

# Noniterative Algorithms for Sensitivity Analysis Attacks

Maha El Choubassi, *Student Member, IEEE*, and Pierre Moulin, *Fellow, IEEE*

## Abstract

Sensitivity analysis attacks constitute a powerful family of watermark “removal” attacks. They exploit a vulnerability in some watermarking protocols: the attacker’s unlimited access to the watermark detector. This paper proposes a mathematical framework for designing sensitivity analysis attacks and focuses on additive spread spectrum embedding schemes. The detectors under attack range in complexity from basic correlation detectors to normalized correlation detectors and maximum likelihood (ML) detectors. The new algorithms precisely estimate and then eliminate the watermark from the watermarked signal. This is done by exploiting geometric properties of the detection boundary and the information leaked by the detector. Several important extensions are presented, including the case of a partially unknown detection function, and the case of constrained detector inputs. In contrast with previous art, our algorithms are noniterative and require at most  $O(n)$  detection operations in order to estimate the watermark, where  $n$  is the dimension of the signal. The cost of each detection operation is  $O(n)$ , hence the algorithms can be executed in quadratic time. The method is illustrated with an application to image watermarking using an ML detector based on a generalized Gaussian model for images.

## Index Terms

Watermarking, security, sensitivity attacks, spread spectrum, generalized Gaussian distribution, maximum likelihood, parametric detector, quantization effects.

This work was supported by NSF under grant CCR 03-25924 and presented in part at the SPIE conference on security, steganography, and watermarking of multimedia contents, San Jose, CA, January 2005.

Both authors are members of Beckman Institute’s Image Formation and Processing Group at the University of Illinois at Urbana-Champaign, 405 N. Mathews, Urbana, IL 61801, USA. Emails: {cel,moulin}@ifp.uiuc.edu, fax: 217-244-8371 (Send correspondence to Maha El Choubassi).

## I. INTRODUCTION

Copyright protection of digital media, together with related applications, has fueled the development of watermarking systems. In many of these applications, security, i.e., the ability to resist intentional attacks, is a core requirement. In this paper, new attacks on spread-spectrum schemes are presented. They belong to a family of attacks called “sensitivity analysis attacks” which are known to be extremely effective for an adversary that has unlimited access to the watermark detector [1]–[9]. In this sense, these attacks are analogous to chosen-cyphertext attacks in cryptography, where the opponent has access to the decryption device but does not know the key [10]. The goal is unauthorized removal of a watermark.

A scenario that is vulnerable to such attacks, is when media players accept both watermarked and unwatermarked copies [3]. Such devices play watermarked commercial digital products as well as unwatermarked products such as home videos. An attacker may then be motivated to remove the watermark from a watermarked copy available to him in order to produce an unlimited number of illegal copies and resell them.

Moreover, in a typical copyright protection watermarking system, the detection algorithm is publicly known. While no one should be able to “remove” the electronic watermark, anyone can detect its presence. In sensitivity analysis attacks, this feature is abused (refer to Figure 1). The attacker makes use of the detector to extract information about the watermark and subsequently “remove” it. The attacker’s goal is to create a pirated copy that is perceptually similar to the original watermarked signal and does not trigger a positive response from the detector. Hence, there is no need to completely remove the watermark: fooling the detector is enough [4].

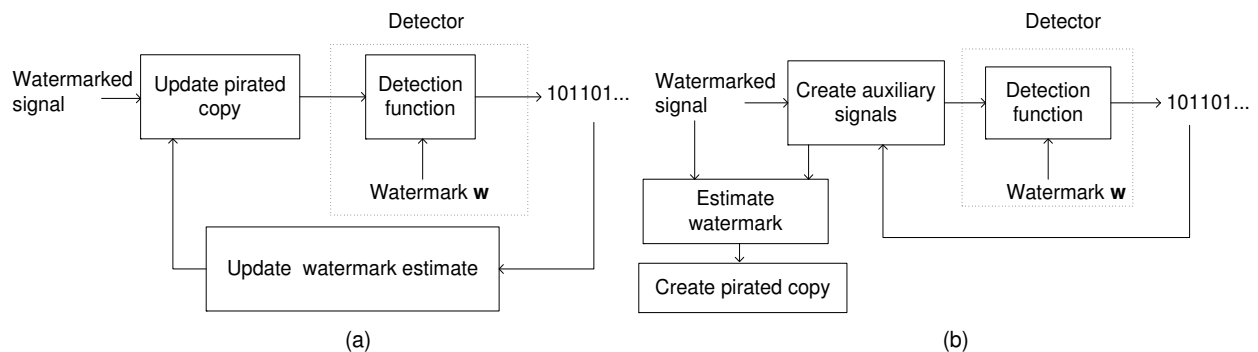


Fig. 1. Detector outputs 0 and 1 indicate watermark absent and present, respectively. The attacker uses the detector as a black box, to estimate the watermark and create a pirated copy. (a) Previous work: the algorithm is iterative and applies to correlation detectors. (b) Our approach: the algorithm is noniterative and applies to a broad family of regular detectors. The pirated copy is constructed in the final step of the algorithm, and triggers the response “0” from the detector.

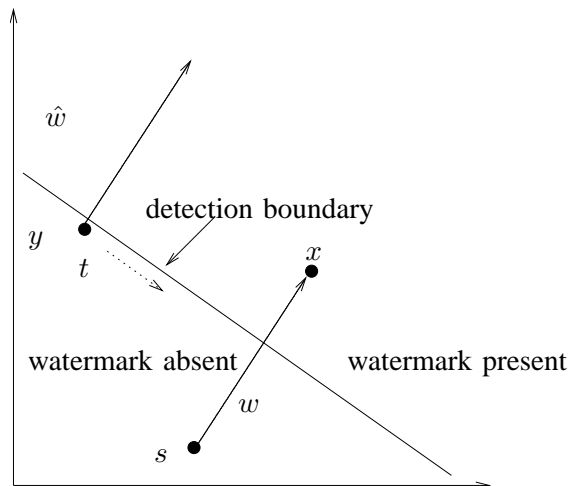


Fig. 2. An illustration of Linnartz's algorithm in the 2-D case.

Sensitivity analysis attacks have been previously addressed by Cox and Linnartz in [1], by Linnartz and Van Dijk [2], by Kalker, Linnartz, and Van Dijk [3], by Tewfik and Mansour [5] and [6], and by Comesãna, Pérez-Freire, and Pérez-González [8].

To our knowledge, Cox and Linnartz [1] were the first to study this problem. They argued that with the aid of a watermark detector and a size- $n$  watermarked image  $\mathbf{x}$ , the attacker should be able to estimate the watermark after  $O(n)$  and not  $O(2^n)$  calls to the detector. Note that sensitivity analysis attacks are only possible because of the repetitive use of the detector. Therefore, if the detection operation itself is of complexity  $O(n)$  (because the detector computes, e.g., a correlation statistic), then the attack method is effectively of complexity  $O(n^2)$ . The attack method in [1] is described at a high level. The attack progressively modifies the watermarked signal into one that is just on the negative side of the decision boundary. For each pixel at a time, the luminance is changed till the detector response changes from watermarked to unwatermarked. At the end of this process, the attacker has a collection of the pixels that largely influence the detector's decision. A correlation type detector is assumed although it is claimed that the attack is still possible with other detectors. In this paper, we explicitly state the steps required by our proposed algorithms to obtain an estimate of the watermark for various detection methods.

For another approach<sup>1</sup> suggested in [2] by Linnartz and Van Dijk, the preliminary step is also to find a signal  $\mathbf{y}$  almost on the decision boundary (see Figure 2). Indeed the basic idea of this approach is to move in the plane tangent to the decision boundary towards  $\mathbf{x}$ . For a correlation detector, the decision

<sup>1</sup>For convenience, this algorithm is denoted as Linnartz's algorithm.

boundary is a hyperplane orthogonal to the watermark  $w$  which can then be estimated and “removed”. For other types of detectors, the attack algorithm requires more iterations. On the contrary, our algorithms are noniterative and they are guaranteed to give a good estimate of the watermark after a finite number of steps. Moreover, we consider real-valued watermarks, while in [2] the watermark is bipolar and only the signs of its components need to be estimated.

Another attack algorithm was proposed in [3] by Kalker et al. Their algorithm is specialized to normalized correlation detectors, is iterative, and considers only bipolar watermarks.

Later, Tewfik and Mansour [6] used the least mean squares (LMS) algorithm to estimate the watermark. For this purpose, the attack requires a sufficient number of signals on the decision boundary. However, the convergence properties of this method remain to be investigated. In the same paper, the authors recommend processing of the detection boundary to make it fractal-like. This is addressed in another paper by the same authors [5]. According to [5], this new boundary cannot be reliably estimated because it is nonparametric.

Finally, another attack was proposed recently by Comesãna et al. in [8]. This attack also uses a numerical method in order to create an unwatermarked signal with minimum Euclidean distortion relative to the watermarked signal  $x$  originally available to the attacker. The numerical method used in [8] is an adaptation of Newton’s method. It is an iterative algorithm, and its computational complexity and convergence properties are currently unknown.

In this paper, we present new algorithms for sensitivity analysis attacks. Table I summarizes the advantages of our new algorithms over the algorithms cited above. The main idea is to exploit the mathematical properties of the detection function and accordingly process the information leaked by the detector to estimate the watermark. For this reason, we study in this paper two general classes of detectors and generate a sensitivity analysis attack algorithm for each class. We first study generalized correlator detectors and provide an algorithm that estimates the watermark in  $n + 1$  steps. Popular detectors in this class are the standard correlation detector, the normalized correlation detector, and the Patchwork detector. Next, we address a broader class of nonlinear detectors, which we call regular detectors. Assuming that the detection boundary is smooth enough, the algorithm locally approximates it by an  $n$ -dimensional hyperplane and obtains the watermark in  $2n + 1$  steps. This class includes a variety of maximum-likelihood (ML) detectors, e.g., based on generalized Gaussian models for the Discrete Cosine Transform (DCT) coefficients of the host image [11].

Next, we study the scenario when a finite set of parameters, such as threshold of the test, or parameters of the ML detector, are unknown to the attacker. We modify our algorithms to fit this scenario and we show

TABLE I  
OUR NEW ALGORITHMS VERSUS PREVIOUS ALGORITHMS.

Algorithm	Characteristics
Cox and Linnartz [1]	Correlation detection assumed, iterative algorithm, real-valued watermark.
Linnartz and Van Dijk [2]	Correlation detection, iterative algorithm, bipolar watermark.
Kalker, Linnartz, and Van Dijk [3]	Normalized correlation detection, iterative algorithm, bipolar watermark.
Tewfik and Mansour [6]	(Iterative) LMS algorithm, real-valued watermark.
Comesãna, Pérez-Freire, and Pérez-González [8]	Iterative based on Newton's method, real-valued watermark.
Our algorithms	Explicit formula for the watermark estimate, noniterative, $O(n)$ detection probes, general/parametric detection methods, real-valued watermark, quantization effects considered.

that their complexity does not increase significantly. Finally, we take into account practical constraints that may be imposed on the detector's input and consequently on the attack algorithm.

