An Information–Theoretic View of Watermark Embedding–Detection and Geometric Attacks

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
merhav@ee.technion.ac.il

Abstract

We propose an information–theoretic approach to watermark embedding and detection. We first introduce our approach to blind embedding/detection in the absence of attacks, and then expand the framework to allow desynchronization and geometric attacks. We prove, in this context, the asymptotic optimality of the exhaustive search approach.

1 Introduction

The problem of watermark embedding and detection/decoding under geometric attacks has received considerable attention throughout the recent years (see, e.g., [1]–[14] and references therein for theoretical aspects). In this work, we raise and examine certain fundamental questions with regard to customary methods of embedding and detection and suggest some new ideas, first, for the most basic setup, even without an attack, and then we extend the scope to include certain classes of desynchronization/geometric attacks.

The most popular approach to watermark embedding and detection has been the following: Denoting by $\boldsymbol{x} = (x_1, \ldots, x_n)$ a block from the covertext source and by $\boldsymbol{w} = (w_1, \ldots, w_n)$ the independent binary (± 1) watermark vector, the watermark embedding rule is normally taken to be additive (linear), i.e., the stegotext vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ is given by

$$\boldsymbol{y} = \boldsymbol{x} + \alpha \boldsymbol{w} \tag{1}$$

or multiplicative, where each component of $\boldsymbol{y}$ is given by

$$y_i = x_i(1 + \alpha w_i), \quad i = 1, \ldots, n, \tag{2}$$

where in both cases, the choice of α controls the tradeoff between quality of the stego–signal (in terms of the distortion relative to the covertext signal $\boldsymbol{x}$) and the detectability of the watermark - the "signal–to–noise" ratio.

Once the linear embedder (1) is adopted, elementary detection theory tells us that the optimal likelihood–ratio detector, assuming a zero–mean, Gaussian, i.i.d. covertext distribution, is a correlation detector, which decides positively (H_1: $\boldsymbol{y} = \boldsymbol{x} + \alpha \boldsymbol{w}$) if the correlation, $\sum_{i=1}^n w_i y_i$, exceeds a certain threshold, and negatively (H_0: $\boldsymbol{y} = \boldsymbol{x}$) otherwise. The reason is that in this case, $\boldsymbol{x}$ simply plays the role of additive noise. In a similar manner, the optimal test for the multiplicative embedder (2) is based on the different variances of the y_i's corresponding to $w_i = +1$ relative to those corresponding to $w_i = -1$, the former being $\sigma_x^2(1+\alpha)^2$, and the latter being $\sigma_x^2(1-\alpha)^2$, where σ_x^2 is the variance of each component of $\boldsymbol{x}$.

While in classical detection theory, the additivity (1), (or somewhat less commonly, the multiplicativity (2)) of the noise is part of the channel model, and hence cannot be controlled, this is not quite the case in watermark embedding, where one has, at least in principle, the freedom to design an arbitrary embedding function $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{w})$, trading off the quality of $\boldsymbol{y}$ and the detectability of $\boldsymbol{w}$. Clearly, for an arbitrary choice of f, the above desctibed detectors are no longer optimal in general.

The problem of finding the optimum watermark embedder f is not trivial: The probabilities of errors of the two kinds (false positive and false negative) corresponding to the likelihood–ratio 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 $\boldsymbol{y}$ and $\boldsymbol{w}$. For example, the energy of $\boldsymbol{y}$, $\sum_{i=1}^n y_i^2$, and the correlation $\sum_{i=1}^n w_i y_i$, which are the sufficient statistics used by the above described correlation detector. Further, if there is reason to suspect that the stegotext might be subjected to a certain cyclic–shift attack (see, e.g., [1]), one might wish to include also correlations between $\boldsymbol{y}$ and the corresponding possible shifted versions of $\boldsymbol{w}$. Other possible statistics are those corresponding to the likelihood–ratio detector of (2), namely, the energies $\sum_{i:\ w_i=+1} y_i^2$, and $\sum_{i:\ w_i=-1} y_i^2$, and so on.

Within the class detectors based on a given set of statistics, we will next show how to find the best embedder and detector which are asymptotically optimum in the Neyman–Pearson sense of trading off the exponents of the error probabilities of the two kinds. In doing so, we will use the techniques of [15] and references therein, and build on them the additional ingredient needed for devising the optimal embedder.

2 Basic Derivation

For the sake of simplicity, let us assume temporarily, that the components of $\boldsymbol{x}$ and $\boldsymbol{y}$ take on values in a finite alphabet $\mathcal{A}$. In the sequel, this assumption will be relaxed and $\mathcal{Y}$ will be allowed to be an infinite set, like the real line. The components of the watermark $\boldsymbol{w}$ will always take on values in $\mathcal{B} = \{+1, -1\}$ as mentioned earlier. Let us further assume that $\boldsymbol{x}$ is drawn from a given memoryless source P.

