Optimal Watermark Embedding and Detection Strategies Under Limited Detection Resources *

Neri Merhav and Erez Sabbag

May 14, 2007

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

Abstract

An information–theoretic approach is proposed to watermark embedding and detection under limited detector resources. First, we consider the attack-free scenario under which asymptotically optimal decision regions in the Neyman-Pearson sense are proposed, along with the optimal embedding rule. Later, we explore the case of zero-mean i.i.d. Gaussian covertext distribution with unknown variance under the attack-free scenario. 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. Finally, these results are extended to the case of memoryless attacks and general worst case attacks. Optimal decision regions and embedding rules are offered, and the worst attack channel is identified.

1 Introduction

The field of information embedding and watermarking has 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]–[4]). 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.

Consider the system depicted in Fig. 1: Let $\boldsymbol{x} = \big(x_1, \ldots, x_n\big)$ denote a covertext sequence emitted from a memoryless source P_X, and let $\boldsymbol{u} = \big(u_1, \ldots, u_n\big)$ denote a watermark sequence available at the embedder and at the detector. Our work focuses on finding the optimal embedding and detection rules for the following binary hypothesis problem: under hypothesis H_1, the stegotext sequence $\boldsymbol{y} = \big(y_1, \ldots, y_n\big)$ is "watermarked" using the embedder $\boldsymbol{y} = f_n(\boldsymbol{x}, \boldsymbol{u})$, while under H_0, $\boldsymbol{y} = \boldsymbol{x}$, i.e, the stegotext sequence in not "watermarked". An attack channel $W_n(\boldsymbol{z}|\boldsymbol{y})$, fed by the stegotext, produces a forgery $\boldsymbol{z}$, which in turn, is observed by the detector. Now, given the forgery sequence $\boldsymbol{z}$ and the watermark sequence $\boldsymbol{u}$, the

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

detector needs to decide whether the forgery is "watermarked" or not. Performance is evaluated under the Neyman-Pearson criterion, namely, minimum false detection probability while the false alarm probability is kept lower than a prescribed level. The problem is addressed under different statistical assumptions: the covertext distribution is known or unknown to the embedder/detector, the attack channel is known to be a memoryless attack or it is a general attack channel, and the watermark sequence is deterministic or random.

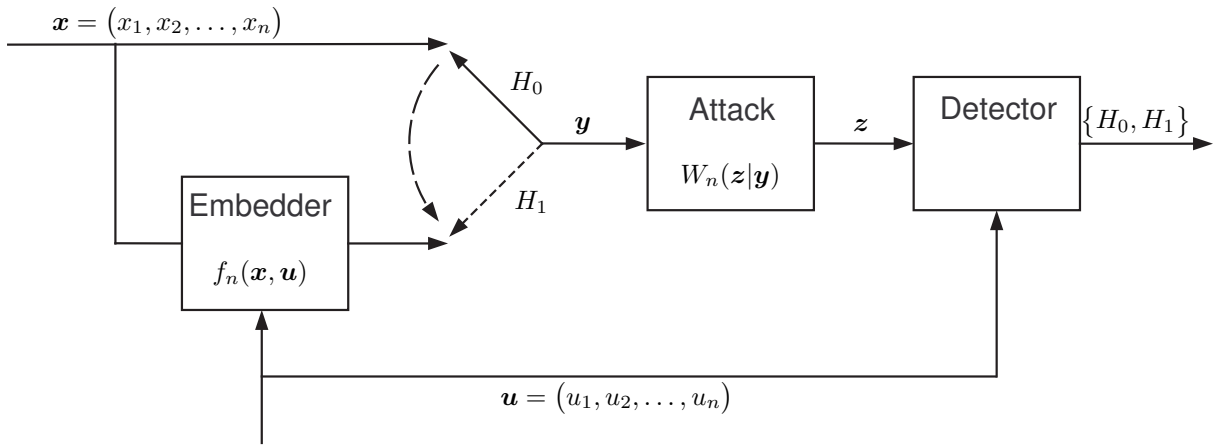

Figure 1: The watermarking and detection problem.

Surprisingly, this problem did not receive much attention in the information theory community. In [5], the problem of universal detection of messages via finite state channel was considered, and an optimal decision rule was proposed for deciding whether the observed sequence is the product of an unknown finite-state channel fed by one of two predefined sequences. Liu and Moulin [6],[7] explored the error exponent of two popular one-bit WM systems: the spread-spectrum scheme and the quantized-index-modulation (QIM) watermarking scheme, under a general additive attack. Bounds and closed form expressions were offered for the error exponents. We note that the setting of [6] is different from ours: here, we are trying to find the best embedder given detection resource under Neyman-Pearson criterion of optimality, while in [6], the performance (the error exponent) of a given embedding schemes and a given source distribution are evaluated under additive attacks. In [8], the problem of embedding/detection was formulated under limited detection resources and the optimal decision region and the optimal embedding rule were offered to the attack-free scenario.

Many researchers from the signal/image processing community (e.g., [2],[3],[9]– [13], [14, Sec.4.2] and references therein) have devoted research efforts to explore the problem of optimal watermark embedding and detection with one common assumption: the watermark embedding rule is normally taken to be additive (linear), i.e., the stegotext vector $\boldsymbol{y}$ is given by

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

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

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

where in both cases, $u_i = \pm 1$, and 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 under the attack free scenario (i.e., $\boldsymbol{z} = \boldsymbol{y}$), assuming a zero–mean, Gaussian, i.i.d. covertext distribution, is a correlation detector, which decides positively (H_1: $\boldsymbol{y} = \boldsymbol{x} + \gamma \boldsymbol{u}$) if the correlation, $\sum_{i=1}^{n} u_i y_i$, 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 (the additive embedding scheme is, in fact, the spread-spectrum modulation technique [15] in which the covertext is treated as an 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 $u_i = +1$ relative to those corresponding to $u_i = -1$, 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 $\boldsymbol{y} = f_n(\boldsymbol{x}, \boldsymbol{u})$, trading off the quality of $\boldsymbol{y}$ and the detectability of $\boldsymbol{u}$. Clearly, for an arbitrary choice of f_n, the above described detectors are no longer optimal in general.

Malvar and Florêncio [16] have noticed that better performance can be gained if γ is chosen as a function of the watermark and the covertext. However, their choice does not lead to the optimal performance as will be shown later. Recently, Furon [17] explored the zero-bit watermark problem using a different setting in which the watermark sequence is a function of the covertext and under a different criterion of optimality.

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.,[11],[12],[18] and references therein), very few works [6],[7] deal with the optimal 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.

The problem of finding the optimum watermark embedder f_n 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_n, are, in general, hard to compute, and a–fortiori hard to optimize in closed form. Moreover, obtaining closed form expressions for the optimal embedder and decision regions when the covertext distribution is unknown is even harder (see Section 2 for more details).

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 $\boldsymbol{z}$ and $\boldsymbol{u}$. For example, the energy of $\boldsymbol{z}$, $\sum_{i=1}^{n} z_i^2$, and the correlation $\sum_{i=1}^{n} u_i z_i$, 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 $\sum_{i:\ u_i=+1} z_i^2$, and $\sum_{i:\ u_i=-1} z_i^2$, and so on. 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 for different settings of the problem.

First, we formulate the embedding and detection problem under the attack free scenario. We devise an asymptotically optimal detector and embedding rule among all detectors which base their decisions on the empirical joint distribution of $\boldsymbol{z}$ and $\boldsymbol{u}$. This modeling assumption, where the detector has access to a limited set of empirical statistics of $\boldsymbol{u}$ and $\boldsymbol{z}$, has two motivations. First, it enables a fair comparison (in terms of detection computational resources) to different embedding/detection methods reported in the literature of WM in which most of the detectors use a similar set of statistics (mostly, correlation and energy) to base their decisions. Second, this approach highlights the tradeoff between detection complexity and performance: Extending the set of statistics on which the detector can base its decisions, might improve the system performance, however, it increases the detector's complexity.

Later, we discuss different aspects of the basic problem, namely, practical issues regarding the implementability of the embedder, universality w.r.t. the covertext distribution, other detector's statistics, and the case where the watermark sequence is random too. These results are obtained by extending the techniques, presented in [5],[19]–[21], which are closely related to universal hypothesis testing problems. We apply these results to a zero-mean i.i.d. Gaussian covertext distribution with unknown variance. We propose a closed-form expression for the optimal embedder, and suggest a lower bound on the false-negative probability error exponent. By analyzing the error exponent of the additive embedder and using the suggested lower bound, we show that the optimal embedder is superior to the customary additive embedder in the exponential sense. Finally, we extend these results to memoryless attack channels and worst-case general attack channels. The worst-attack channel is identified and optimal embedding and detection rules are offered. The model of general worst-case attack channels, treated here, was already considered in the WM literature but in a different context. In [22], general attack channels were considered, where the capacity and random-coding error exponent where derived for the private watermarking game under general attack channels. In [23], the capacity of public watermark game under general attack channels was derived for constant composition codes. This paper is a further development and an extension of [8], [24] and it gives a detailed account for the results of [25].

2 Basic Derivation

We begin with some notation 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). We adopt the following conventions: The minimum (maximum) of a function over an empty set is understood to be ∞ ($-\infty$). The notation $a_n \doteq b_n$, for two positive sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, expresses asymptotic equality in the logarithmic scale, i.e.,

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

Let the vector $\hat{P}_{\boldsymbol{x}} = \big\{\hat{P}_{\boldsymbol{x}}(a),\ a \in \mathcal{X}\big\}$ denotes the empirical distribution induced by a vector $\boldsymbol{x} \in \mathcal{X}^n$, where $\hat{P}_{\boldsymbol{x}}(a) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i = a\}$. The type class $T(\boldsymbol{x})$ is the set of vectors $\tilde{\boldsymbol{x}} \in \mathcal{X}^n$ such that $\hat{P}_{\tilde{\boldsymbol{x}}} = \hat{P}_{\boldsymbol{x}}$. Similarly, the joint empirical distribution induced by $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ is the vector:

$$\hat{P}_{\boldsymbol{xy}} = \Big\{\hat{P}_{\boldsymbol{xy}}(a,b),\ a \in \mathcal{X},\ b \in \mathcal{Y}\Big\}\ , \tag{3}$$

where

$$\hat{P}_{\boldsymbol{xy}}(a,b) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\big\{x_i = a, y_i = b\big\}, \quad x \in \mathcal{X},\ y \in \mathcal{Y}\ , \tag{4}$$

i.e., $\hat{P}_{\boldsymbol{xy}}(a,b)$ is the relative frequency of the pair (a,b) along the pair sequence $(\boldsymbol{x}, \boldsymbol{y})$. Likewise, the type class $T(\boldsymbol{x}, \boldsymbol{y})$ is the set of all pairs $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that $\hat{P}_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} = \hat{P}_{\boldsymbol{xy}}$. The conditional type class $T(\boldsymbol{y}|\boldsymbol{x})$, for given vectors $\boldsymbol{x} \in \mathcal{X}^n$, and $\boldsymbol{y} \in \mathcal{Y}^n$ is the set of all vectors $\tilde{\boldsymbol{y}} \in \mathcal{Y}^n$ such that $T(\boldsymbol{x}, \tilde{\boldsymbol{y}}) = T(\boldsymbol{x}, \boldsymbol{y})$. We denote by $\hat{E}_{\boldsymbol{xy}}(\cdot)$ expectation with respect to empirical joint distribution $\hat{P}_{\boldsymbol{xy}}$. The Kullback-Leibler divergence between two distributions P and Q on $\mathcal{A}$, where $|\mathcal{A}| < \infty$ is defined as

$$\mathcal{D}(P\|Q) = \sum_{a\in\mathcal{A}} P(a) \ln \frac{P(a)}{Q(a)}\ ,$$

with the conventions that $0 \ln 0 = 0$, and $p \ln \frac{p}{0} = \infty$ if $p > 0$. We denote the empirical entropy of a vector $\boldsymbol{x} \in \mathcal{X}^n$ by $\hat{H}_{\boldsymbol{x}}(X)$, where

$$\hat{H}_{\boldsymbol{x}}(X) = -\sum_{a\in\mathcal{X}} \hat{P}_{\boldsymbol{x}}(a) \ln \hat{P}_{\boldsymbol{x}}(a)\ .$$

Other information theoretic quantities governed by empirical distributions (e.g., conditional empirical entropy, empirical mutual information) will be denoted similarly.

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 $\|\boldsymbol{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a}\rangle}$. Let $\mathrm{Vol}\{A\}$ denote the volume of a set $A \subset \mathbb{R}^n$, i.e., $\mathrm{Vol}\{A\} = \int_A d\boldsymbol{x}$. We denote by $\mathrm{sgn}(\cdot)$ the signum function, where $\mathrm{sgn}(x) = \mathbb{1}\{x \geq 0\} - \mathbb{1}\{x < 0\}$.

Throughout this paper, and without essential loss of generality, we assume that the components of $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ all take on values in the same finite alphabet $\mathcal{A}$. In Section 4, the assumption that $\mathcal{A}$ is finite will be dropped, and $\mathcal{A}$ will be allowed to be an infinite set, like the real line. The components of the watermark $\boldsymbol{u}$ 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_X.

Throughout the sequel, until Section 5 (exclusively), we assume that there is no attack, i.e., the channel $W_n(\boldsymbol{z}|\boldsymbol{y})$ is the identity channel:

$$W_n(\boldsymbol{z}|\boldsymbol{y}) = \begin{cases} 1 & , \quad \boldsymbol{z} = \boldsymbol{y} \\ 0 & , \quad \text{else} \end{cases}\ .$$

This is referred to as the *attack-free* scenario. In this scenario, the detector will use $\boldsymbol{y}$ and $\boldsymbol{u}$ to base its decisions.

For a given $\boldsymbol{u} \in \mathcal{B}^n$, 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 $\boldsymbol{u}$ 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

$$P_{fn} = \sum_{\boldsymbol{x}:\ f_n(\boldsymbol{x},\boldsymbol{u})\in\Lambda^c} P_X(\boldsymbol{x}) \tag{5}$$

subject to the following constraints:

(1) Given a certain distortion measure $d_e(\cdot,\cdot)$ and distortion level D_e, the distortion between $\boldsymbol{x}$ and $\boldsymbol{y}$, $d_e(\boldsymbol{x},\boldsymbol{y}) = d_e\big(\boldsymbol{x}, f_n(\boldsymbol{x},\boldsymbol{u})\big)$, does not exceed nD_e.

