Digital watermarking of still images in the presence of de-synchronization attacks: theoretical analysis and development of practical algorithms

WP1: Theoretical analysis

M. Barni
University of Siena

N. Merhav
Israel Institute of Technology

Sponsored by

MiUR

Under contract no RBIN04AC9W

WP1.2 Second report on theoretical analysis

10 September 2008
Preface

This report is a living document that is updated at the end of each reporting period to describe the activity carried out within the theoretical work-package of the FIRB project no RBIN04AC9W. The final version, that will be ready at the end of the third year, will thus contain a comprehensive and self-contained description of the main theoretical results obtained during the project. A similar report will account for the practical part of the project (WP2).

The present version of the report (WP1.2), details the theoretical results obtained during the first 2 years of the project. As such it updates the report WP1.1 that accounted for the activity of the first year.

The structure of the report closely follows the structure of WP1 (devoted to theoretical analysis) of the project. Each chapter of the report corresponds to a task of WP1. Chapters are written so to be as self-contained as possible, however for a better understanding of the project development it is suggested that they are read sequentially. For each chapter (section) the project year during which the activity described within the chapter (section) has been obtained is reported, so to allow an easy understanding of the temporal development of the project.

More specifically, chapter 1 (T1.1) is devoted to the definition of a general theoretical framework upon which the subsequent research of the project will be based. The exact definition of the theoretical framework allowed an exact formulation of the main problems the project will tackle with. A list with such problems is included at the end of chapter T1.1.

The second chapter (T1.2) considers the development of suitable models to define the class of perceptually admissible DAs. Specifically the two-classed of LPCD and MRF-DA attacks are introduced and analyzed from a theoretical point of view. For a practical evaluation of their ability to describe the way the human visual system perceives geometric distortions, and their de-synchronization properties when applied to natural watermarked images, see the WP2.2 report. Chapter T1.3 contains the main theoretical achievements obtained in the first two years of the project, namely the extension of the general framework described in T1.1 to account for several classes of attacks, namely: i) a general discrete attack channel unknown to the detector; ii) the AWGN channel with unknown (universal watermarking) or known (non-universal watermarking) parameters; iii) memoryless, non-Gaussian channels; iv) pixel permutation channel; and v) LPCD channel. Chapter T1.4 considers the computational complexity of the optimum algorithms derived in section T1.3. In particular it focuses on the complexity of the optimum embedding algorithm for the discrete case, since in the continuous (Gaussian) case complexity is not an issue, since a simple close formula for the embedding rule exists. Chapters T1.5 addresses the problem of optimum watermark embedding/detection under security requirement. This activity has started in the last part of the second year, hence for the moment only a precise formulation of the optimization problem under security constraints is available.

The publications the various research lines resulted into are listed at the end of each chapter.
T1.1 - Definition of problems

Project year: Y1

Abstract. The general formulation of the optimum embedding/detection problem is given: such a formulation will represent the background for the future developments of the project. Both the discrete and the continuous cases are considered. For the former case the optimum embedder and detector are derived first by assuming that the pdf of the host signal is known, then in the general case of unknown distribution. The continuous case is treated by referring to a Gaussian host.

1 Introduction

The goal of T1.1 (the first task of WP1) was the definition of a solid theoretical framework upon which the future developments of the project could be built. To do so, we followed the approach outlined in [1], where the problem of optimum watermark embedding and detection is formulated as a hypothesis testing problem. Under certain simplifying assumptions, including the absence of attacks, the memoryless nature of the source and, in the continuous case, the normality of the host signal, the solution of the above problem can be found by relying on the method of types (and its extension to the continuous case) as detailed in [1].

In the next section the theoretical framework introduced in [1] and the main results demonstrated therein are reviewed, then (section 3) such a framework will be used to give a precise formulation of the problems that will be tackled during the rest of the project.

2 Definition of a theoretical framework

Let \( x = (x_1, \ldots, x_n) \) be a block from the covertex source and \( u = (u_1, \ldots, u_n) \) the independent binary \((\pm 1)\) watermark vector. Watermark embedding can be seen as the definition of an embedding rule \( f \) that produces a watermarked sequence \( y = (y_1, \ldots, y_n) \) by starting from \( x \) and \( u \). On the other side, watermark detection is a classical hypothesis testing problem, where the detector has to decide whether a given sequence \( y \) contains a given watermark \( u \) or not.

The problem of finding the optimum watermark embedding function \( f \) (and the corresponding optimum watermark detection rule) is not trivial: the probabilities of errors of the two kinds (false positive and false negative) corresponding to the detector induced by a given \( f \), are, in general, hard to compute, and a fortiori hard to optimize in closed form. Thus, instead of striving to seek the strictly optimum embedder, we take the following approach: suppose that one would like to limit the complexity of the detector by confining its decision to depend on a given set of statistics computed from \( y \) and \( u \). For example, the energy of \( y \), \( \sum_{n=1}^{n} y_i^2 \) and the correlation \( \sum_{n=1}^{n} u_i y_i \), which are the sufficient statistics used by the popular correlation detector. Further, if there is reason to suspect that the stegotext might be subjected to a certain cyclicshift attack (see, e.g. [2]), one might wish to include also correlations between \( y \) and the corresponding possible shifted versions of \( u \). Other possible statistics are those corresponding to the likelihood ratio detector for multiplicative watermarks, namely, the energies \( \sum_{i:u_i=+1} y_i^2 \) and \( \sum_{i:u_i=-1} y_i^2 \), and so on.

Within the class detectors based on a given set of statistics, we want to find the best embedder and detector which are asymptotically optimum in the NeymanPearson sense of trading off the exponents of the error probabilities of the two kinds.
2.1 Basic derivation

For the sake of simplicity, let us start with the discrete case, i.e. let us assume that the components of \( x \) and \( y \) take on values in a finite alphabet \( \mathcal{A} \). The components of the watermark \( u \) will always take on values in \( B = \{+1, -1\} \). Let us further assume that \( x \) is drawn from a given memoryless source \( P \).

For a given \( u \), we would like to devise a decision rule that partitions the space \( \mathcal{A}^n \) of sequences \( \{y\} \) observed by the detector into two complementary regions \( \mathcal{A} \) and \( \mathcal{A}^c \), such that for \( y \in \mathcal{A} \) we decide in favor of \( H_1 \) (watermark \( u \) is present) and for \( y \in \mathcal{A}^c \), we decide in favor of \( H_0 \) (watermark absent: \( y = x \)). Consider the Neyman-Pearson criterion of minimizing the false negative probability

\[
P_{e_1} = \sum_{x : f(x, u) \in \mathcal{A}^c} P(x)
\]

subject to the following constraints:

1. Given a certain distortion measure \( d(\cdot, \cdot) \) and distortion level \( D \), the distortion between \( x \) and \( y \), \( d(x, y) = d(x, f(x, u)) \), does not exceed \( n D \).
2. The false positive probability is upper bounded by

\[
P_{e_2} \triangleq \sum_{y \in \mathcal{A}} P(y) \leq e^{-\lambda n}
\]

where \( \lambda > 0 \) is a prescribed constant.

In other words, we would like to choose \( f \) and \( \Lambda \) so as to minimize \( P_{e_1} \) subject to the constraint that the error exponent of \( P_{e_2} \) would be at least as large as \( \lambda \). As already said we make a simplifying assumption regarding the statistics employed by the detector. Suppose, for example, that we are interested in the class of all detectors which base their decision on the empirical joint distribution of \( y \) and \( u \):

\[
\hat{P}_{xy} = \{\hat{P}_{uy}(u, y), u \in B, y \in \mathcal{A}\}
\]

where

\[
\hat{P}_{uy} = \frac{1}{n} \sum_{i=1}^{n} 1\{u_i = u, y_i = y\}, \quad u \in B, y \in \mathcal{A}
\]

\( 1\{u_i = u, y_i = y\} \) being the indicator function of the event \( \{u_i = u, y_i = y\} \), that is, \( \hat{P}_{uy} \) is the relative frequency of the pair \((u, y)\) along the pair sequence \((u, y)\). Following standard terminology in the information theory literature [3], we define the conditional type class of \( y \) given \( u \), and denote it by \( T(y|u) \), as the set of all sequences \( y' \in \mathcal{A}^n \) such that \( \hat{P}_{uy} = P_{uy} \), that is, the set of all \( y' \) which have the same empirical joint distribution with \( u \) that \( y \) has. The requirement that the decision of the detector depends solely on \( \hat{P}_{uy} \) means that \( \Lambda \) and \( \mathcal{A}^c \) are unions of conditional types classes of \( y \) given \( u \). Now, let \( T(y|u) \subseteq \Lambda \). Then, we have

\[
e^{-\lambda n} \geq \sum_{y' \in \Lambda} P(y')
\]

\[
\geq \sum_{y' \in T(y|u)} P(y')
\]

\[
\geq |T(y|u)| \cdot P(y)
\]

\[
\geq (n + 1)^{-|A|} e^{n R_{uy}(Y|U)} \cdot P(y)
\]

where the first inequality is by the assumed false positive constraint, the second inequality is since \( T(y|u) \subseteq \Lambda \), and the third inequality is due to the fact that all sequences within \( T(y|u) \) are equiprobable under \( P \) as they all have the same empirical distribution, which form the sufficient statistics for
the memoryless source \( P \). In the fourth inequality, we use the lower bound on the cardinality of a conditional type class in terms of the empirical conditional entropy [3], defined as:

\[
\hat{H}_{uy}(Y|U) = - \sum_{u,y} \hat{P}_{uy}(u,y) \ln (\hat{P}_{uy}(y|u))
\]  

(6)

where \( \hat{P}_{uy}(y|u) \) is the empirical conditional probability of \( Y \) given \( U \). Defining now

\[
A_* = \{ y : \ln (P(y)) + n \hat{H}_{uy}(Y|U) + \lambda n - |A| \ln(n+1) \leq 0 \}
\]

(7)

we have actually shown that every \( T(y|u) \) in \( A \) is also in \( A_* \), in other words, if \( A \) satisfies the false positive constraint (2), it must be a subset of \( A_* \). This means that \( A_* \subset A^{c} \) and so the probability of \( A^{c} \) is smaller than the probability of \( A^{*} \), i.e., \( A^{*} \) minimizes \( P_{e1} \) among all \( A^{c} \) corresponding to detectors that satisfy (2). To establish the asymptotic optimality of \( A_* \), it remains to show that \( A_* \) itself has a false positive exponent at least \( \lambda \), which is very easy to show using the techniques of eq. (6) in [4] and references therein. Therefore, we will not include the proof of this fact here. Finally, note also that \( A_* \) bases its decision solely on \( \hat{P}_{uy} \), as required.

While this solves the problem of the optimal detector for a given \( f \), we still have to specify the optimal embedder \( f^* \). Defining \( \Gamma_{u}^{*}(f) \) to be the inverse image of \( A^{c} \) given \( u \), i.e.,

\[
\Gamma_{u}^{*}(f) = \{ x : f(x, u) \in A_u^{c} \}
\]

= \{ x : \ln (P(f(x, u))) + n \hat{H}_{uy, f(x, u)}(Y|U) + \ldots + \lambda n - |A| \ln(n+1) > 0 \}

(8)

then following (1), \( P_{e1} \) can be expressed as

\[
P_{e1} = \sum_{x \in \Gamma_{u}^{*}(f)} P(x)
\]

(9)

Consider now the following embedder:

\[
f^*(x, u) = \arg \min_{y \cdot d(x, y) \leq n \Delta} \left[ \ln (P(y)) + n \hat{H}_{uy}(Y|U) \right]
\]

(10)

where ties are resolved in a arbitrary fashion. It is clear by definition, that \( \Gamma_{u}^{*}(f^*) \subseteq \Gamma_{u}^{*}(f) \) for any other competing \( f \) that satisfies the distortion constraint, and thus \( f^* \) minimizes \( P_{e1} \) subject to the constraints.

### 2.2 Implementability of the embedder

The first impression might be that the minimization in (10) is prohibitively complex as it appears to require an exhaustive search over the sphere \( y : d(x, y) \leq n \Delta \), whose complexity is exponential in \( n \). A closer look, however, reveals that the situation is not that bad. Note that for a memoryless source \( P \),

\[
\ln (P(y)) = -n \left[ \hat{H}_y(Y) + D(P_y||P) \right]
\]

(11)

where \( \hat{H}_y(Y) \) is the empirical entropy of \( y \) and \( D(P_y||P) \) is the divergence between the empirical distribution of \( y \), \( \hat{P}_y \), and the source \( P \). Moreover, if \( d(\cdot, \cdot) \) is an additive distortion measure, i.e.,

\[
d(x, y) = \sum_{i=1}^{n} d(x_i, y_i),
\]

then \( d(x, y)/n \) can be represented as the expected distortion with respect to the empirical distribution of \( x \) and \( y \), \( \hat{P}_{xy} \). Thus, the minimization in (10) becomes equivalent to maximizing \( \hat{H}_{uy}(U; Y) + D(P_y||P) \) subject to \( \hat{E}_{xy} d(X, Y) \leq D \), where \( \hat{H}_{uy}(U; Y) \) denotes the empirical mutual information induced from the joint empirical distribution \( \hat{P}_{uy} \) and \( \hat{E}_{xy} \) denotes the aforementioned expectation with respect to \( \hat{P}_{xy} \). Now, observe that for given \( x \) and \( u \), both
\[ \hat{I}_{uy}(U; Y) + D(\hat{P}_y || P) \] and \( \hat{E}_{xy} d(X, Y) \leq D \) depend on \( y \) only via its conditional type class given \((x, y)\), namely, the conditional empirical distribution \( \hat{P}_{uxy}(y|x, u) \). Once the optimal \( \hat{P}_{uxy}(y|x, u) \) has been found, it does not matter which vector \( y \) is chosen from the corresponding conditional type class \( T(y|x, u) \).

Therefore, the optimization across \( n-v \) vectors in (10) boils down to optimization over empirical conditional distributions, and since the total number of empirical conditional distributions of \( n \) vectors increases only polynomially with \( n \), the search complexity reduces from exponential to polynomial as well. At any rate, this optimization procedure is carried out in a space of fixed dimension, that does not grow with \( n \).

### 2.3 Universality in the coverted text distribution

Thus far we have assumed that the distribution \( P \) is known. In practice, even if it is fine to assume a certain model class, like the model of a memoryless source, the assumption that the exact parameters of \( P \) are known is rather questionable. Suppose then that \( P \) is known to be memoryless but is otherwise unknown. How should we modify our results? First observe, that it would make sense to insist on the constraint (2) for every memoryless source, to be on the safe side. Equivalently, (2) would be replaced by

\[
\max_P P(y) \sum_{y \in \Lambda} P(y) \leq e^{-\lambda n} \tag{12}
\]

where the maximization over \( P \) is across all memoryless source with alphabet \( A \). It is then easy to see that our earlier derivation goes through as before except that \( P(y) \) should be replaced by \( \max_P P(y) \) in all places (see also [4]). Since \( \ln(\max_P P(y)) = -n H_y(Y) \), this means that the modified version of \( \Lambda \) compares the empirical mutual information \( \hat{I}_{uy}(U; Y) \) to the threshold \( \lambda n - |A| \ln(n + 1) \).

By the same token, and in light of the discussion in the previous paragraph, the modified version of the optimal embedder (10) maximizes \( \hat{I}_{xy}(U; Y) \) subject to the distortion constraint (the divergence term now disappears). Both the embedding rule and the detection rule are then based on the idea of maximum mutual information, which is intuitively appealing.

### 2.4 Continuous alphabets

In the previous section, we considered the simple case where the components of both \( x \) and \( y \) take on values in a finite alphabet. It is more common and more natural, however, to model \( x \) and \( y \) as vectors in \( \mathbb{R}^n \). Beyond the fact that, summations should be replaced by integrals, in the analysis of the previous section, this requires, in general, an extension of the method of types [3], used above, to vectors with real valued components (see, e.g. [5],[6],[7]). In a nutshell, a conditional type class, in such a case, is the set of all \( y \)-vectors in \( \mathbb{R}^n \) whose joint statistics with \( u \) have (within infinitesimally small tolerance) prescribed values, and to have a parallel analysis to that of the previous section, we have to be able to assess the exponential order of the volume of the conditional type class.

