Optimal Data-Hiding Strategies for Games with BER Payoffs.

Pedro Comesaña, Fernando Pérez-González, and Félix Balado

Dept. Tecnologías de las Comunicaciones. ETSI Telecom., Universidad de Vigo, 36200 Vigo, Spain
pcomesan@gts.tsc.uvigo.es, fperez@tsc.uvigo.es, fiz@tsc.uvigo.es

Abstract. We analyze three different data hiding methods from a game-theoretic point of view, using the probability of bit error as the payoff. Those data hiding schemes can be regarded to as representatives of three families of methods: spread-spectrum, informed-embedding and hybrid. In all cases, we have obtained theoretical expressions for the BER which are then optimized to derive the strategies for both the attacker and the decoder, assuming that the embedder simply follows point-by-point constraints given by the perceptual mask. Experimental results supporting our analyses are also shown, with examples of watermarking in the spatial domain as well as the DCT domain.

1 Introduction

Some researchers have dealt with game-theoretic aspects of data hiding capacities [1], [2]. However payoffs other than channel capacity are also possible in the data hiding game, as already suggested in [3] and developed in [4]. Here, we build on this idea to determine optimal playing strategies by considering that the bit error rate (BER) for the hidden information defines the payoff in the game.

The main purpose of this paper will be to obtain those optimal strategies for the three main classes of data hiding methods, namely, spread-spectrum, quantization-based and hybrid schemes. We have chosen a representative algorithm of each kind and developed closed-form expressions for the bit error probability which are then used as cost functions for deriving optimal or near-optimal tactics for the decoder and the attacker. To the authors' knowledge, the closest works to ours are those of Eggers and Girod in [2] and Moulin and Ivanovic in [4], compared to which the two main differences are: 1) the game payoff, which is channel capacity in [2] (although specifically optimized for each method) and probability of correct detection (zero-rate spread-spectrum scheme) in [4]; and 2) the agents involved, that are the embedder and the attacker in both mentioned works, whereas here we consider them to be the attacker and the decoder for reasons which will be explained later.[1] Note that while in a data hiding game

[1] The exception is Quantized Projection data hiding, for which both the embedding and decoding strategies are chosen simultaneously (see Section 5.)

involving only an embedder and an attacker, the latter has the final word, this is clearly not the case if we consider that there is optimization at the final decoding stage.

As we have said, three agents generally play the data hiding game: embedder, attacker and decoder, as depicted in Figure 1. Each one has different objectives and constraints which frequently lead to colliding interests. First of all, the embedder is responsible for hiding the information in the host image in a secret way, trying to hinder the estimation of the watermark by the attacker. In most applications, the embedder faces very strict perceptual constraints so as to guarantee that there is no loss of value in the watermarked image.

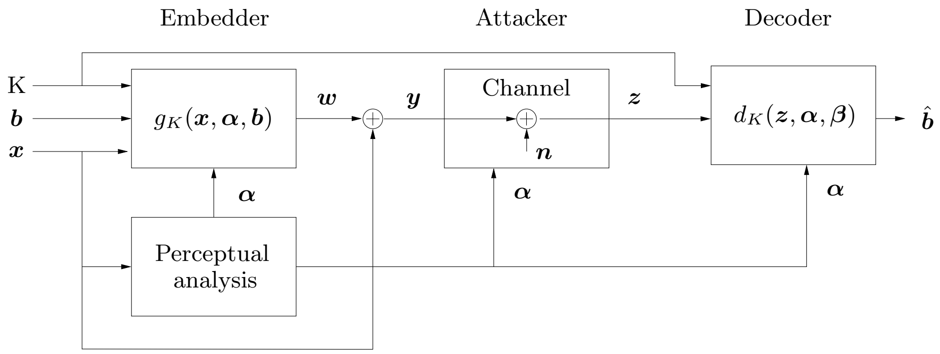

Fig. 1. Data hiding model.

Second, the watermark is manipulated by the attacker. The attacker can play an active role by frustrating or at least making it difficult the decoding process. Needless to say, the attacker's strategy will also be constrained by a distortion measure with some perceptual meaning. In this paper we will consider only additive noise attacks.

Finally, the decoder is in charge of retrieving the information in a reliable way. He/she could have information about the attacker's strategy (this is more likely in unintentional attacks), and in this case, he/she could use this information to improve the performance of the system. Note that the converse could also be true, but we must note once again that the decoder does have the final word.

The paper is structured as follows: Section 2 is devoted to defining the problem and discussing several ways of measuring distortions; optimal strategies are presented in Sect. 3 for spread-spectrum systems, in Sect. 4 for binary dither modulation schemes, and in Sect. 5 for quantized projection methods. Experimental results are shown in Section 6 while Section 7 contains our conclusions and discusses future lines of research.

2 Problem Statement

For notational simplicity, we will assume that host signal samples in any domain given are arranged following a vector, denoted by bold characters. The same notation will be used for matrices.

We will consider here only additive data hiding, that is, given the *host image* $\boldsymbol{x}$ and the watermark $\boldsymbol{w}$, the resulting watermarked image can be written as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w} \tag{1}$$

We will also follow the customary scheme [5], [6] for embedding a particular information bit b_i by using a tile $\mathcal{S}_i \triangleq \{k_1, \cdots, k_{|\mathcal{S}_i|}\}$ of pseudorandomly chosen indices, selecting $L_i = |\mathcal{S}_i|$ samples from $\boldsymbol{x}$ depending on a cryptographic key K, so that this tile will be known to the decoder, but not to the attacker. Also we will assume that tiles $\mathcal{S}_i$ and $\mathcal{S}_j$, for all $i \neq j$ do not overlap. Each one of the possible partitions of the host signal in this way will be denoted as $\mathcal{T}$, and $\mathcal{U}$ will the set containing them. If N bits are going to be hidden, we will define $\mathcal{S} = \bigcup_{i=1}^{N} \mathcal{S}_i$. Throughout this paper we will not elaborate on the possibility of adding an upper coding layer.

