Are you threatening me? Towards smart detectors in watermarking

Mauro Barni[1], Pedro Comesaña-Alfaro[2], Fernando Pérez-González[2], Benedetta Tondi[1]

[1] Dept. Information Theory and Mathematics, University of Siena, Italy
[2] Dept. Teoría de la Señal y Comunicaciones. ETSI Telecom., Universidad de Vigo, 36310 Vigo, Spain

ABSTRACT

We revisit the well-known watermarking detection problem, also known as one-bit watermarking, in the presence of an oracle attack. In the absence of an adversary, the design of the detector generally relies on probabilistic formulations (e.g., Neyman-Pearson's lemma) or on ad-hoc solutions. When there is an adversary trying to minimize the probability of correct detection, game-theoretic approaches are possible. However, they usually assume that the attacker cannot learn the secret parameters used in detection. This is no longer the case when the adversary launches an oracle-based attack, which turns out to be extremely effective. In this paper, we discuss how the detector can learn whether it is being subject to such an attack, and take proper measures. We present two approaches based on different attacker models. The first model is very general and makes minimum assumptions on attacker's beaver. The second model is more specific since it assumes that the oracle attack follows a weel-defined path. In all cases, a few observations are sufficient to the watermark detector to understand whether an oracle attack is on going.

1. INTRODUCTION

Watermarking detection, also known as *zero-rate watermarking* or sometimes *one-bit watermarking*, is one of the most prominent problems in multimedia security, and has received a lot of attention during the past decade. One-bit watermarking differs from its multi-bit counterpart (also referred to as *data hiding*) in that the only information to be extracted from a possibly watermarked object is whether it carries a given watermark or not.

In the absence of an adversary, the standard theory relies on a probabilistic characterization $f_{\mathbf{X}}(\mathbf{x})$ of a typical feature vector $\mathbf{x}$ to formulate the following hypothesis test: given a candidate feature vector $\mathbf{y}$, hypothesis $H_{w,0}$ corresponds to $\mathbf{y}$ being drawn from distribution $f_{\mathbf{X}}(\cdot)$, while hypothesis $H_{w,1}$ corresponds to $\mathbf{y}$ being generated from a vector $\mathbf{x}$ following distribution $f_{\mathbf{X}}(\cdot)$ and later embedding a watermark $\mathbf{w}$. Knowing the distributions of $\mathbf{y}$ under both hypotheses, it is possible to formulate the most powerful statistical test that maximizes the probability P_D of correct detection for a given probability P_F of false alarm, by means of the celebrated Neyman-Pearson lemma. The resulting detector can be subsequently refined by including a noisy channel accounting for possible distortions or manipulations that $\mathbf{y}$ may suffer, as long as a probabilistic description of such channel is available. In those cases where statistical characterizations are lacking, ad-hoc detectors have been proposed which are optimal only in very specific instances, as is the case of the linear correlation detector for additive spread spectrum watermarking, which is only optimal for independent and identically distributed Gaussian hosts and noise.

Moulin and Ivanovic[1] addressed the problem of a smart adversary who tries to minimize the probability of correct detection by modifying $\mathbf{y}$ while satisfying a certain distortion constraint. By taking the presence of an adversary into account, the embedder-detector can try to maximize P_D by carefully designing the embedding and detection functions. This setup naturally leads to a two-player game-theoretic formulation for which Nash

Further author information: (Corresponding author: FPG)
E-mail: fperez@gts.uvigo.es, Telephone: +34 986 812124
barni@dii.unisi.it, pcomesan@gts.uvigo.es, benedettatondi@gmail.com

equilibria can be explicitly found in certain cases. Unfortunately, game-theoretic approaches typically lead to conservative designs, as the pair embedder-detector must account for the worst case of an adversary trying to minimize the performance. In practice, this implies that the achievable P_D for a fixed P_F is significantly reduced even if the adversary is absent (but accounted for).

Furthermore, most existing game-theoretic formulations of the watermarking detection problem optimistically assume that the adversary does not know the secret parameters of the team embedder-detector (e.g., in spread-spectrum watermarking, the watermark itself). In so doing, they disregard the fact that by repeatedly invoking the detector the attacker may learn enough about its shape, so as to put forth very powerful attacks.

More formally, consider the set $\mathcal{Y} \subset \mathbb{R}^L$ of possible feature (column) vectors; then, a typical detector will split $\mathcal{Y}$ into two sets: $\mathcal{R}_{w,0}$ and $\mathcal{R}_{w,1}$, so that if $\mathbf{x} \in \mathcal{R}_{w,i}$, $i = 0, 1$, then hypothesis $H_{w,i}$ is accepted for $\mathbf{x}$. Given a certain object $\mathbf{y} \in \mathcal{Y}$ with some assigned value, the adversary will aim at modifying $\mathbf{y}$ the least possible so that the object retains its value while fooling the detector. Given $\mathbf{y} \in \mathcal{R}_{w,1}$ and a certain real-valued distortion measure $d(\cdot, \cdot)$, the adversary aims at finding some $\mathbf{y}' \in \mathcal{R}_{w,0}$ such that $d(\mathbf{y}, \mathbf{y}')$ is minimal[*]. If the distortion function is convex and has a minimum at $\mathbf{y}' = \mathbf{y}$, and we assume that $\mathcal{R}_{w,1}$ is an open set, then a solution $\mathbf{y}^*$ must lie on the boundary $\delta\mathcal{R}_w$. If the decision function is known to the adversary, then the solution can be either obtained in closed-form or numerically. When the detection function ϕ_w is not fully known, the adversary may try to solve the problem by querying the detector to learn as much as possible about ϕ_w or, better yet, $\delta\mathcal{R}_w$, to later generate $\mathbf{y}^*$. Of course, the feasibility of this solution depends on the number of queries that can be made, as in some cases the system to be attacked will stop accepting them after a number of trials. Attacks based on the information gathered by querying the detector are known as *oracle attacks.*

The simplest oracle attack, the original *sensitivity attack*[2] is suitable when $\delta\mathcal{R}_w$ is a hyperplane. It starts with a vector $\mathbf{y} \in \mathcal{R}_{w,1}$ and modifies it to $\mathbf{z} \in \mathcal{R}_{w,1}$ near the boundary $\delta\mathcal{R}_w$. Then, it works by changing one component of $\mathbf{z}$ at a time and observing the output of the decision function to learn the normal vector that represents the hyperplane. For more complicated decision boundaries, Linnartz and Van Dijk[3] propose an iterative approach which moves along the hyperplane tangent to the decision boundary at $\mathbf{z}$. Another iterative approach is proposed by Mansour and Tewfik[4] where an algorithm akin to the well-known Least Mean Squares (LMS) is applied; this algorithm was shown to be suitable to attack non-linear boundaries. Choubassi and Moulin[5] propose to obtain the normal vector similarly to the sensitivity attack; from this it is immediate to obtain the approximate gradient vector at $\mathbf{z}$. Knowledge of this vector suffices to obtain a good local estimate of the decision function ϕ_w, as long as the adversary knows its form (in this case, the unknown parameters are the watermark samples). Comesaña *et al.*[6,7] present a powerful variant of the sensitivity attack which implements Newton's descent algorithm to iteratively find $\mathbf{y}^*$. The algorithm is completely blind, in the sense that no knowledge of the decision function is assumed; the first and second order local derivatives information required by the iterative algorithm is estimated by querying the detector. The algorithm, termed Blind Newton Sensitivity Attack (BNSA) has been proven very effective in removing the watermark and creating forgeries for a number of existing schemes; moreover, it has been used in the winning strategy in the popular BOWS contest organized by the watermarking community to measure the effectiveness of query-based attacks.[8]

In view of the previous attacks, it seems reasonable to complicate the function ϕ_w. Several works have proposed solutions along this line: Mansour and Tewfik[4] suggest to *fractalize* the decision boundary in an attempt to hamper the use of learning algorithms; Linnartz and Van Dijk[3] propose to *randomize* the boundary so that for points $\mathbf{z}$ close to the boundary, $\phi_w(\mathbf{z})$ is 0 with a certain probability; both countermeasures can be easily overcome by an adversary respectively using the 'envelope' of the fractal boundary or averaging out the boundary randomness. The decision function can be even implemented in zero-knowledge[9] so that the adversary cannot learn anything but the binary output by querying the detector. Strikingly, this minimum disclosure of information (at most one bit per query) is enough for BNSA to work, especially as most existing proposals, with the exception of the one proposed by Troncoso and Pérez-González,[10] use simple decision boundaries.

