Detection in Quantization-Based Watermarking: Performance and Security Issues

Luis Pérez-Freire, Pedro Comesaña-Alfaro, Fernando Pérez-González

Dept. Teoría de la Señal y Comunicaciones, ETSI Telecomunicación
Universidad de Vigo, 36310 Vigo, Spain
{lpfreire, pcomesan, fperez}@gts.tsc.uvigo.es

ABSTRACT

In this paper, a novel method for detection in quantization-based watermarking is introduced. This method basically works by quantizing a projection of the host signal onto a subspace of smaller dimensionality. A theoretical performance analysis under AWGN and fixed gain attacks is carried out, showing great improvements over traditional spread-spectrum-based methods operating under the same conditions of embedding distortion and attacking noise. A security analysis for oracle-like attacks is also accomplished, proposing a sensitivity attack suited to quantization-based methods for the first time in the literature, and showing a trade-off between security level and performance; anyway, this new method offers significant improvements in security, once again, over spread-spectrum-based methods facing the same kind of attacks.

Keywords: detection, quantization-based watermarking, performance, security, sensitivity attack.

1. INTRODUCTION

Although the problem of detection for spread-spectrum-based watermarking has been extensively studied, it has been rarely addressed for quantization-based methods,[1] maybe because the latter are considered to be better suited for data hiding applications, but we will show that quantization-based methods present good performance in detection applications. The problem of detection is usually formulated as a binary hypothesis testing; if $\mathbf{z}$ is the signal at the input of the detector, the two considered hypothesis are the following

$$\mathcal{H}_0 : \mathbf{z} \text{ is not watermarked;}$$
$$\mathcal{H}_1 : \mathbf{z} \text{ contains the watermark } \mathbf{w}.$$

Thus, the the answer of the detector is binary, being the space of messages restricted to $\hat{M} = 0$ when no watermark is found, and $\hat{M} = 1$ when it decides that the received signal is a watermarked one.

The method proposed in this paper is based on a data hiding scheme named QP[2] (Quantized Projection), which belongs to the well known family of *spread transform* methods, originally proposed by Chen and Wornell[3] with their implementation ST-DM (Spread Transform - Dither Modulation). The main difference between QP and ST-DM relies on the formulation of the *unprojection function*; in its simplest form, our method is essentially equivalent to ST-DM, but its general formulation clearly differs from that, as we will see in Section 3.3.

Two different analysis are carried out in this paper, in an attempt at separating clearly the robustness (performance) and security[4] issues. The robustness of our method is measured by the Receiver Operating Characteristic (ROC), showing great improvements over traditional spread-spectrum-based schemes; on the other hand, the security assessment is accomplished in an oracle-attack scenario by measuring the difficulty for an attacker to estimate the secrets of the system. As we will see, the security level (to be defined in Section 4) can be increased by constructing more involved detection regions taking advantage of the features of the generalized QP scheme. The idea of improving the security of a scheme under oracle-attacks by complicating the detection region is not new in watermarking, as it was proposed before in several works[5–8]; however, these approaches were developed under the rationale of asymmetric watermarking scenarios, whereas the method proposed here belongs to the class of symmetric schemes.

Throughout the text, boldface lower-case letters will denote column vectors of length L, whereas boldface capital letters are reserved for matrices, and scalar variables will be denoted by italicized lower-case letters. In

this paper, all vectors will be regarded as zero-mean signals, and more specifically, the host $\mathbf{x}$ and the noise $\mathbf{n}$ signal will be modeled as i.i.d. with variance σ_X^2 and σ_N^2, respectively, in each component. To clearly reflect the influence of the embedding distortion and the noise power in the results, two parameters will be introduced: 1) the Document to Watermak Ratio, defined as $\mathrm{DWR} = 10\log_{10}(\lambda)$, with $\lambda = \sigma_X^2/D_w$, being D_w the embedding distortion; and 2) the Document to Noise Ratio, defined as $\mathrm{DNR} = 10\log_{10}(\xi)$, with $\xi = \sigma_X^2/\sigma_N^2$.

The paper is structured as follows. In Section 2, the basic principles of detection in spread-spectrum (SS) based watermarking are reviewed. Section 3 is devoted to the performance analysis of the proposed scheme under AWGN and fixed gain attacks. In Section 4, some security issues are discussed, and an oracle attack suited to our new scenario is proposed, showing some results. Finally, the conclusions are summarized in Section 5.

2. DETECTION IN SPREAD-SPECTRUM-BASED WATERMARKING

This section briefly reviews the basic principles of detection on spread-spectrum-based methods, in order to better understand differences and similarities with the methods proposed in this paper. The embedding function in classical additive spread spectrum is

$$\mathbf{y} = \mathbf{x} + \mathbf{w} = \mathbf{x} + \gamma\mathbf{v}, \tag{1}$$

being γ a parameter controlling the watermark power, and $\mathbf{v}$ a pseudorandom vector such that $||\mathbf{v}||^2 = L$. The embedding distortion is defined as

$$D_w \triangleq \frac{1}{L}\sum_{k=1}^{L} E\{w^2[k]\}. \tag{2}$$

Clearly, for SS, $D_w = \gamma^2$. In the general case where the signal at the input of the detector, $\mathbf{z}$, may be corrupted by additive white Gaussian noise (AWGN) $\mathbf{n}$, the detector must solve the following hypothesis test

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{z} = \mathbf{x} + \mathbf{n} \\ \mathcal{H}_1 &: \quad \mathbf{z} = \mathbf{x} + \mathbf{w} + \mathbf{n} \end{aligned}$$

where $\mathbf{z}$, $\mathbf{x}$, $\mathbf{w}$ and $\mathbf{n}$ are L-dimensional vectors. The hypothesis test is solved by means of a likelihood ratio. When analyzing spread-spectrum-based methods it is usual to assume that the host signal follows a Gaussian distribution; under this assumption, the likelihood ratio becomes equivalent to performing a cross-correlation between the received signal and the vector $\mathbf{v}$. For instance, under hypothesis $\mathcal{H}_1$:

$$r_z = \frac{1}{L}\sum_{k=1}^{L} z[k]v[k] = \frac{1}{L}\sum_{k=1}^{L} x[k]v[k] + \gamma\frac{1}{L}\sum_{k=1}^{L} v[k]v[k] + \frac{1}{L}\sum_{k=1}^{L} n[k]v[k] = \gamma + \frac{1}{L}(r_x + r_n), \tag{3}$$

with $r_x = \sum_{k=1}^{L} x[k]v[k]$ and $r_n = \sum_{k=1}^{L} n[k]v[k]$. The detection function decides whether the received signal is marked or not by means of a thresholding in the cross-correlation

$$d_{SS}(r_z) \triangleq \begin{cases} 1, & r_z \geq T \\ 0, & r_z < T \end{cases} \tag{4}$$

being T a threshold to adjust the operating point of the detector, and $d_{SS}(r_z) \equiv \hat{M}$. The resulting detection region is parameterized by a hyperplane in a L-dimensional space. The performance of the system is measured by the probabilities of false alarm (P_f) and missed detection (P_m), which are defined as

$$P_f = Pr\{\hat{M} = 1|\mathcal{H}_0\} = Pr\{r_z \geq T|\mathcal{H}_0\} \tag{5}$$

$$P_m = Pr\{\hat{M} = 0|\mathcal{H}_1\} = Pr\{r_z < T|\mathcal{H}_1\} \tag{6}$$

The values of P_f and P_m can be easily calculated, yielding[9]

$$P_f = Q\left(\frac{\sqrt{L}T}{\sqrt{\sigma_X^2 + \sigma_N^2}}\right), \tag{7}$$

$$P_m = Q\left(\frac{\sqrt{L}(\gamma - T)}{\sqrt{\sigma_X^2 + \sigma_N^2}}\right), \tag{8}$$

where $Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{t^2}{2}}dt$.

