IST-2002-507932

ECRYPT

European Network of Excellence in Cryptology

Network of Excellence
Information Society Technologies

D.WVL.1
First Summary Report on Fundamentals

Due date of deliverable: 31st January 2005
Actual submission date: 18th March 2005

Start date of project: 1 February 2004
Duration: 4 years

Lead contractor: Katholieke Universiteit Leuven (KUL)

Revision 2.0

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>PU</th>
<th>PP</th>
<th>RE</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Public</td>
<td>Restricted to other programme participants (including the Commission services)</td>
<td>Restricted to a group specified by the consortium (including the Commission services)</td>
<td>Confidential, only for members of the consortium (including the Commission services)</td>
</tr>
</tbody>
</table>
The work described in this report has in part been supported by the Commission of the European Communities through the IST program under contract IST-2002-507932. The information in this document is provided as is, and no warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.
## Contents

1 Introduction and state of the art

2 “Watermarking security: theory and practice” [CFF05]
   2.1 Attacks Classification based on Diffie-Hellman’s approach
   2.2 Translation of Shannon’s approach to cryptography
   2.3 Security Analysis of Spread Spectrum
   2.4 Practical methods for hacking systems

3 Fundamental definitions

4 Theoretical evaluation of security
   4.1 Spread Spectrum Watermarking
      4.1.1 Known Message Attack
      4.1.2 Comparison with the result in [CFF05]
      4.1.3 Watermarked Only Attack
      4.1.4 Estimated Original Attack
   4.2 Introduction to Side-information-based methods
   4.3 Random codebooks (Costa’s construction)
      4.3.1 Known Message Attack
      4.3.2 Watermarked Only Attack
      4.3.3 Estimated Original Attack
   4.4 Distortion Compensated - Dither Modulation (DC-DM)
      4.4.1 Known Message Attack
      4.4.2 Watermarked Only Attack
      4.4.3 Estimated Original Attack
   4.5 Oracle attack for scalar DC-DM
   4.6 Comparison among spread-spectrum, Costa and scalar DC-DM

5 Real scenarios

6 Conclusions and further work

A Calculation of mutual information for a single observation
   A.1 Mutual information for spread spectrum
A.1.1 Known Message Attack (KMA) ................................. 33
A.2 Mutual Information for Costa’s scheme .......................... 34
  A.2.1 Known Message Attack (KMA) .............................. 34
  A.2.2 Watermarked Only Attack (WOA) ......................... 35
  A.2.3 Estimated Original Attack (EOA) ......................... 35
A.3 Mutual information for scalar DC-DM ........................... 37
  A.3.1 Known Message Attack (KMA) .............................. 37
  A.3.2 Watermarked Only Attack (WOA) ......................... 38

B Calculation of mutual information for multiple observations 39
  B.1 Mutual information for spread spectrum .................... 39
    B.1.1 Known Message Attack (KMA) ............................ 39
  B.2 Mutual information for scalar DC-DM ....................... 39
    B.2.1 Known Message Attack (KMA) ............................ 39
    B.2.2 Watermarked Only Attack (WOA) ....................... 42

C Relationship between Fisher Information Matrix and Entropy 43

D Optimal selection of the distribution for the dither in DC-DM 43
1 Introduction and state of the art

Although a great amount of the watermarking and data-hiding literature deals with the problem of robustness, little has been said about security, and even in this time of relative maturity of watermarking research no consensus has been reached about its definition, and robustness and security continue to be often seen as overlapping concepts. The purpose of this first section is to give an overview of the evolution of research on watermarking security, keeping in mind that this document only deals with a signal processing approach.

First, the notation and a general model for the evaluation of watermarking security will be introduced. The model is depicted in Figures 1-a and 1-b: a message $M$ will be embedded in an original document $X$ (the host), yielding a watermarked vector $Y$. The embedding stage is parameterized by the embedding key $\Theta_e$, and the resulting watermark is $W$. In the detection/decoding stage, the detection key $\Theta_d$ is needed; $\hat{M}$ denotes the estimated message in the case of decoding, and the decision whether the received signal is watermarked or not in the case of detection. Capital letters denote random variables, and bold letters denote vectors. Let us remark the fact that the different embedding and detection keys ($\Theta_e$ and $\Theta_d$, respectively) in Figure 1 are envisaged to encompass general watermarking/data-hiding scenarios, although most of the work in this document deals with scenarios where embedding and detection keys are the same.

Since the beginning, watermarking was identified with security. Perhaps, one of the main reasons for this misconception is the fact that watermarking was conceived as a solution to the problem of illegal copy control. Due to this misconception, researchers focused their efforts during the first years on the design and study of attacks and countermeasures, overlooking the meaning of security in watermarking. As a result, most of the literature deals with the problem of robustness; at most, there was the notion of intentional and non-intentional attacks: those attacks which only imply common signal processing operations (filtering, compression, etc.) were separated from those considered as intentional (sometimes called malicious), aimed at removing or estimating the watermarks. The work in [CL98] shows an example of this type of classification, considering separately the so-called signal transformations (affine transformations, noise addition, compression) and the intentional attacks, introducing at a qualitative level concepts like the sensitivity attack, the statistical averaging attack (which is closely related to the collusion attack [KLM+97]) and attacks based on the availability of embedding devices. The sensitivity attack belongs to the category of oracle attacks, i.e. attacks which exploit the availability of a watermark detector, and it will explained below. The statistical averaging attacks are based on the fact that, if multiple documents (say, images) with the same embedded watermark are available, it is possible to estimate the watermark by averaging all these watermarked documents: if $x_i$ denotes the $i$-th watermarked document, $w$ denotes the watermark, and there are $N$ different watermarked documents, then the sum $Nw + \sum_i x_i$ tends to $Nw$. A similar attack may be performed to estimate the original document when a great number of versions of the same document with different watermarks are available. It is interesting to note that in the classification given in [CL98] there is an underlying notion of security, because the signal transformations can be identified with mere (say, blind) signal processing operations, but intentional attacks go beyond, trying to exploit knowledge of the watermarking algorithms.

In the late nineties, the most popular algorithm for watermarking was spread spectrum
Figure 1: General model for security analysis: embedding (a) and decoding/detection (b)

[CKLS97], considered to provide high robustness in detection besides being highly secure, but contrary to what was believed, it was shown in [CL97] that for a host of length $N$, when a detector which provides binary outputs (i.e. if the watermark is present or not) is available, an attack aimed at estimating the watermark would require a number of tries of order $O(N)$, instead of $O(2^N)$. This result was illustrated with the so-called sensitivity attack which, roughly speaking, consists in the iterative modification of the coefficients of the watermarked vector to estimate the boundary of the detection region by observing the outputs of the detector. For the case of watermarking schemes like spread spectrum, knowledge of the detection region implies knowledge of the watermark, which may be seen as disclosure of the secret key, thus permitting an attacker to forge contents at his will. It could be said that the aforementioned sensitivity attack raised up the problem of security in watermarking: if an algorithm could be broken with such relative ease, then it could be barely thought of as secure. The sensitivity attack was extensively studied for detectors based on linear correlation: in [LvD98], a complete characterization of the problem is given, and even an information-theoretic analysis is performed, measuring the information about the watermark that an attacker can gain by each observation of the detector output (this is a pioneering work in this sense); later, and following the ideas in [LvD98], a practical method for accomplishing the sensitivity attack was proposed in [KLvD98], showing alarmingly good results.

The work in [Mit99], inspired by previous works [Cac98] and [ZFK+98] in the field of steganography, is the first attempt at proposing a theoretical framework for assessing the security of a general watermarking scenario. The two main issues of this paper are the perfect secrecy of the embedded message and the robustness of the embedding, characterizing both in terms of mutual information: in the general model of [Mit99], a message $M$ is embedded in a host $X$ using a secret key $\Theta_e$ which is a pseudorandom sequence known only by authorized users, yielding a watermarked signal $Y$. This signal may undergo attacks which are modeled by a probabilistic channel, resulting in the attacked signal $Z$. If the mutual information $I(M;Y)$ equals 0, then the system is said to achieve perfect secrecy, because no information about the embedded message can be inferred from the observed watermarked signals, whenever the secret key $\Theta_E$ is unknown. On the other hand, robustness is measured by the mutual information $I(M;Z|\Theta_d)$. The notion of perfect secrecy is directly borrowed from the seminal work by Shannon in [Sha49], where the information-theoretic fundamentals of cryptanalysis were established. Under these premises, the secrecy of two simple schemes (permutation modulation and cyclic shift modulation) is evaluated in [Mit99]. However, this approach does not take into account that some information about the secret key may leak from the observations,
giving advantage to the attacker. Summarizing, the major contributions of this paper to the field of watermarking security are:

1. the consideration of the trade-off between robustness and security;

2. the analyzed scenario covers data-hiding, contrarily to all previous works, which focused only on one-bit watermarking;

3. the application, for the first time, of the information-theoretic tools developed by Shannon for the purpose of cryptanalysis.

Most of watermarking methods belong to the so-called category of symmetric schemes, i.e. those watermarking algorithms whose embedding and detection keys are the same; this means that disclosure of the detection key implies a total break of the system, allowing pirates to forge contents at their will. Moreover, bear in mind that estimation of the detection key is relatively easy by means of the sensitivity attack if a detector is available to the attacker, as it is in a great variety of watermarking applications. This is one of the reasons why researchers put their efforts on the design of asymmetric schemes, i.e. those schemes in which the embedding and detection keys do not need to be necessarily the same, in such a way that the impact of attacks revealing the detection key is reduced at a great extent. This and other advantages of asymmetric watermarking schemes are discussed in [FVD01]: in this paper, four asymmetric watermarking methods are analyzed and unified in such a way that the detection function may always be written as a quadratic form. The main conclusion is that, although its robustness is worse than that of symmetric schemes, the security level can be improved because we are passing from linear detection functions to quadratic ones, thus increasing the complexity of oracle attacks (this complexity may be progressively increased by the use of n-th order detection functions). However, a rigorous security analysis of this kind of methods is still a pending issue. The problem of security on asymmetric watermarking was also addressed in [BB04] in a similar way to [FVD01].

After some years of research in the field, it was fairly clear the existence of important problems of security in watermarking, but there was no attempt to clarify the concept of security in this framework. In view of this situation, the work in [Kal01] came to shed some light on the meaning of watermarking security. In a context of robust watermarking, the following definitions are given:

- “Robust watermarking is a mechanism to create a communication channel that is multiplexed into original content”, and whose capacity “degrades as a smooth function of the degradation of the marked content”.

- “Security refers to the inability by unauthorized users to have access to the raw watermarking channel”. Such an access refers to “remove, detect and estimate, write and modify the raw watermarking bits”.

Hence, watermarking security is identified with attacks whose objective is not only the removal of the watermarks, as it is with robustness. The problem with the given definition of security is the fact that it is too general, in such a way that it does not reflect some crucial aspects, one of them concerned with the intentionality of the attacks: from the discussion in that
paper it is inferred that every attack (intentional or not) may result in a threat to security. Notice that if, for instance, any user performs a JPEG compression of an image (e.g. before transmitting it via Internet) resulting in watermark removal, this should not be considered an attack to security, as it is obviously concerned with robustness. In Section 3, this and other aspects which must be encompassed in the definition of watermarking security are discussed.

At least, the former definition of security given in [Kal01] allows to make a clear distinction between cryptography and watermarking, clarifying an usual misconception in the watermarking community, namely that watermarking is often considered as an analog of cryptography because both disciplines are concerned with media security; according to [Kal01] “the purpose of watermarking is to provide a means for transferring bits from sender to receiver; the purpose of cryptography is to provide a security layer on top of the watermarking channel”.

Based on the definitions of security and robust watermarking, a classification of watermark attacks according to different criteria is proposed in [Kal01]. The main classification was already given in [CMB02], and it establishes the division in unauthorized watermark removal, detection (estimation), writing, and modification. Other parameters for classifying attacks are the following:

- The degree of success of the attacks, distinguishing between minimal, partial and full success: for instance, if an estimation attack is only aimed at testing the presence of a watermark this would be a minimal attack, but if the target of the estimation attack is the differentiation between all the possible embedded messages, this would be a full attack.

- The amount of information available to the attacker, considering:
  
  - The number of watermarked objects available to the attacker: he may have access only to a single watermarked object, but sometimes he can manage to obtain multiple documents containing the same watermark, or different watermarked versions of the same document, etc.
  
  - The availability of embedding and/or detection engines, which may help to the attacker to refine his knowledge about the secrets of the system.
  
  - The degree of knowledge of the watermarking algorithms: the attacker may refine his attacks by taking into account his knowledge of the system under attack, exploiting its weaknesses.

- The degree of universality of the attack: the only objective of the attacker may be the removal of the watermark from a certain document, or he may be interested in learning global secrets of the system under attack (the secret key, for instance).

In [F+02], Kalker’s definitions ([Kal01]) of watermarking and Security are reviewed. It also analyzes the idea that security and robustness are different concepts; security has a broader scope, since it does not only deal with watermark removal but also with unauthorized embedding and detection (resembling removal, writing and detection/estimation attacks to security from [Kal01]).