(2) The false positive probability is upper bounded by

$$P_{fp} \stackrel{\Delta}{=} \sum_{\boldsymbol{y}\in\Lambda} P_X(\boldsymbol{y}) \le e^{-\lambda n}, \tag{6}$$

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

In other words, we would like to choose f_n 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.

Clearly, the problem is a classical hypothesis problem (under the Neyman-Pearson criterion of optimality), with the following hypotheses: $H_0 : \boldsymbol{y} = \boldsymbol{x}$ (the covertext is not "marked") and $H_1 : \boldsymbol{y} = f_n(\boldsymbol{x},\boldsymbol{u})$ (the covertext is "marked"). Given f_n and $\boldsymbol{u}$, we can define the conditional distribution of $\boldsymbol{y}$ given the two hypotheses:

$$\begin{aligned} P(\boldsymbol{y}|H_0) &= P_X(\boldsymbol{y}) \ , \\ P(\boldsymbol{y}|H_1) &= \sum_{\boldsymbol{x}:f_n(\boldsymbol{x},\boldsymbol{u})=\boldsymbol{y}} P_X(\boldsymbol{x}) \ . \end{aligned}$$

where $P_X(\boldsymbol{x})$ is the covertext distribution. The optimal test which minimizes the false-negative probability under the Neyman-Pearson criterion of optimality is the likelihood ratio test (LRT) [26, p. 34]:

$$L(\boldsymbol{y}) = \frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_0)} \begin{array}{c} H_1 \\ > \\ < \\ H_0 \end{array} \eta$$

where η is chosen such that

$$P_{fp}\big(f_n,\boldsymbol{u}\big) = \sum_{\boldsymbol{y}:L(\boldsymbol{y})\ge\eta} P_X(\boldsymbol{y}) = e^{-n\lambda} \ . \tag{7}$$

Note that η is a function of λ, f_n and $\boldsymbol{u}$, therefore, we could not find a closed-form expression for η for any general embedding rule and watermark sequence. The false-negative probability associated with the above optimal test is given by

$$P_{fn}(f_n,\lambda,\boldsymbol{u}) = \sum_{\boldsymbol{y}:L(\boldsymbol{y})<\eta} \sum_{\boldsymbol{x}:f_n(\boldsymbol{x},\boldsymbol{u})=\boldsymbol{y}} P_X(\boldsymbol{x}). \tag{8}$$

Now, given a distortion level D_e measured using a distortion function $d_e(\cdot,\cdot)$, we would like to devise an embedder f_n which minimizes the false-negative probability while the distortion between the covertext $\boldsymbol{x}$ and the stegotext $\boldsymbol{y}$ does not exceed nD_e and the false-positive probability is kept lower than $e^{-n\lambda}$, i.e.,

$$f_n^* = \arg \min_{\substack{f_n\,:\,d_e(\boldsymbol{x},f_n(\boldsymbol{x},\boldsymbol{u}))\le nD_e,\forall\boldsymbol{x} \\ P_{fp}(f_n,\boldsymbol{u})\le e^{-n\lambda}}} P_{fn}(f_n,\lambda) \ . \tag{9}$$

The above general problem of finding the optimal embedding rule and detection regions is by no means trivial. The fact that the probabilities of the two kinds of error cannot be expressed in a close form make it very hard to solve this optimization problem and, as far as we know, there is no known solution for it. Moreover, obtaining closed form expressions for the optimal embedder and decision regions when P_X is unknown is even harder.

We therefore make an additional assumption regarding the statistics employed by the detector. Suppose that we limit ourselves to the class of all detectors which base their decisions on certain empirical statistics associated with $\boldsymbol{u}$ and $\boldsymbol{y}$, for example, the empirical joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$, i.e., $\hat{P}_{\boldsymbol{uy}}$. Note that the requirement that the decision of the detector depends solely on $\hat{P}_{\boldsymbol{uy}}$ means that Λ and Λ^c are unions of conditional type classes of $\boldsymbol{y}$ given $\boldsymbol{u}$.

It may seem, at a first glance, that the sequence $\boldsymbol{u}$ is superfluous in the definition of the problem, since it is available to all legitimate parities. However, the presence of the watermark sequence $\boldsymbol{u}$ at the detector provides the detector with a refined version of the statistics of its input (based on the joint empirical statistics of $\boldsymbol{y}$ and $\boldsymbol{u}$) and can be regarded as a secret key shared by both legitimate sides. This additional information at the detector improves the overall performance of the system.

For a given $\lambda > 0$, define

$$\Lambda_* = \left\{ \boldsymbol{y} : \ \ln P_X(\boldsymbol{y}) + n\hat{H}_{\boldsymbol{uy}}(Y|U) + \lambda n - |\mathcal{A}| \ln(n+1) \leq 0 \right\}. \tag{10}$$

The following theorem asserts that Λ_* is asymptotically optimal decision region:

Theorem 1. *(i) $P_{fp}(\Lambda_*) \leq e^{-n(\lambda - \delta_n)}$ where $\lim_{n\to\infty} \delta_n = 0$.*

(ii) For every $\Lambda \subseteq \mathcal{A}^n$ that satisfies $P_{fp}(\Lambda) \leq e^{-n\lambda'}$ for some $\lambda' > \lambda$, we have $\Lambda_*^c \subseteq \Lambda^c$ for all sufficiently large n.

In the above theorem it is argued that Λ_* fulfills the false-positive constraint while minimizes the false-negative probability, i.e., for any decision region Λ which fulfills the false-positive constraint and for any embedding rule $f_n(\boldsymbol{x}, \boldsymbol{u})$ the following holds

$$P_{fn}(\Lambda_*^c) \leq P_{fn}(\Lambda^c) \, . \tag{11}$$

Proof. Let $T(\boldsymbol{y}|\boldsymbol{u}) \subseteq \Lambda$. Then, we have

$$\begin{aligned} e^{-\lambda n} &\geq \sum_{\boldsymbol{y}' \in \Lambda} P_X(\boldsymbol{y}') \\ &\geq \sum_{\boldsymbol{y}' \in T(\boldsymbol{y}|\boldsymbol{u})} P_X(\boldsymbol{y}') \\ &\geq |T(\boldsymbol{y}|\boldsymbol{u})| \cdot P_X(\boldsymbol{y}) \\ &\geq (n+1)^{-|\mathcal{A}|} e^{n\hat{H}_{\boldsymbol{uy}}(Y|U)} \cdot P_X(\boldsymbol{y}) \, , \end{aligned} \tag{12}$$

where the first inequality is by the assumed false positive constraint, the second inequality is since $T(\boldsymbol{y}|\boldsymbol{u}) \subseteq \Lambda$, and the third inequality is due to the fact that all sequences within $T(\boldsymbol{y}|\boldsymbol{u})$ are equiprobable under P_X as they all have the same empirical distribution, which forms the sufficient statistics for the

memoryless source P_X. 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 [27], defined as:

$$\hat{H}_{\boldsymbol{uy}}(Y|U) = -\sum_{u,y} \hat{P}_{\boldsymbol{uy}}(u,y) \ln \hat{P}_{\boldsymbol{uy}}(y|u) \;, \tag{13}$$

where $\hat{P}_{\boldsymbol{uy}}(y|u)$ is the empirical conditional probability of Y given U. We have actually shown that every $T(\boldsymbol{y}|\boldsymbol{u})$ in Λ is also in Λ_*, in other words, if Λ satisfies the false positive constraint (6), it must be a subset of Λ_*. This means that $\Lambda_*^c \subseteq \Lambda^c$ 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 (6). 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 [5, 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 $\hat{P}_{\boldsymbol{uy}}$, as required. □

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

$$\begin{aligned} \Gamma_*^c(f_n) &= \left\{ \boldsymbol{x} : \; f_n(\boldsymbol{x},\boldsymbol{u}) \in \Lambda_*^c \right\} \\ &= \left\{ \boldsymbol{x} : \; \ln P_X(f_n(\boldsymbol{x},\boldsymbol{u})) + n\hat{H}_{\boldsymbol{u},f_n(\boldsymbol{x},\boldsymbol{u})}(Y|U) + \lambda n - |\mathcal{A}| \ln(n+1) > 0 \right\}, \end{aligned} \tag{14}$$

then following eq. (5), P_{fn} can be expressed as

$$P_{fn} = \sum_{\boldsymbol{x} \in \Gamma_*^c(f_n)} P_X(\boldsymbol{x}). \tag{15}$$

Consider now the following embedder:

$$f_n^*(\boldsymbol{x},\boldsymbol{u}) = \text{argmin}_{\boldsymbol{y}:\; d_e(\boldsymbol{x},\boldsymbol{y}) \le nD_e} \left[\ln P_X(\boldsymbol{y}) + n\hat{H}_{\boldsymbol{uy}}(Y|U) \right], \tag{16}$$

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

3 Discussion

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 (16)

The first impression might be that the minimization in (16) is prohibitively complex as it appears to require an exhaustive search over the sphere $\{\boldsymbol{y} : \; d_e(\boldsymbol{x},\boldsymbol{y}) \le nD_e\}$, 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_X,

$$\ln P_X(\boldsymbol{y}) = -n \left[\hat{H}_{\boldsymbol{y}}(Y) + \mathcal{D}(\hat{P}_{\boldsymbol{y}} \| P_X) \right] \;, \tag{17}$$

where $\hat{H}_{\boldsymbol{y}}(Y)$ is the empirical entropy of $\boldsymbol{y}$ and $\mathcal{D}(\hat{P}_{\boldsymbol{y}}\|P_X)$ is the divergence between the empirical distribution of $\boldsymbol{y}$, $\hat{P}_{\boldsymbol{y}}$, and the source P_X. Moreover, if $d_e(\cdot,\cdot)$ is an additive distortion measure, i.e., $d_e(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n d_e(x_i, y_i)$, then $d_e(\boldsymbol{x},\boldsymbol{y})/n$ 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 (16) becomes equivalent to maximizing $[\hat{I}_{\boldsymbol{uy}}(U;Y) + \mathcal{D}(\hat{P}_{\boldsymbol{y}}\|P_X)]$ subject to $\hat{E}_{\boldsymbol{xy}} d_e(X,Y) \leq D_e$, where $\hat{I}_{\boldsymbol{uy}}(U;Y)$ denotes the empirical mutual information induced by the joint empirical distribution $\hat{P}_{\boldsymbol{uy}}$ and $\hat{E}_{\boldsymbol{xy}}$ denotes the aforementioned expectation with respect to $\hat{P}_{\boldsymbol{xy}}$. Now, observe that for given $\boldsymbol{x}$ and $\boldsymbol{u}$, both $[\hat{I}_{\boldsymbol{uy}}(U;Y)+ \mathcal{D}(\hat{P}_{\boldsymbol{y}}\|P_X)]$ and $\hat{E}_{\boldsymbol{xy}} d_e(X,Y) \leq D_e$ depend on $\boldsymbol{y}$ only via its conditional type class given $(\boldsymbol{x},\boldsymbol{u})$, namely, the conditional empirical distribution $\hat{P}_{\boldsymbol{uxy}}(y|x,u)$. Once the optimal $\hat{P}_{\boldsymbol{uxy}}(y|x,u)$ has been found, it does not matter which vector $\boldsymbol{y}$ is chosen from the corresponding conditional type class $T(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$. Therefore, the optimization across n–vectors in (16) 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 $\{P(y|x,u)\}$ (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_X 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_X are known is rather questionable. Suppose then that P_X 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 (6) for *every* memoryless source, to be on the safe side. In other words, eq. (6) would be replaced by

$$\max_{P_X} \sum_{\boldsymbol{y}\in\Lambda} P_X(\boldsymbol{y}) \leq e^{-\lambda n} \,, \tag{18}$$

where the maximization over P_X 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 $P_X(\boldsymbol{y})$ should be replaced by $\max_{P_X} P_X(\boldsymbol{y})$ in all places (see also [5]). Since $\ln \max_{P_X} P_X(\boldsymbol{y}) = -n\hat{H}_{\boldsymbol{y}}(Y)$, this means that the modified version of Λ_* compares the empirical mutual information $\hat{I}_{\boldsymbol{uy}}(U;Y)$ 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 (16) maximizes $\hat{I}_{\boldsymbol{uy}}(U;Y)$ 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 [27, 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 $\{(u,y)\}$. 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 positions, 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: $\boldsymbol{y}$, $\boldsymbol{u}$, and all the $2K$ corresponding cyclic shifts of $\boldsymbol{u}$. Our earlier analysis will carry over provided that the above definition of $\hat{H}_{\boldsymbol{uy}}(Y|U)$ would be replaced the conditional empirical entropy of $\boldsymbol{y}$ given $\boldsymbol{u}$ and all its cyclic shifts. This is different from the exhaustive search (ES) approach (see, e.g., [28]) 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_X, whereas the watermark $\boldsymbol{u}$ is fixed, and hence can be thought of as being deterministic. Another possible setting assumes that $\boldsymbol{u}$ is random as well, in particular, being drawn from another source P_U, 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 $\boldsymbol{u}$ is induced by the randomness of the key. Here, the decision regions Λ and Λ^c 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 $P_X(\boldsymbol{x})P_U(\boldsymbol{u})$. The fact that $\boldsymbol{u}$ is emitted from a memoryless source with a known distribution, makes this model weaker compared to the model treated above in which $\boldsymbol{u}$ is an individual sequence. 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 $T(\boldsymbol{y}|\boldsymbol{u})$ would be replaced by the joint type class $T(\boldsymbol{u}, \boldsymbol{y})$, namely, the set of all *pairs* of sequences $\{(\boldsymbol{u}', \boldsymbol{y}')\}$ that have the same empirical distribution as $(\boldsymbol{u}, \boldsymbol{y})$ (as opposed to the conditional type class which is defined as the set of all such $\boldsymbol{y}$'s for a given $\boldsymbol{u}$). Thus, the corresponding version of Λ_* would be

$$\Lambda_* = \left\{(\boldsymbol{u}, \boldsymbol{y}) : \ \ln P_X(\boldsymbol{y}) + \ln P_U(\boldsymbol{u}) + n\hat{H}_{\boldsymbol{uy}}(U, Y) + \lambda n - |\mathcal{A}| \ln(n+1) \le 0\right\}, \tag{19}$$

where $\hat{H}_{\boldsymbol{uy}}(U, Y)$ is the empirical joint entropy induced by $(\boldsymbol{u}, \boldsymbol{y})$, 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 $|T(\boldsymbol{u}, \boldsymbol{y})|$ in more general situations than it is for $|T(\boldsymbol{y}|\boldsymbol{u})|$. 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 $\{u_i = u, \ y_i = y, \ y_{i-1} = y'\}$, 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 $\{(u_i, y_i)\}$. Another example for the differences between random watermark and deterministic watermark can be seen in Section 6.