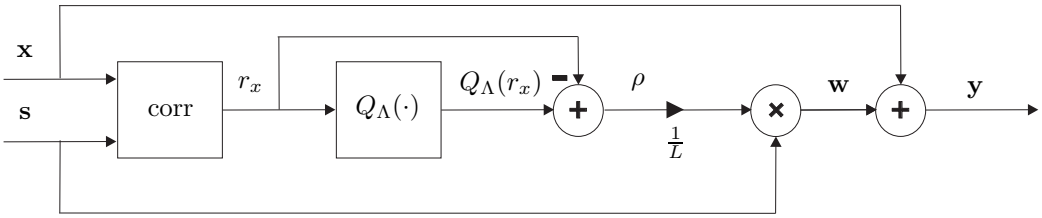


Figure 1. Embedding function in QPD when $M = 1$.

3. QUANTIZATION-BASED DETECTION

Since our method is based on the QP data hiding scheme,[2] we will refer to it as QPD (QP-based Detection) in the sequel. We begin by describing the embedding and detection functions of QPD. Figure 1 depicts the embedding process when the message $M = 1$ is transmitted: first, the correlation between the host signal $\mathbf{x}$ and a pseudorandom vector $\mathbf{s}$ such that $||\mathbf{s}||^2 = L$ is computed yielding a scalar value r_x

$$r_x = \sum_{k=1}^{L} x[k]s[k]. \tag{9}$$

The correlation value r_x is quantized using an Euclidean scalar quantizer $Q_\Lambda(\cdot)$ of step Δ, with its centroids defined by the points in the shifted lattice $\Lambda \triangleq \Delta\mathbb{Z} + \Delta/2$ (the offset $\Delta/2$ is chosen by symmetry reasons). Let $\rho = (Q_\Lambda(r_x) - r_x)$, the watermarked vector is given by $\mathbf{y} = \mathbf{x} + \mathbf{w}$, with $\mathbf{w} = \frac{1}{L}\rho\mathbf{s}$.

For the given projection function and vector $\mathbf{s}$, the embedding distortion (2) is[2] $D_w = \frac{\sigma^2_{R_w}}{L^2}$, being $\sigma^2_{R_w}$ the variance of the quantization error in the projected domain, given by

$$\begin{aligned}\sigma^2_{R_w} &= \int_{-\infty}^{\infty} (Q(r_x) - r_x)^2 f_{R_x}(r_x)dr_x = \sum_{i=-\infty}^{\infty} \int_{i\Delta}^{(i+1)\Delta} \left(i\Delta + \frac{\Delta}{2} - r_x\right)^2 f_{R_x}(r_x)dr_x \\ &= \sum_{i=-\infty}^{\infty} \int_{0}^{\Delta} \left(\frac{\Delta}{2} - r_x\right)^2 f_{R_x}(r_x + i\Delta)dr_x. \end{aligned} \tag{10}$$

The detection function is depicted in Figure 2. In quantization-based detection, the evaluation of the likelihood ratio is difficult, so we propose a suboptimal detection based on the thresholding of the quantization error resulting of applying the quantizer $Q_\Lambda(\cdot)$ to the cross-correlation $r_z = \sum_{k=1}^{L} z[k]s[k]$, that is

$$d_{QPD}(r_z) \triangleq \begin{cases} 1, & |Q_\Lambda(r_z) - r_z| \leq T \\ 0, & |Q_\Lambda(r_z) - r_z| > T \end{cases} \tag{11}$$

where T, as in the case of spread-spectrum, can be varied to adjust the operating point of the detector, and $d_{QPD}(r_z) \equiv \hat{M}$. The resulting detection regions are shown in Figure 3-a. The performance of the system is assessed by means of the probabilities of false alarm (P_f) and missed detection (P_m), which in this case are defined as

$$P_f = Pr\{\hat{M} = 1|\mathcal{H}_0\} = Pr\{|Q_\Lambda(r_z) - r_z| \leq T|\mathcal{H}_0\} \tag{12}$$

$$P_m = Pr\{\hat{M} = 0|\mathcal{H}_1\} = Pr\{|Q_\Lambda(r_z) - r_z| > T|\mathcal{H}_1\} \tag{13}$$

3.1. Performance under AWGN attacks

The to-be-solved hypothesis test is the same as that of Section 2. In order to calculate (12) and (13) we must take into account the probability density function of the received signal $\mathbf{z}$, which will depend on what is the hypothesis

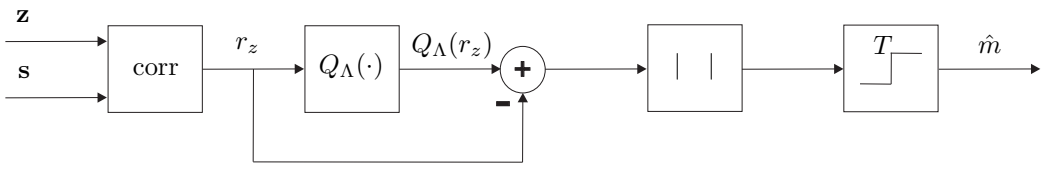


Figure 2. Detection function in QPD.

in force. First, we will accomplish the calculation of the probability of false alarm: under the hypothesis $\mathcal{H}_0$, due to the linearity of the projection function, we can write

$$r_z = \sum_{k=1}^{L} z[k]s[k] = \sum_{k=1}^{L} x[k]s[k] + \sum_{k=1}^{L} n[k]s[k] = r_x + r_n. \tag{14}$$

By resorting to the Central Limit Theorem (CLT) it is possible to show that, for the given projection function and a wide variety of host pdf's, r_x can be accurately modeled by a Gaussian pdf with variance $\sigma^2_{R_x} = L\sigma^2_X$; obviously, the projection of the Gaussian noise is also Gaussian, with variance $\sigma^2_{R_n} = L\sigma^2_N$. Thus, we can conclude that $f_{R_z}(r_z|\mathcal{H}_0) = \mathcal{N}(0, \sigma^2_{R_x} + \sigma^2_{R_n})$. Hence, the probability of false alarm is given by

$$P_f = \sum_{i=-\infty}^{+\infty} \int_{\Delta\left(i+\frac{1}{2}\right)-T}^{\Delta\left(i+\frac{1}{2}\right)+T} f_{R_z}(r_z|\mathcal{H}_0)dr_z = \sum_{i=-\infty}^{+\infty} \left[Q\left(\frac{\Delta\,(i+1/2) - T}{\sqrt{\sigma^2_{R_x} + \sigma^2_{R_n}}} \right) - Q\left(\frac{\Delta\,(i+1/2) + T}{\sqrt{\sigma^2_{R_x} + \sigma^2_{R_n}}} \right) \right]. \tag{15}$$