Nevertheless, some conceptual differences appear between these two works. In [F+02] it is said that “security deals only with intentional attacks, whereas robustness measures the impact
of classical content transformations on the detectability of the watermark” being “inmaterial” for robustness “whether such transformations are intentional or not”. It also continues making a clear distinction between robustness and security: robustness deals with blind attacks (offering a partial break of the watermarking technique used) whereas “security deals with malicious attacks where the adversary has very good knowledge of the watermarking technique being used” (offering a complete break). This is an evolution of the concept of security from [Kal01] approach, since in [Kal01] it is said that “any modification to a watermarked object that causes a watermark detection engine to fail constitutes in principle a watermark removal attack”. Taking this into account, a content transformation which succeeds in fooling the detector would be classified in [Kal01] as a security issue and in [F+02] as a robustness matter.

Furon et al. also translated Kerckhoff’s principle from cryptography to watermarking: all functions (encoding/embedding, decoding/detection, ...) should be declared as public except for a parameter called the secret key. The definition of security level is also obtained as a corollary of Kerckhoff’s principle: “The level of security is the effort (complexity, time, money, ...) the attacker requires to disclose the secret key”, even when it clarifies later that “attacks might exist where the disclosure of the secret key is not needed to achieve the attacker’s goal”.

Based on Diffie-Hellman’s attacks classification for cryptography, an a classification of attacks for watermarking is proposed. They are classified as:

- Only watermarked content attack: the attacker only has access to some watermarked content.
- Watermarked content pair attack: the attacker has access to original content and their corresponding watermarked versions.
- Chosen original content attack: the attacker has access to a watermark embedder.
- Chosen watermarked content attack: the attacker has access to a watermark detector (oracle attack).

Taking into account the resemblance with cryptography, Shannon’s approach is followed also in [F+02] (firstly adopted in [Mit99]). In Shannon’s framework, the adversary knows the algorithms to be used, except for the secret key (a discrete variable), about which he has only some fuzzy a priori knowledge. Shannon measured this ignorance by the entropy of the key. Since the adversary can see some observed messages, he may get a certain knowledge about the key. Therefore, the final ignorance is measured by the entropy of the key given the encrypted messages. For Shannon, an encryption scheme is perfect if the attacker cannot refine his knowledge about the key whatever the observed encrypted messages (null mutual information). A straightforward translation from cryptography to watermarking is introduced, but some problems appear due to the extension of the discrete case (entropy) to the continuous one (differential entropy). A comparison among the attacks is also introduced [F+02], based on the asymptotic growth of the number of tries/realizations needed to estimate the secret key.

For the sensitivity attack (a kind of oracle attack, see [LvD98]), an interesting result is expounded: “at least at the beginning of the sensitivity attack, the information leakage is
proportional to the number of tries”. Finally, some applications and their attacks are introduced: copy protection (especially for DVD), copyright protection, document authentication and document protection.

Another work in the field of watermarking security (focused on the signal processing approach) that cannot be ignored is [BBF03]. One thought shared by the watermarking community is the fact that security requirements strongly depend on the final application, thus at a first glance, each scenario would require a different analysis, so as in the previous work [Kal01], an effort to decouple analysis from application is made. With this purpose, the classification of attacks in four categories which was given in [F⁺02] based on Diffie-Hellman’s work is recovered. Furthermore, a new framework to analyze watermarking security is proposed: it is based on modeling watermarking as a game with some rules; this rules determine which information (parameters of the algorithm, the algorithm itself, etc.) is public; according to this rules, attacks are classified as

- Fair: the attacker only exploits publicly available information. This is a somewhat unrealistic scenario.
- Unfair: the attacker does not observe the rules of the game, so he tries to access all the possible information which can be of help for him.

The unfair scenario raises up the interesting question on what is the best option to the watermarker. One possibility is to keep secret as much information about the algorithm as possible, in order to maximize the difficulty for fair attackers; however, this implies many potential advantages for unfair attackers. On the other hand, if the watermarker wants to minimize the impact of unfair attacks, then all the information must be made public; obviously, this situation gives major advantages to fair attackers. It seems that the optimal solution will have to reach a compromise between the amount of secret and public information. Keeping this discussion in mind, four scenarios are considered in [BBF03], according to the amount of information made publicly available:

1. No knowledge (the so-called security by obscurity scenario): Kerckhoff’s principle is recalled, similarly to [F⁺02].
2. Knowledge of the embedding and detection algorithms: this corresponds with the usual symmetric watermarking scenario.
4. Knowledge of the detection and embedding keys, besides the algorithms: this is the open cards scenario.

The mutual information between the observations and the secret information provides a measure of the unfair knowledge of the attacker (this idea was already used in [F⁺02]). The concept of security level given in [F⁺02] (and inspired by [Sha49]) is recalled: “is the amount of observation, the complexity, the amount of time, or the work that the attacker needs to gather in order to hack a system”. The rest of the work in [BBF03] deals with the analysis of the four proposed scenarios, being the symmetric one the most deeply analyzed.
2 “Watermarking security: theory and practice” [CFF05]

This work should not be ignored in any future research on watermarking security. Indeed, this paper constitutes the main inspiration for the work developed in Section 4. The first point emphasized in [CFF05] is the recognition of the difficulty of distinguishing between security and robustness; they are “neighboring concepts, which are hardly perceived as different”. A significant evolution from [F+02] is that in [CFF05] “the intentionality behind the attack is not enough to make a clear cut between these two concepts”. To define robustness, the authors complete Kalker’s definition in [Kal01], establishing the cause of the degradation of the marked content: “a classical content processing (compression, low filtering, noise addition, geometric attack...)”. To define security the authors resort again to Kalker’s definition in [Kal01], but excluding from the removal attacks “those already encompassed in the robustness category”.

2.1 Attacks Classification based on Diffie-Hellman’s approach

This classification is slightly different from that proposed in [F+02], although it is also based on Diffie-Hellman’s work. In this case, the classification is:

- Watermarked Only Attack (WOA): the opponent has access to some watermarked vector.
- Known Message Attack (KMA): the opponent has access to some watermarked vectors and the associate messages.
- Known Original Attack (KOA): the opponent has access to some watermarked vectors and the corresponding original ones.

Note that chosen watermarked content attack has disappeared from the classification in [F+02], while the only watermarked content is renamed “watermarked only attack“, and the watermarked content pair and chosen original content attack are unified under the “known original attack”. Note that the oracle attack is not included in this classification, since it was a chosen watermarked content attack.

2.2 Translation of Shannon’s approach to cryptography

The authors adapt Shannon’s concept of perfect secrecy to watermarking, coining the term perfect covering, which is achieved when the observation of watermarked vectors does not allow to get knowledge about the watermark or the secret key. The measure proposed by Shannon (for discrete variables) was the equivocation

$$H(\Theta|Y_1,Y_2,\cdots,Y_N) = H(\Theta) - I(\Theta;Y_1,Y_2,\cdots,Y_N)$$

so the information leakage can be measured by the mutual information between the observations and the secret key. Based on this measure, the unicity distance is introduced in watermarking for the first time, meaning the minimum number of observations for which the equivocation becomes null.
The discussion for watermarking uses continuous variables, and in [CFF05] it is claimed that “the entropy (or the conditional entropy) of a continuous random variable does not measure a quantity of information, since, for instance, the equivocation can take positive or non positive values, ruining the concept of unicity distance”. This argument is somewhat questionable, since the differential entropy of a continuous random variable also indicates its variability or the uncertainty about it. In the same way that (discrete) entropy is related to the number of possible values and their probabilities, the (continuous) differential entropy is related to the volume of the typical set [CT91]. The latter may yield negative values ($-\infty$ for a deterministic variable, for example) but it should not be disregarded only due to this. Section 4 deals with this point, where spread-spectrum methods (which are analyzed in [CFF05]) are reviewed from an information-theoretic perspective and new results for quantization-based schemes (Costa and DC-DM) are also introduced.

Due to the avoidance of negative entropies, a different measure was proposed in [CFF05] to solve this problem: the Fisher measure. The Fisher Information Matrix [Fis22], through the Cramér-Rao theorem [Cra99] gives a lower bound on the covariance matrix of an unbiased estimator of a parameter (in watermarking, the secret key). This way, the larger the information leakage is, the more accurate the estimation of the secret key is. Taking this into account, in [CFF05] the auxiliary variable

$$N_o^* = N_o \text{tr}(\text{FIM}(\Theta)^{-1})$$

is introduced, where $N_o$ is the number of observations, and $\text{FIM}(\Theta)$ is the Fisher Information Matrix of the secret key. Applying Cramér-Rao’s inequality, the covariance matrix of the estimate of the parameter when $N_o$ observations are available ($R^N_o$) can be lower-bounded as

$$R^N_o \geq \text{FIM}(\theta)^{-1},$$

in such a way that

$$\text{tr} \left( R^N_o \right) \geq \text{tr} \left( \text{FIM}(\theta)^{-1} \right).$$

If the parameter to be estimated were Gaussian, we could write

$$\frac{\text{tr} \left( R^1 \right)}{N_o} = \text{tr} \left( R^N_o \right),$$

where $R^1$ is the covariance matrix of the estimate of the parameter when only one observation is available. Taking into account (4) and (5), it is possible to write

$$\text{tr} \left( R^1 \right) \geq N_o \text{tr} \left( \text{FIM}(\theta)^{-1} \right) = N_o^*,$$

so $N_o^*$ can be seen as a lower-bound on the variance of the estimation error when only one observation is available. Following this idea, the larger the variance of the error for one observation is, the more secure the system will be.

In [CFF05] this notion is said to be close to the unicity distance, and that is why the security level is also said to be $O(N_o^*)$. In Section 4.1.2 it will be shown that this measure is
neglecting some important parameters as the uncertainty (differential entropy) of the secret key or the watermarked signal. The proof will be based on another paper by Costa [CC84]. Other drawbacks of the use of FIM are its computational complexity (especially for side-informed schemes) and the cumbersome interpretation of some results.

Although a security analysis is also performed for the substitutive method, we will focus on the spread spectrum analysis.

### 2.3 Security Analysis of Spread Spectrum

Applying Bayes’ rule, the authors show that spread-spectrum-based watermarking does not provide perfect covering (so if the attacker has only access to watermarked pieces of content, some information about the watermark signal will leak from these observations).

When studying the KMA case, Gaussianity and independence between the host and the watermark signals is assumed in [CFF05]. The information leakage is shown to be linear with the number of observations. In fact, \( N^* = O(DWR_c) \), where \( DWR_c \) is the Document to Watermark Ratio per carrier.

For the KOA, the problem resembles a blind source separation problem. Following this approach, at best for the opponent, the secret carriers are identified up to a signed permutation ambiguity.

Finally, for the WOA case, the sources (messages to be sent) are unknown and can be regarded as nuisance parameters. These nuisance parameters render the estimation of the secret key less accurate. Moreover, “an ambiguity remains concerning the order and phase of the carriers; the system is only identifiable up to a signed permutation”. In any case the security level is shown to be the same against KMA and WOA.

It is also interesting to bring the classification of the possible hacks given in [CFF05]:

- The pirate discloses the subspace generated by the secret key.
- The pirate discloses the secret vectors up to a signed permutation.
- The pirate discloses the secret vectors.

The quality of the hacked signals will depend on the accuracy of the estimation.

### 2.4 Practical methods for hacking systems

In [CFF05] Principal Component Analysis (PCA) and Independent Component Analysis (ICA) algorithms are used to achieve access to the watermarking communication channel to remove, read or write hidden data. A previous work which used ICA to erase the watermark signal, but not to disclose of the secret parameters is also introduced in [CFF05]. We can also mention here the existence of a simultaneous and independent work ([DD04]), where the subspace generated by the the secret key is estimated with PCA in order to remove the watermark.
Finally, some empirical results are given for side-informed watermarking (Improved Spread Spectrum, Scalar Costa Scheme and Maximized Robustness Embedding). Costa and SCS will be theoretically analyzed in Section 4.

3 Fundamental definitions

In this section, some thoughts about the concept of watermarking security are expounded and some definitions are proposed. First, in order to establish a clear line between robustness and security, the following definitions are put forward for consideration:

- Attacks to robustness\(^1\) are those whose target is to increase the probability of error of the data-hiding channel.

- Attacks to security are those aimed at gaining knowledge about the secrets of the system (e.g. the embedding and/or detection keys).

At first glance, in the definition of attacks to robustness we could have used the concept of channel capacity instead of the probability of error, but this entails some potential difficulties: for instance, an attack consisting on a translation or a rotation of the watermarked signal is only a desynchronization, thus the capacity of the channel is unaffected, but depending on the watermarking algorithm, the detector/decoder may be fooled. Another considerations about security, taking into account the above definitions, are the following:

- About the \textit{intentionality} of the attacks: attacks to security are obviously intentional, but not all intentional attacks are threats to security. For instance, an attacker may perform a JPEG compression to fool the watermark detector because he knows that, under a certain JPEG quality factor, the watermark will be effectively removed. Notice that, independently of the success of his attack, he has learned nothing about the secrets of the system. Hence, security implies intentionality, but the converse is not necessarily true.

- About the \textit{blindness} of the attacks: \textit{blind} attacks are those which do not exploit any knowledge of the watermarking algorithm. Since attacks to security will try to disclose the secret parameters of the watermarking algorithm, it is easy to realize that they can not be blind. On the other hand, a \textit{non-blind} attack is not necessarily targeted at learning the secrets of the system; for instance, in a data-hiding scheme based on binary scalar Dither Modulation (scalar DM), if an attacker adds to each watermarked coefficient a quantity equal to half the quantization step, the communication is completely destroyed because the bit error probability will be 0.5, although the attacker has learned nothing about the secrets of the systems. Hence, security implies \textit{non-blindness}, but the converse is not necessarily true.