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

$$P_{e_1} = \sum_{\boldsymbol{x}:\ f(\boldsymbol{x},\boldsymbol{w}) \in \Lambda^c} P(\boldsymbol{x}) \tag{3}$$

subject to the following constraints:

(1) Given a certain distortion measure $d(\cdot,\cdot)$ and distortion level D, the distortion between $\boldsymbol{x}$ and $\boldsymbol{y}$, $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{x}, f(\boldsymbol{x},\boldsymbol{w}))$, does not exceed nD.

(2) The false positive probability is upper bounded by

$$P_{e_2} \stackrel{\Delta}{=} \sum_{\boldsymbol{y} \in \Lambda} P(\boldsymbol{y}) \le e^{-\lambda n}, \tag{4}$$

where $\lambda > 0$ is a prescribed constant.

In other words, we would like to choose f and Λ 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 λ.

As explained in the Introduction, this problem does not appear to be trivial. We therefore make the additional assumption regarding the statistics employed by the detector. Suppose, for example, that we are interested in the class of all detectors which base their decisions on the empirical joint distribution of $\boldsymbol{y}$ and $\boldsymbol{w}$:

$$\hat{P}_{\boldsymbol{wy}} = \{\hat{P}_{\boldsymbol{wy}}(w,y),\ w \in \mathcal{B},\ y \in \mathcal{A}\} \tag{5}$$

where

$$\hat{P}_{\boldsymbol{wy}}(w,y) = \frac{1}{n}\sum_{i=1}^{n} 1\{w_i = w, y_i = y\}, \quad w \in \mathcal{B},\ y \in \mathcal{A} \tag{6}$$

$1\{w_i = w, y_i = y\}$ being the indicator function of the event $\{w_i = w, y_i = y\}$, that is, $\hat{P}_{\boldsymbol{wy}}(w,y)$ is the relative frequency of the pair (w,y) along the pair sequence $(\boldsymbol{w},\boldsymbol{y})$. Following standard terminology in the information theory literature [16], we define the conditional type class of $\boldsymbol{y}$ given $\boldsymbol{w}$, and denote it by $T(\boldsymbol{y}|\boldsymbol{w})$, as the set of all sequences $\boldsymbol{y}' \in \mathcal{A}^n$ such that $\hat{P}_{\boldsymbol{wy}'} = \hat{P}_{\boldsymbol{wy}}$, that is, the set of all $\boldsymbol{y}'$ which have the same empirical joint distribution with $\boldsymbol{w}$ that $\boldsymbol{y}$ has. The requirement that the decision of the detector depends solely on $\hat{P}_{\boldsymbol{wy}}$ means that Λ and Λ^c are unions of conditional types classes of $\boldsymbol{y}$ given $\boldsymbol{w}$. Now, let $T(\boldsymbol{y}|\boldsymbol{w}) \subseteq \Lambda$. Then, we have

$$\begin{aligned} e^{-\lambda n} &\geq \sum_{\boldsymbol{y}' \in \Lambda} P(\boldsymbol{y}') \\ &\geq \sum_{\boldsymbol{y}' \in T(\boldsymbol{y}|\boldsymbol{w})} P(\boldsymbol{y}') \\ &\geq |T(\boldsymbol{y}|\boldsymbol{w})| \cdot P(\boldsymbol{y}) \\ &\geq (n+1)^{-|\mathcal{A}|} e^{n\hat{H}_{\boldsymbol{wy}}(Y|W)} \cdot P(\boldsymbol{y}). \end{aligned} \tag{7}$$

A few words of explanation are in order at this point: The first inequality is by the assumed false positive constraint, the second inequality is since $T(\boldsymbol{y}|\boldsymbol{w}) \subseteq \Lambda$, and the third inequality is due to the fact that all sequences within $T(\boldsymbol{y}|\boldsymbol{w})$ 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 well known lower bound on the cardinality of a conditional type class in terms of the empirical conditional entropy [16], defined as:

$$\hat{H}_{\boldsymbol{wy}}(Y|W) = -\sum_{w,y} \hat{P}_{\boldsymbol{wy}}(w,y) \ln \hat{P}_{\boldsymbol{wy}}(y|w) \tag{8}$$

where $\hat{P}_{\boldsymbol{wy}}(y|w)$ is the empirical conditional probability of Y given W. Defining now

$$\Lambda_* = \{\boldsymbol{y}: \ \ln P(\boldsymbol{y}) + n\hat{H}_{\boldsymbol{wy}}(Y|W) + \lambda n - |\mathcal{A}|\ln(n+1) \leq 0\}, \tag{9}$$