The pdf of the watermarked signal when hypothesis $\mathcal{H}_1$ is in force is

$$f_{R_z}(r_z|\mathcal{H}_1) = f_{R_y}(r_y) * f_{R_n}(r_n), \tag{16}$$

where $*$ denotes the convolution operator, $f_{R_n}(r_n) = \mathcal{N}(0, \sigma^2_{R_n})$, and $f_{R_y}(r_y)$ is the pdf of the projected watermarked signal in the absence of noise, given by

$$f_{R_y}(r_y) = \sum_{i=-\infty}^{+\infty} \delta(r_y - \Delta(i+1/2))p(c_i), \tag{17}$$

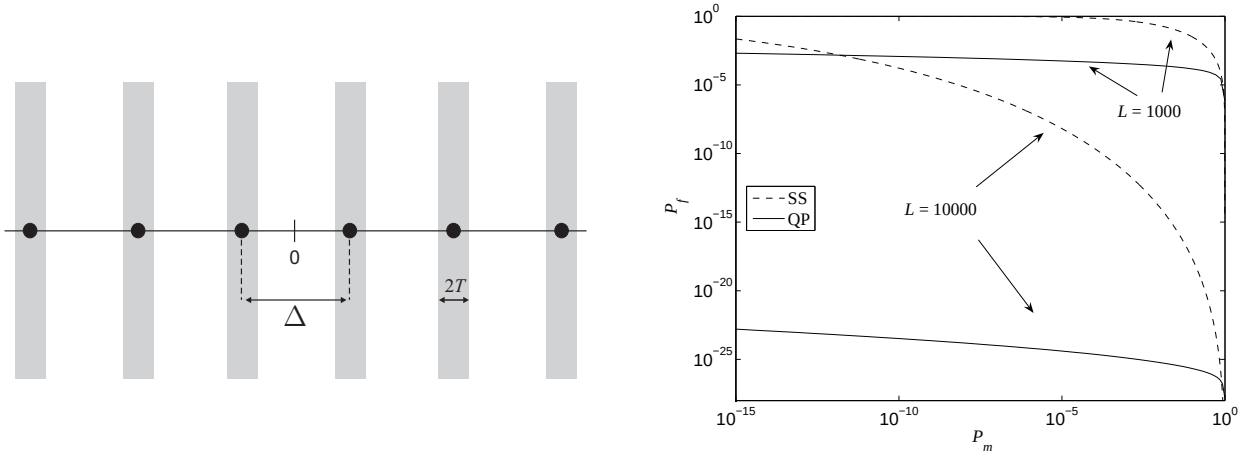
where δ denotes the Dyrac's delta function, and $p(c_i)$ is the probability of the i-th centroid, which under the assumption of Gaussian r_x, is

$$p(c_i) = Q\left(\frac{i\Delta}{\sigma_{R_X}} \right) - Q\left(\frac{(i+1)\Delta}{\sigma_{R_X}} \right). \tag{18}$$

Thus, the probability of missed detection reads as

$$P_m = 1 - \sum_{i=-\infty}^{+\infty} \int_{\Delta\left(i+\frac{1}{2}\right)-T}^{\Delta\left(i+\frac{1}{2}\right)+T} f_{R_z}(r_z|\mathcal{H}_1)dr_z = 1 - \sum_{i=-\infty}^{+\infty} \left[Q\left(\frac{i\Delta - T}{\sigma_{R_n}} \right) - Q\left(\frac{i\Delta + T}{\sigma_{R_n}} \right) \right]. \tag{19}$$

Notice that, in the absence of noise, the probability of missed detection would be null. In Figure 3-b, the ROC for QPD and spread spectrum (SS) are represented, showing that QPD outperforms SS by several orders of magnitude; these results were obtained by setting DWR = DNR = 20 dB, but similar conclusions can be drawn from any other typical values for this variables. Figure 4 shows a comparison between the detection statistics in QPD and SS operating under the same conditions. It can be seen that, for sufficiently large values of L, the probability that r_x is quantized to other centroids than the closest ones to the origin is negligible[2, 10]; the reason for this behavior of QPD is the decreasing of the effective value of the DWR in the projected domain: if we denote by DWR_p the document to watermark ratio in the projected domain, recalling the expressions of D_w and $\sigma^2_{R_X}$ it is easy to realize that $\text{DWR}_p = \text{DWR} - 10\log_{10}(L)$; for instance, for DWR = 30 dB and $L = 1000$,



(a) (b)

Figure 3. Detection regions for QPD (a), and ROC curves for QPD and SS under AWGN attacks, with DWR = DNR = 20 dB and different values of L (b).

we have $\text{DWR}_p = 0$ dB. (which is the case in Figure 4-c). When the two centroids closest to the origin are the only with non-negligible occurrence probability, the detection statistics strongly resemble those of SS, but the host-rejecting feature of QPD reveals its advantage over SS: in SS, the variance of the detection statistics is the same whatever the received signal is watermarked or not, whereas in QPD the variance of the detection statistic under hypothesis $\mathcal{H}_1$ only depends on the noise power.

3.2. Performance under fixed gain attacks

The fixed gain attack accounts for the addition of Gaussian noise plus scaling by an unknown gain factor g. This attack is the Achilles' heel of traditional quantization-based methods, but note that any correlation-based method (as it is also the case of SS) is affected by fixed gain attacks, especially when the decision threshold is not set to 0. The new hypothesis test is the following

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{z} = g\mathbf{x} + \mathbf{n} \\ \mathcal{H}_1: &\quad \mathbf{z} = g(\mathbf{x} + \mathbf{w}) + \mathbf{n} \end{aligned}$$

Clearly, this new scenario encompasses that analyzed in Section 3.1. The probability of false alarm P_f is given by Equation (15), but taking into account that now $\sigma^2_{R_x} = Lg^2\sigma^2_X$. For the calculation of the probability of missed detection, notice that the pdf of the received signal can be written exactly as in Equation (16), but in this case the expression of $f_{R_y}(r_y)$ is

