} + \mathbf{d}_{m_k^1} - \mathbf{d}_{m_k^i}) \mod \Lambda = (\mathbf{t} - \mathbf{d}_l) \mod \Lambda$, for all $k = 1, \dots, N_o$. Substituting in (C.62), we obtain

$$(\mathbf{t} - \mathbf{d}_l) \mod \Lambda \in (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^i} - \mathcal{Z}(\Lambda)) \mod \Lambda, \text{ for all } k = 1, \dots, N_o. \tag{C.63}$$

This proves that, for the message sequences belonging to the equivalence class of $\mathbf{m}^{(1)}$, the intersection (C.61) is always non-empty, since at least the point $(\mathbf{t} - \mathbf{d}_l) \mod \Lambda$ is contained in the intersection.

Step 2)

In this step we prove that the feasible regions of the message sequences that belong to the equivalence class of $\mathbf{m}^{(1)}$ asymptotically converge to $(\mathbf{t} - \mathbf{d}_l) \mod \Lambda$, $l \in \mathcal{M}$.

As shown in Lemma 5.4, the feasible region for the sequences in the equivalence class of $\mathbf{m}^{(1)}$ only differ in their centers: for $i = 1, \dots, p$, $\mathcal{S}_{N_o}(\mathbf{m}^{(i)}) = (\mathcal{S}_{N_o}(\mathbf{m}^{(1)}) - \mathbf{d}_l) \mod \Lambda$. On the other hand, from Lemma 5.1 we know that the feasible region in the KMA case converges almost surely to $\mathbf{t}$ as $N_o \to \infty$. Combining these results with the result of Step 1), it follows that the feasible regions of the sequences in the equivalence class of $\mathbf{m}^{(1)}$ converge to $(\mathbf{t} - \mathbf{d}_l) \mod \Lambda$, $l \in \mathcal{M}$, as $N_o \to \infty$.

Step 3)

Now we consider the sequences not belonging to the equivalence class of $\mathbf{m}^{(1)}$. We will prove that these sequences are all unfeasible for sufficiently large N_o.

Among the elements of this set of sequences, consider a sequence $\mathbf{m}^{(j)}$ which, for some $i = 1, \dots, p$, equals $\mathbf{m}^{(i)}$ for all k but for a certain $k = k_0$. Consider the feasible region of this sequence, which is given by

$$\begin{aligned}\mathcal{S}_{N_o}(\mathbf{m}^{(j)}) &= \left(\bigcap_{k \backslash k_0} (\tilde{\mathbf{y}}_k - \mathbf{d}_{m_k^j} - \mathcal{Z}(\Lambda)) \mod \Lambda \right) \bigcap \left((\tilde{\mathbf{y}}_{k_0} - \mathbf{d}_{m_{k_0}^j} - \mathcal{Z}(\Lambda)) \mod \Lambda \right) \\ &= \mathcal{H}_1 \bigcap \mathcal{H}_2. \end{aligned} \tag{C.64}$$

For all $k \setminus k_0$, $(\mathbf{t} + \mathbf{d}_{m_k^1} - \mathbf{d}_{m_k^i}) \mod \Lambda = (\mathbf{t} - \mathbf{d}_l)$, for some fixed $l \in \mathcal{M}$. Hence, using the result of Step 2),

$$\lim_{N_o \to \infty} \mathcal{H}_1 = (\mathbf{t} - \mathbf{d}_l) \mod \Lambda. \tag{C.65}$$

For $k = k_0$, $(\mathbf{t} + \mathbf{d}_{m_k^1} - \mathbf{d}_{m_k^i}) = (\mathbf{t} - \mathbf{d}_c)$, with $c \neq l$. Therefore, $(\mathbf{t} - \mathbf{d}_c) \mod \Lambda \in \mathcal{H}_2$. Since $\mathcal{Z}(\Lambda) \subset \mathcal{V}(\Lambda_f)$ by assumption, then $(\mathbf{t} - \mathbf{d}_l) \mod \Lambda$, with $c \neq l$, cannot belong to $\mathcal{H}_2$. Thus,

$$\lim_{N_o \to \infty} \mathcal{S}_{N_o}(\mathbf{m}^{(j)}) = \lim_{N_o \to \infty} \mathcal{H}_1 \bigcap \mathcal{H}_2 = \emptyset. \tag{C.66}$$

This shows that the sequences that do not belong to the equivalence class of $\mathbf{m}^{(1)}$ cannot contain in their feasible region any of the points $(\mathbf{t}-\mathbf{d}_l) \mod \Lambda$, $l \in \mathcal{M}$, when $N_o \to \infty$. It only remains to be proved that these points are the only feasible regions when $N_o \to \infty$. This is a consequence of the next result.