The main drawback of adopting intricate decision boundaries is that they are difficult to parameterize and, consequently, to put to work in practice. Quite interestingly, and to the best of our knowledge, a path that

[*] Similarly, the attacker may be interested in solving the converse problem, in which $\mathbf{y} \in \mathcal{R}_{w,0}$ and $\mathbf{y}' \in \mathcal{R}_{w,1}$. The mathematical formulation is identical.

has not been trodden is the use of smart detectors. A smart detector is defined as a detector that is able to *learn from* and *react to* repeated query attacks. Notice that detectors producing a random output close to the boundary are not smart according to the previous definition, because they are not able to determine whether they are being subject to an attack. On the other hand, when the smart detector decides that an oracle attack is being launched, it can take further measures, such as precluding further access to the detector, or conservatively switching to a more convoluted detection function.

This paper addresses the problem of designing a smart detector that understands whether an oracle attack is on going or not. We will focus only on the detection part, leaving the investigation of the countermeasures to be adopted in case of attacks for future research. As we will discuss in Section 2, the design of a smart detector requires that some assumptions are made regarding the strategy adopted by the attacker to launch the oracle attack. Specifically, we will study two different alternatives. In a first case, we will try to minimize the knowledge the detector has on the attacking strategy and will only assume that such a strategy results in an unusually large number of queries close to the detection boundary. This is a reasonable assumption since, to the best of our knowledge, all oracle attacks proposed so far work by looking for queries close to $\delta\mathcal{R}_w$. The second alternative comes from noticing that several attacks (and remarkably those proposed by Choubassi and Moulin[5] and Comesaña et al.[8]) assume the availability of some vector $\mathbf{t}$ in $\mathcal{R}_{w,1}$ such that given $\mathbf{y} \in \mathcal{R}_{w,0}$, there is some $\lambda \in (0,1)$ for which $\lambda\mathbf{t} + (1-\lambda)\mathbf{y} \in \delta\mathcal{R}_w$. This value is found through a binary line search, implying that a few queries are made where the submitted features lie on the same straight line. Both kind of detectors are able to detect whether an oracle attack is going on with only a few observations.

The rest of the paper is organized as follows. In Sect. 2, we give an exact formulation of the problem and highlight the importance of defining a model to describe the attacker's behavior. In Sect. 3, we describe our first oracle attack detector, for which only a few assumptions are made on attacker's behavior. A more sophisticated detector is described in Sect. 4, where we assume that the oracle attack incorporates a line search scheme to find queries lying on $\delta\mathcal{R}_w$. The paper ends in Sect. 5 with some conclusions and highlights for future research.

2. PROBLEM FORMULATION

We are interested in constructing a smart detector that is able to decide whether an adversary is launching an oracle-attack to the system. Notice that we actually seek a *metadetector* that works at a higher level than the watermark detector. To distinguish between both, we will use the subindices w and q to denote the latter and the former, respectively. The metadetector will be based on N consecutive observed queries (L-dimensional column vectors) $\mathbf{y}_1, \cdots, \mathbf{y}_N$, which, for convenience, we will stack in a single column vector $\mathbf{y}^N \in \mathcal{Y}^N \subset \mathbb{R}^{LN}$; this vector is built by locating at its first N components the first component of each $\mathbf{y}_i$, $i = 1, \ldots, N$, then the second component of each query, and so on. The metadetector will output $\phi_q(\mathbf{y}^N) = 0$ if $\mathbf{y}^N$ is deemed a legitimate sequence of queries, and 1 otherwise. Function ϕ_q will partition the feature space $\mathcal{Y}^N$ into two sets $\mathcal{R}_{q,0}$ and $\mathcal{R}_{q,1}$, containing those $\mathbf{y}^N$ for which $\phi_q(\mathbf{y}^N) = 0$ and $\phi_q(\mathbf{y}^N) = 1$, respectively. In general, for a legitimate sequence, individual queries will correspond to both watermarked and non-watermark contents, but they will exhibit no signs of being generated by an adversary. For instance, it is reasonable to assume that individual legitimate queries will be mutually independent.

In practice, two parameters need be taken into account, and fixed depending on the actions that follow a possible adversarial detection. For instance, if the user is blocked from accessing the watermark detector once he/she has been labeled as an adversary, then it is important to set a very low false alarm probability (i.e., the probability that a legitimate user is misclassified), while achieving the largest possible probability of correct detection (i.e., the probability that an adversary is correctly detected). On the other hand, if the purpose of the test is to further monitor the behavior of a potential adversary, then a large probability of correct detection is desired, while trying to make the false alarm probability as small as possible. In the sequel, we will focus on the first case, but the methodology can be straightforwardly adapted to the second one. Formally, we want to fix

$$P_{F,q} = \int_{\mathcal{R}_{q,1}} f_{\mathbf{Y}^N|H_{q,0}}(\mathbf{y}^N|H_{q,0})d\mathbf{y}^N, \qquad (1)$$

where hypothesis $H_{q,0}$ refers to a legitimate sequence of queries. It is important to stress that fixing $P_{F,q} > 0$ alone gives enough degrees of freedom in the partition of $\mathcal{Y}^N$ into $\mathcal{R}_{q,0}$ and $\mathcal{R}_{q,1}$ so that there are an infinite number of detection functions ϕ_q that meet the constraint. One possible and popular choice is to minimize the *volume* of $\mathcal{R}_{q,0}$ by including those $\mathbf{y}^N$ such that $f_{\mathbf{Y}^N|H_{q,0}}(\mathbf{y}^N|H_{q,0})$ is larger. But we must notice that this choice is completely arbitrary in the sense that it does not take into account the adversary's behavior. It is only when we consider this behavior in the form of the alternative hypothesis $H_{q,1}$ that we can attempt to maximize the probability of detection for a fixed false alarm probability (this is done, once again, through Neyman-Pearson's test). To illustrate with an example borrowed from a different field, consider tests of randomness for binary number generators. These tests often reject generators that output sequences with very long runs of ones (or zeros), and it is common belief that this is done on the basis that the probability that a truly random generator outputs such sequence is very small. In fact, for any unbiased, independent-samples generator, *every* possible sequence of length n has the same probability (i.e., 2^{-n}) of occurrence! One may ask then what is the basis for rejecting those sequences with long runs. The answer is that there is often an underlying alternative hypothesis that assumes a generator which outputs dependent samples, but if the generator were *known* to produce independent samples, then it might be unfairly rejected if those sequences show up. We conclude that (probabilistically) modeling the adversary's actions and including them in the alternative hypothesis $H_{q,1}$ is crucial in designing a smart detector that performs as expected, at the price of possibly overlooking some potentially harmful actions. Put it another way, a smart detector requires knowledge of $f_{\mathbf{Y}^N|H_{q,1}}(\mathbf{y}^N|H_{q,1})$, which in turn implies prior assumptions (and statistical modeling) on the adversary's actions. Under this perspective, minimizing the *volume* of $\mathcal{R}_{q,0}$ can be seen to be equivalent to assuming an adversary that generates sequences of queries $\mathbf{y}^N$ that are uniformly distributed over $\mathcal{Y}^N$.