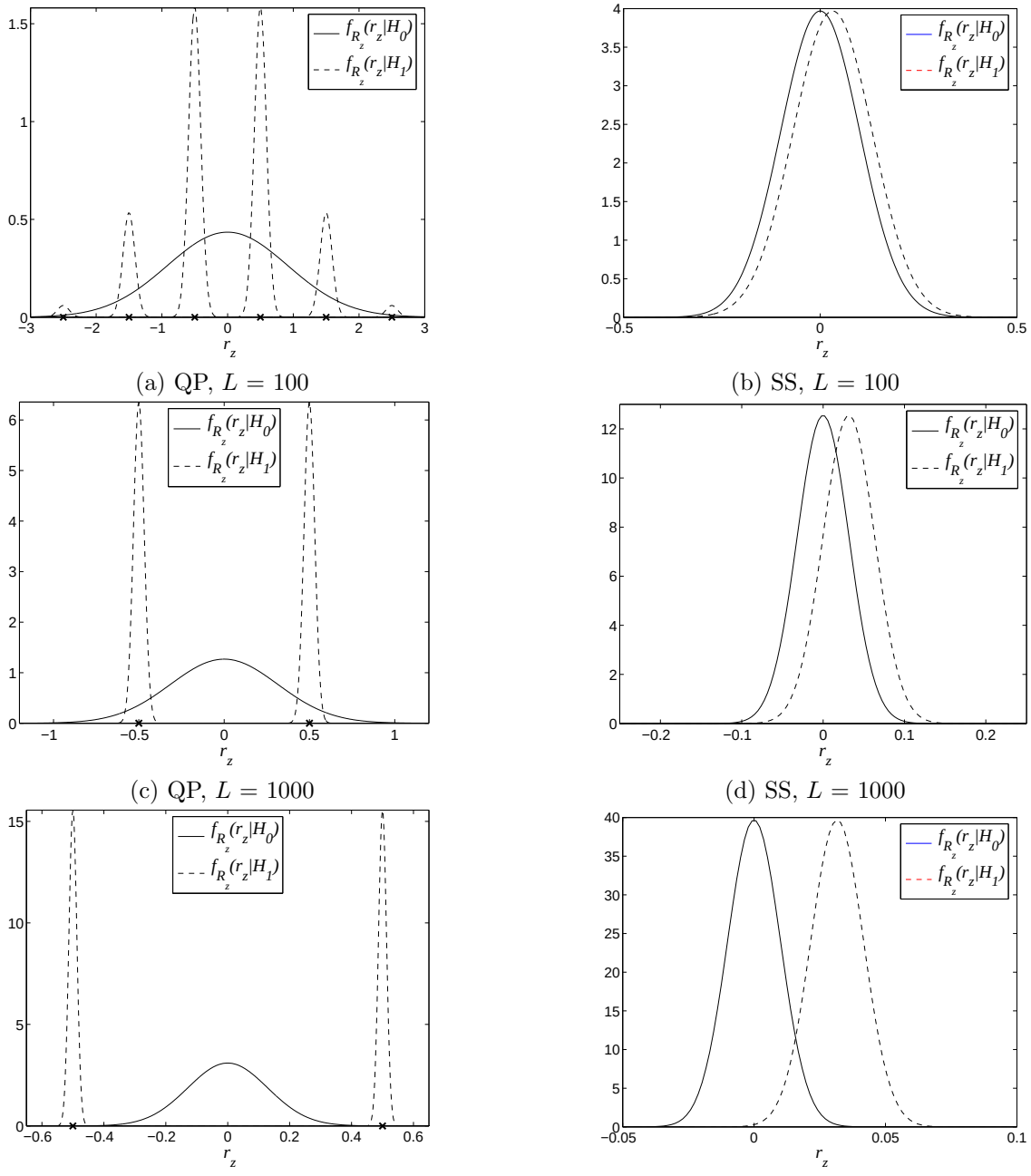
$$f_{R_y}(r_y) = \sum_{i=-\infty}^{+\infty} \delta(r_z - g\Delta(i+1/2))p(c_i), \tag{20}$$

where, again, $p(c_i)$ is given by (18). The probability of missed detection is

$$P_m = \sum_{i=-\infty}^{+\infty} p(c_i)P_{m|c_i}, \tag{21}$$

with $P_{m|c_i} = 1 - P_{d|c_i}$, and

$$\begin{aligned} P_{d|c_i} &= \sum_{k=-\infty}^{+\infty} \int_{\Delta\left(k+\frac{1}{2}\right)-T}^{\Delta\left(k+\frac{1}{2}\right)+T} f_{R_N}(r_z - g\Delta(i+1/2))dr_z \\ &= \sum_{k=-\infty}^{+\infty} \left[Q\left(\frac{\Delta(k+1/2) - T - g\Delta(i+1/2)}{\sigma_{R_n}} \right) - Q\left(\frac{\Delta(k+1/2) + T - g\Delta(i+1/2)}{\sigma_{R_n}} \right) \right]. \end{aligned} \tag{22}$$



(a) QP, $L = 100$ (b) SS, $L = 100$

(c) QP, $L = 1000$ (d) SS, $L = 1000$

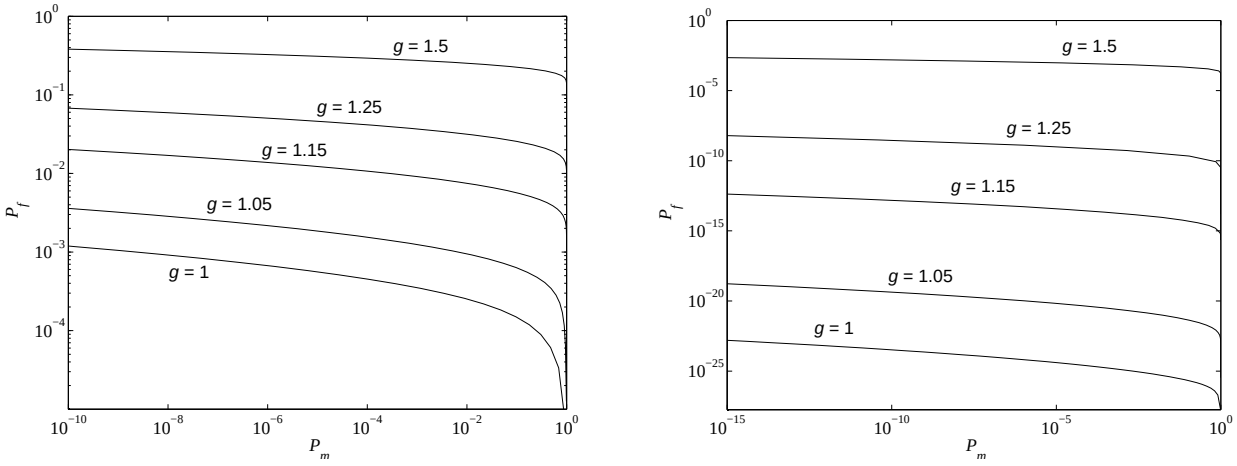
(c) QP, $L = 10000$ (d) SS, $L = 10000$

Figure 4. Comparison of the detection statistics in QPD and SS, for DWR = 30 dB, DNR = 20 dB.

Figure 5 shows the ROC for several values of the gain factor g, evidencing a significant degradation of performance for moderately high values of g. Notice that now, if $g \neq 1$, a null probability of missed detection is no longer guaranteed in the absence of AWGN attacks.

3.3. Generalized QPD

QPD can be generalized[2] so that quantization takes place in a vector subspace. The projection of the host signal into a D-dimensional subspace can be achieved by means of a $L \times D$ projection matrix $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_D)$ as



(b) $L = 1000$ (a) $L = 10000$

Figure 5. ROC for fixed gain attack, with DWR = DNR = 20 dB.

follows

$$\mathbf{r}_x = \mathbf{S}^T\mathbf{x} = (\mathbf{s}_1^T\mathbf{x}, \mathbf{s}_2^T\mathbf{x}, \ldots, \mathbf{s}_D^T\mathbf{x})^T = (r_x[1], r_x[2], \ldots, r_x[D])^T. \tag{23}$$