Claim: Consider any point $\mathbf{z} \in \mathcal{V}(\Lambda)$ such that $\mathbf{z} \neq (\mathbf{t}-\mathbf{d}_l) \mod \Lambda$, $l \in \mathcal{M}$. For any such point, the probability of observing one $\tilde{\mathbf{y}}_k$ such that $\mathbf{z}$ does not belong to any of the regions $(\tilde{\mathbf{y}}_k - \mathbf{d}_i - \mathcal{Z}(\Lambda)) \mod \Lambda$, $i \in \mathcal{M}$, goes to 1 as $N_o \to \infty$.

Sketch of the proof: For fixed $\mathbf{t}$, the support of the observations is $\mathcal{R} \triangleq \bigcup_{l\in\mathcal{M}}(\mathbf{t}+\mathbf{d}_l+\mathcal{Z}(\Lambda)) \mod \Lambda$. Note that $\mathcal{R}$ does not cover the whole Voronoi region $\mathcal{V}(\Lambda)$, since $\mathcal{Z}(\Lambda) \subset \mathcal{V}(\Lambda_f)$ by assumption. Let us denote by $\overline{\mathcal{R}}$ the complement of $\mathcal{R}$ in $\mathcal{V}(\Lambda)$. If $\mathbf{z} \in \overline{\mathcal{R}}$, then it suffices to observe $\tilde{\mathbf{y}}_k = (\mathbf{t}+\mathbf{d}_l) \mod \Lambda$, for some $l \in \mathcal{M}$. On the other hand, if $\mathbf{z} \in \mathcal{R}$, then it suffices to observe $\tilde{\mathbf{y}}_k = (\mathbf{t}+\mathbf{d}_l-\mathbf{e}) \mod \Lambda$, for some $l \in \mathcal{M}$, where $\mathbf{e}$ is the shortest norm vector such that $(\mathbf{z}+\mathbf{e}) \mod \Lambda \in \overline{\mathcal{R}}$. Since for fixed $\mathbf{t}$ the observations are uniformly distributed over $\mathcal{R}$, the probability of observing such $\tilde{\mathbf{y}}_k$ goes to 1 as $N_o \to \infty$. ■

Therefore, any sequence not belonging to the equivalence class of $\mathbf{m}^{(1)}$ is unfeasible for $N_o \to \infty$.

Step 4)

The three previous steps have shown that the only message sequences in $\mathcal{M}^{N_o}$ that have nonnull probability when $N_o \to \infty$ are those belonging to the equivalence class of $\mathbf{m}^{(1)}$. We know, by Lemma 5.4, that all these sequences are equiprobable. Hence, if the attacker has enough computational power for checking an exponentially increasing number of intersections, then his uncertainty about the embedded message sequence becomes $\log(p)$ for $N_o \to \infty$, and the theorem follows.

C.8 Proof of Lemma 5.5

The proof consists in showing that the right hand side of (2.22) is lower bounded by $\log(p)$ when $N_o \to \infty$, i.e.

$$\lim_{N_o\to\infty} H(M_1,\ldots,M_{N_o}|\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o}) - H(M_1,\ldots,M_{N_o}|\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_{N_o},\mathbf{T}) \geq \log(p). \tag{C.67}$$

For the sake of clarity, the proof will be conducted in three steps.

Step 1)

Let us denote by $\mathcal{L}^{N_o} = \mathcal{L}_1 \times \mathcal{L}_2 \times \ldots \times \mathcal{L}_{N_o}$ the set of feasible message sequences for a fair (knowing $\mathbf{T}$) user, given N_o observations. In order to obtain the elements of $\mathcal{L}_i$, consider the a posteriori probability of a certain message m,

$$\Pr(m|\tilde{\mathbf{y}}_i, \mathbf{t}) = \frac{\Pr(m)}{f(\tilde{\mathbf{y}}_i|\mathbf{t})} \cdot \varphi((\tilde{\mathbf{y}}_i - \mathbf{d}_m - \mathbf{t}) \mod \Lambda). \tag{C.68}$$

By recalling the definition of $\varphi(\mathbf{x})$ in (5.18), Eq. (C.68) shows that $m \in \mathcal{L}_i$ iff $\mathbf{d}_m$ belongs to $(\tilde{\mathbf{y}}_i - \mathbf{t} - \mathcal{Z}(\Lambda)) \mod \Lambda$, and that all the elements of $\mathcal{L}_i$ are equiprobable (see example in Figure C.3). Thus, $\mathcal{L}^{N_o}$ is composed of $|\mathcal{L}^{N_o}| = \prod_{i=1}^{N_o} |\mathcal{L}_i|$ equiprobable message sequences.[2] Furthermore, note that $|\mathcal{L}^{N_o}|$ does not depend on the particular realization of $\mathbf{t}$. Hence, the entropy of the embedded message sequence for the fair user, given a particular realization of the observations, is

$$\begin{aligned} H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, \mathbf{t}) &= H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, \mathbf{t} = \mathbf{0}) \\ &= \log(|\mathcal{L}^{N_o}|) = \sum_{i=1}^{N_o} \log(|\mathcal{L}_i|). \end{aligned} \tag{C.69}$$