This paper is organized as follows. Section II describes the notation used in this paper. Section III presents the assumptions made about the attacker. Section IV presents a new algorithm that recovers the exact watermark in  $n + 1$  steps when the detection statistic is the correlation between the signal and the watermark or a function of it. In Section V, another algorithm is derived that applies to the family of regular detectors. Section VI considers parametric detectors, where the attacker does not know some of parameters of the detection function. In Section VII, we take into account the constraints that result when the detector's inputs are digital images. In Section VIII, we present simulation results to ascertain the performance of our algorithms. Finally, conclusions are presented in Section IX.

## II. DEFINITIONS AND NOTATION

All the signals are represented as  $n$ -dimensional vectors. We denote by  $\mathbf{o}$  the zero vector. Let  $\mathbf{s}$  be the original signal,  $\mathbf{x}$  the watermarked signal, and  $\mathbf{w}$  the watermark, an arbitrary element of  $\mathbb{R}^n$ . Let  $\sigma > 0$  be the strength parameter. The watermarked signal is obtained by additive spread spectrum embedding

of the watermark into the original signal <sup>2</sup>:

$$\mathbf{x} = \mathbf{s} + \sigma \mathbf{w}. \quad (1)$$

For simplicity, assume  $\sigma = 1$  and let the strength of embedding be represented in the magnitude of the watermark  $\mathbf{w}$ .

The detection threshold is  $\tau$ . Given a signal  $\mathbf{y}$ , the detector computes a detection statistic  $t(\mathbf{y}, \mathbf{w})$ . Then, the detector compares  $t(\mathbf{y}, \mathbf{w})$  with  $\tau$ . The decision is

$$d(\mathbf{y}, \mathbf{w}) = \begin{cases} 1, & \text{if } t(\mathbf{y}, \mathbf{w}) > \tau \text{ (Watermark present)} \\ 0, & \text{else (Watermark absent)} \end{cases} \quad (2)$$

Given  $\mathbf{w}$ , the set of all  $\mathbf{y}$  such that  $d(\mathbf{y}, \mathbf{w}) = 1$  is the *acceptance region* of the test; the complementary region is the *rejection region*.

### III. ASSUMPTIONS ABOUT THE ATTACKER

The attacker knows the detection function used  $t(\cdot, \cdot)$  and all the system parameters, including the threshold  $\tau$ . He knows neither the watermark  $\mathbf{w}$ , nor the detection statistic  $t(\mathbf{y}, \mathbf{w})$  for any test signal  $\mathbf{y}$ . However he has unlimited access to the detector and has access to a watermarked signal  $\mathbf{x}$ . Therefore, he can design signals  $\mathbf{y}$  and observe the corresponding binary decision  $d(\mathbf{y}, \mathbf{w})$  in (2). (Section VI extends the algorithm to the case of parametric detectors with unknown parameters including  $\tau$ .)

The attack methods derived in Sections IV and V rely on the following subproblem:

- Given a signal  $\mathbf{v}$  and a direction  $\mathbf{d}$ , the attacker needs to estimate a scalar  $\alpha$  such that the signal  $\mathbf{v} + \alpha \mathbf{d}$  is on the detection boundary, i.e.,  $t(\mathbf{v} + \alpha \mathbf{d}, \mathbf{w}) = \tau$ .

In general, we may write  $\alpha = H(\mathbf{v}, \mathbf{d})$ , where the domain of the function  $H$  is a subset of  $\mathbb{R}^n \times \mathbb{R}^n$ .

Consider for instance the correlation statistic  $t(\mathbf{y}, \mathbf{w}) = \mathbf{y} \cdot \mathbf{w}$ ; then

$$\alpha = H(\mathbf{v}, \mathbf{d}) = \frac{\tau - \mathbf{v} \cdot \mathbf{w}}{\mathbf{d} \cdot \mathbf{w}}$$

is defined for all  $\mathbf{v}, \mathbf{d}$ , except on a set of measure zero.

To evaluate  $\alpha$ , the attacker may use any convenient search algorithm, for example binary search. Due to the finite number of steps of the search algorithm, the value of  $\alpha$  obtained is not exactly  $H(\mathbf{v}, \mathbf{d})$ .

<sup>2</sup>In fact the watermark estimation methods studied in this work do not require knowledge of the embedding rule. Instead of (1), one could use an adaptive spread spectrum rule, in which the strength  $\sigma$  varies locally depending on local signal characteristics; or one could apply suitable preprocessing to the host  $\mathbf{s}$  in order to reduce host-signal interference during detection [4], [12]. The watermark removal step, however, depends on the embedding rule.

More accurately, if  $\alpha$  lies in an interval  $I$  of width  $W$ , the minimum number of iterations needed in order to estimate  $\alpha$  with precision  $\kappa > 0$  is

$$Q = \left\lceil \log_2 \frac{W}{\kappa} \right\rceil. \quad (3)$$

The attacker's goal is to produce an estimate  $\hat{\mathbf{w}}$  of  $\mathbf{w}$  and create the pirated copy

$$\hat{\mathbf{s}} = \mathbf{x} - \sigma \hat{\mathbf{w}}. \quad (4)$$

The mean-squared distortion of the pirated copy  $\hat{\mathbf{s}}$  relative to the host signal  $\mathbf{s}$  is  $D_s = \frac{1}{n} \|\mathbf{s} - \hat{\mathbf{s}}\|^2$ . However, recall that the attacker does not know  $\mathbf{s}$ , and only has access to  $\mathbf{x}$ . Since  $\mathbf{x}$  should be perceptually similar to  $\mathbf{s}$ , the attacker may use  $D_a = \frac{1}{n} \|\mathbf{x} - \hat{\mathbf{s}}\|^2$  as an indicator of the perceptual quality of  $\hat{\mathbf{s}}$ .

#### IV. GENERALIZED CORRELATOR DETECTOR

The new approach exploits directly the underlying structure of the detection boundary to estimate the watermark. In Sections IV-A and IV-B below, the simple correlation detection method is used:

$$t(\mathbf{y}, \mathbf{w}) = \mathbf{y} \cdot \mathbf{w}. \quad (5)$$

Then, the detection boundary is an  $n$ -dimensional plane orthogonal to the watermark vector  $\mathbf{w}$ . In particular, Patchwork [13] is an additive spread spectrum embedding scheme with correlation detection method and the algorithms in Sections IV-A and IV-B can be used to defeat it. In Sections IV-C and IV-D, extensions of the basic detection method in (5) are investigated, including normalized correlators and nonlinear pre-whitening correlators.

While deriving the new attack algorithm, several cases should be considered according to the conditions imposed on the detector input. This yields slightly different algorithms.

##### A. Unconstrained Detector Input

In the simplest setup, there is no constraint on the input to the detector. In this case, the attacker selects a set of  $n$  orthonormal vectors  $\mathbf{e}^1, \mathbf{e}^2 \dots \mathbf{e}^n \in \mathbb{R}^n$ . Let  $w_i = \mathbf{e}^i \cdot \mathbf{w}$  be the watermark component along the  $i^{\text{th}}$  unit vector  $\mathbf{e}_i$ . From (5) we have

$$t(\mathbf{e}^i, \mathbf{w}) = w_i. \quad (6)$$

Hence, the attacker just needs to estimate  $t(\mathbf{e}^i, \mathbf{w})$ , the correlation statistic for each  $\mathbf{e}^i$ . For this purpose, it suffices to identify the vector  $\bar{\mathbf{e}}^i = \alpha_i \mathbf{e}^i$  at the intersection of the radial line in direction  $\mathbf{e}^i$  and the

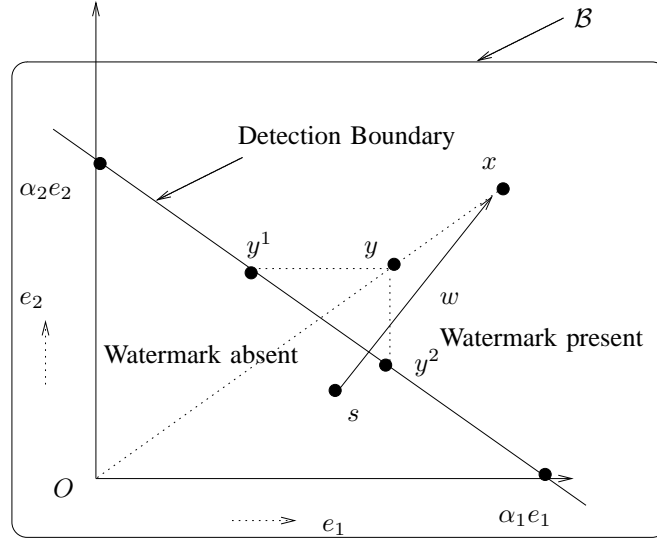


Fig. 3. An illustration of the attack algorithm on correlation detection schemes in the 2-D case.

decision boundary (refer to Figure 3). As described in Section III,  $\alpha_i$  is obtained as  $\alpha_i = H(\mathbf{o}, \mathbf{e}^i)$ . By the linearity property of the dot product, we have  $t(\mathbf{e}^i, \mathbf{w}) = \tau/\alpha_i$ , and therefore from (6) we obtain

$$w_i = \frac{\tau}{\alpha_i}, \quad 1 \leq i \leq n. \quad (7)$$

By executing sufficiently many binary search steps, the attacker obtains an estimate of the watermark vector  $\mathbf{w} = \sum_{i=1}^n w_i \mathbf{e}^i$  with any desired precision. Note that the watermarked signal  $\mathbf{x}$  is not needed at all in this algorithm.

### B. Constrained Detector Input

Often, the input to the detector must belong to a bounded region  $\mathcal{B}$  which is a subset of the Euclidean space  $\mathbb{R}^n$ , as in the case for digitized images or audio. The watermarked signal  $\mathbf{x}$  itself lives in this region. This may preclude using arbitrary orthonormal vectors  $\{\mathbf{e}^i\}$  as test signals as was done in Section IV-A. Therefore, we need a modified strategy for selecting the test signals.

For simplicity, we assume that  $\mathcal{B}$  is a star-shaped region<sup>3</sup>:

$$\mathbf{x} \in \mathcal{B} \Rightarrow \alpha \mathbf{x} \in \mathcal{B}, \quad \forall \alpha \in [0, 1]. \quad (8)$$

<sup>3</sup>This assumption does not hold for instance, when the detector's input is subject to quantization constraints (e.g., a detector that takes only JPEG images). Please see Section VII.

The attacker selects a set of  $n$  orthonormal vectors  $\{\mathbf{e}^i\}$ . He now uses the watermarked signal  $\mathbf{x}$  to create an auxiliary signal  $\mathbf{y} \in \mathcal{B}$  that is not on the decision boundary. Actually,  $\mathbf{y}$  is constructed as a scaled version  $\mathbf{y} = \alpha\mathbf{x}$  of  $\mathbf{x}$ , where  $0 < \alpha \leq 1$ . By our assumption (8), we have  $\mathbf{y} \in \mathcal{B}$ . Then, the scale factor  $\alpha_i$  is selected such that  $\mathbf{y}^i$ , defined below, is on the detection boundary (refer to Figure 3):

$$\mathbf{y}^i \triangleq \mathbf{y} + \alpha_i \mathbf{e}^i, \quad (9a)$$

$$\alpha_i = H(\mathbf{y}, \mathbf{e}^i), \quad (9b)$$

$$\Rightarrow t(\mathbf{y}^i, \mathbf{w}) = \tau, \quad (9c)$$

where  $H(\cdot, \cdot)$  is defined in Section III.