- About the \textit{final purpose} of attacks: many attacks to security constitute a first step towards performing attacks to robustness. This can be easily understood with a simple

\(^1\)Robustness could be defined as the ability of data-hiding system to satisfy certain parameters of performance according to a given application for the specified attacking efforts.
example: an attacker can perform an estimation of the secret pseudorandom sequence used for embedding in a spread-spectrum-based scheme (attack to security); with this estimated sequence, he can attempt to remove the watermark (attack to robustness).

• About the distinction between security and robustness: a watermarking scheme can be extremely secure, in the sense that it is (almost) impossible for an attacker to estimate the secret key(s), but this does not necessarily affect the robustness of the system. For instance, the boundary of the detection region of watermarking algorithms whose decisions are based on linear correlation can be complicated by using, as a decision boundary, a fractal curve [BB04]; this way, security is highly improved since, for example, sensitivity-like attacks are no longer effective because the boundary of the detection region is extremely hard to describe. However, this countermeasure against security attacks does not improve anyway the robustness of the method.

• About the measure of security itself: security must be measured separately from robustness. The following analogy with cryptography may be enlightening in this sense: in cryptography, the objective of the attacker is to disclose the encrypted message, so the security of the system is measured assuming that the communication channel is error-free; otherwise it makes no sense to measure security, since the original message was destroyed both for the attacker and fair users. By taking into account the definition of robustness given at the beginning of this section, the translation of this analogy to the watermarking scenario means that security must be measured assuming that no attacks to robustness occur.

The measure of security proposed here is a direct translation of Shannon’s approach [Sha49] to the case of continuous random variables. Furthermore, we will take into account Kerckhoff’s principle [Ker83], namely that the secrecy of a system must depend only on the secret keys. Security will be evaluated in the two scenarios of Figure 1, but in the concrete case of symmetric watermarking, i.e. \( \Theta_e = \Theta_d = \Theta \).

1. For the scenario depicted in Figure 1-a, security is measured by the mutual information between the observations \( Y \) and the secret key \( \Theta \)

\[
I(Y^1, Y^2, \ldots, Y^{N_o}; \Theta) = h(Y^1, Y^2, \ldots, Y^{N_o}) - h(Y^1, Y^2, \ldots, Y^{N_o} | \Theta) = h(\Theta) - h(\Theta | Y^1, Y^2, \ldots, Y^{N_o}),
\]

where \( h(\cdot) \) stands for differential entropy, and \( Y^n \) denotes the \( n \)-th observation\(^2\). Equivocation is defined as the remaining uncertainty about the key after the observations:

\[
h(\Theta | Y^1, Y^2, \ldots, Y^{N_o}) = h(\Theta) - I(Y^1, Y^2, \ldots, Y^{N_o}; \Theta).
\]

This scenario encompasses attacks concerning the observation of watermarked signals, where it is possible that additional parameters like the embedded message \( M \) or the host \( X \) are also known by the attacker. The model is valid for either side-informed and non-side-informed watermarking/data-hiding schemes.

\(^2\)The observations are independent signals watermarked with the same secret key \( \Theta \)
2. The scenario depicted in Figure 1-b covers the so-called oracle attacks. In this case, the attacker tries to gain knowledge about the secret key $\Theta$ by observing the outputs $M$ of the detector/decoder corresponding to some selected inputs $Y$, so the information leakage is measured by

$$I(\hat{M}^1, \ldots, \hat{M}^{N_o}, Y^1, \ldots, Y^{N_o}; \Theta),$$

where, in this case, $Y^n$ are not necessarily watermarked objects but any arbitrary signal, for instance the result of the iterations of an attacking algorithm.

The translation of Shannon’s approach to the continuous case is straightforward; we only must be careful with the concept of differential entropies, in order to redefine properly the unicity distance for continuous random variables: an attacker will have perfect knowledge of the key when $h(\Theta|Y^1, Y^2, \ldots, Y^{N_o}) = -\infty$. Hence, the security level is the number $N_o$ of observations required to reach the unicity distance. However, since this number is $\infty$ in general, the security level could be measured by the growth-rate of mutual information with the number of observations $N_o$; another possibility is the establishment of a threshold in the value of the equivocation, which is directly related to the minimum error in the estimation of the key:

$$\sigma_E^2 \geq \frac{1}{2\pi e} e^{2h(\Theta|Y)}.$$  

(11)

For an attack based on the key estimate, its probability of success is given by the variance of the estimation error. This way, the security level is defined as the minimum number of observations $N_o^*$ needed to achieve the variance of the estimation error which yields the required probability of success.

For the measure of security to be well defined, at least two of the quantities in (9) must be given, because important information about the security of the system may be masked when only one of those quantities is available:

- The value of $h(\Theta)$ is only the a priori uncertainty about the key, so it does not depend on the system itself.
- The value of $I(Y^1, Y^2, \ldots, Y^{N_o}; \Theta)$ shows the amount of information about the key that leaks from the observations, but a smaller information leakage does not necessarily imply a higher security level: notice that, for example, a deterministic key would yield null information leakage, but the security is also null.
- The value of the equivocation $h(\Theta|Y^1, Y^2, \ldots, Y^{N_o})$ is indicative of the remaining uncertainty about the key, but it does not reflect what is the a priori uncertainty.

4 Theoretical evaluation of security

In this section some theoretical results about the residual entropy of the secret key for spread-spectrum and side-information-based methods will be presented. The notation is borrowed from [CFF05]: $N_v$ will denote the length of the vectors (number of samples in each observation), $N_o$ the number of observations, and $N_c$ the number of carriers (or hidden symbols). After some modifications in the nomenclature described in [CFF05], the following attacks will be analyzed:
• Known Message Attack (KMA): In this case the mutual information between the received signal and the secret key, when the sent message is known by the attacker, should be computed:

\[ I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \Theta | M) = h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | M) - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta, M), \]  

so the residual entropy will be

\[ h(\Theta | \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}, M) = h(\Theta) - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | M) + h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta, M) \]  

(13)

• Watermarked Only Attack (WOA): The mutual information between the observations and the secret key is

\[ I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \Theta) = h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}) - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta) \]

\[ = h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}) - I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; M | \Theta) \]

\[ - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta, M), \]  

and the residual entropy will be

\[ h(\Theta | \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}) = h(\Theta) - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta) + I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; M | \Theta) \]

\[ + h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta, M). \]  

(14)

• Estimated Original Attack (EOA): In this case the following will be computed

\[ I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \Theta | \hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^{N_o}) = h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^{N_o}) \]

\[ - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta, \hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^{N_o}), \]  

\[ \text{where } \hat{\mathbf{X}}^i \triangleq \mathbf{X}^i + \hat{\mathbf{X}}^i \text{ is a estimate of } \mathbf{X}^i \text{ and } \hat{\mathbf{X}}^i \text{ is the estimation error; } \hat{\mathbf{X}}^i \text{ is assumed to have power } E \text{ and to be independent of } \mathbf{X}^i. \]  

The Known Original Attack (KOA) proposed in [CFF05] can be regarded to as a particular case of EOA, where the variance of the original host estimation error is set to 0. On the other hand, when the original host estimation error is \( \sigma_X^2 \), we are in the WOA case, so it can be also seen as particular case of EOA. The attacker could obtain this estimate by averaging several versions of the same host watermarked with different keys, but in order to ensure independence between the key and the estimate, the watermarked version with the to-be-estimated key should not be included in the averaging. Other alternative could be to filter the watermarked signal to compute the estimate of the original host (assuming the resulting signal is independent of the watermark).

Taking into account (16), it is possible to write

\[ h(\Theta | \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}, \hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^{N_o}) = h(\Theta) - h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^{N_o}) \]

\[ + h(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o} | \Theta, \hat{\mathbf{X}}^1, \ldots, \hat{\mathbf{X}}^{N_o}). \]  

(17)

Finally, note that, depending on the method, the secret key could be related to the watermarking scheme parameters (i.e. the spreading sequence in spread-spectrum, the dither sequence in SCS or the codebooks in Costa schemes with random codebooks) through a deterministic function, constructing a Markov chain, in such a way that the attacker could be interested just in estimating the result of this function and not in the secret key itself.
4.1 Spread Spectrum Watermarking

For these methods, \( N_c \) random vectors (the spreading sequences), denoted by \( U_i \), are generated depending on the secret key \( \Theta \). In this way, the embedding function can be written as:

\[
Y_j = X_j + \frac{1}{\sqrt{N_c}} \sum_{i=1}^{N_c} U_i (-1)^{M_j}, \quad 1 \leq j \leq N_o,
\]

with \( Y_j, X_j \) and \( U_i \), \( N_v \)-dimensional vectors. The host is modeled as an i.i.d. Gaussian process, \( X \sim \mathcal{N}(0, \sigma^2_X I_{N_v}) \), and the message letters \( M_i \in \{0,1\} \), being \( Pr\{M_i = 0\} = Pr\{M_i = +1\} = 1/2 \). All of these quantities are assumed to be mutually independent.

4.1.1 Known Message Attack

To compute \( I(Y; U_1, U_2, \ldots, U_{N_c}|M) \) for a generic distribution of \( U_1 \) numerical integration should be used. In Fig. 2 and Fig. 3 the results of numerical integration are shown for both Gaussian and uniform distributions of \( U_1 \) in the scalar case. Those figures show that the information the attacker can not learn (i.e., \( h(U_1|Y) \)) is larger if \( U_1 \) is chosen to be Gaussian.

For \( U_i \) following a Gaussian distribution, the following result is derived in Appendix A.1.1 when the sent symbol is known to the attacker for \( N_v > 1, N_c > 1 \) and \( N_o = 1 \),

\[
I(Y; U_1, U_2, \ldots, U_{N_c}|M) = \frac{N_o}{2} \log \left( 1 + \frac{\sigma^2_U}{\sigma^2_X} \right),
\]

(19)

The result in (19) says that the information that an attacker can obtain is the same whatever the number of carriers, although the entropy of the key is linearly increased (this result applies to a great variety of pdf’s for the key, since by the central limit theorem, the sum of the carriers tends to a Gaussian). This result is also a consequence of the power normalization performed in (18); independently of the number of carriers, the power of the watermark stays constant. In Fig. 2 and 3 the results for \( U_1 \) following Gaussian and uniform distributions when \( N_c = 1 \) are plotted.

In the Appendix, we also analyze the case of one sent bit (\( N_c = 1 \), \( N_o = 1 \), when there are several available observations (\( N_o > 1 \)), all of them watermarked with the same secret key. In App. B.1.1 the following is shown:

\[
I(Y^1, \cdots, Y^{N_o}; U|M) = \frac{1}{2} \log \left( 1 + \frac{N_o \sigma^2_U}{\sigma^2_X} \right).
\]

(20)

This is a remarkable result, as it shows that \( I(Y^1, \cdots, Y^{N_o}; U_1|M_1) \) grows non-linearly with the number of observations, although for large Document to Watermark Ratios (DWR >> 1) and low values of \( N_o \) it grows almost linearly. Moreover, (20) coincides with the capacity of a Gaussian channel with signal power \( \sigma^2_U \) and noise power \( \sigma^2_X/N_o \). This suggests that the best method the attacker should follow for estimating \( U \) is just to average the observations \( Y^i \) (at least this is the case when both the host signal and the watermark are Gaussian distributed). In Fig. 4 the mutual information is compared with its linear version when DWR = 30 dB.
4.1.2 Comparison with the result in [CFF05]

In [CFF05] the security level is said to be $O(N_o^* \sigma^{2})$, where $N_o^* \triangleq N_o \text{tr}(\text{FIM}(\theta)^{-1})$ with FIM($\theta$) the Fisher Information Matrix. In this section we try to link the result obtained in that paper with the one obtained here for spread-spectrum KMA when $N_o = 1$ and $N_c = 1$. In order to do this, we will resort to [CC84].

It is shown in App. C that the FIM will be diagonal with diagonal element $\lambda = 2\pi e e^{-2h(X)/N_v}$, whenever $X$ is Gaussian, in such a way that recalling the definition of security level in [CFF05], the fact that $N_c = 1$ and $N_o = 1$, and assuming $X$ to be Gaussian, we can write

$$\text{tr}(\text{FIM}(\theta)^{-1}) = \frac{N_v}{\lambda} = \frac{N_v e^{2h(Y|\theta)/N_v}}{2\pi e},$$

which for the KMA of Spread Spectrum yields

$$N_o^* = N_o \sigma_X^2,$$

which coincides with the result given in [CFF05] for the same case, except for a term related with the watermark power. This difference can be explained by the normalization of the power of the secret key in [CFF05].

Finally, taking into account that

$$h(\Theta|Y) = h(\Theta) - h(Y) + h(Y|\Theta)$$

it is clear that the measure proposed in [CFF05] can be related to $h(\Theta|Y)$, although the former does not take into account the entropy of the secret key neither the entropy of the...
watermarked signal (in order to take them into account, an additional term should be added). As stated in Sect. 2, both terms are relevant for the analysis of the system, so they should be taken into account. In fact, $h(Y|\Theta)$ for the KMA case is linear with the number of observations, while the mutual information will not increase linearly due to the dependence between observations. The linear approximation is really an upper-bound; the larger the number of observations, the worse this approximation is.