It should be also pointed out that once $\boldsymbol{u}$ is assumed random (say, drawn from a BSS), it is possible to devise a decision rule that is 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

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

rule, obtained using a similar technique, accepts H_1 whenever $\hat{H}_{\boldsymbol{uy}}(U|Y) \leq 1-\lambda$, and the embedder minimizes $\hat{H}_{\boldsymbol{uy}}(U|Y)$ subject to the distortion constraint accordingly.

4 Continuous Alphabet – 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 [27], used above, to vectors with real–valued components (see, e.g., [29],[30],[31]). 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 sufficient statistics with $\boldsymbol{u}$ have (within infinitesimally small tolerance) prescribed values, and to have a parallel analysis to that of the previous section, we have to be able to assess the exponential order of the volume of the conditional type class.

Suppose 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 $\sum_{i=1}^n u_i y_i$. These assumptions are the same as in many theoretical papers in the literature of watermark detection. Then, the conditional empirical entropy $\hat{H}_{\boldsymbol{uy}}(Y|U)$ should be replaced by the empirical differential entropy $\hat{h}_{\boldsymbol{uy}}(Y|U)$, given by [30]:

$$\begin{aligned}
\hat{h}_{\boldsymbol{uy}}(Y|U) &= \frac{1}{2}\ln\left[2\pi e \cdot \min_{\beta}\left(\frac{1}{n}\sum_{i=1}^n (y_i-\beta u_i)^2\right)\right] \\
&= \frac{1}{2}\ln\left[2\pi e\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - \frac{(\frac{1}{n}\sum_{i=1}^n u_i y_i)^2}{\frac{1}{n}\sum_{i=1}^n u_i^2}\right)\right] \\
&= \frac{1}{2}\ln\left[2\pi e\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - (\frac{1}{n}\sum_{i=1}^n u_i y_i)^2\right)\right]. \qquad (20)
\end{aligned}$$

The justification of eq. (20) is as follows: For a given $\epsilon > 0$ define the set

$$T_\epsilon(\boldsymbol{y}|\boldsymbol{u}) = \left\{\tilde{\boldsymbol{y}} \in \mathbb{R}^n : \left|\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \tilde{y}_i^2\right| \leq n\epsilon, \left|\sum_{i=1}^n y_i u_i - \sum_{i=1}^n \tilde{y}_i u_i\right| \leq n\epsilon\right\}. \qquad (21)$$

Similarly as in Lemma 3 [30], it can be shown that

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\ln\Big[\mathrm{Vol}\big\{T_\epsilon(\boldsymbol{y}|\boldsymbol{u})\big\}\Big] = \hat{h}_{\boldsymbol{uy}}(Y|U)\,. \qquad (22)$$

To see this, define an auxiliary channel $\boldsymbol{y} = \beta\boldsymbol{u} + \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}\big(0, \sigma_z^2 I\big)$ (this channel is used only to evaluate $\mathrm{Vol}\{T_\epsilon(\boldsymbol{y}|\boldsymbol{u})\}$ and is not related to the actual distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$, see [30, p. 1262]). By tuning the parameters β and σ_z^2 such that the expectations of $\frac{1}{n}\sum_{i=1}^n y_i^2$ and $\frac{1}{n}\sum_{i=1}^n y_i u_i$ would be $\frac{1}{n}\sum_{i=1}^n \tilde{y}_i^2$ and $\frac{1}{n}\sum_{i=1}^n \tilde{y}_i u_i$, respectively, the set $T_\epsilon(\boldsymbol{y}|\boldsymbol{u})$ has a high probability under the auxiliary channel given $\boldsymbol{u}$. Moreover, any two vectors in $T_\epsilon(\boldsymbol{y}|\boldsymbol{u})$ have conditional pdf's which are exponentially equivalent. Accordingly, using the same technique as in the proof of Lemma 3 in [30, p. 1268] (which is based on these observation) we derive an upper and a lower bound on $\mathrm{Vol}\big\{T_\epsilon(\boldsymbol{y}|\boldsymbol{u})\big\}$. These bounds are

identical in the logarithmic scale, and so,

$$\mathrm{Vol}\{T_\epsilon(\boldsymbol{y}|\boldsymbol{u})\} \doteq e^{n\left[\hat{h}_{\boldsymbol{uy}}(Y|U)+\Delta(\epsilon)\right]}\ , \tag{23}$$

and $\lim_{\epsilon\to 0}\Delta(\epsilon)=0$.

Note that the order in which the limits are taken in (22) is important: We first take the dimension n to infinity, and only then we take ϵ to zero. Mathematically speaking, if ϵ goes to zero for a finite dimension n the volume of $T_\epsilon(\boldsymbol{y}|\boldsymbol{u})$ equals zero. The order of the limits has a practical meaning too. The fact that ϵ is positive for any given dimension means that the detector can calculate the correlation and energy with limited precision. In the absence of such a realistic limitation, one can offer an embedding rule (under the attack-free case and for continuous alphabet) with zero false-negative and false-positive probabilities by designing an embedder with a range having measure zero [2]. This additional limitation that we implicitly impose on the detector, is very natural and it exists in every practical system.

Using the same technique used to evaluate $\hat{h}_{\boldsymbol{uy}}(Y|U)$ in (20), it can easily be shown that

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\ln\Big[\mathrm{Vol}\{T_\epsilon(\boldsymbol{y})\}\Big]=\frac{1}{2}\ln\left(2\pi e\cdot\frac{1}{n}\sum_{i=1}^{n}y_i^2\right)\stackrel{\triangle}{=}\hat{h}_{\boldsymbol{y}}(Y)\ , \tag{24}$$

where

$$T_\epsilon(\boldsymbol{y})=\left\{\tilde{\boldsymbol{y}}\in\mathbb{R}^n:|\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}\tilde{y}_i^2|\le n\epsilon\right\}\ . \tag{25}$$

Therefore, the optimal embedder maximizes

$$\hat{I}_{\boldsymbol{uy}}(U;Y)=-\frac{1}{2}\ln\left(1-\frac{(\frac{1}{n}\sum_{i=1}^{n}u_iy_i)^2}{\frac{1}{n}\sum_{i=1}^{n}y_i^2}\right)\ . \tag{26}$$

or, equivalently, [3] maximizes

$$R(\boldsymbol{u},\boldsymbol{y})\stackrel{\triangle}{=}\frac{\langle\boldsymbol{u},\boldsymbol{y}\rangle^2}{\|\boldsymbol{y}\|^2} \tag{27}$$

subject to the distortion constraint, which in this case, will naturally be taken to be Euclidean, $\sum_{i=1}^{n}(x_i-y_i)^2\le nD_e$. 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, as is implied by the next lemma.

Lemma 1. *The optimal embedding rule under the above setting has the following form:*

$$f_n^*(\boldsymbol{x},\boldsymbol{u})=a\boldsymbol{x}+b\boldsymbol{u}. \tag{28}$$

[2] E.g., the spread-transform dither modulation (STDM) embedder proposed in [32, Sec. V.B] achieves zero false-negative probability under the attack-free scenario because the embedder range has measure zero. We thank M. Barni for drawing our attention to this fact.

[3] Note also that the corresponding detector, which compares $\hat{I}_{\boldsymbol{uy}}(U;Y)$ 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., [28]).

Proof. Clearly, every $\boldsymbol{y} \in \mathbb{R}^n$ can be represented as $\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{u} + \boldsymbol{z}$, where a and b are real valued coefficients and $\boldsymbol{z}$ is orthogonal to both $\boldsymbol{x}$ and $\boldsymbol{u}$ (i.e., $\langle \boldsymbol{u}, \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle = 0$). Now, for any given $\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{u} + \boldsymbol{z}$ such that $\boldsymbol{z} \neq 0$, the vector projected onto the subspace spanned by $\boldsymbol{x}$ and $\boldsymbol{u}$, $\tilde{\boldsymbol{y}} = a\boldsymbol{x} + b\boldsymbol{u}$, achieves a higher squared normalized correlation w.r.t. $\boldsymbol{u}$ than the vector $\boldsymbol{y}$. To see this, consider the following chain of inequalities:

$$\begin{aligned} R(\boldsymbol{u}, \boldsymbol{y}) &= \frac{\langle \boldsymbol{u}, \boldsymbol{y} \rangle^2}{\|\boldsymbol{y}\|^2} \\ &= \frac{\langle \boldsymbol{u}, a\boldsymbol{x} + b\boldsymbol{u} + \boldsymbol{z} \rangle^2}{\langle a\boldsymbol{x} + b\boldsymbol{u} + \boldsymbol{z}, a\boldsymbol{x} + b\boldsymbol{u} + \boldsymbol{z} \rangle} \\ &= \frac{\langle \boldsymbol{u}, a\boldsymbol{x} + b\boldsymbol{u} \rangle^2}{\|a\boldsymbol{x} + b\boldsymbol{u}\|^2 + \|\boldsymbol{z}\|^2} \\ &\leq R(\boldsymbol{u}, \tilde{\boldsymbol{y}}) \ . \end{aligned} \tag{29}$$

In addition, if $\boldsymbol{y}$ fulfills the distortion constraint, then so does the projected vector $\tilde{\boldsymbol{y}}$, i.e.,

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{x}\|^2 &= \|(a-1)\boldsymbol{x} + b\boldsymbol{u} + \boldsymbol{z}\|^2 \\ &= \|(a-1)\boldsymbol{x} + b\boldsymbol{u}\|^2 + \|\boldsymbol{z}\|^2 \\ &\geq \|(a-1)\boldsymbol{x} + b\boldsymbol{u}\|^2 \\ &= \|\tilde{\boldsymbol{y}} - \boldsymbol{x}\|^2 \ . \end{aligned} \tag{30}$$

Therefore, the optimal embedder must have the form $\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{u}$. In summary, given any $\boldsymbol{y}$ that satisfies the distortion constraint, by projecting $\boldsymbol{y}$ onto the subspace spanned by $\boldsymbol{x}$ and $\boldsymbol{u}$, we improve the correlation without violating the distortion constraint.

□

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 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 $\boldsymbol{u}$ (via the joint statistics $\sum_{i=1}^n x_i^2$ and $\sum_{i=1}^n u_i x_i$) rather than being fixed. Therefore, this is *not* a linear embedder.

4.1 Explicit Derivation of the Optimal Embedder

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

$$\begin{aligned} &\max \left[\frac{\left(\frac{1}{n}\sum_{i=1}^n y_i u_i\right)^2}{\frac{1}{n}\sum_{i=1}^n y_i^2} \right] . \\ \text{subject to:} \quad &\sum_{i=1}^n (y_i - x_i)^2 \leq n D_e \end{aligned} \tag{31}$$

Substituting $\boldsymbol{y} = a\boldsymbol{x} + b\boldsymbol{u}$ in eq. (31), gives:

$$\begin{aligned} &\max_{a,b\in\mathbb{R}} \left[\frac{a^2\rho^2 + 2ab\rho + b^2}{a^2\alpha^2 + 2ab\rho + b^2}\right] \\ \text{subject to:} \quad &(a-1)^2\alpha^2 + 2(a-1)b\rho + b^2 \le D \end{aligned} \tag{32}$$

where $\alpha^2 \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^n x_i^2$ and $\rho \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^n x_i u_i$. Note that $\alpha^2 \ge \rho^2$ by Cauchy-Schwarz inequality.

Theorem 2. *The optimal values of* (a, b) *are:*

- *If* $D_e \ge \alpha^2 - \rho^2$*:*

$$a^* = 0 \qquad ; \qquad b^* = \rho + \sqrt{\rho^2 - \alpha^2 + D} \tag{33}$$

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

$$\begin{aligned} a^* &= \arg\max \left\{ t(a) \,\middle|\, a \in \{a_1, a_2, a_3, a_4\} \bigcap R \right\} \\ b^* &= a^* \cdot t(a^*) \end{aligned} \tag{34}$$

where

$$\begin{aligned} t(a) &= \frac{(1-a)\rho + sgn(\rho)\sqrt{D_e - (a-1)^2(\alpha^2 - \rho^2)}}{a} \\ R &= \left[1 - \sqrt{\frac{D_e}{\alpha^2 - \rho^2}}, 1 + \sqrt{\frac{D_e}{\alpha^2 - \rho^2}}\right], \end{aligned} \tag{35}$$

$$\tag{36}$$

and

$$\begin{aligned} a_{1,2} &= \frac{(\alpha^2 - \rho^2)(\alpha^2 - D_e) \pm \sqrt{D\rho^2}\sqrt{(\alpha^2 - \rho^2)(\alpha^2 - D_e)}}{\alpha^2(\alpha^2 - \rho^2)} \\ a_{3,4} &= 1 \pm \sqrt{\frac{D_e}{\alpha^2 - \rho^2}}\,. \end{aligned} \tag{37}$$

The proof is purely technical and therefore is deferred to the Appendix. We note that in the case where $D_e \ll \alpha^2 - \rho^2$, the value of a^* tends to 1, and the value of b^* tends to $\mathrm{sgn}(\rho)\sqrt{D_e}$. Hence, the linear embedder is not optimal even in the case where $D_e \ll \alpha^2$. 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

Since the calculation of the exact false-negative exponent of the optimal optimal embedder is highly non-trivial, in this subsection we derive a lower-bound on this exponent. Later, we show that even this lower bound is by far larger than the exponent of the false-negative probability of the additive embedder. Therefore, the additive embedder is sub-optimal in terms of the exponential decay rate of its false negative probability.