Suppose, for example, that \( x \) is a zero mean Gaussian vector whose covariance matrix is \( \sigma^2 I \), \( I \) being the \( n \times n \) identity matrix and \( \sigma^2 \) is unknown (cf. Subsection 2.3). Let us suppose also that the statistics to be employed by the detector are the energy of \( \sum_{i=1}^n y_i^2 \) and the correlation \( \sum_{i=1}^n u_i y_i \). These assumptions are the same as in many theoretical papers in the literature of watermark detection. Then, the conditional empirical entropy \( \hat{H}_{uy}(Y|U) \) should be replaced by the empirical
differential entropy $\hat{h}_{uy}(Y|U)$, given by [6]:

$$\hat{h}_{uy}(Y|U) = \frac{1}{2} \ln \left[ 2 \pi e \min_a \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - a u_i)^2 \right) \right]$$

$$= \frac{1}{2} \ln \left[ 2 \pi e \left( \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} u_i y_i \right)^2 \right) \right]$$

$$= \frac{1}{2} \ln \left[ 2 \pi e \left( \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} u_i \right)^2 \right) \right] \quad (13)$$

Since

$$\hat{h}_y(Y) = \frac{1}{2} \ln \left( 2 \pi e \frac{1}{n} \sum_{i=1}^{n} y_i^2 \right) \quad (14)$$

the optimal embedder would maximize

$$\hat{I}_{uy}(W; Y) = -\frac{1}{2} \ln \left( 1 - \frac{\left( \frac{1}{n} \sum_{i=1}^{n} u_i y_i \right)^2}{\frac{1}{n} \sum_{i=1}^{n} u_i^2} \right) \quad (15)$$

or, equivalently, maximize $\left( \sum_{i=1}^{n} u_i y_i \right)^2 / \sum_{i=1}^{n} y_i^2$ subject to the distortion constraint, which in this case, will naturally be taken to be the Euclidean one, $\sum_{i=1}^{n} (x_i - y_i)^2 \leq n D$. While our previous discussion regarding optimization over conditional distributions, does not apply directly in the continuous case considered here, it can still be represented as optimization over a finite dimensional space whose dimension is fixed, independently of $n$. In fact, this fixed dimension is 2: every $y \in \mathbb{R}^n$ can be represented as $y = a x + b u + z$, where $a$ and $b$ are real valued coefficients and $z$ is orthogonal to both $x$ and $u$. Now, without loss of optimality, $z$ should be taken to be the zero vector. This is because any non-zero $z$ contributes to the energy of $y$ (the denominator of $\left( \sum_{i=1}^{n} u_i y_i \right)^2 / \sum_{i=1}^{n} y_i^2$) while improving neither the correlation with $u$ (which is the numerator), nor the distance to $x$ (which is the constraint). Thus, the optimal embedding function should be of the form

$$f^*(x, u) = a x + b u \quad (16)$$

and so, it remains only to optimize over two parameters, $a$ and $b$. Upon manipulating this optimization problem, by taking advantage of its special structure, one can further reduce its dimensionality and transform it into a search over one parameter only (for more details see [8]).

## 3 Problem formulation

By referring to the theory illustrated in the previous section, the theoretical problems that are going to be addressed by the project can be formulated as follows.

### 3.1 Optimum embedding/detection in the presence of attacks

Whereas the analysis described in section 2 permits to solve the problem of optimum watermark embedding/detection in the absence of attacks, for the practical applicability of the developed techniques it is of the utmost importance that attacks are taken into account.

**The discrete case** In the discrete case the presence of attacks can be accounted for by noting that the input to the detector is no longer the vector $y$ as before, but another vector, $z = (z_1, \ldots, z_n)$, that is the output of a channel fed by $y$, which we shall denote by $W(z|y)$. For convenience, we will
assume that the components of \( z \) take on values in the same alphabet \( \mathcal{A} \). As we did in section 2, the problem to be solved, then, can be formulated as follows.

For a given \( w \), we must devise an embedding rule \( f(\cdot) \) and a decision rule that partitions the space of sequences \( y \) observed by the detector into two complementary regions \( \Lambda \) and \( \Lambda^c \), such that for \( y \in \Lambda \) we decide in favor of \( H_1 \) (watermark is present) and for \( y \in \Lambda^c \), we decide that the watermark is absent. The definition of the embedding and detection rule has to be made according to the Neyman-Pearson criterion of minimizing the false negative probability subject to a constraint on the distortion level and the false negative error exponent. The difference with respect to the case studied in section 2 is that now the expressions for the false positive and false negative probabilities must take into account the presence of the channel \( W(z|y) \). As to the false positive probability \( P_{e_1} \) two approaches are possible. According to the first one, \( P_{e_1} \) is calculated on the non-attacked sequence, leading to the same optimum detection region found in 2. However the embedding strategy has to be modified to account for the presence of the channel.

The second possibility consists in evaluating \( P_{e_1} \) on the attacked sequence. In this case the shape of the optimum detection region will also change.

The possibility of extending the analysis to the case of an unknown channel (leading to a so to say, universal watermarking scheme) will also be considered.

\[
\text{The continuous case} \quad \text{The theoretical framework for the continuous case introduced in [1] assumes that the host sequence is normally distributed. By following a similar approach we will study the optimum embedding strategy for an additive white gaussian (AWGN) channel, for which the attacked sequence may be expressed as:}
\]

\[
z = y + n, \quad (17)
\]

where \( n \) is an i.i.d. sequence of normally distributed random variables.

### 3.2 The non-Gaussian case

Despite the theoretical interest of the Gaussian case, in practice the host sequence is only rarely normally distributed, for this reason, among the problems addressed by the project, we will consider the extension of the framework described in section 2 to the non-gaussian case. Specifically, the optimum watermark embedding/detection strategies for sequences distributed according to a Laplacian or a Generalized Gaussian distributions will be considered.

### 3.3 Characterization of perceptually admissible DAs

Facing the problem of optimum watermark embedding/detection under DAs first of all requires that a suitable model for this class of attacks is defined. In order to be useful such a model should incorporate some measure of the perceptibility of the distortion when it is applied to a generic image. This is a cumbersome task since the way the human visual system perceives geometric distortion is not known. The approach we will follow is to describe the DA by means of a displacement field, upon which some smoothness constraints are imposed.

### 3.4 Optimum embedding/detection in presence of DAs

Once one or models of DA’s are available, they will be incorporated in the general framework of watermark embedding/detection in presence of attacks derived previously. If possible, both the discrete and the continuous cases will be considered.
3.5 Evaluation of the complexity of the optimum algorithms

A nice property of the optimum embedding algorithms derived in section 2 is that it has a rather low computational complexity. In fact, in the general discrete case such a complexity is polynomial in the size of the alphabet and does not increase with the size of the host sequence. The situation is even more favorable in the gaussian continuous case, where the optimum embedding problem is reduced to a 2-dimensional problem with a very limited complexity. It is one of the aims of the project to check whether these good computational properties are maintained in the more complex scenarios including generic and de-synchronization attacks.

3.6 Security evaluation

As it is becoming evident from a series of recent studies [9–11], robustness and security are conflicting requirements, given that the optimum embedder/detection strategies are defined in terms of watermark robustness (simply measured by means of the error probability), it will be interesting to evaluate the security of the new schemes. A possible approach to do so is to measure the mutual information between the watermark secret (namely the watermark sequence $w$) and the data available to the pirate, which in the simplest case corresponds to the watermarked sequence $y$. Interestingly, the optimum detection region is defined in terms of the mutual information between the watermark and the watermarked sequence, hence providing an easy way to measure the trade-off between security and robustness.

In the same line, another problem that is going to be investigated is the design of an optimum watermarking strategy subject to a given security constraint.
T1.2 - Characterization of admissible DAs

Project year: Y1

Abstract. Despite their importance, only few classes of geometric attacks are considered in the literature, most of which consist of global geometric attacks. The random bending attack contained in the Stirmark benchmark software is the most popular example of local DA, however its de-synchronization effectiveness is rather limited. In order to better describe the class of DAs we introduce two new classes of attacks, namely the LPDC and the MRF-DA attacks. Though the stochastic models behind the two new classes of attacks may not be able to describe all the perceptually admissible geometric attacks, they are general enough to form the basis of a global theory of watermarking in the presence of DAs. In particular the simplicity of the LPDC model makes it a suitable candidate for the subsequent studies of the project.

1 Introduction

In general a geometric distortion can be seen as a transformation of the position of the pixels in the image. It is possible to distinguish between global and local geometric distortions. A global transformation is defined by a mapping analytic function that relates the points in the input image to the corresponding points in the output image. It is defined by a set of operational parameters and performed over all the image pixels. Local distortions, instead, refer to transformations affecting in different ways the position of the pixels of the same image or affecting only part of the image. The random bending attack [12] contained in the Stirmark utility is the most famous example of local geometric transformation.

Global geometric transformations, especially rotation, scaling and translation, have been extensively studied in the watermarking literature given their simplicity and diffusion. Though no perfect solution exists to cope with geometric attacks, DAs based on global transformations can be handled in a variety of ways, including exhaustive search [13, 14], template-based re-synchronization [15–17], self-synchronizing watermarks [18, 19] and watermarking in invariant domains [20]. In all the cases, the proposed solutions rely on the restricted number of parameters specifying the DA. For instance, it is the relatively low cardinality of the set of possible attacks that makes the estimation of the geometric transformation applied by the attacker via exhaustive search or template matching possible (computationally feasible). Thus, recovering from localized attacks is much harder than recovering from a global attack. For this reason in this report (and throughout the whole project) we will focus only on local geometric attacks.

Despite the threats they pose, local geometric transformations have received little attention by the watermarking community. In practice, only the Random Bending Attack (RBA) contained in the Stirmark software has been studied to some extent. However even in this case, the real desynchronization capabilities of RBA are not fully understood, given that as implemented in Stirmark, RBA consists of three modules, only the last one corresponding to a truly local geometric transformation [12].

During the first year of the project, the activity of T1.2 focused on local geometric attacks for still images. In particular, the aim of our research was twofold:

- To introduce two new models to describe local DAs, that extend the class of local geometric attacks for still images;
- To evaluate the effectiveness of the attacks described by the new models and compare them with the classical RBA;
In the next section the two models we introduced, namely LPDC and MRF-DA are described. Before doing that, however, a brief review of the RBA attack implemented in the Stirmark software is given, so to ease its comparison with the new models we developed.

2 Stirmark RBA

In most of the scientific literature, by RBA the corresponding geometric attack implemented in the Stirmark software is meant [21], however such an attack is not a truly local attack since it couples three different geometric transformations applied sequentially, only the last of which corresponds to a local attack.

The first transformation applied by Stirmark is defined by:
\[
\begin{align*}
  x' &= t_{10} + t_{11}x + t_{12}y + t_{13}xy \\
  y' &= t_{20} + t_{21}x + t_{22}y + t_{23}xy
\end{align*}
\]  

(1)

where \(x', y'\) are the new coordinates and \(x, y\) the old ones. In practice this transformation corresponds to moving the four corners of the image into four new positions, and to modify coherently all the other sampling positions. The second step is given by:
\[
\begin{align*}
  x'' &= x' + d_{\text{max}} \sin(\frac{x' \pi}{M}) \\
  y'' &= y' + d_{\text{max}} \sin(\frac{y' \pi}{N})
\end{align*}
\]  

(2)

where \(M\) and \(N\) are the vertical and horizontal dimensions of the image. This transformation applies a displacement which is zero at the border of the image and maximum \((d_{\text{max}})\) in the center. The third step of the Stirmark geometric attack is expressed as:
\[
\begin{align*}
  x''' &= x'' + \delta_{\text{max}} \sin(2 \pi f_x x'') \sin(2 \pi f_y y'') \text{rand}(x'', y'') \\
  y''' &= y'' + \delta_{\text{max}} \sin(2 \pi f_x x'') \sin(2 \pi f_y y'') \text{rand}(x'', y'')
\end{align*}
\]  

(3)

where \(f_x\) and \(f_y\) are two frequencies (usually smaller than 1/20) that depend on the image size, and \(\text{rand}(x'', y'')\) are random numbers in the interval \([1, 2]\). Equation (3) is the only local component of the Stirmark attack since it introduces a random displacement at every pixel position. In the sequel by RBA we will mean only the transformation expressed by equation (3).

3 LPCD

In this section we describe a first new class of DAs, namely LPCD (Local Permutation with Cancellation and Duplication) DAs. We start from the plain LPCD attack, then we pass to the C-LPCD (Constrained LPCD). Finally we consider the multiresolution extension of the above two classes.

3.1 LPCD

By focusing on the 1D-case, let \(y = \{y(1) \ldots y(n)\}\) be a generic signal and let \(z = \{z(1) \ldots z(n)\}\) be the distorted version of \(y\). The LPCD model states that \(z(i) = y(i + \Delta_i)\) where \(\Delta\) is a sequence of i.i.d random variables uniformly distributed in a predefined interval \(I = [\Delta_{\text{max}}, \Delta_{\text{max}}]\). For simplicity we assume that \(\Delta\) can take only integer values in \(I\). In this way the values assumed by the samples of \(z\) are chosen among those of \(y\).

To extend the 1D-LPCD model to the 2-D case, if \(Z(i, j)\) is a generic pixel of the distorted image \(Z\), we let
\[
Z(i, j) = Y(i + \Delta_h(i, j), j + \Delta_v(i, j)),
\]  

(4)

where \(Y\) is the original image and \(\Delta_h(i, j)\) and \(\Delta_v(i, j)\) are i.i.d. integer random variables uniformly distributed in the interval \([-\Delta_{\text{max}}, \Delta_{\text{max}}]\).
3.2 C-LPCD

An important limitation of the LPCD model is the lack of memory. The lack of memory is likely to be a problem from a perceptual point of view: with no constraints on the smoothness of the displacement field there is no guarantee that the set of LPCD distortions is perceptually admissible even by considering very small values of $\Delta_{\text{max}}$.

One way to overcome the limitation of the LPCD model, and to obtain better results from a perceptual point of view, is to require that the sample order, in the 1D case, is preserved (thus introducing memory in the system). In practice, the displacement of each element $i$ of the distorted sequence $z$ is conditioned to the displacement of the element $i-1$ of the same sequence. In formulas, $z(i) = y(i + \Delta_i)$ where $\Delta_i$ is a sequence of i.i.d integer random variables uniformly distributed in the interval $I = \left[ \max (-\Delta_{\text{max}}, \Delta_{i-1} - 1), \Delta_{\text{max}} \right]$. In the sequel we will refer to this new class of DAs as C-LPCD. The 2D extension is obtained by applying the 1D C-LPCD transformation first by rows, then by columns.

A visual inspection conducted on a set of images distorted with the C-LPCD model reveals that changing the value of $\Delta_{\text{max}}$ does not change the intensity of the deformation. This phenomenon can be explained by resorting to the theory of Markov Chains. For simplicity let us focus again on the 1-D case. It is possible to design a Markov chain whose states correspond to the possible sizes of the 1-D case. It is possible to design a Markov chain whose states correspond to the possible sizes of the interval $I = \left[ \max (-\Delta_{\text{max}}, \Delta_{i-1} - 1), \Delta_{\text{max}} \right]$. By choosing for example $\Delta_{\text{max}} = 2$, the state space $S$ is $\{1, 2, 3, 4\}$ where the state $S_i$ corresponds to a size of $I$ equal to $i + 1$, i.e. state 2 corresponds to the set $I = \{0, 1, 2\}$ of size 3, while state 4 corresponds to $I = \{-2, -1, 0, 1, 2\}$ of size 5 and it is the maximum size that $I$ can have (with $\Delta_{\text{max}} = 2$).

In a more general case, given $\Delta_{\text{max}}$, the maximum size of $I$ is equal to $N = 2\Delta_{\text{max}} + 1$ and the transition matrix of the Markov chain (whose size is $2\Delta_{\text{max}} \times 2\Delta_{\text{max}}$) is:

$$P = \begin{bmatrix}
\frac{1}{4} & \frac{1}{2} & 0 & \cdots & 0 \\
\frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & \cdots \\
\frac{1}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{2\Delta+1} & \frac{1}{2\Delta+1} & \frac{1}{2\Delta+1} & \cdots & \cdots & \frac{1}{2\Delta+1}
\end{bmatrix}$$

(5)

where $\Delta = \Delta_{\text{max}}$ and each element of the matrix $p_{ij}$ is the transition probability of going from state $i$ to state $j$. From Markov chain theory [22], we know that if a Markov chain is aperiodic and irreducible then there exists a unique limit distribution that can be found by solving the following system:

$$\begin{cases}
\pi_j = \sum_{i \in S} \pi_i p_{ij} & \forall j \in S \\
\sum_{j \in S} \pi_j = 1
\end{cases}$$

(6)

where $\pi_i$ is the probability to be in state $i$. It is easy to show that in our case the chain is irreducible and aperiodic thus it is possible to find a unique limit distribution by solving the system in equation (6) that now is equivalent to:

$$\begin{cases}
\pi_1 = \sum_{i=1}^{2\Delta} \pi_i \cdot \frac{1}{2\Delta+1} \\
\pi_j = \sum_{i=j-1}^{2\Delta} \pi_i \cdot \frac{1}{2\Delta+1} & \text{with } j \text{ going from } 2 \text{ to } 2\Delta - 1 \\
\pi_{2\Delta} = \pi_{2\Delta-1} \cdot \frac{1}{2\Delta+1} + \pi_{2\Delta} \cdot \frac{2}{2\Delta+1} \\
\sum_{j=1}^{2\Delta} \pi_j = 1
\end{cases}$$
After some algebraic manipulations it is possible to find the limit probability distribution of the states:

\[
\begin{cases}
\pi_1 = \frac{1}{2^{\Delta+2}} \\
\pi_j = \frac{1}{2^{\Delta+2}} \frac{1}{\sum_{k=0}^{\Delta+2} \pi_k} \pi_1 & \text{with } j \text{ going from } 2 \text{ to } 2\Delta - 1 \\
\pi_{2\Delta} = \frac{2\Delta+1}{2^{2\Delta+1}} \pi_1
\end{cases}
\]  

(7)

Figure (1.a) shows the limit probability of the states (i.e., the probability to be in state \(i\) after \(n\) steps when \(n \to \infty\)) as a function of \(N\). Interestingly, regardless of the value of \(N\) and therefore of \(\Delta_{\text{max}}\) states 1 and 2 are by far the most probable states (\(\pi_1 = \pi_2\) except when \(\Delta_{\text{max}} = 1\)). This explains why changing the value of \(\Delta_{\text{max}}\) does not change the perceptual quality of the image. In fact even by increasing the value of \(\Delta_{\text{max}}\) states 1 and 2 are the most probable states, however state 2 corresponds to the interval \(I = \{0, 1, 2\}\) of size 3 so even by increasing the value of \(N\) the model tends to prefer small displacements. Furthermore it is easy to verify that the states reach their limit probability distribution already for very small values of \(n\).