This is done for every  $i \in \{1, 2, \dots, n\}$ . If  $\mathbf{y}$  is selected inside  $\mathcal{B}$  but far enough from the boundary of  $\mathcal{B}$ , it is guaranteed that the signals  $\mathbf{y}^i$  will belong to  $\mathcal{B}$  (see Figure 3).

Using the linearity property of the dot product again, (6), (9a) and (9c) imply

$$\tau = t(\mathbf{y}, \mathbf{w}) + \alpha_i w_i, \quad 1 \leq i \leq n. \quad (10)$$

Moreover, from (5) we have

$$t(\mathbf{y}, \mathbf{w}) = \mathbf{y} \cdot \mathbf{w} = \sum_{i=1}^n y_i w_i. \quad (11)$$

Substituting (11) into (10), we obtain

$$\sum_{j=1, j \neq i}^n y_j w_j + (y_i + \alpha_i) w_i = \tau, \quad 1 \leq i \leq n.$$

This is a linear system of  $n$  equations with  $n$  unknowns. Normally, solving such a system would require  $O(n^3)$  operations. However, the special structure of this system reduces the number of operations to  $n + 1$ , as shown below.

From (10), we have

$$w_i = \frac{\tau - t(\mathbf{y}, \mathbf{w})}{\alpha_i}, \quad 1 \leq i \leq n. \quad (12)$$

Multiplying both sides of this equation by  $y_i$ , summing from 1 to  $n$ , and substituting the sum into the right side of (11), we obtain

$$t(\mathbf{y}, \mathbf{w}) = (\tau - t(\mathbf{y}, \mathbf{w})) \sum_{i=1}^n \frac{y_i}{\alpha_i}, \quad (13)$$

which yields the value of the correlation statistic,

$$t(\mathbf{y}, \mathbf{w}) = \frac{\tau \sum_{i=1}^n y_i / \alpha_i}{1 + \sum_{i=1}^n y_i / \alpha_i}. \quad (14)$$

TABLE II  
CORRELATION DETECTION ALGORITHM.

1	Use $\mathbf{x}$ to construct $\mathbf{y}$ near the decision boundary but not on it.
2	Construct $n$ signals $\mathbf{y}^i = \mathbf{y} + \alpha_i \mathbf{e}^i$ on the decision boundary.
3	Compute $t(\mathbf{y}, \mathbf{w})$ from (14).
4	Estimate the watermark by replacing $t(\mathbf{y}, \mathbf{w})$ , $\alpha_i$ and $y_i$ in (12).

Hence, the attacker first uses (14) to compute  $t(\mathbf{y}, \mathbf{w})$ , then (12) to compute  $w_i$  for  $1 \leq i \leq n$ , and finally obtains  $\mathbf{w} = \sum_{i=1}^n w_i \mathbf{e}^i$ . The algorithm is summarized in Table II.

After  $n+1$  steps, an estimate  $\hat{\mathbf{w}}$  of the watermark  $\mathbf{w}$  is obtained and is used to construct the pirated copy  $\hat{\mathbf{s}}$  as indicated in Section III. Note that the attacker's unlimited access to the detector is what enables him to estimate the scale factors  $\alpha_i$ ,  $i \in \{1, \dots, n\}$ . As explained in Section III, the binary search algorithm can be used for this purpose. If  $\alpha_i$  lies in an interval  $I$  of width  $W$ , the minimum number of iterations needed in order to estimate  $\alpha_i$  with precision  $\kappa > 0$  is  $Q$  in (3). Hence, the algorithm requires  $Qn$  detection operations in order to estimate the watermark. However, the detection operation itself has linear complexity in  $n$ , the length of the signal  $\mathbf{x}$ . Therefore, the algorithm has  $O(Qn^2)$  complexity. Moreover, the algorithm is noniterative in the sense that in order to estimate  $\{\alpha_i\}$  with precision  $\kappa > 0$  and hence  $\mathbf{w}$ ,  $Qn$  operations are required exactly.

### C. Function of the Correlation Statistic

Let us consider the following detection statistic:

$$t(\mathbf{y}, \mathbf{w}) = F(\mathbf{y} \cdot \mathbf{w}, \mathbf{y}), \quad (15)$$

where  $F(\cdot, \cdot)$  is a general function mapping  $\mathbb{R} \times \mathbb{R}^n$  to  $\mathbb{R}$ . In other words,  $\mathbf{w}$  affects the detector output only via the scalar quantity  $\mathbf{y} \cdot \mathbf{w}$ . Note that since  $\mathbf{y}$  is known to the attacker, he can view  $t(\mathbf{y}, \mathbf{w})$  as a function of the scalar unknown  $\mathbf{y} \cdot \mathbf{w}$ . We assume that  $F(\cdot, \mathbf{y}^i)$  is invertible for the test signals  $\{\mathbf{y}^i\}$  defined in (9a), and denote by  $F^{-1}(\cdot, \mathbf{y}^i)$  the inverse function.

Of course the simple correlation statistic used earlier in this section is a particular case of (15), with  $F(\mathbf{y} \cdot \mathbf{w}, \mathbf{y}) = \mathbf{y} \cdot \mathbf{w}$ . Another particular case is the normalized correlation statistic [4], which is used in Kalker's algorithm [3]:

$$t(\mathbf{y}, \mathbf{w}) = \frac{\mathbf{y} \cdot \mathbf{w}}{\|\mathbf{y}\| \|\mathbf{w}\|}, \quad (16)$$

where the function  $F(\mathbf{y} \cdot \mathbf{w}, \mathbf{y}) = \frac{\mathbf{y} \cdot \mathbf{w}}{\|\mathbf{y}\| \|\mathbf{w}\|}$  is invertible for all  $\mathbf{y}$ , with  $F^{-1}(f, \mathbf{y}) = f \|\mathbf{y}\| \|\mathbf{w}\|$ . This is true if  $\|\mathbf{w}\|$  is known to the attacker. If this is not the case, then redefine the detection function as

$$t(\mathbf{y}, \mathbf{w}) = \frac{\mathbf{y} \cdot \mathbf{w}}{\|\mathbf{y}\|}, \quad (17)$$

and let the threshold be  $\tau' = \tau \|\mathbf{w}\|$ . Again, this function belongs to the family of detection functions considered in this section. However, the threshold  $\tau'$  is unknown to the attacker, a scenario studied in Section VI and proven not to affect the complexity of the attack.

We now ask: under which conditions can the watermark  $\mathbf{w}$  be restored in  $O(n)$  detection operations. The attacker creates a signal  $\mathbf{y}$  anywhere in  $\mathbb{R}^n$  except  $\mathbf{o}$  and the decision boundary. Then, similarly to the algorithm proposed in Section IV-B, he constructs signals  $\mathbf{y}^i$ ,  $1 \leq i \leq n$ , on the decision boundary, i.e.,  $F(\mathbf{y}^i \cdot \mathbf{w}, \mathbf{y}^i) = \tau$ . Using (6), (9b), and (15), and the linearity of the dot product, we obtain

$$F(\mathbf{y} \cdot \mathbf{w} + \alpha_i w_i, \mathbf{y}^i) = \tau, \quad 1 \leq i \leq n. \quad (18)$$

The  $n$  equations given in (18) form a nonlinear system in  $n$  unknowns. The system can however be transformed into a linear system under the invertibility assumption on  $F$  above. From (18) we obtain

$$\mathbf{y} \cdot \mathbf{w} + \alpha_i w_i = F^{-1}(\tau, \mathbf{y}^i), \quad 1 \leq i \leq n.$$

This system can be solved similarly to that in Section IV-B, and therefore an estimate  $\hat{\mathbf{w}}$  of the watermark is obtained in  $n + 1$  steps, and  $O(n)$  detection operations.

#### D. Nonlinear “Pre-Whitened” Correlator

In this section, we study a class of detectors that attempt to remove host signal interference prior to correlation with the watermark [14]. First the detector estimates the host signal  $\mathbf{s}$  by  $\tilde{\mathbf{s}}(\mathbf{y})$ , then it subtracts the estimate from  $\mathbf{y}$  before correlating with  $\mathbf{w}$ . The detection function is

$$t(\mathbf{y}, \mathbf{w}) = (\mathbf{y} - \tilde{\mathbf{s}}(\mathbf{y})) \cdot \mathbf{w}. \quad (19)$$

If the estimator is linear in  $\mathbf{y}$ , the detection function in (19) reduces to

$$\begin{aligned} t(\mathbf{y}, \mathbf{w}) &= (G\mathbf{y}) \cdot \mathbf{w} \\ &= \mathbf{y} \cdot (G^t \mathbf{w}), \end{aligned}$$

where the superscript  $t$  denotes matrix transpose.

By our assumptions in Section III, the attacker knows the matrix  $G$ . He may use the algorithm described in Section IV-B to estimate  $G^t \mathbf{w}$  as  $\hat{\mathbf{w}}_g$ . If  $G$  is invertible, the estimate of  $\mathbf{w}$  is obtained as

$$\hat{\mathbf{w}} = (G^{-1})^t \hat{\mathbf{w}}_g.$$

More generally, if the estimator,  $\tilde{s}(\mathbf{y})$ , is nonlinear, the detection function in (19) takes the form

$$t(\mathbf{y}, \mathbf{w}) = f(\mathbf{y}) \cdot \mathbf{w}, \quad (20)$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a nonlinear transformation, known to the attacker. Consider the case when  $f(\cdot)$  is invertible. In order to obtain an estimate for the watermark  $\mathbf{w}$ ,  $n$  signals  $\{\mathbf{y}^i\}$  are generated on the detection boundary:

$$f(\mathbf{y}^i) \cdot \mathbf{w} = \tau, \quad 1 \leq i \leq n. \quad (21)$$

The attacker generates, for each  $i \in \{1, \dots, n\}$ , a new signal  $\mathbf{z}^i = \mathbf{y} + \alpha_i \mathbf{e}^i$  such that the inverse signal  $\mathbf{y}^i = f^{-1}(\mathbf{z}^i)$  is on the detection boundary. A slight variation of the mapping  $H(\cdot, \cdot)$  defined in Section III is used to evaluate the scalar  $\alpha_i$ .

$$\begin{aligned} \mathbf{z}^i &\triangleq \mathbf{y} + \alpha_i \mathbf{e}^i, \\ \mathbf{y}^i &= f^{-1}(\mathbf{z}^i), \\ \mathbf{z}^i \cdot \mathbf{w} &= \tau, \quad 1 \leq i \leq n. \end{aligned} \quad (22)$$

The system (22) can be solved using the algorithm of Section IV-B.