The lower bound will be obtained by exploring the performance of a sub-optimal embedder of the form $\boldsymbol{y} = \boldsymbol{x} + \mathrm{sgn}(\rho)\sqrt{D_e}\boldsymbol{u}$, which we name the *sign embedder*. This embedder is obtained by setting $a = 1$ in (28)(note that 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 (26), the false-positive probability can be expressed as

$$\begin{aligned} P_{fp} = \Pr\Big\{\hat{I}_{\boldsymbol{uy}}(U;Y) > T \mid H_0\Big\} &= \Pr\Big\{\hat{\rho}^2_{\boldsymbol{uy}} > 1 - e^{-2T} \mid H_0\Big\} \\ &= 2\Pr\Big\{\hat{\rho}_{\boldsymbol{uy}} > \sqrt{1 - e^{-2T}} \mid H_0\Big\} \end{aligned}$$

where $\hat{\rho}_{\boldsymbol{uy}} = \frac{\langle \boldsymbol{u}, \boldsymbol{y}\rangle}{\|\boldsymbol{u}\| \cdot \|\boldsymbol{y}\|}$ is the normalized correlation between $\boldsymbol{u}$ and $\boldsymbol{y}$. Because under H_0 $\boldsymbol{Y} = \boldsymbol{X}$, and because of the radial symmetry of the pdf of $\boldsymbol{X}$, we can conclude that for large n [33, p. 295]:

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

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

$$\begin{aligned} -\lambda &\geq \ln(\sin\theta) \\ T &\geq -\frac{1}{2}\ln\big[1 - \cos^2\big(\arcsin(e^{-\lambda})\big)\big] = \lambda , \end{aligned} \tag{38}$$

where the last equality was obtained using the fact that $\cos\big(\arcsin(x)\big) = \sqrt{1 - x^2}$. Hence, setting $T = \lambda$ ensures a false positive probability not greater than $e^{-n\lambda}$ for large n. Define the false-negative exponent of the sign embedder

$$E^{se}_{fn} \stackrel{\triangle}{=} \lim_{n\to\infty} -\frac{1}{n}\ln P_{fn} \tag{39}$$

where the false-negative probability is given by

$$P_{fn} = \Pr\Big\{\hat{I}_{\boldsymbol{uy}}(U;Y) \leq \lambda \mid H_1\Big\} = \Pr\Big\{\hat{\rho}^2_{\boldsymbol{uy}} \leq 1 - e^{-2\lambda} \mid H_1\Big\}. \tag{40}$$

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

$$E^{se}_{fn}(\lambda, D_e) = \begin{cases} 0 & , \quad \frac{D_e e^{-2\lambda}}{1 - e^{-2\lambda}} \leq \sigma^2 \\ \frac{1}{2}\Big[\frac{D_e e^{-2\lambda}}{\sigma^2(1 - e^{-2\lambda})} - \ln\Big(\frac{D_e e^{-2\lambda}}{\sigma^2(1 - e^{-2\lambda})}\Big) - 1\Big] & , \quad else \end{cases} \tag{41}$$

The proof, which is mainly technical, is deferred to the Appendix. Let us explore some of the properties of $E^{se}_{fn}(\lambda, D_e)$. First, it is clear that $E^{se}_{fn}(0, D_e) = \infty$ (the detector output is constantly H_1) since $\hat{\rho}^2_{\boldsymbol{uy}} \geq 0$. In addition, $E^{se}_{fn}(\lambda, 0) = 0$ ($\boldsymbol{y} = \boldsymbol{x}$ and therefore does not contain any information on $\boldsymbol{u}$). For a given D_e, $E^{se}_{fn}(\lambda, D_e) = 0$ for $\lambda \geq \frac{1}{2}\ln\big(1 + \frac{D_e}{\sigma^2}\big)$.

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 2, in the case where $D_e \geq \alpha^2 - \rho^2$,

[4] It is well-known [33, p. 293] that $A_n(\theta) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n}{2}\right)} \int_0^\theta \sin^{(n-2)}(\varphi) d\varphi$ and $A_n(\pi) = 2A_n(\pi/2)$.

the optimal embedder can completely "erase" the covertext and therefore achieves a zero false negative probability. We use this property to improve the performance by introducing sub-optimum embedder which outperforms the sign embedder. Since $D_e \geq \alpha^2 \geq \alpha^2 - \rho^2$, the following embedding rule is obtained: $y = a\boldsymbol{x} + b\boldsymbol{u}$ where

$$(a,b) = \begin{cases} (0, \rho + \sqrt{\rho^2 - \alpha^2 + D_e}) & , \quad D_e \geq \alpha^2 \\ (1, \mathrm{sgn}(\rho)\sqrt{D_e}) & , \quad \text{else} \end{cases} . \tag{42}$$

This embedder, which is an improved version of the sign embedder (but still sub-optimal), erases the covertext in the cases where $D_e \geq \alpha^2$ (to keep the embedding rule a function of one parameter, we chose to "erase" the covertext only if $D_e \geq \alpha^2$). Its performance is presented in the following Corollary:

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

$$E(\lambda, D_e) = \begin{cases} 0 & , \quad D_e \leq \sigma^2 \\ \frac{1}{2}\left[\frac{D_e}{\sigma^2} - \ln\left(\frac{D_e}{\sigma^2}\right) - 1\right] & , \quad else \end{cases} ; \tag{43}$$

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

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 $D_e \geq \sigma^2$ (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 2, 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_e and σ. The point where the two graphs break apart is $\lambda = \frac{1}{2}\ln(2)$. From this point on, the improved sign embedder achieves a fixed value of $0.5(D_e/\sigma^2 - \ln(D_e/\sigma^2) - 1)$.

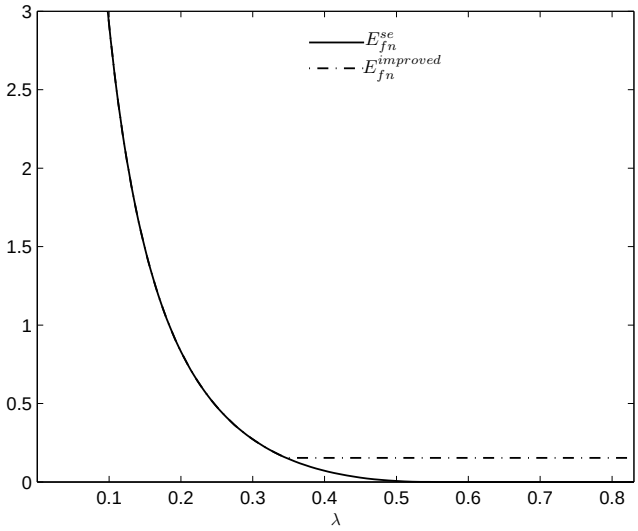

Figure 2: Error exponents of the sign embedder and its improved version for $\sigma^2 = 1$ and $D_e = 2$.

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 $\boldsymbol{y} = \boldsymbol{x} + \sqrt{D_e}\boldsymbol{u}$, 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

$$P_{fp} = \Pr\left\{\hat{\rho}_{\boldsymbol{uy}} > T \middle| H_0\right\} = \Pr\left\{\frac{\langle \boldsymbol{u}, \boldsymbol{x}\rangle}{\|\boldsymbol{u}\| \cdot \|\boldsymbol{x}\|} > T\right\} = \frac{A_n(\theta)}{A_n(\pi)} \doteq e^{n\ln(\sin\theta)}\,, \tag{44}$$

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

$$T \quad \geq \quad \cos\left[\arcsin\left(e^{-\lambda}\right)\right] = \sqrt{1 - e^{-2\lambda}}\,, \tag{45}$$

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

$$\Psi_1(r) \quad \stackrel{\triangle}{=} \quad \arccos\left[\frac{\sqrt{D_e}(T^2 - 1) + T\sqrt{r - D_e(1 - T^2)}}{\sqrt{r}}\right] \tag{46}$$

and define the false-negative exponent of the additive embedder

$$E_{fn}^{ae} \stackrel{\triangle}{=} \lim_{n\to\infty} -\frac{1}{n}\ln P_{fn}, \tag{47}$$

where the false-negative probability is given by

$$P_{fn} = \Pr\left\{\hat{\rho}_{\boldsymbol{uy}} \leq \sqrt{1 - e^{-2\lambda}} \middle| H_1\right\}. \tag{48}$$

Theorem 4. *The false negative exponent of the additive embedder is given by*

$$E_{fn}^{ae}(\lambda, D_e) = \min\left\{E_1(\lambda, D_e), E_2(\lambda, D_e)\right\} \tag{49}$$

where,

$$\begin{aligned} E_1(\lambda, D_e) \quad &= \quad \min_{D_e e^{-2\lambda} < r \leq \frac{D_e e^{-2\lambda}}{1 - e^{-2\lambda}}} \frac{1}{2}\left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 2\ln\sin\left(\Psi_1(r)\right) - 1\right] \\ E_2(\lambda, D_e) \quad &= \quad \begin{cases} 0 & , \quad \frac{D_e e^{-2\lambda}}{1 - e^{-2\lambda}} \leq \sigma^2 \\ \frac{1}{2}\left[\frac{D_e e^{-2\lambda}}{(1 - e^{-2\lambda})\sigma^2} - \ln\left(\frac{D_e e^{-2\lambda}}{(1 - e^{-2\lambda})\sigma^2}\right) - 1\right] & , \quad \textit{else} \end{cases} \end{aligned} \tag{50}$$

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

Let us examine some of the properties of $E_{fn}^{ae}(\lambda, D_e)$. It is easy to see that $E_{fn}^{ae}(\lambda, D_e) \leq E_2(\lambda, D_e) = E_{fn}^{se}(\lambda, D_e)$, 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 $E_{fn}^{ae}(0, D_e) < \infty$. To see why this is the case let us look at

$$E_1(0, D_e) \quad = \quad \min_{r > D_e} f(r) \tag{51}$$

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 $r > D_e$, 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 $\boldsymbol{u}$ is not zero. Clearly, for a given D_e, $E_{fn}^{ae}(\lambda, D_e) = 0$ for $\lambda \geq \frac{1}{2}\ln\left(1 + \frac{D_e}{\sigma^2}\right)$. Numerical calculations show that this happens even for smaller values of λ, however, the exact smallest value of λ for which $E_{fn}^{ae}(\lambda, D_e) = 0$ is hard to find. In Figures 3, 4 and 5 we compare the two embedding strategies by plotting their exponents as a functions of σ^2/D_e.

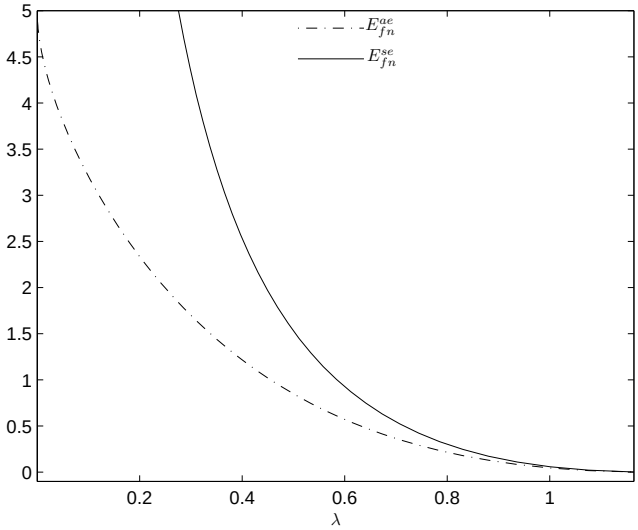

Figure 3: Error exponents of the two embedding strategies ($\sigma^2/D = .1$)

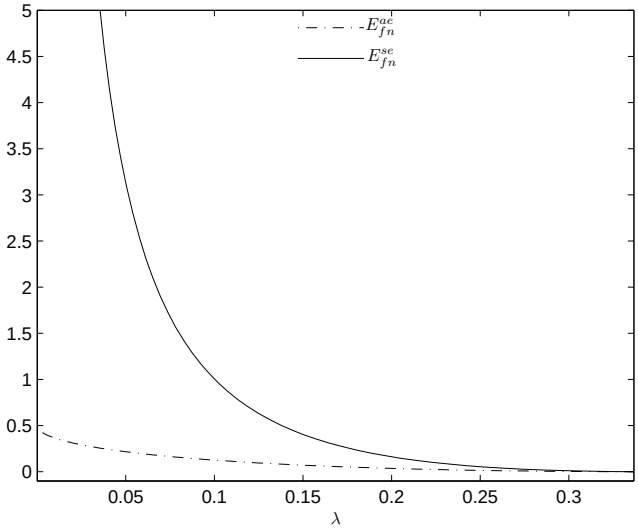

Figure 4: Error exponents of the two embedding strategies ($\sigma^2/D = 1$)

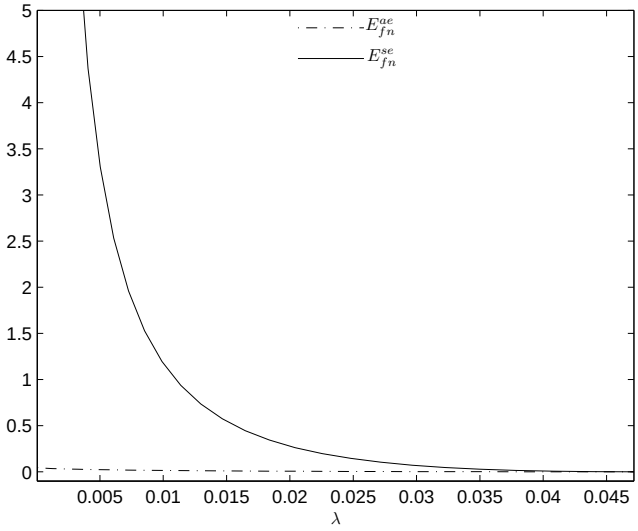

Figure 5: Error exponents of the two embedding strategies ($\sigma^2/D = 10$)

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 $\boldsymbol{u}$ is non-negative, the additive embedder and the sign embedder achieve the same performance. However, when the correlation between $\boldsymbol{x}$ and $\boldsymbol{u}$ 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 $\boldsymbol{u}$ (while the detector compares the normalized correlation $\hat{\rho}_{\boldsymbol{yu}}$ between $\boldsymbol{y}$ and $\boldsymbol{u}$ to a given threshold), however, these 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 $\lambda = 0$. In this case, the sign embedder achieves $E_{fn}^{se}(0, D_e) = \infty$ while $E_{fn}^{ae}(0, D_e)$ is finite since the probability of embedded vectors $\boldsymbol{Y}$ for which $\hat{\rho}_{\boldsymbol{yu}} < 0$ 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 $\mathrm{sgn}(\rho)$.