To avoid this undesirable effect and to allow the model to generate a larger variety of displacement fields, we modified the Markov chain by changing the transition probabilities among the states in order to give a greater probability to the transitions that result in a larger interval \(I\). A way to do this is to assign the same probability (equal to \(\frac{\Delta_{\text{max}}}{2^{\Delta+2}}\)) to the transitions that cause a decrease of the size of \(I\), corresponding to the elements \(i, j\) with \(i = 1, ..., \Delta_{\text{max}}\) and \(j = 1, ..., i\) of the transition matrix, and to assign all the remaining probability mass, equal to \(1 - \sum_{j=1}^{\Delta_{\text{max}}} p_{ij}\), to the transition corresponding to the element \(i, j\) with \(i = 1, ..., \Delta_{\text{max}}\) and \(j = i + 1\), i.e. the transition whose effect is to enlarge the interval \(I\). The corresponding transition matrix becomes:

\[
P = \begin{bmatrix}
\frac{1}{2^{\Delta+1}} & \frac{2\Delta}{2^{\Delta+1}} & 0 & \cdots & 0 \\
\frac{2\Delta}{2^{\Delta+1}} & \frac{1}{2^{\Delta+1}} & \frac{\Delta - 1}{2^{\Delta+1}} & 0 & \cdots \\
\frac{\Delta - 2}{2^{\Delta+1}} & \frac{\Delta - 1}{2^{\Delta+1}} & \frac{1}{2^{\Delta+1}} & \frac{\Delta - 2}{2^{\Delta+1}} & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{2^{\Delta+1}} & \frac{1}{2^{\Delta+1}} & \frac{1}{2^{\Delta+1}} & \cdots & \frac{2\Delta}{2^{\Delta+1}}
\end{bmatrix}
\]  

(8)

Figure (1.b) shows the limit probability distribution of the states versus \(\Delta_{\text{max}}\) for the new Markov chain.

By comparing figures (1.a) and (1.b) it is evident that with the modified Markov Chain is possible to obtain larger displacement fields because regardless of the value of \(\Delta_{\text{max}}\) all the states have almost the same limit probabilities. In the sequel, when we consider the C-LPCD model, we will always refer to the improved version with the modified Markov chain.

The extension of the C-LPCD model to the 2D case is obtained applying the 1D algorithm by rows to obtain the horizontal displacement field \(\Delta_h(i, j)\), and by columns for the vertical displacements \(\Delta_v(i, j)\).

### 3.3 Multiresolution extension

To further improve the C-LPCD and LPCD models making the distortion less perceptible, we considered a multiresolution version of these attacks, whereby the DAs are applied at different resolutions to obtain the global displacement field: a low resolution displacement field is first generated, then a full size displacement is built by means of bicubic interpolation. The full resolution field is applied to the original image to produce the distorted image.

More specifically the multiresolution models consist of two steps. Let \(S \times S\) be the size of the image (for sake of simplicity we assume \(S\) is a power of 2). To apply the LPCD (or C-LPCD) model...
at the $L$–th level of resolution two displacement fields $\delta_h(i, j)$ and $\delta_v(i, j)$ with size $\frac{S}{2^L} \times \frac{S}{2^L}$ are generated. Then the full resolution fields $\Delta_h(i, j)$ and $\Delta_v(i, j)$ are built by means of bicubic interpolation. Note that in this way non-integer displacement values are introduced \textsuperscript{1}. The full resolution displacement fields $\Delta_h$ and $\Delta_v$ are used to generate the warped image $Z$ as follows:

\[
Z(i, j) = Y(i - \Delta_h(i, j), j - \Delta_v(i, j)).
\]  

\text{(9)}

As opposed to the original version of LPCD and C-LPCD, however, the presence non-integer displacements is now possible due to the bicubic interpolation. To account for this possibility, whenever the displacement vector points to non integer coordinates of the original image, the gray level of the

\textsuperscript{1} It is still possible to obtain integer displacements by applying a nearest neighbor interpolation instead of a bicubic one (of course at the expense of the smoothness of the displacement field).
attacked image $Z(i, j)$ is computed by means of bilinear interpolation. While the above interpolation does not have a significant impact on the visual quality of the attacked image, the possible introduction of new gray levels that were not present in the original image, complicates the LPCD and C-LPCD models.

### 3.4 Cardinality evaluation

A measure of the difficulty of coping with a given type of DA is given by the cardinality of the attack class. In fact, the larger the DA space the more difficult will be to recover the synchronization between the embedded and the detector, both in terms of complexity and accuracy. As a matter of fact, it is possible to show [14, 23] that as long as the cardinality of the DAs is subexponential, the exhaustive search of the watermark results in asymptotically optimum watermark detection with no loss of accuracy with regard to false detection probability. By contrast, when the size of the DA is exponential, simply considering all the possible distortions may not be a feasible solution both from the point of view of computational complexity and detection accuracy [23]. In order to evaluate the cardinality of the classes of DAs the perceptual impact of LPCD and C-LPCD must be taken into account.

Thus we first found the limits of the model parameters by means of perceptual considerations, then we estimated the cardinality of the various classes of LPCD DAs.

Let us observe that from a perceptual point of view LPCD DAs have a different behavior for different values of $N$ and for different levels of resolution $L$, in particular the image quality increases if the attacks are applied to a lower level of resolution (larger $L$) but, at the same time, the number of possible distortions decreases.

In T2.1 (see the report about the activity carried out in WP2) both subjective and objective tests were performed to establish the sensitivity of the human visual system to the geometric distortions introduced by the LPCD model as a function of the control parameters $N$ and $L$. In this way we were able to identify the range of values of the control parameters that do not affect image quality: for each level of resolution the maximum value of $N$ that can be used while keeping the distortion invisible was found.

We now use these considerations to estimate the cardinality of the class of LPCD DAs. For the LPCD model the number of possible admissible geometric distortions is simply equal, neglecting boundary effects, to $(N S^2 L)N S^2 L$, where $S$ is the size of the image. Then if we consider an $512 \times 512$ image, and we take into account the perceptual analysis carried out in T2.1, we obtain $2 \times 10^{89}$ different attacked images.

With regard to the C-LPCD model, we need to refer again to the theory of Markov chains. Let us consider the 1D case and the graph of the Markov chain describing the C-LPCD model. It is possible to construct a matrix $A$ of zeroes and ones, where $A_{i,j} = 1$ if in the graph there is an edge going from node $i$ to node $j$ and zero otherwise. The number of paths of length $n$ that start from node $i$ and end into node $j$ is given by the $(i, j)$ entry of the matrix $A^n$. Then, to evaluate the number of paths of length $n$ we only need to sum all the elements of the $A^n$ matrix. For large $n$ the computation of $A^n$ may not be computationally feasible. In such a case we may resort to a result of graph theory stating that the exponential growth rate of the number of paths of length $n$ in the graph is $e^{\lambda_{max} n} N$, where $\lambda_{max}$ is the largest eigenvalue of $A$. In the C-LPCD case, the practical values of $n$ are not very large, for instance for a $512 \times 512$ image, with $L = 5$ we have $n = 16$, then we can easily compute the matrix $A^n$ and derive the exact size of the C-LPCD class of attacks. Specifically, by remembering that the 2D extension of C-LPCD is obtained by applying the 1D C-LPCD DA first by rows and then by columns, we obtain the results reported in table 1.

With the above approach we were able to count all the distortions that it is possible to generate with the C-LPCD model. Nevertheless, the occurrence of a particular distortion configuration depends of the Markov chain transition matrix and is not constant for all the configurations. Thus, for a more appropriate evaluation of the cardinality of C-LPCD DAs, we need to refer to the entropy rate of the corresponding Markov chain. In this context the following result from information theory [24]
is useful: let \( \{X_i\} \) be a stationary Markov chain with stationary distribution \( \mu \) and transition matrix \( P \), then the entropy rate is
\[
H(X) = - \sum_{ij} \mu_i p_{ij} \log p_{ij}.
\] (10)

The knowledge of the entropy rate of the Markov chain and the Asymptotic Equipartition Property (AEP) [24] help us to find the number of possible distortions that can be generated with a so defined Markov chain, since it asymptotically corresponds to the number of typical sequences, i.e. \( 2^{nH} \).

After some algebraic manipulations, we find that in the case of C-LPCD with \( N = 5 \) and \( L = 5 \), \( H(X) \) is approximately equal to 1.4881 bits and the number of different distortions that is possible to generate is \( 2^{1.4881} \approx 4.76 \cdot 10^{14} \). In the same way, in the case of C-LPCD with \( N = 7 \) and \( L = 6 \), it is possible to generate \( 2^{6.6055} \approx 8.53 \cdot 10^{30} \) different distortions. By looking at Table 1 we can see that, as we expected, the cardinality of C-LPCD evaluated by considering the entropy rate of the Markov chain (second row) is much smaller than the number of possible distortions (first row).

<table>
<thead>
<tr>
<th>LPCD</th>
<th>C-LPCD ( L=5 ) (-N5)</th>
<th>C-LPCD ( L=6 ) (-N5)</th>
<th>C-LPCD ( L=6 ) (-N7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cardinality</td>
<td>( 2.93 \times 10^9 )</td>
<td>( 1.54 \times 10^{10} )</td>
<td>( 1.54 \times 10^{11} )</td>
</tr>
<tr>
<td>( 2^{nH} )</td>
<td>( 2.93 \times 10^9 )</td>
<td>( 4.76 \times 10^{14} )</td>
<td>( 8.53 \times 10^{30} )</td>
</tr>
</tbody>
</table>

Table 1. Cardinality evaluation of the LPCD attacks: in the first row the number of possible distortions is reported, the second row refers to the number of typical sequences.

4 MRF-DA

A problem with the C-LPCD model is that it does not take into account the 2D nature of images, since it is based on a 1D Markov chain. To overcome this limitation we introduced a new class of DAs based on the theory of Markov random fields [25]. We will refer to this new class of attacks as MRF-DA.

4.1 Model description

A random field \( F = \{F_1, F_2, \ldots, F_m\} \) is a family of random variables defined on a set \( S \), in which each random variable \( F_i \) takes a value \( f_i \) in \( L \).

\( F \) is said to be a Markov random field (MRF) on \( S \) with respect to a neighborhood system \( N \) if and only if the two following conditions are satisfied:

\[
P(f) > 0, \quad \forall f \in L^m \quad \text{(positivity)} \quad (11)
\]

\[
P(f_i | F_{S-i}) = P(f_i | f_{N_i}), \quad \forall i \in S \quad \text{(Markov property)} \quad (12)
\]

where \( f = \{f_1, \ldots, f_m\} \) is a configuration of \( F \) (corresponding to a realization of the field), \( P(f) \) is the joint probability \( P(F_1 = f_1, \ldots, F_m = f_m) \) of the joint event \( F = f \), and

\[
f_{N_i} = \{f_{i'} \} \quad \forall i' \in N_i \quad (13)
\]

denotes the set of values at the sites neighboring \( i \), i.e. the neighborhood \( N \) centered at position \( i \). The positivity is due to technical reasons, since it is a necessary condition if we want that the Hammersley-Clifford theorem (see below) holds [26].
To exploit MRFs characteristics in a practical way we need to refer to the Hammersley-Clifford theorem [25] for which the probability distribution of a MRF has the form of a Gibbs distribution, i.e.:

\[ P(f) = Z^{-1} \times e^{-\frac{1}{T}U(f)} \]  

where \( Z \) is a normalizing constant called the partition function, \( T \) is a constant called the temperature and \( U(f) \) is the energy function. The energy function

\[ U(f) = \sum_{c \in C} V_c(f) \]  

is a sum of cliques potentials \( V_c(f) \) over all possible cliques \( C \), where a clique \( c \) is defined as a subset of neighboring sites in \( S \). Thus the value of \( V_c(f) \) depends on the local configuration on the clique \( c \). The practical value of the theorem is that it provides a simple way of specifying the joint probability. \( P(f) \) measures the probability of the occurrence of a particular configuration: the more probable configurations are those with lower energies.

In our case, we can model geometric attacks with a random field \( F \) defined on the set \( S \) of the image pixels. The value assumed by each random variable represents the displacement associated to a particular pixel. Specifically, for each pixel \((x, y)\) we have two values for the two directions \( x \) and \( y \), for this reason each variable \( F_i \) is assigned a displacement vector \( f_i = (f_x, f_y) \in \mathbb{L} \times \mathbb{L} \). The advantage brought by MRF theory is that by letting the displacement field of a generic point \((x, y)\) of the image depend on the displacement fields of the other points of its neighborhood (let us indicate this set with the notation \( N(x, y) \)), we can automatically impose that the resulting displacement field is smooth enough to avoid annoying geometrical distortions.

![Fig. 2. Structure of a first order neighborhood system and corresponding pair-sites cliques.](image)

As we said, an MRF is uniquely determined once the Gibbs distribution and the neighborhood system are defined. In the approach proposed here, for each pixel \((x, y)\) only four neighbors of first order and the corresponding four pair-site cliques are considered, as described in figure (2). The potential function we used is a bivariate normal distribution expressed by:

\[ V((x,y),(\tilde{x},\tilde{y})) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ - \frac{(f_x - f_{\tilde{x}})^2}{2\sigma_x^2} - \frac{(f_y - f_{\tilde{y}})^2}{2\sigma_y^2} \right\} \]  

where \( f_x \) and \( f_y \) are the components of the displacement vector \( f_{(x,y)} \) associated to the pixel \((x, y)\), \((\tilde{x}, \tilde{y})\) is a point belonging to the 4-neighborhood of \((x, y)\), \( f_{\tilde{x}} \) and \( f_{\tilde{y}} \) are the \( x,y \) components of the displacement vector \( f_{(\tilde{x},\tilde{y})} \) associated to the pixel \((\tilde{x}, \tilde{y})\) and \( \sigma_x \) and \( \sigma_y \) are the two components of the standard deviation vector \( \sigma \) (these values are controlled by perceptual constraints).

A typical application of MRF in the image processing field is to recover the original version of an image (or a motion vector field) by relying on a noisy version of the image. By assuming that the original image can be described by means of a MRF, the above problem is formulated as a Maximum
a Posteriori estimation problem. Thanks to the Hammersley-Clifford theorem this corresponds to an energy minimization problem that is usually solved by applying an iterative relaxation algorithm to the noisy version of the image \[26\]. The problem we have to face here, however, is slightly different. We simply want to generate a displacement field according to the Gibbs probability distribution defined by equation (14) and the particular potential function expressed in (16).

To do so, the displacement field is initialized by assigning to each pixel \((x, y)\) in the image a displacement vector \(f_{(x,y)}\) generated randomly (and independently on the other pixels) in \(\mathbb{L} \times \mathbb{L}\) with \(\mathbb{L} = \{ f \in \mathbb{Z} : -c \leq f \leq c \}\) (the value of \(c\) is determined by following perceptual considerations). This initial random field is treated as a noisy version of an underlying displacement field obeying the MRF-DA model. The MRF-DA field is then obtained by applying an iterative smoothing algorithm to the randomly generated field. More specifically, the technique we used visits all the points of the displacement field and updates their values through the Iterated Conditional Mode (ICM) algorithm detailed in \[26\]. When the ICM algorithm starts, each pixel \((x, y)\) of the displacement field is randomly visited and its displacement vector updated by minimizing the potential function (16). Specifically, a local minimum is sought by letting

\[
\mathbf{f}_{\text{opt}}(x,y) = \arg \min_{f \in (\mathbb{L} \times \mathbb{L})} \sum_{(\tilde{x},\tilde{y}) \in N(x,y)} V((x,y),(\tilde{x},\tilde{y}))
\]

Note that in the above equation the displacements of the pixels in the neighborhood of \((x, y)\) are fixed, hence resulting in a local minimization of the Gibbs potential. After that each pixel is visited and the corresponding displacement updated, a new iteration starts. The algorithm ends when no new modification is introduced for a whole iteration, which is usually the case after 7-8 iterations.

As for the LPCD DAs, we also considered a multiresolution version of the MRF-DA, where the full resolution version of the the displacement field is built by interpolating the displacement field obtained by applying the MRF-DA at a resolution level \(L\). In figures (3.a) and (3.b) two examples of displacement fields generated with the MRF-DA model are shown, using respectively the parameters \(L = 6 \sigma = (1,1) c = 6\) and \(L = 4 \sigma = (7,7) c = 18\). With MRF-DA it is possible to obtain larger displacement vectors than with the LPCD attacks (due to the high value of the \(c\) parameter), while keeping the distortion invisible thanks to the ability of the iterative conditional mode to generate a very smooth field, as we can see from figure (3).

Regarding the cardinality evaluation of this new class of DAs, in principle all the displacement fields are allowed, with the most annoying distortions corresponding to very low probabilities (and thus very large Gibbs potential). In order to evaluate the cardinality of the MRF-DA class a first step could be to calculate the entropy rate of the field. However this is a prohibitive task given that no technique is known to calculate the entropy rate of even the simplest MRFs.

