OPTIMAL STRATEGIES FOR SPREAD-SPECTRUM AND QUANTIZED-PROJECTION IMAGE DATA HIDING 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
email: pcomesan@gts.tsc.uvigo.es, fperez@tsc.uvigo.es, fiz@tsc.uvigo.es

ABSTRACT

We analyze spread-spectrum and quantization projection data hiding methods from a game-theoretic point of view, using the bit error rate (BER) as the payoff, and assuming that the embedder simply follows point-by-point constraints given by a perceptual mask, whereas for the attacker an MSE-like constraint is imposed. The optimal attacking and decoding strategies are obtained by making use of a theorem that in addition states that those strategies constitute an equilibrium of the game. Experimental results supporting our analyses are also shown.

1. INTRODUCTION

Although the mere existence of a game played by the embedder and the attacker was recognized since the very inception of watermarking and data hiding, it was not until very recently that this idea was formalized by Moulin et al. [1], [2] who introduced the concept of data hiding games and analyzed them from an information theoretic point of view. This formulation allows to derive optimal strategies which are then of use for establishing the true limits of data hiding. Some other researchers have also dealt with game-theoretic aspects of data hiding capacities [3], [4]; however payoffs other than channel capacity are also possible in the data hiding game, as already suggested in [1] and developed in [5]. 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 rationale behind this choice is that even though capacity measures are of great importance when establishing theoretical bounds, practical data hiding algorithms require strategies which may be different from those afforded by capacity considerations. In fact, as we will later show, different algorithms represent different games and, consequently, the optimal playing strategies also differ.

Three agents generally play the data hiding game: embedder, attacker and decoder. Each one has different objectives and constraints that frequently lead to colliding interests which have been already discussed in depth in [1]. As it is widely recognized, the most severe restriction faced by the embedder is that of invisibility. In our case, we will assume that this is driven by a perceptual mask $\boldsymbol{\alpha}$ that indicates the maximum allowed watermark energy that produces the least noticeable modification of the corresponding sample of the host image. If $\boldsymbol{x}$ denotes the samples (or coefficients in a transform domain) of the host image, arranged for convenience in vector form, $\boldsymbol{w}$ is the watermark and $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}$ denotes the watermarked image, then we will write the invisibility restriction as the following set of point-by-point constraints

$$E\{|y_i - x_i|^2\} = E\{w_i^2\} \leq \alpha_i^2, \text{ for all } i \in \mathcal{S} \qquad (1)$$

with $\mathcal{S}$ the set of pixels (or coefficients) devoted to data-hiding purposes. Now, it can be seen that (1) allows very little flexibility in choosing the embedder's strategy: except for rare cases, the optimum will be achieved when all the watermark coefficients take their extremal values in (1). For this reason, we will leave the embedder out of the game, although it must be stressed that since the characteristics of the human visual system (HVS) exploited here do not sweep all known properties, it would be plausible to let the embedder play with additional degrees of freedom, a consideration that will be left for future research.

The main purpose of this paper is to obtain the optimal strategies for two classes of data hiding methods, namely, spread-spectrum and quantized-projection schemes. We have developed closed-form expressions for the bit error probability which are then used as cost functions for deriving optimal tactics for the decoder and the attacker. To the authors' knowledge, the closest works to ours are those of Eggers and Girod in [4] and Moulin et al. in [5], compared to which the two main differences are: 1) the game payoff, which is channel capacity in [4] (although specifically optimized for each method) and probability of correct detection (zero-rate spread-spectrum scheme) in [5]; 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.

2. PRELIMINARIES

We will follow the customary scheme [6] for dividing the set $\mathcal{S}$ into N non-overlapping sets $\mathcal{S}_i$, $i = 1, \cdots, N$, each

Work partially funded by the *Xunta de Galicia* under projects PGIDT01 PX132204PM and PGIDT02 PXIC32205PN, and the CYCIT project AMULET, reference TIC2001-3697-C03-01.

of size L, through a key-driven pseudorandom permutation. Therefore, a total of $M \triangleq N \cdot L$ samples are employed. Each set $\mathcal{S}_i$ is devoted to conveying a particular bit $b_i \in \{\pm 1\}$. Moreover, 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 $\boldsymbol{x}$. 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 $\sigma^2_{n_i}$, $i \in \mathcal{S}$.

Another working hypothesis is that the vector of perceptual masks $\boldsymbol{\alpha}$ is available not only to the embedder, but also to the attacker and to the decoder. The decoder uses a certain decoding function parameterized by some weights vector $\boldsymbol{\beta}$ to produce the decoded vector $\hat{\boldsymbol{b}}$. Then, the BER for the i-th bit is just $P_e(i) = P\{\hat{b_i} \neq b_i\}$, and the games consist in the successive maximization/ minimization of $P_e = \sum_k P_e(i)/N$ by respectively the attacker and the decoder and viceversa, i.e. $\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\sigma}_n} P_e$, $\max_{\boldsymbol{\sigma}_n} \min_{\boldsymbol{\beta}} P_e$.

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. If there does not exist a saddle-point, the playing order is relevant and the solution to the maximin (minimax) problems allows to establish upper (lower) bounds to the BER performance. However, as we will see, our problems do admit an equilibrium.