## V. REGULAR DETECTORS

In this section, the vulnerability of general decision rules to sensitivity analysis attacks is investigated. For this purpose, detection statistics  $t(\mathbf{y}, \mathbf{w})$  other than the simple correlation statistic  $\mathbf{y} \cdot \mathbf{w}$  and its extensions  $F(\mathbf{y} \cdot \mathbf{w}, \mathbf{y})$  and  $f(\mathbf{y}) \cdot \mathbf{w}$  considered in Section IV are addressed. In particular, we assume that the detection boundary satisfies second-order regularity conditions and can be locally approximated by a hyperplane. Under these regularity conditions, we are still able to produce an accurate estimate of the watermark in quadratic time.

### A. Assumptions on Detector

Let us consider the general decision statistic  $t(\mathbf{y}, \mathbf{w})$ , and define the gradient mapping  $\mathbf{g} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  as follows:

$$\mathbf{g}(\mathbf{y}, \hat{\mathbf{w}}) \triangleq \nabla_{\mathbf{y}} t(\mathbf{y}, \hat{\mathbf{w}}). \quad (23)$$

Our first assumption is (8): the feasible region  $\mathcal{B}$  for the detector input is star-shaped. Assume that the watermark  $\mathbf{w}$ , the watermarked signal  $\mathbf{x}$ , and the scaled signal  $\mathbf{y} = \alpha \mathbf{x}$  defined in (25) below satisfy the following properties:

(A1)  $t(\mathbf{o}, \mathbf{w}) < \tau < t(\mathbf{x}, \mathbf{w})$ , i.e., the origin  $\mathbf{o}$  belongs to the *rejection region* and the watermarked signal  $\mathbf{x}$  to the *acceptance region*.

(A2) There exists  $\eta > 0$  such that  $t(\cdot, \mathbf{w})$  is twice continuously differentiable in the  $n$ -dimensional  $L^2$ -ball of radius  $\sqrt{n\eta}$  centered at  $\mathbf{y}$ :

$$\mathcal{B}_{\mathbf{y}}(\eta) = \{\hat{\mathbf{y}} : \|\mathbf{y} - \hat{\mathbf{y}}\| \leq \sqrt{n\eta}\}.$$

Moreover, the absolute eigenvalues of the Hessian  $\nabla_{\hat{\mathbf{y}}}^2 t(\hat{\mathbf{y}}, \mathbf{w})$  are upper-bounded by

$$\bar{\lambda} < \infty \quad (24)$$

for all  $\hat{\mathbf{y}} \in \mathcal{B}_{\mathbf{y}}(\eta)$ . (Note that  $\bar{\lambda}$  generally depends on  $\mathbf{w}$ ,  $\mathbf{y}$ , and  $\eta$ .)

(A3) There exists  $\epsilon > 0$  such that the gradient mapping  $\mathbf{g}(\mathbf{y}, \cdot)$  of (23) is invertible over the  $L^2$ -ball of radius  $\sqrt{n\epsilon}$  centered at  $\mathbf{w}$ :

$$\mathcal{B}_{\mathbf{w}}(\epsilon) = \{\hat{\mathbf{w}} : \|\mathbf{w} - \hat{\mathbf{w}}\| \leq \sqrt{n\epsilon}\}.$$

### B. Algorithm

By our assumptions in Section III, the attacker knows the function  $t(\cdot, \cdot)$  and the threshold  $\tau$ . Given the watermarked signal  $\mathbf{x}$ , he may then implement the following steps:

- 1) Evaluate the scale factor  $\alpha = H(\mathbf{o}, \mathbf{x})$  such that the signal

$$\mathbf{y} = \alpha \mathbf{x} \quad (25)$$

lies on the decision boundary:

$$t(\mathbf{y}, \mathbf{w}) = \tau. \quad (26)$$

This is possible because of (A1) and our assumption (8) that the region  $\mathcal{B}$  is star-shaped.

- 2) Select an orthonormal set of vectors  $\{\mathbf{e}^i\}_{i=1,\dots,n}$  and  $n$  small positive numbers  $\epsilon_1, \dots, \epsilon_n$ . For each  $1 \leq i \leq n$ , if needed, flip the sign of  $\epsilon_i$  such that the signal

$$\bar{\mathbf{y}}^i = \mathbf{y} + \epsilon_i \mathbf{e}^i \quad (27)$$

lies in the *acceptance region* (see Figure 4). This signal is scaled to produce a signal  $\mathbf{y}^i$  on the detection boundary:

$$\mathbf{y}^i = \alpha_i \bar{\mathbf{y}}^i, \quad (28)$$

$$t(\mathbf{y}^i, \mathbf{w}) = \tau, \quad (29)$$

$$\alpha_i = H(\mathbf{o}, \bar{\mathbf{y}}^i), \quad 1 \leq i \leq n.$$



TABLE III  
ATTACK ON REGULAR DETECTORS.

1	Scale $\mathbf{x}$ and obtain $\mathbf{y}$ on the decision boundary.
2	Construct $n$ signals $\bar{\mathbf{y}}^i = \mathbf{y} + \epsilon_i \mathbf{e}^i$ in the <i>acceptance</i> region near $\mathbf{y}$ .
3	Scale these signals and obtain $n$ signals $\mathbf{y}^i = \alpha_i \bar{\mathbf{y}}^i$ on the decision boundary.
4	Solve (38) for $\beta$ .
5	Estimate $\mathbf{g}(\mathbf{y}, \mathbf{w})$ by replacing $\alpha_i$ , $\epsilon_i$ and $\beta$ in (36).
6	Estimate the watermark by substituting $\mathbf{g}(\mathbf{y}, \mathbf{w})$ into (37).

Substituting the expressions (27) and (28) for  $\bar{\mathbf{y}}^i$  and  $\mathbf{y}^i$  into (31), we obtain

$$\mathbf{d}^i = \alpha_i \bar{\mathbf{y}}^i - \mathbf{y} = (\alpha_i - 1)\mathbf{y} + \alpha_i \epsilon_i \mathbf{e}^i. \quad (35)$$

Taking the dot product of  $\mathbf{d}^i$  in (35) with  $\mathbf{g}(\mathbf{y}, \mathbf{w})$  and using (32), (33), and (34), we obtain

$$g_i(\mathbf{y}, \mathbf{w}) \simeq \frac{1 - \alpha_i}{\alpha_i \epsilon_i} \beta, \quad 1 \leq i \leq n. \quad (36)$$

By assumption (A3), the watermark  $\mathbf{w}$  is recoverable from the gradient vector  $\mathbf{g}(\mathbf{y}, \mathbf{w})$ . Denoting the inverse function by  $\mathbf{g}^{-1}(\mathbf{y}, \cdot)$ , and using  $\mathbf{g}(\mathbf{y}, \mathbf{w}) = \sum_{i=1}^n g_i(\mathbf{y}, \mathbf{w}) \mathbf{e}^i$ , we obtain

$$\mathbf{w} = \mathbf{g}^{-1} \left( \mathbf{y}, \sum_{i=1}^n g_i(\mathbf{y}, \mathbf{w}) \mathbf{e}^i \right). \quad (37)$$

At this point, the attacker has selected  $\{\epsilon_i\}$  and evaluated  $\{\alpha_i\}$ . Using (36), he can now estimate the  $n$  components of  $\mathbf{g}(\mathbf{y}, \mathbf{w})$  up to a scaling factor  $\beta$ . Therefore, the  $n \times n$  system (32) can be solved for  $\mathbf{g}(\mathbf{y}, \mathbf{w})$  up to the factor  $\beta$  in  $n$  steps instead of  $O(n^3)$  steps. To compute  $\beta$ , we substitute (36) into (37) and then (37) into (26), and obtain a nonlinear equation with a single unknown  $\beta$ :

$$t \left( \mathbf{y}, \mathbf{g}^{-1} \left( \mathbf{y}, \beta \sum_{i=1}^n \frac{1 - \alpha_i}{\alpha_i \epsilon_i} \mathbf{e}^i \right) \right) \simeq \tau. \quad (38)$$

Since the attacker knows the mapping  $\mathbf{g}^{-1}(\cdot, \cdot)$ , he can numerically solve (38) for  $\beta$ . Then he can obtain  $\{g_i(\mathbf{y}, \mathbf{w})\}$  from (36) and  $\mathbf{w}$  from (37).

It should be noted that (38) may be hard to solve and may have more than one solution, depending on the nature of the detection statistic  $t(\cdot, \cdot)$ . Moreover, unless the decision boundary is a hyperplane in the neighborhood of  $\mathbf{y}$ , the local linearization (32) is only an approximation. Yet as illustrated in Section VIII-C, by selecting appropriate scalars  $\epsilon_i$ , the watermark can be *almost exactly* estimated. Table III summarizes the steps of the algorithm.

Finally, we comment on the complexity of this algorithm. As mentioned above, the scalars used by the algorithm are  $\alpha$ ,  $\alpha_i$ , and  $\epsilon_i$  for  $i \in \{1, \dots, n\}$ . First, the attacker estimates the scalar  $\alpha \in (0, 1)$  by the binary search algorithm with precision  $0 < \kappa < 1$ . Then, he sets the magnitude of the scalars  $\epsilon_i$  to a small value. Their signs are selected such that  $\bar{\mathbf{y}}^i = \mathbf{y} + \epsilon_i \mathbf{e}^i$  belongs to the *acceptance* region. Next, the scalars  $\alpha_i \in (0, 1)$  for  $i \in \{1, \dots, n\}$  are estimated in a similar way to  $\alpha$ . The required number of steps for each estimation is

$$Q = \left\lceil \log_2 \frac{1}{\kappa} \right\rceil.$$

Therefore, the algorithm requires  $(Q + 1)n + Q$  detection operations. Since each such operation is linear in  $n$ , the dimension of the signal, the algorithm has  $O((Q + 1)n^2)$  complexity. Moreover, the algorithm is noniterative, since a good estimate of the watermark  $\mathbf{w}$  is obtained exactly after  $(Q + 1)n + Q$  detection operations.

### C. Application to Generalized Gaussian Hosts

Let us apply the algorithm of Section V-B to ML detectors, assuming that the host signal  $\mathbf{s}$  is distributed according to the generalized Gaussian distribution (GGD):

$$f_{\mathbf{s}}(\mathbf{s}) = A \cdot \exp \left( - \sum_{i=1}^n |cs_i|^\mu \right),$$

where  $c$  is a scale parameter, and  $A$  is a normalizing constant. Given an input signal  $\mathbf{z}$ , the log likelihood ratio statistic, scaled by  $c^{-\mu}$ , is equal to

$$t(\mathbf{z}, \mathbf{w}) \triangleq c^{-\mu} \ln \frac{f_{\mathbf{s}}(\mathbf{z} - \mathbf{w})}{f_{\mathbf{s}}(\mathbf{z})} = \sum_{i=1}^n |z_i|^\mu - |z_i - w_i|^\mu. \quad (39)$$

In (39),  $z_i$ ,  $1 \leq i \leq n$ , are the components of  $\mathbf{z}$ . This detector was first used for watermark detection by Hernández et al [15].

We assume that  $B = \mathbb{R}^n$ , so assumption (8) holds. A necessary condition for assumption (A1) in Section V-A to hold is that the threshold  $\tau$  exceeds  $t(\mathbf{o}, \mathbf{w}) = - \sum_{i=1}^n |w_i|^\mu$ .

