CCIT Report #564
December 2005

Optimal Watermark Embedding and Detection Strategies under Limited Detection Resources *

Neri Merhav and Erez Sabbag

December 7, 2005

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

Abstract

We propose an information–theoretic approach to the watermark embedding and detection under limited detector resources. First, we present asymptotically optimal decision regions in the Neyman–Pearson sense. We expand these results to the case of zero-mean i.i.d. Gaussian covertext distribution with unknown variance. For this case, we propose a lower bound on the exponential decay rate of the false–negative probability and prove that the optimal embedding and detecting strategy is superior to the customary linear, additive embedding strategy in the exponential sense.

1 Introduction

Information embedding and watermarking have become a very active field of research in the last decade, both in the academic community and in the industry, due to the need of protecting the vast amount of digital information available over the Internet and other data storage media and devices (see, e.g.,[1]–[5], and references therein). Watermarking (WM) is a form of embedding information secretly in a host data set (e.g., image, audio signal, video, etc.). In this work, we raise and examine certain fundamental questions with regard to customary methods of embedding and detection and suggest some new ideas for the most basic setup.

The most popular approach to watermark embedding and detection has been the following (see, e.g., [2],[6],[4, sec. 4.2] and references therein): 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,

*This research was supported by the Israel Science Foundation (grant no. 223/05).

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

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

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

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

where in both cases, the choice of γ 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: $\boldsymbol{y} = \boldsymbol{x} + \gamma \boldsymbol{w}$) 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 $\sigma_x^2(1+\gamma)^2$, and the latter being $\sigma_x^2(1-\gamma)^2$, 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 described detectors are no longer optimal in general.

The problem of finding the optimum watermark embedder f, for reliable WM detection, 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. 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.

While many papers in the literature addressed the problem of computing the performance of different embedding and detection strategies and plotting their receiver operating characteristics (ROC) for different values of the problem dimension n (see, e.g.,[7]–[9] and references therein), to the best of our knowledge, no reported work deals with the asymptotic behavior of the two kinds of error probabilities, i.e., the exponential decay rate of the two kind of the error probabilities as n tends to infinity.

Within the class of detectors based on a given set of statistics, we present the optimal (in the Neyman–Pearson sense) embedder and its corresponding detector. In doing so, we will extend the techniques, presented in [10] and references therein, to devising the optimal embedder. For the sake of simplicity, we will analyze the performance of the attack free scenario. Nevertheless, this analysis can easily be extended to certain classes of attacks (e.g., attacks that can be represented as a memoryless channel) as will be explained later.

The remainder of the paper is organized as follows: In the next section, the problem is formulated and the main results are presented. In Section 3, we discuss some aspects and different scenarios of the problem. In Section 4, we address the Gaussian case where we present the optimal embedder and suggest a lower-bound on the false-negative error exponent. In addition, we show that the optimum embedder is superior to the linear embedder, by analyzing their error exponents.

2 Basic Derivation

We begin with some notations and definitions. Throughout this work, capital letters represent scalar random variables (RVs), and specific realizations of them are denoted by the corresponding lowercase letters. Random vectors of dimension n will be denoted by bold-face letters. The notation $\mathbb{1}\{A\}$, where A is an event, will designate the indicator function of A (i.e.,$\mathbb{1}\{A\} = 1$ if A occurs and $\mathbb{1}\{A\} = 0$ otherwise). The notion $a_n \doteq b_n$ for two positive sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ expresses asymptotic equality in the logarithmic scale, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{a_n}{b_n}\right) = 0.$$

For two vectors, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, the Euclidean inner product is defined as $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{n} a_i \cdot b_i$ and the L_2-norm of a vector is defined as $\|\boldsymbol{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle}$.

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 $\mathcal{A}$ 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

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

subject to the following constraints:

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

(2) The false positive probability is upper bounded by

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

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

In other words, we would like to choose f and Λ so as to minimize P_{fn} subject to a distortion constraint and the constraint that the exponential decay rate of P_{fp} 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 $\boldsymbol{w}$:

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

where

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

i.e, $\hat{P}_{\boldsymbol{wy}}(w,y)$ 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 [11], 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

$\hat{P}_{\boldsymbol{w}\boldsymbol{y}'} = \hat{P}_{\boldsymbol{w}\boldsymbol{y}}$, 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 $\hat{P}_{\boldsymbol{w}\boldsymbol{y}}$ 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

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

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 [11], defined as:

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

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

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

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_{fn} 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 [10, 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 $\hat{P}_{\boldsymbol{w}\boldsymbol{y}}$, 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.,

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

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

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

Consider now the following embedder:

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

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_{fn} 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 $\{\boldsymbol{y} : \ d(\boldsymbol{x}, \boldsymbol{y}) \le nD\}$, 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,

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

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 (26) becomes equivalent to maximizing $[\hat{I}_{\boldsymbol{wy}}(W;Y) + D(\hat{P}_{\boldsymbol{y}} \| P)]$ subject to $\hat{E}_{\boldsymbol{xy}} d(X,Y) \le D$, 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 $\hat{E}_{\boldsymbol{xy}} d(X,Y) \le D$ 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 (26) 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. 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

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

where the maximization over P is across all memoryless sources 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 [10]). 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 $\lambda n - |\mathcal{A}|\ln(n+1)$ (the divergence term now disappears). By the same token, and in light of the discussion in the previous paragraph, the modified version of the optimal embedder (26) maximizes $\hat{I}_{\boldsymbol{wy}}(W;Y)$ subject to the distortion constraint. Both the embedding rule and the detection rule are then based on the idea of *maximum mutual information*, which is intuitively appealing. For more on this idea and its use as a universal decoding rule see [11, Sec. 2.5].

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 uncertainty 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., [12]) to confront such desynchronization attacks. Note, however, that this works as long as K is fixed and does not grow with n.

3.4 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 possible 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_{fn} and P_{fp} 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

$$\Lambda_* = \Big\{(\boldsymbol{w}, \boldsymbol{y}): \ \ln P(\boldsymbol{y}) + \ln Q(\boldsymbol{w}) + n\hat{H}_{\boldsymbol{wy}}(W, Y) + \lambda n - |\mathcal{A}| \ln(n+1) \le 0\Big\}, \qquad (15)$$

where $\hat{H}_{\boldsymbol{wy}}(W, Y)$ is the empirical joint entropy induced by $(\boldsymbol{w}, \boldsymbol{y})$, and the derivation of the optimal embedder is accordingly.[1] 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

[1] 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.

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 asymptotically 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 technique, accepts H_1 whenever $\hat{H}_{\boldsymbol{wy}}(W|Y) \leq 1-\lambda$, and the embedder minimizes $\hat{H}_{\boldsymbol{wy}}(W|Y)$ subject to the distortion constraint accordingly.