On the other hand, it is often the case that a full statistical characterization of the adversary's behavior is not possible or that the resulting Neyman-Pearson detector is not practical. Both problems can be somewhat circumvented by constraining the shape of the detection region, imposing that the detector bases its analysis on some particular property. This is usually achieved by a dimensionality reduction, as it is done for instance in[11, 12] , where the detector (therein called a *detector with limited resources*) is forced to make its decision by relying on first order statistics only. To succeed, the sought properties must be such that it should be possible to effectively discriminate between the legitimate actions and those of the adversary. In this sense, the mere selection of the properties implicitly amounts to making assumptions on the adversary's behavior, to the extent that the remaining degrees of freedom after the dimensionality reduction together with $P_{F,q}$ may be enough to completely determine the metadetector. Generally, the price to be paid for the simplicity of this property-based metadector is its suboptimality.

In this paper we will consider the two approaches outlined in this section. The first one, which is discussed in Section 3, assumes that as a consequence of an oracle attack the number of queries falling very close to the detection boundary is very high. Thus, this metadetector bases the decision on a single property: the distance to $\delta\mathcal{R}_w$, which for additive spread spectrum can be evaluated after reducing the problem to one dimension by projecting the query onto the watermark subspace. The second approach, discussed in Section 4, models adversarial queries that are located on a straight line, as it occurs when a line search is being used to locate points on $\delta\mathcal{R}_w$. The model used in this case is richer, as no dimensionality reduction is applied, but also impractical unless further simplifications are made. Fortunately, the formalization of the metadetector easily leads to suggestions on how those simplifications should be made.

3. DETECTION BASED ON THE CLOSENESS TO THE BOUNDARY

A unifying characteristic of all oracle attacks proposed so far in the field of watermarking (but not only; a similar observation, for instance, holds for the attack to spam filters proposed by Lowd and Meek[13]) is that they rely on the discovery of queries that are very close to the detection boundary. In this section we show how a simple yet powerful oracle attack detector can be built by relying only on this assumption. We will do so for a particularly simple watermarking system; however, a similar approach can be used for systems with more complicated detection regions. Specifically, we will assume that an additive spread spectrum watermarking method is used, for which $\mathbf{x}_w = \mathbf{x} + \gamma\mathbf{w}$, where $\mathbf{x}_w$ is the watermarked signal, γ defines the watermark strength, and the watermark sequence $\mathbf{w}$ takes values in $\{-1, +1\}^L$. A simple detector for additive spread spectrum watermarking relies on the correlation between the to-be-checked sequence $\mathbf{y}$ and the watermark sequence $\mathbf{w}$:

$$\rho = \langle \mathbf{y}, \mathbf{w} \rangle. \tag{2}$$

From basic watermarking theory[14] , we know that ρ can be modeled as a Gaussian random variable[†] with mean $\mu_{\rho|0} = 0$ under $H_{w,0}$ and $\mu_{\rho|1} = \gamma L$ under $H_{w,1}$. As to the variance of ρ, in the noiseless case we have $\sigma_\rho^2 = L\sigma_X^2$, under both hypothesis.

In the above system, $\delta\mathcal{R}_w$ is a hyperplane, as in the system considered by the original sensitivity attack[2,6,7] . More specifically, the decision function $\phi_w(\cdot)$ splits the space in region $\mathcal{R}_{w,0} = \{\mathbf{y} : \langle \mathbf{y}, \mathbf{w} \rangle \leq T\}$ and $\mathcal{R}_{w,1} = \overline{\mathcal{R}_{w,0}}$, where T is the decision threshold determined by the maximum false alarm $P_{F,w}^*$, that is by the equation

$$P_{F,w}^* = \int_T^\infty f_\rho(\rho|H_{w,0})d\rho. \tag{3}$$

The oracle attack detector relies on the definition of a narrow strip across the hyperplane $\langle \mathbf{y}, \mathbf{w} \rangle = T$, namely :

$$\mathcal{A} = \{\mathbf{y} : T - \Delta < \langle \mathbf{y}, \mathbf{w} \rangle < T + \Delta\}, \tag{4}$$

where Δ determines the width of the stripe. The assumption behind the oracle attack detector is that a dishonest user will query the detector with an unusually large number of vectors falling within $\mathcal{A}$. Given a vector with N observations (queries), we define a test in which the metadetector makes a decision based on the number of $\mathbf{y}_i$'s $\in \mathcal{A}$. More precisely, the test is defined by the following decision function:

$$\phi_q(\mathbf{y}^N) = \begin{cases} 0 & \text{if } K < \alpha \cdot N \\ 1 & \text{if } K \geq \alpha \cdot N, \end{cases} \tag{5}$$

where $K = \#\{i : \mathbf{y}_i \in \mathcal{A}\} = n_y(\mathcal{A})$ and α is a given percentage of occurrences. For a fixed number of queries (that is, for a fixed N), we must set Δ and α in such a way to satisfy the constraint on the false positive error probability. Then, by assuming a given attack strategy, we will be able to compute the missed detection probability, thus determining both the significance level and the power of the test.

3.1 False alarm probability

In order to compute $P_{F,q}$ we must define a proper model for normal queries. In order to make the analysis mathematically tractable, we will adopt a Gaussian model. More complicated models can be treated by means of numerical analysis.

Definition 3.1 (Model of legitimate queries). *We consider that legitimate users can send two kinds of queries, corresponding to watermarked and non-watermarked signals. We model the former by $\mathcal{N}(\mathbf{w}, \sigma_{X_i}^2 \boldsymbol{I}_{L\times L})$, where the watermark $\mathbf{w}$ is known at the detector. On the other hand, the non-watermarked signals are assumed to follow a $\mathcal{N}(\mathbf{0}, \sigma_{X_i}^2 \boldsymbol{I}_{L\times L})$. In both cases the variance may be different for each query, so we write $\sigma_{X_i}^2 \in \mathbb{R}^+$ with $i = 1, \ldots, N$. Query signals are mutually independent.*

We will also find it useful to define the indicator $\mathbf{s}$ whose components are

$$s_i \triangleq \begin{cases} 1, & \text{if } \mathbf{y}_i \text{ is watermarked} \\ 0, & \text{otherwise} \end{cases},$$

where $i = 1, \ldots N$, and the corresponding random variable S_i.

Our goal is to compute $P_{F,q} = P(n_y(\mathcal{A}) \geq \alpha \cdot N | H_{q,0})$. To start with, we need to evaluate the probability that an observation $\mathbf{Y}_i$ made by a legitimate user falls inside $\mathcal{A}$. We will do so by assuming that all queries have the same variance. Therefore, index i can be neglected in our computations. Specifically, we can write:

$$P(\mathbf{Y} \in \mathcal{A}|H_{q,0}) = P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 0)p_S(0) + P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 1)p_S(1), \tag{6}$$

[†] The Gaussianity of ρ derives either from a Gaussianity assumption on $\mathbf{x}$ or from the application of the central limit theorem.