5 Conclusions

The two models we developed constitute a good basis for the analysis of watermarking in presence of a very general class of DAs. From a theoretical point of view the LPCD model will constitute the basis for the development of an optimum embedder/detector capable of coping with geometric attacks. The basis for such a theory are described later on in the section detailing the activity of T1.3. The perceptual aspects linked to both LPCD and MRF-DA have also been analyzed in WP2, and the results we have got are described in the corresponding report, where the de-synchronization efficacy of such attacks is tested on real images and real watermarking systems.


Fig. 3. Examples of displacement fields generated with MRF DA’s: (a) MRF with \( L = 6 \), \( \sigma = 1 \) and \( c = 6 \); (b) MRF with \( L = 4 \), \( \sigma = 7 \) and \( c = 18 \).


T1.3 - Optimum embedding/detection under DA

Project years: Y1,Y2

Abstract. The theoretical framework defined in T1.1 is extended so to consider several kinds of attacks, thus answering some of the questions set by the project. First the discrete case is considered deriving the form of the optimum embedder/detection strategy under a universal attack channel. Then the Gaussian continuous case is addressed for which the optimum embedding strategy for an AWGN channel is devised, both the universal and non-universal cases are considered. The extension to a few non-Gaussian (still memoryless channels) is also analyzed. Finally, geometric attacks are incorporated in the model.

1 Introduction

The theoretical framework developed in T1.1 may be generalized in several interesting ways. Before DAs are introduced into the picture, it is necessary that the optimum embedding/detection problem is solved for a set up including the presence of so-to-say conventional attacks. The activity carried out so far considered both the discrete and the continuous (Gaussian) cases. The results we have got are described in sections 2, 3, 4 and 5. We also started considering the case of geometric attacks. A first solution addressing a very simple, unrealistic, model for DAs was found as detailed in 6. A possible approach to cope with more sophisticated attacks, such as the LPCD ones, has also been devised (see section 7), which will constitute the basis for the future activity of the project.

2 Optimum embedding under a universal attack channel (year: Y1)

The first problem considered in T1.2 was the derivation of the optimum embedding and detection strategies for discrete sequences under the presence of an unknown (to the detector) attack, whose only limitation is the matching of a given distortion constraint. To illustrate the solution we developed, let us adopt the same notation used in T1.1. Specifically, Let \(x\) be an \(n\)-long cover-sequence drawn from an i.i.d. source \(X\) with finite alphabet \(\mathcal{X}\). Let \(y\) be the watermarked sequence (still taking on values in the finite alphabet \(\mathcal{X}\)). Let \(u = (u_1, u_2, \ldots, u_n)\) be the watermark, with \(u_i \in \{-1, +1\}\). Let \(\Lambda\) and \(\Lambda_c\) be the acceptance and rejection regions of the watermark detector.

As in T1.1, we consider the problem of finding the optimum acceptance region \(\Lambda\) and the optimum embedding rule \(f(x, u)\) according to the Neyman-Pearson criterion of minimizing the false negative probability subject to a given maximum false positive rate. Specifically the false detection probability is defined as the probability that a non-marked, non-attacked, sequence \(x\) belongs to the acceptance region \(\Lambda\). As in T1.1, we assume that the detector is based on \(\hat{P}_{xy}\), the empirical joint distribution of \(x\) and \(y\).

We consider the case of a universal covertext distribution, accordingly the constraint on the false detection probability takes the form:

\[
P_{fp} = \max_{P} \sum_{y \in \mathcal{Y}} P(y) \leq e^{-\lambda n}.
\]  

Given that \(P_{fp}\) is computed on the non-attacked sequences, the derivation of the optimum detection region is the same as in [1]. The optimum acceptance region \(\Lambda_*\) is then defined as:

\[
\Lambda_* = \left\{ y : n \hat{I}_{yu}(Y;U) > n\lambda - |\mathcal{X}| \ln(n+1) \right\}
\]
where $\hat{I}_{yu}(Y;U)$ is the empirical mutual entropy between the sequences $y$ and $u$.

As to the distortion constraints, we adopt a maximum distance measure, since this assumption results in a more easily implementable embedder. Specifically, the embedding function must obey the following constraint:

$$d(x, y) = \max_{i} |x_i - y_i| = \|x - y\|_\infty \leq D_e.$$  (3)

With regard to the attack, we model it as a memoryless channel characterized by a transfer probability $W(Z|Y)$ such that:

$$W(z|y) = 0, \forall (z, y) \in \mathcal{X} \times \mathcal{X} : |z - y| > D_a.$$  (4)

Apart from this constraint nothing else is known about the specific attack the watermark will undergo.

**Theorem 1.** Let $W$ be the set of all the attack channels satisfying equation (4). Let $u$ (the watermark) and $x$ (the covertext) be given. Let $\Lambda_*$ be defined as in (2). Let the sets $\mathcal{P}_u$ and $\mathcal{P}_a$ be defined as:

$$\mathcal{P}_u = \left\{ P_{y|xu}(Y|X, U) : P_{y|xu}(y|x, u) = 0 \forall (y, x) : |y - x| > D_e \right\},$$  (5)

$$\mathcal{P}_a = \left\{ P_{zy}|u|(Z|Y, U) : P_{zy}|u|(z|y, u) = 0 \forall (z, y) : |z - y| > D_a \right\}.$$  (6)

where we assume that $D_e \leq D_a$.

Then the embedding function that minimizes the missed detection probability works by choosing at random a sequence $y$ such that:

$$\hat{P}_{y|xu}(Y|X, U) = \arg \max_{P_{y|xu}(Y|X, U) \in \mathcal{P}_u} \left\{ \min_{P_{zy}|u|(Z|Y) \in \mathcal{P}_a} \frac{\hat{I}_{zu}(Y;U)}{\ln(n+1)} \right\}.$$  (7)

**Proof.** The embedder must choose the watermarked sequence $y$ so that the probability of missing the watermark is minimized. Given that the attack channel is not known, it makes sense to minimize the maximum missed detection probability over all the admissible channels. In formula:

$$f^*(x, u) = \arg \min_{y : d(y|x) \leq D_e} \left\{ \max_{W(Z|Y) \in W} \sum_{z \in \mathcal{X}} W(z|y) \right\}.$$  (8)

where $\mathcal{X}^e$ is the complementary set of $\mathcal{X}$. By reasoning as in [1] (section 3.5) it is easy to show that the above embedding rule is equivalent to the following minimization problem, where the minimization is performed over all the possible empirical conditional distributions $\hat{P}_{y|xu}(Y|X, U)$:

$$\hat{P}_{y|xu}(Y|X, U) = \arg \max_{P_{y|xu}(Y|X, U) \in \mathcal{P}_u} \left\{ \min_{W(Z|Y) \in W} \left\{ \min_{P_{zy}|u|(Z|Y) \in \mathcal{P}_a} \frac{\hat{I}_{zu}(Y;U)}{\ln(n+1)} \right\} \right\}.$$  (9)

$^1$We assume that the output of the channel takes on values in the same alphabet of the input.
By exchanging the minimization over $W(Z|Y)$ with that over $\hat{P}_{zyu}(Z|Y, U)$ the above equation can be rewritten as:

$$
\hat{P}_{ywu}(Y|X,U) = \arg \max_{\hat{P}_{ywu}(Y|X,U) \in P_a} \left\{ \min_{\hat{P}_{ywu}(Z|Y,U) \text{ such that } n I_{zuy}(Z;U|Y)} \left\{ \min_{W(Z|Y) \in W} \left[ I_{zuy}(Z;U|Y) + \sum_{a \in X} \hat{P}_y(a) D(\hat{P}_{zy}(Z|Y = a) || W(Z|Y = a)) \right] \right\} \right\} \quad (10)
$$

We now observe that the minimization over $\hat{P}_{zyu}(Z|Y, U)$ can be limited to the set $P_a$. In fact, if $\hat{P}_{zyu}(Z|Y, U)$ is not in $P_a$ it will exist an $a$ and a $z$ for which $\hat{P}_{zy}(z|a) \neq 0$ and $W(z|a) = 0$ leading to $D(\hat{P}_{zy}(Z|Y = a) || W(Z|Y = a)) = \infty$. It is evident then that such a $\hat{P}_{zyu}(Z|Y, U)$ can not minimize the objective function of equation (10)$^2$. The above observation leads us to re-write equation (10) as:

$$
\hat{P}_{ywu}(Y|X,U) = \arg \max_{\hat{P}_{ywu}(Y|X,U) \in P_a} \left\{ \min_{\hat{P}_{ywu}(Z|Y,U) \text{ such that } n I_{zuy}(Z;U|Y)} \left\{ \min_{W(Z|Y) \in P_a} \left[ I_{zuy}(Z;U|Y) + \sum_{a \in X} \hat{P}_y(a) D(\hat{P}_{zy}(Z|Y = a) || W(Z|Y = a)) \right] \right\} \right\} \quad (11)
$$

For any $\hat{P}_{zyu}(Z|Y, U) \in P_a$ it surely exists a $W(Z|Y) \in W$ such that $W(Z|Y = a) = \hat{P}_{zy}(Z|Y = a)$, given that $I_{zuy}(Z;U|Y)$ does not depend on $W(Z|Y)$ the minimization over $W(Z|Y)$ will always result in zeroing the divergence term of equation (11), yielding

$$
\hat{P}^*_y(Y|X,U) = \arg \max_{\hat{P}_{ywu}(Y|X,U) \in P_a} \left\{ \min_{\hat{P}_{ywu}(Z|Y,U) \text{ such that } n I_{zuy}(Z;U|Y)} \left[ I_{zuy}(Z;U|Y) \right] \right\} \quad (12)
$$

thus proving the theorem.

2.1 Comments

In addition to the theoretical interest of above result, the form of the optimum embedder summarized in (7) has a great practical interest since by restricting the search over $\hat{P}_{ywu}(Y|X,U)$ to $P_e$ and that over $\hat{P}_{zyu}(Z|Y, U)$ to $P_a$ significantly simplifies the optimization problem since the number of variables over which the minimization is performed is greatly reduced. We believe that with some proper simplifications and/or modifications the result of theorem 1, can constitute the basis of a practical watermarking scheme for real signals, e.g. still images.

3 Optimum embedding for AWGN channels (year: Y1)

The computation of the error exponent of the false negative error probability and the form of the optimum embedder when an additive white Gaussian noise (AWGN) is added to the watermarked signals is the second extension of the general framework described in T1.1 that we considered. As in the discrete case, we assumed that the false positive error probability is computed by considering the original, non-marked, sequence. Under this assumption, the form of the optimum detector does

---

$^2$ One may wonder whether the actual minimum is lower than $\infty$. This is indeed the case, since under the assumption $D_e \leq D_a$ the missed detection probability is strictly larger than zero.
not change, being again based on the correlation of the received signal and the watermark and on the energy of the received signal. The threshold fixed for obtaining a probability of false positive smaller or equal than $e^{-\lambda n}$ is also unaltered, given that no noise is added to the non-watermarked sequence.

Let us now formalize the behavior of the embedder. Let $u$ be the watermark and $x$ the host sequence. The embedder will produce a watermarked sequence $y = x + w$, where $w$ is the so called watermarking signal, i.e. the vector added to $x$ to watermark it. Note that in general $w$ will depend both on $u$ and $x$. We are interested in determining the optimum choice of $w$, where optimality corresponds to the minimization of the false negative error exponent ($E_{fn}$) in the presence of an AWGN attack. The layout we will follow for our derivation is the following: we fix $w$ corresponds to the minimization of the false negative error exponent ($E_{fn}$) and then we maximize it to find the optimum watermarking signal $w$.

In order to simplify the subsequent analysis it is convenient to introduce an ad hoc coordinate system for which the only components of $u$, $x$ and $w$ are along the first three coordinate axes. To do so, let $\{e_i\}_{i=1}^n$ be the unitary vectors defining the coordinate system. Given $u$, $x$ and $w$, we derive $e_1$, $e_2$ and $e_3$, by applying the Gram-Schmidt orthogonalization procedure to $u$, $x$ and $w$, and fix the other directions arbitrarily. As a consequence of the above choice we have: $u = (u_1, 0, 0 \ldots)$, $x = (x_1, x_2, 0 \ldots)$, $w = (w_1, w_2, w_3, 0 \ldots)$ and $y = (x_1 + w_1, x_2 + w_2, w_3, 0 \ldots)$, while all the components of the noise sequence $z$ will in general be different than 0. We are now ready to evaluate $E_{fn}$. To start with we can say that a false negative happens whenever

$$
\frac{(x_1 + w_1 + z_1)^2}{(x_1 + w_1 + z_1)^2 + (x_2 + w_2 + z_2)^2 + (w_3 + z_3)^2 + \sum_{j=4}^n z_j^2} < \cos^2(\beta),
$$

where $w_1^2 + w_2^2 + w_3^2 \leq nD_e$, and

$$
x_1^2 = n \cdot r \cdot \sin^2(\alpha),
$$
$$
x_2^2 = n \cdot r \cdot \cos^2(\alpha),
$$

being $r \triangleq \frac{|x|}{n}$, and $\alpha \triangleq \arcsin \left( \frac{<x,u>}{||x|| \cdot ||u||} \right)$. In order to also make explicit the growth of the embedding distortion when the number of dimensions is increased, we define $v = \frac{w}{\sqrt{n}}$, so $||v||^2 \leq D_e$. In this way, a false negative happens if

$$
(x_1 + \sqrt{n}v_1 + z_1)^2 \left( \frac{1}{\cos^2(\beta)} - 1 \right) = (x_2 + \sqrt{n}v_2 + z_2)^2 - (\sqrt{n}v_3 + z_3)^2
$$
$$
= (\sqrt{n}r \sin(\alpha) + \sqrt{n}v_1 + z_1)^2 \left( \frac{1}{\cos^2(\beta)} - 1 \right)
$$
$$
- \left[ \sqrt{n}r \cos(\alpha) + \sqrt{n}v_2 + z_2 \right]^2 - (\sqrt{n}v_3 + z_3)^2 < \sum_{j=4}^n z_j^2 = (n - 3)q,
$$

where $q \triangleq \frac{1}{n-3} \sum_{j=4}^n z_j^2$. Defining

$$
T_1 \triangleq (\sqrt{n} \sin(\alpha) + v_1)^2 \left( \frac{1}{\cos^2(\beta)} - 1 \right) - \left[ \sqrt{n} \cos(\alpha) + v_2 \right]^2 - v_3^2,
$$

and

$$
T_2 \triangleq -[z_1^2 + 2z_1(\sqrt{n} \sin(\alpha) + \sqrt{n}v_1)] \left( \frac{1}{\cos^2(\beta)} - 1 \right) + z_2^2
$$
$$
+ 2z_2 \left[ \sqrt{n} \cos(\alpha) + \sqrt{n}v_2 \right] + z_3^2 + 2\sqrt{n}v_3z_3,
$$

the condition for a false negative can be rewritten as

$$
nT_1 < (n - 3)q + T_2,$$
Therefore, we can write the probability of false positive as

\[ q > \frac{nT_1}{n-3} - \frac{T_2}{n-3}. \]

Recalling that the probability density function (pdf) of \(\frac{(n-3)Q}{\sigma^2} \) is a \(\chi^2\) distribution of \(n - 3\) degrees of freedom, we can write

\[ f_Q(q) = \left(\frac{1}{2}\right)^{\frac{(n-3)/2}{2}} \frac{1}{\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{(n-3)q}{\sigma^2}\right)^{\frac{(n-3)/2-1}{2}} e^{-\frac{(n-3)q}{2\sigma^2}}, \]

where \(q \geq 0\). For the random variable \(R\), corresponding to the squared Euclidean norm of the host signal normalized by the number of dimensions, \(\frac{e^R}{\sigma^2}\) follows a \(\chi^2\) distribution with \(n\) degrees of freedom, so we can write

\[ f_R(r) = \left(\frac{1}{2}\right)^{\frac{n/2}{2}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{nr}{\sigma^2}\right)^{\frac{(n)/2-1}{2}} e^{-\frac{nr}{2\sigma^2}}. \]

Furthermore, if we denote by \(\Psi\) the random variable whose samples are the values of \(\alpha\), we can see that the probability distribution of \(\Psi\) is

\[ P(\Psi \leq \alpha) = 1 - \frac{A_n(\pi/2 - \alpha)}{2A_n(\pi/2)}, \]

being \(A_n(\theta)\) the surface area of the \(n\)-dimensional spherical cap cut from a unit sphere about the origin by a right circular cone of half angle \(\theta\), i.e.

\[ A_n(\theta) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n}{2}+1\right)} \int_0^\theta \sin^{(n-2)}(\varphi) d\varphi. \]

From the last formula, it is easy to see that the pdf of \(\Psi\) is given by

\[ f_\Psi(\alpha) = \frac{\partial P(\Psi \leq \alpha)}{\partial \alpha} = \frac{2\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n}{2}+1\right)} \cos^{n-2}(\alpha). \]

Therefore, we can write the probability of false positive as

\[
P_{fn} = \left. \frac{\pi/2}{\alpha=-\pi/2 r=0 z_3=-\infty z_2=-\infty z_1=-\infty q=\max(0, \frac{nT_1}{n-3} - \frac{T_2}{n-3})} \right| \left[ \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} \int_0^{\frac{1}{2}} \left(\frac{1}{2}\right)^{\frac{(n-3)/2}{2}} \right. \\
\left. \frac{1}{\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{(n-3)q}{\sigma^2}\right)^{\frac{(n-3)/2-1}{2}} e^{-\frac{(n-3)q}{2\sigma^2}} \right. \\
\left. \frac{1}{\sqrt{2\pi} \sigma^2} \frac{1}{\sqrt{2\pi} \sigma^2} \frac{1}{\sqrt{2\pi} \sigma^2} \frac{1}{\sqrt{2\pi} \sigma^2} \right. \\
\left. \left(\frac{1}{2}\right)^{\frac{n/2}{2}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{nr}{\sigma^2}\right)^{\frac{(n)/2-1}{2}} e^{-\frac{nr}{2\sigma^2}} \frac{2\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n}{2}+1\right)} \cos^{n-2}(\alpha) \right. \\
\left. \frac{dqdz_1dz_2dz_3d\alpha}. \right. \]

Given that all the involved functions in the last formula are infinitely differentiable, and considering that \(\lim_{n \to \infty} \frac{nT_1}{n-3} - \frac{T_2}{n-3} = T_1\), since \(T_2\) grows sublinearly with the dimensionality of the
where the notation $T_1(r, \alpha)$ tries to make evident the dependence of $T_1$ with those variables.