If the function in (39) is differentiable for the signals  $\mathbf{z}$  and  $\mathbf{w}$ , the gradient  $\mathbf{g}(\mathbf{z}, \mathbf{w})$  exists:

$$\begin{aligned} g_i(\mathbf{z}, \mathbf{w}) &= \frac{\partial}{\partial z_i} (|z_i|^\mu - |z_i - w_i|^\mu), \\ &= \mu (\operatorname{sgn}(z_i) |z_i|^{\mu-1} - \operatorname{sgn}(z_i - w_i) |z_i - w_i|^{\mu-1}), \quad 1 \leq i \leq n. \end{aligned} \quad (40)$$

For  $\mu > 1$ , the gradient of (40) exists for all  $\mathbf{z}$  and  $\mathbf{w}$ . However, for  $\mu \leq 1$ ,  $\mathbf{g}(\mathbf{z}, \mathbf{w})$  exists for the signals  $\mathbf{z}$  and  $\mathbf{w}$  if and only if  $z_i \neq 0$  and  $z_i \neq w_i$  for all  $i \in \{1, \dots, n\}$ . This condition holds almost everywhere (a.e.) on  $B \times \mathbb{R}^n$ . Therefore, the gradient  $\mathbf{g}(\mathbf{z}, \mathbf{w})$  exists a.e.

Let us denote  $g_i(\mathbf{z}, \mathbf{w})$  by  $\gamma(z_i, w_i)$  since it is a function of  $z_i$  and  $w_i$  only. For any given  $z_i$ ,  $\gamma(z_i, \cdot)$  is invertible, as shown below. The inverse function is denoted by  $\gamma^{-1}(z_i, \cdot)$ .

When  $\mu = 2$ , the GGD detector is a detector equivalent to the simple correlation detector in (5). The test statistic of (5) is multiplied by 2 and the energy term  $\|\mathbf{w}\|^2$  is subtracted:

$$t(\mathbf{z}, \mathbf{w}) = 2\mathbf{z} \cdot \mathbf{w} - \|\mathbf{w}\|^2. \quad (41)$$

Equation (40) yields  $\gamma(z_i, w_i) = 2w_i$ , and therefore,  $\mathbf{g}(\mathbf{z}, \mathbf{w}) = 2\mathbf{w}$ .

For the more general case when  $\mu$  is not necessarily equal to 2, (40) implies

$$\text{sgn}(z_i - w_i) = -\text{sgn}\left(\frac{\gamma(z_i, w_i)}{\mu} - \text{sgn}(z_i)|z_i|^{\mu-1}\right),$$

and

$$|z_i - w_i| = \left| \frac{\gamma(z_i, w_i)}{\mu} - \text{sgn}(z_i)|z_i|^{\mu-1} \right|^{\frac{1}{\mu-1}}.$$

Therefore, for each  $z_i$  and  $g_i = \gamma(z_i, w_i)$ , we have

$$\begin{aligned} w_i &= \gamma^{-1}(z_i, g_i) \\ &= z_i + \text{sgn}\left(\frac{g_i}{\mu} - \text{sgn}(z_i)|z_i|^{\mu-1}\right) \left| \frac{g_i}{\mu} - \text{sgn}(z_i)|z_i|^{\mu-1} \right|^{\frac{1}{\mu-1}}. \end{aligned} \quad (42)$$

Hence, the watermark  $\mathbf{w}$  is recoverable, given the gradient  $\mathbf{g}(\mathbf{z}, \mathbf{w})$  and the signal  $\mathbf{z}$  and the GGD detector satisfies assumption (A3). Let us now check assumption (A2). The Hessian matrix for the detection statistic  $t(\mathbf{z}, \mathbf{w})$  in (39) is diagonal. Thus its eigenvalues coincide with the diagonal entries

$$\begin{aligned} \lambda_i &= \frac{\partial^2}{\partial z_i^2} t(\mathbf{z}, \mathbf{w}) \\ &= \frac{\partial^2}{\partial z_i^2} (|z_i|^\mu - |z_i - w_i|^\mu) = \mu(\mu - 1) (|z_i|^{\mu-2} - |z_i - w_i|^{\mu-2}), \quad 1 \leq i \leq n. \end{aligned}$$

We have  $|\lambda_i| < \infty$  for  $\mu > 2$  and  $\forall \mathbf{z} \in B$  and  $\forall \mathbf{w} \in \mathbb{R}^n$ . We have  $|\lambda_i| < \infty$  for  $\mu < 2$ , only when  $z_i \neq 0$  and  $z_i \neq w_i$  for all  $i \in \{1, \dots, n\}$ , a condition that is satisfied a.e. on  $B \times \mathbb{R}^n$ . Hence,  $t(\mathbf{z}, \mathbf{w})$  is twice differentiable a.e. Hence, recalling (24), if  $\mathbf{y}$  and  $\mathbf{w}$  are selected from some probability distribution that is continuous with respect to the Lebesgue measure,  $Pr[\bar{\lambda} = \infty]$  vanishes as  $\eta \rightarrow 0$ .

The algorithm constructs the signals  $\mathbf{y}$  and  $\mathbf{y}^i$ ,  $1 \leq i \leq n$ , from  $\mathbf{x}$  as described in the previous section. For  $\mu > 2$ ,  $\bar{\lambda} < \infty$  since all the signals of interest belong to a bounded region of the space,  $B_{\mathbf{y}}(\eta)$ . For  $1 < \mu < 2$ , we have

$$\bar{\lambda} = \mu(\mu - 1) \max(z_{min}^{\mu-2}, d_{min}^{\mu-2})$$

where

$$z_{min} \triangleq \min_{\mathbf{z} \in B_{\mathbf{y}}(\eta)} \min_{1 \leq i \leq n} |z_i|,$$

and

$$d_{min} \triangleq \min_{\mathbf{z} \in B_{\mathbf{y}}(\eta)} \min_{1 \leq i \leq n} |z_i - w_i|,$$

where (for small  $\eta$ ), the probability that  $z_{min} = 0$  or  $d_{min} = 0$  is very small. Note that  $\mathbf{y}$  is an auxiliary signal on the decision boundary. If for this signal,  $z_{min}$  and/or  $d_{min}$  are too low, i.e.,  $\bar{\lambda}$  is too high, the attack may fail, i.e., not result in a pirated copy  $\hat{\mathbf{s}}$  with low distortion. In this case, the attacker can just generate another signal<sup>5</sup>  $\mathbf{y}$  on the decision boundary. Hence, (A2) is satisfied with high probability.

Therefore in the cases where all the assumptions are satisfied, the algorithm is used to generate the signal  $\mathbf{y}$  and the scalars  $\alpha_i$  and  $\epsilon_i$ ,  $i \in \{1, \dots, n\}$ , as described in Section V-A. Equation (42) gives the expression of the  $i^{th}$  watermark component  $w_i$  in terms of the  $i^{th}$  gradient component  $g_i(\mathbf{y}, \mathbf{w})$ . From (36), we obtain an approximation for  $\frac{1}{\beta} g_i(\mathbf{y}, \mathbf{w})$ . Substituting (36) into (42) and using the fact the signal  $\mathbf{y}$  lies on the boundary, we obtain

$$\begin{aligned} t(\mathbf{y}, \mathbf{w}) &= \tau \\ \sum_{i=1}^n |y_i|^\mu - |y_i - w_i|^\mu &= \tau \\ f(\beta) &\simeq \tau, \end{aligned} \tag{43}$$

where we have defined

$$f(\beta) = \sum_{i=1}^n \left\{ |y_i|^\mu - \left| \frac{(1 - \alpha_i)\beta}{\alpha_i \epsilon_i} \frac{\beta}{\mu} - \text{sgn}(y_i) |y_i|^{\mu-1} \right|^{\frac{\mu}{\mu-1}} \right\}, \quad \beta \in \mathbb{R}. \tag{44}$$

We are interested in studying the existence and the number of roots of the equation  $f(\beta) = \tau$ .

Let us assume temporarily that  $\tau = 0$ , corresponding to a Bayes test with equal priors on  $H_0$  and  $H_1$  and zero/one cost assignment. The function  $f(\beta)$  satisfies the following properties, which are illustrated in Figure 5:

- 1)  $f$  is continuous.
- 2)  $f$  is concave since it is the sum of  $n$  concave functions of the form  $g_i(\beta) = c_i - |a_i \beta + b_i|^p$ , where  $a_i$ ,  $b_i$ , and  $c_i$  are real numbers and  $p > 1$ .
- 3) Equation (44) evaluated at  $\beta = 0$  results in  $f(0) = 0$ . Hence,  $\beta = 0$  is a solution of (43).
- 4)  $f(\beta) \rightarrow -\infty$  as  $|\beta| \rightarrow \infty$ .

<sup>5</sup>For instance, a random vector,  $\mathbf{v}$ , can be added to  $\mathbf{x}$ , and the algorithm of Table III is applied to  $\mathbf{x} + \mathbf{v}$ .

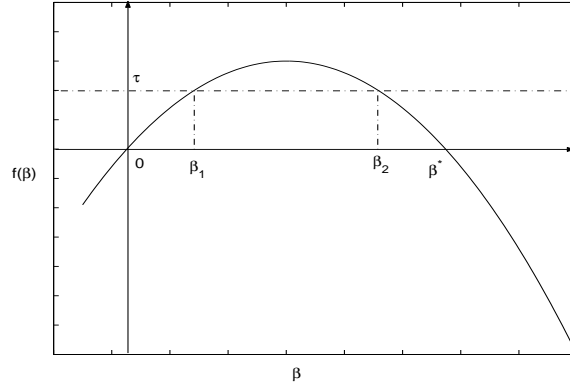


Fig. 5. An example of a function with similar properties as  $f(\beta)$ .

Therefore, the function  $f(\beta)$  must cross the  $\beta$ -axis at a *unique* location  $\beta = \beta^* > 0$ . Note that the first solution,  $\beta = 0$ , is degenerate, since it results in  $\hat{\mathbf{g}}(\mathbf{y}, \mathbf{w}) = 0$  in (36) and (40), we obtain  $\hat{\mathbf{w}} = 0$ . The next step of the attacker is to use this  $\beta^*$  in (36) in order to estimate the  $n$  components  $g_i = g_i(\mathbf{z}, \mathbf{w})$ . Next, (42) is used to estimate the components  $w_i$  of  $\mathbf{w}$  for  $1 \leq i \leq n$ . In summary, the watermark is recovered in  $2n + 1$  steps. Using this estimate, the attacker computes the pirated copy  $\hat{\mathbf{s}}$  using (4).

In the general case when the threshold is  $\tau > 0$ , the roots of the equation  $f(\beta) = \tau$  are  $\beta_1$  and  $\beta_2$  shown in Figure 5. Since  $\beta_1$  and  $\beta_2$  are continuous functions of  $\tau$ ,  $\beta_2$  is always the root that should be selected by the attacker. Note that if  $\tau > \max_{\beta} f(\beta)$ , the equation  $f(\beta) = \tau$  no longer has roots.

## VI. PARAMETRIC DETECTORS

As stated in Section III, the threshold, the detection function, and all its parameters are known to the attacker who uses this knowledge together with his access to the detector in order to estimate  $\mathbf{w}$ . Some schemes attempt to improve security by keeping a few parameters secret. Intuitively, we cannot expect such an approach to be successful. In this section, we extend the algorithms of Sections IV and V to defeat such schemes. The complexity of the algorithms is not significantly increased.