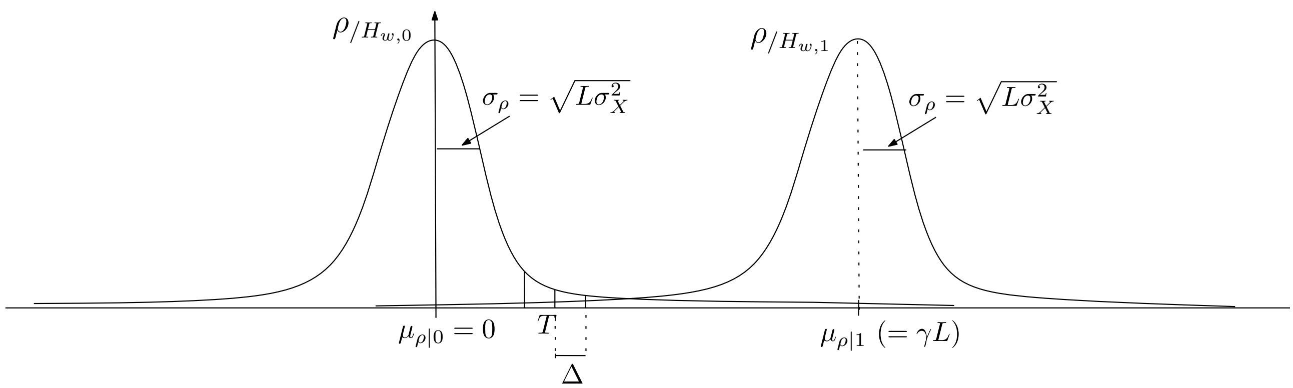

Figure 1. Typical behavior of the probability distributions of the queries under $H_{w,0}$ and $H_{w,1}$. Once the correlation with the watermark is computed, the queries are modeled by Gaussian random variables and the strip across the hyperplane boils down to a segment.

where $p_S(0)$ and $p_S(1)$ are the *a priori* probabilities of having a watermarked or non-watermarked signal under the null hypothesis (legitimate queries). The above probabilities can be redefined as a function of the correlation ρ, since $\mathbf{Y} \in \mathcal{A}$ when $|\rho - T| \leq \Delta$. Consequently, we have:

$$P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 0) = P(|\rho - T| \leq \Delta|H_{q,0}, S = 0) = \frac{1}{\sqrt{2\pi\sigma_\rho^2}} \int_{T-\Delta}^{T+\Delta} e^{\frac{-\rho^2}{2\sigma_\rho^2}} d\rho,$$
$$= Q\left(\frac{T-\Delta}{\sigma_\rho}\right) - Q\left(\frac{T+\Delta}{\sigma_\rho}\right)^{\ddagger}, \tag{7}$$

and similarly,

$$P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 1) = P(|\rho - T| \leq \Delta|H_{q,0}, S = 1) = \frac{1}{\sqrt{2\pi\sigma_\rho^2}} \int_{T-\Delta}^{T+\Delta} e^{\frac{-(\rho-\gamma L)^2}{2\sigma_\rho^2}} d\rho,$$
$$= Q\left(\frac{\gamma L - (T+\Delta)}{\sigma_\rho}\right) - Q\left(\frac{\gamma L - (T-\Delta)}{\sigma_\rho}\right). \tag{8}$$

In the following we will indicate with p_Δ the probability that a query $\mathbf{Y}$ falls inside $\mathcal{A}$.

In a typical watermarking system $\mu_{\rho|1} \gg \sigma_\rho$ and $T \lesssim 5\sigma_\rho$ (for $P^*_{F,w} = 10^{-6}$ we have $T = 4.8\sigma_\rho$), while Δ will be a fraction of σ_ρ. For this reason $P(|\rho - T| \leq \Delta|H_{q,0}, S = 1) < P(|\rho - T| \leq \Delta|H_{q,0}, S = 0)$ and then, for the sake of simplicity, we can take $p_\Delta = P(|\rho - T| \leq \Delta|H_{q,0}, S = 0)$ (this results in a light overestimation of $P_{F,q}$). The situation is depicted in Figure 1. The probability of having K out of N queries in $\mathcal{A}$ can now be obtained by resorting to the formula of repeated Bernoulli trials:

$$P(n_y(\mathcal{A}) = K|H_{q,0}) = \binom{N}{K} p_\Delta^K (1 - p_\Delta)^{N-K}, \tag{9}$$

and then (we assume for simplicity that αN is an integer number):

$$P_{F,q} = \sum_{K=\alpha N}^{N} \binom{N}{K} p_\Delta^K (1 - p_\Delta)^{N-K}. \tag{10}$$

For large N, using Stirling's approximation for the binomial coefficient, we could derive upper and lower bounds for $P_{F,q}$. However, as we will see later, small values of N are be sufficient to build an effective test, so we can actually compute the value of (10).

[‡] $Q(\cdot)$ is the Q-function (tail probability of the standard normal distribution): $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$.

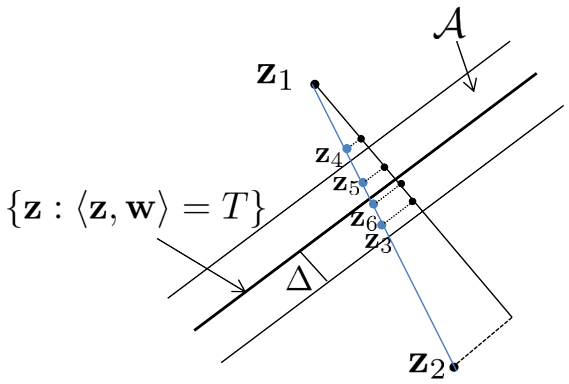

Figure 2. Graphical illustration of the binary line search procedure. The width of the strip Δ is exaggerated for graphical purposes.

Given the target false alarm probability of the metatest $P^*_{F,q}$, we must choose Δ and α such that

$$P_{F,q} = \sum_{K=\alpha N}^{N} \binom{N}{K} p_\Delta^K (1 - p_\Delta)^{N-K} \leq P^*_{F,q}. \tag{11}$$

Clearly, there are many combinations of parameters Δ and α which lead to the same value of $P_{F,q}$. The optimum choice can be made by considering the missed detection probability $P_{M,q}$, for which we need to specify the behavior of the attacker (i.e., the query model under hypothesis $H_{q,1}$).

3.2 Missed detection probability (under a line search attack)

In order to calculate $P_{M,q}$ we need to characterize the queries under $H_{q,1}$. Specifically, hereafter we assume that the attacker performs a binary line search for estimating a point of the hyperplane. Roughly speaking line search works as follows: given two queries corresponding to a non-watermarked and a watermarked signal, the attacker apply a bisection algorithm on the line identified by the two queries until he finds a point that is arbitrarily close to the boundary. Let $\mathbf{z}_1 \in \mathcal{R}_{w,0}$ be the first of such queries and $\mathbf{z}_2 \in \mathcal{R}_{w,1}$ the second one. The number of queries required to fall inside $\mathcal{A}$ and then remaining always in $\mathcal{A}$, is at most $\lceil \log_2(|\rho_2 - \rho_1|/\Delta) \rceil$, where $\rho_i = \langle \mathbf{z}_i, \mathbf{w} \rangle$. Figure 2 illustrates the situation.