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. In Section 6, we will discuss general worst-case 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 $W_n(\boldsymbol{z}|\boldsymbol{y})$ as is shown in Fig. 1. For convenience, we will assume that the components of $\boldsymbol{z}$ take on values in the same alphabet $\mathcal{A}$, which will be assumed again to be finite, as in Sections 2 and 3. Thus,

the operation of the attack, which in general may be stochastic, is thought of as a channel. Denoting the channel output marginal by $Q(\boldsymbol{z}) = \sum_{\boldsymbol{y}} P_X(\boldsymbol{y}) W_n(\boldsymbol{z}|\boldsymbol{y})$, the analysis of this case is, in principle, the same as before.

Assuming, for example, that Q is memoryless (which is the case when both P_X and W_n are memoryless, i.e., $W_n(\boldsymbol{z}|\boldsymbol{y}) = \prod_{i=1}^n W(z_i|y_i)$ for some discrete memoryless channel $W : \mathcal{A} \to \mathcal{A}$), then Λ_* is as in Section 2, except that P_X, Y, and $\boldsymbol{y}$ should be replaced by Q, Z and $\boldsymbol{z}$, respectively. The optimal embedder then becomes

$$f_n^*(\boldsymbol{x}, \boldsymbol{u}) = \mathrm{argmin}_{\{\boldsymbol{y}:\ d_e(\boldsymbol{x},\boldsymbol{y}) \le nD_e\}} \sum_{\boldsymbol{z} \in \Lambda_*^c} W_n(\boldsymbol{z}|\boldsymbol{y}), \tag{52}$$

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

$$\begin{aligned} \Lambda_*^c &= \Big\{\boldsymbol{z} : \ln Q(\boldsymbol{z}) + n\hat{H}_{\boldsymbol{zu}}(Z|U) + n\lambda - |\mathcal{A}| \ln(n+1) > 0\Big\} &(53)\\ &= \Big\{\boldsymbol{z} : -n\hat{I}_{\boldsymbol{zu}}(Z;U) - n\mathcal{D}\big(\hat{P}_{\boldsymbol{z}} \| Q\big) + n\lambda - |\mathcal{A}| \ln(n+1) > 0\Big\}\ , &(54)\end{aligned}$$

where $\hat{P}_{\boldsymbol{z}}$ is the empirical distribution of $\boldsymbol{z}$. Evidently, eq. (52) is not a convenient formula to work with. Therefore, let us try to simplify (52). For a given $\boldsymbol{y}$, let us rewrite (52) as follows:

$$\begin{aligned} \sum_{\boldsymbol{z} \in \Lambda_*^c} W_n(\boldsymbol{z}|\boldsymbol{y}) &= \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u}) \subseteq \Lambda_*^c} \ \sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})} W_n(\boldsymbol{z}'|\boldsymbol{y}) \\ &= \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u}) \subseteq \Lambda_*^c} \big|T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})\big| W_n(\boldsymbol{z}|\boldsymbol{y})\ . &(55)\end{aligned}$$

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

$$W_n(\boldsymbol{z}'|\boldsymbol{y}) = e^{-n\left[\hat{H}_{\boldsymbol{yz}}(Z|Y) + \sum_{a \in \mathcal{A}} \hat{P}_{\boldsymbol{y}}(a) \mathcal{D}\left(\hat{P}_{\boldsymbol{yz}}(Z|Y=a) \| W(Z|Y=a)\right)\right]}. \tag{56}$$

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

$$|T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})| \doteq e^{n\hat{H}_{\boldsymbol{uyz}}(Z|Y,U)}, \tag{57}$$

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

$$\begin{aligned} \hat{P}^*_{\boldsymbol{uxy}}(Y|X,U) = \arg \max_{\substack{\hat{P}_{\boldsymbol{uxy}}(Y|X,U):\\ \hat{E}_{\boldsymbol{xy}} d_e(X,Y) \le D_e}} \Bigg\{ \min_{\substack{\hat{P}_{\boldsymbol{uyz}}(Z|Y,U):\\ \hat{I}_{\boldsymbol{uz}}(Z;U) + \mathcal{D}(\hat{P}_{\boldsymbol{z}} \| Q) \le \lambda}} \Big[\hat{I}_{\boldsymbol{uyz}}(Z;U|Y) \\ + \sum_{a \in \mathcal{A}} \hat{P}_{\boldsymbol{y}}(a) \mathcal{D}\big(\hat{P}_{\boldsymbol{yz}}(Z|Y=a) \big\| W(Z|Y=a)\big)\Big]\Bigg\} \end{aligned} \tag{58}$$

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

6 General Attack Channel

In this section we extend the results of the previous sections to include general attack channels subject to a distortion criterion.

Consider a covertext sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ emitted from a memoryless source P_X as before. Let $d_a : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_+$ denote another bounded single-letter distortion measure. An attacker subject to distortion level D_a w.r.t. d_a is a channel W_n, fed by a stegotext $\boldsymbol{y}$ and which produces a forgery $\boldsymbol{z}$ such that

$$d_a(\boldsymbol{y}, \boldsymbol{z}) \stackrel{\triangle}{=} \sum_{i=1}^{n} d_a(y_i, z_i) \leq nD_a \quad \forall (\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{A} \times \mathcal{A}. \tag{59}$$

We denote the set of attack channels which satisfy (59) by $\mathcal{W}_n(D_a)$.

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

$$P_{fn} \stackrel{\triangle}{=} \max_{W_n \in \mathcal{W}_n(D_a)} P_{fn}\big(f_n, \Lambda, W_n\big) \tag{60}$$

where

$$P_{fn}\big(f_n, \Lambda, W_n\big) \stackrel{\triangle}{=} \sum_{\boldsymbol{z} \in \Lambda^c} \left[\sum_{\boldsymbol{y} \in \mathcal{A}^n} \left(\sum_{\boldsymbol{x}: f_n(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{y}} P_X(\boldsymbol{x}) \right) W_n(\boldsymbol{z}|\boldsymbol{y}) \right] , \tag{61}$$

and $P_X(\boldsymbol{x}) = \prod_{i=1}^{n} P_X(x_i)$, subject to the following constraints:

(1) The distortion between $\boldsymbol{x}$ and $\boldsymbol{y}$ does not exceed nD_e.

(2) The false positive probability is upper bounded by

$$P_{fp} \stackrel{\triangle}{=} \max_{W_n \in \mathcal{W}_n(D_a)} P_{fp}\big(\Lambda, W_n\big) \leq e^{-n\lambda} , \tag{62}$$

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

$$P_{fp}\big(\Lambda, W_n\big) \stackrel{\triangle}{=} \sum_{\boldsymbol{z} \in \Lambda} \left(\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) W_n(\boldsymbol{z}|\boldsymbol{y}) \right) . \tag{63}$$

In other words, we would like to choose an embedder f_n and a decision region Λ so as to minimize P_{fn} subject to a distortion constraint (between the covertext and the stegotext) and the constraint that the exponential decay rate of P_{fp} would be at least as large as λ, for *any* attack channel in $\mathcal{W}_n(D_a)$.

Similarly as in Section 2, we focus on the class of detectors which base their decisions on the empirical joint distribution of $\boldsymbol{z}$ and $\boldsymbol{u}$.

6.1 Strongly Exchangeable Attack Channels

First, we restrict the set of attack channels to be strongly exchangeable channels (the exact definition will be given in the sequel). Later, this restriction will be dropped, and the attack channel will be allowed to

be any member of $\mathcal{W}_n(D_a)$. However, in this case random watermarks (rather than deterministic ones) must be considered.

The use of strongly exchangeable channels in the context of general attack channels was proposed in [22], where Somekh-Baruch and Merhav showed (in another context) that the worst strongly exchangeable attack channel is as bad as the worst general attack channel, while strongly exchangeable channels are much easier to analyze. In the sequel, we will adjust the proof technique proposed in [22] to fit our needs.

Definition 1. *A strongly exchangeable channel W_n is one that satisfies for all $\boldsymbol{y} \in \mathcal{A}^n, \boldsymbol{z} \in \mathcal{A}^n$*

$$W_n(\boldsymbol{z}'|\boldsymbol{y}') = W_n(\boldsymbol{z}|\boldsymbol{y}), \qquad \forall (\boldsymbol{y}', \boldsymbol{z}') \in T(\boldsymbol{y}, \boldsymbol{z}) \ .$$

Denote the set of all strongly exchangeable channels that operate on n-tuples by $\mathcal{C}_n^{ex}$ and let $\mathcal{W}_n^{ex}(D_a) = \mathcal{W}_n(D_a) \cap \mathcal{C}_n^{ex}$.

Define

$$W_n^*(\boldsymbol{z}|\boldsymbol{y}) = \frac{c_n(\boldsymbol{y})}{|T(\boldsymbol{z}|\boldsymbol{y})|} \mathbb{1}\left\{d_a(\boldsymbol{y}, \boldsymbol{z}) \le nD_a\right\} \ , \tag{64}$$

where, $c_n(\boldsymbol{y}) = \left[\sum_{\boldsymbol{z}: d_a(\boldsymbol{y},\boldsymbol{z}) \le nD_a} \frac{1}{|T(\boldsymbol{z}|\boldsymbol{y})|}\right]^{-1}$ [22, p. 543]. Clearly, $W_n^* \in \mathcal{W}_n^{ex}(D_a)$. Note that $c_n(\boldsymbol{y})$ equals to the reciprocal of the number of conditional types $T(\boldsymbol{z}|\boldsymbol{y})$ such that $d_a(\boldsymbol{y}, \boldsymbol{z}) \le nD_a$ [22, p. 543] which implies that $(n+1)^{-\mathcal{A}^2} \le c(\boldsymbol{y}) \le 1$. Hence, $c_n(\boldsymbol{y})$ is at most polynomial in n.

Define

$$\Lambda_* = \left\{ \boldsymbol{z} \ : \ \hat{I}_{\boldsymbol{z}\boldsymbol{u}}(Z;U) + \min_{\hat{P}_{\boldsymbol{y}}: \hat{E}_{\boldsymbol{y}\boldsymbol{z}} d_a(Y,Z) \le D_a} \mathcal{D}\big(\hat{P}_{\boldsymbol{y}} \| P_X\big) \ge \frac{|\mathcal{A}| \ln(n+1)}{n} + \lambda \right\} \ . \tag{65}$$

Lemma 2. *(i) For every $W_n \in \mathcal{W}_n^{ex}(D_a)$,*

$$P_{fp}(\Lambda_*, W_n) \le e^{-n(\lambda - \delta_n)}$$

where $\lim_{n\to\infty} \delta_n = 0$.

(ii) For any $\Lambda \subseteq \mathcal{A}^n$ that satisfies

$$P_{fp}(\Lambda, W_n) \le e^{-n\lambda'} \qquad \forall W_n \in \mathcal{W}_n^{ex}(D_a)$$

for some $\lambda' > \lambda$, then $\Lambda_*^c \subseteq \Lambda^c$ for all sufficiently large n.

Proof. Let $T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda$. Then, we have

$$\begin{aligned}
e^{-n\lambda} &\ge \max_{W_n \in \mathcal{W}_n^{ex}(D_a)} P_{fp}(\Lambda, W_n) \\
&= \max_{W_n \in \mathcal{W}_n^{ex}(D_a)} \sum_{\boldsymbol{z} \in \Lambda} \left(\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) W_n(\boldsymbol{z}|\boldsymbol{y}) \right) \\
&\ge \sum_{\boldsymbol{z} \in \Lambda} \left(\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) W_n^*(\boldsymbol{z}|\boldsymbol{y}) \right) \\
&= \sum_{T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda} \sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{u})} \left(\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) W_n^*(\boldsymbol{z}'|\boldsymbol{y}) \right) \\
&= \sum_{T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda} \sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{u})} Q^*(\boldsymbol{z}') \ ,
\end{aligned} \tag{66}$$

where $Q^*(\boldsymbol{z}) \stackrel{\triangle}{=} \sum_{\boldsymbol{y}\in\mathcal{A}^n} P_X(\boldsymbol{y})W_n^*(\boldsymbol{z}|\boldsymbol{y})$. Now,

$$\begin{aligned}
Q^*(\boldsymbol{z}) &= \sum_{\boldsymbol{y}\in\mathcal{A}^n} P_X(\boldsymbol{y})W_n^*(\boldsymbol{z}|\boldsymbol{y}) \\
&= \sum_{T(\boldsymbol{y}|\boldsymbol{z})\subset\mathcal{A}^n} \sum_{\boldsymbol{y}'\in T(\boldsymbol{y}|\boldsymbol{z})} P_X(\boldsymbol{y}')W_n^*(\boldsymbol{z}|\boldsymbol{y}') \\
&= \sum_{T(\boldsymbol{y}|\boldsymbol{z})\subset\mathcal{A}^n} \sum_{\boldsymbol{y}'\in T(\boldsymbol{y}|\boldsymbol{z})} P_X(\boldsymbol{y}')\frac{c_n(\boldsymbol{y}')}{|T(\boldsymbol{z}|\boldsymbol{y})|}\mathbb{1}\{d_a(\boldsymbol{y}',\boldsymbol{z})\le nD_a\} \\
&\doteq \sum_{T(\boldsymbol{y}|\boldsymbol{z})\subset\mathcal{A}^n} |T(\boldsymbol{y}|\boldsymbol{z})|e^{-n\left[\hat{H}\boldsymbol{y}(Y)+\mathcal{D}(\hat{P}\boldsymbol{y}\|P_X)\right]}e^{-n\hat{H}\boldsymbol{y}\boldsymbol{z}(Z|Y)}c_n(\boldsymbol{y})\mathbb{1}\{d_a(\boldsymbol{y},\boldsymbol{z})\le nD_a\} \\
&\doteq \sum_{T(\boldsymbol{y}|\boldsymbol{z})\subset\mathcal{A}^n} e^{-n\left[\hat{H}\boldsymbol{y}(Y)+\mathcal{D}(\hat{P}\boldsymbol{y}\|P_X)-\hat{H}\boldsymbol{y}\boldsymbol{z}(Y|Z)+\hat{H}\boldsymbol{y}\boldsymbol{z}(Z|Y)\right]}c_n(\boldsymbol{y})\mathbb{1}\{d_a(\boldsymbol{y},\boldsymbol{z})\le nD_a\} \\
&\doteq \exp\left\{-n\left[\hat{H}\boldsymbol{z}(Z)+\min_{\hat{P}\boldsymbol{y}:\hat{E}\boldsymbol{y}\boldsymbol{z}^{d_a}(Y,Z)\le D_a}\mathcal{D}(\hat{P}\boldsymbol{y}\|P_X)\right]\right\}, \qquad (67)
\end{aligned}$$

where the last equality stems from the fact that $c_n(\boldsymbol{y})$ is polynomial in n.