### A. Unknown Threshold

We begin by showing that keeping the value of the threshold secret does not make the watermarking scheme more secure.

1) *Generalized Correlator Detector*: As in Section IV, the main idea of sensitivity analysis attacks is to make use of the unlimited access to the detector in order to obtain information about the watermark

$\mathbf{w}$ . This is done by creating auxiliary signals on the detection boundary, resulting in an  $n \times n$  system of equations of the form

$$\mathbf{y}^i \cdot \mathbf{w} = \tau, \quad i \in \{1, \dots, n\}. \quad (45)$$

The unknowns are the vector  $\mathbf{w}$  and the parameter  $\tau$ .

*Claim 1:* If  $(\mathbf{w}_0, \tau_0)$  is a solution of the system (45), then so is  $(c\mathbf{w}_0, c\tau_0)$  for any  $c \in \mathbb{R}$ .

Therefore the attacker cannot recover the exact watermark and threshold. In fact, the attacker is not concerned about the threshold, he is only interested in producing a good estimate of the watermark and a good signal in the rejection region. Although the threshold is unknown, the attacker can still estimate the watermark up to a scalar.

Define the normalized watermark

$$\mathbf{w}' = \frac{1}{\tau} \mathbf{w}.$$

Then (45) may be viewed as a linear system of  $n$  equations in the  $n$  unknowns  $w'_i, 1 \leq i \leq n$ :

$$\mathbf{y}^i \cdot \mathbf{w}' = 1, \quad i \in \{1, \dots, n\}. \quad (46)$$

This is exactly the same problem as the one considered in Section IV-B, with threshold equal to 1. From (12) and (14), we obtain  $\mathbf{w}'$  as follows:

$$w'_i = \frac{w_i}{\tau} = \frac{1}{\alpha_i \left(1 + \sum_{k=1}^n \frac{y_k}{\alpha_k}\right)}, \quad \forall i \in \{1, \dots, n\}.$$

Having  $\mathbf{w}'$ , we can construct the projection  $\mathbf{x}'$  of the watermarked signal  $\mathbf{x}$  onto the boundary. Since  $\mathbf{x}' - \mathbf{x}$  is orthogonal to the boundary, we have

$$\mathbf{x}' = \mathbf{x} + c\mathbf{w}', \quad \text{for some constant } c \in \mathbb{R}.$$

Since  $\mathbf{x}'$  is on the boundary, we also have

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}' &= \tau \\ \mathbf{w}' \cdot \mathbf{x}' &= 1 \\ \mathbf{w}' \cdot \mathbf{x} + c\|\mathbf{w}'\|^2 &= 1 \\ c &= \frac{1 - \mathbf{w}' \cdot \mathbf{x}}{\|\mathbf{w}'\|^2}. \end{aligned}$$

Therefore, the projection of the watermarked signal  $\mathbf{x}$  on the boundary is given by

$$\mathbf{x}' = \mathbf{x} + \frac{1 - \mathbf{w}' \cdot \mathbf{x}}{\|\mathbf{w}'\|^2} \mathbf{w}'.$$

2) *Host Available at the Detector*: In the analysis we made so far, we assumed blind detection. It turns out that the case when the host signal is available at the detector is just the same as the case of blind detection with unknown threshold considered in this section. Assume again that the detector is the correlation detector. If the detector knows the host  $\mathbf{s}$ , the test takes the form

$$t(\mathbf{y}, \mathbf{w}) = (\mathbf{y} - \mathbf{s}) \cdot \mathbf{w} \underset{<}{\underset{>}{\geq}} \tau,$$

which is equivalent to

$$\mathbf{y} \cdot \mathbf{w} \underset{<}{\underset{>}{\geq}} \tau',$$

where  $\tau' = \tau + \mathbf{s} \cdot \mathbf{w}$ . Both the host signal  $\mathbf{s}$  and the watermark  $\mathbf{w}$  are unknown to the attacker, therefore  $\tau'$  is unknown to him also, and we are back into the problem of the previous section: estimate the watermark in case of blind correlation detection and unknown parameter  $\tau'$ .

3) *Regular Detectors*: The family of regular detectors was introduced in Section V. Here, we have to estimate two unknowns,  $\beta$  and  $\tau$ . Therefore, we need one more equation in addition to (38). For this purpose, an auxiliary signal  $\mathbf{u}$  is generated from the watermarked signal  $\mathbf{x}$  on the decision boundary:

$$t\left(\mathbf{u}, \mathbf{g}^{-1}\left(\mathbf{y}, \beta \sum_{i=1}^n \frac{1 - \alpha_i}{\alpha_i \epsilon_i} \mathbf{e}^i\right)\right) \simeq \tau. \quad (47)$$

We can solve for  $\beta$  by subtracting (47) from (38) and finding the root of the equation

$$t\left(\mathbf{y}, \mathbf{g}^{-1}\left(\mathbf{y}, \beta \sum_{i=1}^n \frac{1 - \alpha_i}{\alpha_i \epsilon_i} \mathbf{e}^i\right)\right) - t\left(\mathbf{u}, \mathbf{g}^{-1}\left(\mathbf{y}, \beta \sum_{i=1}^n \frac{1 - \alpha_i}{\alpha_i \epsilon_i} \mathbf{e}^i\right)\right) \simeq 0. \quad (48)$$

Next, we substitute the estimated  $\beta$  into (38) and obtain an estimate of  $\tau$ . Recall from Section V-C that for GGD hosts with  $\mu = 2$ , the detector is a correlator, the boundary is a hyperplane<sup>6</sup>, and the gradient in (40) is equal to  $2\mathbf{w}$ . Therefore the magnitude of  $\mathbf{w}$  is proportional to  $\beta$  and the attacker knows the watermark up to its magnitude (see (36)). But neither  $\|\mathbf{w}\|$  nor  $\tau$  are recoverable by Claim 1. In this case, any  $\beta \in \mathbb{R}$  is a valid root for (48), as expected. The attacker can also follow the method in Section VI-A.1 for an estimate of the direction of  $\mathbf{w}$ .

### B. General Parametric Detector

In this section, we give the general steps for a sensitivity analysis attack with  $p$  unknown parameters. Let  $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_p^*)$  be the  $p$ -length vector of parameters. The threshold  $\tau$  may be one of these parameters. The detection function depends on  $\boldsymbol{\theta}$ . Denote the difference between this function and  $\tau$  as

<sup>6</sup>Hence, all the approximate equalities are exact for this special case.

$f_{\theta}(\mathbf{z}, \mathbf{w})$ , where  $\mathbf{z} \in \mathbb{R}^n$  is the input to the detector and  $\mathbf{w} \in \mathbb{R}^n$  is the watermark. Therefore the decision boundary is given by the equation  $f_{\theta}(\mathbf{z}, \mathbf{w}) = 0$ .

If the parameter-vector  $\theta^*$  was known by the attacker, one of the algorithms described in Sections IV and V could be used to estimate the watermark by generating  $n$  signals,  $\mathbf{y}^i$ ,  $1 \leq i \leq n$ , on the detection boundary. When  $\theta^*$  is unknown in addition to the watermark  $\mathbf{w}$ , the attacker can just generate additional signals  $\mathbf{z}^i$ ,  $1 \leq i \leq q$ , on the detection boundary, i.e.,

$$f_{\theta^*}(\mathbf{z}^i, \mathbf{w}) = 0, \quad 1 \leq i \leq q, \quad (49)$$

where  $q \geq p$ . For any candidate  $p$ -vector  $\theta$ , an estimate of the watermark<sup>7</sup> can be obtained using one of the algorithms in Sections IV and V. Let  $\mathbf{w}(\theta)$  be this estimate. We propose the following strategy for the attacker: find  $\hat{\theta}$  that minimizes the cost function

$$J_q(\theta) = \sum_{i=1}^q |f_{\theta}(\mathbf{z}^i, \mathbf{w}(\theta))| \geq 0. \quad (50)$$

Then the attacker's estimate of the watermark is  $\mathbf{w}(\hat{\theta})$ .

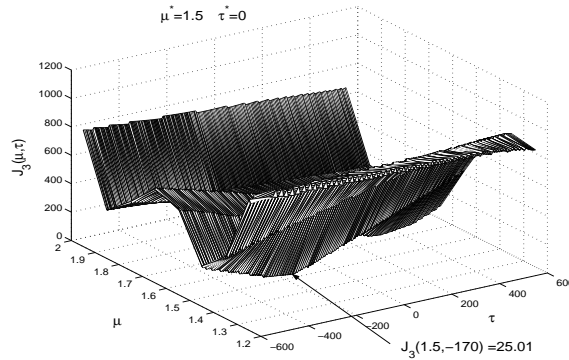


Fig. 6. Cost function  $J_q(\theta)$  with  $q = 3$  and two unknown parameters  $\mu$  and  $\tau$ .

According to the theory of Sections IV and V,  $\mathbf{w}(\theta^*)$  can in principle be a perfect estimate of the watermark, i.e.,  $\mathbf{w}(\theta^*) = \mathbf{w}$ . Then the cost function in (50) is minimized at  $\theta^*$ , i.e.,  $J_q(\theta^*) = 0$  due to (49). If the cost function  $J_q(\cdot)$  admits a single global minimum, then  $\hat{\theta}$  coincides with  $\theta^*$ , and the attacker's strategy is guaranteed to recover the watermark. In practice, the cost function  $J(\theta)$  may have

<sup>7</sup>Not necessarily a good estimate if  $\theta$  differs from  $\theta^*$ .

multiple local minima, so we use a multistart optimization procedure to seek a global minimum<sup>8</sup>. Note that in practice the signals  $\mathbf{z}^i$ ,  $1 \leq i \leq q$ , may not be exactly on the detection boundary but very close to it. For this reason and because of the nonperfect accuracy of the algorithms of Sections IV and V,  $\mathbf{w}(\boldsymbol{\theta}^*)$  is only approximately equal to  $\mathbf{w}$ .

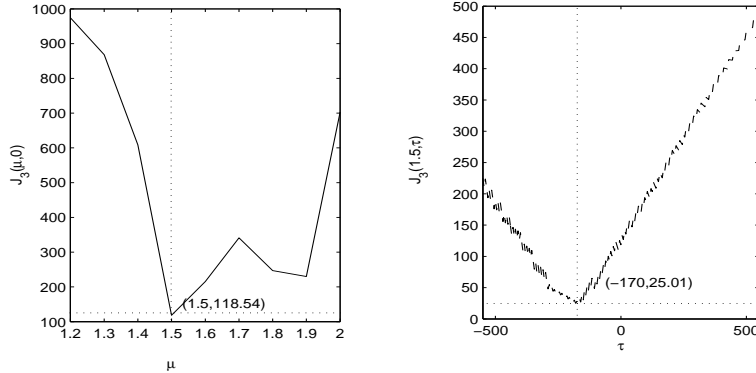


Fig. 7. Left: Cost function with  $\mu$  the only unknown parameter. Right: Cost function with  $\tau$  the only unknown parameter.