Then, under $H_{q,1}$, after N observations ($N \geq \log_2(|\rho_2 - \rho_1|/\Delta)$), we have $K \geq N - \log_2(|\rho_2 - \rho_1|/\Delta)$. The oracle attack detection test yields a correct decision (i.e., in favor of $H_{q,1}$) whenever:

$$\log_2 \frac{|\rho_2 - \rho_1|}{\Delta} \leq (1 - \alpha)N. \tag{12}$$

According to equation (12), the test succeeds if the distance between the initial queries along $\mathbf{w}$ is not too large. In order to understand the probability of this event, we need to determine the statistics of the difference $\rho_2 - \rho_1$. To do so we notice that, to launch his attack, the attacker needs to know two sequences $\mathbf{z}_1$ and $\mathbf{z}_2$, the former belonging to $\mathcal{R}_{w,0}$, the second to $\mathcal{R}_{w,1}$. With regard to $\mathbf{z}_1$, the attacker can query the detector with a few non-watermarked sequences until he finds one which belongs to $\mathcal{R}_{w,0}$ (the search will be extremely fast, since the probability that a non-watermarked sequence belongs to $\mathcal{R}_{w,0}$ is very high). As to $\mathbf{z}_2$, this is the watermarked sequence that the attacker would like to attack and which belongs to $\mathcal{R}_{w,1}$ with high probability. The statistics of ρ_1, then, is obtained by conditioning to $H_{w,0}$ and to the event that a non-watermarked sequence belongs to $\mathcal{R}_{w,0}$, that is $f_{\rho_1}(\rho_1) = \mathcal{N}(0, \sigma_\rho^2)$ conditioned to $\rho_1 \leq T$. By construction, the probability that ρ_1 is smaller than T under $H_{w,0}$ is very close to 1, hence we can neglect the conditioning event and assume that $\rho_1 \sim \mathcal{N}(0, \sigma_\rho^2)$. In the same way, we can say that $\rho_2 \sim \mathcal{N}(\gamma L, \sigma_\rho^2)$. By the independence of $\mathbf{z}_1$ and $\mathbf{z}_2$, $\rho_2 - \rho_1$ is still a Gaussian random variable, with mean value equal to $\mu_{\rho|1}$ and variance equal to $2\sigma_\rho^2$.

Given the characteristics of the watermarking system (in terms of γ and L), and given the parameters α and Δ ensuring that $P_{F,q} \leq P^*_{F,q}$, the above equation permits to derive $P_{M,q}$ as follows:

$$\begin{aligned} P_{M,q} &= 1 - Pr\left\{(\rho_2 - \rho_1) \in [-\Delta \cdot 2^{(1-\alpha)N}, +\Delta \cdot 2^{(1-\alpha)N}]\right\} \\ &= 1 - \frac{1}{\sqrt{2\pi\sigma_\rho^2}} \int_{-\Delta \cdot 2^{(1-\alpha)N}}^{\Delta \cdot 2^{(1-\alpha)N}} e^{\frac{-(\rho-\gamma L)^2}{4\sigma_\rho^2}} d\rho. \end{aligned} \tag{13}$$

Since $\mu_{\rho|1} \gg \sigma_\rho$, from equation (13) we can write

$$P_{M,q} = 1 - \frac{1}{\sqrt{2\pi\sigma_\rho^2}} \int_{-\infty}^{\Delta \cdot 2^{(1-\alpha)N}} e^{\frac{-(\rho-\gamma L)^2}{4\sigma_\rho^2}} d\rho = Q\left(\frac{\Delta \cdot 2^{(1-\alpha)N} - \gamma L}{\sqrt{2\sigma_\rho^2}}\right). \tag{14}$$

Equation (14) enables to optimize the values of α and Δ. As expected, doubling the width of the strip has the same effect on $P_{M,q}$ as decreasing α of $1/N$. Be aware, however, that while for the computation of the false positive error probability we did not make any additional assumption on the behavior of the attacker, the optimization with respect to $P_{M,q}$, is valid only under the line search model.

Example 1: Let us take DWR = 20dB ($\gamma^2 = 10^{-2}\sigma_X^2$), $L = 2 \cdot 10^4$ and $T = 4.5\sigma_\rho$ leading to $P_{F,w} = 3.34 \cdot 10^{-6}$, $\sigma_\rho^2 = L\sigma_X^2 = 2 \cdot 10^4 \sigma_X^2$ and $\mu_{\rho|1} = \gamma L = 2 \cdot 10^3 \sigma_X$, which are rather typical values for an additive spread spectrum watermarking system. Then, consider the following setting: $N = 8$, $\alpha = 1/4$, and $\Delta = 0.33\sigma_\rho$; we obtain:

$$\begin{aligned} p_\Delta =& Q\left(\frac{T-\Delta}{\sigma_\rho}\right) - Q\left(\frac{T+\Delta}{\sigma_\rho}\right) \\ =& Q\left(\frac{4.5\sigma_\rho - 0.33\sigma_\rho}{\sigma_\rho}\right) - Q\left(\frac{4.5\sigma_\rho + 0.33\sigma_\rho}{\sigma_\rho}\right) = Q(4.17) - Q(4.83) \cong 1.46 \cdot 10^{-5}, \end{aligned} \tag{15}$$

finally, yielding, $P_{F,q} = 6 \cdot 10^{-9}$. At the same time, from equation (14) we have $P_{M,q} = 2.4 \cdot 10^{-7}$.

The result of the above computation for different values of N is reported in Table 1. More in general, we may look for the pair (Δ/σ_ρ, α) that for a given value of N and approximately the same $P_{F,q}$, results in the minimum $P_{M,q}$. An example of such an approach is given in Table 2 for $N = 8$. In this case the minimum of the missed detection probability is achieved when $\Delta = 1.19\sigma_\rho$ and $\alpha N = 3$.

Table 1. Behavior of $P_{F,q}$ and $P_{M,q}$ (line search case) for various values of N ($\Delta/\sigma_\rho = 0.33$, $\alpha = 0.25$).

	N					
	5	6	7	8	9	10
$P_{F,q}$	$2 \cdot 10^{-9}$	$3.2 \cdot 10^{-9}$	$4.4 \cdot 10^{-9}$	$6 \cdot 10^{-9}$	$2.5 \cdot 10^{-13}$	$3.7 \cdot 10^{-13}$
$P_{M,q}$	$1 - 10^{-16}$	$1 - 10^{-10}$	0.9925	$2.4 \cdot 10^{-7}$	$2.4 \cdot 10^{-7}$	$5 \cdot 10^{-89}$

Table 2. Values of $P_{M,q}$ for all the possible choices of the pair (Δ/σ_ρ, α) for which $P_{F,q}$ is $6 \cdot 10^{-9}$ ($N = 8$).

	(Δ/σ_ρ, α)							
	$(2.4 \cdot 10^{-5}, 1/8)$	$(0.33, 2/8)$	$(1.19, 3/8)$	$(1.76, 4/8)$	$(2.18, 5/8)$	$(2.54, 6/8)$	$(2.7, 7/8)$	$(3.18, 8/8)$
$P_{M,q}$	$1 - 10^{-23}$	$2.4 \cdot 10^{-7}$	$2.6 \cdot 10^{-65}$	$6.7 \cdot 10^{-24}$	0.0075	0.9967	$1 - 10^{-10}$	$1 - 10^{-14}$

3.3 A simplified test

A simple variant of the test considered in the previous section is obtained by letting the detector decide in favor of $H_{q,1}$ if even one single query falls inside $\mathcal{A}$. Accordingly, given the observation vector $\mathbf{y}^{N'}$ the decision function ϕ_q is defined as:

$$\phi_q(\mathbf{y}^{N'}) = \begin{cases} 1 & \text{if } \exists i : \mathbf{y}_i \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}, \tag{16}$$

which is equivalent to the previous test with $\alpha = 1/N'$. Obviously, the width of region $\mathcal{A}$ in this case, let us denote it by Δ' must be much smaller than in the previous section. In addition Δ' should be linked to the minimum accuracy required by the attacker in his attempt to find a point on the boundary of the detection region. That is, we implicitly assume that the maximum error that the attacker admits for the estimation of the boundary is lower than Δ' (otherwise it would be very difficult for the smart detector to reveal the presence of the attack). The error probabilities of the two types can still be computed by relying on the analysis carried out in the previous section, by letting $\alpha = 1/N'$.