we have actually shown that every $T(\boldsymbol{y}|\boldsymbol{w})$ in Λ is also in Λ_*, in other words, if Λ satisfies the false positive constraint (4), it must be a subset of Λ_*. This means that $\Lambda_*^c \subset \Lambda^c$ and so the probability of Λ_*^c is smaller than the probability of Λ^c, i.e., Λ_*^c minimizes P_{e_1} among all Λ^c corresponding to detectors that satisfy (4). To establish the asymptotic optimality of Λ_*, it remains to show that Λ^* itself has a false positive exponent at least λ, which is very easy to show using the techniques of [15, eq. (6)] and references therein. Therefore, we will not include the proof of this fact here. Finally, note also that Λ_* bases its decision solely on $\hat{P}_{\boldsymbol{wy}}$, 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 $\Gamma_*^c(f)$ to be the inverse image of Λ_*^c given $\boldsymbol{w}$, i.e.,

$$\begin{aligned}\Gamma_*^c(f) &= \{\boldsymbol{x}: \ f(\boldsymbol{x},\boldsymbol{w}) \in \Lambda_*^c\} \\ &= \{\boldsymbol{x}: \ \ln P(f(\boldsymbol{x},\boldsymbol{w})) + n\hat{H}_{\boldsymbol{w},f(\boldsymbol{x},\boldsymbol{w})}(Y|W) + \lambda n - |\mathcal{A}|\ln(n+1) > 0\}, \end{aligned} \tag{10}$$

then following eq. (3), P_{e_1} can be expressed as

$$P_{e_1} = \sum_{\boldsymbol{x}\in\Gamma_*^c(f)} P(\boldsymbol{x}). \tag{11}$$

Consider now the following embedder:

$$f^*(\boldsymbol{x},\boldsymbol{w}) = \text{argmin}_{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\leq nD}\left[\ln P(\boldsymbol{y}) + n\hat{H}_{\boldsymbol{wy}}(Y|W)\right], \tag{12}$$

where ties are resolved in an arbitrary fashion. Then, it is clear by definition, that $\Gamma_*^c(f^*) \subseteq \Gamma_*^c(f)$ for any other competing f that satisfies the distortion constraint, and thus f^* minimizes P_{e_1} subject to the constraints.

3 A Few Important Comments

In this section, we pause to discuss a few important aspects of our basic results, as well as possible modifications that might be of theoretical and practical interest.

3.1 Implementability of the Embedder (12)

The first impression might be that the minimization in (12) is prohibitively complex as it appears to require an exhaustive search over the sphere $\{\boldsymbol{y}: \ d(\boldsymbol{x},\boldsymbol{y}) \leq nD\}$, 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,

$$\ln P(\boldsymbol{y}) = -n[\hat{H}_{\boldsymbol{y}}(Y) + D(\hat{P}_{\boldsymbol{y}}\|P)] \tag{13}$$

where $\hat{H}_{\boldsymbol{y}}(Y)$ is the empirical entropy of $\boldsymbol{y}$ and $D(\hat{P}_{\boldsymbol{y}}\|P)$ is the divergence between the empirical distribution of $\boldsymbol{y}$, $\hat{P}_{\boldsymbol{y}}$, and the source P. Moreover, if $d(\cdot,\cdot)$ is an additive distortion measure, i.e., $d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n d(x_i,y_i)$, then $d(\boldsymbol{x},\boldsymbol{y})/n$ can be represented as the expected distortion with respect to the empirical distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$, $\hat{P}_{\boldsymbol{xy}}$. Thus, the minimization in (12) becomes equivalent to maximizing $[\hat{I}_{\boldsymbol{wy}}(W;Y)+D(\hat{P}_{\boldsymbol{y}}\|P)]$ subject to $\hat{E}_{\boldsymbol{xy}}d(X,Y) \leq D$, where $\hat{I}_{\boldsymbol{wy}}(W;Y)$ denotes the empirical mutual information induced from the joint empirical distribution $\hat{P}_{\boldsymbol{wy}}$ and $\hat{E}_{\boldsymbol{xy}}$ denotes the aforementioned expectation with respect to $\hat{P}_{\boldsymbol{xy}}$. Now, observe that for given $\boldsymbol{x}$ and $\boldsymbol{w}$, both $[\hat{I}_{\boldsymbol{wy}}(W;Y) + D(\hat{P}_{\boldsymbol{y}}\|P)]$ and $\hat{E}_{\boldsymbol{xy}}d(X,Y) \leq D$ depend on $\boldsymbol{y}$ only via its conditional type class given $(\boldsymbol{x},\boldsymbol{w})$, namely, the conditional empirical distribution $\hat{P}_{\boldsymbol{wxy}}(y|x,w)$. Once the optimal $\hat{P}_{\boldsymbol{wxy}}(y|x,w)$ has been found, it does not matter which vector $\boldsymbol{y}$ is chosen from the corresponding conditional type class $T(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$. Therefore, the optimization across n–vectors in (12) boils down to optimization over empirical conditional distributions, and since the total number of empirical conditional distrubtions of n–vectors increases only polynomially with n, the search complexity reduces from exponential to polynomial as well. In practice, one may not perform such an exhaustive search over the discrete set of empirical distributions, but apply an optimization procedure in the continuous space of conditional distributions $\{P(y|x,w)\}$ (and then approximate the solution by the closest feasible empirical distribution). At any rate, this optimization procedure is carried out in a space of fixed dimension, that does not grow with n.