4.1.3 Watermarked Only Attack

Due to the symmetry of the pdf’s, it is possible to conclude that the components of the vector $Y$ are still mutually independent, so for $N_e = 1$ and a single observation, we can write

$$I(Y;U_1) = N_v I(Y_i;U_{1,i}) = N_v (h(Y_i) - h(Y_i|U_{1,i}))$$

(24)

$$= N_v (h(Y_i|M = 0) - h(Y_i|U_{1,i})).$$

(25)

In order to determine this for a generic distribution of $U_1$, numerical integration should be used. In Fig. 2 the results of numerical integration are shown. Once again, the information the attacker can not learn ($h(U_1|Y)$) is larger for the shown cases when $U_1$ is chosen to be Gaussian. Therefore, assuming $U_1$ to be Gaussian, we can write

$$I(Y;U_1) = N_v \left( \frac{1}{2} \log \left( 2\pi e (\sigma_x^2 + \sigma_u^2) \right) - h(Y_i|U_{1,i}) \right).$$

(26)

The rightmost term of (26) must still be numerically computed.
When DWR $<< 1$ we can easily analyze the asymptotic behavior of the mutual information taking into account $h(Y) \approx h(U)$ and $h(Y|U) \approx h(X) + \log(2)$, yielding

\[
I(Y; U) = h(U) - h(X) + \log(2), \tag{27}
\]
\[
I(Y; U|M) = h(U) - h(X). \tag{28}
\]

This explains and quantifies the gap between the WOA and KMA cases, which is exactly $\log(2) = 0.69$ nats. Nevertheless, note that a very small DWR is not practical, since it would yield useless watermarked images. This case has been introduced here only to shed some light into the general behavior of the mutual informations. On the other hand, to compute the gap between a Gaussian and a uniform distribution for $U$, $h(U)$ will be determined in both cases for a constant variance $\sigma_U^2$,

\[
h(U_{Gauss}) - h(U_{unif}) = \frac{1}{2} \log(2\pi e\sigma_U^2) - \frac{1}{2} \log(12\sigma_U^2) = \frac{1}{2} \log \left( \frac{\pi e}{6} \right) = 0.1765, \tag{29}
\]

which will be the asymptotic gap (in residual entropy terms) between the Gaussian and uniform cases for both known and unknown messages (see Fig. 2) when DWR $>> 1$, since for a large DWR both $I(Y; U)$ and $I(Y; U|M)$ are approximately 0.

For $N_c$ carriers we have, similarly to the KMA case, the following mutual information:

\[
I(Y; U_1, U_2, \ldots, U_{N_c}) = N_v I(Y_i; U_{1,i}, U_{2,i}, \ldots, U_{N_c,i}) \tag{30}
\]
\[
= N_v \left( \sum_{j=1}^{N_c} I(Y_i; U_{j,i}|U_{j-1,i}, \ldots, U_{1,i}) \right) \tag{31}
\]
\[
= N_v \left( h(Y_i) - h(Y_i|U_{1,i}, \ldots, U_{N_c,i}) \right) \tag{32}
\]
\[
= N_v \left[ \frac{1}{2} \log(2\pi e(\sigma_x^2 + \sigma_u^2)) - h(Y_i|U_{1,i}, \ldots, U_{N_c,i}) \right], \tag{33}
\]
where the second term of (33) must be numerically computed again.

The case of one sent bit ($N_c = 1$), $N_v = 1$, and several available observations ($N_o > 1$) needs very expensive numerical computations. Practical computations demand reducing the number of available observations to a very small value; in that case, the mutual information will be in the linear region, so no knowledge is obtained about the growth of the mutual information for large values of $N_o$.

4.1.4 Estimated Original Attack

In this case, the attacker will have access to an estimate of the original host signal, with some estimation error denoted by $\tilde{X}$, which is assumed to be i.i.d. Gaussian with variance $E$, in such a way that we can write

$$I(Y; U | X + \tilde{X}) = h(Y | X + \tilde{X}) - h(Y | X + \tilde{X}, U)$$

(34)

Assuming $\sigma^2_X >> E$, $\tilde{X}$ will be almost orthogonal (and therefore independent) to $X + \tilde{X}$, so

$$I(Y; U | X + \tilde{X}) \approx h(MU - \tilde{X}) - h(MU - \tilde{X} | U).$$

(35)

This situation is equivalent to that described in 4.1.3, but replacing $\sigma^2_X$ by $E$, so it is possible to use Fig. 2 for obtaining numerical results, using the Estimation error to Watermark Ratio (EWR), instead of the DWR, in the horizontal axis. When the estimate is perfect, i.e. $\sigma^2_{\tilde{X}} = 0$, the mutual information approaches infinity.

4.2 Introduction to Side-information-based methods

In this section a novel theoretical security analysis will be presented for side-information-based methods. These schemes are based on [Cos83] (where a random codebook is assumed) and its implementations, which try to use structured codebooks ([CW01, EG02, EZ04]). In Fig. 5 the considered framework is represented. The codebook is a function of $\Theta$ and is denoted by $U = f(\Theta)$. Depending on the sent message $m$, one coset in the codebook will be chosen, namely $U_m = g(U, m)$. Taking into account the $N_v$-dimensional host signal $X$ and the distortion compensation parameter $\alpha$, the encoder will look for a sequence $U = h(U_m, X)$ belonging to $U_m$ such that

$$|(U - \alpha X)^t X| \leq \delta,$$

(36)

for some arbitrarily small $\delta$. In practical implementations (like QIM or SCS) this is implemented by quantizing $\alpha X$ to the nearest $U \in U_m$. In any case, the watermark signal will be $W = U - \alpha X$ and the attacker will have access to $Y = X + W$. Finally, the decoder will observe $Z = X + W + N$, where $N$ is the channel noise.

4.3 Random codebooks (Costa’s construction)

For the sake of simplicity, in this subsection we will focus on the analysis of this system when a single observation is available. We will also assume $X$, $W$ and $N$ to be i.i.d. random vectors with distributions $N(0, \sigma^2_X I_{N_v})$, $N(0, P I_{N_v})$ and $N(0, \sigma^2_N I_{N_v})$, respectively.
4.3.1 Known Message Attack

Since the knowledge of the secret key and the sent symbol implies the knowledge of the coset in the codebook (i.e., $U_m$), we can write

$$I(Y;\Theta|M) = h(Y) - I(Y;M) - h(Y|U_M).$$

(37)

In App. A.2.1, we show that if $\alpha > 0.2$, then

$$I(Y;\Theta|M) = \frac{N_v}{2} \log \left[ \frac{P + \sigma_X^2}{(1-\alpha)^2\sigma_X^2} \right].$$

(38)

so

$$h(\Theta|Y,M) = h(\Theta) - \frac{N_v}{2} \log \left[ \frac{P + \sigma_X^2}{(1-\alpha)^2\sigma_X^2} \right].$$

(39)

This is a quite reasonable result. The larger the DWR is, the higher the residual entropy, because the host signal is making it difficult to estimate the secret key. In the same way, the larger $\alpha$, the smaller the residual entropy will be, since the self-noise is reduced and the estimation becomes easier. In Figs. 6 and 7 theoretical results are plotted for different values of $\alpha$.

4.3.2 Watermarked Only Attack

Again, the knowledge of the secret key and the sent symbol implies the knowledge of the coset in the codebook (i.e., $U_m$). Therefore (14) can be rewritten as

$$I(Y;\Theta) = h(Y) - I(Y;M|\Theta) - h(Y|U_M).$$

(40)

In App. A.2.2, it is shown that under the same conditions as in the WOA case (i.e., $\alpha > 0.2$)

$$I(Y;\Theta) = \frac{N_v}{2} \log \left[ \frac{(P + \sigma_X^2)\left\{P\sigma_X^2(1-\alpha)^2 + \sigma_N^2(P + \alpha^2\sigma_X^2)\right\}}{P(P + \sigma_X^2 + \sigma_N^2)(1-\alpha)^2\sigma_X^2} \right].$$

(41)

The power of the channel noise is reducing the maximum reliable rate, so it is making easier the attack (this will be further explained in App. A.2.2). Be aware that when $\sigma_N^2 = 0$, or equivalently, when the maximum reliable rate is achieved, then the uncertainty about the sent symbol is maximum, which complicates the attacker’s work, yielding $I(Y;\Theta) = 0$ (perfect secrecy).
In any case, using (41) we can write

$$h(\Theta|\mathbf{Y}) = h(\Theta) - \frac{N_v}{2} \log \left[ \frac{(P + \sigma_X^2) \{ P \sigma_X^2 (1 - \alpha)^2 + \sigma_N^2 (P + \alpha^2 \sigma_X^2) \}}{P (P + \sigma_X^2 + \sigma_N^2)(1 - \alpha)^2 \sigma_X^2} \right].$$

(42)

In Figs. 8, 9 and 10 theoretical results are plotted. They depend on the DWR, the WNR (Watermark to Noise Ratio) and $\alpha$. Since $I(\mathbf{Y}; \Theta)$ depends on the transmission rate and this depends in turn on the WNR, the WNR has been fixed in order to plot the results. Some conclusions could be:

- The larger $\alpha$, the larger the mutual information; this is explained because a smaller self-noise power is introduced.
- The larger DWR, the smaller the mutual information; it seems to be intuitive that the host-interference will make difficult the estimation of the key.
- The larger the WNR, the smaller the mutual information, because the embedder could achieve a higher reliable rate, so the attacker will have a greater uncertainty about the sent symbol, which makes more difficult his/her job.

### 4.3.3 Estimated Original Attack

In App. A.2.3 it is shown that if $\alpha > 0.2$, then

$$I(\mathbf{Y}; \Theta|\hat{\mathbf{X}}) \approx \frac{N_v}{2} \log \left[ \frac{(P + E) \{ PE(1 - \alpha)^2 + \sigma_N^2 (P + \alpha^2 E) \}}{P (P + E + \sigma_N^2)(1 - \alpha)^2 E} \right],$$

(43)
so

$$h(\Theta|Y, \hat{X}) \approx h(\Theta) - \frac{N_v}{2} \log \left[ \frac{(P + E) \{PE(1 - \alpha)^2 + \sigma_X^2(P + \alpha^2E)\}}{P(P + E + \sigma_X^2)(1 - \alpha)^2E} \right].$$

(44)

Therefore, when the attacker has perfect knowledge of the original host signal, $E = 0$ so $I(Y; \Theta|\hat{X}) = \infty$. The minimum value of the mutual information corresponds to $\alpha = 0$.

It can be seen that (43) is equivalent to (41) but replacing $\sigma_X^2$ by $E$. For that reason, Figs. 8, 9 and 10 are still valid, but replacing the DWR by EWR. In fact, when an estimate of the original host signal is available, the actual host signal can be thought of as being on a sphere centered at that estimate and with squared radius equal to the variance of the estimation error. Since such a shift should not modify the results, this problem must be equivalent to having the host on a sphere with squared radius equal to the variance of the estimation error, but centered at the origin (so $\sigma_X^2$ should be replaced by $E$). Nevertheless, note that the codebook is not designed for this scenario, but for the original one (where the host signal is completely unknown), so the analysis performed in App. A.2.3 is still pertinent.

4.4 Distortion Compensated - Dither Modulation (DC-DM)

We will focus on the scalar version of DC-DM [CW01] (also known as Scalar Costa Scheme, SCS [EBTG03]), for two reasons: first, for simplicity of the analysis, and second, because it provides the fundamental insights into quantization-based methods and illustrates well the theoretical results obtained for Costa’s setup in Section 4.3. The embedding function for scalar DC-DM is given by

$$y_k = x_k + \alpha (Q_{\Lambda}(x_k) - x_k),$$

(45)
Figure 8: $I(Y; \Theta)$ for Costa in nats vs. DWR in dB, for different values of $\alpha$. WNR = 0 dB

where $y_k$ and $x_k$ are samples of the $N_v$-dimensional vectors $Y$ and $X$, which are assumed to be independent and identically distributed (i.i.d.), $Q_{\Lambda_i}(\cdot)$ is an Euclidean quantizer with uniform step size $\Delta$ and its centroids defined by the shifted lattice $\Lambda_i$, according to the to-be-transmitted message symbol $m_i$:

$$\Lambda_i = \Delta \mathbb{Z} + m_i \frac{\Delta}{|\mathcal{M}|},$$

being $|\mathcal{M}|$ the number of different symbols, and finally, $\alpha$ is the distortion compensation parameter. Randomization of the codebook is achieved by means of subtractive dithering [Sch64] using a random parameter $d$, as follows:

$$y_k = x_k + \alpha (Q_{\Lambda_i}(x_k + d_k) - x_k - d_k).$$

The dither signal $d$ may be any function of the secret key $\theta$, i.e. $d = f(\theta)$. If function $f$ is unknown, the only observation of watermarked vectors will not provide any information about $\theta$, thus the target of the attacker is to disclose the dither signal used for embedding, or equivalently the location of the centroids. The security level of the system will depend, obviously, on the statistical distribution of the dither. We show in Appendix D that the entropy of the watermarked signal $Y$ only depends on the modulo-$\Delta$ version of the dither, and furthermore the distribution which maximizes the equivocation is the uniform over the quantization bin; thus, hereafter we will assume that $D \sim U(-\Delta/2, \Delta/2)$ with i.i.d samples. The common assumption when analyzing quantization-based methods is that the host pdf can be modeled as uniform inside each quantization bin and the host variance is large enough to assume that all the centroids are equally likely to occur; this is equivalent to assuming that DWR = $\infty$, so we will refer to this in the sequel as the flat-host assumption. One of the consequences of the flat-host assumption is that the pdf of $Y$ is periodic with period $\Delta$, thus
we can restrict our attention to the modulo-$\Delta$ version of $Y$ without any loss of information (the flat-host assumption is valid for most practical cases of interest, as it was shown in [PFPGV04]). The security weaknesses of this scheme under the above assumptions will be addressed in the following. When necessary, we will compare our theoretical results with those obtained numerically by considering finite DWR’s.