At the embedder's side, a perceptual mask vector $\boldsymbol{\alpha}$ is computed from the host signal $\boldsymbol{x}$. For practical effects, we will consider that $\boldsymbol{x}$ is an image in the spatial or the DCT domain. This vector $\boldsymbol{\alpha}$ indicates the maximum allowed watermark energy that produces the least noticeable modification of the corresponding sample of $\boldsymbol{x}$.

Consequently, we will regard the watermark $\boldsymbol{w}$ as being produced from the desired information vector $\boldsymbol{b}$ by using a certain function, $\boldsymbol{w} = g_K(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{b})$, where K is a secret key. For notational simplicity we will assume that $b_i \in \{\pm 1\}$, $i = 1, \cdots, N$. It must be also taken into account that the function $g_K(\cdot)$ depends on the specific watermarking scheme, as we will later confirm.

We will assume an additive probabilistic noise channel for modeling attacks. Therefore, the image at the decoder's input $\boldsymbol{z}$ can be written as $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{n} = \boldsymbol{x} + \boldsymbol{w} + \boldsymbol{n}$, where $\boldsymbol{n}$ is noise independent of $\boldsymbol{x}$. Therefore we are not considering all the possible range of attacks (think of *JPEG* as an example). By virtue of the pseudorandom choice of the indices in $\mathcal{S}$ we may assume that the samples in $\boldsymbol{n}$ are also mutually independent, with zero mean and variances $\sigma_{n_i}^2$, $i \in \mathcal{S}$.

The decoder uses a certain decoding function $\hat{\mathbf{b}} = d_K(\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ are some weights used to improve the decoding process. Then, the BER for the i-th bit is just $P_e(i) = P\{\hat{b_i} \neq b_i\}$, and the game consists in the maximization/minimization of $P_e = \sum_k P_e(i)/N$ by respectively the attacker and the decoder, .i.e.

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\sigma}_n} P_e, \quad \max_{\boldsymbol{\sigma}_n} \min_{\boldsymbol{\beta}} P_e. \tag{2}$$

The game has a pure (deterministic) equilibrium if the minimax solution equals the maximin one at a given BER value (called the value of the game) for some deterministic optimal values $\boldsymbol{\sigma}_n^*$ and $\boldsymbol{\beta}^*$. Then, the payoff function is said to have a saddle-point at $(\boldsymbol{\sigma}_n^*, \boldsymbol{\beta}^*)$. If this happens, the order in which the agents play the game is indifferent as neither the attacker nor the decoder want to deviate from the most conservative option marked by the saddle-point. Nevertheless, the order is relevant if there does not exist at least one saddle-point.

A crucial issue in our development are the constraints the embedder and the attacker must verify. A certain trade-off between the mathematical suitability of the Mean Square Error (MSE), that it is arguably inadequate for data hiding [6], and perceptual adequateness is achieved by an MSE-like condition imposed on each set of coefficients devoted to a particular information bit. For instance, the attacker constraints would read as

$$\frac{1}{L_i}\sum_{j\in\mathcal{S}_i} E\{|z_j - y_j|^2\} = \frac{1}{L_i}\sum_{j\in\mathcal{S}_i} \sigma_{n_j}^2 \leq D_c(i), \text{ for all } i \in \{1,\cdots,N\} \tag{3}$$

for some specified positive quantities $D_c(i)$, $i = 1, \cdots, N$. The problem with this constraint is that the attacker is supposed to know which coefficients are devoted to the same bit, what depends on the cryptographic key K.

Also, the perceptual masks can be directly mapped into a set of point-by-point constraints

$$E\{|y_i - x_i|^2\} = E\{w_i^2\} \leq \alpha_i^2, \text{ for all } i \in \mathcal{S}. \tag{4}$$

Even though it is a less flexible strategy than the previous one, it will be used to restrict the embedding power.

Finally, it is useful, mainly for comparison purposes, to define the *watermark-to-noise ratio* (WNR) as the ratio (in decibels) between the total energy devoted to the watermark and that devoted to the distortion, that is,

$$\text{WNR} \triangleq 10\log_{10}\left(\frac{\sum_{k\in\mathcal{S}} \text{E}\{w_k^2\}}{\sum_{k\in\mathcal{S}} \sigma_{n_k}^2}\right) \tag{5}$$

3 Spread-spectrum.

Given the assumptions of Section 2, spread-spectrum methods compute the watermark to be embedded as

$$w_j = g_K(x_j, \alpha_j, b_k) = b_k \alpha_j s_j, \text{ for all } j \in \mathcal{S}_k, k \in \{1, \cdots, N\} \tag{6}$$

where s_k is a pseudorandom sequence generated using a pseudonoise generator initialized to a state which depends on the value of K, with $E\{s_k\} = 0$ and $E\{s_k^2\} = 1$, so that (4) is satisfied. Here, we will assume the simplest distribution of this kind, that is, $s_k \in \{\pm 1\}$.

Before decoding, a transformation is applied to the received signal. The more widespread of these transformations is a simple linear projection onto one dimension:

$$r_i = \sum_{j\in\mathcal{S}_i} \beta_j s_j z_j, \; i \in \{1, \cdots, N\} \tag{7}$$

which is similar to the correlation receiver applied in spread-spectrum communications, but replacing $\boldsymbol{\alpha}$ with a more general weighting vector $\boldsymbol{\beta}$. It can be shown that, under some practical assumptions, the optimal decoder is equivalent to a bit-by-bit hard-decision maker with the threshold located at the origin [5]. Then, the output of the decoder, known the partition $\mathcal{T}$, is

$$\hat{b_i} = \mathrm{sign}\,(r_i|\,\mathcal{T}), i \in \{1, \cdots, N\} \tag{8}$$