The watermark in the projected subspace is given by

$$\mathbf{r}_w = Q_D(\mathbf{r}_x) - \mathbf{r}_x, \tag{24}$$

where Q_D is any D-dimensional quantizer. The vector $\mathbf{w}$ is obtained by projecting $\mathbf{r}_w$ onto another subspace with the same dimensionality that the host vector, i.e. L, by means of a $L \times D$ *unprojection matrix* $\mathbf{U}$

$$\mathbf{w} = \mathbf{U}\mathbf{r}_w, \tag{25}$$

so the embedding distortion in this case is

$$D_w = \frac{1}{L}\sum_{k=1}^{L} E\left\{\left(\sum_{i=1}^{D} u_i[k]r_w[i]\right)^2\right\}, \tag{26}$$

where $u_i[k]$ is the k-th element in the i-th column of $\mathbf{U}$. Matrices $\mathbf{S}$ and $\mathbf{U}$ can be whatever matrices which fulfill the following condition, necessary to ensure that $\mathbf{S}^T\mathbf{w} = \mathbf{r}_w$

$$\mathbf{S}^T\mathbf{U} = \mathbf{I}_D, \tag{27}$$

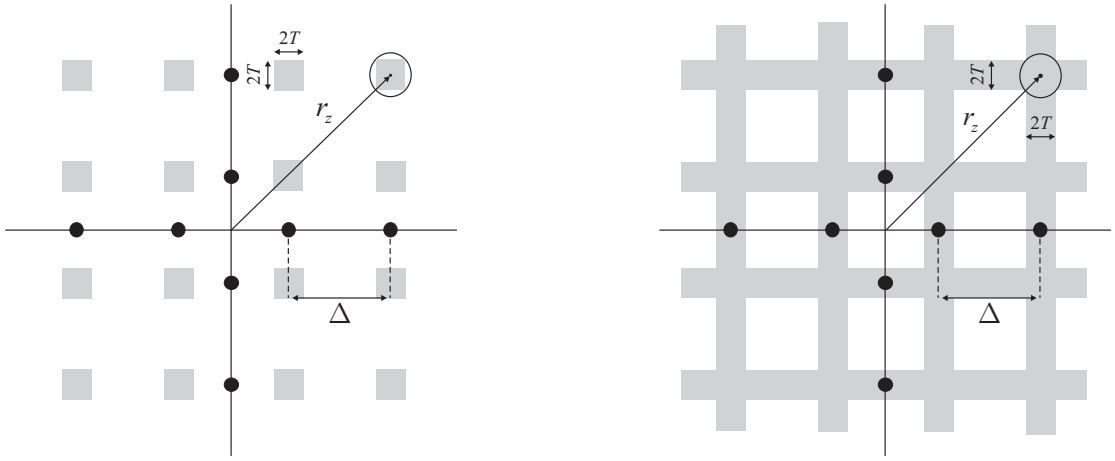
where $\mathbf{I}_D$ is the D-th order identity matrix. Clearly, there exist infinite matrices $\mathbf{U}$ which fulfill that condition, but we are interested in finding the matrix $\mathbf{U}$ such that, given $\mathbf{S}$, the embedding distortion (26) is minimized. Such matrix is given by

$$\mathbf{U} = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}. \tag{28}$$

The condition (27) can be easily satisfied if $\mathbf{S}$ is chosen as a matrix with orthogonal columns. Furthermore, the use of a matrix $\mathbf{S}$ with these characteristics considerably simplifies the theoretical analysis which will be carried out in the following; to this end, we choose the simplest form of orthogonality, which consists of dividing the set of indices $\mathcal{L} = \{1, 2, \ldots, L\}$ in D non-overlapping subsets $\mathcal{L}_i$ of cardinality L/D^*, in such a way that

$$s_i[k] = \begin{cases} m[k], & k \in \mathcal{L}_i \\ 0, & \text{otherwise,} \end{cases} \tag{29}$$

*We assume that L/D is an integer value



(a) AND region (b) OR region

Figure 6. Different detection regions (shaded) in 2-D quantization.

where $\mathbf{m}$ is any L-dimensional vector such that, for the resulting $\mathbf{s}_i$, $||\mathbf{s}_i||^2 = L/D$. Finally, for the given matrix $\mathbf{S}$, the unprojection matrix is simply given by $\mathbf{U} = \frac{D}{L}\mathbf{S}$. The quantization of $\mathbf{r}_x$ may be performed by means of any D-dimensional quantizer; for our analysis we will consider the simplest approach, i.e. the use of a quantizer consisting of the D-Cartesian product of a uniform scalar quantizer. Due to the particular structure imposed to matrix $\mathbf{S}$, we have that[†] $\mathbf{R}_z \sim \mathcal{N}\left(\mathbf{0}, \frac{L}{D}\sigma_X^2\mathbf{I}_D\right)$ and $\mathbf{R}_n \sim \mathcal{N}\left(\mathbf{0}, \frac{L}{D}\sigma_N^2\mathbf{I}_D\right)$. The D-dimensional detection regions depend on the formulation of the detection function, thus providing an additional degree of freedom. We propose two different detection functions, defined as follows

$$d_{AND}(\mathbf{r}_z) \triangleq \begin{cases} 1, & |Q_\Lambda(r_z[i]) - r_z[i]| \leq T \ \text{ for all } i \\ 0, & \text{otherwise} \end{cases}$$

$$d_{OR}(\mathbf{r}_z) \triangleq \begin{cases} 1, & |Q_\Lambda(r_z[i]) - r_z[i]| \leq T \text{ for at least one } i \\ 0, & \text{otherwise.} \end{cases}$$

Note that the d_{AND} function yields detection regions (namely, AND regions) consisting of D-dimensional non-overlapping hypercubes, but the OR regions resulting from the d_{OR} function are more involved. As an illustrative example, the resulting detection regions for $D = 2$ are illustrated in Figure 6. Needless to say that these two detection functions are equivalent in scalar quantization. The probabilities of false alarm and missed detection for the D-dimensional AND region are given by

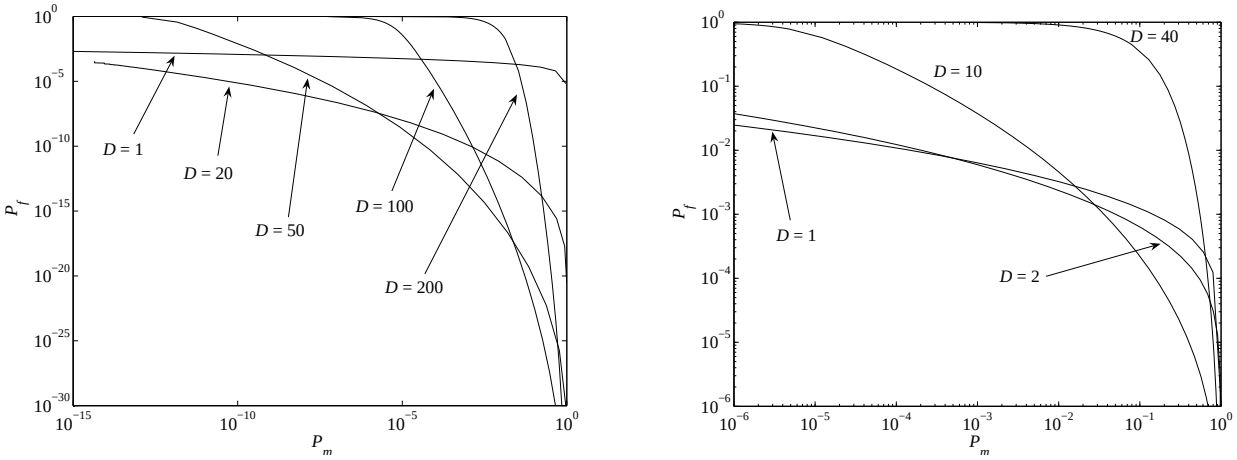
$$P_{f,AND}(D) = P_f^D, \qquad P_{m,AND}(D) = 1 - (1 - P_m)^D, \tag{30}$$

whereas for the OR region are the following

$$P_{f,OR}(D) = 1 - (1 - P_f)^D, \qquad P_{m,OR}(D) = P_m^D. \tag{31}$$

In (30) and (31), P_f and P_m stand for the expressions in the unidimensional case, namely (12) and (13), respectively. The OR function provides poorer performance, but it may be interesting from a security perspective, as we will see in Section 4. Figure 7 shows the resulting ROC's for different dimensionalities using the AND function; it can be seen that, given one of the probabilities, P_f or P_m, there exists an optimum number of dimensions D^* which minimizes the other probability. The calculation of D^* is rather involved, since it depends on the DWR and the DNR, and it will not be accomplished in this paper. Although in our analysis we have assumed the particular form of orthogonality given by (29), the results are essentially the same for any general

[†]Note that the value of L/D must be large enough to assure the validity of the CLT.