### 4.4.1 Known Message Attack

This is the simplest case to analyze. When only one watermarked vector is observed, we have that the mutual information is simply given by (see Section A.3.1 in Appendix A)

$$I(Y; D | M) = N_v I(Y_i; D_i | M_i) = N_v (\log(\Delta) - \log((1 - \alpha)\Delta)) = -N_v \log(1 - \alpha) \text{ nats}, \quad (47)$$

where $Y_i$ denotes the $i$-th component of vector $Y$, so the equivocation is

$$h(D | Y, M) = h(D | M) - I(Y; D | M) = N_v (\log(\Delta) + \log((1 - \alpha)\Delta)) \text{ nats}. \quad (48)$$

Figure 11 shows the result for the mutual information when $N_v = 1$. For the general case of $N_o$ observations and $\alpha \geq 0.5$ we have (see Section B.2.1 in Appendix B)

$$I(Y^1, Y^2, \ldots, Y^{N_o}; D | M^1, M^2, \ldots, M^{N_o}) = N_v \left( -\log(1 - \alpha) + \sum_{i=2}^{N_o} \frac{1}{i} \right) \text{ nats}, \quad (49)$$

It can be seen in Figure 12 that the first observations provide most of the information about the key, and the growth-rate of the mutual information after 8 or 10 observations is very small. Figure 13 shows numerical results for $\alpha < 0.5$ and 10 observations.
Bear in mind that this results were derived assuming DWR = ∞; however, Figure 14 shows that (for \( N_v = 1 \)) there is no noticeable difference when the DWR is above 30 dB (even for DWR = 10 dB, the difference is almost negligible). When considering finite DWR’s, the mutual informations must be numerically computed, since no closed-form expressions exist for the involved pdf’s. Details for the exact computation of these pdf’s can be found in [PFPGV04].

### 4.4.2 Watermarked Only Attack

In this case, the only information at hand for the attacker is the watermarked vector. Thus, we must calculate the mutual information \( I(Y; D) \), which for one observation \( (N_o = 1) \) results in

\[
I(Y; D) = N_v \left( \log(\Delta) - h(Y_i|D_i = 0) \right) \text{ nats ,}
\]

It is always possible to obtain a theoretical expression for \( h(Y_i|D_i = 0) \); however, for the sake of simplicity we only show the result for the binary case, i.e. \( M_i = \{0, 1\} \) (see Section A.3.2 in Appendix A for details)

\[
h(Y_i|D_i = 0) = \begin{cases} 
\log(2(1 - \alpha)\Delta) & \text{for } \alpha \geq 1/2 \\
\log((1 - \alpha)\Delta) \frac{(1 - 2\alpha)}{1 - \alpha} + \log(2(1 - \alpha)\Delta) \frac{\alpha}{(1 - \alpha)} & \text{for } \alpha < 1/2 
\end{cases}
\]

With the above expressions, derivation of the equivocation is straightforward. In Figure 11, results for 2 and 3 transmitted symbols are shown. It can be seen that when \( \alpha = 0.5 \) the information leakage is null (perfect secrecy); this is because the pdf of the host and that of
the watermarked signal are the same. When $\alpha < 0.5$ the information leakage is very small due to overlappings between adjacent centroids. For the case of multiple observations and $\alpha \geq 0.75$ we have (see Section B.2.2 in Appendix B for details)

$$I(Y^1, Y^2, \ldots, Y^N_o; D) = -\log(1 - \alpha) - \log(2) + \frac{N_o}{2} \sum_{i=2}^{N_o} \frac{1}{i}$$ (52)

$$= I(Y^1, Y^2, \ldots, Y^N_o; D|M_1, M_2, \ldots, M^N_o) - \log(2) \text{ nats} \ . (53)$$

Then, the loss with respect to the KMA case is exactly $\log(2)$ nats (i.e. one bit). For $\alpha < 0.75$ the results are purely numerical.

Finally, it can be seen in Figure 14 that, similarly to the KMA case, the results under the flat-host assumption are sufficiently accurate above moderately high DWR’s. Nevertheless, note that perfect secrecy is no longer achieved when the DWR is very small, even though $\alpha = 0.5$.

4.4.3 Estimated Original Attack

Now, we consider that the attacker has access to watermarked vectors and he manages to build an estimation of the host. Specifically, the estimate of the host is of the form $\hat{X} \triangleq X + \tilde{X}$, being $\tilde{X}$ the estimation error, which we assume to be i.i.d. with its components uniformly distributed on an interval of width $b$, i.e. $\tilde{X}_i \sim U(-b/2, b/2)$. We have to compute the following mutual information:

$$I(Y; D|\hat{X}) = N_v I(Y_i; D_i|\hat{X}_i) = N_v \left( h(Y_i|\hat{X}_i) - h(Y_i|D_i, \hat{X}_i) \right) . \quad (54)$$
A closed-form expression for (54) cannot be obtained, but the following approximation for a binary signaling scheme is tight for small $b$:

$$I(Y_i; D_i|\hat{X}_i) \approx h(R) - \log(2b(1-\alpha)\Delta),$$

with $R = S + T$, being $S$ and $T$ uniform, independent random variables such that

$$S \sim U(-b/2, b/2), \quad T \sim U(-\alpha\Delta/2, \alpha\Delta/2).$$

The approximation (55) is actually a lower bound, whose tightness can be checked in Figure 15, where several plots of the mutual information are represented considering different accuracies in the estimation. In the limiting case when the host is perfectly estimated, we have

$$h(Y_i|D_i, \hat{X}_i) = h(Y_i|D_i, X) = -\infty,$$

so $I(Y; D|\hat{X}) = \infty \forall \alpha$. This result makes sense because when $\alpha = 1$, the exact location of the centroids is revealed after only one observation.

### 4.5 Oracle attack for scalar DC-DM

In this case the attacker is trying to estimate the dither $D$ from the output of the decoder, so it should be measured how much information he/she can obtain by observing the output of the decoder $\hat{M}$, taking into account that he/she knows its input. Hence,

$$I(\hat{M}, Y; D) = h(D) - h(D|\hat{M}, Y).$$

If $\text{DWR} \gg 1$ and $|\mathcal{M}| = 2$, $f_{D|\hat{M}, Y}(d|\hat{M} = \hat{m}, Y = y) = U[-\Delta/4, \Delta/4]$, whereas $f_D(d) = U[-\Delta/2, \Delta/2]$, so

$$I(\hat{M}, Y; D) = \log(2).$$
If the components of $D$ are i.i.d. and the same applies to the components of $Y$, and furthermore both vectors are mutually independent, then we could write

$$I(\hat{M}, Y; D) = N_v \log(2). \quad (59)$$

A decoder which gives binary outputs provide, at most, one bit of information about the key per observation. That information rate can be achieved with a dichotomy algorithm, which reduces the interval where the dither is with the rule $2^{-N_o}$. Therefore,

$$I(\hat{M}^1, \hat{M}^2, \ldots, \hat{M}^{N_o}, Y^1, Y^2, \ldots, Y^{N_o}; D) = N_v N_o \log(2). \quad (60)$$

This can be reached by following the next algorithm for each sample (in this case, for the $i$-th):

```plaintext
lower = x_i;
upper = x_i + Delta/2;
for (iteration = 1; iteration < N_o; iteration++)
    center = (lower + upper)/2;
    if (dec(lower) == dec(medium))
        lower = center;
    else
        upper = center;
end
estimator_i = (lower + upper)/2;
```

where `dec()` is the decoding function, whose output is $\hat{M}$; `estimator_i` is an estimate of the bound between the decoding regions for the $i$-th sample, so $\hat{D}_i =$
mod(- estimator_i + (-1)^{dec(lower)} \Delta/4 + \Delta/2, \Delta/2 - \Delta/4 + \Delta/2). This algorithm could be followed for any other value of |\mathcal{M}|, since the technique to estimate the boundary of the decoding regions is straightforwardly extended. Nevertheless, it will be optimal only when |\mathcal{M}| = 2.

4.6 Comparison among spread-spectrum, Costa and scalar DC-DM

In Fig. 16 the results obtained in this section are compared for the three studied methods in the KMA case, showing a large resemblance between Costa’s and true (numerically obtained) DC-DM results. Moreover, it shows that, when compared under the same conditions, the information leakage for the informed embedding methods is larger than those of spread spectrum. However, note that the security level is not only given by the information leakage but depends also on the entropy of the secret key (see the discussion about (9) in Sect. 3).

In Fig. 17 the residual entropy is also plotted for the studied methods when the watermark power is set to 1 (P = 1) and the WNR = 0 dB. For DC-DM, the larger \alpha, the smaller the residual entropy. Nevertheless, for Costa the curves for \alpha = 0.5 and \alpha = 0.3 intersect; this is due to the dependence h(\Theta) with the DWR and \alpha, since the number of needed centroids depends on these parameters.

5 Real scenarios

Following, some scenarios are introduced and the possible attacks are enumerated:\footnote{Note that the WOA attack is possible in all scenarios.}
Figure 15: Mutual information for scalar DC-DM, with EOA attack, $N_v = 1$ and infinite DWR

**Copy protection**: the detector is available in a great amount of devices so oracle-attacks are possible. Furthermore, the embedded message may be known in many cases (for instance in the copy-protection mechanism DVD messages [BCK+99] copy free, copy once, copy never) so KMA attacks make sense.

**Document authentication**: the receiver must be able to reliably identify the source of a document. It could be performed with fragile, semi-fragile and robust watermarks. KMA makes sense, since the message could be known by the receiver in order to check the authenticity.

**Proof of ownership for IPR protection**: whenever the detector is only available to the watermarker the oracle-attacks are not possible, but when a trusted third party takes part in the process, oracle-attacks could be performed. In this last case also KOA (and maybe KMA) attacks make sense.

**Monitoring of broadcast of distributed images**, for both people metering (evaluation of the broadcast audience) and tracking of illegal exploitation (related to IPR protection). Apart of WOA, in the people metering scenario the oracle attack could be possible.

**Fingerprinting**: the main application is pirate tracing. Cooperative attack is straightforward by averaging, so EOA attacks are possible.

**Information side channel**: the attacks make sense only when the information is private.

**Video watermarking**: In [DD05] the problem of collusion attack in the context of video watermarking is addressed, comparing at a qualitative level several frame-by-frame embedding strategies from a security point of view.
ECRYPT — European NoE in Cryptology

6 Conclusions and further work

In this work, a historical overview of the evolution of the security concept in watermarking and its related attacks is introduced, from seminal works ([CL98], [Mit99]) to an up-to-date state of the art ([BBF03]). Special attention was paid to [CFF05], a very interesting work where security concept is also reviewed, a classification of attacks was presented and quantitative conclusions can be achieved.

Taking into account this background and the complexity of this question, fundamental definitions and a distinction between security and robustness were proposed in the present work, discussing some side aspects of these concepts such as intentionality, blindness or final purpose. A different security measure is also proposed, adapting that introduced by Shannon in the discrete case ([Sha49]) for cryptography systems, which is based on the entropy of the secret key when observations of watermarked signals are available. Main watermarking methods (spread-spectrum, Costa with random codebooks and DC-DM) are analyzed following this measure for the 3 attacks described in [CFF05]. This is also the first time that a security analysis of informed watermarking methods is introduced. From the analysis, the following conclusions are drawn:

- The information leakage of SS is smaller than that of quantization-based schemes. Host-rejection makes it possible to increase the transmission rate, as well as makes easier the secret key estimation by the attacker. In this sense, host-interference is desirable to hide information about the secret key. The problem with SCS is the structured codebook, due to the reduced entropy of the secret key.

- The greater the transmission-rate, the better the secrecy, since the attacker has a greater uncertainty about the sent symbol.
The growth of mutual information is non-linear and decreasing with the number of observations.

The security level of SS depends on the DWR; Costa and DC-DM are independent of the DWR for sufficiently high DWR’s.

The real scenarios where these attacks could appear have been briefly discussed.

Future research lines:

- Theoretical analyses for several observations ($N_o > 1$) as well as oracle attacks should be studied, since here these problems are analyzed only for some particular cases.

- The methods based on quantization after projection (as Spread Transform-Dither Modulation [CW01] or Quantized Projection [PGBH03]) also deserve a security analysis.

- Another interesting approach could be that dealing with the security problem from a game theoretic point of view. The role of the attacker could be to extract as much information as possible to devise an attack (to robustness), and in turn the role of the embedder would be to design both a secure and robust algorithm. This game-theoretic approach clearly links security and robustness.