In the spatial domain case the watermarked image $\boldsymbol{y}$ could undergo a linear filtering operation as a way of reducing the host-interference power at the decoder. This can be represented by means of a spatial-varying and noise independent filtering. Wiener filtering [7] is included in this category, since the host signal power usually is much greater than the noise power (at least if the attacked signal is to remain valuable), so Wiener filter's coefficients are not going to be modified in a significant way by the addition of noise. We can represent this situation by a $M \times M$ matrix that will be denoted by $\boldsymbol{H}$, so that the filtered host image would become $\boldsymbol{x}_f \triangleq \boldsymbol{Hx}$. As it was shown in [7], the observation vector $\boldsymbol{r}$ when $\mathcal{T}$ is known can be modeled as the output of an additive white Gaussian noise (AWGN) channel, $r_{i|\mathcal{T}} = a_{i|\mathcal{T}} b_i + u_{i|\mathcal{T}}$, $i \in \{1, \cdots, N\}$, where

$$a_{i|\mathcal{T}} = \sum_{k \in \mathcal{S}_i|\mathcal{T}} \beta_k h_{k,k} \alpha_k, \;\; i = 1, \cdots, N \tag{9}$$

and $u_{1|\mathcal{T}}, \cdots, u_{N|\mathcal{T}}$ are samples of an i.i.d. zero-mean Gaussian random process with variance

$$\sigma^2_{u_{i|\mathcal{T}}} = \sum_{k \in \mathcal{S}_i|\mathcal{T}} \beta_k^2 \left[x^2_{f_k} + \sum_{l=1}^{M} h^2_{k,l} \left(\alpha_l^2 + \sigma^2_{n_l} \right) - h^2_{k,k} \alpha_k^2 \right], \quad i = 1, \cdots, N \tag{10}$$

being M the length of the host signal. Since $\mathcal{T}$ is generated by K, we will assume the attacker does not know it. So he/she will try to maximize the probability of error considering the averaged channel, whose statistics for the case of uniform partitions are

$$a = \sum_{\forall \mathcal{T} \in \mathcal{U}} \mathrm{E}(r_i|\mathcal{T}) \Pr(\mathcal{T}) = \frac{1}{N} \sum_{k=2}^{M} \beta_k h_{k,k} \alpha_k \tag{11}$$

$$\begin{aligned} \sigma_u^2 &= \sum_{\forall \mathcal{T} \in \mathcal{U}} \mathrm{Var}(r_i|\mathcal{T}) \Pr(\mathcal{T}) + \sum_{\forall \mathcal{T} \in \mathcal{U}} \mathrm{E}^2(r_i|\mathcal{T}) \Pr(\mathcal{T}) - \left(\sum_{\forall \mathcal{T} \in \mathcal{U}} \mathrm{E}(r_i|\mathcal{T}) \Pr(\mathcal{T}) \right)^2 \\ &= \frac{1}{N} \sum_{k=1}^{M} \beta_k^2 \left[x^2_{f_k} + \sum_{l=1}^{M} h^2_{k,l} \left(\alpha_l^2 + \sigma^2_{n_l} \right) - h^2_{k,k} \alpha_k^2 \right] \\ &+ \frac{N-1}{N^2} \sum_{k=1}^{M} \beta_k^2 h^2_{k,k} \alpha_k^2 \end{aligned} \tag{12}$$

and since N typically will be large, $(N-1)/N^2$ can be substituted by $1/N$.

From (8), and recalling we are taking into account the averaged channel

$$\overline{P_e} = Q\left(\frac{a}{\sigma_u}\right) \tag{13}$$

with $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{\tau^2}{2}} d\tau$, so from the attacking point of view, the objective will be to maximize the partition-averaged signal to noise ratio given by

$$\overline{\mathrm{SNR}} \triangleq \frac{a}{\sigma_u} \tag{14}$$

while from the decoding point of view, the objective will be to maximize the per-pulse signal to noise ratio given to be

$$\mathrm{SNR}_i \triangleq \frac{a_{i|\mathcal{T}}}{\sigma_{u_{i|\mathcal{T}}}} \tag{15}$$

for all $i \in \{1, \cdots, N\}$, since the decoder knows the partition which is being used, so he/she knows the probability of error for this partition is

$$P_e = \frac{1}{N} \sum_{i=1}^{N} Q\left(\frac{a_{i|\mathcal{T}}}{\sigma_{u_{i|\mathcal{T}}}}\right) \tag{16}$$

Optimal Decoding Weights for a Known Attack Distribution. First, we will consider the case in which the attacking-noise distribution is known and determine the optimal decoding weights vector $\boldsymbol{\beta}^*$ that minimizes the BER in (16). Substituting (9-10) into (15) and inverting the result, we obtain the noise-to-signal ratio that the decoder should *minimize*:

$$\mathrm{NSR}_i = \frac{\sum_{j\in\mathcal{S}_i} \beta_j^2 \left[x_{f_j}^2 + \sum_{l=1}^{M} h_{j,l}^2 \left(\alpha_l^2 + \sigma_{n_l}^2\right) - h_{j,j}^2 \alpha_j^2\right]}{\left(\sum_{j\in\mathcal{S}_i} \beta_j h_{j,j} \alpha_j\right)^2}, \ \forall\, i = 1, \cdots, N \tag{17}$$

The problem can be solved in a general form to yield the following optimal weights

$$\beta_i^* = \frac{K h_{i,i} \alpha_i}{x_{f_i}^2 + \sum_{l=1}^{M} h_{i,l}^2 \left(\alpha_l^2 + \sigma_{n_l}^2\right) - h_{i,i}^2 \alpha_i^2}, \ i \in \mathcal{S} \tag{18}$$

with K any positive constant.

Optimal Attack for Known Decoding Weights. In the case that the attacker knows the decoding weights vector $\boldsymbol{\beta}$, his/her problem becomes that of maximizing the $\overline{\mathrm{NSR}}$ in (14) subject to an imperceptibility constraint. It can be proven that for a MSE distortion constraint the optimal attack would imply concentrating all the distortion in those coefficients with the largest values of $\tau_j = \sum_{k=1}^{M} \beta_k^2 h_{k,j}^2$. Note that this strategy will likely produce visible results (see Section 6) and clearly shows that constraining just the MSE may lead to impractical attacks.