3.5 Attacks

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

$$f^*(\boldsymbol{x}, \boldsymbol{w}) = \mathrm{argmin}_{\{\boldsymbol{y}:\ d(\boldsymbol{x},\boldsymbol{y}) \leq nD\}} \sum_{\boldsymbol{z} \in \Lambda_*^c} W(\boldsymbol{z}|\boldsymbol{y}), \tag{16}$$

for the redefined version of Λ_*^c which is given by:

$$\begin{aligned} \Lambda_*^c &= \Big\{\boldsymbol{z} : \ln Q(\boldsymbol{z}) + n\hat{H}_{\boldsymbol{zw}}(Z|W) + n\lambda - |\mathcal{A}| \ln(n+1) > 0\Big\} &(17)\\ &= \Big\{\boldsymbol{z} : -n\hat{I}_{\boldsymbol{zw}}(Z;W) - nD\big(\hat{P}_{\boldsymbol{z}} \| Q\big) + n\lambda - |\mathcal{A}| \ln(n+1) > 0\Big\}\ , &(18)\end{aligned}$$

where $\hat{P}_{\boldsymbol{z}}$ is the empirical distribution of $\boldsymbol{z}$. Evidently, this is not a convenient formula to work with. Therefore, let us try to simplify (16). For a given $\boldsymbol{y}$ let us rewrite (16) as follows:

$$\begin{aligned} \sum_{\boldsymbol{z} \in \Lambda_*^c} W(\boldsymbol{z}|\boldsymbol{y}) &= \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w}) \subseteq \Lambda_*^c} \ \sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w})} W(\boldsymbol{z}'|\boldsymbol{y}) \\ &= \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w}) \subseteq \Lambda_*^c} \big|T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w})\big| W(\boldsymbol{z}|\boldsymbol{y}) &(19)\end{aligned}$$

It is easy to show that for a given $\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w})$ and a memoryless channel $W(\boldsymbol{z}|\boldsymbol{y})$ the probability of $\boldsymbol{z}'$ given $\boldsymbol{y}$ is given by the following expression:

$$W(\boldsymbol{z}'|\boldsymbol{y}) = e^{-n\left[\hat{H}_{\boldsymbol{zy}}(Z|Y)+\sum_{a\in\mathcal{A}}\hat{P}_{\boldsymbol{y}}(a)D\left(\hat{P}_{\boldsymbol{zy}}(Z|Y=a)\|W(Z|Y=a)\right)\right]} \tag{20}$$

Using the fact that the cardinality of $T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w})$ is given by

$$|T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w})| \doteq e^{n\hat{H}_{\boldsymbol{zyw}}(Z|Y,W)}, \tag{21}$$

we conclude that $f^*(\boldsymbol{x},\boldsymbol{w}) \in T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$, where $T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$ corresponds to the following conditional empirical distribution:

$$\hat{P}^*_{\boldsymbol{yxw}}(Y|X,W) = \arg \max_{\substack{\hat{P}_{\boldsymbol{yxw}}(Y|X,W):\\ \hat{E}_{\boldsymbol{xy}}d(X,Y)\le D}} \left\{ \min_{\substack{\hat{P}_{\boldsymbol{zyw}}(Z|Y,W):\\ \hat{I}_{\boldsymbol{zw}}(Z;W)+D(\hat{P}_{\boldsymbol{z}}\|Q)\le\lambda}} \left[\hat{I}_{\boldsymbol{zwy}}(Z;W|Y) + \sum_{a\in\mathcal{A}}\hat{P}_{\boldsymbol{y}}(a)D\big(\hat{P}_{\boldsymbol{zy}}(Z|Y=a)\big\|W(Z|Y=a)\big)\right]\right\} \tag{22}$$

i.e., for a given $\boldsymbol{w}$ and $\boldsymbol{x}$ we search for the empirical distribution $\hat{P}_{\boldsymbol{yxw}}(Y|X,W)$ which maximize the exponent of the false negative probability dictated by the dominating conditional type $T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w})$ in Λ_*^c. Once the optimal empirical distribution $\hat{P}^*_{\boldsymbol{yxw}}(Y|X,W)$ has been found, it does not matter which vector $\boldsymbol{y}$ is chosen the corresponding conditional type $T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$.

4 Continuous Alphabets – the Gaussian Case

In the previous sections, 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 [11], used above, to vectors with real–valued components (see, e.g., [13],[14],[15]). 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 that $\boldsymbol{x}$ is a zero–mean Gaussian vector whose covariance matrix is $\sigma^2 I$, I being the $n \times n$ identity matrix, and 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 [14]:

$$\begin{aligned}\hat{h}_{\boldsymbol{wy}}(Y|W) &= \frac{1}{2}\ln\left[2\pi e\cdot\min_{a}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-aw_i)^2\right)\right]\\ &= \frac{1}{2}\ln\left[2\pi e\left(\frac{1}{n}\sum_{i=1}^{n}y_i^2-\frac{(\frac{1}{n}\sum_{i=1}^{n}w_iy_i)^2}{\frac{1}{n}\sum_{i=1}^{n}w_i^2}\right)\right]\\ &= \frac{1}{2}\ln\left[2\pi e\left(\frac{1}{n}\sum_{i=1}^{n}y_i^2-(\frac{1}{n}\sum_{i=1}^{n}w_iy_i)^2\right)\right]. \end{aligned} \tag{23}$$

The justification of eq. (23) is as follows: as was done in the proof of Lemma 3 in [14], we define an auxiliary channel $\boldsymbol{y} = a\boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{z} \sim \mathcal{N}\big(0, \sigma_z^2 I\big)$ and σ_z is unknown. Then, we calculate an upper and lower bounds on $T_\epsilon(\boldsymbol{y}|\boldsymbol{w}) = \big\{\tilde{\boldsymbol{y}} \in \mathbb{R}^n : |\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n}\tilde{y}_i^2| \le n\epsilon, |\sum_{i=1}^{n} y_i w_i - \sum_{i=1}^{n}\tilde{y}_i w_i| \le n\epsilon\big\}$. The value of $\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\ln \mathrm{Vol}\big\{T_\epsilon(\boldsymbol{y}|\boldsymbol{w})\big\}$ equals to (23). Since [2]

$$\hat{h}_{\boldsymbol{y}}(Y) = \frac{1}{2}\ln\left(2\pi e\cdot\frac{1}{n}\sum_{i=1}^{n}y_i^2\right)\ , \tag{24}$$

the optimal embedder maximizes

$$\hat{I}_{\boldsymbol{wy}}(W;Y) = -\frac{1}{2}\ln\left(1-\frac{(\frac{1}{n}\sum_{i=1}^{n}w_iy_i)^2}{\frac{1}{n}\sum_{i=1}^{n}y_i^2}\right)\ , \tag{25}$$

or, equivalently, [3] maximizes $(\sum_{i=1}^{n} w_i y_i)^2 / \sum_{i=1}^{n} y_i^2$ subject to the distortion constraint, which in this case, will naturally be taken to be Euclidean, $\sum_{i=1}^{n}(x_i - y_i)^2 \le nD$. While our discussion in Subsection 3.1, regarding optimization over conditional distributions, does not apply directly to 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. To see this, note that 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

[2] It is easy to show that $\hat{h}_{\boldsymbol{y}}(Y) = \lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\mathrm{Vol}\{T_\epsilon(\boldsymbol{y})\}$ where $T_\epsilon(\boldsymbol{y}) = \big\{\tilde{\boldsymbol{y}} \in \mathbb{R}^n : |\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n}\tilde{y}_i^2| \le n\epsilon\big\}$, and $\mathrm{Vol}\{\cdot\}$ means the volume of a set in $\mathbb{R}^n$.

[3] 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., [12]).

