Asymptotically optimum embedding strategy for one-bit watermarking under Gaussian attacks*

Pedro Comesaña[1], Neri Merhav[2] and Mauro Barni[3]

[1] Signal Theory and Communications Dept., University of Vigo, Vigo, Spain
[2] EE Dept., Technion - I.I.T., Haifa, Israel
[3] Dept. of Information Engineering, University of Siena, Siena, Italy

ABSTRACT

The problem of asymptotically optimum watermark detection and embedding has been addressed in a recent paper by Merhav and Sabbag where the optimality criterion corresponds to the maximization of the false negative error exponent for a fixed false positive error exponent. In particular Merhav and Sabbag derive the optimum detection rule under the assumption that the detector relies on the second order statistics of the received signal (universal detection under limited resources), however the optimum embedding strategy in the presence of attacks and a closed formula for the negative error exponents are not available. In this paper we extend the analysis by Merhav and Sabbag, by deriving the optimum embedding strategy under Gaussian attacks and the corresponding false negative error exponent. The improvement with respect to previously proposed embedders are shown by means of plots.

1. INTRODUCTION

About one decade ago the watermarking and data hiding researching community was surprised by the rediscovery of an important result by Costa,[1] and its application to watermarking field.[2,3] Costa's result is that the capacity of the additive white Gaussian noise (AWGN) channel with an additional independent interfering signal, known non–causally to the transmitter, is not reduced by the lack of availability of this knowledge at the decoder. When applied in the realm of watermarking and data hiding systems, this means that host signal (playing the role of the interfering signal), should not actually be considered as an interference since the embedder (the transmitter) can incorporate its knowledge upon generating the watermarked signal (the codeword). The methods based on that paradigm, usually known as *side-informed* methods, can even asymptotically eliminate (under some particular conditions) the interference due to the host signal, that was previously believed to be inherent to any watermarking system.

From the rediscovery of Costa's result, numerous works can be found in the literature proposing practical implementations of this paradigm for the so-called multibit watermarking,[3–6] where the decoder estimates which message among several messages in a given message set has been transmitted. Nevertheless, far less attention has been devoted to the problem of deciding on the presence or absence of a given watermark in the observed signal. In fact, for many of the recent works that deal with this binary hypothesis testing problem, usually known as one-bit (or also zero-bit) watermarking, the watermarking signal is not dependent on the the host signal.[7–9] To the best of our knowledge, exceptions to this statement are the works due to Cox *et al.*,[2,10] Liu and Moulin,[11]

*This work was supported by the Italian Ministry for University and Research, under FIRB project no. RBIN04AC9W; *Xunta de Galicia* under projects FACTICA 07TIC012322PR and Competitive research units program Ref. 150/2006; European Comission through the IST Programme under Contract IST-2002-507932 ECRYPT. ECRYPT disclaimer: The information in this paper is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Further author information: (Corresponding author: P.C.)
P.C.: E-mail: pcomesan@gts.tsc.uvigo.es, Telephone: +34-986812683
N.M.: E-mail: merhav@ee.technion.ac.il, Telephone: +972-(4)-829-4737
M.B.: E-mail: barni@dii.unisi.it, Telephone: +39-0577234621

Merhav and Sabbag[12] and Furon.[13] In the next few paragraphs, we briefly describe the main results in these works.

Cox et. al.[2, 10]

In their paper, Cox *et. al.*[2] introduce the paradigm of watermarking as a communication system with side information at the embedder. Based on this paradigm and considering a statistical model for the attacking vectors (based on the correlation between the noise and the watermarked signal), the authors propose a detection rule based on the Neyman-Pearson test. Nevertheless, the detection region obtained this way is replaced by the union of two hypercones; mathematically $\frac{|\mathbf{q}^t \cdot \mathbf{u}|}{||\mathbf{q}|| ||\mathbf{u}||} \geq \tau(\alpha)$, with $\mathbf{q}$ being the received signal, $\mathbf{u}$ the watermark, $(\mathbf{x})^t$ the transpose of $\mathbf{x}$, $\mathbf{q}^t \cdot \mathbf{u}$ the inner product of $\mathbf{q}$ and $\mathbf{u}$, α the bound on the probability of false positive, and $\tau(\alpha)$ the decision threshold, which will be a function of α. In a successive paper, Miller *et al.* also compare the performance of the strategy derived there with other typical embedding strategies.[10]

Liu and Moulin[11]

In Liu and Moulin's paper[11] the error exponents of both false positive and false negative probabilities are studied for the one-bit watermarking problem, both for Additive Spread Spectrum (Add-SS) and a Quantization Index Modulation (QIM) based techniques. In this case, the constraint on the embedding distortion is imposed on the mean Euclidean norm of the watermarking signal, and the unwatermarked signals are also considered to be attacked. For Additive Spread Spectrum the exact expressions for the error exponents of both false positive and false negative are computed; for QIM the authors provide bounds to those values. The obtained results show that although the error exponents for QIM are larger than those obtained for public Add-SS (where the host signal is not available at the detector), they are still smaller that those computed for private Add-SS (where the host signal is also available at the detector side). This seems to indicate that the interference due to the host is not completely removed.