Optimal Attack When the Decoder Follows the Optimum Strategy. Now, suppose that the decoder knows which distribution the attacker is using, so that he/she employs the optimal strategy as derived in Sect. 3. In this case, the best an attacker can do is to minimize (14) after replacing β_i with (18), while satisfying a certain distortion constraint. Therefore, making the assignments $p_j^2 = x_{f_j}^2 + \sum_{l=1}^{M} h_{j,l}^2 \left(\alpha_l^2 + \sigma_{n_l}^2\right) - h_{j,j}^2 \alpha_j^2$, $q_j = h_{j,j}\alpha_j$ and $t_j = 0$ the attacker has to minimize

$$\overline{\mathrm{SNR}} = \frac{\left(\sum_{k=1}^{M} \beta_k q_k\right)^2}{N\left[\sum_{k=1}^{M} \beta_k^2 p_k^2 + \beta_k^2 q_k^2\right]} = \frac{\left(\sum_{k=1}^{M} \frac{q_k^2}{p_k^2}\right)^2}{N\left[\sum_{k=1}^{M} \frac{q_k^2}{p_k^2} + \frac{q_k^4}{p_k^4}\right]} . \tag{19}$$

Since $p_j^2 \gg q_j^2$ we may neglect the second term in the denominator, so we can reformulate the problem as the minimization of

$$\varphi = \sum_{k=1}^{M} \frac{q_k^2}{p_k^2} = \sum_{k=1}^{M} \frac{h_{k,k}^2 \alpha_k^2}{m_k^2 + \sum_{l=1}^{M} h_{k,l}^2 \sigma_{n_l}^2} , \tag{20}$$

where $m_k = x_{f_k}^2 + \sum_{l=1}^{M} h_{k,l}^2 \alpha_l^2 - h_{k,k}^2 \alpha_k^2$. Unfortunately, a close look at (20) reveals that each particular noise sample exerts influence on several terms of the sum, thus making it difficult the interpretation of the solution. Aiming at producing meaningful results, for the remaining of this section we will make the simplification $\boldsymbol{H} = \mathrm{diag}(h_{1,1}, \cdots, h_{M,M})$ which is reasonable in many practical situations: as an example we have closely studied Wiener filtering and made the whole numerical optimization taking into account all the values of $h_{k,l}$ [8], [9]. The results are virtually the same as those we obtained with the proposed simplification. The explanation is based on the fact that the central element of the filter is much larger than the others, so the influence of the latter on the optimization is very small. So (20) becomes $\varphi = \sum_{k=1}^{M} \frac{\alpha_k^2}{\frac{x_{f_k}^2}{h_{k,k}^2} + \sigma_{n_k}^2}$ and (18) simplifies to $\beta_i = \frac{k\alpha_i h_{i,i}}{x_{f_i}^2 + h_{i,i}^2 \sigma_{n_i}^2}$. As in the previous section, the attack is constrained to meet a condition for the maximum allowed distortion introduced in the image, that is $D_c = \frac{1}{M}\sum_{j=1}^{M} \sigma_{n_j}^2$ and also it must verify $\sigma_{n_j}^2 \le L \cdot D_c$. This last condition tries to avoid the effect of assigning all the power to a few coefficients. One host image coefficient should not be assigned more power than the averaged power dedicated to each bit. In this case it can be shown that the optimal attacking distribution is

$$\sigma_{n_i}^{*\,2} = \min\left[\frac{D_c}{N}, \left(\xi\alpha_i - \frac{x_i^2}{h_{i,i}^2}\right)^+\right], \text{ for all } 1 \le i \le M \tag{21}$$

where $(x)^+ \triangleq \max\{x, 0\}$, and ξ is a suitably chosen parameter so that

$$\frac{1}{M}\sum_{i=1}^{M} \min\left[L \cdot D_c, \left(\xi\alpha_i - \frac{x_i^2}{h_{i,i}^2}\right)^+\right] = D_c \tag{22}$$

Although the analyzed problem is very different, this is quite similar to the expression obtained in [4] in which after getting the diagonalization by the KLT, the eigenvalues of the covariance matrix of the noise, and therefore the elements of this matrix, are

$$\sigma_{n_i}^{*\ 2} = \left(\xi_2 \alpha_i - \sigma_{x_i}^2\right)^+ \tag{23}$$

where $\sigma_{x_i}^2$ is the variance of x_i and ξ_2 a constant such that

$$\frac{1}{M}\sum_{i=1}^{M} \left(\xi_2 \alpha_i - \sigma_{x_i}^2\right)^+ = D_c \tag{24}$$

4 Distortion-Compensated Dither Modulation.

Informed embedding watermarking or, equivalently, quantization-based methods, are based on hiding information by constructing a set of vector quantizers $\boldsymbol{Q_b}(\cdot)$, each representing a different codeword $\boldsymbol{b}$. So, given a host vector $\boldsymbol{x}$ and an information codeword $\boldsymbol{b}$, the embedder constructs the watermarked vector $\boldsymbol{y}$ by simply quantizing $\boldsymbol{x}$ with $\boldsymbol{Q_b}(\cdot)$, i.e. $\boldsymbol{y} = \boldsymbol{Q_b}(\boldsymbol{x})$ [10]. We will only consider here one of the simplest implementations of quantization-based schemes, carried out by means of uniform dithered quantizers, which has been named Distortion-Compensated Dither Modulation (DC-DM).

In binary DC-DM the watermark is obtained as

$$w_j = g_K(x_j, \alpha_j, b_k) = \nu_k e_j, \text{ for all } j \in \mathcal{S}_k, k \in \{1, \cdots, N\} \tag{25}$$

i.e. the L_i-dimensional quantization error $\boldsymbol{e}_i \triangleq \boldsymbol{Q}_b(\boldsymbol{x}_i) - \boldsymbol{x}_i$ weighted by an optimizable constant ν_i, $0 < \nu_i \leq 1$, with $i = 1, \cdots, N$. Consequently, we will have

$$\boldsymbol{y}_i = \boldsymbol{Q}_b(\boldsymbol{x}_i) - (1 - \nu_i)\boldsymbol{e}_i, \ i = 1, \cdots, N \tag{26}$$

When $\nu_i = 1$, we have the uncompensated (i.e., pure DM) case.