3.2 Universality in the Covertext 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 then make sense to insist on the constraint (4) for *every* memoryless source, to be on the safe side. Equivalently, eq. (4) would be replaced by

$$\max_P \sum_{\boldsymbol{y}\in\Lambda} P(\boldsymbol{y}) \le e^{-\lambda n} \tag{14}$$

where the maximization over P is across all memoryless source with alphabet $\mathcal{A}$. It is then easy to see that our earlier derivation goes through as before except that $P(\boldsymbol{y})$ should be replaced by $\max_P P(\boldsymbol{y})$ in all places (see also [15]). Since $\ln \max_P P(\boldsymbol{y}) = -n\hat{H}_{\boldsymbol{y}}(Y)$, this means that the modified version of Λ_* compares the empirical mutual information $\hat{I}_{\boldsymbol{wy}}(W;Y)$ to the threshold $\lambda n - |\mathcal{A}|\ln(n+1)$. By the same token, and in light of the discussion in the previous paragraph, the modified version of the optimal embedder (12) maximizes $\hat{I}_{\boldsymbol{wy}}(W;Y)$ 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.

3.3 Other Detector Statistics

In the previous section, we focused on the class of detectors that base their decision on the empirical joint distribution of pairs of letters $\{(w,y)\}$. What about classes of detectors that base their decisions on larger (and more refined) sets of statistics? It turns out that such extensions are possible as long as we are able to assess the cardinality of the corresponding conditional type class. For example, suppose that the stegotext is suspected to undergo a desynchronization attack that cyclically shifts the data by k points, where k lies in some uncertainy region, say, $\{-K,-K+1,\ldots,-1,0,1,\ldots,K\}$. Then, it would make sense to allow the detector depend on the joint distribution of $2K+2$ vectors: $\boldsymbol{y}$, $\boldsymbol{w}$, and all the $2K$ corresponding cyclic shifts of $\boldsymbol{w}$. Our earlier analysis will carry over provided that the above definition of $\hat{H}_{\boldsymbol{wy}}(Y|W)$ would be replaced the conditional empirical entropy of $\boldsymbol{y}$ given $\boldsymbol{w}$ and all its cyclic shifts. This is different from the exhaustive search (ES) approach (see, e.g., [1]) to confront such desynchronization attacks. Note, however, that this works as long as K is fixed and does not grow with n. We will discuss the case where K grows with n later on in Section 4.

3.4 Continuous Alphabets

In the previous section, we considered, for convenience, the simple case where the components of both $\boldsymbol{x}$ and $\boldsymbol{y}$ take on values in a finite alphabet. It is more common and more natural, however, to model $\boldsymbol{x}$ and $\boldsymbol{y}$ 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 [16], used above, to vectors with real–valued components (see, e.g., [17],[18],[19]). In a nutshell, a conditional type class, in such a case, is the set of all $\boldsymbol{y}$–vectors in $\mathbb{R}^n$ whose joint statistics with $\boldsymbol{w}$ 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 $\boldsymbol{x}$ is a zero–mean Gaussian vector whose covariance matrix is $\sigma^2 I$, I being the $n \times n$ identity matrix an σ^2 is unknown (cf. Subsection 3.2). 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 $\sum_{i=1}^n w_i y_i$. These assumptions are the same as in many theoretical papers in the literature of watermark detection. Then, the conditional empirical entropy $\hat{H}_{\boldsymbol{wy}}(Y|W)$ should be replaced by the empirical differential entropy $\hat{h}_{\boldsymbol{wy}}(Y|W)$, given by [18]:

$$\begin{aligned}
\hat{h}_{\boldsymbol{wy}}(Y|W) &= \frac{1}{2}\ln\left[2\pi e \cdot \min_a \left(\frac{1}{n}\sum_{i=1}^n (y_i - aw_i)^2\right)\right] \\
&= \frac{1}{2}\ln\left[2\pi e\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - \frac{(\frac{1}{n}\sum_{i=1}^n w_i y_i)^2}{\frac{1}{n}\sum_{i=1}^n w_i^2}\right)\right] \\
&= \frac{1}{2}\ln\left[2\pi e\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - (\frac{1}{n}\sum_{i=1}^n w_i y_i)^2\right)\right]. \qquad (15)
\end{aligned}$$

Since

$$\hat{h}_{\boldsymbol{y}}(Y) = \frac{1}{2}\ln\left(2\pi e \cdot \frac{1}{n}\sum_{i=1}^n y_i^2\right) \qquad (16)$$

the optimal embedder would maximize

$$\hat{I}_{\boldsymbol{wy}}(W;Y) = -\frac{1}{2}\ln\left(1 - \frac{(\frac{1}{n}\sum_{i=1}^n w_i y_i)^2}{\frac{1}{n}\sum_{i=1}^n y_i^2}\right) \qquad (17)$$

or, equivalently,[1] maximize $(\sum_{i=1}^n w_i y_i)^2 / \sum_{i=1}^n y_i^2$ subject to the distortion constraint, which in