Merhav and Sabbag[12]

Merhav and Sabbag[12] dealt with the problem of one-bit watermarking from an information theoretic point of view. The authors look for the optimal embedder and detector, in the sense of minimizing the probability of false negative while the exponential decay rate of the probability of false positive is larger than or equal to a given constant λ, under a certain limitation on the kind of empirical statistics gathered by the detector. Another important feature of Merhav and Sabbag's analysis is that the statistics of the host signal are assumed to be unknown (thus making this analysis very practical given that in real systems this is often the case). For the continuous case, it is shown that a detector based on the normalized correlation between the received signal $\mathbf{y}$ and the deterministically known watermark $\mathbf{u}$, i.e. $\frac{(\mathbf{u}^t \cdot \mathbf{y})^2}{||\mathbf{y}||^2 \cdot ||\mathbf{u}||^2}$, is optimal for the case of Gaussian distributed host and attack–free scenario. Merhav and Sabbag also derive the optimal embedding strategy for the no-attack case and find a lower bound on the false negative error exponent in this case. Furthermore, the optimization problem is reduced to an easily implementable 2D problem yielding a very simple embedding rule. In the same paper, Merhav and Sabbag study the scenario where the watermarked signal is attacked, using the concept of strongly exchangeable attack channels. In this case, however, closed formulas for the error exponents and the optimum embedding rule are not available due to the complexity of the obtained optimizations.

Furon[13]

In a very recent work[13] Furon uses the Pitman-Noether theorem[14] to derive the form of the best detector for a given embedding function, and the best embedding function for a given detection function; by combining both of them, a differential equation is obtained, that the author refers to as *"fundamental equation of zero-bit watermarking"*. Compared with the framework introduced by Merhav and Sabbag, two important differences must be highlighted:

- For Furon[13] the watermarking signal is constrained to be just a function of the host signal which is scaled to yield a given embedding distortion. This means that the direction of the watermarking signal can not be changed as a function of the allowed embedding distortion.

- One of the conditions that must be verified in order to apply the Pitman-Noether theorem is that the power of the watermarking signal has to go to zero when the dimensionality is increased. In fact, the author hypothesizes that this is the reason why neither the absolute normalized correlation nor the normalized correlation are solutions to the fundamental equation.

In this paper we extend the analysis carried out by Merhav and Sabbag to derive the optimum embedding strategy for Gaussian host and Additive White Gaussian attack. In order to do so, an exact expression for the false negative error exponent is derived in Section 2; then, this expression is maximized as a function of the watermarking signal in Section 3. As a particular case of the obtained result, the error exponent of the noiseless case is computed in Section 4, comparing the currently proposed exact results with previous bounds in the literature. Finally, the main conclusions of this work are summarized in Section 5.

2. PROPOSED FRAMEWORK AND COMPUTATION OF FALSE NEGATIVE PROBABILITY

We denote scalar random variables with capital letters (e.g., X) and their realizations with lowercase letters (e.g. x). The same notation convention applies to n-dimensional random vectors and their outcomes, denoted by bold letters (e.g. $\mathbf{X}$, $\mathbf{x}$). The ith component of a vector $\mathbf{X}$ is denoted as X_i. The probability density function (pdf) of a random variable A is denoted by $f_A(a)$.

Let $\mathbf{u}$ be the watermark and $\mathbf{x}$ the host sequence, both of them n-dimensional vectors; $\mathbf{u}$ is considered to be a binary vector whose components belong to $\{-1, +1\}$, while the components of $\mathbf{x}$ are real-valued. The embedder will produce a watermarked sequence $\mathbf{y} = \mathbf{x} + \mathbf{w}$, with $\mathbf{w}$, the so called watermarking signal, being a function of $\mathbf{x}$ and $\mathbf{u}$. The embedder must keep the embedding distortion $d_e(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||^2$ within a prescribed limit, i.e., $d_e(\mathbf{x}, \mathbf{y}) \leq nD_e$, where $D_e > 0$ is the maximum allowed distortion per dimension. Note that this embedding distortion constraint must be satisfied for every value of the host signal and watermark. The watermarked signal could be manipulated by an attacker trying to make the watermark detection more difficult. In this paper, the attacker is modeled by a discrete–time additive white Gaussian channel whose noise variance is σ_Z^2; the attacking signal is denoted by $\mathbf{z}$, and the attacked signal, which will be the detector observation, by $\mathbf{s} = \mathbf{y} + \mathbf{z}$, that takes values in $\mathbb{R}^n$.

$\mathbb{R}^n$ is partitioned into two complementary regions, Λ and Λ^c, where in the former (usually known as detection region), the detector decides that the watermark is present and in the latter, it decides that it is absent. In this work we assume that the detector knows the watermark signal $\mathbf{u}$, whereas it does not know the host signal $\mathbf{x}$. The design of the detection region was studied by Merhav and Sabbag.[12]

The performance of one-bit watermarking is usually measured using the *false positive* and *false negative* probabilities, defined as the probability of deciding that the received signal is watermarked when it is not (P_{fp}), and the probability of deciding that the received signal is not watermarked when it is actually watermarked (P_{fn}), respectively. As n grows without bound, these probabilities normally decay exponentially rapidly. Accordingly, the corresponding exponential decay rates of these error probabilities, i.e. the *error exponents*, are defined as

$$E_{fp} \triangleq \lim_{n\to\infty} -\frac{1}{n} \ln P_{fp},$$
$$E_{fn} \triangleq \lim_{n\to\infty} -\frac{1}{n} \ln P_{fn}.$$

The main aim of this paper is to determine the optimum choice of the watermarking signal $\mathbf{w}$ for the case of i.i.d. Gaussian host with variance σ_X^2, where optimality corresponds to the maximization of the false negative error exponent (E_{fn}) in the presence of an i.i.d. AWGN attack $\mathbf{z}$ with variance σ_Z^2, subject to a guaranteed false positive exponent of at least λ, where λ is a prescribed positive real. Note that our setup assumes that the attacker knows whether the signal he is observing is watermarked or not; this seems to be a reasonable assumption, as he usually knows the source of the observed signals, being able to guess if the observed signal is watermarked or not. Practical examples of this discussion could be attacker's personal photo database, where he knows that the photos are not watermarked, on one hand, and the photos available in the webpage of a news agency, that will be watermarked with a very high probability, on the other hand. Therefore, the attacker will focus his attack only on the watermarked contents. Given that the detection rule proposed by Merhav and Sabbag[12] is derived as a function of the false positive probability, therefore depending on the statistics of non-watermarked contents, it will not be modified when just the watermarked signals are attacked. The mentioned rule is based on the comparison of squared normalized correlation between the received signal and the watermark, i.e. $\frac{(\mathbf{s}^t \cdot \mathbf{u})^2}{||\mathbf{s}||^2 ||\mathbf{u}||^2}$, with a threshold that depends on the probability of false positive. The methodology we follow for our derivation is the following: we fix $\mathbf{w}$, then we calculate E_{fn}, and finally we maximize it to find the optimum watermarking signal $\mathbf{w}$.