Clearly, for any $\boldsymbol{z}' \in T(\boldsymbol{z})$ the following holds

$$\begin{aligned}
Q(\boldsymbol{z}) &= \sum_{\boldsymbol{y}\in\mathcal{A}^n} P_X(\boldsymbol{y})W_n(\boldsymbol{z}|\boldsymbol{y}) \\
&= \sum_{\pi(\boldsymbol{y})} P_X(\pi(\boldsymbol{y}))W_n(\pi(\boldsymbol{z})|\pi(\boldsymbol{y})) \\
&= Q(\pi(\boldsymbol{z})) \\
&= Q(\boldsymbol{z}'), \qquad (68)
\end{aligned}$$

where the second equality is because $W_n \in \mathcal{W}_n^{ex}(D_a)$ and $\pi(\cdot)$ is a permutation of $\{1,\ldots,n\}$ such that $\boldsymbol{z}' = \pi(\boldsymbol{z})$. Hence $Q^*(\boldsymbol{z}') = Q^*(\boldsymbol{z})\ \ \forall \boldsymbol{z}' \in T(\boldsymbol{z})$. Following (66), we get

$$\begin{aligned}
e^{-n\lambda} &\ge \sum_{T(\boldsymbol{z}|\boldsymbol{u})\subseteq\Lambda} |T(\boldsymbol{z}|\boldsymbol{u})|Q^*(\boldsymbol{z}) \\
&\ge |T(\boldsymbol{z}|\boldsymbol{u})|Q^*(\boldsymbol{z}) \\
&\ge |T(\boldsymbol{z}|\boldsymbol{u})|\exp\left\{-n\left[\hat{H}\boldsymbol{z}(Z)+\min_{\hat{P}\boldsymbol{y}:\hat{E}\boldsymbol{y}\boldsymbol{z}^{d_a}(Y,Z)\le D_a}\mathcal{D}(\hat{P}\boldsymbol{y}\|P_X)\right]\right\} \\
&\ge \exp\left\{-n\left[\hat{H}\boldsymbol{z}(Z)-n\hat{H}\boldsymbol{z}\boldsymbol{u}(Z|U)+\min_{\hat{P}\boldsymbol{y}:\hat{E}\boldsymbol{y}\boldsymbol{z}^{d_a}(Y,Z)\le D_a}\mathcal{D}(\hat{P}\boldsymbol{y}\|P_X)\right]\right\}(n+1)^{-|\mathcal{A}|} \\
&= \exp\left\{-n\left[\hat{I}\boldsymbol{z}\boldsymbol{u}(Z;U)+\min_{\hat{P}\boldsymbol{y}:\hat{E}\boldsymbol{y}\boldsymbol{z}^{d_a}(Y,Z)\le D_a}\mathcal{D}(\hat{P}\boldsymbol{y}\|P_X)\right]\right\}(n+1)^{-|\mathcal{A}|}. \qquad (69)
\end{aligned}$$

In the same spirit as in the attack-free scenario, we have shown that every $T(\boldsymbol{z}|\boldsymbol{u})$ in Λ is also in Λ_*. Therefore, $\Lambda_*^c \subseteq \Lambda^c$ 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 (62). It remains to show that Λ_* itself has a false positive exponent which is at least as large as λ for sufficiently large n.

Clearly, for any attack channel $W_n \in \mathcal{W}_n^{ex}(D_a)$,

$$\begin{aligned} W_n(\boldsymbol{z}|\boldsymbol{y}) &= \frac{\sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{y})} W_n(\boldsymbol{z}'|\boldsymbol{y})}{|T(\boldsymbol{z}|\boldsymbol{y})|} \\ &= \frac{W\big(T(\boldsymbol{z}|\boldsymbol{y})|\boldsymbol{y}\big)}{|T(\boldsymbol{z}|\boldsymbol{y})|} \\ &\leq \frac{1}{|T(\boldsymbol{z}|\boldsymbol{y})|} \mathbb{1}\{d_a(\boldsymbol{y},\boldsymbol{z}) \leq nD_a\} \,, \end{aligned} \tag{70}$$

where the first equality is because $W_n(\boldsymbol{z}'|\boldsymbol{y}) = W_n(\boldsymbol{z}|\boldsymbol{y}) \ \ \forall \boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{y})$. Moreover, similarly as in (67), combined with the fact that $c(\boldsymbol{y})$ is polynomial in n implies that

$$\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) \frac{\mathbb{1}\big\{d_a(\boldsymbol{y},\boldsymbol{z}') \leq nD_a\big\}}{|T(\boldsymbol{z}'|\boldsymbol{y})|} \doteq \exp\left\{-n\left[\hat{H}_{\boldsymbol{z}}(Z) + \min_{\hat{P}_{\boldsymbol{y}}:\hat{E}_{\boldsymbol{y}\boldsymbol{z}}\, d_a(Y,Z) \leq D_a} \mathcal{D}(\hat{P}_{\boldsymbol{y}} \| P_X)\right]\right\} . \tag{71}$$

Using (70) and (71), it follows that Λ_* indeed fulfills the false-positive constraint for any attack channel $W_n \in \mathcal{W}_n^{ex}(D_a)$:

$$\begin{aligned} \max_{W_n \in \mathcal{W}_n^{ex}(D_a)} P_{fp}(\Lambda_*, W_n) &= \max_{W_n \in \mathcal{W}_n^{ex}(D_a)} \sum_{\boldsymbol{z} \in \Lambda_*} \left(\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) W_n(\boldsymbol{z}|\boldsymbol{y}) \right) \\ &\leq \sum_{T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda_*} \sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{u})} \left(\sum_{\boldsymbol{y} \in \mathcal{A}^n} P_X(\boldsymbol{y}) \frac{\mathbb{1}\big\{d_a(\boldsymbol{y},\boldsymbol{z}') \leq nD_a\big\}}{|T(\boldsymbol{z}'|\boldsymbol{y})|} \right) \\ &= \sum_{T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda_*} \sum_{\boldsymbol{z}' \in T(\boldsymbol{z}|\boldsymbol{u})} \left[\exp\left\{ -n \left(\hat{H}_{\boldsymbol{z}}(Z) + \min_{\hat{P}_{\boldsymbol{y}}:\hat{E}_{\boldsymbol{y}\boldsymbol{z}}\, d_a(Y,Z) \leq D_a} \mathcal{D}(\hat{P}_{\boldsymbol{y}} \| P_X) \right) \right\} \right] \\ &= \sum_{T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda_*} e^{n\hat{H}_{\boldsymbol{u}\boldsymbol{z}}(Z|U)} \left[\exp\left\{ -n \left(\hat{H}_{\boldsymbol{z}}(Z) + \min_{\hat{P}_{\boldsymbol{y}}:\hat{E}_{\boldsymbol{y}\boldsymbol{z}}\, d_a(Y,Z) \leq D_a} \mathcal{D}(\hat{P}_{\boldsymbol{y}} \| P_X) \right) \right\} \right] \\ &\dot{\leq} \sum_{T(\boldsymbol{z}|\boldsymbol{u}) \subseteq \Lambda_*} \exp\left\{ -n\hat{I}_{\boldsymbol{u}\boldsymbol{z}}(Z;U) \right\} \exp\left\{ -n \min_{\hat{P}_{\boldsymbol{y}}:\hat{E}_{\boldsymbol{y}\boldsymbol{z}}\, d_a(Y,Z) \leq D_a} \mathcal{D}(\hat{P}_{\boldsymbol{y}} \| P_X) \right\} \\ &\leq (n+1)^{|\mathcal{A}|} e^{-n\lambda} \\ &\doteq e^{-n(\lambda - \delta_n)} \,, \end{aligned} \tag{72}$$

where $\delta_n = \frac{|\mathcal{A}| \ln(n+1)}{n} \to 0$ as $n \to \infty$. □

Our next step is to find an embedder which minimizes the probability of false negative under the given decision region for any attack channels $W_n \in \mathcal{W}_n^{ex}(D_a)$. Following Section 5, the optimal embedder can be written as follows:

$$f_n^*(\boldsymbol{x},\boldsymbol{u}) = \arg \min_{\boldsymbol{y}:d_e(\boldsymbol{x},\boldsymbol{y}) \leq nD_e} \max_{W_n \in \mathcal{W}_n^{ex}(D_a)} \sum_{\boldsymbol{z} \in \Lambda_*^c} W_n(\boldsymbol{z}|\boldsymbol{y}) \,. \tag{73}$$

Lemma 3. *For any attack channel $W_n \in \mathcal{W}_n^{ex}(D_a)$, the optimal embedder f_n^* which minimizes the false-negative probability can be expressed in the following manner:*

$$f_n^*(\boldsymbol{x},\boldsymbol{u}) = \boldsymbol{y}, \qquad \boldsymbol{y} \in T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u}) \tag{74}$$

where $T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$ corresponds to the following conditional empirical distribution:

$$\hat{P}_{\boldsymbol{uxy}}(Y|X,U) = \arg \max_{\substack{\hat{P}_{\boldsymbol{uxy}}(Y|X,U):\\ \hat{E}_{\boldsymbol{xy}}d_e(X,Y)\le D_e}} \left\{ \min_{\substack{\hat{P}_{\boldsymbol{uyz}}(Z|Y,U):\\ \hat{I}_{\boldsymbol{uz}}(Z;U)+\min_{\hat{P}_{\boldsymbol{y}}(Y):\hat{E}_{\boldsymbol{y}}d_a(Y,Z)\le D_a} D(\hat{P}_{\boldsymbol{y}}\|P_X)<\lambda}} \hat{I}_{\boldsymbol{uyz}}(Z;U|Y) \right\}. \tag{75}$$

Proof. For a given $\boldsymbol{y}\in\mathcal{A}^n$,

$$\begin{aligned}
\max_{W_n\in\mathcal{W}_n^{ex}(D_a)} \sum_{\boldsymbol{z}\in\Lambda_*^c} W_n(\boldsymbol{z}|\boldsymbol{y}) &= \max_{W_n\in\mathcal{W}_n^{ex}(D_a)} \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})\subseteq\Lambda_*^c} \sum_{\boldsymbol{z}'\in T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})} W_n(\boldsymbol{z}|\boldsymbol{y}) \\
&\le \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})\subseteq\Lambda_*^c} \sum_{\boldsymbol{z}'\in T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})} |T(\boldsymbol{z}|\boldsymbol{y})|^{-1}\mathbb{1}\{d_a(\boldsymbol{y},\boldsymbol{z}')\le nD_a\} \\
&\le \sum_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})\subseteq\Lambda_*^c} |T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})|\cdot|T(\boldsymbol{z}|\boldsymbol{y})|^{-1}\mathbb{1}\{d_a(\boldsymbol{y},\boldsymbol{z}')\le nD_a\} \\
&\doteq \max_{T(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{u})\subseteq\Lambda_*^c} e^{-n\hat{I}_{\boldsymbol{uyz}}(Z;U|Y)}.
\end{aligned} \tag{76}$$

Therefore $f_n^*(\boldsymbol{x},\boldsymbol{u})\in T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$, where $T^*(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$ corresponds to the conditional empirical distribution (75). □

Note that the optimal embedder and the optimal decision rule correspond to the case where the detector and the embedder are tuned to the worst possible channel W_n^*. To extend the above results to general attack channels (i.e., channels that are members of $\mathcal{W}_n(D_a)$ rather than $\mathcal{W}_n^{ex}(D_a)$) we must consider the random watermark setting (cf. Subsection 3.4). The reason for this will be made clear in the sequel.

6.2 Random Watermarks and General Attack Channels

In the spirit of Subsection 3.4, from this point on, we will use the model in which $\boldsymbol{u}$ is random as well, in particular, being drawn from another source P_U, independently of $\boldsymbol{x}$, normally, the binary symmetric source (BSS). In this case, the decision regions Λ and Λ^c will be defined as subsets of $\mathcal{A}^n\times\mathcal{B}^n$ and the probabilities of error P_{fn} and P_{fp} will be defined, again, as the corresponding summations of products $P_X(\boldsymbol{x})P_U(\boldsymbol{u})$.

The corresponding version of Λ_*, proposed for strongly exchangeable attack, channels would be:

$$\Lambda_{**}\stackrel{\triangle}{=}\left\{(\boldsymbol{z},\boldsymbol{u}):\ \hat{I}_{\boldsymbol{zu}}(Z;U)+\mathcal{D}\big(\hat{P}_{\boldsymbol{u}}\|P_U\big)+\min_{\hat{P}_{\boldsymbol{y}}:\hat{E}_{\boldsymbol{yz}}d_a(Y,Z)\le D_a}\mathcal{D}\big(\hat{P}_{\boldsymbol{y}}\|P_X\big)\ge\frac{|\mathcal{A}|\ln(n+1)}{n}+\lambda\right\}. \tag{77}$$

Theorem 5. *(i) For every $W_n\in\mathcal{W}_n(D_a)$,*

$$P_{fp}(\Lambda_{**},W_n)\le e^{-n(\lambda-\delta_n)},$$

where $\lim_{n\to\infty}\delta_n=0$.