[1] Note also that the corresponding detector, which compares $\hat{I}_{\boldsymbol{wy}}(W;Y)$ to a threshold, is equivalent to a correlation detector, which compares the (absolute) correlation to a threshold that depends on the energy of $\boldsymbol{y}$, rather than a fixed threshold (see, e.g., [1]).

this case, will naturally be taken to be Euclidean, $\sum_{i=1}^{n}(x_i - y_i)^2 \leq nD$. While our discussion in Subsection 3.1, 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 $\boldsymbol{y} \in \mathbb{R}^n$ can be represented as $\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{w} + \boldsymbol{z}$, where a and b are real valued coefficients and $\boldsymbol{z}$ is orthogonal to both $\boldsymbol{x}$ and $\boldsymbol{w}$. Now, without loss of optimality, $\boldsymbol{z}$ should be taken to be the zero vector. This is because any non-zero $\boldsymbol{z}$ contributes to the energy of $\boldsymbol{y}$ (the denominator of $(\sum_{i=1}^{n} w_i y_i)^2 / \sum_{i=1}^{n} y_i^2$) while improving neither the correlation with $\boldsymbol{w}$ (which is the numerator), nor the distance to $\boldsymbol{x}$ (which is the constraint). Thus, the optimal embedding function should be of the form

$$f^*(\boldsymbol{x}, \boldsymbol{w}) = a\boldsymbol{x} + b\boldsymbol{w}, \tag{18}$$

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 (the details are omitted here).

Going back to the openning discussion in the Introduction, this seems to be very close to the linear embedder (1) that is so customarily used (with one additional degree of freedom allowing also scaling of $\boldsymbol{x}$). A closer look, however, reveals that this is not quite the case because the optimal values of a and b depend here on $\boldsymbol{x}$ and $\boldsymbol{w}$ (via the joint statistics $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} w_i x_i$) rather than being fixed. Therefore, this is *not* a linear embedder!

Finally, if we want our detector to take into account joint statistics of $\boldsymbol{y}$ with several (cyclic) shifts of $\boldsymbol{w}$, as discussed earlier, the corresponding empirical differential entropy (cf. the first line of eq. (15)) will be based on the norm of the projection error of $\boldsymbol{y}$ over the space spanned by $\boldsymbol{w}$ and its cyclic shifts, rather than space spanned by $\boldsymbol{w}$ alone, as in eq. (15). Of course, the derivation of the optimal embedder will be more involved.

3.5 Random Watermarks

Thus far, our model assumption was that $\boldsymbol{x}$ emerges from a probabilistic source P, whereas the watermark $\boldsymbol{w}$ is fixed, and hence can be thought of as being deterministic. Another possibile setting assumes that $\boldsymbol{w}$ is random as well, in particular, being drawn from another source Q, independently of $\boldsymbol{x}$, normally, the binary symmetric source (BSS). This situation may arise, for example, when security is an issue and then the watermark is encrypted. In such a case, the randomness of $\boldsymbol{w}$ is induced by the randomness of the key. In this case, the decision regions Λ and Λ^* will be defined as subsets of $\mathcal{A}^n \times \mathcal{B}^n$ and the probabilities of errors P_{e_1} and P_{e_2} will be defined, of course, as the corresponding summations of products $P(\boldsymbol{x})Q(\boldsymbol{w})$. Although this model is somewhat weaker, it can be analyzed for more general classes of detectors. This is because the role of the conditional type class $T(\boldsymbol{y}|\boldsymbol{w})$, would be replaced by the joint type class $T(\boldsymbol{w},\boldsymbol{y})$, namely, the set of all *pairs* of sequences $\{(\boldsymbol{w}',\boldsymbol{y}')\}$ that have the same empirical distribution as $(\boldsymbol{w},\boldsymbol{y})$ (as opposed to the conditional type class which is defined as the set of all such $\boldsymbol{y}$'s for a given $\boldsymbol{w}$). Thus, the corresponding version of Λ^* would be

$$\Lambda_* = \{(\boldsymbol{w},\boldsymbol{y}): \ \ln P(\boldsymbol{y}) + \ln Q(\boldsymbol{w}) + n\hat{H}_{\boldsymbol{w}\boldsymbol{y}}(W,Y) + \lambda n - |\mathcal{A}|\ln(n+1) \le 0\}, \tag{19}$$

where $\hat{H}_{\boldsymbol{w}\boldsymbol{y}}(W,Y)$ is the empirical joint entropy induced by $(\boldsymbol{w},\boldsymbol{y})$, and the derivation of the optimal embedder is accordingly.[2] The advantage of this model, albeit somewhat weaker, is that it is easier to assess $|T(\boldsymbol{w},\boldsymbol{y})|$ in more general situations than it is for $|T(\boldsymbol{y}|\boldsymbol{w})|$. For example, if $\boldsymbol{x}$ is a first order Markov source, rather than i.i.d., and one is then naturally interested in the statistics formed by the frequency counts of triples $\{w_i = w,\ y_i = y,\ y_{i-1} = y'\}$, then there is no known expression for the cardinality of the corresponding conditional type class, but it is still possible to assess the size of the joint type class in terms of the empirical first-order Markov entropy of the pairs $\{(w_i, y_i)\}$.