$\boldsymbol{w}$ (which is the numerator), nor the distance to $\boldsymbol{x}$ (which is the constraint). Thus, the optimal embedding function should be of the form

$$f^*(\boldsymbol{x}, \boldsymbol{w}) = a\boldsymbol{x} + b\boldsymbol{w}, \tag{26}$$

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 are in Subsection 4.1).

Going back to the opening discussion in the Introduction, at first glance, 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.

We note that this embedding and detection strategy is also optimal in the case of Gaussian memoryless attack of the form $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{N}$ where $N \sim \mathcal{N}\big(0, \sigma_N^2 I_{n\times n}\big)$ and σ_N is unknown. The detector should maximize the normalized-correlation between $\boldsymbol{Z}$ and $\boldsymbol{W}$ the embedder should employ the embedding rule which will be presented in the following section.

4.1 Explicit Derivation of the Optimal Embedder

In this section, we present a closed-form expression for the optimal embedder. As was shown in the previous section, the following optimization problem should be solved:

$$\begin{aligned} &\max \left[\frac{\left(\frac{1}{n}\sum_{i=1}^n y_i w_i\right)^2}{\frac{1}{n}\sum_{i=1}^n y_i^2} \right]. \\ \text{subject to:} \quad &\sum_{i=1}^n (y_i - x_i)^2 \le nD \end{aligned} \tag{27}$$

Substituting $\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{w}$ in eq. (27), gives:

$$\begin{aligned} &\max_{a,b\in\mathbb{R}} \left[\frac{a^2\rho^2 + 2ab\rho + b^2}{a^2\alpha^2 + 2ab\rho + b^2} \right] \\ \text{subject to:} \quad &(a-1)^2\alpha^2 + 2(a-1)b\rho + b^2 \le D \end{aligned} \tag{28}$$

where $\alpha^2 \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^n x_i^2$ and $\rho \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^n x_i w_i$. We note that $\alpha^2 \ge \rho^2$ which stems from the Cauchy-Schwartz inequality.

Theorem 1. *The optimal values of* (a, b) *are:*

- *If* $D \geq \alpha^2 - \rho^2$*:*

$$(a^*, b^*) = (0, \rho + \sqrt{\rho^2 - \alpha^2 + D}) \tag{29}$$

- *If* $D < \alpha^2 - \rho^2$*:*

$$\begin{aligned} a^* &= \arg\max \left\{ t(a) \;\middle|\; a \in \{a_1, a_2, a_3, a_4\} \bigcap R \right\} \\ b^* &= a^* \cdot t(a^*) \end{aligned} \tag{30}$$

where $t(a) = \frac{(1-a)\rho + sgn(\rho)\sqrt{D-(a-1)^2(\alpha^2-\rho^2)}}{a}$, $R = \left[1 - \sqrt{\frac{D}{\alpha^2-\rho^2}}, 1 + \sqrt{\frac{D}{\alpha^2-\rho^2}}\right]$,
$a_{1,2} = \frac{(\alpha^2-\rho^2)(\alpha^2-D) \pm \sqrt{D\rho^2}\sqrt{(\alpha^2-\rho^2)(\alpha^2-D)}}{\alpha^2(\alpha^2-\rho^2)}$ *and* $a_{3,4} = 1 \pm \sqrt{\frac{D}{\alpha^2-\rho^2}}$.

The proof is purely technical and therefore is deferred to the Appendix. We note that in the case where $D \ll \alpha^2 - \rho^2$, the value of a^* tends to 1, and the value of b^* tends to $\mathrm{sgn}(\rho)\sqrt{D}$. Hence, the linear embedder is not optimal even in the case where $D \ll \alpha^2$. We will next use the above values to devise a lower bound on the exponential decay rate of the false-negative probability of the optimal embedder, and then compare it to an upper bound on the false negative exponent of the linear embedder.

4.2 Lower Bound to the False Negative Error Exponent of the Optimal Embedder

We derive the lower-bound on the exponent of the false-negative probability of the optimum embedder by exploring the performance of a sub-optimal embedder of the form $\boldsymbol{y} = \boldsymbol{x} + \mathrm{sgn}(\rho)\sqrt{D}\boldsymbol{w}$, which we name the *sign embedder*. This embedder is obtained by setting $a = 1$ in (26), where this value is in the allowable range R of a. We assume that $\boldsymbol{X} \sim \mathcal{N}\big(0, \sigma^2 I\big)$. First, we calculate a threshold value T which always guarantees a false-positive exponent not smaller than λ. Using the proposed detector (25), the false-positive probability can be expressed as

$$\begin{aligned} P_{fp} = \Pr\left\{ \hat{I}_{\boldsymbol{wy}}(W;Y) > T \;\middle|\; H_0 \right\} &= \Pr\left\{ \hat{\rho}^2_{\boldsymbol{wy}} > 1 - e^{-2T} \;\middle|\; H_0 \right\} \\ &= 2\Pr\left\{ \hat{\rho}_{\boldsymbol{wy}} > \sqrt{1 - e^{-2T}} \;\middle|\; H_0 \right\} \end{aligned}$$

where $\hat{\rho}_{\boldsymbol{wy}} = \frac{\langle \boldsymbol{w}, \boldsymbol{y} \rangle}{\|\boldsymbol{w}\| \cdot \|\boldsymbol{y}\|}$ is the normalized correlation between $\boldsymbol{w}$ and $\boldsymbol{y}$. Because $\boldsymbol{Y} = \boldsymbol{X}$ under H_0 and because of the radial symmetry of the PDF of $\boldsymbol{X}$, we can conclude that for large n [16, pp. 295]:

$$P_{fp} = 2\frac{A_n(\theta)}{A_n(\pi)} \doteq e^{n \ln(\sin\theta)} \ ,$$

where $A_n(\theta)$ [4] is the surface area of the n-dimensional spherical cap cut from a unit spare about the origin by a right circular cone of half angle $\theta = \arccos\left(\sqrt{1-e^{-2T}}\right)$ $(0 < \theta \leq \pi/2)$. Since we required that $P_{fp} \leq e^{-n\lambda}$, then $\ln(\sin\theta)$ must not exceed $-\lambda$, which means that

$$-\lambda \geq \ln(\sin\theta) \tag{31}$$

$$T \geq -\frac{1}{2}\ln\left[1-\cos^2\left(\arcsin(e^{-\lambda})\right)\right] = \lambda \tag{32}$$

where the last equality was obtained using the fact that $\cos\left(\arcsin(x)\right) = \sqrt{1-x^2}$. Hence, setting $T = \lambda$ ensures a false positive probability not greater than $e^{-n\lambda}$ for large n. Note that for every $\lambda > 0$, T is non-negative and does not depend on a specific embedder, since, under H_0, $\boldsymbol{Y}$ does not contain any watermark. Define the false-negative exponent of the sign-embedder

$$E_{fn}^{se} \stackrel{\triangle}{=} \lim_{n\to\infty} -\frac{1}{n}\ln P_{fn} \tag{33}$$

where the false-negative probability is given by

$$P_{fn} = \Pr\left\{\hat{I}_{\boldsymbol{wy}}(W;Y) \leq \lambda \mid H_1\right\} = \Pr\left\{\hat{\rho}^2_{\boldsymbol{wy}} \leq 1-e^{-2\lambda} \mid H_1\right\}. \tag{34}$$