In order to simplify the subsequent analysis, it is convenient to introduce an ad hoc coordinate system for which the only components of $\mathbf{u}$, $\mathbf{x}$ and $\mathbf{w}$ are along the first three coordinate axes. In particular, given $\mathbf{u}$, $\mathbf{x}$ and $\mathbf{w}$, we apply the Gram-Schmidt orthogonalization procedure to $\mathbf{u}$, $\mathbf{x}$ and $\mathbf{w}$, and fix the other directions arbitrarily. By remembering that the optimum detection region derived in Merhav and Sabbag's paper[12] corresponds to the set of vectors forming an angle lower than a limit angle β (with $\beta \leq \arcsin(e^{-\lambda})$) with the watermark direction $\mathbf{u}$, the optimum detection rule can be written as

$$\frac{(x_1 + w_1 + z_1)^2}{(x_1 + w_1 + z_1)^2 + (x_2 + w_2 + z_2)^2 + (w_3 + z_3)^2 + \sum_{j=4}^{n} z_j^2} < \cos^2(\beta), \tag{1}$$

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

$$x_1^2 = n \cdot r \cdot \sin^2(\alpha), \tag{2}$$
$$x_2^2 = n \cdot r \cdot \cos^2(\alpha), \tag{3}$$

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

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

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

Defining

$$T_1 \triangleq (\sqrt{r}\sin(\alpha) + v_1)^2 \left(\frac{1}{\cos^2(\beta)} - 1\right) - \left[\sqrt{r}\cos(\alpha) + v_2\right]^2 - v_3^2, \tag{5}$$

and

$$\begin{aligned} T_2 \quad \triangleq \quad & -[z_1^2 + 2z_1(\sqrt{nr}\sin(\alpha) + \sqrt{n}v_1)] \left(\frac{1}{\cos^2(\beta)} - 1\right) + z_2^2 \\ + \quad & 2z_2\left[\sqrt{nr}\cos(\alpha) + \sqrt{n}v_2\right] + z_3^2 + 2\sqrt{n}v_3 z_3, \end{aligned} \tag{6}$$

the condition for a false negative can be rewritten as

$$nT_1 < (n-3)q + T_2, \tag{7}$$

or equivalently

$$q > \frac{nT_1}{n-3} - \frac{T_2}{n-3}. \tag{8}$$

Recalling that pdf of $\frac{(n-3)Q}{\sigma_Z^2}$ is a χ^2 distribution of $n-3$ degrees of freedom, we can write

$$f_Q(q) = \begin{cases} \left(\frac{1}{2}\right)^{(n-3)/2} \frac{1}{\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{(n-3)q}{\sigma_Z^2}\right)^{\left(\frac{n-3}{2}-1\right)} e^{-\frac{(n-3)q}{2\sigma_Z^2}}, & \text{if } q \geq 0 \\ 0, & \text{elsewhere} \end{cases}. \tag{9}$$

For the random variable R, corresponding to the squared Euclidean norm of the host signal normalized by the number of dimensions, $\frac{nR}{\sigma_X^2}$ follows a χ^2 distribution with n degrees of freedom, so we can write

$$f_R(r) = \begin{cases} \left(\frac{1}{2}\right)^{n/2} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{nr}{\sigma_X^2}\right)^{\left(\frac{n}{2}-1\right)} e^{-\frac{nr}{2\sigma_X^2}}, & \text{if } r \geq 0 \\ 0, & \text{elsewhere} \end{cases}. \tag{10}$$

Furthermore, if we denote by Ψ the random variable whose samples are the values of α, we can see that the probability distribution of Ψ is

$$P(\Psi \leq \alpha) = 1 - \frac{A_n(\pi/2 - \alpha)}{2A_n(\pi/2)}, \tag{11}$$

where $A_n(\theta)$ 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 θ, i.e.

$$A_n(\theta) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^\theta \sin^{(n-2)}(\varphi) d\varphi. \tag{12}$$

From the last formula, it is easy to see that the pdf of Ψ is given by

$$f_\Psi(\alpha) = \frac{\partial P(\Psi \leq \alpha)}{\partial \alpha} = \frac{2\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \cos^{n-2}(\alpha). \tag{13}$$

Therefore, we can write the probability of false positive as

$$\begin{aligned} P_{fn} \;=\; & \int_{\alpha=-\pi/2}^{\pi/2} \int_{r=0}^{+\infty} \int_{z_3=-\infty}^{+\infty} \int_{z_2=-\infty}^{+\infty} \int_{z_1=-\infty}^{+\infty} \int_{q=\max(0, \frac{nT_1}{n-3} - \frac{T_2}{n-3})}^{+\infty} \left(\frac{1}{2}\right)^{(n-3)/2} \\ & \frac{1}{\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{(n-3)q}{\sigma_Z^2}\right)^{\left(\frac{n-3}{2}-1\right)} e^{-\frac{(n-3)q}{2\sigma_Z^2}} \frac{e^{-\frac{z_1^2}{2\sigma_Z^2}}}{\sqrt{2\pi\sigma_Z^2}} \frac{e^{-\frac{z_2^2}{2\sigma_Z^2}}}{\sqrt{2\pi\sigma_Z^2}} \frac{e^{-\frac{z_3^2}{2\sigma_Z^2}}}{\sqrt{2\pi\sigma_Z^2}} \\ & \left(\frac{1}{2}\right)^{n/2} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{nr}{\sigma_X^2}\right)^{\left(\frac{n}{2}-1\right)} e^{-\frac{nr}{2\sigma_X^2}} \frac{2\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \cos^{n-2}(\alpha)\, dq dz_1 dz_2 dz_3 dr d\alpha. \end{aligned} \tag{14}$$