(a) DNR = 20 dB (b) DNR = 10 dB

Figure 7. ROC for AND region and different dimensionalities of the projection subspace, with DWR = 20 dB and $L = 1000$.

case of orthogonality. Figure 8-a compares the theoretical results with numerical ones by means of Monte Carlo simulations using matrices $\mathbf{S}$ whose elements are i.i.d. Gaussian random variables[‡].

Orthogonality in the columns of $\mathbf{S}$ is desirable to achieve the best performance possible; however, as we will discuss in Section 4, the introduction of a certain correlation between the components of the projected signal may be interesting from a security point of view. The obtention of a specified covariance matrix in the projected domain is very simple: let $\mathbf{R}$ be the $D \times D$ desired covariance matrix whose eigen-decomposition is

$$\mathbf{R} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T. \tag{32}$$

If $\mathbf{S}$ is an orthonormal matrix and we use the projection matrix defined as

$$\mathbf{S}' \triangleq \mathbf{S}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}^T, \tag{33}$$

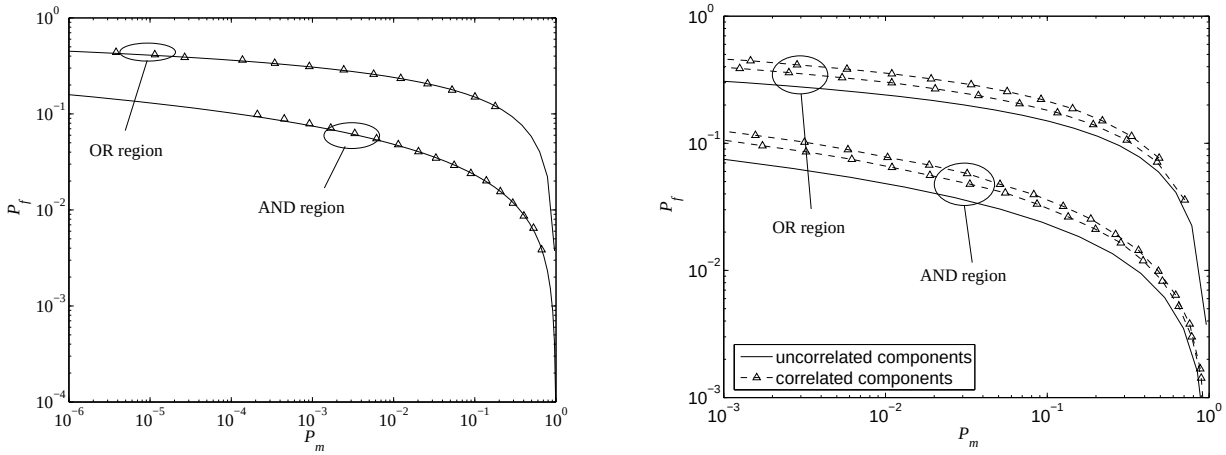
then it is easy to show that $\mathbf{R}_{\mathbf{r}_x} = E\{\mathbf{r}_x\mathbf{r}_x^T\} = \sigma_X^2\mathbf{R}$, with $\mathbf{r}_x = \mathbf{S}'^T\mathbf{x}$. Figure 8-b shows numerical results by means of Monte Carlo simulations for different degrees of correlation between components in 2-D quantization, with stronger correlations yielding poorer performance.

4. SECURITY

The concept of security in watermarking is still diffuse, in the sense that there is no full agreement on what are the relevant variables to assess the security of a system. For people working in the field of cryptography it is well known the Kerckhoff's principle,[11] which claims that the security of a cryptographic system must rely solely on the secret key; the translation of this principle to our scenario implies that all parameters of the algorithm (Δ, T, L, D and the chosen detection function, in our case) are publicly known, with exception of the projection matrix $\mathbf{S}$, which can be made pseudorandom to play the role of the secret key. This way, the security analysis of this section will deal with attacks (obviously intentional) whose aim is to gain knowledge of the secret matrix $\mathbf{S}$. This approach allows to make a clear distinction between attacks to security and attacks to robustness, which were addressed in Section 3. Another approach to the assessment of watermarking security relying on these concepts can be found in.[4]

Although a great amount of threats to security exist, much of the researchers' attention has been paid to the category of *oracle attacks*, i.e. those attacks which try to exploit the availability of detection devices (usually in

[‡]Note that, with i.i.d. Gaussian elements, the columns of $\mathbf{S}$ will be almost perfectly orthogonal for moderately high values of L.



(a) **S** with orthogonal columns (b) **S** with non-orthogonal colums

Figure 8. ROC with 2-dimensional quantization, for DWR = DNR = 20 dB, and $L = 100$. Solid lines and points stand for theoretical and empirical values (with Gaussian random vectors), respectively.

the form of a black box). We will focus on the so-called *sensitivity attack*,[12] which consists in modifying the signal at the detector input in a component-by-component basis, in order to estimate the points on the boundary of the detection region; when this region is given by a hyperplane, which is the case of SS detection based on linear correlation (4), the attack succeeds when L different points of the boundary are correctly estimated; moreover, this is equivalent to estimating the secret vector $\mathbf{v}$ of Equation (1). The complexity of a successful attack for SS-based methods have been shown to be $O(L)$.[13] Most of the proposed approaches in the literature[5, 7, 8] to overcome this problem rely on the construction of more involved detection regions, much of them parameterized in a non-linearly fashion, which in many cases may yield hard-to-implement schemes. What we propose here is to take advantage of the properties of the generalized QPD scheme described in Section 3.3 to construct detection regions based on the combination of linearly parameterized regions. The advantage of this approach is that the detection region can be progressively complicated just by increasing the dimensionality of the projected subspace, but still allowing the design of implementable schemes from a practical point of view.

It its important to note that classical sensitivity attacks performed for spread spectrum will not work well for QPD, since that algorithm is designed assuming that there is only a single boundary to estimate. For example, in the simplest form of QPD, with scalar quantization ($D = 1$), the detection function strongly resembles that of SS, but the boundary of the detection region in QPD (Figure 3-a) is multiple, since it is defined by shifted versions of the hyperplane orthogonal to the secret vector $\mathbf{s}$. In this section, a new sensitivity attack suited to this new scenario is proposed. The main steps of the algorithm are outlined in the following lines.

The proposed algorithm tries to estimate a projection vector in the AND region case or a linear combination of the projection vectors in the OR region case. In fact, in both cases it tries to compute the minimum-normed vector which modifies the watermarked vector in such a way that the output of the detector is $\hat{M} = 0$.