Theorem 2. *The false-negative exponent of the sign-embedder is given by*

$$E_{fn}^{se}(\lambda, D) = \begin{cases} 0 & , \quad \frac{De^{-2\lambda}}{1-e^{-2\lambda}} \leq \sigma^2 \\ \frac{1}{2}\left[\frac{De^{-2\lambda}}{\sigma^2(1-e^{-2\lambda})} - \ln\left(\frac{De^{-2\lambda}}{\sigma^2(1-e^{-2\lambda})}\right) - 1\right] & , \quad else \end{cases} \tag{35}$$

The proof, which is mainly technical, is deferred to the Appendix. Let us explore some of the properties of $E_{fn}^{se}(\lambda, D)$. First, is is clear that $E_{fn}^{se}(0, D) = \infty$ (the detector output is constantly H_1) since $E_{fn}^{se}(\lambda, D)$ is monotonically increasing in $\frac{e^{-2\lambda}}{1-e^{-2\lambda}}$. In addition, $E_{fn}^{se}(\lambda, 0) = 0$ ($\boldsymbol{y} = \boldsymbol{x}$ and therefore does not contain any information on $\boldsymbol{w}$). For a given D, $E_{fn}^{se}(\lambda, D) = 0$ for $\lambda \geq \frac{1}{2}\ln\left(1+\frac{D}{\sigma^2}\right)$.

The exact value of the optimal exponent achieved when the optimal embedder is employed is too involved to calculate. However, we can use some of the properties of the optimal embedder to improve the lower bound on the optimal exponent. According to Theorem 1, in the case where $D \geq \alpha^2 - \rho^2$, the optimal embedder can completely "erase" the covertext and therefore achieves a zero false negative probability. We use this property to improve the performance of the sign embedder. This leads to the following embedding rule: $y = a\boldsymbol{x} + b\boldsymbol{w}$ where

$$(a, b) = \begin{cases} (0, \rho + \sqrt{\rho^2 - \alpha^2 + D}) & , \quad D \geq \alpha^2 \\ (1, \mathrm{sgn}(\rho)\sqrt{D}) & , \quad else \end{cases} \tag{36}$$

This embedder, which is an improved version of the sign embedder, erases the covertext in the cases where $D \geq \alpha^2$. Its performance are presented in the following Corollary:

[4] It is well-known [16, pp. 293] that $A_n(\theta) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)}\int_0^\theta \sin^{(n-2)}(\varphi)d\varphi$ and $A_n(\pi) = 2A_n(\pi/2)$.

Corollary 1. *For $\lambda > \frac{1}{2}\ln 2$, the false negative exponent of the improved sign embedder is given by:*

$$E(\lambda, D) = \begin{cases} 0 & , \quad D \leq \sigma^2 \\ \frac{1}{2}\left[D - \ln(D) - 1\right] & , \quad else \end{cases} \tag{37}$$

otherwise, the false-negative exponent equals to $E_{fn}^{se}(\lambda, D)$.

The proof is deferred to the Appendix. The fact that the optimal embedder can offer a positive false-negative exponent for every value of λ is not surprising due to its ability to erase the covertext, which leads to zero probability of false-negative. Although the improved sign embedder can offer a tighter lower bound, the improvement is made only in the case where $D \geq \sigma^2$ (though it is not known a priori to the embedder). Nevertheless, it emphasizes the true potential of the optimal embedder and the fact that the sign embedder is truly inferior to the optimal embedder. In Figure 1, the false negative exponent of the sign embedder and the false negative exponent of the improved embedder are plotted as functions of λ for a given values of D and σ. The point where the two graphs break apart is $\lambda = 1/2\ln(2)$. From this point on, the improved sign embedder achieves a fixed value of $0.5(D - \ln(D) - 1)$.

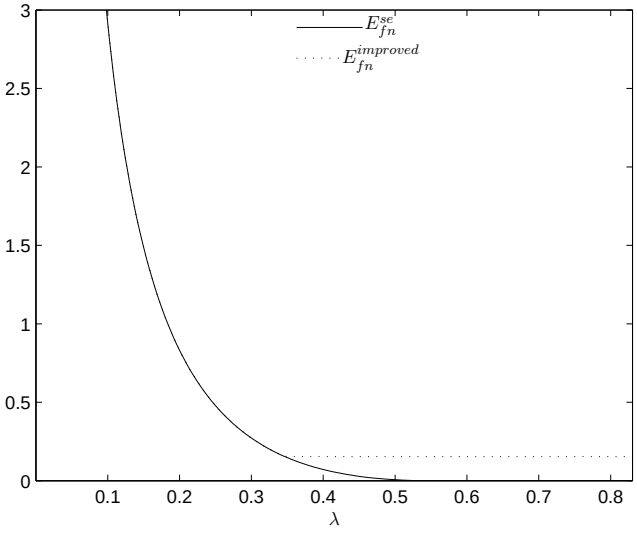

Figure 1: Error exponents of the sign-embedder and its improved version for $\sigma^2 = 1$ and $D = 2$.

4.3 Comparison to the Additive Embedder

Our next goal is to calculate the exponent of the false-negative probability of the linear additive embedder $\boldsymbol{y} = \boldsymbol{x} + \sqrt{D}\boldsymbol{w}$, where a normalized correlation detector is employed. Again, we first

calculate a threshold value used by the detector which ensures a false-positive probability not greater than $e^{-n\lambda}$. The false positive probability is given by

$$P_{fp} = \Pr\left\{\hat{\rho}_{\boldsymbol{wy}} > T \middle| H_0\right\} = \Pr\left\{\frac{\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{\|\boldsymbol{w}\| \cdot \|\boldsymbol{x}\|} > T\right\} = \frac{A_n(\theta)}{A_n(\pi)} \doteq e^{n\ln(\sin\theta)}\,, \tag{38}$$

where $\theta = \arccos(T)$ $(0 < \theta \leq \pi/2)$. The second equality is due to the fact that $\boldsymbol{Y} = \boldsymbol{X}$ under H_0 and the third equality is due to the radial symmetry of the PDF of $\boldsymbol{X}$. Then, $\ln(\sin\theta) \leq -\lambda$ implies:

$$T \;\geq\; \cos\left[\arcsin\left(e^{-\lambda}\right)\right] = \sqrt{1 - e^{-2\lambda}} \tag{39}$$

and therefore, by letting $T = \sqrt{1 - e^{-2\lambda}}$ ensures a false-positive probability not greater than $e^{-n\lambda}$. Note that $\lambda \geq 0$ implies that T must be non-negative. Define the false-negative exponent of the additive-embedder

$$E_{fn}^{ae} \stackrel{\triangle}{=} \lim_{n\to\infty} -\frac{1}{n}\ln P_{fn}, \tag{40}$$

where the false-negative probability is given by

$$P_{fn} = \Pr\left\{\hat{\rho}_{\boldsymbol{wy}} \leq \sqrt{1 - e^{-2\lambda}} \middle| H_1\right\}. \tag{41}$$

Theorem 3. *The false negative exponent of the additive embedder is given by*

$$E_{fn}^{ae}(\lambda, D) = \min\left\{E_1(\lambda, D), E_2(\lambda, D)\right\} \tag{42}$$

where,

$$\begin{aligned} E_1(\lambda, D) &= \min_{De^{-2\lambda} < r \leq \frac{De^{-2\lambda}}{1-e^{-2\lambda}}} \frac{1}{2}\left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 2\ln\sin\left(\Psi_1(r)\right) - 1\right] \\ E_2(\lambda, D) &= \begin{cases} 0 & , \quad \frac{De^{-2\lambda}}{1-e^{-2\lambda}} \leq \sigma^2 \\ \frac{1}{2}\left[\frac{De^{-2\lambda}}{(1-e^{-2\lambda})\sigma^2} - \ln\left(\frac{De^{-2\lambda}}{(1-e^{-2\lambda})\sigma^2}\right) - 1\right] & , \quad \textit{else} \end{cases} \end{aligned} \tag{43}$$

and $E_{fn}^{ae}(\lambda, D) < E_{fn}^{se}(\lambda, D)$ *for* $\frac{De^{-2\lambda}}{1-e^{-2\lambda}} > \sigma^2$.