(ii) For any $\Lambda\subseteq\mathcal{A}^n\times\mathcal{B}^n$ that satisfies

$$P_{fp}(\Lambda,W_n)\le e^{-n\lambda'}\qquad\forall W_n\in\mathcal{W}_n(D_a)$$

for some $\lambda'>\lambda$, then $\Lambda_{**}^c\subseteq\Lambda^c$ for all sufficiently large n.

To prove the above theorem in the case of general attack channels, we first need to ensure that the probability of false positive under Λ_{**} will be smaller than $e^{-n\lambda}$ for any attack channel in$\mathcal{W}_n(D_a)$. We use an argument, which was used in [22, Lemma 4], to prove that the worst strongly exchangeable attack channel is as bad as the worst general channel, and therefore we can reuse the results of Lemma 2. For the sake of completeness, we will rephrase the argument and adjust it to our problem.

Proof. Given a general attack channel $W_n \in \mathcal{W}_n(D_a)$, let π denote a permutation of $\{1,\ldots,n\}$ and let $W_n^\pi(\boldsymbol{z}|\boldsymbol{y}) \stackrel{\triangle}{=} W_n\big(\pi(\boldsymbol{z})|\pi(\boldsymbol{y})\big)$. Clearly,

$$\tilde{W}_n(\boldsymbol{z}|\boldsymbol{y}) = \frac{1}{n!}\sum_\pi W_n^\pi(\boldsymbol{z}|\boldsymbol{y})$$

is a strongly exchangeable channel. For a given $W_n \in \mathcal{W}_n(D_a)$, let the false-positive probability under Λ be

$$P_{fp}(\Lambda, W_n) \stackrel{\triangle}{=} \sum_{\boldsymbol{u}} P_U(\boldsymbol{u}) \sum_{\boldsymbol{z}\in\Lambda(\boldsymbol{u})} \left(\sum_{\boldsymbol{y}\in\mathcal{A}^n} P_X(\boldsymbol{y}) W_n(\boldsymbol{z}|\boldsymbol{y}) \right) , \tag{78}$$

where $\Lambda(\boldsymbol{u}) = \big\{\boldsymbol{z} : (\boldsymbol{z},\boldsymbol{u}) \in \Lambda\big\}$. Recall that any decision region Λ is a union of joint type classes $\big\{T(\boldsymbol{u},\boldsymbol{z})\big\}$. Since $P_{fp}(\Lambda, W_n)$ is affine in W_n, we can see that

$$\begin{aligned}
\frac{1}{n!}\sum_\pi P_{fp}(\Lambda, W_n^\pi) &= \frac{1}{n!}\sum_\pi\sum_{\boldsymbol{u}} P_U(\boldsymbol{u}) \sum_{\boldsymbol{z}\in\Lambda(\boldsymbol{u})} \left(\sum_{\boldsymbol{y}} P_X(\boldsymbol{y}) W_n^\pi(\boldsymbol{z}|\boldsymbol{y}) \right) \\
&= \sum_{\boldsymbol{u}} P_U(\boldsymbol{u}) \sum_{\boldsymbol{z}\in\Lambda(\boldsymbol{u})} \left[\sum_{\boldsymbol{y}} P_X(\boldsymbol{y}) \left(\frac{1}{n!}\sum_\pi W_n^\pi(\boldsymbol{z}|\boldsymbol{y}) \right) \right] \\
&= P_{fp}\left(\Lambda, \frac{1}{n!}\sum_\pi W_n^\pi \right) .
\end{aligned} \tag{79}$$

Now, for a given permutation π,

$$\begin{aligned}
P_{fp}(\Lambda, W_n^\pi) &= \sum_{\boldsymbol{u}} P_U(\boldsymbol{u}) \sum_{\boldsymbol{z}\in\Lambda(\boldsymbol{u})} \left(\sum_{\boldsymbol{y}} P_X(\boldsymbol{y}) W_n\big(\pi(\boldsymbol{z})|\pi(\boldsymbol{y})\big) \right) \\
&= \sum_{\boldsymbol{u}} P_U\big(\pi(\boldsymbol{u})\big) \sum_{\boldsymbol{z}\in\Lambda(\pi(\boldsymbol{u}))} \left(\sum_{\boldsymbol{y}} P_X\big(\pi(\boldsymbol{y})\big) W_n\big(\pi(\boldsymbol{z})|\pi(\boldsymbol{y})\big) \right) \\
&= \sum_{\boldsymbol{u}} P_U(\boldsymbol{u}) \sum_{\boldsymbol{z}\in\Lambda(\boldsymbol{u})} \left(\sum_{\boldsymbol{y}} P_X(\boldsymbol{y}) W_n(\boldsymbol{z}|\boldsymbol{y}) \right) \\
&= P_{fp}(\Lambda, W_n)
\end{aligned} \tag{80}$$

where the second equality follows since $\boldsymbol{z} \in \Lambda(\boldsymbol{u}) \Rightarrow \pi(\boldsymbol{z}) \in \Lambda(\pi(\boldsymbol{u}))$ (and that is because Λ is a union of joint type classes $\{T(\boldsymbol{u},\boldsymbol{z})\}$) and the third equality follows from the fact that P_X and P_U are memoryless which implies that $P_X\big(\pi(\boldsymbol{y})\big) = P_X(\boldsymbol{y})$ and $P_U\big(\pi(\boldsymbol{y})\big) = P_U(\boldsymbol{y})$.

From (79) and (80), we get that for any Λ,

$$\begin{aligned} P_{fp}\left(\Lambda, \frac{1}{n!}\sum_{\pi} W_n^{\pi}\right) &= \frac{1}{n!}\sum_{\pi} P_{fp}(\Lambda, W_n^{\pi}) \\ &= \frac{1}{n!}\sum_{\pi} P_{fp}(\Lambda, W_n) \\ &= P_{fp}(\Lambda, W_n)\ . \end{aligned} \tag{81}$$

Therefore, for any Λ,

$$\max_{W_n\in\mathcal{W}_n(D_a)} P_{fp}(W_n,\Lambda) = \max_{W_n\in\mathcal{W}_n^{ex}(D_a)} P_{fp}(W_n,\Lambda)\ . \tag{82}$$

Hence, the worst general attack channel is not worse than the worst strongly exchangeable channel, and therefore we can confine our search to the set of strongly exchangeable channels under which Λ_{**}, defined in (77), is optimal. Using a similar proof of Lemma 2, it is easy to show that indeed under Λ_{**} the false-positive probability is not greater than $\exp\big\{-n(\lambda-\delta_n)\big\}$, where $\lim_{n\to\infty}\delta_n = 0$. □

Note that the summation over $\boldsymbol{u}$ (and the fact that any Λ is a union of types) enabled us the use of this argument, which might suggest that for a deterministic watermark, a general attack channel is worse than the worst strongly exchangeable channel. However, this channel might be dependent on the watermark sequence which is not available to the attacker. This is exactly the reason why random watermark setting is considered in the general attack scenario.

Once again, it is easy to verify that Λ_{**} does not violate the false-positive probability constraint under general attack channel while minimizing the false-negative probability.

We now proceed to find the optimal embedder. The false-negative probability for a given attack channel W_n, embedder f_n, and decision region Λ can be written as follow

$$P_{fn}\big(f_n,\Lambda,W_n\big) = \sum_{\boldsymbol{u}\in\mathcal{B}^n} P_U(\boldsymbol{u}) P_{fn}\big(f_n,\Lambda(\boldsymbol{u}),W_n\big), \tag{83}$$

where

$$P_{fn}\big(f_n,\Lambda(\boldsymbol{u}),W_n\big) = \sum_{\boldsymbol{z}\in\Lambda^c(\boldsymbol{u})}\sum_{\boldsymbol{y}\in\mathcal{A}^n}\left(\sum_{\boldsymbol{x}:f_n(\boldsymbol{x},\boldsymbol{u})=\boldsymbol{y}} P_X(\boldsymbol{x})\right) W_n(\boldsymbol{z}|\boldsymbol{y})\ . \tag{84}$$

Corollary 2. *For any attack channel $W_n \in \mathcal{W}_n(D_a)$, the optimal embedder f_n^{**} which minimizes the false-negative probability is the embedder defined in (74).*

Proof. Clearly, for any $\boldsymbol{u}\in\mathcal{B}^n$

$$\begin{aligned} \min_{\substack{f_n(\boldsymbol{x},\boldsymbol{u}):\\ d_e(\boldsymbol{x},\boldsymbol{y})\le nD_e}} \max_{W_n\in\mathcal{W}_n(D_a)} P_{fn}(f_n,\Lambda_{**}(\boldsymbol{u}),W_n) &\ge \min_{\substack{f_n(\boldsymbol{x},\boldsymbol{u}):\\ d_e(\boldsymbol{x},\boldsymbol{y})\le nD_e}} \max_{W_n\in\mathcal{W}_n^{ex}(D_a)} P_{fn}(f_n,\Lambda_{**}(\boldsymbol{u}),W_n) \\ &= \max_{W_n\in\mathcal{W}_n^{ex}(D_a)} P_{fn}(f_n^*,\Lambda_{**}(\boldsymbol{u}),W_n)\ , \end{aligned} \tag{85}$$

but on the other hand

$$\begin{aligned} \min_{\substack{f_n(\boldsymbol{x},\boldsymbol{u}):\\ d_e(\boldsymbol{x},\boldsymbol{y})\le nD_e}} \max_{W_n\in\mathcal{W}_n(D_a)} P_{fn}(f_n,\Lambda_{**},W_n) &\le \max_{W_n\in\mathcal{W}_n(D_a)} P_{fn}(f_n^*,\Lambda_{**},W_n) \\ &= \max_{W_n\in\mathcal{W}_n^{ex}(D_a)} P_{fn}(f_n^*,\Lambda_{**},W_n)\ , \end{aligned} \tag{86}$$

where the last equality can easily be obtained from the above argument when applied to embedders which use a certain conditional type $T(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$ to produce the stegotext (as f_n^*). Therefore, the optimal embedder in the case of a general attack channel is f_n^*, proposed in Theorem 3. □

Note that from (74), (77) the false-negative error exponent can be expressed in a closed form using the method of types [27].

6.3 Discussion

In this section, we extended the basic setup presented in Section 2 to the case of general attack channels. First, we solved the problem for the case where the watermark sequence is deterministic under strongly exchangeable channels. Then, we treated the case of general attack channels, but, we had to assume that the watermark sequence $\boldsymbol{u}$ is random too. However, this should not surprise us. Clearly, for a given watermark, the worst attack channel is dependent on the watermark (although it is not known to the attacker). In this case, the attacker can imitate the detector operation: first, it decides which hypothesis is more likely (using a similar decision rule used by the detector). Then, it can try to "push" the stegotext in the wrong direction causing a false detection. A similar behavior can be seen in the case of a random watermark message $\boldsymbol{u}$ and a deterministic covertext sequence $\boldsymbol{x}$. If $d_e = d_a$ and $D_a \geq D_e$, the worst channel (which does depend on the covertext $\boldsymbol{x}$) is the following: if $\boldsymbol{y} \neq \boldsymbol{x}$ (hypothesis H_1) then $\boldsymbol{z} = \boldsymbol{x}$, i.e., the channel completely erases the message, otherwise (hypothesis H_0) the channel tries to "push" $\boldsymbol{y}$ to Λ. In this case, both the false-negative probability and the false-positive probability might converge to one. The reason for that is rooted in the fact that the set of attack channels has not been limited. In Subsection 6.1, we restricted the class of attack channels to be a strongly exchangeable channel and got non-trivial results. Other limitations may be imposed on the attack channels (e.g., blockwise memoryless, finite-state channels) if meaningful results ought to be obtained.

Note that the worst attack strategy W_n^* is independent of λ, the covertext distribution P_X, and even the embedder strategy and its distortion level D_e (assuming that the embedder use a certain type $T(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{u})$ to produce the stegotext). The attack strategy is only dependent on the allowable distortion level D_a. Therefore, the embedding strategy can be designed assuming that the worst attack channel is present. This can be useful in evaluating the performance (in terms of false-negative probability) of suboptimal embedders.

Appendix

Proof of Theorem 2. First, we explore the case where $a = 0$, i.e., $\boldsymbol{y} = b\boldsymbol{u}$. Substituting $a = 0$ in the constraint of eq. (32), we get that $b^2 - 2\rho b + (\alpha^2 - D_e) \leq 0$. The fact that b is a real number implies that the discriminant of $(b^2 - 2\rho b + (\alpha^2 - D_e))$ is non-negative which leads to $\rho^2 - (\alpha^2 - D_e) \geq 0$, or $D \geq \alpha^2 - \rho^2$. This corresponds to the case where the stegotext includes *only* a fraction of $\boldsymbol{u}$ without violating the distortion constraint. In this case, the false-negative 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 $D_e < \alpha^2 - \rho^2$ 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 (32) by dividing the numerator and denominator by a^2:

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

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

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

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

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

consequently,

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

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

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

After differentiating with respect to a and equating to zero, we get

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

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

Proof of Theorem 3. It is easy to show that under H_1

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

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

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

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

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

and $\Gamma(\cdot)$ denotes the Gamma function. Now, given α^2, D_e and a threshold value $\tau \stackrel{\triangle}{=} 1 - e^{-2\lambda}$, let us find the range of ρ for which $\hat{\rho}^2_{\boldsymbol{uy}} \leq \tau$, i.e.,

$$\hat{\rho}^2_{\boldsymbol{uy}}(\rho) \stackrel{\triangle}{=} \frac{(|\rho| + \sqrt{D_e})^2}{(|\rho| + \sqrt{D_e})^2 + (\alpha^2 - \rho^2)} \leq \tau \quad . \tag{A-12}$$

The function $\hat{\rho}^2_{\boldsymbol{uy}}(\rho)$ is symmetric with respect to the ρ axis, monotonically increasing in $|\rho|$ and attains its minimum value $\frac{D_e}{D_e+\alpha^2}$ at $\rho = 0$. Hence, for $\alpha^2 < \frac{D_e(1-\tau)}{\tau}$, $\hat{\rho}^2_{\boldsymbol{uy}}$ is greater than τ. After solving (A-12) with respect to ρ and using the fact that $\tau \leq 1$, we get that $|\hat{\rho}_{\boldsymbol{uy}}| \leq \sqrt{\tau}$ implies that $|\rho| \leq \sqrt{D_e}(\tau - 1) + \sqrt{D_e\tau^2 + \tau\alpha^2 - \tau D}$ as long as $\alpha^2 \geq \frac{D_e(1-\tau)}{\tau}$. Define

$$\Theta(r) \stackrel{\triangle}{=} \arccos\left[\frac{\sqrt{D_e}(\tau - 1) + \sqrt{D_e\tau^2 + \tau r - \tau D}}{\sqrt{r}}\right] \tag{A-13}$$