Table 3. Probabilities of error of the simplified test with $\Delta' = 0.01\sigma_\rho$.

	N'							
	5	6	7	8	9	10	11	12
$P_{F,q}$	$1.6 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2.2 \cdot 10^{-6}$	$2.5 \cdot 10^{-6}$	$2.8 \cdot 10^{-6}$	$3.2 \cdot 10^{-6}$	$3.5 \cdot 10^{-6}$	$3.8 \cdot 10^{-6}$
$P_{M,q}$	$1 - 10^{-23}$	$1 - 10^{-22}$	$1 - 10^{-21}$	$1 - 10^{-19}$	$1 - 10^{-16}$	$1 - 10^{-10}$	0.961	$2.3 \cdot 10^{-6}$

3.3.1 Performance of the simplified test

If we assume that the general and the simplified tests rely on the same number of queries ($N' = N$) and that they must satisfy the same constraint on the false positive probability, it is easy to show that the missed detection error probability that we obtain with the simplified test is much higher than that obtain with the general test. Indeed, in the simplified case, fixing N' and a maximum value for $P_{F,q}$ already defines Δ'. On the contrary, with the general test, we can exploit the additional degree of freedom provided by α to achieve the target $P^*_{F,q}$ and minimize $P_{M,q}$. Let us consider the following example: assume that $N' = N = 8$ and that $P_{F,q} \simeq 10^{-6}$ (which is a rather large value). The constraint on $P_{F,q}$ forces us to use a very small Δ', that is $\Delta' \leq 0.01\sigma_\rho$ ($P_{F,q} = 1 - (1 - p_{0.01\sigma_\rho})^8 \cong 2.5 \cdot 10^{-6}$). As a consequence, we get $P_{M,q} = 1 - 10^{-19}$.

Significantly better results can be obtained with the general test. In this case, in fact, even by using the non-optimized values used in Example 1 ($\Delta = 0.33\sigma_\rho$ and $\alpha = 0.25$), for which $P_{F,q} \cong 6 \cdot 10^{-9}$, the missed detection error probability would be $P_{M,q} = 2.4 \cdot 10^{-7}$.

Of course, this does not mean that the simplified test can not be applied; however, larger values of N' must be used. In Table 3 we report the missed detection error probability of the simplified test for various values of N' (and the same setting used in the previous example). As it can be seen, we need $N' = 12$ in order to obtain a small $P_{M,q}$.

4. LINE SEARCH DETECTION

In this section, we used the knowledge that the attacker uses a line search, to develop a more powerful (even if more complicated) oracle attack detection test. Of course this comes at the price of a loos of generality since the test can be applied only when the attacker adopts a line search strategy for its attack. Note that in the previous section we exploited the same knowledge to evaluate the missed detection probability. The test, however, would remain valid even for other attacking strategies, assuming that, at a certain point, they result in an unnaturally high number of queries close to the boundary.

To design our new test, we need to exploit the knowledge of the attacker strategy to refine the model of illegitimate queries.

Definition 4.1 (Model of illegitimate queries). *Query signals corresponding to illegitimate users are noisy convex combinations of a watermarked signal, which we denote by* $\mathbf{Z}_1 \sim \mathcal{N}(\mathbf{w}, \sigma^2_{Z_1}\boldsymbol{I}_{L\times L})$*, and a non-watermarked signal, denoted by* $\mathbf{Z}_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2_{Z_2}\boldsymbol{I}_{L\times L})$*. Specifically, the i-th query can be written as* $\mathbf{Y}_i = \psi_i\mathbf{Z}_1 + (1-\psi_i)\mathbf{Z}_2 + \mathbf{N}_i$*, where* $0 \leq \psi_i \leq 1$*,* $\mathbf{N}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2_N)$*,* $i = 1, \ldots, N$*. Since* $\mathbf{N}_i$ *is typically used for modelling quantization effects in a transformed domain, we will assume that: 1)* $\mathbf{N}_i$ *is independent of* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$*; 2) the* $\mathbf{N}_i$*'s are mutually independent, and identically distributed; and 3)* σ^2_N *is known by the detector. Furthermore,* $\sigma^2_{Z_i} \in \mathbb{R}^+$*,* $i = 1, 2$*, and* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *are mutually independent.*

As in the previous section, the null hypothesis can be formalized as $H_{q,0} : \mathbf{Y}^N \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, where each component of the mean vector $\boldsymbol{\mu}_0$ is set to $\mathbf{0}$ or to the corresponding component of $\mathbf{w}$, depending if the query corresponds to a watermarked signal or not, and the covariance matrix Σ_0 is a diagonal matrix whose elements are obtained by repeating L times the vector $(\sigma^2_{X_1}, \cdots, \sigma^2_{X_N})$. Similarly, the alternative hypothesis is $H_{q,1} : \mathbf{Y}^N \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, where the mean vector $\boldsymbol{\mu}_1$ contains the mean vectors $\psi_i \mathbf{w}$, $i = 1, \ldots, N$, and the covariance matrix Σ_1 is a block diagonal matrix, with block-size $N \times N$, which is repeated L times. Therefore, the null hypothesis distribution can be parameterized by $\mathbf{s}$ and $\sigma^2_{X_i}$, $i = 1, \ldots, N$, and the alternative hypothesis distribution by $\psi_i \in [0, 1]$, $i = 1, \ldots, N$, $\sigma^2_{Z_1}$, and $\sigma^2_{Z_2}$.

4.1 Detector

Due to the presence of the nuisance parameters mentioned before, both the null hypothesis and the alternative hypothesis are *composite hypotheses*; therefore, Neyman-Pearson criterion can not be directly applied to the considered detection problem. Instead, we will use the so called *generalized likelihood ratio test* (GLRT), defined as

$$\Lambda(\mathbf{y}^N) = 2 \log \left(\max_{\boldsymbol{\theta}_1} f_{\mathbf{Y}^N | H_{q,1}}(\mathbf{y}^N | \boldsymbol{\theta}_1) \right) - 2 \log \left(\max_{\boldsymbol{\theta}_0} f_{\mathbf{Y}^N | H_{q,0}}(\mathbf{y}^N | \boldsymbol{\theta}_0) \right) \gtrless \tau,$$

where $\boldsymbol{\theta}_0$ is the concatenation of $\mathbf{s}$ and $\sigma^2_{X_i}$, $i = 1, \ldots, N$, and $\boldsymbol{\theta}_1$ is the concatenation of ψ_i, $i = 1, \ldots, N$, $\sigma^2_{Z_1}$, and $\sigma^2_{Z_2}$.

Exploiting the nature of the pdfs involved in the current problem, one can write

$$\begin{aligned} \Lambda(\mathbf{y}^N) &= \min_{\boldsymbol{\theta}_0 \in \{0,1\}^N \times (\mathbb{R}^+)^N} \left[(\mathbf{y}^N - \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\mathbf{y}^N - \boldsymbol{\mu}_0) + \log(|\Sigma_0|) \right] \\ &\quad - \min_{\boldsymbol{\theta}_1 \in [0,1]^N \times (\mathbb{R}^+)^2} \left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|) \right]. \end{aligned} \tag{17}$$

Therefore, the only remaining issue is to determine the value of the decision threshold τ in order to verify that the probability of false alarm when the GLRT is used is smaller than or equal to a target value $P^*_{F,q}$, i.e., $P_{F,q} = P(\Lambda(\mathbf{Y}^N) \geq \tau | H_{q,0}) \leq P^*_{F,q}$.