Let us examine some of the properties of $E_{fn}^{ae}(\lambda, D)$. It is easy to see that $E_{fn}^{ae}(\lambda, D) \leq E_2(\lambda, D) = E_{fn}^{se}(\lambda, D)$, i.e., the upper bound on the additive-embedder exponent serves as a lower bound on the optimal-embedder exponent. It is clear that $E_{fn}^{ae}(\lambda, 0) = 0$ since $E_{fn}^{ae}(\lambda, 0) \leq E_{fn}^{se}(\lambda, 0) = 0$. In contrast to the sign-embedder, it turns out that $E_{fn}^{ae}(0, D) < \infty$. To see why this is the case let us look at

$$E_1(0, D) \;=\; \min_{r > D} f(r) \tag{44}$$

where $f(r) = \frac{1}{2}\left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 2\ln\sin\left(\Psi_1(r)\right) - 1\right]$. Now, since $f(r)$ is finite for $r > D$ the minimum value of $f(r)$ must be finite too. This is the case where the threshold value equals to zero and the probability that there is an embedded vector $\boldsymbol{Y}$ with negative correlation to $\boldsymbol{w}$ is not zero. Clearly, for a given D, $E_{fn}^{ae}(\lambda, D) = 0$ for $\lambda \geq \frac{1}{2}\ln\left(1 + \frac{D}{\sigma^2}\right)$. Numerical calculations show that this happens even from smaller values of λ, however, the exact smallest value of λ for which $E_{fn}^{ae}(\lambda, D) = 0$ is hard to find. In Figure 2, Figure 3 and Figure 4, we compare the two embedding strategies by plotting their exponents for different values of σ^2/D.

4.4 Discussion

When we take a closer look at the results, the fact the sign embedder achieves a better performance should not surprise us. Clearly, when the correlation between $\boldsymbol{x}$ and $\boldsymbol{w}$ is non-negative, the additive embedder and the sign embedder achieve the same performance. However, when the correlation between $\boldsymbol{x}$ and $\boldsymbol{w}$ is negative (this happens in probability 1/2 due to the radial symmetry of the PDF of the covertext) this is not true anymore. In this case, the additive embedder tries to maximize the correlation ρ between the covertext $\boldsymbol{x}$ and the watermark $\boldsymbol{w}$ (while the detector compares the normalized correlation $\hat{\rho}_{\boldsymbol{yw}}$ between $\boldsymbol{y}$ and $\boldsymbol{w}$ to a given threshold), however, the efforts are turned to the wrong direction. Contrary to the additive embedding scheme, the sign embedder tries to maximize the absolute value of the correlation ρ while the detector compares the absolute value of the normalized correlation to a given threshold. In this case, the sign embedder tries to minimize the correlation ρ. This difference is best exemplified in the case where the $\lambda = 0$. In this case, the sign embedder achieves $E_{fn}^{se}(0, D) = \infty$ while $E_{fn}^{ae}(0, D)$ is finite since the probability of embedded vectors $\boldsymbol{Y}$ for which $\hat{\rho}_{\boldsymbol{yw}} < 0$ is not zero.

We note that although the sign embedder is suboptimal, it achieves a much better performance than the additive embedder with a slight increase in its complexity which is due to the calculation of $\mathrm{sgn}(\boldsymbol{x}, \boldsymbol{w})$.

Appendix

Proof of Theorem 1. First, we explore the case where $a = 0$, i.e., $\boldsymbol{y} = b\boldsymbol{w}$. Substituting $a = 0$ in the constraint of eq. (28), we get that $b^2 - 2\rho b + (\alpha^2 - D) \leq 0$. The fact that b is a real number implies that the discriminant of $(b^2 - 2\rho b + (\alpha^2 - D))$ is non-negative which leads to $\rho^2 - (\alpha^2 - D) \geq 0$, or $D \geq \alpha^2 - \rho^2$. This corresponds to the case where the stegotext includes *only* a fraction of $\boldsymbol{w}$ without violating the distortion constraint. In this case, the false-negative

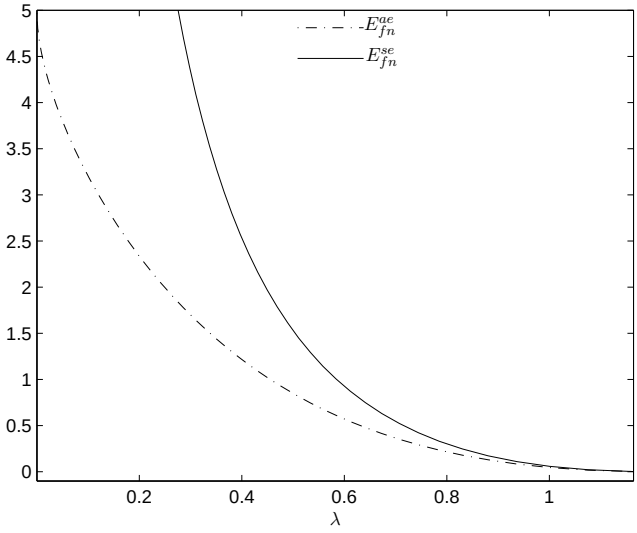

Figure 2: Error exponents of the two embedding strategies ($\sigma^2/D = .1$)

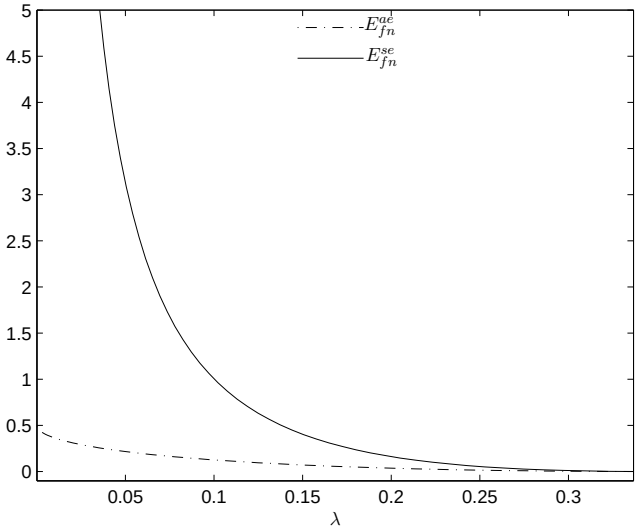

Figure 3: Error exponents of the two embedding strategies ($\sigma^2/D = 1$)

probability is zero (the distortion constraint is so loose, it allows to "erase" the covertext). In the following case, we can choose $b^* = \rho + \sqrt{\rho^2 - \alpha^2 + D}$ as the optimal solution. From now on, we assume that $D < \alpha^2 - \rho^2$ which means that $a = 0$ is not a legitimate solution. Let us assume that $\rho \geq 0$. Define $t \stackrel{\triangle}{=} b/a$, and rewrite (28) by dividing the numerator and denominator by a^2:

$$\begin{aligned} & \max_{t \in \mathbb{R}} f(t) \\ \text{subject to:} \quad & a^2 t^2 + 2(a-1)a\rho t + (a-1)^2 \alpha^2 \leq D \end{aligned} \tag{A-1}$$