Considering that $\lim_{n\to\infty} \frac{nT_1}{n-3} - \frac{T_2}{n-3} = T_1$, since T_2 grows sublinearly with the dimensionality of the problem, we can replace the inner integral as below:

$$\begin{aligned} P_{fn} \;=\; & \int_{\alpha=-\pi/2}^{\pi/2} \int_{r=0}^{+\infty} \int_{z_3=-\infty}^{+\infty} \int_{z_2=-\infty}^{+\infty} \int_{z_1=-\infty}^{+\infty} \int_{q=\max(0,T_1)}^{+\infty} \left(\frac{1}{2}\right)^{(n-3)/2} \\ & \frac{1}{\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{(n-3)q}{\sigma_Z^2}\right)^{\left(\frac{n-3}{2}-1\right)} e^{-\frac{(n-3)q}{2\sigma_Z^2}} \frac{e^{-\frac{z_1^2}{2\sigma_Z^2}}}{\sqrt{2\pi\sigma_Z^2}} \frac{e^{-\frac{z_2^2}{2\sigma_Z^2}}}{\sqrt{2\pi\sigma_Z^2}} \frac{e^{-\frac{z_3^2}{2\sigma_Z^2}}}{\sqrt{2\pi\sigma_Z^2}} \\ & \left(\frac{1}{2}\right)^{n/2} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{nr}{\sigma_X^2}\right)^{\left(\frac{n}{2}-1\right)} e^{-\frac{nr}{2\sigma_X^2}} \frac{2\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \cos^{n-2}(\alpha)\, dq dz_1 dz_2 dz_3 dr d\alpha. \end{aligned} \tag{15}$$

Using the multidimensional version of Laplace's approximation,[15] one can conclude that

$$\begin{aligned}\lim_{n\to\infty} -\frac{1}{n}\ln(P_{fn}) &= \min_{(q,r,\alpha)\in[\max(0,T_1(r,\alpha)),\infty)\times[0,\infty)\times[-\pi/2,\pi/2]} \frac{1}{2}\left[\frac{q}{\sigma_Z^2} - \ln\left(\frac{q}{\sigma_Z^2}\right) - 1\right] \\ &+ \frac{1}{2}\left[\frac{r}{\sigma_X^2} - \ln\left(\frac{r}{\sigma_X^2}\right) - 1\right] - \ln\left[\cos\left(\alpha\right)\right], \end{aligned} \tag{16}$$

where the notation $T_1(r,\alpha)$ tries to make evident the dependence of T_1 with those variables.

3. FALSE POSITIVE ERROR EXPONENT AND OPTIMAL WATERMARKING SIGNAL COMPUTATION

In order to devise the optimum embedding strategy we need to maximize the error exponent reported in equation (16) as a function of v_1, v_2 and v_3. Let us start by considering the dependence of the last formula with α; on one hand, it is straightforward to see that $-\ln[\cos(\alpha)]$ is minimized when $\alpha = 0$. On the other hand, T_1 also depends on α; given that the embedder is interested in maximizing T_1, since in that way it is making smaller the interval where the objective function can be minimized with respect to q, and changing the sign of any component of the watermark does not affect the embedding distortion, it is straightforward to see that the sign of v_1 and v_2 will be such that $v_1\sin(\alpha) \geq 0$, and $v_2\cos(\alpha) \leq 0$. Therefore $T_1(r,\alpha)$ is symmetrical with respect to α, and its minimum is reached for $\alpha = 0$. Summarizing, the minimum of (16) is obtained for $\alpha = 0$. This implies that (16) can be rewritten as

$$\begin{aligned}\lim_{n\to\infty} -\frac{1}{n}\ln(P_{fn}) &= \min_{(q,r)\in[\max(0,T_1(r)),\infty)\times[0,\infty)} \frac{1}{2}\left[\frac{q}{\sigma_Z^2} - \ln\left(\frac{q}{\sigma_Z^2}\right) - 1\right] \\ &+ \frac{1}{2}\left[\frac{r}{\sigma_X^2} - \ln\left(\frac{r}{\sigma_X^2}\right) - 1\right]. \end{aligned} \tag{17}$$

Given that the objective function is convex with respect to (r,q), and the global minimum is at (σ_X^2, σ_Z^2), the result of (17) will be 0 if $(\sigma_Z^2, \sigma_X^2) \in [\max(0, T_1(r)), \infty) \times [0,\infty)$, and, in any other case, the minimum will be in the boundary of that region, i.e., the points of the form $(T_1(r), r)$, with $r \geq 0$.

So far we have not paid attention to the choice of the optimal watermarking signal $(w_1^*, w_2^*, w_3^*, 0, \cdots, 0)$. The first question to be answered concerning this problem is the role that the watermarking signal plays in the optimization described by (17). In this case, it is easy to see that the only influence of $\mathbf{w}^*$ (or equivalently $\mathbf{v}^*$) on that formula is through T_1. In other words, the embedder will choose the watermarking signal in order to maximize T_1, since in that way it will reduce the region where q can take values; of course this will imply an increase on the obtained error exponent. Given that the considered value of α is 0, as it was explained before, T_1 can be now written like

$$T_1 = v_1^2\left(\frac{1}{\cos^2(\beta)} - 1\right) - \left[\sqrt{r} + v_2\right]^2 - v_3^2, \tag{18}$$

that the embedder is trying to maximize constrained to

$$v_1^2 + v_2^2 + v_3^2 \leq D_e. \tag{19}$$

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

$$T_1 = v_1^2\left(\frac{1}{\cos^2(\beta)} - 1\right) - \left[\sqrt{r} - \sqrt{D_e - v_1^2}\right]^2. \tag{20}$$

[†]Note that two solutions are possible for v_2, namely $v_2 = \pm\sqrt{D_e - v_1^2}$. Here we took the negative one, since, as we noted before, v_2 and $cos(\alpha)$ must have opposite signs and given that $-\pi/2 \leq \alpha \leq \pi/2$ $cos(\alpha)$ is always positive.