First of all, we will study the dependence of the last formula with $\alpha$; on one hand, it is straightforward to see that $-\ln \left[ \cos(\alpha) \right]$ is minimized when $\alpha = 0$. On the other hand, $T_1$ also depends on
given that the embedder is interested in maximizing $T_1$, since in this way he is making the interval where the objective function can be minimized with respect to $q$ smaller, and changing the sign of any component of the watermark does not affect the embedding distortion, it is straightforward to see that the sign of $v_1$ and $v_2$ will be such that $v_1 \sin(\alpha) \geq 0$, and $v_2 \cos(\alpha) \leq 0$. Therefore $T_1(r, \alpha)$ is symmetrical with respect to $\alpha$, and its minimum is reached for $\alpha = 0$. Summarizing, the minimum of (14) is obtained for $\alpha = 0$. This implies that (14) can be rewritten as

$$
\lim_{n \to \infty} \frac{-1}{n} \ln(P_{fn}) = \min_{(q, r) \in [\max(0, T_1(r)), \infty] \times [0, \infty)} \frac{1}{2} \left[ \frac{q}{\sigma_Z^2} - \ln \left( \frac{q}{\sigma_Z^2} \right) - 1 \right] + \frac{1}{2} \left[ \frac{r}{\sigma_X^2} - \ln \left( \frac{r}{\sigma_X^2} \right) - 1 \right].
$$

(15)

Given that the objective function is convex with respect to $(r, q)$, and the global minimum is at $(\sigma_X^2, \sigma_Z^2)$, the result of (15) will be 0 if $(\sigma_X^2, \sigma_Z^2) \in [\max(0, T_1(r)), \infty] \times [0, \infty)$, and, in any other case, the minimum will be in the boundary of that region, i.e., the points of the form $(T_1(r), r)$, with $r \geq 0$.

So far we did not consider the choice of the optimal watermarking signal $(w_1^*, w_2^*, w_3^*, 0, \ldots, 0)$. The first question to be answered concerning this problem is the role that the watermarking signal plays in the optimization described by (15). In this case, it is easy to see that the only influence of $w^\ast$ (or equivalently $v^\ast$) on that formula is through (15). In other words, the embedder will choose the watermarking signal in order to maximize $T_1$, since in that way he will reduce the region where $q$ can take values; of course this will imply an increase of the error exponent. Given that the considered value of $\alpha$ is 0, as it was explained before, $T_1$ can be written as:

$$
T_1 = v_1^2 \left( \frac{1}{\cos^2(\beta)} - 1 \right) - \left[ \sqrt{r} + v_2 \right]^2 - v_3^2,
$$

that the embedder is trying to maximize constrained to

$$
v_1^2 + v_2^2 + v_3^2 \leq D_e.
$$

It is obvious that any component of the watermarking signal along $v_3$ will decrease $T_1$, and will also reduce the power available for spending in the other dimensions; therefore, $v_3^\ast = 0$. On the other hand, $T_1$ is monotonically increasing with $v_1^2$, so its maximum will be achieved when $v_1^2 + v_2^2 = D_e$, allowing us to express $T_1$ as

$$
T_1 = v_1^2 \left( \frac{1}{\cos^2(\beta)} - 1 \right) - \left[ \sqrt{r} - \sqrt{D_e - v_1^2} \right]^2.
$$

Computing the derivative of $T_1$ with respect to $v_1$, one obtains

$$
\frac{\partial T_1}{\partial v_1} = 2v_1 \left( \frac{1}{\cos^2(\beta)} - \frac{\sqrt{r}}{\sqrt{D_e - v_1^2}} \right),
$$

which is equal to 0 in the following cases

$$
\begin{cases}
0 & v_1 = 0 \\
-\sqrt{D_e - r \cos^4(\beta)} & v_1 = -\sqrt{D_e - r \cos^4(\beta)} \\
\sqrt{D_e - r \cos^4(\beta)} & v_1 = \sqrt{D_e - r \cos^4(\beta)}
\end{cases}
$$

Note that two solutions are possible for $v_2$, namely $v_2 = \pm \sqrt{D_e - v_1^2}$. Here we took the negative one, since, as we noted before, $v_2$ and $\cos(\alpha)$ must have opposite signs and given that $-\pi/2 \leq \alpha \leq \pi/2 \cos(\alpha)$ is always positive.
Considering the second derivative, it is easy to see that for \( v_1^* = \pm \sqrt{D_e - r \cos^4(\beta)} \) one obtains relative (in this case in fact they are absolute) maxima of \( T_1 \), yielding a value of \( v_2^* = -\sqrt{r \cos^2(\beta)} \), and a corresponding value of \( T_1 = D_e \tan^2(\beta) - r \sin^2(\beta) \). This gives us a first threshold for obtaining positive error exponents: if \( T_1 \leq 0 \), then the optimization in (15) is performed on the region \([0, \infty) \times [0, \infty)\), so any pair \((\sigma_Z^2, \sigma_X^2)\), even with \( \sigma_Z^2 = 0 \), will be in the allowed region, yielding an error exponent equal to 0. The condition for this not to happen is \( r \leq \frac{D_e}{\cos(\beta)} \).

When \( \alpha = 0 \), which as it was previously discussed is the case that asymptotically sums up most of probability, we can express the two components of the watermarked signal \( y \),

\[
y_1 = \pm \sqrt{n} \sqrt{D_e - r \cos^4(\beta)},
y_2 = \sqrt{n} \sqrt{1 - \cos^2(\beta)}.
\]

It is interesting to note that only the component of the watermarking signal in the dimension of the watermark is increased when \( D_e \) is increased. In fact, this embedding strategy can be seen as a subtraction of part of the interference due to the host signal \( x \), and then spending the remaining embedding distortion in increasing the component of the watermarked signal in the direction of the watermark. For example, when \( D_e = r \cos^2(\beta) \) the watermarked signal is the intersection of the boundary of the detection region and the perpendicular vector to that boundary that goes through \( x \).

Once we have obtained the optimal embedder for the noisy case, and we have an upper bound on \( r \) for obtaining positive error exponents, we can go back to the optimization of (15) when \((\sigma_Z^2, \sigma_X^2)\) does not belong to the region \([\max(0, T_1(r)), \infty) \times [0, \infty)\). The optimization is performed over the points of the form \((T_1(r), r) = (D_e \tan^2(\beta) - r \sin^2(\beta), r)\), with \( 0 \leq r \leq \frac{D_e}{\cos(\beta)} \). The derivative of (15) with respect to \( r \) takes the value

\[
\frac{1}{2} \left( -\frac{1}{r} + \frac{1}{\sigma_X^2} + \frac{\cos^2(\beta)}{D_e - r \cos^2(\beta)} - \frac{\sin^2(\beta)}{\sigma_Z^2} \right),
\]

being (15) piecewise convex in \((0, D_e / \cos^2(\beta))\), and \((D_e / \cos^2(\beta), \infty)\). Due to the constraints previously introduced, we are interested in the minimum in the interval \((0, D_e / \cos^2(\beta))\), which is achieved when

\[
r^* = \left( D_e \sigma_Z^2 + 2 \sigma_Z^2 \sigma_X^2 \cos^2(\beta) - D_e \sigma_X^2 \sin^2(\beta) \right.
- \sqrt{D_e^2 \sigma_Z^2 + 4 \sigma_Z^2 \sigma_X^2 \cos^4(\beta) - 2D_e^2 \sigma_Z^2 \sigma_X^2 \sin^2(\beta)^2 + D_e^2 \sigma_X^2 \sin^4(\beta)}

\left. \left( 2(\sigma_Z^2 \cos^2(\beta) - \sigma_X^2 \cos^2(\beta) \sin^2(\beta) \right)^{-1} \right).
\]

Replacing \( r \) by \( r^* \) in the definition of \( T_1(r) \) we get he value of \( q^* \), and replacing both of them in (15), the error exponent for the case of AWGN and optimal embedder is obtained:

\[
q^* = \left[ 2D_e \sigma_Z^2 + \sqrt{16 \sigma_Z^4 \sigma_X^4 \cos^2(\beta)^4 + D_e^2 [2 \sigma_Z^2 - \sigma_X^2 (1 - \cos(2\beta))]^2} \right] \tan^2(\beta)
\]

\[ - 2 \sigma_X^2 \sin^2(\beta) (2 \sigma_Z^2 + D_e \tan^2(\beta)) \left[ 4 (\sigma_Z^2 - \sigma_X^2 \sin^2(\beta)) \right]^{-1}, \]

\[
E_{fn} = \frac{1}{2} \left[ \frac{q^*}{\sigma_Z^2} - \ln \left( \frac{q^*}{\sigma_Z^2} \right) - 1 \right] + \frac{1}{2} \left[ \frac{r^*}{\sigma_X^2} - \ln \left( \frac{r^*}{\sigma_X^2} \right) - 1 \right].
\]

(16)
Particularly interesting is the case when $\sigma^2_Z = 0$, yielding

$$r^* = \frac{D_e}{\cos^2(\beta)},$$
$$q^* = 0,$$
$$E_{fn} = \frac{1}{2} \left[ \frac{D_e}{\sigma^2_X \cos^2(\beta)} - \ln \left( \frac{D_e}{\sigma^2_X \cos^2(\beta)} \right) - 1 \right].$$

In Figs. 1, 2 and 3 one can see the behaviour of the obtained error exponent of false negative as a function of the different parameters involved in its computation.

![Graph](image)

**Fig. 1.** Error exponent of probability of false alarm, as a function of $\lambda$, for several powers of AWGN. $\sigma^2_X = 1$ and $D_e = 2$.

### 3.1 Remarks

Thanks to the results obtained so far in the scope of this project the exact value of the error exponents of the false positive and false negative probabilities can be computed, both for the noiseless and AWGN attack scenarios, when the optimal detector with constrained complexity introduced in [8] is used. Furthermore, the optimal embedding strategy for both cases is derived, providing a new and surprising one-bit watermark algorithm. These results will be very useful for analyzing the impact of the desynchronization attacks, the very final objective of the project, as they will allow to obtain upper bounds to both the probabilities of false negative and false positive. Considering these bounds, one could conclude how harmful this family of attacks really is.

The activity described in this section led to the following publications.

Fig. 2. Error exponent of probability of false alarm, as a function of $\sigma_Z$, for several embedding distortions. $\sigma_X^2 = 1$ and $\lambda = 0.1$.

Fig. 3. Error exponent of probability of false alarm, as a function of $\sigma_X$, for several embedding distortions. $\sigma_X^2 = 1$ and $\lambda = 0.1$. 
4 Optimum embedding for AWGN channels with known statistics (year: Y2)

For the scenario where the variances of both the host signal and the noise are known by the embedder and the detector the previous derivation completely changes. First of all, we will have to consider the modifications on the optimal detection statistic, and define accordingly the detection region by comparing this new statistic with a detection threshold. The second step will be the computation of the mentioned threshold in order to verify the false positive error exponent constraint. Only then we will be able to derive the probability of false negative, and apply the necessary asymptotic approximations for obtaining the false negative error exponent, and study what will be the embedding strategy maximizing this number.

4.1 The optimal detection statistic

From [8] we know that the detection should be based on the comparison of \( \hat{I}_{uy}(U;Y) + D(\hat{P}_y||P_X) \) with a threshold, in such a way that if for a particular value of the received signal \( y \) the previous expression is larger than that threshold the watermark will said to be present, and absent otherwise. Be aware that the difference with the universal case comes from the second term, the Kullback-Leibler Divergence between the empirical pdf of the received signal \( y \), and the pdf of the original host signal \( X \). Given that

\[
\hat{I}_{uy}(U;Y) = -\frac{1}{2} \ln \left( 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{u_i y_i}{\sigma^2_X} \right),
\]

and

\[
D(\hat{P}_y||P_X) = D(\mathcal{N}(0,a)||\mathcal{N}(0,\sigma_X^2)) = \frac{1}{2} \ln \left( \frac{\sigma_X^2}{a} \right) + \frac{a}{2 \sigma_X^2} - \frac{1}{2},
\]

where \( a \triangleq \frac{1}{n} \sum_{i=1}^{n} y_i^2 \), the detection statistic in this case is given by

\[
-\frac{1}{2} \ln \left( \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} u_i y_i \right)^2 \right) + \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \frac{1}{2}.
\]

It is straightforward to see that the previous derivation can be modified for the case where the non-watermarked contents are attacked by the addition of Gaussian noise of variance \( \sigma_U^2 \), replacing the previous formula by

\[
-\frac{1}{2} \ln \left( \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} u_i y_i \right)^2 \right) + \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \frac{1}{2}.
\]

(17)

4.2 Derivation of the Detection Threshold as a function of the false positive error exponent.

We now have to determine the detection threshold \( T \) \((17)\) should be compared to, in order to achieve a false positive error exponent larger or equal than \( \lambda \).

In order to do so, we will rewrite \((17)\) as

\[
g(r_1, r_2) = -\frac{1}{2} \ln \left( \frac{r_1}{\sigma_X^2 + \sigma_U^2} \right) + \frac{r_1}{2(\sigma_X^2 + \sigma_U^2)} + \frac{r_2}{2(\sigma_X^2 + \sigma_U^2)} - \frac{1}{2},
\]
where the first component of the received signal \( y \) is that in the direction of the watermark \( u \), \( r_1 \overset{\Delta}{=} \frac{1}{n} \sum_{i=2}^{n} y_i^2 \) and \( r_2 = \frac{1}{n} y_1^2 \). Given that \( \frac{nR_1}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} \) is a \( \chi^2 \) distribution with \( n-1 \) degrees of freedom, its probability density function (pdf) is

\[
f_{R_1}(r_1) = \begin{cases} \frac{n}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} \left( \frac{1}{2} \right)^{(n-1)/2} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left( \frac{nR_1}{2(\sigma_\hat{X}^2 + \sigma_\hat{U}^2)} \right)^{(n-1)/2} e^{-\frac{nR_1}{2(\sigma_\hat{X}^2 + \sigma_\hat{U}^2)}}, & \text{if } r_1 \geq 0 \\ 0 & \text{elsewhere} \end{cases}
\]

For the sake of notational simplicity, we will define

\[
l(x) \overset{\Delta}{=} - \frac{1}{2} \ln \left( \frac{x}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} \right) + \frac{x}{2(\sigma_\hat{X}^2 + \sigma_\hat{U}^2)} - 1.5
\]

and \( H \overset{\Delta}{=} l(R_1) \). Given that \( l(x) \) is a convex function, with its minimum at \( \sigma_\hat{X}^2 + \sigma_\hat{U}^2 \), equal to 0, whenever \( x > 0 \), we can define the two inverse functions of \( l(\cdot) \), namely \( l^{-1}_1(\cdot) \) and \( l^{-1}_2(\cdot) \), providing solutions in the intervals \([0, \sigma_\hat{X}^2 + \sigma_\hat{U}^2] \) and \((\sigma_\hat{X}^2 + \sigma_\hat{U}^2, \infty) \) respectively. Therefore, the pdf of \( H \) is

\[
f_H(h) = \begin{cases} \frac{1}{2} \left( \frac{1}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} \right)^{(n-1)/2} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left( \frac{nR_1}{2(\sigma_\hat{X}^2 + \sigma_\hat{U}^2)} \right)^{(n-1)/2} e^{-\frac{nR_1}{2(\sigma_\hat{X}^2 + \sigma_\hat{U}^2)}}, & \text{if } h \geq 0 \\ 0 & \text{elsewhere} \end{cases}
\]

On the other hand, let us define \( M \overset{\Delta}{=} R_2 / 2(\sigma_\hat{X}^2 + \sigma_\hat{U}^2) \); the corresponding pdf is

\[
f_M(m) = \begin{cases} \frac{\sqrt{n}e^{-n m}}{\sqrt{m}} & \text{if } m \geq 0 \\ 0 & \text{elsewhere} \end{cases}
\]

and \( g(R_1, R_2) = H + M \). Given that \( H \) and \( M \) are independent random variables, the probability of false positive can be written as

\[
P_{fp} = \int_{Th}^{\infty} \int_{0}^{\infty} f_H(h-c) f_M(c) dc dh.
\]

In order to compute the error exponent of this probability, we use Laplace’s approximation, obtaining

\[
E_{fp} = \min_{(k,c,h) \in \{1,2\} \times [0,\infty) \times [\max(c,Th),\infty)} \left[ \frac{1}{2} \left( \frac{l_k^{-1}(h-c)}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} - \ln \left( \frac{l_k^{-1}(h-c)}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} \right) \right) - 1 \right] + c. \tag{19}
\]

Nevertheless, due to the definition of \( l(\cdot) \), it is possible to write

\[
\frac{1}{2} \left( \frac{l_k^{-1}(h-c)}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} - \ln \left( \frac{l_k^{-1}(h-c)}{\sigma_\hat{X}^2 + \sigma_\hat{U}^2} \right) \right) - 1 = l \left[ l_k^{-1}(h-c) \right] = h - c,
\]

for both the possible values of \( k \), i.e. \( k \in \{1, 2\}; \) therefore, (19) is tantamount to

\[
E_{fp} = \min_{(k,c,h) \in \{1,2\} \times [0,\infty) \times [\max(c,Th),\infty)} h = Th.
\]

This result establishes a clear relation between the detection threshold \( Th \) on the detection statistic (17), and the false positive error exponent \( \lambda \). Indeed, the main result of this analysis is that both values should coincide, i.e., \( Th = \lambda \).
4.3 False Negative Error Exponent Derivation.