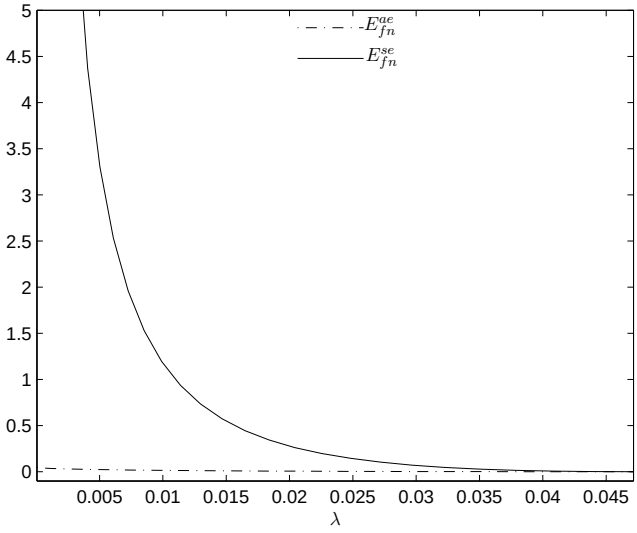

Figure 4: Error exponents of the two embedding strategies ($\sigma^2/D = 10$)

where

$$f(t) = \frac{(t+\rho)^2}{(t+\rho)^2 + (\alpha^2 - \rho^2)}$$

It is easy to show that maximizing $f(t)$ is equivalent to maximizing t. Since t is a real number, the discriminant of $\left[a^2t^2 + 2(a-1)a\rho t + (a-1)^2\alpha^2 - D\right]$ must be non-negative, i.e.,

$$\Delta = 4a^2\left[D - (a-1)^2(\alpha^2 - \rho^2)\right] \geq 0 \tag{A-2}$$

which leads to

$$1 - \sqrt{\frac{D}{\alpha^2 - \rho^2}} \leq a \leq 1 + \sqrt{\frac{D}{\alpha^2 - \rho^2}}. \tag{A-3}$$

Hence, a must be in the range $R \stackrel{\triangle}{=} \left[\, 1 - \sqrt{\frac{D}{\alpha^2-\rho^2}}, 1 + \sqrt{\frac{D}{\alpha^2-\rho^2}}\,\right]$. Let us rewrite the constraint as follows,

$$[at + (a-1)\rho]^2 + (a-1)^2(\alpha^2 - \rho^2) - D \leq 0 \tag{A-4}$$

consequently,

$$\frac{(1-a)\rho - \sqrt{D - (a-1)^2(\alpha^2-\rho^2)}}{a} \leq t \leq \frac{(1-a)\rho + \sqrt{D - (a-1)^2(\alpha^2-\rho^2)}}{a} \tag{A-5}$$

Our next step will be to maximize the upper bound on t in the allowable range of a.

$$\arg\max_{a \in R}\ t(a) \tag{A-6}$$

where

$$t(a) = \frac{(1-a)\rho + \sqrt{D-(a-1)^2(\alpha^2-\rho^2)}}{a} \,. \tag{A-7}$$

Differentiating with respect to a and equating to zero, we get

$$a_{1,2} = \frac{(\alpha^2-\rho^2)(\alpha^2-D) \pm \sqrt{D\rho^2}\sqrt{(\alpha^2-\rho^2)(\alpha^2-D)}}{\alpha^2(\alpha^2-\rho^2)} \,. \tag{A-8}$$

Accordingly, the optimal value of a and b are

$$(a^*, b^*) = \left(\arg\max\left\{t(a)|a \in \{a_1,a_2,a_3,a_4\} \bigcap R\right\}, a^* \cdot t(a^*)\right) \tag{A-9}$$

where $a_{3,4} = 1 \pm \sqrt{\frac{D}{\alpha^2-\rho^2}}$. The same results are obtained for the case where $\rho < 0$. □

Proof of Theorem 2. It is easy to show that under H_1

$$\hat{\rho}^2_{\boldsymbol{wy}} = \frac{\left(|\rho| + \sqrt{D}\right)^2}{\left(|\rho| + \sqrt{D}\right)^2 + (\alpha^2-\rho^2)} \,, \tag{A-10}$$

where α^2 and ρ are functions of the random vector $\boldsymbol{X}$. By conditioning on α^2, we can express the false-negative probability as

$$P_{fn} = \int_0^\infty \Pr\left\{\hat{\rho}^2_{\boldsymbol{wy}} \le 1 - e^{-2\lambda} \;\middle|\; H_1, \alpha^2 = r\right\} \cdot p_{\alpha^2}(r)dr, \tag{A-11}$$

where $(n\alpha^2/\sigma^2)$ is χ^2 distributed with n degrees of freedom and the probability density function for the χ^2 distribution with n degrees of freedom is given by

$$p_{\chi^2_n}(z) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} z^{n/2-1} e^{-n/2}, \qquad z \ge 0$$

where $\Gamma(\cdot)$ denotes the Gamma function. Now, given α^2, D and a threshold value $\tau \triangleq 1 - e^{-2\lambda}$, let us find the range of ρ for which $\hat{\rho}^2_{\boldsymbol{wy}} \le \tau$:

$$\hat{\rho}^2_{\boldsymbol{wy}}(\rho) = \frac{(|\rho| + \sqrt{D})^2}{(|\rho| + \sqrt{D})^2 + (\alpha^2-\rho^2)} \le \tau$$
$$\rho^2 + 2|\rho|\sqrt{D}(1-\tau) + (D - \tau D - \tau\alpha^2) \le 0$$

the function $\hat{\rho}^2_{\boldsymbol{wy}}(\rho)$ is symmetric with respect to the ρ axis, monotonically increasing in ρ and attains its minimum value $\frac{D}{D+\alpha^2}$ at $\rho = 0$. Hence, for $\alpha^2 < \frac{D(1-\tau)}{\tau}$, $\hat{\rho}^2_{\boldsymbol{wy}}$ is greater than τ. After solving the equation with respect to ρ and using the fact that $\tau \le 1$, we get that $|\hat{\rho}_{\boldsymbol{wy}}| \le \sqrt{\tau}$ implies that $|\rho| \le \sqrt{D}(\tau-1) + \sqrt{D\tau^2 + \tau\alpha^2 - \tau D}$ as long as $\alpha^2 \ge \frac{D(1-\tau)}{\tau}$. Define

$$\Theta(r) \triangleq \arccos\left[\frac{\sqrt{D}(\tau-1) + \sqrt{D\tau^2 + \tau r - \tau D}}{\sqrt{r}}\right] \tag{A-12}$$