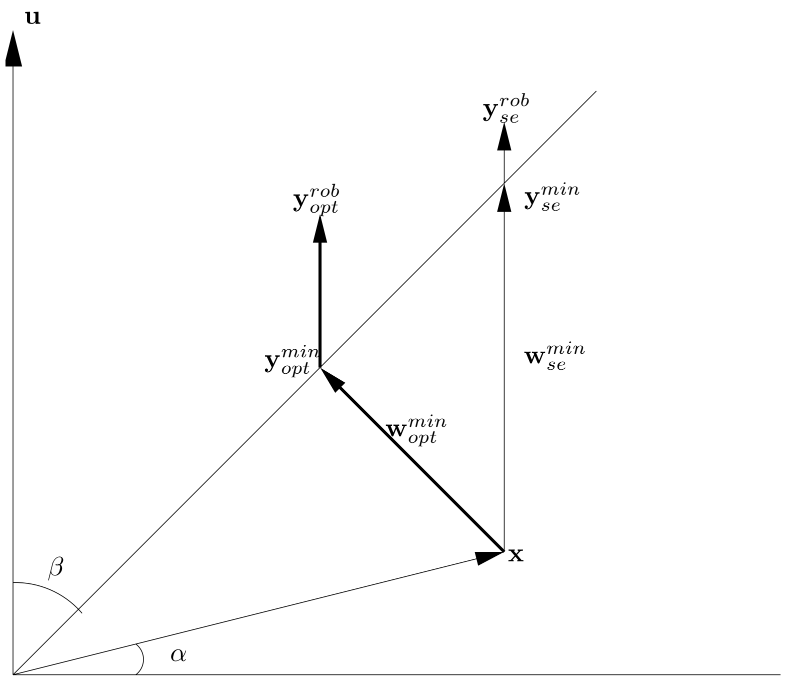

Figure 1. Geometrical interpretation of the problem, and comparison between the sign-embedder solution and the optimal one. $\mathbf{w}_{opt}^{min}$ and $\mathbf{w}_{se}^{min}$ denote the minimum norm watermarking signals that produce signals in the detection region, for both the optimal embedder and the sign embedder, respectively. The corresponding watermarked signals are $\mathbf{y}_{opt}^{min}$ and $\mathbf{y}_{se}^{min}$. Furthermore, one can see the watermarked signals for the optimal embedder and the sign embedder when part of the embedding distortion can be used to gain some robustness to noise (denoted by $\mathbf{y}_{opt}^{rob}$ and $\mathbf{y}_{se}^{rob}$).

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

$$\frac{\partial T_1}{\partial v_1} = 2v_1 \left(\frac{1}{\cos^2(\beta)} - \frac{\sqrt{r}}{\sqrt{D_e - v_1^2}} \right), \tag{21}$$

which is equal to 0 in the following cases

$$\begin{cases} v_1 = 0 \\ v_1 = -\sqrt{D_e - r\cos^4(\beta)} \\ v_1 = \sqrt{D_e - r\cos^4(\beta)} \end{cases}. \tag{22}$$

Considering the second derivative, it is easy to see that for $v_1^* = \pm\sqrt{D_e - r\cos^4(\beta)}$ one obtains local (in this case in fact they are global) maxima of T_1, yielding a value of $v_2^* = -\sqrt{r}\cos^2(\beta)$, and a corresponding value of $T_1 = D_e \tan^2(\beta) - r\sin^2(\beta)$. This gives us a first threshold for obtaining positive error exponents: if $T_1 \le 0$, then the optimization in (17) is performed on the region $[0,\infty) \times [0,\infty)$, so any pair (σ_Z^2, σ_X^2), even with $\sigma_Z^2 = 0$, will be in the allowed region, yielding an error exponent equal to 0. The condition for this not to happen is $r \le \frac{D_e}{\cos^2(\beta)}$.

When $\alpha = 0$, which, as previously discussed, is the case that asymptotically sums up most of probability, we can express the two components of the watermarked signal $\mathbf{y}$,

$$y_1 = \pm\sqrt{n}\sqrt{D_e - r\cos^4(\beta)}, \tag{23}$$

$$y_2 = \sqrt{nr}[1 - \cos^2(\beta)]. \tag{24}$$

Interestingly, the watermarked signal lies in the plane defined by the watermark $\mathbf{u}$ and the host signal $\mathbf{x}$ (a similar conclusion was reached by Merhav and Sabbag[12]in the attack-free case). The geometrical interpretation of the embedding strategy is particularly interesting: the embedder spends part of the embedding distortion scaling down the host signal (for reducing its interference), and then introduces as much energy as possible in the direction of the watermark. In fact this is the reason why only the first component of the watermarked signal depends on the allowed embedding distortion D_e. For the sake of illustration we compare the optimal embedding and the sign-embedder. For the latter the watermarked signal is given by $\mathbf{y}_{se} = \mathbf{x} + \text{sign}(\mathbf{x}^t \cdot \mathbf{u})\sqrt{D_e}\mathbf{u}$, so the watermarking signal can be written as $\mathbf{w}_{se} = \text{sign}(\mathbf{x}^t \cdot \mathbf{u})\sqrt{D_e}\mathbf{u}$. Both strategies can be compared in Fig. 1, where it is easy to see that the proposed strategy is that of minimizing the embedding distortion necessary for obtaining a watermarked signal. Be aware that the proposed embedding technique could not be described by the work due to Furon,[13] as in that case the watermarking signal direction is just a function of the host signal, and it is scaled for obtaining the desired distortion.