In order to do so, we define

$$\begin{aligned} \Lambda_1(\mathbf{y}^N) &\triangleq \min_{(\sigma^2_{X_1}, \ldots, \sigma^2_{X_N}) \in (\mathbb{R}^+)^N} \left[(\mathbf{y}^N - \hat{\mathbf{S}} \mathbf{w}^N)^T \Sigma_0^{-1} (\mathbf{y}^N - \hat{\mathbf{S}} \mathbf{w}^N) + \log(|\Sigma_0|) \right] \\ &\quad - \min_{\boldsymbol{\theta}'_1 \in [0,1]^{2N} \times (\mathbb{R}^+)^2} \left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|) \right], \end{aligned} \tag{18}$$

where in the first minimization $\mathbf{w}^N$ denotes the counterpart of $\mathbf{Y}^N$ when $\mathbf{Y}_i$ is replaced by $\mathbf{w}$, vector $\hat{\mathbf{s}}$ is the ML estimate of $\mathbf{s}$, and $\hat{\mathbf{S}}$ is a diagonal matrix with the elements of $\hat{\mathbf{s}}$ repeated L times. On the other hand, in the second optimization in (18) the first N values in $[0, 1]$ are used as ψ_i in the parameterization of $\boldsymbol{\mu}_1$, while the second N values are used as ψ_i in the parameterization of Σ_1. This decoupling, which somewhat serves to simplify the optimization problem, also produces an expanded domain that in turn may lead to a smaller value of the objective function; for this reason, $\Lambda_1(\mathbf{y}^N) \geq \Lambda(\mathbf{y}^N)$.

It is illustrative to interpret the term $(\mathbf{y}^N - \hat{\mathbf{S}} \mathbf{w}^N)^T \Sigma_0^{-1} (\mathbf{y}^N - \hat{\mathbf{S}} \mathbf{w}^N)$ in (18) as a weighted Euclidean distance between the observations and the mean vector that results after making the decision on whether each query corresponds to a watermarked signal or not. Similarly, the term $(\mathbf{y}^N - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{y}^N - \boldsymbol{\mu}_1)$ can be seen as a weighted Euclidean distance. By decomposing $\boldsymbol{\Sigma}_1 = \mathbf{U}^T \mathbf{D} \mathbf{U}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{U}$ is unitary, one can transform the observations and mean vectors so that a diagonal weight matrix is recovered. Interestingly, under hypothesis $\mathcal{H}_{q,1}$ when $\sigma^2_N \ll \sigma^2_{Z,1}, \sigma^2_{Z,2}$, and the true $\boldsymbol{\Sigma}_1$ is used, such diagonalization will lead to only two significant values in each N consecutive elements in the main diagonal of $\mathbf{D}$ (recall that the diagonal of $\mathbf{D}$ will be the repetition of L such blocks). This reflects the fact that the queries are roughly located on a straight line. Therefore, for the proper values of ψ_i in $\boldsymbol{\Sigma}_1$, the second optimization in (18) can be seen as a sort of line fitting. This suggests solving such problem in practice by replacing $\boldsymbol{\Sigma}_1$ by the sample covariance matrix, so that the orthogonal decomposition corresponds to Principal Component Analysis (PCA), which provides the only significant direction.

Under $H_{q,0}$, $\mathbf{Y}^N - \boldsymbol{\mu}_1$ corresponds to $\mathbf{X}_i - \hat{\psi}_i \mathbf{w}$ when $\mathbf{Y}_i$ is not watermarked, and to $\mathbf{X}_i + (1 - \hat{\psi}_i)\mathbf{w}$ when $\mathbf{Y}_i$ is watermarked, where $\hat{\psi}_i = (\boldsymbol{\theta}'_1)_i$, $1 \leq i \leq N$; we will write these relationships as $\mathbf{Y}^N - \boldsymbol{\mu}_1 = \mathbf{X}^N + (\mathbf{S} - \hat{\Psi})\mathbf{w}^N$, where $\mathbf{X}^N$ is constructed by stacking the $\mathbf{X}_i$, and $\mathbf{S}$ and $\hat{\Psi}$ are diagonal matrices with the elements of $\mathbf{s}$ and $\hat{\boldsymbol{\psi}}$ repeated L times. Then, the value of $\Lambda_1(\mathbf{y}^N)$ is

$$\begin{aligned} &\min_{(\sigma^2_{X_1},\ldots,\sigma^2_{X_N}) \in (\mathbb{R}^+)^N} \left[(\mathbf{x}^N + [\mathbf{S} - \hat{\mathbf{S}}]\mathbf{w}^N)^T \Sigma_0^{-1} (\mathbf{x}^N + [\mathbf{S} - \hat{\mathbf{S}}]\mathbf{w}^N) + \log(|\Sigma_0|) \right] \\ &- \min_{\boldsymbol{\theta}'_1 \in [0,1]^{2N} \times (\mathbb{R}^+)^2} \left[(\mathbf{x}^N + (\mathbf{S} - \hat{\Psi})\mathbf{w}^N)^T \Sigma_1^{-1} (\mathbf{x}^N + (\mathbf{S} - \hat{\Psi})\mathbf{w}^N) + \log(|\Sigma_1|) \right]. \end{aligned} \tag{19}$$

Let $\hat{\Psi}^*$ be the diagonal matrix obtained from the first N components of the value of $\boldsymbol{\theta}'_1$ that minimizes the second target function in Λ_1. Also, for a given indicator matrix $\mathbf{S}' \neq \mathbf{S}$, let $\boldsymbol{\Delta}$ be the diagonal matrix that has ones in those positions where $\mathbf{S}'$ and $\mathbf{S}$ differ, and zeros otherwise. Then, for this matrix $\mathbf{S}'$ it can be seen that the diagonal matrix $(\hat{\Psi}^*)'$ obtained from the first N components of $\boldsymbol{\theta}'_1$ that minimizes (19) is $(\mathbf{I} - \boldsymbol{\Delta})\hat{\Psi}^*$. In fact it can be shown that $(\mathbf{x}^N + (\mathbf{S} - \hat{\Psi}^*)\mathbf{w}^N)^T \Sigma_1^{-1} (\mathbf{x}^N + (\mathbf{S} - \hat{\Psi}^*)\mathbf{w}^N) = ((\mathbf{x}^N)' + (\mathbf{S}' - (\hat{\Psi}^*)')\mathbf{w}^N)^T \Sigma_1^{-1} ((\mathbf{x}^N)' + (\mathbf{S}' - (\hat{\Psi}^*)')\mathbf{w}^N)$, where $(\mathbf{x}^N)' \triangleq (\mathbf{I} - 2\boldsymbol{\Delta})\mathbf{x}^N$. In other words, due to symmetry of the pdf of $\mathbf{X}_i$ with respect to componentwise sign changes, the pdf of the second term of $\Lambda_1(\mathbf{Y}^N)$ is indeed independent of $\mathbf{s}$.