The uniform quantizers $\boldsymbol{Q}_{-1}(\cdot)$ and $\boldsymbol{Q}_1(\cdot)$ are such that the corresponding centroids are the points in the lattices

$$\Lambda_{-1} = 2(\Delta_1 \mathbb{Z}, \cdots, \Delta_L \mathbb{Z})^T + \boldsymbol{d}_i \tag{27}$$

$$\Lambda_1 = 2(\Delta_1 \mathbb{Z}, \cdots, \Delta_L \mathbb{Z})^T + \boldsymbol{d}_i + (\Delta_1, \cdots, \Delta_L)^T \tag{28}$$

where $\boldsymbol{d}_i$ is an arbitrary (possibly key-dependent) vector. Since the presence of a known offset $\boldsymbol{d}_i$ in the lattices will not modify the results, we will suppose that $\boldsymbol{d}_i = \boldsymbol{0} \triangleq (0, \cdots, 0)^T$.

If the quantization step in each dimension is small enough, we can consider that the quantization error $\boldsymbol{e}_i$ in each dimension will be uniformly distributed in $[-\Delta_k, \Delta_k]$, being $2\Delta_k$ the quantization step. From (25), this in turn implies that the watermark is also uniformly distributed in a hyperrectangle. Thus, the embedding distortion in each dimension will be $D_{w_k} = \nu_i^2 \Delta_k^2 / 3$, for all $k \in \mathcal{S}_i$.

Decoding is implemented as

$$\hat{b}_i = \arg\min_{-1,1}\left\{\left(\boldsymbol{z}_j - \boldsymbol{Q}_{b_j}(\boldsymbol{z}_j)\right)^T \boldsymbol{B}_j \left(\boldsymbol{z}_j - \boldsymbol{Q}_{b_j}(\boldsymbol{z}_h)\right)\right\},\ i = 1,\cdots,N \quad (29)$$

where $\boldsymbol{B}_j \triangleq \text{diag}\left(\beta_{j1}/\Delta_{j1}^2,\cdots,\beta_{jL_j}/\Delta_{jL_j}^2\right)$ being the β_i some weights the decoder will use to improve decoding. Following the discussion on the decision regions made in [11], we can assume without loss of generality that a symbol $b_i = -1$ is sent, and that $\boldsymbol{x}_i$ is such that $Q_{-1}(\boldsymbol{x}_i) = \mathbf{0}$. Let $P_e(i)$ denote the bit error probability conditioned to the transmission of the i-th bit, $i = 1,\cdots,N$. Then, assuming that all the bits sent are equiprobable, we can write

$$P_e = \frac{1}{N}\sum_{i=1}^{N} P_e(i) \quad (30)$$

so we will be interested in computing the bit error probability for the i-th bit. However, for the remaining of this section and for the sake of notational simplicity we will drop the subindex i whenever there is no possible confusion.

Let $\boldsymbol{u} \triangleq \boldsymbol{n} - (1-\nu)\boldsymbol{e}$, then $\boldsymbol{z} = \boldsymbol{u}$. Recalling that $\boldsymbol{e}$ has independent components, $e_k \sim U(-\Delta_k, \Delta_k)$, it follows that the random vector $\boldsymbol{u}$ will also have independent components, each having pdf

$$f_{u_k}(u_k) = \begin{cases} f_{n_k}(u_k) * \frac{1}{(1-\nu)} f_{e_k}(u_k/(1-\nu)), & 0 < \nu < 1 \\ f_{n_k}(u_k), & \nu = 1 \end{cases} \quad (31)$$

being $f_{n_k}(\cdot)$ and $f_{e_k}(\cdot)$ the marginal pdf's of respectively the noise and the quantization error components of each dimension. We will find useful to define an auxiliary variable $v_j \triangleq u_j/\Delta_j$

4.1 Approximate Computation of the Bit Error Probability

If the noise pdf is symmetric with respect to the coordinate planes, then both $\boldsymbol{u}$ and $\boldsymbol{v}$ will inherit this symmetry. In that case, we can concentrate our analysis in the positive orthant $\mathcal{O}$, so we can upperbound $P_e(i)$ as [11]

$$\begin{aligned} P_e(i) \le P_s(i) &\triangleq P\left\{\boldsymbol{v}_i^T \boldsymbol{B}_i \boldsymbol{v}_i > (\boldsymbol{v}_i - (1,\cdots,1)^T)^T \boldsymbol{B}_i (\boldsymbol{v}_i - (1,\cdots,1)^T) | \boldsymbol{v}_i \in \mathcal{O}\right\} \\ &= P\left\{\sum_{k\in\mathcal{S}_i} \beta_k v'_k > \frac{1}{2}\sum_{k\in\mathcal{S}_i}\beta_k\right\} \end{aligned} \quad (32)$$

being $v'_k \triangleq |v_k|$ (since we are considering only $\mathcal{O}$), with pdf given by

$$f_{v'_k}(v'_k) \triangleq \begin{cases} 2\Delta_k f_{u_k}(v'_k \Delta_k), & u'_k > 0 \\ 0, & \text{otherwise} \end{cases},\ k \in \mathcal{S}_i \quad (33)$$

If we define

$$t_i = \sum_{k\in\mathcal{S}_i} \beta_k v'_k,\ i = 1,\cdots,N \quad (34)$$

then the pdf of the random variable t_i will be the convolution of L independent random variables with pdf $f_{v'_k}(v'_k/\beta_k)/\beta_k$, for all $k \in \mathcal{S}_i$, and from (32) $P_s(i)$ can be obtained by integrating its tail from $\sum_{k\in\mathcal{S}_i} \beta_k/2$. Moreover, by virtue of the central limit theorem (CLT), as $L \to \infty$, $f_{t_i}(t_i)$ will tend to a normal distribution. Then, for very large L, t_i can be approximated by a Gaussian, so

$$P_s(i) \approx Q\left(\frac{\sum_{k\in\mathcal{S}_i} \beta_k/2 - \sum_{k\in\mathcal{S}_i} \beta_k \mathrm{E}\{v'_k\}}{\sqrt{\sum_{k\in\mathcal{S}_i} \beta_k^2 \mathrm{Var}\{v'_k\}}}\right) \tag{35}$$

for all $i = 1, \cdots, N$. As discussed in [6], the CLT-based approximation applied to highly-skewed pdf's results in a very slow convergence. In any case, it may be reasonable to use (35) as the functional to be maximized (minimized) by the attacker (decoder). The exact value of P_e can be obtained [12], but it leads to a cumbersome expression that hinders the solution.