It should be also pointed out that once $\boldsymbol{w}$, is assumed random (say, drawn from a BSS), it is possible to devise a decision rule that it astymptotically optimum for an *individual* covertext sequence, i.e., to drop the assumption that $\boldsymbol{x}$ emerges from a probabilistic source of a known model. The resulting decision rule, obtained using a similar techique, accepts H_1 whenever $\hat{H}_{\boldsymbol{w}\boldsymbol{y}}(W|Y) \le 1-\lambda$, and the embedder minimizes $\hat{H}_{\boldsymbol{w}\boldsymbol{y}}(W|Y)$ subject ot the distortion constraint accordingly.

[2] Note that in the universal case (where both P and Q are unknown), this leads again to the same empirical mutual information detector as before.

4 Geometric Attacks

Let us now extend the setup to include attacks. We first discuss attacks in general and then confine our attention to geometric attacks.

The case of attack is characterized by the fact that the input to the detector is no longer the vector $\boldsymbol{y}$ as before, but another vector, $\boldsymbol{z} = (z_1, \ldots, z_n)$, that is the output of a channel fed by $\boldsymbol{y}$, which we shall denote by $W(\boldsymbol{z}|\boldsymbol{y})$. For convenience, we will assume that the components of $\boldsymbol{z}$ take on values in the same alphabet $\mathcal{A}$. Thus, the operation of the attack, which in general may be stochastic, is thought of as a channel. Denoting the channel output marginal $Q(\boldsymbol{z}) = \sum_{\boldsymbol{y}} P(\boldsymbol{y})W(\boldsymbol{z}|\boldsymbol{y})$, the analysis of this case is, in principle, the same as before. Assuming, for example, that Q is memoryless (which is the case when by P and W are memoryless), then Λ_* is as in Section 2, except that P, Y, and $\boldsymbol{y}$, should be replaced by Q, Z and $\boldsymbol{z}$, respectively. The optimal embedder then becomes

$$f^*(\boldsymbol{x}, \boldsymbol{w}) = \text{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y}) \leq nD\}} \sum_{\boldsymbol{z} \in \Lambda_*^c} W(\boldsymbol{z}|\boldsymbol{y}), \tag{20}$$

for the redefined version of Λ_*.

We now turn to the more 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 $\boldsymbol{z}$ as a (randomly chosen) permutation of $\boldsymbol{y}$ (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 $\{\pi_1, \ldots, \pi_M\}$ that the attack channel may apply, using the following mechanism: First, a random integer J is drawn, say, uniformly over $\{1, \ldots, M\}$, independently of $\boldsymbol{y}$, and then the attacker produces $\boldsymbol{z} = \pi_J(\boldsymbol{y})$, i.e., $\boldsymbol{z}$ is the result of the operation of the permutation π_J on the input $\boldsymbol{y}$. Thus,

$$W(\boldsymbol{z}|\boldsymbol{y}) = \frac{1}{M}\sum_{j=1}^{M} 1\{\boldsymbol{z} = \pi_j(\boldsymbol{y})\}. \tag{21}$$

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 Λ^* is asymptotically optimum. Let $\Lambda_*(j)$ denote the set of all $\boldsymbol{z}$ such that $\pi_j^{-1}(\boldsymbol{z}) \in \Lambda_*$ (π_j^{-1} always exists), where here Λ_* is again as in Section 2. 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 $\boldsymbol{y} = \pi_j^{-1}(\boldsymbol{z})$ and we are back to the case without attack, as in Section 2. We now show that the ES approach of applying Λ_* to all inverse permutations performs asymptotically as well as this genie–aided detector. In particular, let us define

$$\begin{aligned}
\Lambda_{ES} &= \bigcup_{j=1}^{M} \Lambda_*(j) \\
&= \left\{\boldsymbol{z} : \ln P(\boldsymbol{z}) + n \min_j \hat{H}_{\boldsymbol{w},\pi_j^{-1}(\boldsymbol{z})}(Y|W) + \lambda n - |\mathcal{A}| \ln(n+1) \le 0\right\} \qquad (22)
\end{aligned}$$