In order to remove the dependence of the first optimization with respect to $\mathbf{s}$, we exploit that $|s_i - \hat{s}_i| \leq 1$, $1 \leq i \leq N$, to define

$$\begin{aligned} \Lambda_2(\mathbf{x}^N) \quad \triangleq \quad & \min_{(\sigma^2_{X_1},\ldots,\sigma^2_{X_N}) \in (\mathbb{R}^+)^N} \left[(|\mathbf{x}^N| + |\mathbf{w}^N|)^T \Sigma_0^{-1} (|\mathbf{x}^N| + |\mathbf{w}^N|) + \log(|\Sigma_0|) \right] \\ & - \min_{\boldsymbol{\theta}'_1 \in [0,1]^{2N} \times (\mathbb{R}^+)^2} \left[(\mathbf{x}^N - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{x}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|) \right], \end{aligned}$$

where $|\mathbf{x}|$ denotes, with some abuse of notation, the componentwise absolute value, i.e., $(|\mathbf{x}|)_i = |x_i|$. Note that $P(\Lambda_2(\mathbf{x}^N) \geq \tau) \geq P(\Lambda_1(\mathbf{x}^N + \mathbf{S}\mathbf{w}^N) \geq \tau)$ for any τ and $\mathbf{s}$.

Finally, we will calculate τ in order to verify that $P(\Lambda_2(\mathbf{X}^N) \geq \tau) = P^*_{F,q}$, implying that

$$P_{F,q} \quad = \quad P(\Lambda(\mathbf{Y}^N) \geq \tau | H_{q,0}) \leq P(\Lambda_1(\mathbf{Y}^N) \geq \tau | H_{q,0}) \leq P(\Lambda_2(\mathbf{X}^N) \geq \tau) = P^*_{F,q}.$$

Now, the distribution of $\Lambda_2(\mathbf{X}^N)$ does not depend on $\mathbf{s}$. Consequently, we do not need to assume any *a priori* probability of a legitimate query to correspond to a watermarked or non-watermarked content. Based on $\Lambda_2(\mathbf{X}^N)$, the threshold τ can be found by using Monte Carlo simulations.

In order to evaluate the performance of the approach presented in this section, we will consider $\sigma^2_{X_i} \sim U(80, 120)$ and $\sigma^2_W = 1$ (so DWR $= 20$ dB, similarly to the previous section), $\sigma^2_N = 10$, $L = 2 \cdot 10^4$, $N = 3$, and under $H_{q,1}$ $\boldsymbol{\psi} = (0, 0.5, 1)$. By using Monte Carlo simulations we obtain $\mathrm{E}(\Lambda_2 | H_{q,0}) \approx -1.156 \cdot 10^5$, $\mathrm{Var}(\Lambda_2 | H_{q,0}) \approx 1.775 \cdot 10^8$, $\mathrm{E}(\Lambda | H_{q,1}) \approx 2.816 \cdot 10^4$, $\mathrm{Var}(\Lambda | H_{q,1}) \approx 1.674 \cdot 10^6$. If we model both the distribution of Λ_2 under $H_{q,0}$, and that of Λ under $H_{q,1}$ to be Gaussian (which is reasonable, given the large value of L), then the decision threshold for $P^*_{F,q} = 10^{-20}$ is $\tau = 7807$, yielding a probability of missed detection $P_{M,q} = Q(15.74) \approx 4.34 \cdot 10^{-56}$. This result illustrates the good behavior of the proposed approach.

5. CONCLUSIONS

As a new way to contrast oracle attacks against one-bit watermarking, we have introduced the concept of smart detector, i.e., a detector that first tries to understand whether an attack is on going and then reacts properly. We have investigated the possibility of developing an effective metatest to see whether the queries submitted to the detector provide enough evidence that an attack is in place. We did so for two different settings making different assumptions on the amount of knowledge the detector has on the query pattern used by the attacker. In both cases, the developed detectors permit to reveal the presence of an oracle attack by observing very few queries, even if, expectedly, the more knowledge is available to the detector the more effective the test. It goes without saying that several aspects need further attention. First of all, we should take into account the possibility that the attacker adopts a query pattern explicitly thought to minimize its detectability by means of a smart detector, which in turn could define its strategy by assuming that proper countermeasures are taken by the attacker. To understand who will finally win this kind of tug of war, the smart detector problem can be casted in a game-theoretic framework following the path suggested by Barni and Pérez-González[15] . Secondly, the most suitable strategy to be undertaken once an oracle attack is detected should be defined.

We conclude by observing that, even if we focused on oracle attacks launched against a watermarking system, the analysis presented in this paper goes well beyond the watermarking problem: as discussed by Barni and Pérez-González,[15] binary decision can be considered as on of the core problems in *adversarial signal processing*, so the advances presented here find their application in other fields, such as biometrics, network intrusion detection, forensics, reputation systems, and many more.

REFERENCES

1. P. Moulin and I. Ivanovic, "The Zero-Rate Spread-Spectrum Watermarking Game," *IEEE Transactions on Signal Processing*, vol. 51, pp. 1098–1117, April 2003.
2. I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *IEEE International Conference on Image Processing, ICIP'97*, vol. 3, (Santa Barbara, California, USA), pp. 3–6, October 1997.
3. J. P. M. G. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *2nd International Workshop on Information Hiding, IH'98* (D. Aucsmith, ed.), vol. 1525 of *Lecture Notes in Computer Science*, (Portland, OR, USA), pp. 258–272, Springer Verlag, April 1998.
4. M. F. Mansour and A. H. Tewfik, "LMS-based attack on watermark public detectors," in *IEEE International Conference on Image Processing, ICIP'02*, vol. 3, (Rochester, NY, USA), pp. 649–652, September 2002.
5. M. E. Choubassi and P. Moulin, "Noniterative Algorithms for Sensitivity Analysis Attacks," *IEEE Transactions on Information Forensics and Security*, vol. 2, pp. 113–126, June 2007.
6. P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "The return of the sensitivity attack," in *4th International Workshop, IWDW'05* (M. Barni, I. Cox, T. Kalker, and H. J. Kim, eds.), vol. 3710 of *Lecture Notes in Computer Science*, (Siena, Italy), pp. 260–274, Springer Verlag, September 2005.
7. P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind Newton sensitivity attack," in *IEE Proceedings on Information Security*, vol. 153, pp. 115–125, IET, September 2006.
8. P. Comesaña and F. Pérez-González, "Breaking the BOWS watermarking system: key guessing and sensitivity attacks," *EURASIP Journal on Information Security*, vol. 2007, 2007.
9. A. Adelsbach and A.-R. Sadeghi, "Zero-knowledge watermark detection and proof of ownership," in *4th International Workshop on Information Hiding, IH'01* (I. S. Moskowitz, ed.), vol. 2137 of *Lecture Notes in Computer Science*, (Pittsburgh, PA, USA), pp. 273–288, Springer Verlag, April 2001.
10. J. Troncoso-Pastoriza and F. Pérez-González, "Zero-knowledge watermark detector robust to sensitivity attacks," in *Proceedings of the 8th workshop on Multimedia and security*, (Portland, OR, USA), pp. 97–107, ACM, April 2006.
11. N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Transactions on Information Theory*, vol. 54, pp. 255–274, January 2008.
12. M. Barni and B. Tondi, "The source identification game: an information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 450–463, March 2013.
13. D. Lowd and C. Meek, "Adversarial learning," in *13th ACM SIGKKD International Conference Knowledge Discovery and Data Mining European*, (San Jose, California, USA), pp. 641–647, August 2005.
14. M. Barni and F. Bartolini, *Watermarking Systems Engineering.* Signal Processing and Communications, Marcel Dekker, 2004.
15. M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP'13*, (Vancouver, Canada), pp. 8682 – 8686, May 2013.