Optimal Decoding Weights for a Known Attack Distribution Recalling that the $Q(\cdot)$ function is monotonically decreasing, it follows that $P_s(i)$ is minimized when its argument—that is, the signal to noise ratio SNR_i— is maximized. Then, the optimal weights can be found by differentiating

$$\begin{aligned}\frac{\partial \mathrm{SNR}_i}{\partial \beta_j} &= \left(\frac{1}{2} - \mathrm{E}\{v'_j\}\right)\left(\sqrt{\sum_{k\in\mathcal{S}} \beta_k^2 \mathrm{Var}\{v'_k\}}\right) \\ &- \left(\frac{1}{2}\sum_{k\in\mathcal{S}} \beta_k - \sum_{k\in\mathcal{S}} \beta_k \mathrm{E}\{v'_k\}\right)\left(\beta_j \ \mathrm{Var}\{v'_j\}\right)\left(\sum_{k} \beta_k^2 \mathrm{Var}\{v'_k\}\right)^{-1/2}\end{aligned} \tag{36}$$

and setting to zero, which yields

$$\beta_j^* = \frac{\left(\frac{1}{2} - \mathrm{E}\{v'_j\}\right)}{\mathrm{Var}\{v'_j\}} \cdot \frac{\left(\sum_{k\in\mathcal{S}_i} \beta_k^2 \mathrm{Var}\{v'_k\}\right)}{\left(\sum_{k\in\mathcal{S}_i} \beta_k \left(\frac{1}{2} - \mathrm{E}\{v'_k\}\right)\right)}, \ j \in \mathcal{S}_i \tag{37}$$

The second factor in (37) is an irrelevant constant, since $\boldsymbol{\beta}^*$ can be scaled without any impact on performance. It is worth mentioning that β_j^* in (37) can take negative values, which are due to large noise values in certain dimensions. The result here obtained implies that for those dimensions it is profitable to *subtract* the corresponding square distance terms in (29). This is reasonable if one thinks, for instance, of a noise pdf uniform in $[-3\Delta_j/2, 3\Delta_j/2]$. When the modular transformation is applied and the absolute value is taken, the resulting pdf has a mean with value 1.75/3. If this held for all dimensions and the decoder were not using β^*, he/she would get $P_e > 0.5$.

Optimal Attack for Known Decoding Weights. In this case, it is easy to verify that the procedure of building the Lagrangian and equating its derivatives to zero leads to a system of nonlinear equations, which requires numerical methods for solving it. Since this does not shed any light on the strategy that the attacker should follow, we will not develop it here.

Optimal Attack When the Decoder Follows the Optimal Strategy. Now, the question is to decide what is the optimal attack when the decoder is using the optimum $\boldsymbol{\beta}$. This problem is quite difficult to solve even in the simplest cases. In fact, we will concentrate on the case where the attacker knows $\mathcal{T}$, the attack consists on uniform noise in each dimension with distribution $[-\eta_k \Delta_k, \eta_k \Delta_k]$ ($[-\eta_k, \eta_k]$ once it has been normalized by Δ_k) for all $k \in \mathcal{S}$, and there is no distortion compensation (pure DM case). So, replacing $\boldsymbol{\beta}$ by its optimal value in the argument of (35), the attacker has to minimize

$$\mathrm{SNR}_i = \sum_{k \in \mathcal{S}_i} \frac{\left(\frac{1}{2} - \mathrm{E}\{v'_k\}\right)^2}{\mathrm{Var}\{v'_k\}} = \sum_{k \in \mathcal{S}_i} \frac{3\,(1-\eta_k)^2}{\eta_k^2},\ i = 1, \cdots, N \tag{38}$$

constrained to $\sum_{k \in \mathcal{S}_i} \frac{\Delta_k^2 \eta_k^2}{3} \leq D_c(i)$, $i = 1, \cdots, N$.

Using the Lagrange multipliers technique, we may proceed to differentiate the unconstrained functional with respect to η_i and equate to zero to get

$$\frac{(\eta_i - 1)\,\eta_i^2 - (\eta_i - 1)^2\,\eta_i}{\eta_i^4} + \lambda_k \eta_i \Delta_i^2 = 0, \text{ for all } i \in \mathcal{S}_k, \quad k = 1, \cdots, N \tag{39}$$

So even in this simple case, the following fourth order equation has to be solved for every η_i, $i \in \mathcal{S}_k$,

$$\lambda_k \eta_i^4 \Delta_i^2 + \eta_i - 1 = 0 \tag{40}$$

Equation (40) gives a hint on the complexity of the problem for DC-DM, because in such case the noise due to distortion compensation is combined with the additive noise from the attacker.

5 Quantized Projection.

In the Quantized Projection method [6], the set of samples $\mathcal{S}_i$ assigned to one bit b_i, is projected onto one dimension obtaining the variable r_{x_i}, which is later quantized with a uniform scalar quantizer with step $2\Delta_i$. Hence, centroids of the decision cells associated to $\hat{b}_i = 1$ and $\hat{b}_i = -1$ are respectively given by the unidimensional lattices Λ_{-1} and Λ_1 in (27-28), with $d_i = -\Delta_i/2$ due to symmetry considerations on the pdf of the host signal projection. The projection of $\boldsymbol{y}_i$ is written as

$$r_{y_i} = \sum_{k \in \mathcal{S}_i} y_k s_k \beta_k, \quad i \in \{1, \cdots, N\} \tag{41}$$

where $\boldsymbol{s}$ is a key-dependent pseudorandom vector verifying $E\{s_k\} = 0$ and $E\{s_k^2\} = 1$, and r_{y_i} must be a centroid belonging to one of the two former lattices depending on the transmitted bit.