In order to illustrate this method, we will consider the GGD detector of Section V-C with the fixed coefficient  $\mu^* = 1.5$  and threshold  $\tau^* = 0$ . The watermarked signal has length  $n = 1024$ . Figure 6 illustrates the case when  $q = 3$  and both  $\mu$  and  $\tau$  are unknown to the attacker ( $p = 2$ ). The cost function  $J_3(\mu, \tau)$  is minimized at  $\mu = 1.5$  and  $\tau = -170$ . If only one of these parameters was unknown to the attacker, then  $p$  is equal to one ( $\boldsymbol{\theta} = \mu$  or  $\boldsymbol{\theta} = \tau$ ) and the minimization problem is one dimensional, hence simpler. To the left of Figure 7, the cost function  $J_3(\mu, \tau^*)$  is shown when  $\mu$  is the only unknown parameter. Similarly, the cost function  $J_3(\mu^*, \tau)$  is presented to the right of Figure 7. Note that the sharpness of the minimum of the cost function increases with  $q$ .

In conclusion, the algorithm succeeds in obtaining a perfect estimate of  $\mu$  since the cost function is minimized at  $\mu^* = 1.5$ . The estimated normalized threshold is  $\frac{1}{n}\hat{\tau} = -0.167$  instead of  $\frac{1}{n}\tau^* = 0$ . Observe that the purpose of the attacker is to estimate the watermark. The threshold  $\tau$  is only used to solve for the parameter  $\beta$  in (38). For  $\tau^* = 0$ , the solution to (43) is  $\beta^* = 2445$ , while for  $\hat{\tau} = -170$  it is  $\hat{\beta} = 2535 \approx \beta^*$ . The normalized correlation  $\rho$  between the watermark and the estimated watermark is equal to 0.988 for  $\tau^* = 0$  and to 0.983 for  $\hat{\tau} = -170$ . Figure 8 shows that  $\rho$  is quite high for a wide range of  $\beta$ .

<sup>8</sup>Depending on the nature of the cost function, the global minimum might or might not be found by the optimization algorithm.

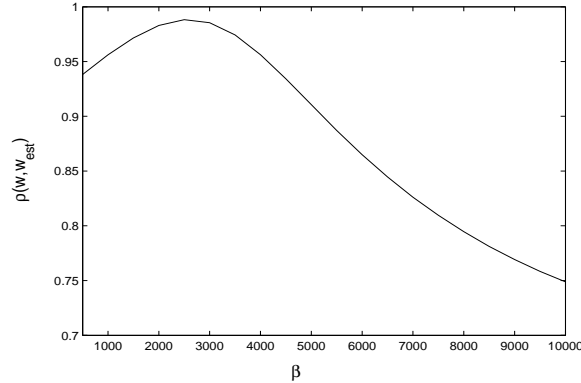


Fig. 8. Normalized correlation between  $w$  and  $\hat{w}$  versus  $\beta$ .

## VII. QUANTIZATION EFFECTS

In order to estimate the watermark  $w$ , the attacker uses the watermarked signal  $x$  to create new signals such as  $y$ ,  $y^i$ , and  $\bar{y}^i$ ,  $\forall i \in \{1, 2, \dots, n\}$ , constructed in Sections IV and V. In practice the detector's input signals are restricted to a region  $\mathcal{B} \subset \mathbb{R}^n$ , therefore the newly created signals have also to belong to  $\mathcal{B}$ . In order to illustrate the concepts, we consider JPEG compressed images [16]. In JPEG image compression, the DCT coefficients of an image are scaled, quantized with integer accuracy, and encoded. Once quantized, these coefficients become integers in the range  $\{-1023, \dots, +1023\}$ . So in this case, the region  $\mathcal{B}$  is  $\{-1023, \dots, +1023\}^n$ , the intersection of the lattice  $\mathbb{Z}^n$  with the hypercube

$$\mathcal{B}_c = [-1023, 1023]^n.$$

Depending on the detection function, a suitable attack algorithm is picked from Sections IV and V and is applied to the quantized, scaled DCT coefficients of the image, components of the signal  $x$ .

Although it might appear that these restrictions make the attacker's task harder, our algorithms can be modified to satisfy these input constraints. The effects of this modification on the performance depend on the nature of the constraints. We first assume that the restriction region is bounded but still connected. Later, we add the constraint of  $\mathcal{B}$  being discrete also. Due to lack of space, we will briefly illustrate the main results (see Table IV). For details, please refer to [9].

In Section IV-B, we described how the basic correlation detector can be modified to account for the constraint that the input belongs to a star-shaped region. A similar extension applies to the generalized correlation detectors of Sections IV-C and IV-D [9]. When we have the additional constraint that the inputs are vectors of integers, i.e.,  $\mathcal{B}$  is discrete, all the auxiliary signals needed by the algorithms of

TABLE IV  
LIMITATIONS IMPOSED BY THE PROPERTIES OF THE RESTRICTION REGION  $\mathcal{B}$ .

Bounded but connected detector's input domain $\mathcal{B}$	Bounded but discrete $\mathcal{B}$
Arbitrary perturbations are not allowed. <b>No loss in performance.</b> Generalized correlator algorithm successfully adapted. No problem for the regular detectors' algorithm, signals occupy small region in $\mathcal{B}$ .	Conservative condition for a successful adaptation for the Generalized correlator algorithm [9]. For the regular detectors' algorithm, case will be studied in future work.

Section IV are quantized to have integer components. Due to quantization, some of these signals may even lie outside the region  $\mathcal{B}_c$ , i.e., have magnitude larger than 1023. Let  $\mathcal{I} = \{i : \mathbf{y}^i \in \mathcal{B}_c, 1 \leq i \leq n\}$  be the index set of the auxiliary signals that belong to  $\mathcal{B}_c$ . Still these signals might not be in  $\mathcal{B}$ . In this case, they are approximated by signals in  $\mathcal{B}$  closest to them and only the watermark components,  $w_i$ , with  $i \in \mathcal{I}$  are estimated using (12). The estimates of the other components are set to zero. Although the attacker may not obtain a perfect estimate of  $\mathbf{w}$ , he may still succeed in removing the watermark resulting in a signal  $\hat{\mathbf{s}}$  in the *rejection region* and with good perceptual quality as shown in VIII. Note that as the quantization gets finer, it is more likely that all auxiliary signals lie in  $\mathcal{B}_c$ .

In the more general case of a regular detector, the main idea of the algorithm is to find a signal  $\mathbf{y}$  on the detection boundary, and  $2n$  signals,  $\bar{\mathbf{y}}^i$  and  $\mathbf{y}^i$ , in a small neighborhood of  $\mathbf{y}$  so that the detection boundary in this neighborhood can be approximated by a hyperplane. The construction of these signals is not affected when boundedness is imposed on the signals input to the detector, and hence there is *no loss in the performance* of the algorithm. The case requiring these signals to take integer values needs further study in order to justify the approximation of the region occupied by the signals  $\mathbf{y}$  and  $\mathbf{y}^i \forall i \in \{1, \dots, n\}$  by a hyperplane.

### VIII. NUMERICAL RESULTS

In this section, we verify the effectiveness of our algorithms by applying them to the three grayscale JPEG images of Figure 9:

- The  $256 \times 256$  *Lena* image.
- A  $128 \times 128$  image, cropped from the original *Lena* image.
- A  $64 \times 64$  image, also cropped from the original *Lena* image.

The quantized DCT coefficients are in the range  $\{-1023, \dots, +1023\}$ . We assume that the detector accepts only images in the JPEG format. Additional implementation details can be found in [9].



Fig. 9. The three test images used: one  $256 \times 256$ , one  $128 \times 128$ , and one  $64 \times 64$ .

### A. Watermark Embedding

In the previous sections, all the signals including the watermark were treated as length  $n$  vectors for mathematical convenience. To describe the simulation results, it is more convenient to use a 2-D representation. In JPEG compression, the image is divided into  $8 \times 8$  blocks, and the 2-D DCT transform of each block is quantized and encoded. We select 13 mid frequencies for watermark embedding, as depicted in Figure 10. In each block, 7 components are chosen randomly and are sampled from  $\{\pm 2\}$  with equal probability. The remaining 6 components are sampled from  $\{\pm 6\}$  also with equal probability. Note that the  $L_p$  norms are the same for all watermarks generated in this way. In particular, the energy per nonzero watermark component is fixed and is equal to 18.7692.

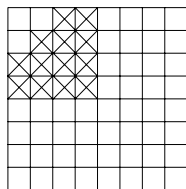


Fig. 10. An  $8 \times 8$  block of the watermark. The squares marked with  $\times$  correspond to the nonzero components of the watermark. All the other components are set to zero.

### B. Correlation Detector

First we study the simple correlation detector of (5). The results of the algorithm are illustrated in Table V, where  $n$  denotes the number of watermarked pixels in the image. The embedding distortion per

TABLE V  
RESULTS FOR ATTACK ON CORRELATION DETECTOR

Image size	$n$	Time	$1.709 \times 10^{-6} n^2$	$D_e$	$D_a$	$D_s$	$\rho$	$\tau$	$d_s$	$d_{\hat{s}}$
$64 \times 64$	832	0.73 s	1.2 s	18.77	18.77	0	1	7808	3198	3198
$128 \times 128$	3328	17.6 s	18.9 s	18.77	18.77	0	1	31232	13704	13704
$256 \times 256$	13312	302.7 s	302.6 s	18.77	18.77	0	1	124928	-312	-312
$256 \times 256$	13312	186.6 s	302.6 s	18.77	22.62	2.62	0.94	124928	-19652	-27844

sample is

$$D_e = \frac{\|\mathbf{x} - \mathbf{s}\|^2}{n},$$

the attack distortion per sample is

$$D_a = \frac{\|\hat{\mathbf{s}} - \mathbf{x}\|^2}{n},$$

and the distortion between the pirated copy  $\hat{\mathbf{s}}$  and the original signal  $\mathbf{s}$  per sample is

$$D_s = \frac{\|\hat{\mathbf{s}} - \mathbf{s}\|^2}{n}.$$

The normalized correlation between the original watermark  $\mathbf{w}$  and the estimated one  $\hat{\mathbf{w}}$  is given as

$$\rho = \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}.$$

Finally, the detection coefficients,  $t(\mathbf{s}, \mathbf{w})$  and  $t(\hat{\mathbf{s}}, \mathbf{w})$  (see (5)), corresponding to the original unwatermarked signal  $\mathbf{s}$  and the estimated signal  $\hat{\mathbf{s}}$  are denoted by  $d_s$  and  $d_{\hat{s}}$ , respectively. The algorithm of Section IV-B was modified as described in Section VII and used to attack these images. Table V shows the results of four experiments using four different realization of the random watermark, and three different image sizes. We note from the first three rows of Table V that the algorithm succeeds at exactly estimating the original image with perfect correlation between the actual watermark and the estimated one,  $\rho = 1$ . However, as mentioned in Section VII, since the feasible region  $\mathcal{B}$  is discrete, the algorithm is not always guaranteed to produce a perfect estimate of the watermark. The fourth row shows an example where the algorithm cannot recover the original image exactly. For 7168 components of the auxiliary signal  $\mathbf{y}$ , the corresponding  $\mathbf{y}^i$  signals lie outside the feasible region  $\mathcal{B}$ , and hence are not valid inputs to the detector. However, one should note that while the algorithm did not manage to completely remove the watermark and recover the original signal  $\mathbf{s}$ , the estimated watermark is very close to the original one:  $\rho = 0.9409$ . Moreover, the constructed signal  $\hat{\mathbf{s}}$  lies in the *rejection* region and is perceptually similar to the original