For convenience, let us apply the Gram–Schmidt orthogonalization procedure to the vectors $u$, $x$ and $w$, and then select the remaining $n - 3$ orthonormal basis functions for $\mathbb{R}^n$ in an arbitrary manner. After transforming to the resulting coordinate system, the above vectors have the forms $u = (\sqrt{n}, 0, 0, \ldots, 0)$, $x = (x_1, x_2, 0, \ldots, 0)$, $w = (w_1, w_2, w_3, 0, \ldots, 0)$ and $y = (x_1 + w_1, x_2 + w_2, w_3, 0, \ldots, 0)$, while all the components of the noise sequence $z$ will remain, in general, non–null. In this way, we can define

$$
x_1^2 = n \cdot r \cdot \sin^2(\alpha),
$$

$$
x_2^2 = n \cdot r \cdot \cos^2(\alpha),
$$

$$
y \triangleq \frac{||x||^2}{n},
$$

$$
\alpha \triangleq \arcsin \left( \frac{\langle x, u \rangle}{||x|| \cdot ||u||} \right),
$$

$$
v \triangleq \frac{w}{\sqrt{n}},
$$

$$
q \triangleq \frac{1}{n} \sum_{j=4}^{n} z_j^2,
$$

$$
t_1 \triangleq \frac{z_1}{\sqrt{n}},
$$

$$
t_2 \triangleq \frac{z_2}{\sqrt{n}},
$$

$$
t_3 \triangleq \frac{z_3}{\sqrt{n}}.
$$

Therefore, we can rewrite (17) as

$$
l \left( \left( \sqrt{r} \cdot \cos(\alpha) + v_2 + t_2 \right)^2 + [v_3 + t_3]^2 + q \right) + \frac{\left( \sqrt{r} \cdot \sin(\alpha) + v_1 + t_1 \right)^2}{2(\sigma_X^2 + \sigma_U^2)}.
$$

Considering that a false negative event happens whenever the last formula is lower than $\lambda$, we can write that condition as

$$
l(T_2 + q) < T_1,
$$

where

$$
T_1 \triangleq \lambda - \frac{\left( \sqrt{r} \cdot \sin(\alpha) + v_1 + t_1 \right)^2}{2(\sigma_X^2 + \sigma_U^2)},
$$

and

$$
T_2 \triangleq \left[ \sqrt{r} \cdot \cos(\alpha) + v_2 + t_2 \right]^2 + [v_3 + t_3]^2.
$$

Note that whenever $T_1 \leq 0$, $l(\cdot)$ will not have inverse, therefore indicating that a false negative will not happen. In fact, a false negative will happen when

$$
l_1^{-1}(T_1) < T_2 + q < l_2^{-1}(T_1).
$$

This allows us to obtain the interval where $q$ should be contained in order to have a false negative event, given by

$$
\max(l_1^{-1}(T_1) - T_2, 0) < q < \max(l_2^{-1}(T_1) - T_2, 0).
$$
Therefore, the probability of false negative can be written as

\[
P_{fn} = \int_{\alpha=-\pi/2}^{\pi/2} \int_{\tau_0}^{\tau_1} \int_{\tau_2}^{\tau_3} \int_{t_0}^{t_1} \int_{T_0}^{T_1} \frac{n}{\sigma^2} \left( \frac{1}{2} \right)^{(n-3)/2} \frac{1}{T^2} \left( \frac{nq}{\sigma^2} \right) \left( \frac{n-3}{2} \right) e^{-nq \sqrt{n}e^{-\frac{nq}{2\sigma^2}}} \sqrt{2\pi\sigma^2} \left( \int_{t_0}^{t_1} \right) \left( \int_{T_0}^{T_1} \right) \cos^{-2}(\alpha) \ dq \ dz \ dz_2 \ dz_3 \ dz_4 \ dz_5.
\]

Applying the multidimensional version of Laplace’s approximation, we can conclude that

\[
E_{fn} = \lim_{n \to \infty} -\frac{1}{n} \ln(P_{fn})
= \min_{(q,t_1,t_2,t_3,r,\alpha) \in \left(\max \left( l_1^{-1}(T_1)-T_2,0 \right) \right) \times \left( \min \left( l_1^{-1}(T_1)-T_2,0 \right) \right) \times \left( -\infty,\infty \right) \times \left( -\infty,\infty \right) \times \left( 0,\infty \right) \times \left[ -\pi/2,\pi/2 \right]} \left\{ \frac{1}{2} \left[ \frac{q}{\sigma^2} - \ln \left( \frac{q}{\sigma^2} \right) \right] - 1 \right\} + \frac{1}{2} \left[ \frac{r}{\sigma^2} - \ln \left( \frac{r}{\sigma^2} \right) \right] - 1
+ \frac{t_1^2}{2\sigma^2} + \frac{t_2^2}{2\sigma^2} \right\} - \ln \left\{ \cos(\alpha) \right\}.
\] (20)

4.4 Two different strategies.

In order to obtain the solution to (20), we must consider that the target function is convex, with minimum value equal to 0, achievable at \((q,t_1,t_2,t_3,r,\alpha) = (\sigma Z, 0, 0, 0, \sigma Z, 0)\). For given values of \(r\) and \(\alpha\), the embedder will look for the watermarking signal \(w^*\) solution to the following problem,

\[
w^*(t_1, t_2, t_3) = \arg \max_{|w| \leq D} \min_{q \in \mathbb{R}} \left\{ \frac{1}{2} \left[ \frac{q}{\sigma^2} - \ln \left( \frac{q}{\sigma^2} \right) \right] - 1 \right\},
\] (21)

with

\[
t_q = \left( \max \left( l_1^{-1}(T_1)-T_2,0 \right) \right) \times \left( \min \left( l_1^{-1}(T_1)-T_2,0 \right) \right) \times \left( -\infty,\infty \right) \times \left( -\infty,\infty \right) \times \left( -\infty,\infty \right) \times \left[ -\pi/2,\pi/2 \right].
\]

and where we have done explicit the dependency of both \(T_1\) and \(T_2\) with \(t_1, t_2, t_3, w, r\) and \(\alpha\). In fact, considering the watermarking signal we could rewrite (20) as

\[
E_{fn} = \min_{(r,\alpha) \in (0,\infty) \times [-\pi/2,\pi/2]} \max_{|w| \leq D} \min_{q \in \mathbb{R}} \left\{ \frac{1}{2} \left[ \frac{q}{\sigma^2} - \ln \left( \frac{q}{\sigma^2} \right) \right] - 1 \right\} + \frac{1}{2} \left[ \frac{r}{\sigma^2} - \ln \left( \frac{r}{\sigma^2} \right) \right] - 1
+ \frac{t_1^2}{2\sigma^2} + \frac{t_2^2}{2\sigma^2} + \frac{t_3^2}{2\sigma^2} \right\} - \ln \left\{ \cos(\alpha) \right\}.
\]

Going back to (21), due to the convexity of the target function, that has its minimum equal to 0 at \(q = \sigma Z\), the watermarking signal design involves obtaining an interval of possible values of \(q\),

\[
\left( \max \left( l_1^{-1}(T_1)-T_2,0 \right) \right) \times \left( \min \left( l_1^{-1}(T_1)-T_2,0 \right) \right) \times \left( -\infty,\infty \right) \times \left( -\infty,\infty \right) \times \left( -\infty,\infty \right) \times \left[ -\pi/2,\pi/2 \right].
\]

such that it does not include \(\sigma Z\). In fact, in case that the previous interval does not contain \(\sigma Z\), the solution to the minimization in (21) is just the bound of the previous interval closer to \(\sigma Z\), independently of the other bound.
Two strategies are possible now: either maximizing the lower bound of the mentioned interval, i.e., \( \max \left( T_1 - 1, 0 \right) \), or minimizing the upper one, \( \max \left( T_2 - 1, 0 \right) \). In the first case the embedder will try to minimize both \( T_1 \) and \( T_2 \), whereas in the second one the embedder will try to minimize \( T_1 \) but maximize \( T_2 \). In any case, one should be aware that these two strategies depend on the value of \( t_1 \), \( t_2 \), and \( t_3 \), which in fact also affect the error exponent, making really complicated the full optimization process.

5 Optimum embedding for non-Gaussian channels (year: Y2)

In this section we extend the analysis carried out in the previous sections to the case of a hist signal following a generalized Gaussian distribution and no attacks present. To do so, consider the GG model for the covertext distribution:

\[
f_c(x) = \frac{\nu c}{\Gamma(1/\nu)} e^{-\frac{|x|^\nu}{\nu}} \quad \nu \geq 1, c > 0
\]  

(22)

where \( \nu = 1 \) corresponds to the Laplacian case and \( \nu = 2 \) corresponds to the Gaussian case. It is assumed that \( \nu \) is known, but \( c \) (the scaling parameter) is unknown.

Suppose that we are interested in a detector which bases its decision on the following statistics:

\[
\frac{1}{n} \sum_{i=1}^{n} |y_i|^\nu \text{ and } \frac{1}{n} \sum_{i=1}^{n} w_i y_i
\]  

(23)

For given \( w = (w_1, \ldots, w_n) \), \( y = (y_1, \ldots, y_n) \), and \( \epsilon > 0 \), let

\[
T_\epsilon(y|w) = \left\{ \tilde{y} : \left| \frac{1}{n} \sum_{i=1}^{n} |\tilde{y}_i|^\nu - \frac{1}{n} \sum_{i=1}^{n} |y_i|^\nu \right| \leq \epsilon, \left| \frac{1}{n} \sum_{i=1}^{n} w_i \tilde{y}_i - \frac{1}{n} \sum_{i=1}^{n} w_i y_i \right| \leq \epsilon \right\}
\]  

(24)

Let

\[
h(y) = -\frac{1}{n} \log \max \left( \prod_{i=1}^{n} f_c(y_i) \right) = \ln \left( 2 \left( e \nu \frac{1}{n} \sum_{i=1}^{n} |y_i|^\nu \right)^{1/\nu} \right)
\]  

(25)

and

\[
h(y|w) \triangleq \frac{1}{n} \ln \text{Vol} \left( T_\epsilon(y|w) \right)
\]  

(26)

Then, as we have seen in [8], the detector that accepts the \( H_1 \) hypothesis (watermark present) iff \( I(w, y) \geq h(y) - h(y|w) \geq \lambda \) is asymptotically optimum among all detectors that guarantee \( P_{FA} \leq e^{\lambda n} \) for all \( c > 0 \).

Thus, to make this decision rule implementable, we must have a good estimate of \( h(y|w) \). To this end, consider now an auxiliary memoryless channel \( P(y|w) = \prod_{i=1}^{n} P(y_i|w_i) \), where the component channel, \( P(y|w) \), is defined by

\[
P(y|w) = \frac{e^{-a|y|^{\nu} - bwy}}{\int_{-\infty}^{\infty} e^{-a|y|^{\nu} - bwy'} dy'}
\]  

(27)

where the denominator is
\[ f(a, b) \triangleq \int_{-\infty}^{\infty} e^{-a|y|^\nu - b|w|y} \, dy \]
\[ = \int_{0}^{\infty} e^{-(a^{1/\nu})^\nu (e^{by} + e^{-by})} \, dy \]
\[ = 2 \sum_{k=0}^{\infty} \frac{b^{2k}}{(2k)!} \int_{0}^{\infty} e^{-(a^{1/\nu})^\nu y^{2k}} \, dy \]
\[ = 2 \sum_{k=0}^{\infty} \frac{b^{2k}}{(2k)!} \int_{0}^{\infty} e^{-x \left( \frac{x}{a} \right)^{2k/\nu}} \cdot \frac{1}{\nu a^{1/\nu}} \left( \frac{x}{a} \right)^{1/\nu - 1} \, dx \]
\[ = \frac{2}{\nu a^{1/\nu}} \sum_{k=0}^{\infty} \frac{(b/a^{1/\nu})^{2k}}{(2k)!} \int_{0}^{\infty} e^{-x \left( \frac{2k+1}{\nu} \right)/\nu - 1} \, dx \]
\[ = \frac{2}{\nu a^{1/\nu}} \sum_{k=0}^{\infty} \frac{(b/a^{1/\nu})^{2k}}{(2k)!} \Gamma \left( \frac{2k+1}{\nu} \right). \]  

(28)

Then, it is possible to show (using probabilistic arguments) that for small \( \epsilon \) and large \( n \):

\[ h(y|w) \approx \min_{a,b} \left[ a \frac{1}{n} \sum_{i=1}^{n} |y_i|^{\nu} + b \frac{1}{n} \sum_{i=1}^{n} w_i y_i + \ln f(a, b) \right] \]  

(29)

For the Gaussian case \((\nu = 2)\), we have:

\[ h(y) = \frac{1}{2} \ln \left( 2\pi e \frac{1}{n} \sum_{i=1}^{n} y_i^2 \right) \]  

(30)

and

\[ f(a, b) = \frac{1}{\sqrt{\pi a}} \sum_{k=0}^{\infty} \frac{(b/\sqrt{a})^{2k}}{(2k)!} \Gamma \left( k + \frac{1}{2} \right) \]
\[ = \sqrt{\frac{\pi}{a}} \sum_{k=0}^{\infty} \frac{(b^2/(2a))^k}{(2k)!} \cdot 1 \cdot 3 \cdot \ldots \cdot 2k - 1 \]
\[ = \sqrt{\frac{\pi}{a}} \sum_{k=0}^{\infty} \frac{(b^2/(2a))^k}{2 \cdot 4 \cdot \ldots \cdot (2k)} \]
\[ = \sqrt{\frac{\pi}{a}} \sum_{k=0}^{\infty} \frac{(b^2/(2a))^k}{2^k \cdot k!} \]
\[ = \sqrt{\frac{\pi}{a}} \sum_{k=0}^{\infty} \frac{(b^2/(4a))^k}{k!} \]
\[ = \sqrt{\frac{\pi}{a}} e^{b^2/(4a)} \]  

(31)
which when substituted into 30, yields
\[
h(y|w) = \frac{1}{2} \ln \left[ 2\pi e \left( \frac{1}{n} \sum_{i=1}^{2} y_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} w_i y_i \right)^2 \right) \right]
\] (32)
and resulting detector agrees with the one in [8]. As for the Laplacian case ($\nu = 1$), we have
\[
h(y) = \ln \left( 2e \cdot \frac{1}{n} \sum_{i=1}^{n} |y_i| \right)
\] (33)
and
\[
f(a, b) = \frac{2}{a} \sum_{k=0}^{\infty} \left( \frac{b/a}{2k} \right)^{2k} \Gamma(2k + 1)
\]
\[
= \frac{2}{a} \sum_{k=0}^{\infty} \left( \frac{b}{a} \right)^{2k}
\]
\[
= \frac{2}{a} \left( 1 - b^2/a^2 \right)
\]
\[
= \frac{2a}{a^2 - b^2}
\] (34)

As to the detector, two cases must be considered: (i) $\frac{1}{n} \sum_{i=1}^{n} |y_i| = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$, or (ii) $\frac{1}{n} \sum_{i=1}^{n} |y_i| > \frac{1}{n} \sum_{i=1}^{n} w_i y_i$. The first case happens if either $w_i = sgn(y_i)$ for all $i$ or $w_i = -sgn(y_i)$ for all $i$. It is obvious that in this case the volume of $T_\epsilon(y|w)$ is a fraction $1/2^n$ of the volume of $\{ \tilde{y} : \frac{1}{n} \sum_{i=1}^{n} |\tilde{y}_i| - \frac{1}{n} \sum_{i=1}^{n} |y_i| \leq \epsilon \}$, which behaves like $e^{nh(y)}$ because the sign of all the components of $y$ are restricted. Thus, $I(w|y) = \ln 2$, which is the maximum possible value, and so, the detector should accept $H_1$ for any reasonable choice of $\lambda$.