- If the methods were classified as $\epsilon$-secure, indicating $I(Y^1, \cdots, Y^{N_o}; \Theta) < \epsilon$, it could be studied what would be a good value for $\epsilon$. That value should obviously depend on $h(\Theta)$ and the desired variance of the estimation error.
Acknowledgments

Thanks are due to Dr. Teddy Furon, for his valuable comments on the first version of this deliverable.
A Calculation of mutual information for a single observation

A.1 Mutual information for spread spectrum

A.1.1 Known Message Attack (KMA)

For a single observation \((N_o = 1)\) and \(N_c = 1\), we have

\[
I(Y; U_1|M) = \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} I(Y_i; U_{1,j}|M, Y_{i-1}, \ldots, Y_1, U_{1,j}, \ldots, U_{1,1})
\]

(61)

\[
= \sum_{i=1}^{N_v} I(Y_i; U_{1,i}|M, Y_{i-1}, \ldots, Y_1)
\]

(62)

\[
= \sum_{i=1}^{N_v} I(Y_i; U_{1,i}|M)
\]

(63)

\[
= N_v I(Y_i; U_{1,i}|M),
\]

(64)

where (62) follows from the fact that \(Y_i\) and \(U_{1,j}\) are independent \(\forall i \neq j\); (63) follows from the independence between the components of \(Y\) given the message, and (64) follows from the fact that \(Y\) and \(U_1\) are i.i.d. processes. The theoretical expression for (64) is easy to calculate:

\[
I(Y_i; U_{1,i}|M) = I(Y_i; U_{1,i}|M = 0) = h(Y_i|M = 0) - h(Y_i|M = 0, U_{1,i}),
\]

(65)

where \(h(Y_i|M = 0)\) will obviously depend on the distribution of \(U_{1,i}\). Assuming \(U_1\) to be Gaussian \(U_1 \sim \mathcal{N}(0, \sigma_U^2 I_{N_v})\), we can write

\[
I(Y_i; U_{1,i}|M) = h(\mathcal{N}(0, \sigma_X^2 + \sigma_U^2)) - h(\mathcal{N}(0, \sigma_X^2)) = \frac{1}{2} \log \left( 1 + \frac{\sigma_U^2}{\sigma_X^2} \right).
\]

(66)

Next, the case of multiple carriers is analyzed. When \(N_c > 1\), we can write

\[
I(Y; U_1, U_2, \ldots, U_{N_c}|M) = N_v I(Y_i; U_{1,i}, U_{2,i}, \ldots, U_{N_c,i}|M)
\]

(67)

\[
= N_v \left( h(Y_i|M) - h(Y_i|U_{1,i}, \ldots, U_{N_c,i}, M) \right)
\]

(68)

\[
= N_v \left( h \left( X_i + \sum_{j=1}^{N_c} \frac{1}{\sqrt{N_c}} U_{j,i} \right) - h(X_i) \right)
\]

(69)

\[
= N_v \left( h(\mathcal{N}(0, \sigma_X^2 + \sigma_U^2)) - h(\mathcal{N}(0, \sigma_X^2)) \right)
\]

(70)

\[
= \frac{N_v}{2} \log \left( 1 + \frac{\sigma_U^2}{\sigma_X^2} \right),
\]

(71)
A.2 Mutual Information for Costa’s scheme

A.2.1 Known Message Attack (KMA)

In this appendix the mutual information between the received signal and the secret key when the sent message is known by the attacker will be computed:

\[
I(Y; \Theta|M) = h(Y|M) - h(Y|\Theta, M) = h(Y) - I(Y; M) - h(Y|U_M). \tag{72}
\]

Studying the second term,

\[
I(Y; M) = h(Y) - h(Y|M), \tag{73}
\]

it can be seen to be 0 whenever \( f_Y(y) = f_Y|_M(y|M = m) \) for all the possible values of \( m \). Taking into account that \( Y = f_1(\Theta, M, X) \) this will be true in several cases. For example, if \( U_M \) is a lattice shifted by a random variable uniform over its Voronoi region (as in [EZ04]) since the value of that random variable is not known by the attacker, the former equality is verified and \( I(Y; M) = 0 \). This will be also the case when \( U_M \) is a random codebook ([Cos83]); the attacker could know exactly all the \( u \)’s in \( \mathcal{U} \), but if he/she does not know the value of \( M \) corresponding to each of them, the best he/she can do is to apply his/her a priori knowledge about \( P(M = m) \). So \( I(Y; M) = 0 \) again. Nevertheless, in the general case \( 0 \leq I(Y; M) \leq I(Y; M|\Theta) \).

To compute \( h(Y|U_M) \) we will focus on the implementations using random codebooks. In those schemes every \( u \) in \( U_M \) has the same probability of being chosen. Taking into account we can write \( y = c\ u + u^\perp \) and all the variables are Gaussian, the samples of \( y \) will be very close to a sphere with radius \( \sqrt{N_v \text{Var}\{U^\perp\}} \) centered at some \( c u_o \), so when \( N_v \) is large enough these spheres will be disjoint if

\[
\frac{\text{Var}\{U^\perp\}}{c^2} < P. \tag{74}
\]

This is true for any Document to Watermark Ratio (DWR) if \( \alpha > 0.2 \). Assuming this holds,

\[
h(Y|U_M) = h(Y|U) + \log(|U_M|). \tag{75}
\]

Concerning \( \log(|U_M|) \) we can write

\[
|U_M| \approx \frac{e^{f(U,Z)}}{e^{N_e R}} = e^{f(U,X)} = \left( \frac{P + \alpha^2\sigma^2_X}{P} \right)^{N_v/2}. \tag{76}
\]

On the other hand,

\[
h(Y|U) = \frac{N_v}{2} \log \left[ 2\pi e \frac{(1 - \alpha)^2 P\sigma^2_X}{P + \alpha^2\sigma^2_X} \right], \tag{77}
\]

so,

\[
h(Y|U_M) = \frac{N_v}{2} \log \left[ 2\pi e \frac{(1 - \alpha)^2 P\sigma^2_X}{P + \alpha^2\sigma^2_X} \right] + \frac{N_v}{2} \log \left[ \frac{P + \alpha^2\sigma^2_X}{P} \right] = \frac{N_v}{2} \log \left[ 2\pi e (1 - \alpha)^2 \sigma^2_X \right]. \tag{78}
\]

\(^4\) \( y \) can be written as a sum of a scaled version of \( u \) and other random variable independent of \( u \) (denoted by \( u^\perp \)). The scale parameter will be \( c = \frac{P + \alpha^2\sigma^2_X}{P + \alpha^2\sigma^2_X} \).
Note that this value is just an upper bound when the spheres described above are not disjoint, i.e., when (74) is not verified. However, a value of $\alpha = 0.2$ is the optimal one for a WNR $= -6$ dB, so the condition $\alpha > 0.2$ will be verified in most practical situations.

Therefore, we can write

\[
I(Y; \Theta|M) = \frac{N_v}{2} \log \left[ 2\pi e (P + \sigma_X^2) \right] - \frac{N_v}{2} \log \left[ 2\pi e (1 - \alpha)^2 \sigma_X^2 \right] = \frac{N_v}{2} \log \left[ \frac{P + \sigma_X^2}{(1 - \alpha)^2 \sigma_X^2} \right]. \tag{79}
\]

\section*{A.2.2 Watermarked Only Attack (WOA)}

In this case, the mutual information between the observations and the secret key is

\[
I(Y; \Theta) = h(Y) - h(Y|\Theta) = h(Y) - I(Y; M|\Theta) - h(Y|\Theta, M) = h(Y) - I(Y; M|\Theta) - h(Y|\mathcal{U}_M). \tag{80}
\]

The only term that it has not been analyzed yet is $I(Y; M|\Theta)$. This is the reliable rate that can be reached when the codebook is known. Note that when $I(Y; M|\Theta)$ is increased, $I(Y; \Theta)$ is decreased; this is due to the increase in the uncertainty of the sent symbol, which complicates the attacker’s work. In [Cos83] this is shown to be

\[
I(Y; M|\Theta) = \frac{N_v}{2} \log \left[ \frac{P(P + \sigma_X^2 + \sigma_N^2)}{P \sigma_X^2 (1 - \alpha)^2 + \sigma_N^2 (P + \alpha^2 \sigma_X^2)} \right]. \tag{81}
\]

So in this case, assuming again $\alpha > 0.2$, we can write

\[
I(Y; \Theta) = \frac{N_v}{2} \log \left[ 2\pi e (P + \sigma_X^2) \right] - \frac{N_v}{2} \log \left[ \frac{P(P + \sigma_X^2 + \sigma_N^2)}{P \sigma_X^2 (1 - \alpha)^2 + \sigma_N^2 (P + \alpha^2 \sigma_X^2)} \right] - \frac{N_v}{2} \log \left[ 2\pi e (1 - \alpha)^2 \sigma_X^2 \right] = \frac{N_v}{2} \log \left[ \frac{(P + \sigma_X^2) \{ P \sigma_X^2 (1 - \alpha)^2 + \sigma_N^2 (P + \alpha^2 \sigma_X^2) \}}{P(P + \sigma_X^2 + \sigma_N^2)(1 - \alpha)^2 \sigma_X^2} \right]. \tag{82}
\]

\section*{A.2.3 Estimated Original Attack (EOA)}

In this Appendix we compute

\[
I(Y; \Theta|\hat{X}) = h(Y|\hat{X}) - h(Y|\Theta, \hat{X}), \tag{83}
\]

where $\hat{X} \triangleq X + \hat{X}$ is an estimate of $X$ and $\hat{X}$ is the estimation error; $\hat{X}$ is assumed to be i.i.d. Gaussian with power $E$ and independent of $X$. In this way, we can write

\[
h(Y|\hat{X}) = h(X + W|X + \hat{X}) < h(W - \hat{X}). \tag{84}
\]

In fact, if $\sigma_X^2 >> E$, $\hat{X}$ will be almost independent of $X + \hat{X}$ and

\[
h(Y|\hat{X}) \approx h(W - \hat{X}) = \frac{N_v}{2} \log \left[ 2\pi e (P + E) \right]. \tag{85}
\]
For the rightmost term in (83), we can write
\[ h(Y|\Theta, \hat{X}) = I(Y; M|\Theta, \hat{X}) + h(Y|\Theta, M, \hat{X}). \]  
(86)
Adapting the achievable rate from [Cos83],
\[ I(Y; M|\Theta, \hat{X}) = I(U; Z|\hat{X}) - I(U; X|\hat{X}), \]  
(87)
where \( I(\mathbf{U}; \mathbf{Z}|\mathbf{\hat{X}}) = h(\mathbf{Z}|\mathbf{\hat{X}}) - h(\mathbf{Z}, \mathbf{\hat{X}}), \) with
\[ h(\mathbf{Z}|\mathbf{X}) = h(\mathbf{X} + \mathbf{W} + \mathbf{N}|\mathbf{\hat{X}}) \approx \frac{N_v}{2} \log(2\pi e (E + \sigma^2_{N})), \]  
(88)
where it has been assumed \( \sigma^2_{X} >> E. \) On the other hand, \( \mathbf{Z} \) conditioned on \( \mathbf{U} \) and \( \mathbf{\hat{X}} \) will be a Gaussian variable, so the computation of its entropy is done by simply determining its variance. Therefore, we can write
\[ Y_{\mathbf{X}} = c_{\mathbf{X}} U_{\mathbf{X}} + U^\perp_{\mathbf{X}} \]  
(89)
where the notation implies that \( \mathbf{\hat{X}} \) is given. Since \( U_{\mathbf{X}} = W + \alpha X_{\mathbf{X}}, \) \( \text{Var}(U_{\mathbf{X}}) = \text{Var}(U|\hat{X}) = \text{Var}(W + \alpha(X + \hat{X}) - \alpha X|X + \hat{X}) \approx P + \alpha^2 E, \) where we have assumed that \( \sigma^2_{X} >> E; \) in the same way, \( U^\perp_{\mathbf{X}} = X_{\mathbf{X}}(1 - c_{\mathbf{X}} \alpha) + W(1 - c_{\mathbf{X}}), \) so \( \text{Var}(U^\perp_{\mathbf{X}}) \approx E(1 - c_{\mathbf{X}} \alpha)^2 + P(1 - c_{\mathbf{X}})^2. \) Therefore, \( c_{\mathbf{X}} \) must verify
\[ P + E = c_{\mathbf{X}}^2 (P + \alpha^2 E) + E(1 - c_{\mathbf{X}} \alpha)^2 + P(1 - c_{\mathbf{X}})^2, \]  
(90)
so
\[ c_{\mathbf{X}}^2 = \frac{P + \alpha E}{P + \alpha^2 E}. \]  
(91)
Taking this into account,
\[ \text{Var}(Z|\mathbf{\hat{X}}, \mathbf{U}) = \text{Var}(Y_{\mathbf{X}}|U_{\mathbf{X}}) + \text{Var}(N) = \text{Var}(U^\perp_{\mathbf{X}}) + \text{Var}(N) \approx \frac{PE(1 - \alpha)^2}{P + \alpha^2 E} + \sigma^2_{N}, \]  
(92)
so we can write
\[ h(Z|\mathbf{\hat{X}}, \mathbf{U}) \approx \frac{N_v}{2} \log \left( 2\pi e \left[ \frac{PE(1 - \alpha)^2}{P + \alpha^2 E} + \sigma^2_{N} \right] \right). \]  
(93)
Going back to (87), we should compute
\[ I(\mathbf{U}; \mathbf{X}|\mathbf{\hat{X}}) = h(U|\mathbf{\hat{X}}) - h(U|\mathbf{X}, \mathbf{\hat{X}}) = h(W + \alpha X|\hat{X}) - h(W + \alpha X|X, \hat{X}) \approx \frac{N_v}{2} \log \left( \frac{P + \alpha^2 E}{P} \right) \]  
(94)
Finally, the last needed term is \( h(Y|\Theta, M, \mathbf{\hat{X}}). \) Under the same assumption made in App. A.2.1 \( (\alpha > 0.2), \) we obtain it as
\[ h(Y|\Theta, M, \mathbf{\hat{X}}) = h(Y|\mathbf{\tilde{U}}_M, \mathbf{\hat{X}}) = h(Y|U, \mathbf{\hat{X}}) + \log(|\mathbf{\tilde{U}}_M|) \]  
(95)
where \(|\mathcal{U}_M\hat{X}|\) is the number of centroids associated with symbol \(M\) needed to verify the watermark power restriction when \(\hat{X}\) is given, and \(h(Y|U,\hat{X})\) coincides with \(h(Z|U,\hat{X})\) (see (93)) when \(\sigma^2_N = 0\). It can be shown that