The algorithm consists in the addition of an attacking vector $\boldsymbol{\alpha}$ to the watermarked signal $\mathbf{y}$. This $\boldsymbol{\alpha}$ is initially set to an observation of a random variable which is iteratively modified. In the k-th iteration, one of L_β possible modification vectors $\boldsymbol{\beta}_{k \bmod L_\beta}$[§], scaled by τ_k (τ_0 is a parameter of the algorithm), is chosen, and added and subtracted to the attacking vector, yielding $\boldsymbol{\gamma}_k^+ = \boldsymbol{\alpha} + \tau_k \boldsymbol{\beta}_{k \bmod L_\beta}$ and $\boldsymbol{\gamma}_k^- = \boldsymbol{\alpha} - \tau_k \boldsymbol{\beta}_{k \bmod L_\beta}$. The two resulting vectors are scaled to be in the boundary of the detection region:

$$\boldsymbol{\delta}_k^+ = a_k^+ \boldsymbol{\gamma}_k^+ \tag{34}$$

with a_k^+ the minimum positive number such that $d_{AND}(\mathbf{y} + \boldsymbol{\delta}_k^+) = 0$ or $d_{OR}(\mathbf{y} + \boldsymbol{\delta}_k^+) = 0$, and equivalently for $\boldsymbol{\delta}_k^-$. If $\min(||\boldsymbol{\delta}_k^+||, ||\boldsymbol{\delta}_k^-||) < ||\boldsymbol{\alpha}||$ (but always in the first iteration), then $\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\rho} = \{\boldsymbol{\delta}_k^+, \boldsymbol{\delta}_k^-\}} ||\boldsymbol{\rho}||$.

[§]In our implementation we have chosen that the modification is component-by-component, but it could also be done by choosing a random vector for the modification.

If $\boldsymbol{\alpha}$ is not modified for any $\boldsymbol{\beta}_{k \bmod L_\beta}$, the value of τ_k is reduced[¶]; otherwise, $\tau_{k+1} = \tau_k$. The value of τ_0 will determine the rate of convergence of the algorithm. The computation of a_k^+ (or a_k^-) is performed by a first step multiplying the $\boldsymbol{\gamma}_k^+$ ($\boldsymbol{\gamma}_k^-$) by a constant (2 in the implementation) until the detector output is $\hat{M} = 0$. After that, the boundary is estimated by a bisection algorithm.

When successful, the algorithm provides one vector $\mathbf{y} + \boldsymbol{\alpha}$ at the boundary of the detection region which is at minimum distance of the watermarked vector $\mathbf{y}$. For the AND region, the vector $\boldsymbol{\alpha}$ is co-linear with one of the columns of $\mathbf{S}$; for the OR region, $\boldsymbol{\alpha}$ is a linear combination of the columns of $\mathbf{S}$. Note that, due to the nature of the algorithm, the addition of a random dither signal in the projected domain will not provide any additional security to the system.

To quantitatively assess the security of our system, we must first define some measures:

- For the AND region, we define

$$\eta_{AND}(\boldsymbol{\alpha}, \mathbf{S}) \triangleq \max_{i=1,\ldots,D} \left\{ \frac{|\boldsymbol{\alpha}^T \mathbf{s}_i|}{||\boldsymbol{\alpha}|| \cdot ||\mathbf{s}_i||} \right\}, \tag{35}$$

  where $\mathbf{s}_i$ is the i-th column of the projection matrix $\mathbf{S}$. Notice that (35) can be interpreted as the cosine of the minimum angle formed by $\boldsymbol{\alpha}$ and the columns of $\mathbf{S}$, and gives a measure of the co-linearity between $\boldsymbol{\alpha}$ and the column of $\mathbf{S}$ that the algorithm tried to estimate. Similarly, we can define for the OR region

$$\eta_{OR}(\boldsymbol{\alpha}, \mathbf{S}) \triangleq \max_{i=1,\ldots,2^D} \left\{ \frac{|\boldsymbol{\alpha}^T \mathbf{v}_i|}{||\boldsymbol{\alpha}|| \cdot ||\mathbf{v}_i||} \right\}, \tag{36}$$

  where $\mathbf{v}_i$ is one of the 2^D possible linear combinations obtained by multiplying the columns of $\mathbf{S}$ by ± 1.

- Let us also define

$$\varphi(\boldsymbol{\alpha}) \triangleq 20 \log_{10} \left(\frac{||\boldsymbol{\alpha}||}{K} \right), \tag{37}$$

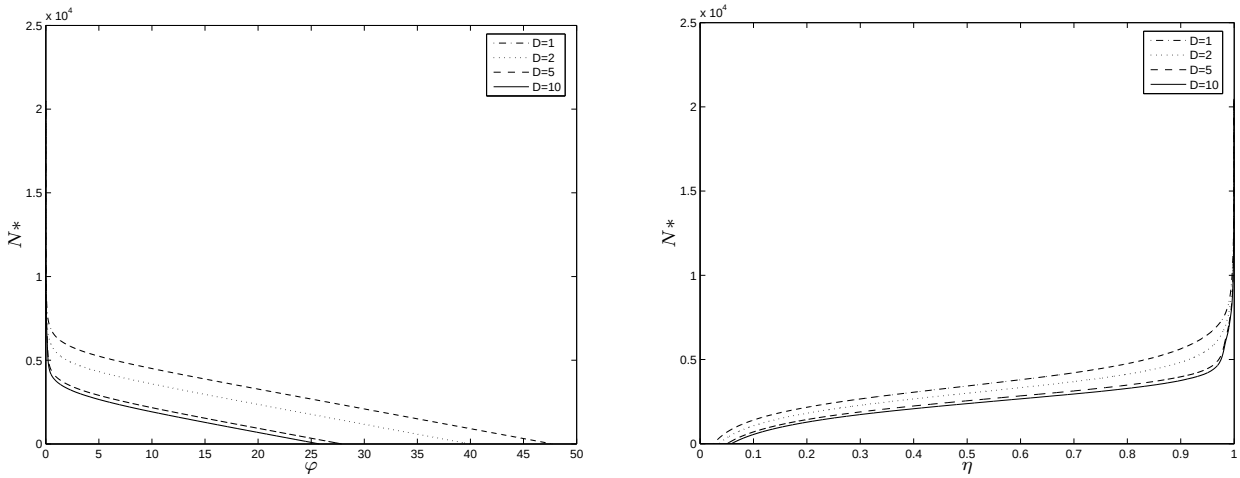
  where K is the norm of the minimum-normed vector which moves the marked signal outside of the detection region. Note that, for the AND region, $K = T$, but for the OR region we have $K = \sqrt{D}T$. Equation (37) is related to the saving in the power necessary to perform a successful attack (i.e. generating an unwatermarked signal) when the output of the algorithm is taken into account.

With the definitions given by (35), (36) and (37), we can give two alternative, although related, definitions of *security level*:

1. the number of iterations N^* of the algorithm which are necessary to bring the value of (35) or (36), depending on the considered detection function, above a certain threshold;
2. the number of iterations N^* of the algorithm necessary to bring the value of (37) below a certain threshold.