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

$$\begin{aligned}
P_{e_2} &= \sum_{\boldsymbol{z}\in\Lambda_{ES}} \sum_{\boldsymbol{y}} P(\boldsymbol{y}) W(\boldsymbol{z}|\boldsymbol{y}) \\
&= \sum_{\boldsymbol{z}\in\Lambda_{ES}} \frac{1}{M} \sum_{j=1}^{M} P(\pi_j^{-1}(\boldsymbol{z})) \\
&= \sum_{\boldsymbol{z}\in\Lambda_{ES}} P(\boldsymbol{z}) \\
&\le \sum_{j=1}^{M} \sum_{\boldsymbol{z}\in\Lambda_*(j)} P(\boldsymbol{z}) \\
&= \sum_{j=1}^{M} \sum_{\{\boldsymbol{z}:\ \pi_j^{-1}(\boldsymbol{z})\in\Lambda_*\}} P(\boldsymbol{z}) \\
&= \sum_{j=1}^{M} \sum_{\{\boldsymbol{z}:\ \pi_j^{-1}(\boldsymbol{z})\in\Lambda_*\}} P(\pi_j^{-1}(\boldsymbol{z})) \\
&= \sum_{j=1}^{M} \sum_{\boldsymbol{y}\in\Lambda_*} P(\boldsymbol{y}) \\
&= M \cdot \sum_{\boldsymbol{y}\in\Lambda_*} P(\boldsymbol{y}) \qquad (23)
\end{aligned}$$

which decays exponentially rapidly at the rate of $e^{-\lambda n}$ since $\sum_{\boldsymbol{y}\in\Lambda_*} P(\boldsymbol{y})$ decays at such a rate and M is assumed sub–exponential. Thus, Λ_{ES} satisfies the false positive constraint (4). As for the error probability of the first kind, note that $\Lambda_{ES}^c = \bigcap_{j=1}^{M} \Lambda_*^c(j)$ which means that $\Lambda_{ES}^c \subseteq \Lambda_*^c(j)$ for all $j = 1, \ldots, M$. Thus, P_{e_2} 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

$$\begin{aligned} f^*(\boldsymbol{x},\boldsymbol{w}) &= \text{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\le nD\}} \sum_{\boldsymbol{z}\in\Lambda_{ES}^c} W(\boldsymbol{z}|\boldsymbol{y}) \\ &= \text{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\le nD\}} \sum_{\boldsymbol{z}\in\Lambda_{ES}^c} \sum_{j=1}^{M} 1\{\boldsymbol{z}=\pi_j(\boldsymbol{y})\} \\ &= \text{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\le nD\}} \sum_{j=1}^{M} 1\{\pi_j(\boldsymbol{y})\in\Lambda_{ES}^c\}. \end{aligned} \tag{24}$$

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

$$\begin{aligned} f'(\boldsymbol{x},\boldsymbol{w}) &= \text{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\le nD\}}[\ln P(\boldsymbol{y}) + n\max_j\min_i \hat{H}_{\boldsymbol{w},\pi_i^{-1}(\pi_j(\boldsymbol{y}))}(Y|W)] \\ &= \text{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\le nD\}}[\ln P(\boldsymbol{y}) + n\max_j\min_i \hat{H}_{\pi_i(\boldsymbol{w}),\pi_j(\boldsymbol{y})}(Y|W)]. \end{aligned} \tag{25}$$

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

$$1\{\exists j:\ \pi_j(\boldsymbol{y})\in\Lambda_{ES}^c\} \le \sum_{j=1}^{M} 1\{\pi_j(\boldsymbol{y})\in\Lambda_{ES}^c\} \le M\cdot 1\{\exists j:\ \pi_j(\boldsymbol{y})\in\Lambda_{ES}^c\}. \tag{26}$$

Now, let us denote

$$P_{e_1}(f) = \sum_{\boldsymbol{x}} P(\boldsymbol{x})\cdot\frac{1}{M}\sum_{j=1}^{M} 1\{\pi_j(f(\boldsymbol{x},\boldsymbol{w}))\in\Lambda_{ES}^c\} \tag{27}$$

and

$$P'_{e_1}(f) = \sum_{\boldsymbol{x}} P(\boldsymbol{x})\cdot\frac{1}{M}\cdot 1\{\exists j:\ \pi_j(f(\boldsymbol{x},\boldsymbol{w}))\in\Lambda_{ES}^c\}. \tag{28}$$