It follows that

$$\begin{aligned}\Pr\left\{\hat{\rho}^2_{\boldsymbol{wy}} \le \tau \;\middle|\; H_1, \alpha^2 = r\right\} &= \Pr\left\{\rho^2 \le \left[\sqrt{D}(\tau-1) + \sqrt{D\tau^2 + \tau\alpha^2 - \tau D}\right]^2 \;\middle|\; H_1, \alpha^2 = r\right\} \\ &= 1 - \Pr\left\{\rho^2 > \left[\sqrt{D}(\tau-1) + \sqrt{D\tau^2 + \tau\alpha^2 - \tau D}\right]^2 \;\middle|\; H_1, \alpha^2 = r\right\} \\ &= 1 - 2\frac{A_n\big(\Theta(r)\big)}{A_n(\pi)} \doteq 1 - e^{n\ln\sin\left(\Theta(r)\right)}.\end{aligned}$$

We note that $\Pr\left\{\hat{\rho}^2_{\boldsymbol{wy}} \le \tau \middle| H_1, \alpha^2\right\} = 0$ for α^2 in the range $\left[0, \frac{D(1-\tau)}{\tau}\right]$. Therefore,

$$\begin{aligned}P^{(n)}_{fn} &= \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_{\frac{D(1-\tau)}{\tau}}^{\infty} \left[1 - e^{n\ln\sin\left(\Theta(r)\right)}\right] e^{-\frac{nr}{2\sigma^2}} \left(\frac{nr}{\sigma^2}\right)^{\frac{n-2}{2}} dr \\ &= \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \left[\int_{\frac{D(1-\tau)}{\tau}}^{\infty} \frac{\sigma^2}{r} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr - \int_{\frac{D(1-\tau)}{\tau}}^{\infty} e^{n\ln\sin\Theta(r)} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr\right]\end{aligned} \tag{A-13}$$

Our next step is to evaluate the exponential decay rate of (A-13). It is easy to see that the first integral of (A-13) has a slower exponential decay rate and therefore dictates the overall decay rate. To evaluate the exponential decay rate of $P^{(n)}_{fn}$ as $n \to \infty$ we use Laplace's method for integrals [17, Ch.4]. Therefore, we need to find the slowest exponential decay rate of the integrant in the limits of the integral. It is easy to show that

$$\lim_{n\to\infty} \frac{1}{n} \ln\left[\frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)}\right] = \frac{1}{2} \tag{A-14}$$

and therefore the overall exponent is given by

$$E^{se}_{fn}(\tau, D) = \min_{r \ge \frac{D(1-\tau)}{\tau}} \frac{1}{2}\left[\frac{r}{\sigma^2} - \ln(r/\sigma^2) - 1\right]. \tag{A-15}$$

Since $\left[r/\sigma^2 - \ln(r/\sigma^2) - 1\right]$ is monotonically increasing in r in the range $[\frac{D(1-\tau)}{\tau}, \infty)$ the minimum of (A-15) is obtained at $r = \frac{D(1-\tau)}{\tau}$. Hence, the false-negative exponent of the sign-embedder is given by

$$E^{se}_{fn}(\tau, D) = \begin{cases} 0 & , \quad \frac{D(1-\tau)}{\tau} \le \sigma^2 \\ \frac{1}{2}\left[\frac{D(1-\tau)}{\tau\sigma^2} - \ln\left(\frac{D(1-\tau)}{\tau\sigma^2}\right) - 1\right] & , \quad \text{else} \end{cases} \tag{A-16}$$

Setting $\tau = 1 - e^{-2\lambda}$ achieves (35).

□

Proof of Corollary 1. Since the false-negative probability of the improved embedder (36) is zero for $\alpha^2 \le D$ we can rewrite the integral (A-13) for the case where $\frac{1-\tau}{\tau} \le 1$ (or $\lambda \ge 1/2\ln 2$)

where the lower limit equals to D (and does not depend on λ) as following:

$$P_{fn}^{(n)} = \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_D^\infty \left[1 - e^{n \ln \sin\left(\Theta(r)\right)}\right] e^{-\frac{nr}{2\sigma^2}} \left(\frac{nr}{\sigma^2}\right)^{\frac{n-2}{2}} dr \tag{A-17}$$

optimizing using Laplace method as done in the proof of Theorem 2 leads to (37). □

Proof of Theorem 3. Given $\lambda > 0$, the false-negative probability is given by

$$P_{fn} = \Pr\left\{\hat{\rho}_{\boldsymbol{wy}} \leq \sqrt{1 - e^{-2\lambda}} \middle| H_1\right\} , \tag{A-18}$$

where the normalized correlation, under H_1, is given by

$$\hat{\rho}_{\boldsymbol{wy}} = \frac{\rho + \sqrt{D}}{\sqrt{\alpha^2 + 2\sqrt{D}\rho + D}} \quad < \quad T \, . \tag{A-19}$$

The function $\hat{\rho}_{\boldsymbol{wy}}(\rho)$ achieves its minimum at $\rho = -\frac{\alpha^2}{\sqrt{D}}$. Since $\rho \in [-\alpha, \alpha]$ we conclude that in the case where $\alpha^2 \geq D$, $\hat{\rho}_{\boldsymbol{wy}} < T$ implies that $\rho < \sqrt{D}(T^2 - 1) + T\sqrt{\alpha^2 - D(1 - T^2)}$ ($\hat{\rho}_{\boldsymbol{wy}}(\rho)$ is monotonically increasing in ρ, and $\hat{\rho}_{\boldsymbol{wy}}(-\alpha) = -1$). If $(1 - T^2)D \leq \alpha^2 < D$, $\hat{\rho}_{\boldsymbol{wy}} < T$ implies that $\sqrt{D}(T^2 - 1) - T\sqrt{\alpha^2 - D(1 - T^2)} \leq \rho \leq \sqrt{D}(T^2 - 1) + T\sqrt{\alpha^2 - D(1 - T^2)}$. For $\alpha^2 < (1 - T^2)D$, $\hat{\rho}_{\boldsymbol{wy}} \geq T$ for all $\rho \in [-\alpha, \alpha]$. Define

$$\Psi_1(r) \triangleq \arccos\left[\frac{\sqrt{D}(T^2 - 1) + T\sqrt{r - D(1 - T^2)}}{\sqrt{r}}\right] \tag{A-20}$$

$$\Psi_2(r) \triangleq \arccos\left[\frac{\sqrt{D}(T^2 - 1) - T\sqrt{r - D(1 - T^2)}}{\sqrt{r}}\right] \tag{A-21}$$

We need to pay attention to the point $r_0 = \frac{D(1-T^2)}{T^2}$ in which $\Psi_1(r_0) = \pi/2$. Beyond that point ($r > r_0$), the probability of false-negative given $\alpha^2 = r$ goes to one as n tends to infinity. Therefore, the false-negative probability can be written as follows: In the case where $\frac{1-T^2}{T^2} > 1$ (or $\lambda < \frac{1}{2}\ln(2)$)

$$\begin{aligned} P_{fn}^{(n)} = \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \Bigg[& \int_{D(1-T^2)}^{D} \frac{\sigma^2}{r} \left(e^{n \ln \sin\left(\Psi_1(r)\right)} - e^{n \ln \sin\left(\Psi_2(r)\right)}\right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{D}^{\frac{D(1-T^2)}{T^2}} \frac{\sigma^2}{r} e^{n \ln \sin\left(\Psi_1(r)\right)} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{\frac{D(1-T^2)}{T^2}}^{\infty} \frac{\sigma^2}{r} \left(1 - e^{n \ln \sin\left(\Psi_1(r)\right)}\right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \Bigg] \end{aligned} \tag{A-22}$$

The first integral in (A-22) represent the false-negative probability when both $\Psi_1(r)$ and $\Psi_2(r)$ are greater than $\pi/2$. In this case, we need to subtract the areas of two caps, i.e., $\frac{A_n(\pi - \Psi_1(r)) - A_n(\pi - \Psi_2(r))}{A_n(\pi)}$.