Following the procedure in [6] it is straightforward to show that $P_e(i)$ that can be approximated by

$$P_e(i) \approx 2Q\left(\frac{\Delta_i}{2\sigma_{r_{n_i}}}\right) = 2Q\left(\frac{\tau_i\left(\sum_{j\in\mathcal{S}_i}\alpha_j\beta_j\right)}{2\sqrt{\sum_{j\in\mathcal{S}_i}\sigma_{n_j}^2\beta_j^2}}\right), \quad i \in \{1,\cdots,N\} \tag{42}$$

where $\tau_i \in [\sqrt{3}, 2]$ is a function that depends on the ratio $\frac{\sigma_{r_{x_i}}}{\Delta_i}$, and consequently also on $\boldsymbol{\beta}$, although in a weaker way. Therefore, as $Q(\cdot)$ is monotonic, we have to maximize (minimize) the argument of this function in (42).

Optimal Decoding Weights for a Known Attack Distribution. If we assume that τ_i does not depend on $\boldsymbol{\beta}$ (in fact, there is only a weak dependence), it can be proven that the optimal weights becomes

$$\beta_j^* = \frac{K\alpha_j}{\sigma_{n_j}^2}, \text{ for all } j \in \mathcal{S}_i, \quad i \in \{1,\cdots,N\} \tag{43}$$

being K any constant.

Optimal Attack for Known Decoding Weights. In this case, we are in the same situation as in Section 3, so all the considerations made there are perfectly valid here. All the attacking power will be concentrated in those coefficients with the largest values of β_k^2.

Optimal Attack When the Decoder Follows the Optimum Strategy. If we follow a strategy similar to the one described in Section 3, assuming the attacker does not know the actual partition, we will have a expression like (19), but now with $p_j = \sigma_{n_j}$, $q_j = \alpha_j$, $t_j = 0$. In this case it is not so clear that $p_j \gg q_j$. In fact, for WNR>0, $q_j > p_j$. Therefore the same simplification cannot be done and the problem requires to be solved by numerical optimization. To show empirical results, we have studied also the case when the attacker knows the partition, and the solution is $\sigma_{n_j}^2 = \xi_i\alpha_j$, for all $j \in \mathcal{S}_i, i \in \{1,\cdots,N\}$, where $\xi_i = L\cdot D_c(i)/\left(\sum_{j\in\mathcal{S}_i}\alpha_j\right), i \in \{1,\cdots,N\}$.

6 Experimental Results

We show next the results of applying the strategies derived along previous sections to real data. In the figures that follow, symbols refer to empirical (MonteCarlo) simulations, while lines show theoretical results. Empirical data come from the gray-scale *Lena* image (256×256), for which the spatial perceptual mask $\boldsymbol{\alpha}$ has been computed using the method detailed in [7], except for the DC-DM scheme where, for illustrative purposes, we have chosen to work in the DCT domain, with a perceptual mask that has been obtained as in [13].

First, in Figure 2 the P_e's resulting when different strategies are considered for spread-spectrum (Section 3) are shown. Watermarking has been performed in the spatial domain with Wiener filtering prior to decoding and 50 pixels per bit ($L = 50$) have been used. Three cases are analyzed: first, the noise variance $\sigma^2_{n_j}$ at each sample is made proportional to α^2_j and $\boldsymbol{\beta} = C\boldsymbol{\alpha}$, with C any positive constant; second, the attack is the same as in the previous case but the optimal decoding weights β^* are employed; finally, the plot labeled as "worst attack" refers to the case where the attacker follows his/her optimal strategy knowing that the decoder also uses the optimal decoding weights. In all cases, the theoretical results lie close to the empirical ones, although for those where the optimal $\boldsymbol{\beta}^*$ is used the difference is larger.

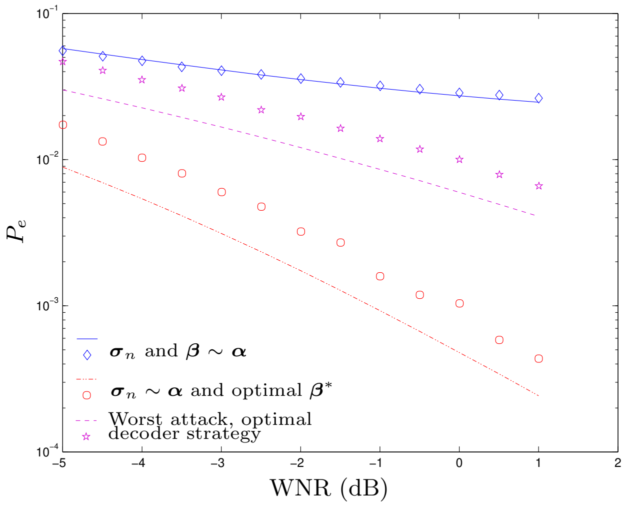

Fig. 2. BER versus WNR for spread-spectrum (L=50) showing three different attacking/decoding strategies.

The cases depicted in Figures 3 correspond to the binary DC-DM method where, as mentioned, watermarking is done in the DCT domain. The distortion compensating parameter ν is set to 0.7 (see Section 4). In order to establish a meaningful case for the experiments, we have selected uniform noise proportional to the quantization step that results when a JPEG quality factor of 80 is selected. For both Figures we have set $L = 10$ (notice that now less samples are needed when compared to spread-spectrum in order to achieve similar BER's for identical WNR's). Two scenarios are depicted in Figure 3: in the first case, each sample, say the j-th, is scaled by Δ_j at the decoder but no further weighting (i.e., $\beta_j = 1$) is considered; in the second plot, the optimal $\boldsymbol{\beta}^*$ that follows from applying the results from Section 4.1 is used. The fact that in this second case the empirical results lie above the theoretical ones may be surprising at first sight since the latter have been obtained with Eq. (35) which was said to be an upper bound to P_e. The explanation to this phenomenon is that in such case some β_j^* take negative values which affect the validity of the CLT approximation (see Section 4.1). Note that as we have less noise (i.e., the WNR increases), it becomes more unlikely to have negative values of $\boldsymbol{\beta}^*$ (since the average value of v_k' decreases), so the theoretical curve and the empirical results get much closer.