As for the case (ii), let us denote
\[
U = \frac{1}{n} \sum_{i=1}^{n} |y_i|
\]
and
\[
V = \frac{1}{n} \sum_{i=1}^{n} w_i y_i.
\]
Then our goal is to minimize
\[
aU + bV + \ln(2a) - \ln(a + b) - \ln(a - b)
\]
over all real $a$ and $b$ with $a > |b|$. Taking derivatives w.r.t. $a$ and $b$, we get two equations:
\[
U + \frac{1}{a} - \frac{1}{a + b} - \frac{1}{a - b} = 0
\] (35)
\[
V - \frac{1}{a + b} + \frac{1}{a - b} = 0
\] (36)
The first equation can be written as
\[
\frac{a^2 + b^2}{a(a^2 - b^2)} = U.
\]
From the second equation, we can easily find that \( a = \sqrt{b^2 - 2b/V} \), which when substituted into the first equation, yields

\[
U = \frac{1 - bV}{\sqrt{b^2 - 2b/V}}.
\]

Squaring both sides, we get a quadratic equation whose solutions are

\[
b = \frac{1}{V}(1 \pm \frac{U}{\sqrt{U^2 - V^2}}).
\]

Of the two solutions, only the one with the minus sign, namely,

\[
b = \frac{1}{V}(1 - \frac{U}{\sqrt{U^2 - V^2}}),
\]

guarantees that \( b \) and \( V \) have opposite sign, a necessary and sufficient condition for \( a = \sqrt{b^2 - 2b/V} \) to be larger than \( |b| \). On substituting this solution of \( b \) into the expression \( a = \sqrt{b^2 - 2b/V} \), we get:

\[
a = \frac{1}{\sqrt{U^2 - V^2}}.
\]

Finally, substituting the resulting \( a \) and \( b \) into (30), we get

\[
h(y|w) = \ln\left(\frac{e^{V^2}}{U - \sqrt{U^2 - V^2}}\right)
\]

which gives

\[
I(w|y) = \ln[2\theta(\theta - \sqrt{\theta^2 - 1})] = \ln\left(\frac{2\theta}{\theta + \sqrt{\theta^2 - 1}}\right)
\]

where \( \theta \triangleq |U/V| \). As this is a monotonically non-increasing function of \( \theta \), then this is detector is equivalent to one that compares \( \theta \) to a threshold.

6 Optimum embedding/detection under permutation attacks (year: Y1)

After considering the optimum watermarking strategy in the presence of so-to-say conventional attacks, we turn the attention to the main goal of the project, i.e. the derivation of the optimum embedding and detection strategies in the presence of DAs. To start with, we defined the general framework and derived the solution for a scenario where the class of possible DAs is a subset of all the permutations of pixel positions. A more realistic set up will be sketched in the next section, where the class of LPCD attacks is considered.

The case of attack is characterized by the fact that the input to the detector is no longer the vector \( y \), but another vector, \( z = (z_1, z_2, \ldots, z_n) \), that is the output of a channel fed by \( y \), which we shall denote by \( W(z|y) \). For convenience, we will assume that the components of \( z \) take on values in the same alphabet \( A \). Thus, the operation of the attack, which in general may be stochastic, is thought of as a channel. Denoting the channel output marginal \( Q(z) = \sum_y P(y) W(z|y) \), the analysis of this case is, in principle, the same as before.

Let us now turn the attention to the specific case of geometric attack channels. A geometric attack creates, in general, a transformation of the coordinates of the stegotext signal, or image. Assuming that no information is lost by such a transformation, we will simply think of \( z \) as a (randomly chosen) permutation of \( y \) (e.g., a cyclic shift by a random amount). Let us assume then the following model.

There is a known set of \( M \leq n! \) possible permutations \( \{\pi_1, \pi_2, \ldots, \pi_M\} \) that the attack channel may apply, using the following mechanism: first, a random integer \( J \) is drawn, say, uniformly over
\{1, 2, \ldots, M\}, independently of \(y\), and then the attacker produces \(z = \pi_j(y)\), i.e., \(z\) is the result of the operation of the permutation \(\pi_j\) on the input \(y\). Thus,

\[ W(z|y) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}\{z = \pi_j(y)\} \quad (39) \]

Assuming that \(P\) is i.i.d. as before, we now argue that as long as \(M\) grows slower than exponentially in \(n\), the exhaustive search (ES) approach applied to \(\Lambda_*\) is asymptotically optimum. Let \(A_*(j)\) denote the set of all \(z\) such that \(\pi_j^{-1}(z) \in A_*\) (\(\pi_j^{-1}\) always exists), where here \(A_*\) is again as in Section 2.1. Consider a genie-aided detector which is informed of the realization \(j\) of the random variable \(J\). Clearly, such a detector can apply the inverse transformation \(y = \pi_j^{-1}(z)\) and we are back to the case without attack, as in Section 2.1. We now show that the ES approach of applying \(\Lambda_*\) to all inverse permutations performs asymptotically as well as this genie-aided detector. In particular, let us define

\[ A_{ES} = \bigcup_{j=1}^{M} A_*(j) \]

\[ = \left\{ z : \ln \left( P(z) \right) + n \min_j \hat{H}_{\pi_j^{-1}(z)}(Y|W) + \ldots + \lambda n - |A| \ln(n+1) \leq 0 \right\} \quad (40) \]

where we have used the fact that \(P(\pi_j^{-1}(z)) = P(z)\) as \(P\) is memoryless. Now, the probability of error of the second kind is bounded by

\[ P_{e_2} = \sum_{z \in A_{ES}} \sum_y P(y) W(z|y) \]

\[ = \sum_{z \in A_{ES}} \frac{1}{M} \sum_{j=1}^{M} P(\pi_j^{-1}(z)) \]

\[ = \sum_{z \in A_{ES}} P(z) \]

\[ \leq \sum_{j=1}^{M} \sum_{z \in A_*(j)} P(z) \]

\[ = \sum_{j=1}^{M} \sum_{\{z : \pi_j^{-1}(z) \in A_*\}} P(z) \]

\[ = \sum_{j=1}^{M} \sum_{\{z : \pi_j^{-1}(z) \in A_*\}} P(\pi_j^{-1}(z)) \]

\[ = \sum_{j=1}^{M} \sum_{y \in A_*} P(y) \]

\[ = M \cdot \sum_{y \in A_*} P(y) \quad (41) \]

which decays exponentially at the rate of \(e^{-\lambda n}\) since \(\sum_{y \in A_*} P(y)\) decays at such a rate and \(M\) is assumed subexponential. Thus, \(A_{ES}\) satisfies the false positive constraint (2). As for the error probability of the first kind, note that \(A'_{ES} = \bigcup_{j=1}^{M} A'_c(j)\) which means that \(A'_{ES} \subseteq A'_c(j)\) for
all \( j = 1, 2, \ldots, M \). Thus, \( P_{e_1} \) of the ES detector is smaller than that of the genie-aided detector whatever the realization of \( J \) may be.

The corresponding optimum embedder will now be

\[
f^*(x, u) = \arg \min_{\{y : d(x, y) \leq nD\}} \sum_{z \in A_{ES}} W(z | y)
= \arg \min_{\{y : d(x, y) \leq nD\}} \sum_{j=1}^{M} \sum_{j=1}^{M} \{z = \pi_j(y)\}
= \arg \min_{\{y : d(x, y) \leq nD\}} \sum_{j=1}^{M} \{\pi_j(y) \in A_{ES}'\}
\]

Evidently, this is not a convenient formula to work with. A reasonable approach would be to approximate \( \sum_{j=1}^{M} \{\pi_j(y) \in A_{ES}'\} \) by \( \{\exists j : \pi_j(y) \in A_{ES}'\} \) and the corresponding embedder \( f' \) is

\[
f'(x, u) = \arg \min_{\{y : d(x, y) \leq nD\}} \{\|\log P(y) + n \max_i \hat{H}_{\pi_i^{-1}(\pi_j(y))}(Y|W)\| \}
= \arg \min_{\{y : d(x, y) \leq nD\}} \{\|\log P(y) + n \max_i \hat{H}_{\pi_i}(u, \pi_j(y))(Y|W)\| \}
\]

We now argue that the approximation of \( f^* \) by \( f' \) causes \( P_{e_1} \) to grow by a factor of \( M \) at most, and hence it does not affect the exponential decay rate. To see this, first observe that

\[
1\{\exists j : \pi_j(y) \in A_{ES}'\} \leq \sum_{j=1}^{M} 1\{\pi_j(y) \in A_{ES}'\} \leq M \cdot \{\exists j : \pi_j(y) \in A_{ES}'\}
\]

Now, let us denote

\[
P_{e_1}(f) = \sum_x P(x) \cdot \frac{1}{M} \sum_{j=1}^{M} 1\{\pi_j(f(x, u)) \in A_{ES}'\}
\]

and

\[
P'_{e_1}(f) = \sum_x P(x) \cdot \frac{1}{M} \cdot 1\{\exists j : \pi_j(f(x, u)) \in A_{ES}'\}
\]

Then, following (43), we have \( P_{e_1}(f) \leq M P'_{e_1}(f) \leq M P_{e_1}(f) \) for every \( f \), and since \( f' \) minimizes \( P'_{e_1}(f) \), we also have:

\[
P_{e_1}(f') \leq M P'_{e_1}(f') \leq M P'_{e_1}(f^*) \leq M P_{e_1}(f^*)
\]

which proves this argument. The computation associated with \( f' \) is still quite involved in general. However, if the set of permutations \( \{\pi_j\} \) forms a group (e.g., the set of all \( M = n \) cyclic shifts), then things can be substantially simplified: let the group operation \( \ast \) be given by the rule \( \pi_i(\pi_j(\cdot)) = \pi_{i \ast j}(\cdot) \), and let the inverse of \( i \), denoted \( i^{-1} \), be induced by \( \pi_{i^{-1}}(\cdot) \). Then, \( \pi_{i^{-1}}(\pi_j(\cdot)) = \pi_{i^{-1} \ast j}(\cdot) \), and so,

\[
\max_i \min_j \hat{H}_{\pi_i^{-1}(\pi_j(y))}(Y|W) = \max_i \min_j \hat{H}_{\pi_i^{-1} \ast j(y)}(Y|W)
= \max_j \min_k \hat{H}_{\pi_k(y)}(Y|W)
= \min_k \hat{H}_{\pi_k(y)}(Y|W)
\]
Now, in the double minimization
\[
\min_{\{y : d(y, y') \leq nD\}} \min_k \left[ \ln P(y) + n \tilde{H}_{u, \pi_k}(y|W) \right]
\]
implemented by the embedding function \( f' \), the order of the minimizations can be interchanged, and then, the minimization over \( y \) can be carried out first. For every given \( k \), the complexity of this minimization is as described in T1.1. The overall complexity will then be proportional to \( M \) (due to the additional minimization over \( k \)), and hence sub-exponential in \( n \).

7 Optimum embedding/detection under LPCD attacks: a possible approach

(year: Y1)

In order to make a step towards a set up that is closest to reality, we have considered the optimum watermarking strategy in the presence of a non-constrained LPCD attack. As we show later, the model is simple enough to allow for a derivation of the optimum detection strategy, whereas the definition of the optimum embedder needs still to be studied.

To start with let us observe that assuming an LPCD attack, the channel input sequence \( y = (y_1, \ldots, y_n) \) is mapped to a forgery \( z = (z_1, \ldots, z_n) \) under a channel \( W(z|y) \) defined as follows (neglecting border effects):
\[
W(z|y) = \prod_{i=1}^{n} W(z_i|y_i^{i+K}), \quad (48)
\]
where: (i) \( y_i^j \), for \( i \leq j \) denotes \( (y_i, y_{i+1}, \ldots, y_j) \) (and a similar notation convention will apply to other signals), (ii) the notation \( \Delta_{\text{max}} \) was replaced by \( K \) (just for convenience), and (iii) according to LPCD definition:
\[
W(z_i|y_i^{i+K}) = \frac{1}{2K+1} \sum_{k=-K}^{K} 1\{z_i = y_{i-k}\}, \quad (49)
\]
thus \( W(z_i|y_i^{i+K}) \) assigns the same probability, \( 1/(2K+1) \), and independently, to all possible values of \( k \in \{-K, -K+1, \ldots, K\} \) and picks \( z_i = y_{i-k} \). However, we will allow any probability assignment \( W(z_i|y_i^{i+K}) \), not necessarily the uniform one, and moreover, will not assume that it is known a-priori. Likewise, the probability law of \( y \) will not be assumed known either, except from the fact that it is memoryless.

An equivalent representation of this model is obtained by defining \( u_i = (y_{i-K}^{i+K}) \). Here, if \( \{y_i\} \) are i.i.d., then \( \{u_i\} \) is a first-order Markov process. Also, the channel \( W \) from \( u = (u_1, \ldots, u_n) \) to \( y \) is obviously memoryless according to (48). Thus, \( z \) is governed by a hidden Markov process:
\[
Q(z) = \sum_u \prod_{i=1}^{n} [P(u_i|u_{i-1})W(z_i|u_i)]. \quad (50)
\]
As usually, we would like to keep the false-positive probability below some exponent \( 2^{-\lambda n} \). Let us consider detectors that base their sufficient statistics on joint empirical distributions of \( \ell \)-blocks \( \{(u_i^{i+\ell}) : y_i^{i+\ell}\} n/\ell - 1 \), assuming that \( \ell \) divides \( n \). It makes sense to choose \( \ell \) significantly smaller than \( n \) (to gather enough statistics), but significantly larger than \( K \), if possible. Suppose that a genie provides us with a subsequence of the channel inputs \( \tilde{u} = (u_1, u_{\ell+1}, u_{2\ell+1}, \ldots, u_n) \), in addition to \( z \). Let \( \Omega \) be a decision region where the genie–aided detector decides that the given watermark
Thus, another version of our decision rule detects the watermark of yet another asymptotically small term empirical conditional mutual information, $H$.

Note that the terms added to $\Lambda$ (and hence improved) in terms of the conditional Lempel-Ziv length function $\Lambda_{\ell} ^a$ follows from Lemma 2 in [28]. Thus, any decision region that satisfies the FP constraint

$$2^{-\lambda n} \geq \max_Q \sum_{(z, \tilde{u}) \in \Omega} Q(z, \tilde{u})$$

$$= \max_Q \sum_{(z, \tilde{u}) \in \Omega} \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1})$$

$$\geq |T_\ell(z|w, \tilde{u})| \cdot \max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1})$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

$$\geq \left( \frac{n}{\ell} + 1 \right)^{-A^*} \cdot 2^{n/\ell} \cdot (\max_Q \prod_{i=0}^{n/\ell-1} Q(z_{i\ell+1}^{i\ell+\ell}, u_{i\ell+\ell}|u_{i\ell+1}))$$

where $T_\ell(z|w, \tilde{u}) \subset \Omega$ is the conditional type of $z$ given $w$ and $\tilde{u}$ in the superalphabet of $\ell$-vectors, $H_\ell(Z^\ell|W^\ell, \tilde{U})$ is the corresponding empirical conditional entropy, $I(\tilde{U};Z^\ell|W^\ell)$ is the associated empirical conditional mutual information, $A$ is the alphabet of $y$, and the last inequality (with $\epsilon _\ell \to 0$ as $\ell \to \infty$) follows from Lemma 2 in [28]. Thus, any decision region that satisfies the FP constraint is a subset of

$$A = \left\{ z : \frac{1}{\ell}H_\ell(Z^\ell|W^\ell) + \frac{1}{n} \log \max_Q Q(z) \leq -\lambda + \frac{(2K + 1) \log |A|}{\ell} + \epsilon _\ell + \frac{|A|^\ell}{n} \log (\frac{n}{\ell} + 1) \right\} .$$

(52)

Note that the terms added to $-\lambda$ on the right-hand side are negligible when $\ell >> K$ and $n >> \ell$.

Now, as is shown in [28] (Lemma 1), the empirical conditional entropy is further lower bounded (and hence improved) in terms of the conditional Lempel-Ziv length function $U(z|w)$ at the expense of yet another asymptotically small term $\epsilon ^*(n, \ell)$, which tends to zero as $n \to \infty$ for any fixed $\ell$. Thus, another version of our decision rule detects the watermark $w$ whenever $y$ falls in the region:

$$A' = \left\{ z : \frac{1}{n}U(z|w) + \frac{1}{n} \log \max_Q Q(z) \leq -\lambda + \frac{(2K + 1) \log |A|}{\ell} + \epsilon _\ell + \epsilon ^*(n, \ell) + \frac{|A|^\ell}{n} \log (\frac{n}{\ell} + 1) \right\} .$$

(53)

The embedder corresponding to $A$, should be designed using considerations in the spirit of those that are described in [8]. However, here unlike the memoryless case, this seems to be computationally non-trivial, as it requires a search over exponentially many sequences, thus leaving open the search for a computationally manageable embedding strategy.

Part of the above findings have been published in...
T1.4 - Evaluation of complexity

Project years: Y2

Abstract. The computational complexity of the optimum embedding algorithm for the discrete case is analyzed both in the presence and absence of attacks. Possible suboptimum solutions are proposed to speed up the computation. With the proposed solutions the problem of optimum watermark embedding is feasible for sources with alphabets of moderate size.