Then, following eq. (26), we have $P_{e_1}(f)\le MP'_{e_1}(f)\le MP_{e_1}(f)$ for every f, and since f' minimizes $P'_{e_1}(f)$, we also have:

$$P_{e_1}(f')\le MP'_{e_1}(f')\le MP'_{e_1}(f^*)\le MP_{e_1}(f^*), \tag{29}$$

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 $\star$ be given by the rule $\pi_i(\pi_j(\cdot)) = \pi_{i\star j}(\cdot)$, and let the inverse of i, denoted i^{-1}, be induced by $\pi_i^{-1}(\cdot)$. Then, $\pi_i^{-1}(\pi_j(\cdot)) = \pi_{i^{-1}\star j}(\cdot)$, and so,

$$\begin{aligned}\max_j \min_i \hat{H}_{\boldsymbol{w},\pi_i^{-1}(\pi_j(\boldsymbol{y}))}(Y|W) &= \max_j \min_i \hat{H}_{\boldsymbol{w},\pi_{i^{-1}\star j}(\boldsymbol{y})}(Y|W) \\ &= \max_j \min_k \hat{H}_{\boldsymbol{w},\pi_k(\boldsymbol{y})}(Y|W) \\ &= \min_k \hat{H}_{\boldsymbol{w},\pi_k(\boldsymbol{y})}(Y|W). \qquad (30)\end{aligned}$$

Now, in the double minimization $\min_{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y})\leq nD} \min_k [\ln P(\boldsymbol{y}) + n\hat{H}_{\boldsymbol{w},\pi_k(\boldsymbol{y})}(Y|W)]$, implemented by the embedding function f', the order of the minimizations can be interchanged, and then, the minimization over $\boldsymbol{y}$ can be carried out first. For every given k, the complexity of this minimization is as described in Subsections 3.1 and 3.4. The overall complexity will then be proportional to M (due to the additional minimization over k), and hence subexponential in n.

Similar techniques can be applied to the case of informed detection, which is the case where the detector has access to $\boldsymbol{x}$ in addition to $\boldsymbol{z}$ and $\boldsymbol{w}$. The performance will, of course, improve, in general.

References

[1] M. Barni, "Effectiveness of exhaustive search and template matching against watermark desynchronization," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 158–161, February 2005.

[2] R. Eslami, J. R. Deller, Jr., and H. Radha, "On the detection of multiplicative watermarks in the wavelet and DCT domains," submitted to *ISIT 2005*.

[3] J. Lichtenauer, I. Setyawan, T. Kalker, and R. Lagendijk, "Exhaustive geoemtrical search and the false positive watermark detection probability," *Proc. SPIE, Security and Watermarking of Multimedia Contents V*, vol. 5020, pp. 203–214, January 2003.

[4] V. Licks, F. Ourique, R. Jordan, and F. Perez–Gonzales, "The effect of random jitter attack on the bit error rate performance of spatial domain image watermarking," *2003 IEEE Proc. International Conference on Image Processing*, vol. 2, pp. II-455–448, September 2003.

[5] P. Moulin and A. Ivanovic, "The Fisher information game for optimal design of synchronization patterns in blind watermarking," *Proc. 2001 IEEE International Conference on Image Processing*, vol. 2, pp. 550–553, October 2001.

[6] P. Moulin, A. Briassouli, and H. Malvar, "Detection–theoretic analysis of desynchronization attacks in watermarking," *Proc. 14th International Conference on Digital Signal Processing*, vol. 1, pp. 77–84, July 2002.

[7] A. Briassouli and P. Moulin, "Detection–theoretic analysis of warping attacks in spread–spectrum watermarking," *Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. III-53–56, April 2003.

[8] S. Baudry, P. Nguyen, and H. Maître, "Some theoretical bounds on the capacity of watermarking channels with geometrical distortions," *2001 IEEE Proc. International Conference on Image Processing*, vol. 3, pp. 995–998, October 2001.

[9] V. Licks and R. Jordan, "Geometric attacks on image watermarking systems: a survey," available on-line at: [http://www.eece.unm.edu/∼vlicks/docs/survey.pdf], submitted, 2003.

[10] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, "A new decoder for the optimum recovery of nonadditive watermarks," *IEEE Trans. on Image Processing*, vol. 10, pp. 755–766, May 2001.

[11] P. Amin and K. P. Subbalakshmi, "Rotation and cropping resilient data hiding with Zernike moments," Proc. ICIP, 2004.

[12] A. Briassouli and M. G. Strintzis, "Locally optimum nonlinearities for DCT watermark detection," *IEEE Trans. on Image Processing*, vol. 13, no. 12, pp. 1604–1617, December 2004 (also, available on-line: [http://vision.ai.uiuc.edu/∼briassou/zmnl_qn_final.pdf]).

[13] V. Licks, F. Ourique, R. Jordan, and G. Heileman, "Performance loss of dirty–paper codes in additive white Gaussian noise and jitter channels," *Proc. IEEE Workshop on Statistical Signal Processing*, St. Louis, U.S.A., 2003.

[14] Y. Xin, S. Liao, and M. Pawlak, "Geometrically robust image watermarking on a circular domain," available on–line at: [http://www.ee.umanitoba.ca/∼pawlak/papers/Imaging/Watermarking.pdf].

[15] N. Merhav, "Universal detection of messages via finite–state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242–2246, September 2000.

[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press 1981.

[17] N. Merhav, "On the estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 5, pp. 1109–1114, September 1989.

[18] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1261–1269, July 1993.

[19] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1953-1967, November 1994.