Once the optimum embedder has been found it is possible to compute the false negative error exponent of the optimum embedder and compare it with the previous results available in the literature. Specifically the expression we obtained is as follows:

$$E_{fn} = \frac{1}{2}\left[\frac{q^*}{\sigma_Z^2} - \ln\left(\frac{q^*}{\sigma_Z^2}\right) - 1\right] + \frac{1}{2}\left[\frac{r^*}{\sigma_X^2} - \ln\left(\frac{r^*}{\sigma_X^2}\right) - 1\right]. \tag{25}$$

with

$$\begin{aligned} r^* = {} & \Big(D_e\sigma_Z^2 + 2\sigma_Z^2\sigma_X^2\cos^2(\beta) - D_e\sigma_X^2\sin^2(\beta) \\ & -\sqrt{D_e^2\sigma_Z^4 + 4\sigma_Z^4\sigma_X^4\cos^4(\beta) - 2D_e^2\sigma_Z^2\sigma_X^2\sin(\beta)^2 + D_e^2\sigma_X^4\sin^4(\beta)}\Big) \\ & \Big(2(\sigma_Z^2\cos^2(\beta) - \sigma_X^2\cos^2(\beta)\sin^2(\beta))\Big)^{-1}, \end{aligned} \tag{26}$$

$$\begin{aligned} q^* = {} & \Bigg[\left(2D_e\sigma_Z^2 + \sqrt{16\sigma_Z^4\sigma_X^4\cos(\beta)^4 + D_e^2\left[2\sigma_Z^2 - \sigma_X^2(1-\cos(2\beta))\right]^2}\right)\tan^2(\beta) \\ & - 2\sigma_X^2\sin^2(\beta)\left(2\sigma_Z^2 + D_e\tan^2(\beta)\right)\Bigg]\left[4\left(\sigma_Z^2 - \sigma_X^2\sin^2(\beta)\right)\right]^{-1}. \end{aligned} \tag{27}$$

In Figs. 2, 3 and 4 the behavior of the best achievable error exponent of the false negative error probability as a function of the different parameters involved in its computation is depicted. As it was intuitively expected, the false negative error exponent decreases when the false positive error exponent λ, the attacking signal variance σ_Z^2 or the host signal variance σ_X^2 increase, while it increases for increasing values of D_e.

4. NOISELESS CASE

As a particular case of the studied framework we can derive the false negative error exponent for the noiseless case ($\sigma_Z^2 = 0$). Computing the limit when σ_Z^2 goes to 0 in (26), (27) and (25) is easy to see that in the noiseless case

$$\lim_{\sigma_Z^2 \to 0} r^* = \frac{D_e}{\cos^2(\beta)} = \frac{D_e}{1 - e^{-2\lambda}}, \tag{28}$$

$$\lim_{\sigma_Z^2 \to 0} \frac{q^*}{\sigma_Z^2} = 1 \tag{29}$$

$$\lim_{\sigma_Z^2 \to 0} E_{fn} = \begin{cases} 0, & \text{, if } \frac{D_e}{1-e^{-2\lambda}} \le \sigma_X^2 \\ \frac{1}{2}\left[\frac{D_e}{\sigma_X^2(1-e^{-2\lambda})} - \ln\left(\frac{D_e}{\sigma_X^2(1-e^{-2\lambda})}\right) - 1\right] & \text{, elsewhere} \end{cases}. \tag{30}$$

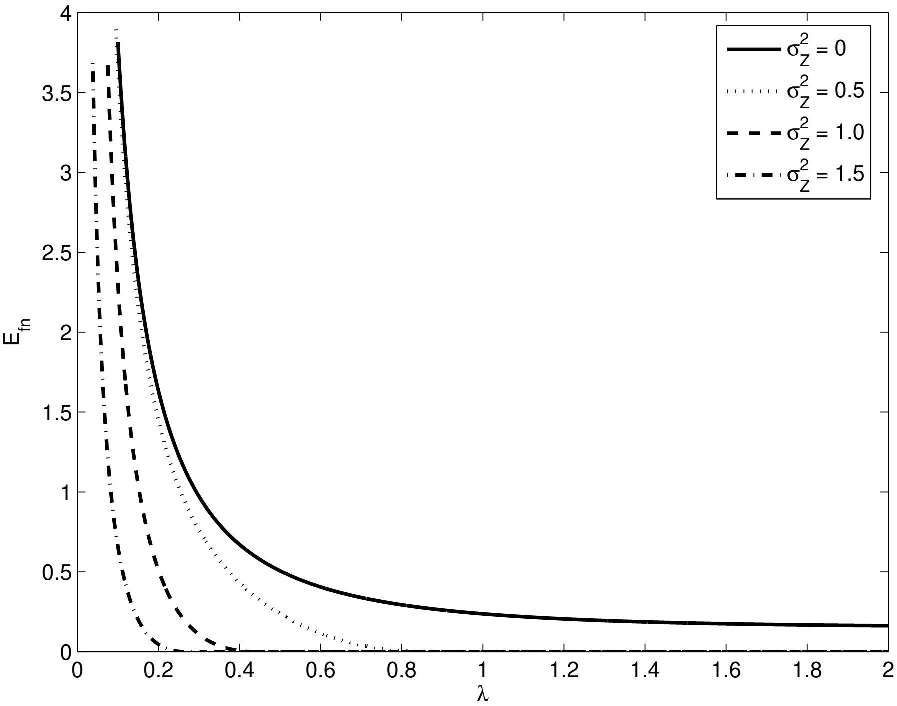

Figure 2. Error exponent of probability of false alarm, as a function of λ, for several powers of AWGN. $\sigma_X^2 = 1$ and $D_e = 2$.