\[
\log(|\mathcal{U}_M\hat{X}|) = I(U;X|\hat{X})
\]

which has already been derived in (94). Summarizing, \(I(Y;\Theta|\hat{X})\) will be

\[
I(Y;\Theta|\hat{X}) \approx \frac{N_v}{2} \log \left[ \frac{2\pi e(P + E)}{P-E(1-\alpha)^2 + \sigma^2_N (P + \alpha^2 E)} \right] \\
- \frac{N_v}{2} \log \left[ 2\pi e(1-\alpha)^2 E \right] \\
= \frac{N_v}{2} \log \left[ \frac{(P + E) \{ P-E(1-\alpha)^2 + \sigma^2_N (P + \alpha^2 E) \}}{P(P + E + \sigma^2_N)(1-\alpha)^2 E} \right],
\]

which is (82), but replacing \(\sigma^2_X\) by \(E\). This is explained because the uncertainty about the host signal, which makes difficult the attack, is reduced, being \(\hat{X}\) (with power \(E\)) the only unknown component (in the WOA case, it was \(X\), with power \(\sigma^2_X\)). Following this idea, WOA could be also seen as a particular case of EOA where the power of the estimation error is just \(\sigma^2_X\). Notice in any case that in several equations we have assumed \(\sigma^2_X \gg E\) to ensure the independence between \(\hat{X}\) and \(\tilde{X}\). Seeing the final result and the equivalent one for WOA, this condition does not seem to be critical, perhaps because the cancellation of this dependence between different terms.

A.3 Mutual information for scalar DC-DM

A.3.1 Known Message Attack (KMA)

In this case, the following equalities hold

\[
I(Y;D|M) = \sum_{i=1}^{N_v} I(Y_i;D_i|Y_{i-1},\ldots,Y_1,M) \\
= \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} I(Y_i;D_j|Y_{i-1},\ldots,Y_1,D_{j-1},\ldots,D_1,M) \\
= \sum_{i=1}^{N_v} I(Y_i;D_i|M_i) = N_v I(Y_i;D_i|M_i),
\]

where (98) and (99) follow from the chain rule for mutual informations [CT91], and (100) follows from the fact that the pairs \(Y_i, D_i\) and \(Y_i, M_i\) are independent \(\forall i \neq j\), and furthermore \(Y_i, D_i, M_i\) are i.i.d. processes. From the definition of mutual information we have

\[
I(Y_i;D_i|M_i) = h(Y_i|M_i) - h(Y_i|D_i,M_i) \\
= h(Y_i|M_i = 0) - h(Y_i|D_i = 0, M_i = 0),
\]
where (102) follows from the flat-host assumption introduced in Section 4.4, thus (102) is a valid result whatever the distribution of $D_i$ and $M_i$. Furthermore, due to this assumption, the entropies of (102) can be easily calculated by resorting to one period of the watermarked signal, resulting in

$$I(Y_i; D_k|M_i) = \log(\Delta) - \log((1 - \alpha)\Delta) = -\log(1 - \alpha).$$

(A.3.2) Watermarked Only Attack (WOA)

By reasoning as in the KMA case we can write

$$I(Y; D) = N_v I(Y_i; D_i) = N_v (\log(\Delta) - h(Y_i|D_i))$$

for the general case of $N_o$ observed watermarked vectors. It is always possible to obtain a theoretical expression for (104); however, for the sake of simplicity, we will calculate it here only for the case of binary signaling ($M_i = \{0, 1\}$).

It is not difficult to realize that $h(Y_i) = \log(\Delta)$ as in the KMA case (actually, this is true for $|M|$-ary signaling). For the term $h(Y_i|D_i)$ we have again that $h(Y_i|D_i) = h(Y_i|D_i = 0)$. Thus, we can write

$$I(Y_i; D_i) = \log(\Delta) - h(Y_i|D_i = 0),$$

(105)

For calculating $h(Y_i|D_i = 0)$ we must distinguish two cases:

1. If $\alpha \geq 0.5$ there is no overlapping between the pdf’s associated to adjacent centroids, so the pdf of $Y_i$ conditioned on the key is formed by two uniform pdf’s centered at each centroid, thus the conditional entropy is simply given by $h(Y_i|D_i = 0) = \log((1 - \alpha)\Delta) + \log(2)$.

2. If $\alpha < 0.5$ we must take into account that the uniform pdf’s centered at each centroid may overlap, so the pdf of $Y_i$ conditioned on $D = 0$ is given by

$$f_{Y_i}(y_i|D_i = 0) = \begin{cases} \frac{1}{(1-\alpha)\Delta}, & \text{for } -\frac{\Delta}{4} - (0.5 - \alpha)\frac{\Delta}{2} \leq y_i < -\frac{\Delta}{4} + (0.5 - \alpha)\frac{\Delta}{2} \\ \frac{1}{2(1-\alpha)\Delta}, & \text{for } \frac{\Delta}{4} - (0.5 - \alpha)\frac{\Delta}{2} \leq y_i < \frac{\Delta}{4} + (0.5 - \alpha)\frac{\Delta}{2} \\ \frac{1}{2(1-\alpha)\Delta}, & \text{for } -\frac{\Delta}{4} + (0.5 - \alpha)\frac{\Delta}{2} \leq y_i < -\frac{\Delta}{4} - (0.5 - \alpha)\frac{\Delta}{2} \\ \frac{1}{2(1-\alpha)\Delta}, & \text{for } \frac{\Delta}{4} + (0.5 - \alpha)\frac{\Delta}{2} \leq y_i < \frac{\Delta}{4} - (0.5 - \alpha)\frac{\Delta}{2} \end{cases}$$

(106)

Finally, we can write

$$h(Y_i|D_i = 0) = \begin{cases} \log(2(1 - \alpha)\Delta), & \text{for } \alpha \geq 1/2 \\ \log((1 - \alpha)\Delta)\left(\frac{1-2\alpha}{1-\alpha}\right) + \log(2(1 - \alpha)\Delta)\left(\frac{\alpha}{(1-\alpha)}\right), & \text{for } \alpha < 1/2 \end{cases}$$

(107)
B Calculation of mutual information for multiple observations

B.1 Mutual information for spread spectrum

B.1.1 Known Message Attack (KMA)

When $N_v = 1$, there are several available observations ($N_o > 1$) watermarked with the same secret key and there is one bit to be sent in each observation ($N_c = 1$) which we will assume without loss of generality to be the same for all the observations, it can be seen that the covariance matrix of $(Y^1, \ldots, Y^{N_o})$, denoted by $R_Y$, becomes

$$R_Y = \begin{pmatrix}
\sigma_X^2 + \sigma_U^2 & \sigma_X^2 & \cdots & \sigma_U^2 \\
\sigma_X^2 & \sigma_X^2 + \sigma_U^2 & \cdots & \sigma_U^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_X^2 & \sigma_X^2 & \cdots & \sigma_X^2 + \sigma_U^2
\end{pmatrix},$$

(108)

so its entropy is (see [CT91])

$$h(Y^1, \ldots, Y^{N_o}) = \frac{1}{2} \log \left((2\pi e)^{N_o} |R_Y|\right) = \frac{1}{2} \log \left((2\pi e)^{N_o} \left[\frac{N_o \sigma_U^2}{\sigma_X^2} + 1\right] \sigma_X^{2N_o}\right),$$

(109)

and we can write

$$I(Y^1, \ldots, Y^{N_o}; U_1|M_1) = \frac{1}{2} \log \left(1 + \frac{N_o \sigma_U^2}{\sigma_X^2}\right).$$

(110)

B.2 Mutual information for scalar DC-DM

B.2.1 Known Message Attack (KMA)

Assuming that the flat-host assumption introduced in Section 4.4 is valid, we will use the modulo-$\Delta$ version of the pdf of $Y$, hence $-\Delta/2 \leq Y < \Delta/2$. Without loss of generality, we consider that the transmitted symbol is the same ($M = 0$) in the $N_o$ observations (the general case does not change the result). In the following, $Y^{N_o}_k$ will denote a vector of $N_o$ observations of the $k$-th component of $Y$, and $Y_{k,i}$ will be the $i$-th observation of that component. For the sake of simplicity, $D_k = \Theta_k$, so:

$$I(Y^{N_o}_k; \Theta_k|M^{N_o}_k) = \sum_{i=1}^{N_o} \left(h(Y_{k,i}|M^{N_o}_k, Y_{k,i-1}, \ldots, Y_{k,1}) - h(Y_{k,i}|\Theta_k, M^{N_o}_k, Y_{k,i-1}, \ldots, Y_{k,1})\right)$$

$$= \sum_{i=1}^{N_o} \left(h(Y_{k,i}|M^{N_o}_k, Y_{k,i-1}, \ldots, Y_{k,1}) - \log((1 - \alpha)\Delta)\right)$$

(111)

$$= \sum_{i=1}^{N_o} I(Y_{k,i}; \Theta_k|M^{N_o}_k, Y_{k,i-1}, \ldots, Y_{k,1}).$$

(112)
The problem here is the calculation of the conditional entropy $h(Y_k|I_k^{N_o}, Y_{k,i-1}, \ldots, Y_{k,1})$, since we must know the pdf of the $i$-th observation conditioned on the previous ones. In order to illustrate the dependence between observations, consider the simple case of 2 observations:

$$f(y_1, y_2) = \int_\Theta f(y_1, y_2|\theta)g(\theta)d\theta = \int_\Theta f(y_1|\theta)f(y_2|\theta)g(\theta)d\theta,$$

with $g(\theta) = U(-\Delta/2, \Delta/2)$ and $f(y_1|\theta) = U(y_1 - (1 - \alpha)\Delta, y_1 + (1 - \alpha)\Delta)$. This means that, with only one observation, the possible values for $\Theta$ have been reduced from $[0, \Delta)$ to $[y_1 - (1 - \alpha)\Delta/2, y_1 + (1 - \alpha)\Delta/2)$, thus $I(Y_{k,2}; \Theta_k|M_k^2, Y_{k,1}) \leq I(Y_{k,1}; \Theta_k|M_k^1)$, with equality if and only if $\alpha = 0$. In any case, Equation (114) can be rewritten as

$$f(y_1, y_2) = \frac{1}{(1 - \alpha)^2\Delta^2} \text{rect} \left( \frac{y_2}{(1 - \alpha)\Delta} \right) \ast \text{rect} \left( \frac{y_2 - y_1}{(1 - \alpha)\Delta} \right),$$

being $\text{rect}(x) = 1$ for $|x| \leq 0.5$ and $\text{rect}(x) = 0$ for $|x| > 0.5$. Finally, the conditional pdf can be obtained as $f(y_2|y_1) = f(y_1, y_2)/f(y_1)$. In the general case of $N_o$ observations:

$$f(y_1, y_2, \ldots, y_{N_o}) = \int_{\Theta_{int}} \prod_{i=1}^{N_o} f(y_i|\theta)g(\theta)d\theta = \frac{1}{(1 - \alpha)^{N_o - 1}\Delta^{N_o}} \int_{\Theta_{int}} f(y_{N_o}|\theta)d\theta,$$

with

$$\Theta_{int} = \theta \in (-\Delta/2, \Delta/2] \text{ such that } f(y_i|\theta) \neq 0 \forall i < N_o - 1,$$

(the region of integration may be composed of disjoint intervals, in general), and the conditional entropy of (112) can be calculated as

$$h(Y_i|M^{N_o}, Y_1, \ldots, Y_{i-1}) = \int h(Y_i|M^{N_o}, Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})dy_1 \ldots dy_{i-1}.$$

The integration limits in (117) can be specialized for $\alpha \geq 1/2$, resulting in

$$\Theta_{int} = \left\{ \begin{array}{ll}
\max_i \{y_i - \frac{(1 - \alpha)\Delta}{2}\}, \min_i \{y_i + \frac{(1 - \alpha)\Delta}{2}\} & \text{if } |y_i - y_j| < (1 - \alpha)\Delta \forall i, j < N_o \\
0 & \text{otherwise}
\end{array} \right.$$

(119)

For $\alpha > 1/2$ there is no overlapping between adjacent centroids, thus the following expression for the calculation of the pdf of $Y_i$ conditioned on the previous observations is valid, but for