Step 2)
Let us denote by $\mathcal{P}^{N_o}$ the set of feasible message sequences when $\mathbf{T}$ is not known in advance. Lemma 5.4 tells us that $\mathcal{P}^{N_o}$ contains, at least, the sequences of $\mathcal{L}^{N_o}$ plus all the sequences that fulfill the equivalence relation (5.63) with those of $\mathcal{L}^{N_o}$. Taking into account the particular form of the equivalence classes for Construction A (see (5.63)), we can write $\mathcal{P}^{N_o} \supseteq \mathcal{K}^{N_o}$, where

$$\mathcal{K}^{N_o} \triangleq \bigcup_{i=1}^{|\mathcal{L}^{N_o}|} \bigcup_{j=0}^{p-1} (\mathbf{m}^{(i)} + j \cdot \mathbf{1}) \mod p, \text{ with } \mathbf{m}^{(i)} \in \mathcal{L}^{N_o}.$$

Notice that all the elements of $\mathcal{K}^{N_o}$ are equiprobable. Thus, the equivocation about the embedded message for an unfair user and a particular realization of the observations can be bounded from below as

$$H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) \geq \log(|\mathcal{K}^{N_o}|). \tag{C.70}$$

The cardinality of $\mathcal{K}^{N_o}$ depends on the number of equivalence classes in $\mathcal{L}^{N_o}$ as

$$|\mathcal{K}^{N_o}| = p \cdot |\mathcal{G}^{N_o}|, \tag{C.71}$$

where $\mathcal{G}^{N_o}$ is the quotient space under the equivalence relation (5.63), i.e. $\mathcal{G}^{N_o} \triangleq \mathcal{L}^{N_o}/\sim$. The determination of $|\mathcal{G}^{N_o}|$ is the subject of the next lemma.

[2] Notice that in the case of $\alpha = 0$, we have $\mathcal{L}_i = \mathcal{M}\ \forall\ i$ equiprobably, i.e. null communication rate. When $\alpha > 0$, a positive communication rate is always achievable.

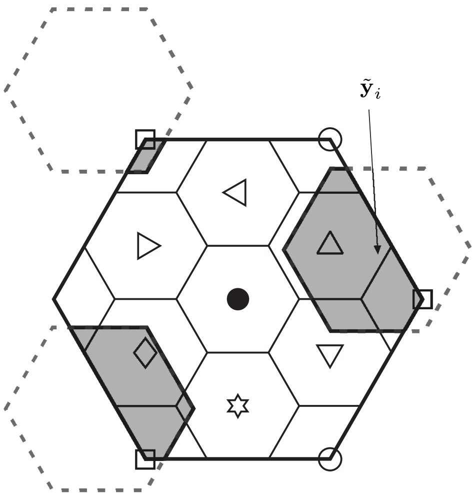


Figure C.2: Illustration of the region $(\tilde{\mathbf{y}}_i - \mathbf{t} - \mathcal{Z}(\Lambda)) \mod \Lambda$ for the lattice code of Figure 5.1(a) with $\alpha = 0.5$. Notice that 3 different messages are feasible in this case.

Lemma C.3 (Cardinality of the quotient space). For $\mathcal{L}^{N_o}$ with $|\mathcal{L}_i| \leq \lceil \frac{p}{2} \rceil$, $\forall\, i = 1, \ldots, N_o$, consider the equivalence relation defined in (5.63) and the quotient space under this equivalence relation, $\mathcal{G}^{N_o} = \mathcal{L}^{N_o} / \sim$. The cardinality of the quotient space is given by

$$|\mathcal{G}^{N_o}| = \sum_{k=1}^{N_o} \left(\prod_{i=k+1}^{N_o} |\mathcal{L}_i| \prod_{i=1}^{k-1} (|\mathcal{L}_i| - 1) \right). \tag{C.72}$$

Proof:

Let us assume, without loss of generality, that $\mathcal{L}_i = \{0, \ldots, |\mathcal{L}_i| - 1\}$. The number of equivalence classes can be obtained in closed form by taking into account the geometrical structure that the equivalence relation (5.63) defines in $\mathcal{M}^{N_o}$, which is illustrated in Figure C.3 for $N_o = 2$, $p = 5$, $|\mathcal{L}_1| = |\mathcal{L}_2| = 3$. The 25 dots in the shaded area represent $\mathcal{M}^{N_o}$. The nine circled dots in the lower left corner represent $\mathcal{L}^{N_o}$, whereas the remaining circled dots are identical modulo p to $\mathcal{L}^{N_o}$. The elements of $\mathcal{L}^{N_o}$ that belong to the same equivalence class fall in the same diagonal (taking into account the modulo-p reduction). Thus, the number of equivalence classes is determined by the number of different diagonals in $\mathcal{M}^{N_o}$ that hit the points in $\mathcal{L}^{N_o}$. As in the figure, under the condition $|\mathcal{L}_i| \leq \lceil \frac{p}{2} \rceil$, $\forall\, i = 1, \ldots, N_o$, this number is given by the number of elements of $\mathcal{L}^{N_o}$ contained in the N_o-dimensional hyperplanes defined by $\{m_i = 0,\ i = 1, \ldots, N_o\}$. In other words, it is the number of sequences in $\mathcal{L}^{N_o}$ which contain zeros in any of their coordinates. A simple calculus reveals that $|\mathcal{G}^{N_o}|$ is given by Eq. (C.72).

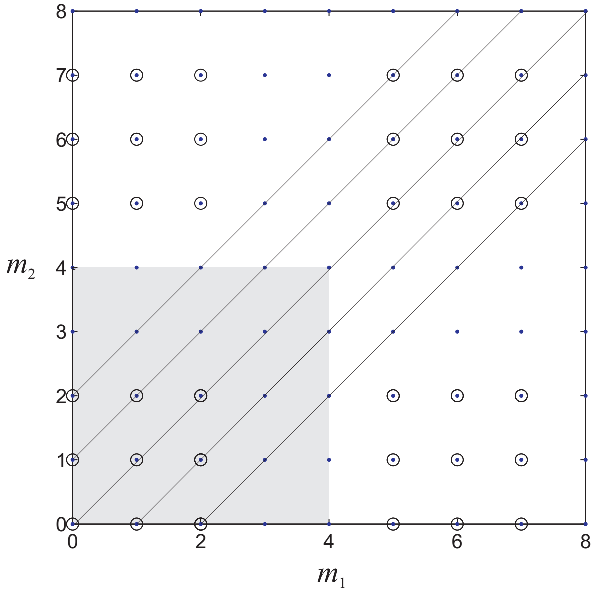


Figure C.3: Geometrical interpretation of the equivalence classes involved in the proof of Lemma C.3 for $N_o = 2$, $\mathcal{L}_1 = \mathcal{L}_2 = \{0, 1, 2\}$, and $p = 5$. The circled dots represent $\mathcal{L}^{N_o}$ and its modulo-p equivalent sequences according to (5.63). The points falling on the same diagonal line belong to the same equivalence class.

■

Step 3)
By combining (C.69) and (C.70), we have

$$H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}) - H(M_1, \ldots, M_{N_o}|\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_o}, \mathbf{t} = \mathbf{0}) \geq \log\left(\frac{|\mathcal{K}^{N_o}|}{|\mathcal{L}^{N_o}|}\right). \tag{C.73}$$

If $|\mathcal{L}_i| \leq \lceil \frac{p}{2} \rceil \ \forall\, i = 1, \ldots, N_o$, using (C.71) and the result of Lemma C.3 we can write

$$\begin{aligned}\log\left(\frac{|\mathcal{K}^{N_o}|}{|\mathcal{L}^{N_o}|}\right) &= \log\left(p \cdot \frac{|\mathcal{G}^{N_o}|}{|\mathcal{L}^{N_o}|}\right) = \log\left(p \cdot \sum_{k=1}^{N_o} \frac{\prod_{i=1}^{k-1}(|\mathcal{L}_i| - 1)}{\prod_{i=1}^{k}|\mathcal{L}_i|}\right) \\ &= \log(p) + \log\left(1 - \prod_{k=1}^{N_o}\left(\frac{|\mathcal{L}_k| - 1}{|\mathcal{L}_k|}\right)\right).\end{aligned} \tag{C.74}$$

The values of $|\mathcal{L}_k|$ above depend on each particular realization of the observations. Thus, in order to obtain the bound to (2.22) we need to average (C.74) over the observations:

$$\begin{aligned} g(N_o) &\geq \log(p) + E\left[\log\left(1 - \prod_{k=1}^{N_o}\left(\frac{|\mathcal{L}_k| - 1}{|\mathcal{L}_k|}\right)\right)\right] \\ &\geq \log(p) + \log\left(1 - \left(1 - \lceil p/2 \rceil^{-1}\right)^{N_o}\right), \end{aligned} \tag{C.75}$$

where for the second inequality we have set $|\mathcal{L}_k| = \lceil \frac{p}{2} \rceil$, $\forall\, k$, in the computation of the expectation. Lemma 5.5 follows by realizing that the right hand side of (C.75) goes to $\log(p)$ for $N_o \to \infty$.