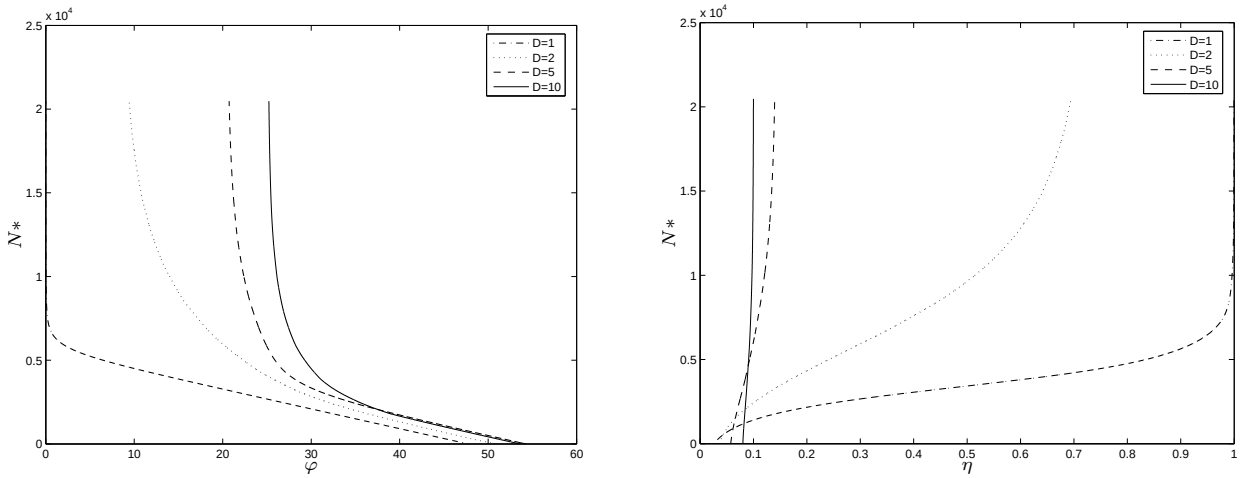
Figure 9 shows the results for the AND regions: as can readily be seen, the larger D, the better the estimate for a given number of iterations of the algorithm. This seems to be somewhat contradictory, but bear in mind that the probability of finding a correct projection direction is increasing with D. Anyway, the algorithm succeeds in finding a correct direction regardless the value of D. The results for the OR region are represented in Figure 10. The situation with respect to the AND region is radically different: the estimation becomes more difficult for increasing values of D, and even for a large number of iterations, the estimation is really poor, which can be attributed to the characteristics of the to-be-minimized function (the norm of vector $\boldsymbol{\alpha}$). At this point, we want to remark that the classical sensitivity attack (i.e., intended for SS) applied to QPD in the simplest case of $D = 1$ yielded really poor results ($\eta_{AND} = \eta_{OR} < 0.1$).

[¶]In our implementation, $L_\beta = L$ and $\tau_{k+1} = 0.7\tau_k$.



(a) (b)

Figure 9. Security levels for the AND region ($L = 1024$).



(a) (b)

Figure 10. Security levels for the OR region ($L = 1024$).

Be aware that the proposed algorithm provides a method to unwatermark watermarked signals, but it is not sufficient to generate valid watermarked signals, because the it is only aimed at disclosing one secret vector or a linear combination of all of them, depending on the detection function. To disclose the whole projection matrix, the algorithm should be modified to search only in the remaining possible directions. For example, considering an AND detection function, if the attacker knows that the projection matrix is orthogonal, once one column is disclosed, the algorithm proposed here can be used to estimate the other ones by restricting the modifying vectors to directions orthogonal with the already disclosed one. Of course, this simple adaptation of the algorithm is not suitable to perform a complete attack when OR functions are considered. Taking into account the previous adaptation, we want to note that the use of non-orthogonal projection matrices can be interesting to improve security.

5. CONCLUSIONS AND FURTHER WORK

In this paper, a novel method for detection in quantization-based watermarking has been presented and analyzed. The method is based on the quantization of a projection function applied to the host signal, and it can be adapted

to the requirements of a wide variety of scenarios by an appropriate selection of its parameters, which allow to select the desired levels of robustness and security, but keeping always in mind the existence of a trade-off between these two quantities, in the sense that their simultaneous maximization, although desirable, is not possible. The most remarkable features of this new method are its great improvement in performance compared to traditional spread spectrum methods, and the high security level against oracle-like attacks provided by the the combination of the OR detection function with a moderately high number of dimensions in the projected domain.

Although the results presented in this paper are very promising, a lot of work remains to be done. Concerning robustness, the optimization of the number of dimensions of the projected subspace, and the study of the threshold in QPD in terms of the Neyman-Pearson criterion appear as interesting future lines to explore. Finally, concerning the security aspects, we are mainly interested in the refinement of the proposed sensitivity attack for the estimation of the whole projection matrix, so as to provide a more exact security level.

ACKNOWLEDGMENTS

This work was partially funded by *Xunta de Galicia* under projects PGIDT04 TIC322013PR and PGIDT04 PXIC32202PM; MEC project DIPSTICK, reference TEC2004-02551/TCM; FIS project G03/185, and European Comission through the IST Programme under Contract IST-2002-507932 ECRYPT. ECRYPT disclaimer: the information in this document reflects only the authors' views, is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

REFERENCES

1. T. Liu and P. Moulin, "Error exponents for one-bit watermarking," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, **3**, pp. 65–68, 6-8 April 2003.
2. F. Pérez-González, F. Balado, and J. R. Hernández, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Transactions on Signal Processing* **51**, pp. 960–980, April 2003.
3. B. Chen and G. Wornell, "Quantization Index Modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory* **47**, pp. 1423–1443, May 2001.
4. F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: theory and practice," *IEEE Transactions on Signal Processing* , 2005. To appear.
5. J. J. Eggers, J. K. Su, and B. Girod, "Public key watermarking by eigenvectors of linear transforms," in *European Signal Processing Conference (EUSIPCO)*, (Tampere, Finland), September 2000.
6. T. Furon and P. Duhamel, "An asymmetric public detection watermarking technique," in *Proc. of the third Int. Workshop on Information Hiding*, A. Pfitzmann, ed., pp. 88–100, Springer Verlag, (Dresden, Germany), September 1999.
7. T. Furon, B. Macq, N. Hurley, and G. Silvestre, "JANIS: Just Another n-order Side-Informed Watermarking Scheme," in *International Conference on Image Processing (ICIP)*, **2**, pp. 22–25 September, (Rochester, NY, USA), 6-8 April 2002.
8. M. F. Mansour and A. H. Tewfik, "Secure detection of public watermarks with fractal decision boundaries," in *European Signal Processing Conference (EUSIPCO)*, (Toulouse, France), 2002.
9. M. Barni and F. Bartolini, *Watermarking Systems Engineering*, Signal Processing and Communications, Marcel Dekker, 2004.
10. L. Pérez-Freire, F. Pérez-González, and S. Voloshinovskiy, "Revisiting quantization-based data-hiding: exact analysis and results," *IEEE Transactions on Signal Processing* , 2004. Submitted.
11. A. Kerckhoff, "La cryptographie militaire," *Journal des sciences militaires* **9**, pp. 5–38, January 1883.
12. I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *IEEE International Conference on Image Processing ICIP'97*, **3**, (Santa Barbara, California, U.S.A.), October 1997.
13. T. Kalker, J. P. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *IEEE Int. Conf. on Image Processing, ICIP'98*, pp. 425–429, (Chicago, IL, USA), October 1998.