---

5We will obviate the subindex $k$ in the following discussion, for simplicity of notation.
a location parameter:

\[
f_{Y_i}(y_i|M, y_{i-1}, \ldots, y_1) = \begin{cases} 
K \left( \frac{(1-\alpha)\Delta}{2} + a + y \right), & \text{for } -\frac{(1-\alpha)\Delta}{2} - a < y < -\frac{(1-\alpha)\Delta}{2} + a \\
K2a, & \text{for } -\frac{(1-\alpha)\Delta}{2} + a < y < \frac{(1-\alpha)\Delta}{2} - a \\
K \left( \frac{(1-\alpha)\Delta}{2} + a - y \right), & \text{for } \frac{(1-\alpha)\Delta}{2} - a < y < \frac{(1-\alpha)\Delta}{2} + a 
\end{cases}
\]

with \( K = ((1-\alpha)^2a)^{-1} \), and 2a is the volume of \( \Theta_{\text{int}} \). Since a location parameter does not change entropy, we can write

\[
h(Y_i|M^{N_o}, y_{i-1}, \ldots, y_1) = - \int_{\frac{-\Delta}{2} + a}^{\frac{\Delta}{2} - a} K \left( \frac{(1-\alpha)\Delta}{2} + a + y \right) \log \left( K \left( \frac{(1-\alpha)\Delta}{2} + a + y \right) \right) dy
\]

\[
- \int_{\frac{-\Delta}{2} - a}^{\frac{\Delta}{2} + a} K2a \log(K2a) dy
\]

\[
- \int_{\frac{-\Delta}{2} - a}^{\frac{\Delta}{2} + a} K \left( \frac{(1-\alpha)\Delta}{2} + a - y \right) \log \left( K \left( \frac{(1-\alpha)\Delta}{2} + a - y \right) \right) dy
\]

\[
= \frac{a}{(1-\alpha)\Delta} + \log((1-\alpha)\Delta) \text{ nats}.
\]

Substituting (121) into (118), we obtain

\[
h(Y_i|M^{N_o}, Y_{i-1}, \ldots, Y_1) = \log((1-\alpha)\Delta) + \frac{1}{(1-\alpha)\Delta}E_{f(y_{i-1}, \ldots, y_1)}[a],
\]

with

\[
a = \frac{1}{2}(1-\alpha)\Delta + \min\{y_1, \ldots, y_{i-1}\} - \max\{y_1, \ldots, y_{i-1}\},
\]

accordingly to (119). Hence, the conditional entropy depends only on the mean value of the integration volume.

Analytical evaluation of (122) with the exact distribution of the received samples \( y_i \) is extremely difficult, and numerical computation is very demanding, in such a way that numerical results with reasonable precision can only be obtained for a few observations, for reasons of computational burden. However, we have found an alternative and simpler formulation of the problem that allows for the obtention of theoretical results; this alternative formulation is based on the assumption that the received samples \( Y_i \) are all independent but uniformly distributed around an unknown centroid \( \theta \): \( Y_i \sim U(\theta - (1-\alpha)\Delta/2, \theta + (1-\alpha)\Delta/2) \). Under this assumption, let us define the random variable

\[
X \triangleq \min(Y_1, Y_2, \ldots, Y_{N_o}) - \max(Y_1, Y_2, \ldots, Y_{N_o}).
\]

The pdf of \( X \) for \( N \) observations can readily be shown to be

\[
f_X(x) = N(N - 1) \frac{(-x)^{N-2}}{((1-\alpha)\Delta)^N} [(1-\alpha)\Delta + x],
\]

with \( x \in (-1-\alpha)\Delta, 0] \). Hence, the mean value of \( a \) results in

\[
E_{f_X}[a] = \frac{(1-\alpha)\Delta}{2} \left( 1 - \frac{N - 1}{N + 1} \right),
\]
and substituting it in (122), after some algebra, we finally obtain the following expression for the conditional entropy

\[ h(Y_i|M^N, Y_{i-1}, \ldots, Y_1) \approx \log((1 - \alpha)\Delta) + \frac{1}{i} \text{ nats}, \text{ for } i > 1. \]  

(126)

When \( N \to \infty \), the entropy (126) clearly tends to \( \log((1 - \alpha)\Delta) \), and consequently, the growth of the mutual information with the number of observations tends to 0. Substituting (126) in Eq. (112), we have the final expression for the mutual information

\[ I(Y^N; \Theta|M^N) = -\log(1 - \alpha) + \sum_{i=2}^{N} \frac{1}{i} \text{ nats}. \]  

(127)

We have compared this theoretical result to that obtained numerically with the exact distribution of the received samples. Numerical results are subject to errors due to the discretization of the problem; however, for a small number of observations (we have tried up to 5), the number of points can be increased, reducing the difference between the numerical results and the theoretical ones to less than \( 10^{-8} \).

**B.2.2 Watermarked Only Attack (WOA)**

For the WOA case, we must take into account that the observations may be associated to any of the possible cosets. For instance, with 2 observations and binary signaling \( (M = \{0,1\}) \) we have:

\[
 f(y_i|\theta) \sim U \left( \theta - \frac{(1 - \alpha)\Delta}{2}, \theta + \frac{(1 - \alpha)\Delta}{2} \right) Pr\{M = 0\} \\
 + U \left( \theta - \frac{(1 - \alpha)\Delta}{2} + \frac{\Delta}{2}, \theta + \frac{(1 - \alpha)\Delta}{2} + \frac{\Delta}{2} \right) Pr\{M = 1\}. 
\]

For \( \alpha > 0.75 \) there is no overlapping between the adjacent cosets, so the conditional pdf's can be expressed as the sum of two pdf's like (120), separated \( \Delta/2 \). Under these assumptions, it is easy to realize that

\[ h(Y_i|M_i, Y_{i-1}, \ldots, Y_1) = \log((1 - \alpha)\Delta) + \frac{1}{i} + \log(2) \text{ nats}, \]  

(128)

and the terms of the mutual information will be of the form

\[ I(Y_i : \Theta|Y_{i-1}, \ldots, Y_1) = \log((1 - \alpha)\Delta) + \frac{1}{i} + \log(2) - \log(2(1 - \alpha)\Delta) = \frac{1}{i} \text{ nats}, \]  

(129)

hence the mutual information in the WOA case for \( \alpha > 0.75 \) is given by

\[ I(Y^N; \Theta) = -\log(1 - \alpha) - \log(2) + \sum_{i=2}^{N} \frac{1}{i} = I(Y^N; \Theta|M^N) - \log(2) \text{ nats}. \]  

(130)
C  Relationship between Fisher Information Matrix and Entropy

The \((i, i)\)-th component of the Fisher Information Matrix can be written as

\[
\text{FIM}_{ii}(\theta) = \int f(y; \theta) \left( \frac{\partial}{\partial \theta_i} \log f(y; \theta) \right)^2 dy. \tag{131}
\]

If \(Y = X + \Theta\) (this is the case for SS-KMA, SCS-KMA and SCS-WOA), with the samples in \(X\) i.i.d,

\[
\text{FIM}_{ii}(\theta) = \int f_X(y - \Theta) \left( \frac{\partial}{\partial \theta_i} \log f_X(y - \Theta) \right)^2 dy \tag{132}
\]

\[
= \int f_X(y_i - \theta_i) \left( \frac{\partial}{\partial y_i} \log f_X(y_i - \theta_i) \right)^2 dy_i \tag{133}
\]

\[
= \int f_X(y_i - \theta_i) \left( \frac{\partial}{\partial x_i} \log f_X(x_i) \right)^2 dx_i = J(X_i) = \lambda, \tag{134}
\]

where \(\lambda\) is the value of the diagonal elements of the FIM. Following the same procedure, for the off-diagonal elements, we have

\[
\text{FIM}_{ij}(\theta) = \int f_X(x_i) \frac{\partial}{\partial x_i} \log f_X(x_i) dx_i \int f_X(x_j) \frac{\partial}{\partial x_j} \log f_X(x_j) dx_j, \tag{135}
\]

so if \(X \sim \mathcal{N}(0, \sigma_X^2 I_{N_e})\) this implies \(\text{FIM}_{ij}(\theta) = 0\), for all \(i \neq j\), so \(\text{FIM}(\theta)\) will be diagonal with identical diagonal values. In any case, the trace of the Fisher Information Matrix is

\[
J(X) = \sum_{j=1}^{N_e} J(X_i) = \mathbb{E} \left[ \frac{||\nabla f_X||^2}{f_X^2} \right] = N_e \lambda, \tag{136}
\]

as it was shown in [CC84]. Another result from [CC84] is

\[
J(X) = \mathbb{E} \left[ \frac{||\nabla f_X||^2}{f_X^2} \right] \geq N_e 2\pi e e^{-2h(X)/N_e}, \tag{137}
\]

so \(\lambda \geq 2\pi e e^{-2h(X)/N_e}\) reaching the equality whenever \(X\) is Gaussian. Take into account that \(\theta\) is a deterministic translation parameter of \(f_Y(y)\) in [CC84], so it does not modify \(h(Y)\).

D  Optimal selection of the distribution for the dither in DC-DM

First, we show that for scalar DC-DM

\[
I(Y; D) = I(Y; D \mod \Delta). \tag{140}
\]
Notice that the randomized embedding function (46) can be written as

\[ y_k = x_k + \alpha \left( Q_{\Lambda_{i,d}}(x_k) - x_k \right), \]

with

\[ \Lambda_{i,d} = \Delta Z + m_i \frac{\Delta}{|M|} - d = \Delta Z + m_i \frac{\Delta}{|M|} - d \mod \Delta, \]

(141)

where the second equality in (141) follows from the periodicity of the scalar lattice \( \Delta Z \). We will assume without loss of generality that \( Y \) and \( D \) are scalars. Let \( f_D(d) \) and \( f_Y(y|D = d) \) denote the pdf of the secret key and the pdf of the watermarked signal conditioned on the dither, respectively. Taking into account (141), we have that

\[ f_Y(y|D = d) = f_Y(y|D = d + i\Delta) \ \forall \ i, \]

thus, after the change of variable \( t = d \) it is possible to write

\[
\begin{align*}
  f_Y(y) &= \int f_Y(y|T = t)f_T(t)dt = \int_{t=0}^{t=\Delta} f_Y(y|T = t) \sum_{i=-\infty}^{\infty} f_T(t + i\Delta)dt \\
  &= \int_{t=0}^{t=\Delta} f_Y(y|T = t)f_{T \mod \Delta}(t)dt,
\end{align*}
\]

(142)

where \( f_{T \mod \Delta}(t) \) is the pdf of the modulo-\( \Delta \) version of \( T \), hence equality (140) immediately follows. Note that this result is valid whatever the distribution of the host and the DWR.

Now, we consider what is the best choice for the dither from a security point of view. For simplicity of notation we define the random variable \( Z \triangleq D \mod \Delta \). It is a known fact that the uniform distribution maximizes the entropy in an interval \([CT91]\), but the watermarker is interested in maximizing the equivocation

\[ h(Z|Y) = h(Z) - h(Y) + h(Y|Z). \]

(143)

In the following discussion we will make use of the flat-host assumption introduced in Section 4.4, thus we will consider that \(-\Delta/2 \leq Y < \Delta/2\). We have that \( h(Y|Z) = h(Y|Z = z) \ \forall \ z \), thus the rightmost term of (143) does not depend on the distribution of \( Z \). Then, we must find \( f_Z(z) \) such that \( \{h(Z) - h(Y)\} \) is maximum.

\[
\max_{f_Z(z)} \{h(Z) - h(Y)\}.
\]

Let us define a random variable \( V \) such that \( f_V(v) \triangleq f_Y(y|Z = 0) \). Under the flat-host assumption we have that \( f_Y(y) = f_V(v) \circ f_Z(z) \), where \( \circ \) denotes cyclic convolution over \([-\Delta/2, \Delta/2]\). Hence, the maximization problem can be rewritten as

\[
\max_{f_Z(z)} \{h(Z) - h(V \oplus Z)\},
\]

where \( \oplus \) denotes modulo-\( \Delta \) sum. We have the following lemma:

**Lemma:** \( h(Z) \leq h(V \oplus Z) \), with equality if \( Z \sim U(-\Delta/2, \Delta/2) \).

**Proof:** Consider that \( f_{\tilde{V}}(\tilde{v}) \) is the periodic extension of \( f_V(v) \) over \( n \) bins, properly scaled to ensure that \( f_{\tilde{V}}(\tilde{v}) \) is still a valid pdf, i.e.

\[
f_{\tilde{V}}(\tilde{v}) = \frac{1}{n} \sum_{i=-n/2}^{n/2-1} f_V(v + i\Delta),
\]
and that the same applies for $f_{\tilde{Z}}(\tilde{z})$. Now, define $\tilde{Q} \triangleq \tilde{V} + \tilde{Z}$. $\tilde{Q}$ will be also periodic with period $\Delta$ in $n - 2$ bins. Notice that

$$h(\tilde{Z}) = h(Z) + \log(n).$$  \hfill (144)

Assuming that for sufficiently large $n$ we can neglect the border effects, if we denote by $Q$ the modulo-$\Delta$ version of $\tilde{Q}$, we have that

$$h(\tilde{Q}) = h(Q) + \log(n).$$  \hfill (145)

We know that $h(\tilde{Z}) \leq h(\tilde{Q})$ [CT91], hence by (144) and (145) we have $h(Z) \leq h(V \oplus Z)$. To achieve equality it is sufficient to choose $Z$ such that $Z \sim U(-\Delta/2, \Delta/2)$. □

The proof of the lemma shows that the uniform over the quantization bin maximizes the equivocation.
References