TABLE VI  
RESULTS FOR ATTACKS ON ML DETECTOR USING GGD HOST SIGNAL MODEL

Image size	$n$	Time	$2.69 \times 10^{-6}n^2$	$D_e$	$D_a$	$D_s$	$\rho$	$\beta^*$	$d_s$	$d_{\hat{s}}$
$64 \times 64$	832	22.3 s	18.5 s	18.77	18.60	0.0037	0.99991	2766	-3198.2	-3152.5
$128 \times 128$	3328	314.2 s	296.6 s	18.77	18.64	0.0047	0.99988	13648	-9833.7	-9659.13
$256 \times 256$	13312	4744.7 s	4745.8 s	18.77	18.71	0.0263	0.9993	71173	-56159.5	-55591.2

signal  $\mathbf{s}$  (see the last row of Table V). Therefore, the algorithm succeeds at “removing” the watermark. Moreover, the algorithm’s complexity is truly  $O(n^2)$ , as evidenced by the excellent linear least-squares fit of running time to  $n^2$ . The difference in execution time for the last two rows of Table V is due to the fact that in the last row, the algorithm required 7168 fewer iterations. To see which case is more typical, we ran 120 independent experiments and observed that for 93.33% of these experiments, the correlation between the true and estimated watermarks was greater than 0.93.

### C. ML detector with Generalized Gaussian Host Model

Next we consider the GGD detector of (39). We apply the attack algorithm of Section V-C to our three test images. The results are shown in Table VI.

The threshold  $\tau$  is zero and the detector uses fixed parameter  $\mu^* = 1.5$ . The nonzero root of (44) is given by  $\beta^*$  in Table VI. The value of the scalars  $\epsilon_i$  in (27) is set to 0.05. Note that the normalized correlation  $\rho$  is *almost equal to one*, despite the non-exactness of (38). However with  $|\epsilon_i| = 0.0005$ , our algorithm is less stable:  $\rho$  is in the order of 0.8 for the  $128 \times 128$  image and 0.7 for the  $256 \times 256$  one. In fact,  $|\epsilon_i|$  should be neither too large nor too small. On one hand, small  $|\epsilon_i|$  is desirable to justify the linearization implicit in (32). On the other hand, if  $|\epsilon_i|$  is too small, other approximation errors will be amplified because  $|\epsilon_i|$  is in the denominator of (38).

Therefore, the algorithm produces an *almost perfect* estimate of the watermark and succeeds at “removing” it by generating an image  $\hat{\mathbf{s}}$  perceptually similar to the original image in the *rejection* region. Note that this algorithm is slower than the correlation detection algorithm because of the more complex nature of the detector. The algorithm is still of order  $O(n^2)$ .

## IX. CONCLUSION

In this paper, we considered sensitivity analysis attacks on additive spread spectrum schemes. In such attacks, the attacker benefits from the availability of a watermarked signal  $\mathbf{x}$  and a watermark detector.

TABLE VII

ALGORITHM TO USE DEPENDING ON THE DETECTION FUNCTION.

Detection function, $t(\mathbf{y}, \mathbf{w})$ , and assumptions about the attacker	Algorithm to use
<ul style="list-style-type: none"> <li>✓ <math>\mathbf{y} \cdot \mathbf{w}</math>, <math>F(\mathbf{y} \cdot \mathbf{w}, \mathbf{y})</math> invertible for given <math>\mathbf{y}</math>, <math>(\mathbf{y} - \hat{\mathbf{s}}(\mathbf{y})) \cdot \mathbf{w}</math>, <math>\tau</math> and all parameters known.</li> <li>✓ Generalized correlator detector and <math>\tau</math> unknown, or <math>(\mathbf{y} - \mathbf{s}) \cdot \mathbf{w}</math>.</li> <li>✓ Generalized smooth detector, <math>\tau</math> and all parameters known.</li> <li>✓ Generalized smooth detector, <math>\tau</math> unknown.</li> <li>✓ Generalized correlator or smooth detector, finite number of unknown parameters.</li> </ul>	<ul style="list-style-type: none"> <li>✓ Variations of generalized correlator detectors’ algorithm, see Table II.</li> <li>✓ Variation of generalized correlator detectors’ algorithm, see Table II and Section VI-A.1.</li> <li>✓ See Table III, for GGD detector, see Section V-C also.</li> <li>✓ See Section VI-A.3 and Table III, for GGD detector, see Section V-C also.</li> <li>✓ See Section VI-B.</li> </ul>

By probing the detector repetitively, his goal is to derive a new signal that “fools” the detector with minimum possible distortion to  $\mathbf{s}$ . We derived new sensitivity attack algorithms that exploit the nature of the detection method and reliably estimate the watermark (refer to Table VII). Once the watermark is estimated, it is “removed” by inverting the embedding function. The set of detection methods vulnerable to such attacks is quite wide. It includes the simple correlation detection method, the normalized correlation detection method, the Patchwork method, the generalized Gaussian host detection method, and any other method that obeys the assumptions specified in Sections IV and V. We also considered the case when a finite number of parameters is unknown by the attacker and showed that this does not improve the security of the watermarking scheme. Most often, only  $O(n)$  detection operations are required to break these schemes whether these parameters are known or not by the attacker. We have also extended our basic algorithms so they can cope with restrictions on input signals that are commonly encountered. For instance, the signals are restricted to bounded regions in Euclidean space, and subject to quantization constraints.

The results of this paper establish the lack of security of one of the most used embedding schemes (additive spread spectrum) and several of its variations. In contrast, high dimensional quantization index modulation schemes (QIM) with randomized lattices present great challenges to attackers [17]. The potential vulnerability of constrained QIM schemes, e.g., scalar QIM, is a topic of current research [8].

**Acknowledgments.** We thank the reviewers for comments and suggestions that have significantly improved this paper.

#### APPENDIX

The derivations of Step 3 in Section V-B are given here. For each  $i = 1, 2, \dots, n$ , using Taylor's remainder theorem, we expand the function  $t(\cdot, \mathbf{w})$  around  $\mathbf{y}$ :

$$t(\mathbf{y}^i, \mathbf{w}) = t(\mathbf{y}, \mathbf{w}) + \mathbf{d}^i \cdot \mathbf{g}(\mathbf{y}, \mathbf{w}) + \eta_i, \quad (51)$$

where  $\mathbf{d}^i = \mathbf{y}^i - \mathbf{y}$  and  $|\eta_i| \leq \frac{\bar{\lambda}}{2} \|\mathbf{d}^i\|^2$  owing to assumption (A2). For small enough  $\|\mathbf{d}^i\|$ , the second-order terms in (51) can be neglected.

Applying the triangle inequality to (35), we obtain

$$\|\mathbf{d}^i\| \leq |\alpha_i - 1| \|\mathbf{y}\| + |\alpha_i \epsilon_i|. \quad (52)$$

When  $\epsilon_i$  tends to zero, the signal  $\bar{\mathbf{y}}^i$  in (27) converges to  $\mathbf{y}$  and consequently  $\alpha_i$  converges to 1. By (52),  $\|\mathbf{d}^i\|$  converges to zero also. Therefore,  $\|\mathbf{d}^i\|$  is made small enough by selecting arbitrarily small  $\epsilon_i$ . Taking this into consideration and substituting (26) and (29) into (51), we obtain

$$|\mathbf{d}^i \cdot \mathbf{g}(\mathbf{y}, \mathbf{w})| \leq \frac{\bar{\lambda}}{2} \|\mathbf{d}^i\|^2, \quad 1 \leq i \leq n. \quad (53)$$

Neglecting the higher-order terms in (51) is equivalent to locally approximating the decision boundary in the neighborhood of the signals  $\mathbf{y}$  and  $\mathbf{y}^i$ ,  $i \in \{1, \dots, n\}$ , by a hyperplane as shown in (32).

#### REFERENCES

- [1] I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *Proc. International Conference on Image Processing (ICIP)*, only CD version of proceedings available, Santa Barbara, CA, 1997.
- [2] J. P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proceedings of the Workshop of Information Hiding*, Portland, OR, April 1998, pp. 258-272.
- [3] T. Kalker, J. P. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *Proc. International Conference on Image Processing (ICIP)*, vol. 1, pp. 425-429, Chicago, IL, October 1998.
- [4] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. San Francisco: Morgan Kaufmann, 2001.
- [5] A. Tewfik and M. Mansour, "Secure watermark detection with nonparametric decision boundaries," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, May 2002, pp. 2089-2092.
- [6] A. Tewfik and M. Mansour, "LMS-based attack on watermark public detectors," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Rochester, NY, September 2002, pp. 649-652.
- [7] M. El Choubassi and P. Moulin, "A new sensitivity analysis attack," in *Proc. SPIE Conf.*, San Jose, CA, January 2005, pp. 734-745.

- [8] P. Comesãna, L. Pérez-Freire, and F. Pérez-González, "The return of the sensitivity attack," in *Proc. International Workshop on Digital Watermarking*, Siena, Italy, September, 2005, pp. 260-274.
- [9] M. El Choubassi, "Novel algorithms for sensitivity analysis attacks," Master thesis, University of Illinois at Urbana-Champaign, IL, ECE Department, Dec. 2005. Available from [www.ifp.uiuc.edu/~cel](http://www.ifp.uiuc.edu/~cel)
- [10] G. D. R. Stinson, *Cryptography, Theory and Practise*. Boca Raton, Florida: CRC Press, 1995.
- [11] F. Müller, "Distribution shape of two-dimensional DCT coefficients of natural images," *Electron. Lett.*, vol. 29, no. 22, pp. 1935-1936, Oct. 1993.
- [12] H. Malvar and D. Florêncio, "Improved spread spectrum: a new modulation technique for robust watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 898-905, Apr. 2003.
- [13] W. Bender, D. Gruhl, N. Marimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, pp. 313-336, 1996.
- [14] S. Voloshynovskiy, S. Pereira and T. Pun, "Attacks on digital watermarks: classification, estimation-based attacks, and benchmarks," *IEEE Communications Magazine*, pp. 2-10, Aug. 2001.
- [15] J. R. Hernández, M. Amado, and F. Pérez-González, "DCT-Domain Watermarking Techniques for Still Images: Detector Performance Analysis and a New Structure," *IEEE Trans. Signal Processing*, vol. 9, no. 1, pp. 55-68, Jan. 2000.
- [16] W. B. Pennebaker and J. L. Mitchell, *The JPEG Still Image Data Compression Standard*. New York, NY: Van Nostrand Reinhold, 1993.
- [17] P. Moulin and R. Koetter, "Data-Hiding Codes," (tutorial paper), in *Proceedings IEEE*, Vol. 93, No. 12, pp. 2083-2127, Dec. 2005.