It follows that

$$\begin{aligned}
\Pr\left\{\hat{\rho}^2_{\boldsymbol{uy}} \leq \tau \;\middle|\; H_1, \alpha^2 = r\right\} &= \Pr\left\{\rho^2 \leq \left[\sqrt{D_e}(\tau - 1) + \sqrt{D_e\tau^2 + \tau\alpha^2 - \tau D}\right]^2 \;\middle|\; H_1, \alpha^2 = r\right\} \\
&= 1 - \Pr\left\{\rho^2 > \left[\sqrt{D_e}(\tau - 1) + \sqrt{D_e\tau^2 + \tau\alpha^2 - \tau D}\right]^2 \;\middle|\; H_1, \alpha^2 = r\right\} \\
&= 1 - 2\frac{A_n\big(\Theta(r)\big)}{A_n(\pi)} \quad ,
\end{aligned}$$

where

$$\frac{A_n\big(\Theta(r)\big)}{A_n(\pi)} \doteq e^{n \ln \sin\left(\Theta(r)\right)} \, .$$

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

$$\begin{aligned}
P^{(n)}_{fn} &= \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_{\frac{D_e(1-\tau)}{\tau}}^\infty \left[1 - e^{n \ln \sin\left(\Theta(r)\right)}\right] e^{-\frac{nr}{2\sigma^2}} \left(\frac{nr}{\sigma^2}\right)^{\frac{n-2}{2}} dr \\
&= \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \left[\int_{\frac{D_e(1-\tau)}{\tau}}^\infty \frac{\sigma^2}{r} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr - \int_{\frac{D_e(1-\tau)}{\tau}}^\infty e^{n \ln \sin \Theta(r)} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr\right] \, .
\end{aligned} \tag{A-14}$$

Our next step is to evaluate the exponential decay rate of (A-14). It is easy to see that the first integral of (A-14) has a slower exponential decay rate and therefore dictates the overall decay rate. To evaluate

the exponential decay rate of $P_{fn}^{(n)}$ as $n \to \infty$ we use Laplace's method for integrals [5]. 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

$$\lim_{n\to\infty} \frac{1}{n} \ln \left[\frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \right] = \frac{1}{2} \ , \tag{A-15}$$

and therefore the overall exponent is given by

$$E_{fn}^{se}(\tau, D_e) = \min_{r \geq \frac{D_e(1-\tau)}{\tau}} \frac{1}{2} \left[\frac{r}{\sigma^2} - \ln(r/\sigma^2) - 1 \right]. \tag{A-16}$$

The function $g(r) = \left[r/\sigma^2 - \ln(r/\sigma^2) - 1 \right]$, $r \in (0, \infty)$ achieves its minimum at $r = \sigma^2$ and $g(\sigma^2) = 0$. Therefore, in the case where $\frac{D_e(1-\tau)}{\tau} \leq \sigma^2$, $E_{fn}^{se}(\tau, D_e) = 0$. Otherwise, the minimum of (A-16) is obtained at $r = \frac{D_e(1-\tau)}{\tau}$. Hence, the false-negative exponent of the sign embedder is given by

$$E_{fn}^{se}(\tau, D_e) = \begin{cases} 0 & , \quad \frac{D_e(1-\tau)}{\tau} \leq \sigma^2 \\ \frac{1}{2} \left[\frac{D_e(1-\tau)}{\tau\sigma^2} - \ln \left(\frac{D_e(1-\tau)}{\tau\sigma^2} \right) - 1 \right] & , \quad \text{else} \end{cases} \tag{A-17}$$

Setting $\tau = 1 - e^{-2\lambda}$ achieves (41).

□

Proof of Corollary 1. Since the false-negative probability of the improved embedder (42) is zero for $\alpha^2 \leq D_e$ we can rewrite the integral (A-14) for the case where $\frac{1-\tau}{\tau} \leq 1$ (or $\lambda \geq 1/2 \ln 2$) where the lower limit equals to D_e (and does not depend on λ) as following:

$$P_{fn}^{(n)} \quad = \quad \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_{D_e}^{\infty} \left[1 - e^{n \ln \sin \left(\Theta(r) \right)} \right] e^{-\frac{nr}{2\sigma^2}} \left(\frac{nr}{\sigma^2} \right)^{\frac{n-2}{2}} dr \ . \tag{A-18}$$

Optimizing using Laplace method as done in the proof of Theorem 3 leads to (43). □

Proof of Theorem 4. Given $\lambda > 0$, the false-negative probability is given by

$$P_{fn} = \Pr \left\{ \hat{\rho}_{\boldsymbol{uy}} \leq \sqrt{1 - e^{-2\lambda}} \middle| H_1 \right\} \ , \tag{A-19}$$

where the normalized correlation, under H_1, is given by

$$\hat{\rho}_{\boldsymbol{uy}} = \frac{\rho + \sqrt{D_e}}{\sqrt{\alpha^2 + 2\sqrt{D_e}\rho + D}} \quad < \quad T \ . \tag{A-20}$$

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

$$\sqrt{D_e}(T^2 - 1) - T\sqrt{\alpha^2 - D(1 - T^2)} \leq \rho \leq \sqrt{D_e}(T^2 - 1) + T\sqrt{\alpha^2 - D(1 - T^2)} \ .$$

[5] Laplace's method is a general technique for obtaining the asymptotic behavior of integrals of the form $I(x) = \int_a^b f(t) e^{x\Phi(t)} dt$ as $x \to \infty$. In this case $c \in [a, b]$, the maximum of $\Phi(t)$ in the interval $[a, b]$, dictates the asymptotic behavior of the integral (assuming that $f(c) \neq 0$), or in the above case:

$$\lim_{n\to\infty} -\frac{1}{n} \ln \left[\int_a^b f(t) e^{-n\Phi(t)} dt \right] = \min_{t \in [a,b]} \Phi(t) \ .$$

See [34, Sec. 6.4], [35, Ch.4] for more information.

Otherwise, for $\alpha^2 < (1-T^2)D_e$, $\hat{\rho}_{\boldsymbol{uy}} \geq T$ for all $\rho \in [-\alpha, \alpha]$. Define

$$\Psi_1(r) \triangleq \arccos\left[\frac{\sqrt{D_e}(T^2-1)+T\sqrt{r-D(1-T^2)}}{\sqrt{r}}\right], \tag{A-21}$$

$$\Psi_2(r) \triangleq \arccos\left[\frac{\sqrt{D_e}(T^2-1)-T\sqrt{r-D(1-T^2)}}{\sqrt{r}}\right]. \tag{A-22}$$

We need to pay attention to the point $r_0 = \frac{D_e(1-T^2)}{T^2}$ 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 $\lambda < \frac{1}{2}\ln(2)$)

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

The first integral in (A-23) represents 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-23) stems from the fact that for $r \geq D_e$ the false-negative probability (given $\alpha^2 = r$) equals to $\frac{A_n(\pi-\Psi_1(r))}{A(\pi)}$. The last integral in (A-23) 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 $\lambda \geq \frac{1}{2}\ln(2)$)

$$\begin{aligned} P_{fn}^{(n)} = \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \Bigg[& \int_{D_e(1-T^2)}^{\frac{D_e(1-T^2)}{T^2}} \frac{\sigma^2}{r}\left(e^{n\ln\sin\left(\Psi_1(r)\right)} - e^{n\ln\sin\left(\Psi_2(r)\right)}\right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{\frac{D_e(1-T^2)}{T^2}}^{D_e} \frac{\sigma^2}{r}\left(1 - e^{n\ln\sin\left(\Psi_1(r)\right)} - e^{n\ln\sin\left(\Psi_2(r)\right)}\right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{D_e}^{\infty} \frac{\sigma^2}{r}\left(1 - e^{n\ln\sin\left(\Psi_1(r)\right)}\right) e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \Bigg]. \end{aligned} \tag{A-24}$$

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 $D_e(1-T^2) \leq r \leq D(1-T^2)/T^2$ implies that

$$\begin{aligned} P_{fn} \doteq \frac{(1/2)^{\frac{n}{2}} n^{\frac{n-2}{2}}}{\Gamma(n/2)} \Bigg[& \int_{D_e(1-T^2)}^{\frac{D_e(1-T^2)}{T^2}} \frac{\sigma^2}{r} e^{n\ln\sin\left(\Psi_1(r)\right)} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \\ & + \int_{\frac{D_e(1-T^2)}{T^2}}^{\infty} \frac{\sigma^2}{r} e^{-\frac{nr}{2\sigma^2}} e^{\frac{n}{2}\ln(r/\sigma^2)} dr \Bigg]. \end{aligned} \tag{A-25}$$

Again, using the Laplace's method for integrals [35, Ch.4] we can conclude that

$$E_{fn}^{ae}(T, D_e) = \min\left\{E_1(T, D_e), E_2(T, D_e)\right\}, \tag{A-26}$$

where,

$$E_1(T, D_e) = \min_{D_e(1-T^2)<r\leq \frac{D_e(1-T^2)}{T^2}} \frac{1}{2}\left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 2\ln\sin\left(\Psi_1(r)\right) - 1\right], \tag{A-27}$$

$$E_2(T, D_e) = \min_{r>\frac{D_e(1-T^2)}{T^2}} \frac{1}{2}\left[\frac{r}{\sigma^2} - \ln\left(\frac{r}{\sigma^2}\right) - 1\right]. \tag{A-28}$$

$E_2(T, D_e)$ is given by

$$E_2(T, D_e) = \begin{cases} 0 & , \quad \frac{D_e(1-T^2)}{T^2} \leq \sigma^2 \\ \frac{1}{2}\left[\frac{D_e(1-T^2)}{T^2\sigma^2} - \ln\left(\frac{D_e(1-T^2)}{T^2\sigma^2}\right) - 1\right] & , \quad \text{else} \end{cases}. \tag{A-29}$$

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

$$f(r) = \frac{r}{2\sigma^2} - \frac{1}{2}\ln\left(\frac{r}{\sigma^2}\right) - \ln\sin\left(\Psi_1(r)\right) - \frac{1}{2} \tag{A-30}$$

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

$$E_1(T, D_e) \leq f\left(\frac{D_e(1-T^2)}{T^2}\right) = E_2(T, D_e). \tag{A-31}$$

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

$$f'\left(\frac{D_e(1-T^2)}{T^2}\right) = \frac{1}{2}\left[1 - \frac{T^2\sigma^2}{D_e(1-T^2)}\right] > 0 \tag{A-32}$$

hence, $f(r)$ is monotonically increasing in small neighborhood of $\frac{D_e(1-T^2)}{T^2}$, and therefore $E_1(T, D_e) < E_2(T, D_e)$. This fact leads to the conclusion that $E_{fn}^{ae}(\lambda, D_e) < E_{fn}^{se}(\lambda, D_e)$. The exact value of $E_1(T, D_e)$ 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] 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.

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

[6] T. Liu and P. Moulin, "Error exponents for watermarking game with squared-error constraints," in *Proceedings of International Symposium on Information Theory, (ISIT '03)*, Yokohama, Japan, July 2003.

[7] ——, "Error exponents for one-bit watermarking," in *Proceedings of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 3, Apr. 2003, pp. 65–68.

[8] N. Merhav, "An information-theoretic view of watermark embedding-detection and geometric attacks," June 2005, presented at WaCha '05, Barcelona, Spain. [Online]. Available: http://www.ee.technion.ac.il/people/merhav/papers/p98.pdf

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

[10] J. Linnartz, T. Kalker, and G. Depovere, "Modelling the false alarm and missed detection rate for electronic watermarks," in *Information Hiding: Second International Workshop, IH'98*, Portland, Oregon, USA, Apr. 1998, p. 329.

[11] 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.

[12] 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 '00)*, vol. 3, 2000, pp. 1–4.

[13] C. Podilchuk and E. Delp, "Digital watermarking: Algorithms and applications," *IEEE Signal Processing Mag.*, vol. 18, no. 4, pp. 33–46, July 2001.

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

[15] F. Hartung, J. Su, and B. Girod, "Spread spectrum watermarking: Malicious attacks and counter-attacks," in *Proceedings of SPIE Vol. 3657, Security and Watermarking of Multimedia Contents*, San Jose, CA, Jan. 1999, pp. 147–158.

[16] H. S. Malvar and D. A. F. Florêncio, "Improved spread spectrum: a new modulation technique for robust watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 898–905, Apr. 2003.

[17] T. Furon, "A constructive and unifying framework for zero-bit watermarking," *submitted to IEEE Trans. Information Forensics and Security*, 2006.

[18] 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.

[19] M. Gutman, "Asymptotically optimal classification for multiple tests withempirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[20] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a markov chain and universal datacompression," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1014–1019, Sept. 1989.

[21] ——, "Estimating the number of states of a finite-state source," *IEEE Trans. Inform. Theory*, vol. 38, no. 1, pp. 61–65, Jan. 1992.

[22] A. Somekh-Baruch and N. Merhav, "On the error exponent and capacity games of private watermarking systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 537–562, Mar. 2003.

[23] ——, "On the capacity game of public watermarking systems," *IEEE Trans. Inform. Theory*, vol. 50, no. 3, pp. 511–524, Mar. 2004.

[24] E. Sabbag and N. Merhav, "Optimal watermark embedding and detection strategies under limited detection resources," in *Proc. Int. Symp. on Information Theory (ISIT'06)*, Seattle, USA, 2006, pp. 173–177.

[25] ——, "Optimal watermark embedding and detection strategies under general worst case attacks," *Accepted to Int. Symp. on Information Theory (ISIT'07)*, June 2007.

[26] H. L. Van-Trees, *Detection, Estimation and Modulation Theory-Volume I.* New-York: John Wiely & Sons, 1968.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[33] 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.

[34] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers.* New York: McGraw-Hill, 1978.

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