In view of (30) it is interesting to note that as long as $D_e > \sigma_X^2$, E_{fn} will be larger than 0 for any value of λ; in fact, on those conditions, the asymptotic value of E_{fn} when λ goes to infinity is

$$\frac{1}{2}\left[\frac{D_e}{\sigma_X^2} - \ln\left(\frac{D_e}{\sigma_X^2}\right) - 1\right], \tag{31}$$

coinciding with the result of [2. Corollary 1].

On the other hand, when $D_e \leq \sigma_X^2$ another interesting point which reflects the goodness of the proposed strategy is the computation of the range of values of λ where $E_{fn} > 0$ can be achieved. In this case, the condition to be verified is

$$\frac{D_e}{1 - e^{-2\lambda}} > \sigma_X^2, \tag{32}$$

implying that

$$\lambda < -\frac{1}{2}\ln\left(1 - \frac{D_e}{\sigma_X^2}\right) = \lambda_1, \text{ for } D_e \leq \sigma_X^2, \tag{33}$$

whereas for the sign embedder studied by Merhav and Sabbag[12] the values of λ which provide $E_{fn} > 0$ are those such that

$$\frac{D_e}{\sigma_X^2} > \frac{1 - e^{-2\lambda}}{e^{-2\lambda}}, \tag{34}$$

or, equivalently,

$$\lambda < -\frac{1}{2}\ln\left(\frac{\sigma_X^2}{D_e + \sigma_X^2}\right) = \lambda_2, \text{ for all } D_e. \tag{35}$$

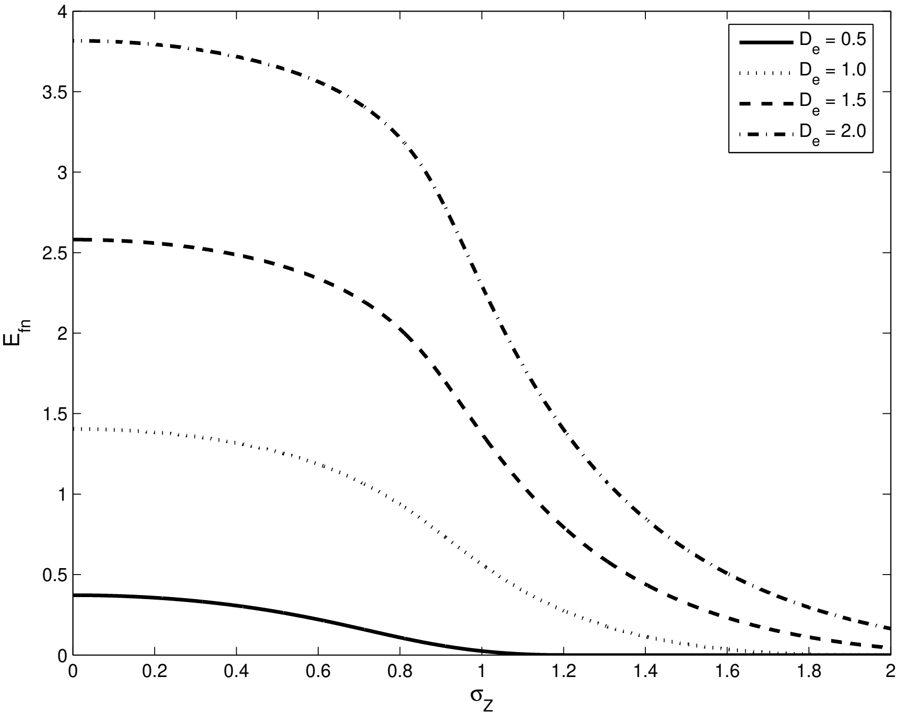

Figure 3. Error exponent of probability of false alarm, as a function of σ_Z, for several embedding distortions. $\sigma_X^2 = 1$ and $\lambda = 0.1$.

Given that $\lambda_1 > \lambda_2$, larger values of error exponents of probability of false negative can be achieved for $E_{fn} > 0$ in the current case. Finally, in Fig. 5 we can compare the bounds to the false negative error exponent for the noiseless case proposed by Merhav and Sabbag,[12] with its optimal value derived in the current work.

5. CONCLUSIONS

In this paper we extend the results presented by Merhav and Sabbag[12] on the computation of the optimum one-bit watermarking system, and the corresponding false negative and false positive error exponents, when the resources available at the detector are limited and the host signal is Gaussian. We consider the case where the watermarked signal is added an AWGN attacking signal, and compute in that case both the optimal embedding strategy and the resulting false negative error exponents, constrained to verify a false positive error exponent. The noiseless scenario can be seen as a particular case of this framework, so the obtained results can be particularized for it; doing so, we compute for the first time the optimum false negative error exponent for the noiseless problem previously studied in the literature, and compare it with the lower bounds proposed by Merhav and Sabbag.[12]

REFERENCES

1. M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory* **29**, pp. 439–441, May 1983.
2. I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE* **87**, pp. 1127–1141, July 1999.
3. B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory* **47**, pp. 1423–1443, May 2001.
4. M. Ramkumar and A. N. Akansu, "Signaling methods for multimedia steganography," *IEEE Transactions on Signal Processing* **52**, pp. 1100–1111, April 2004.