1 Introduction

This chapter strongly relies on the analysis carried out in section 2 where the optimum embedding strategy for a generic channel is derived (for the discrete case). Specifically, we discuss the computational burden associates with the solution of the min-max problem described by equation 7 in chapter T1.2. Hereafter we will use the same symbolism adopted in chapter T1.2.

The problem of finding the optimal probability matrix to generate a watermarked image can be tackled by using an optimization procedure which modify the probabilities trying to achieve the maximum mutual information.

In the sequel we focus on the watermarking of a grey level image with 8 bits per pixels. Even if a generic 8-bit image contains 256 symbols (values from 0 to 255) and the $P_{yx|u}^*(Y|X,U)$ should be composed by $256 \times 256 \times 2 = 131072$ values, the parameter $D_e$ in equations (3) and (5) - chapter T1.2 - determines the effective number of non-zeros elements involved in the minimization. At the same time, the complexity of the inner optimization depends on the parameter $D_a$ introduced, since it influences the number of variables of the inner optimization cycle.

The problem is that even if the simplifications introduced due to the constraints imposed by $D_e$ and $D_a$ greatly reduce the search space, the case where attacks to the watermarked image are considered still remains very difficult to solve.

2 Local Optimization

Many local optimizers where tried in order to see if we could find a good compromise between solution optimality and computational complexity, and here is the outcome of our experiments:

- Genetic Algorithms (too many variables generated in every step → memory problems)
- Pattern Search (local extrema tend to stop the algorithm too early → not acceptable results)
- Simplex Method (unable to include non linear constraints)
- Integer Programming (the probabilities are real values)

Finally we decided to use a general purpose algorithm to find a local minimum of a constrained nonlinear multivariable function. We chose an implementation that includes three alternative methods:

1. Active Set
2. Interior Point
3. Trust Region Reflective
For problems without prior knowledge of the Hessian of the objective function, large-scale algorithms like Trust Region Reflective cannot be used. These strategies achieve better probability of obtaining a global solution instead of local ones, but require that the Hessian is fully evaluated at every iteration; since our objective function based on the mutual information is highly non linear, we were unable to use such large-scale optimizer and we had to rely on alternative methods like Active Set or Interior Point.

The first one uses a Sequential Quadratic Programming procedure that solves a quadratic programming subproblem at each iteration and uses an estimate of the Hessian of the Lagrangian with the BFGS formula as in [29]; the line search procedure uses a merit function similar to that proposed by [30].

The second one is based on a Trust Region Method that solve nonlinear programming problems [31], but in our tests it obtained similar solutions as the first one, but in a longer time, so we decided to use the Active Set algorithm.

3 Benchmarks

Since an analytical expression of the computational complexity of the optimum embedders cannot be directly derived from equations (7) and (12) of chapter T1.2, we implemented both objective functions and then we used the optimizer to record the time taken to find solutions depending on the number of symbols ($N$) and the maximum embedding/attacking distortions ($D_e$ and $D_a$).

We also developed a graphical user interface in order to visualize results and make the parameter tuning easier (Figure 1).

For all the experiments, the watermark has been generated from an uniform distribution of 2 symbols (binary signal), while the host signal was generated from an uniform distribution of $N$ symbols ($N = 256$ in the case of 8-bit images).

For completeness, here is a list of the other parameters of the optimizer:

- Maximum iterations = 50
- Convergence criterion = $10^{-6}$
- Maximum SQP sub-iterations = 50

3.1 No-attacks case

When attacks are not considered - see equation (7) in chapter T1.2 - the complexity is reduced, hence we could make many practical experiments to evaluate the computational complexity of the problem.

The size of the watermark and the host were fixed to $256 \times 256$, while the number of symbols $N$ has been varied from 5 to 200; this experiment has been repeated twice, the first time with embedding distortion $D_e = 1$, the second time with $D_e = 2$.

In order to empirically evaluate the complexity of the problem we tried to fit several models to the measured execution times and we found that a good fit was given by a cubic polynomial function for both cases, so we can conclude that the overall complexity of the algorithm is $\approx O(n^3)$.

Even if the complexity rapidly increase with the number of symbols to be considered (i.e. the number of grey levels of the images), the Active Set algorithm was able to find a solution (local minimum) even if we chose to use the whole set of 256 symbols, so this kind of approach revealed to be feasible also in real scenarios.

3.2 Attacks case

When attacks are considered, the problem is far more complex, since every step of the outer minimization includes an inner step of minimization as in (12) chapter T1.2.
In this case we experienced that the computational cost of finding a solution exceeds the limits of modern hardware even with a moderate number of symbols. Figure 3 shows the measured times with symbols ranging from 5 to 50.

Even for this case we tried to find the best fit in order to empirically evaluate the computational requirements and the smaller error was given by a 4th degree function, so we can conclude that the algorithm that considers attacks has complexity $\approx O(n^4)$.

We could not provide additional measurements for problems with more than 50 symbols because of the too long times needed to optimize the matrix.

**Approximated solutions:** Since a complete solution for our problem was too hard to find we thought of some approximate solutions to be used as a tradeoff or as a starting point for the optimization with whole symbol set.

The idea is to consider the probabilities as a set of independent blocks to be separately optimized: to explain how, we observe that if we fix one symbol of the watermark, $\hat{P}_{y|x|u}(y|x,u)$ becomes a bidimensional square matrix with $N \times N$ elements.
Fig. 2. The measured times needed to insert the watermark into an host of $N$ symbols considering a maximum embedding distortion $D_e$ (no attacks to the watermarked image are considered); the superimposed plots shows the best fit obtained by a cubic polynomial function.

Fig. 3. The measured times needed to insert the watermark into an host of $N$ symbols considering a maximum embedding distortion $D_e = D_a = 1$ (attacks to the watermarked image are considered in this case); the superimposed plot shows the best fit obtained by a $4^{th}$ degree polynomial function.

Each plane of the full matrix represents the probabilities $\hat{P}_{YX|U}(Y \equiv i|X = j, U = k)$, where $k = 1$ or $k = 2$ depending on which of the two slices we want to consider. Imposing the embedding
distortion constraint $D_e$ as in (3), means that for each symbol in the host image $X$ (the rows of the matrix) not all the symbols can be used in the watermarked image $Y$ (the columns of the matrix).

Depending on the value of $D_e$, more or less non-zero values accumulates on both sides of the main diagonal and if we subdivide this matrix into square blocks, we note that the greater the block, the more non-zero values fall outside these blocks: if these values are neglected, the optimization procedure can be run independently on each block. Of course in this way only a suboptimal solution is found whose validity must be carefully studied. As it can be easily noted the number of neglected variables depends on the size of the blocks that are treated independently. On the other side, the larger the blocks the higher the complexity: what we need then is to find a good balance between the sub-optimality of the solution and execution time.

Every column $j$ of the slice $k$ of the matrices express $\hat{P}_{X,U}(Y|X=j, U=k)$, i.e. the probability of generating allowed symbols given a host symbol $j$ and a watermark symbol $k$ (1 or 2 in our case), thus we must ensure that these values respect the probability constraint $\sum_i \hat{P}_{X,U}(Y = i|X = j, U = k) = 1$. For example, if we take $N = 12$, the valid block size can be 2, 3, 4, 6 or 12: ignoring the limit cases (2 $\rightarrow$ too many approximation errors, 12 $\rightarrow$ the full case discussed previously), the other values represent compromises between speed and accuracy.

Figure 4 shows the effect of separating one slice of the matrix into blocks of size 3, 4, 6 while varying the embedding distortion $D_e$ from 1 to 2 (the blocks are darker squares, the 'X' are the non-zero variables to be optimized, while the 'O' are the ignored variables (fixed to 0 in our case)).

![Figure 4](image-url)

**Fig. 4.** The non-zero variables (X) and the approximation errors (O) when the block size is 3 (first row), 4 (second row), 6 (third row) and $D_e = 1$ (left column), $D_e = 2$ (right column).
Some additional efforts were made trying to investigate if performance would benefit from this approximation, but the outcome was that the errors introduced by the blocks were too high to be ignored. We also tried to use the approximate solution as a starting point for the full problem, but the speed gain did not exceed 10%–15%.

More experiments will be made in the future to effectively evaluate the validity of the proposed approximated solution.
T1.5 - Optimum trade-off between error probability, complexity and security

Project years: Y2

Abstract. The role of security within the scenarios studied in the previous sections is considered. A preliminary analysis regarding the security measures to be used to evaluate the security of a given watermarking scheme and the influence of the false positive/negative error exponents on such measures is made.

1 Basic definitions and problem formulation

In the data hiding literature it is possible to distinguish between two kinds of attacks:

– Attacks to robustness: the attacker tries to impede the communication between the embedder and the decoder/detector, by using off-the-shelf signal processing tools, e.g. geometrical attacks, Wiener filtering, compression, transcoding, addition of AWGN, etc..

– Attacks to security: the attacker tries to get knowledge of some secret parameter, usually named the secret key of the data hiding system, in order to develop a more sophisticated attack. For example, with the information the attacker has got about the secret key, he could try to modify a watermarked content in order to yield a non-watermarked decision at the detector, or the other way around, he could try to supplant a content provider, modifying a non-watermarked content in order to give a watermarked decision at the detector.

Although so far we have just focused on the study and design of optimal embedding strategies when geometrical and AWGN attacks are considered (i.e., we considered two particular examples of attacks to robustness), the weakness of most of data hiding methods against attacks to security has been made patent in the last years, raising the need to study the peculiarities of the proposed methods against this kind of attacks.

In this section we will introduce some measures that can be used order to quantify the weakness of a watermarking method against security attacks. Given that in most cases the attacker will have access to a set of contents, not knowing a priori if a given content is watermarked or not, a first step in security attacks is to provide some measure, that can help the attacker to decide if the content he is studying is watermarked or not. The measure most widely used when one wants to quantify the similarity between two possible distributions is the Kullback-Leibler Divergence (KLD), defined as

\[ D(P_X||P_Y) = \sum_{x \in X} P_X(x) \log \left( \frac{P_X(x)}{P_Y(x)} \right); \]

the smaller the KLD, the more similar the considered pdfs are. In our particular case, the KLD will be used for comparing the empirical pdf of the received signal \( y \), denoted by \( \hat{P}_y \), with the pdf of the original host \( X \), denoted by \( P_X \), so \( D(\hat{P}_y||P_X) \) will be our measuring function. Based on that measure, the attacker will fix a threshold, in such a way that if the obtained KLD is larger than that threshold, then the received signal \( y \) will not be said to follow the distribution of \( X \), or in other words, \( y \) will be considered to be watermarked.

Obviously a trade-off will exist when establishing the aforementioned threshold between the confidence on the fact of the considered content is really watermarked, and the number of contents that the attacker will use in the estimate of the secret key:
– The larger the threshold, the more confident the attacker will be about the fact that the selected signals are watermarked, i.e. the aforementioned KLD exceeding the threshold, are indeed watermarked. On the other hand, the number of contents verifying that condition will be reduced, thus the attacker will have a lower number of observations to estimate the secret key.

– On the other hand, the smaller the threshold, the larger the number of observations the attacker will have to estimate the secret key (as he will use all those signals that following the mentioned test are said to be watermarked). Nevertheless, a larger portion of signals in the set will not be watermarked, misleading the estimation process.

The attacker should carefully study the above trade-off in order to fix a threshold that allows him to reduce the secret key estimation error.

Once the attacker has separated a set of probably watermarked contents from the full set of observations, he will use such signals to estimate the watermark, which in the watermarking schemes considered so far constitutes the secret key. A useful measure for quantifying how much the attacker can learn about the watermark is given by the mutual information between the watermark and the considered observations \[ I(Y^1, Y^2, \ldots, Y^N, W) \], where \( Y^j \) denotes the \( j \)th observation. Of course, the larger the mutual information, the more the attacker can learn of the watermark. Nevertheless, one should take into account that this mutual information only gives an upper bound to the information leaking from the observations, i.e. it says what is the best the attacker can do. Real attacks require the design of the estimation algorithm. The variance of the estimation error will depend on the goodness of the particular algorithm. A lower bound to that variance is given by

\[ \sigma^2 \geq \frac{1}{2\pi e^2} h(W|Y^1, Y^2, \ldots, Y^N), \]

where \( h(W|Y^1, Y^2, \ldots, Y^N) \) is the residual entropy of the watermarked given the observations, which is obviously related to the information leakage as

\[ h(W|Y^1, Y^2, \ldots, Y^N) = h(W) - I(Y^1, Y^2, \ldots, Y^N; W). \]

In the cases considered in this report the watermark is i.i.d. Gaussian distributed with variance \( \sigma_w^2 \), hence \( h(W) \) is given by

\[ h(W) = \frac{n}{2} \log \left( 2\pi e \sigma_w^2 \right). \]

From a practical point of view, several techniques have been proposed in the literature for the estimation of the secret key of Additive Spread-Spectrum (Add-SS) or Improved Spread-Spectrum (ISS) techniques. Given that the nature of the scheme proposed in Section 3 is similar to those methods, it seems reasonable to think that similar strategies could be followed in order to attack the current method. Most of the attacking methods proposed for Add-SS and ISS are based on the use of Blind Source Separation methods \([11, 10]\), as Principal Component Analysis and Independent Component Analysis. These techniques basically estimate the direction of the observations with the maximum/minimum variance, as it is likely the direction of the watermark, and they have been shown to be very useful for the estimation of the watermark for the aforementioned Add-SS and ISS methods.

According to the reasoning above, both from a information theoretic and practical point of view, the following conclusions can be derived about the influence on security of the basic parameters of the watermarking system described in Section 3:

– Host signal variance: as in most watermarking and data hiding schemes, the larger the host variance, the smaller the information leakage, and therefore the more secure the system. This is due to the fact that in the embedding methods where the interference due to the host is not completely rejected, the larger the variance of that signal, the larger its interference, making more complicated not only the transmission of information, but also the estimation of the watermark.
Embedding distortion: again, as it was already observed in the literature for other watermarking and data hiding methods, the larger the embedding distortion, the easier the estimation of the watermark. It is in fact easy to see that a larger embedding distortion does not only make watermark detection easier, but it also gives clues of the direction where the watermarking signal is embedded (and therefore about the watermark signal itself).

False positive error exponent (or equivalently, the angle of the hypercones constituting of the detection region): given that the embedding strategy of the method introduced in Section 3 depends on the false positive error exponent, through the angle of the hypercone constituting the detection region, it is reasonable to think the security of the scheme will depend on this parameter. Nevertheless, there does not seem to be a trivial answer to the question of what value of \( \lambda \) will minimize or maximize the information leakage (quantify as the mutual information described above). In fact, for large false positive error exponents (small angle of the hypercones), the embedding strategy proposed in Section 3 implies that most of the embedding distortion is spent reducing the interference due to the host (or equivalently, reducing the power in the dimensions orthogonal to the watermark); for sure this will make attacks easier. On the other hand, for small false positive error exponents (large angle of the hypercones), most of the embedding distortion will be devoted to move the signal in the direction of the watermark, which is also good for the attacker. Therefore, as it was previously said, an easy intuitive solution to the problem of how the false positive error exponent influence on the security of the proposed watermarking system does not exist, and it should be analyzed case by case.

2 Optimum embedding/detection under A security constraint

A question that has not been answered yet is how much the security constraints reduce the performance of a watermarking system, or in other words, what is the price, in terms of false negative error exponent, that has to be paid in order to have a watermarking system that does not leak information about the secret key (or with the leaked information smaller that a given quantity). This problem is intrinsically related to the derivation of the optimal embedding strategy under this additional constraint (note that the derivation of the optimum detector remains unaltered, as the security constraint does not affect to the non-watermarked - original contents).

The embedding problem in this scenario can be formulated as

\[
\hat{P}_{y|x,u}^* = \arg\max_{P_{y|x,u}} \sum_{y \in Y} P_{y|x,u}(y|x,u) \max_{I(U;Y) \leq \epsilon} \hat{I}_{uy}(U;Y),
\]

where the first constraint bounds the embedding distortion, the second one bounds the false positive error exponent, and the third is the security constraint, with \( \epsilon \geq 0 \). Concerning \( I(U;Y) \), it can be written like

\[
I(U;Y) = H(Y) - H(Y|U);
\]

where \( H(Y|U) \) can be computed by taking into account that

\[
H(Y|U) = H(P_{Y|U}(y|u)) = H\left(\sum_x P_X(x) \frac{1}{|T^*(y|x,u)|}\right)
\]

with the inner part of the rightmost term being proportional to

\[
\sum_x P_X(x) 2^{-nH_{y|x,u}},
\]
whereas $H(Y)$ can be written as

$$H(Y) = H\left(\sum_u P_U(u) \sum_x P_X(x) \frac{1}{|T^*(y|x,u)|}\right);$$

given that

$$P_X(x) = 2^{-n[H_u(X)+D(\hat{P}_u||P_x)]},$$

and we can rewrite the second sum as a sum over the conditional type $T(x|u)$, with cardinality $2^n H_{au}(x|u)$, the inner term of the entropy in the rightmost part of (2) is proportional to

$$\sum_u 2^{-n} \sum_{T(x|u)} 2^{-n[I_{au}(U;X)+D(\hat{P}_u||P_X)]} 2^{-n \hat{H}_{yau}^*(y|x,u)},$$

showing the dependency of (1) with $\hat{P}_{yau}^*(y|x,u)$. 