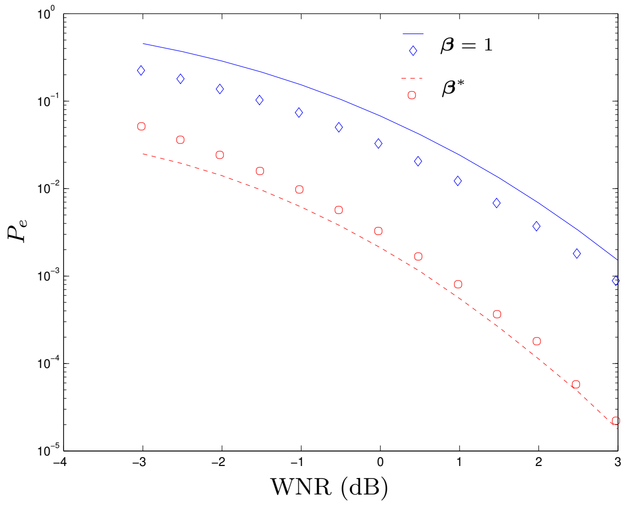

Fig. 3. BER versus WNR for DC-DM (L=10, ν = 0.7) and JPEG noise when no weights are used, and for the optimal weighting.

Finally figure 4 shows a similar comparison for the case considered in Section 5. The decoding weights are set so that $\boldsymbol{\beta} = \boldsymbol{\alpha}$, and the optimal attack for this case is compared to an attack consisting in using noise variances $\sigma_{n_k}^2$ proportional to α_k.

7 Conclusions and Future Research

As a conclusion of this paper, one aspect that clearly requires further study is that of distortion constraints and their relationship with optimal strategies. For instance, as it can be checked in Sects. 3 and 5, the optimal attack will end up in a visible attacked image. Whether this image keeps some of its original value is a moot question that largely depends on the final application scenario.

Related to this, we can think of the problem where the embedder has an active role (as we have already done in QP), and does not just generate the

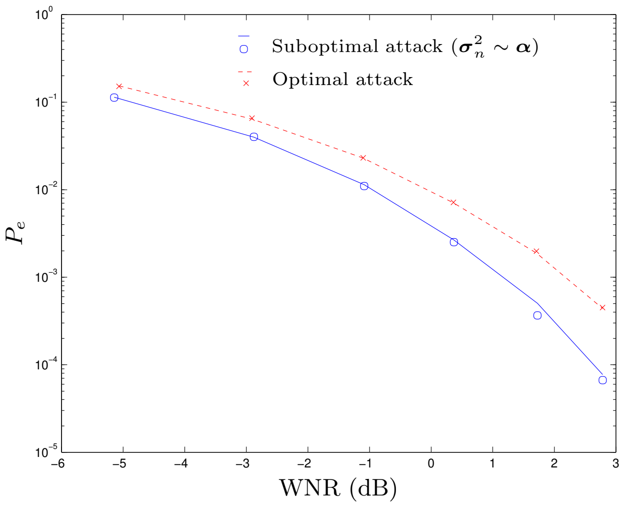

Fig. 4. BER versus WNR corresponding to the optimal and suboptimal attacks for QP when the attacker knows the decoder weights (L=10)

watermarked image independently of the possible attacks. In any way, the distortion introduced by the embedder has to be extremely small; in that regard, we can assume that the attacker has always more freedom to making it difficult the decoding process.

References

1. A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 48, pp. 1639–1667, June 2002.
2. J. J. Eggers and B. Girod, *Informed Watermarking.* Kluwer Academic Publishers, 2002.
3. P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. on Information Theory*, 2003.
4. P. Moulin and A. Ivanovic, "The zero-rate spread-spectrum watermarking game," *IEEE Trans. on Signal Processing*, 2003.
5. J. R. Hernández and F. Pérez-González, "Statistical analysis of watermarking schemes for copyright protection of images," *Proceedings of the IEEE*, vol. 87, pp. 1142–1166, July 1999. Special Issue on Identification and Protection of Multimedia Information.
6. 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 Trans. on Signal Processing*, 2003. Special Issue "Signal Processing for Data Hiding in Digital Media & Secure Content Delivery".
7. J. R. Hernández, F. Pérez-González, J. M. Rodríguez, and G. Nieto, "Performance analysis of a 2D-multipulse amplitude modulation scheme for data hiding and

watermarking of still images," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 510–524, May 1998.
8. M. E. Vázquez-Méndez, *Análisis y control óptimo de problemas relacionados con la dispersión de contaminantes*. PhD thesis, Universidade de Santiago de Compostela, 1999.
9. J. Herskovits, "Feasible direction interior-point technique for nonlinear optimization," *Journal of optimization theory and applications*, 1998.
10. B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory*, vol. 47, pp. 1423–1443, May 2001.
11. F. Pérez-González and F. Balado, "Nothing but a kiss: A novel and accurate approach to assessing the performance of multidimensional distortion-compensated dither modulation," in *Proc. of the 5th International Workshop on Information Hiding*, Lecture Notes in Computer Science, (Noorwijkerhout, The Netherlands), Springer-Verlag, October 2002.
12. F. Pérez-González, P. Comesaña, and F. Balado, "Dither-modulation data hiding with distortion-compensation: Exact performance analysis and an improved detector for jpeg attacks," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, (Barcelona, Spain), September 2003.
13. 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. on Image Processing*, vol. 9, pp. 55–68, January 2000.