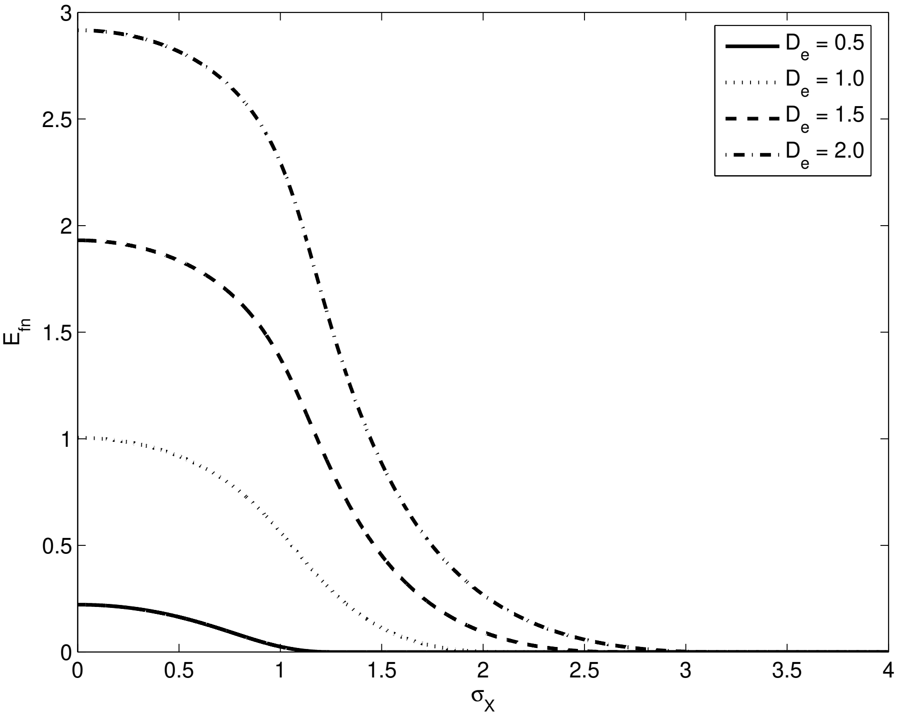

Figure 4. Error exponent of probability of false alarm, as a function of σ_X, for several embedding distortions. $\sigma_N^2 = 1$ and $\lambda = 0.1$.

5. A. Abrardo and M. Barni, "Informed watermarking by means of orthogonal and quasi-orthogonal dirty paper coding," *IEEE Transactions on Signal Processing* **53**, pp. 824–833, February 2005.
6. F. Pérez-González, C. Mosquera, M. Barni, and A. Abrardo, "Rational dither modulation: a high-rate data-hiding method invariant to gain attack," *IEEE Transactions on Signal Processing* **53**, pp. 3960–3975, October 2005.
7. X. Huang and B. Zhang, "Statistically robust detection of multiplicative spread-spectrum watermarks," *IEEE Transactions on Information Forensics and Security* **2**, pp. 1–13, March 2007.
8. M. Noorkami and R. M. Mersereau, "A framework for robust watermarking of H.264-encoded video with controllable detection performance," *IEEE Transactions on Information Forensics and Security* **2**, pp. 14–23, March 2007.
9. W. Liu, L. Dong, and W. Zeng, "Optimum detection for spread-spectrum watermarking that employs self-masking," *IEEE Transactions on Information Forensics and Security* **2**, pp. 645–654, December 2007.
10. M. L. Miller, I. J. Cox, and J. A. Bloom, "Informed embedding: Exploiting image and detector information during watermark insertion," in *IEEE International Conference on Image Processing (ICIP)*, **3**, pp. 1–4, (Vancouver, BC, Canada), September 2000.
11. T. Liu and P. Moulin, "Error exponents for one-bit watermarking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **3**, pp. 65–68, (Hong Kong), April 2003.
12. N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Transactions on Information Theory* **54**, pp. 255–274, January 2008.
13. T. Furon, "A constructive and unifying framework for zero-bit watermarking," *IEEE Transactions on Information Forensics and Security* **2**, pp. 149–163, June 2007.
14. H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer Texts in Electrical Engineering, 2nd ed., 1994.
15. R. Wong, *Asymptotic Approximations of Integrals*, SIAM, 2001.

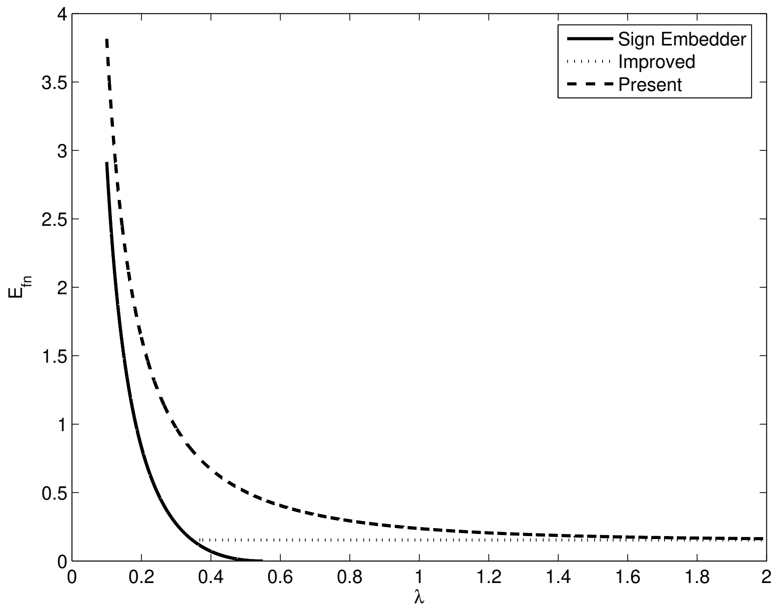

Figure 5. Comparison of the errors exponents obtained for the sign embedder described by Merhav and Sabbag,[12] its improved version, and the technique presented in this work. $\sigma_X^2 = 1$ and $D_e = 2$.