The second integral in (A-22) stems from the fact that for $r \geq D$ the false-negative probability (given $\alpha^2 = r$) equals to $\frac{A_n(\pi - \Psi_1(r))}{A(\pi)}$. The last integral in (A-22) stems from the fact that the false-negative probability (given $\alpha^2 = r$) equals to $1 - \frac{A(\Psi_1(r))}{A(\pi)}$. In a similar way, in the case where $\frac{1-T^2}{T^2} \leq 1$ (or $\lambda \geq \frac{1}{2}\ln(2)$)

$$\begin{aligned} P_{fn}^{(n)} = \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \Bigg[& \int_{D(1-T^2)}^{\frac{D(1-T^2)}{T^2}} \frac{\sigma^2}{r} \left(e^{n \ln \sin\left(\Psi_1(r)\right)} - e^{n \ln \sin\left(\Psi_2(r)\right)} \right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{\frac{D(1-T^2)}{T^2}}^{D} \frac{\sigma^2}{r} \left(1 - e^{n \ln \sin\left(\Psi_1(r)\right)} - e^{n \ln \sin\left(\Psi_2(r)\right)} \right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{D}^{\infty} \frac{\sigma^2}{r} \left(1 - e^{n \ln \sin\left(\Psi_1(r)\right)} \right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \Bigg] \end{aligned} \tag{A-23}$$

Since we are interested in the exponential decay rate (to the first order), the slowest exponent dictates the overall exponential behavior. Therefore, the fact that $\sin\left(\Psi_1(r)\right) > \sin\left(\Psi_2(r)\right)$ for $D(1-T^2) \leq r \leq D(1-T^2)/T^2$ implies that

$$\begin{aligned} P_{fn} \doteq \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \Bigg[& \int_{D(1-T^2)}^{\frac{D(1-T^2)}{T^2}} \frac{\sigma^2}{r} e^{n \ln \sin\left(\Psi_1(r)\right)} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{\frac{D(1-T^2)}{T^2}}^{\infty} \frac{\sigma^2}{r} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \Bigg] . \end{aligned} \tag{A-24}$$

Again, using the Laplace's method for integrals [17, Ch.4] we can conclude that

$$E_{fn}^{ae}(T, D) = \min\left\{ E_1(T, D), E_2(T, D) \right\} , \tag{A-25}$$

where,

$$E_1(T, D) = \min_{D(1-T^2) < r \leq \frac{D(1-T^2)}{T^2}} \frac{1}{2} \left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 2 \ln \sin\left(\Psi_1(r)\right) - 1 \right] \tag{A-26}$$

$$E_2(T, D) = \min_{r > \frac{D(1-T^2)}{T^2}} \frac{1}{2} \left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 1 \right] . \tag{A-27}$$

$E_2(T, D)$ is given by

$$E_2(T, D) = \begin{cases} 0 & , \quad \frac{D(1-T^2)}{T^2} \leq \sigma^2 \\ \frac{1}{2}\left[\frac{D(1-T^2)}{T^2\sigma^2} - \ln\left(\frac{D(1-T^2)}{T^2\sigma^2}\right) - 1 \right] & , \quad \text{else} \end{cases} \tag{A-28}$$

Since $T^2 = 1 - e^{-2\lambda}$, then $E_2(\lambda, D) = E_{fn}^{se}(\lambda, D)$ and therefore $E_{fn}^{ae}(\lambda, D) \leq E_{fn}^{se}(\lambda, D)$. Our next step will be to prove that $E_1(T, D) < E_2(T, D)$ when $\frac{D(1-T^2)}{T^2} > \sigma^2$ (otherwise, $E_{fn}^{ae}(T, D) = 0$). Define

$$f(r) = \frac{r}{2\sigma^2} - \frac{1}{2}\ln\left(\frac{r}{\sigma^2}\right) - \ln \sin\left(\Psi_1(r)\right) - \frac{1}{2} \tag{A-29}$$

$f(r)$ is a continuous, non-negative function in the range $D(1-T^2) < r \leq \frac{D(1-T^2)}{T^2}$. Clearly,

$$E_1(T,D) \leq f\left(\frac{D(1-T^2)}{T^2}\right) = E_2(T,D). \tag{A-30}$$

In addition, $f'(r)$ is continuous in the above range. It can easily be shown that

$$f'\left(\frac{D(1-T^2)}{T^2}\right) = \frac{1}{2}\left[1 - \frac{T^2\sigma^2}{D(1-T^2)}\right] > 0 \tag{A-31}$$

hence, $f(r)$ is monotonically increasing in small neighborhood of $\frac{D(1-T^2)}{T^2}$, and therefore $E_1(T,D) < E_2(T,D)$. This fact leads to the conclusion that $E_{fn}^{ae}(\lambda,D) < E_{fn}^{se}(\lambda,D)$. The exact value of $E_1(T,D)$ is cumbersome and therefore will not be presented. □

References

[1] R. Anderson and F. Petitcolas, "On the limits of stenography," *IEEE J. Select. Areas Commun.*, vol. 16, no. 4, pp. 474–481, May 1998.

[2] F. Petitcolas, R. Anderson, and M. Kuhn, "Information hiding – a survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, July 1999.

[3] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. IEEE*, vol. 87, no. 7, pp. 1127–1141, July 1999.

[4] M. Barni and F. Bartolini, *Watermarking Systems Engineering: Enabling Digital Assets Security and Other Applications.* Marcel Dekker, 2004.

[5] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563–593, Mar. 2003.

[6] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE*, vol. 87, no. 7, pp. 1079–1107, July 1999.

[7] J. Hernandez and F. Perez-Gonzalez, "Statistical analysis of watermarking schemes for copyright protection of images," *Proc. IEEE*, vol. 87, no. 7, pp. 1142–1166, July 1999.

[8] M. L. Miller, I. J. Cox, and J. A. Bloom, "Informed embedding: Exploiting image and detector information during watermark insertion," in *International Conference on Image Processing Processing (ICIP) 2000*, vol. 3. IEEE, 2000, pp. 1–4.

[9] M. L. Miller and J. A. Bloom, "Computing the probability of false watermark detection," in *IH '99: Proceedings of the Third International Workshop on Information Hiding*. London, UK: Springer-Verlag, 2000, pp. 146–158.

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

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

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

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

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

[15] 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, Nov. 1994.

[16] A. D. Wyner, "A bound on the number of distinguishable functions which are time-limited and approximately band-limited," *SIAM Journal on Applied Mathematics*, vol. 24, no. 3, pp. 289–297, May 1973.

[17] N. G. de Bruijn, *Asymptotic Methods in Analysis*, 3rd ed. North-Holland publishing company, 1970.