Regarding attacks, as we said, they are limited to additive noise; moreover, Mean Square Error (MSE) constraints will be taken into account. As noted in [6], the main drawback of MSE is that unacceptably high local distortions are not ruled out, since they can be globally compensated. A certain trade-off between mathematical suitability and perceptual adequateness is achieved by an MSE-like condition imposed on each set of coefficients devoted to a particular information bit. The attacker constraints then read as

$$\frac{1}{L}\sum_{j\in\mathcal{S}_i} E\{|z_j - y_j|^2\} = \frac{1}{L}\sum_{j\in\mathcal{S}_i} \sigma^2_{n_j} \leq D_c(i), \qquad (2)$$

for some specified positive quantities $D_c(i)$, and for all $i = 1, \cdots, N$. Note that this obviously assumes that the attacker knows the partitions. Although the more general case in which the attacker does not have access to the partitions is more involved and for clarity is not pursued here, it can be shown [7] that the *form* of the solution is essentially the same.

For comparison purposes it is useful 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,

$$\text{WNR} \triangleq 10\log_{10}\left(\frac{\sum_{j\in\mathcal{S}} \mathrm{E}\{w_j^2\}}{\sum_{j\in\mathcal{S}} \sigma^2_{n_j}}\right) \qquad (3)$$

Finally, we state a Theorem which constitutes the basis for deriving optimal attacking and decoding strategies. The proof is omitted due to its length and can be found elsewhere [7]. First, we need some definitions: let $\mathcal{P}$ denote the N-dimensional ball centered at the origin and with radius R, whereas $\mathcal{J}$ is any set of integer indices with cardinality N. Also, let us introduce the function $\varphi : \mathcal{P} \times \mathbb{R}^N \to \mathbb{R}^+$ defined as

$$\varphi(\boldsymbol{p}, \boldsymbol{\beta}) \triangleq \frac{\sum_{j\in\mathcal{J}} \beta_j^2 (t_j^2 + p_j^2)}{\left(\sum_{j\in\mathcal{J}} \beta_j q_j\right)^2} \qquad (4)$$

with $\boldsymbol{q}, \boldsymbol{t}$ arbitrary vectors in $\mathbb{R}^N$. Also let $(x)^+ \triangleq \max\{x, 0\}$.

Theorem 1 *The vectors $\boldsymbol{p}^*$ and $\boldsymbol{\beta}^*$ with components*

$$(p_j^*)^2 = (\xi q_j - t_j^2)^+, \ \textit{for all } j \in \mathcal{J} \qquad (5)$$

$$\beta_j^* = \begin{cases} \frac{K_2 q_j}{t_j^2}, & \textit{if } j \in \mathcal{J}_0 \\ \frac{K_2}{\xi}, & \textit{otherwise} \end{cases} \qquad (6)$$

where the constant $\xi \in \mathbb{R}$ is the solution to the equation $\sum_{j\in\mathcal{J}} (\xi q_j - t_j^2)^+ = R^2$, K_2 is a nonzero real constant, and $\mathcal{J}_0 \subset \mathcal{J}$ is the set of indices for which the right hand side of (5) is negative; satisfy

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{p}} \varphi(\boldsymbol{p}, \boldsymbol{\beta}) = \varphi(\boldsymbol{p}^*, \boldsymbol{\beta}^*) = \max_{\boldsymbol{p}} \min_{\boldsymbol{\beta}} \varphi(\boldsymbol{p}, \boldsymbol{\beta}). \qquad (7)$$

3. STRATEGIES FOR SPREAD-SPECTRUM.

Given the assumptions in the previous sections, spread-spectrum methods compute the watermark to be embedded as $w_j = b_i \alpha_j s_j$, for all $j \in \mathcal{S}_i$, $i \in \{1, \cdots, N\}$, where s_j is a key-dependent pseudorandom sequence satisfying $E\{s_j\} = 0$ and $E\{s_j^2\} = 1$, so that (1) holds. Here, we will assume the simplest distribution, that is, $s_j \in \{\pm 1\}$. As it is well-known [8], the simplest receiver is based on the cross-correlating decoder which constructs the set of statistics

$$r_i = \sum_{j\in\mathcal{S}_i} \beta_j s_j z_j, \ i \in \{1, \cdots, N\} \qquad (8)$$

cascaded with a bit-by-bit hard decisor, i.e., $\hat{b_i} = \text{sign}(r_i)$, $i \in \{1, \cdots, N\}$. Note that the main difference with the decoder considered in [8] is that the vector $\boldsymbol{\alpha}$ has been replaced by a more general vector $\boldsymbol{\beta}$ suitable for a proper optimization.

In the case that the watermarked image $\boldsymbol{y}$ has undergone a linear filtering operation, which we suposse invariant with the noise, as a way of reducing the host-interference power at the decoder, 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 $\boldsymbol{x}_f \triangleq \boldsymbol{H}\boldsymbol{x}$. As it was shown in [8], the observation vector $\boldsymbol{r}$ can be now modeled as the output of an additive white Gaussian noise (AWGN) channel, $r_i = a_i b_i + u_i$, $i \in \{1, \cdots, N\}$, where

$$a_i = \sum_{k \in \mathcal{S}_i} \beta_k h_{k,k} \alpha_k,$$

and $u_1, \cdots, u_N$ are samples of an i.i.d. zero-mean Gaussian random process with variance

$$\sigma^2_{u_i} = \sum_{j \in \mathcal{S}_i} \beta_j^2 \left[x^2_{f_j} + \sum_{k=1}^{M} h^2_{j,k} \left(\alpha_k^2 + \sigma^2_{n_k} \right) - h^2_{j,j} \alpha_j^2 \right].$$

Recalling that the information bits are assumed to be equiprobable and that we are using a hard decisor, we can write

$$P_e = \frac{1}{N} \sum_{i=1}^{N} Q(a_i / \sigma_{u_i}) \tag{9}$$

with $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{\tau^2}{2}} d\tau$. Aiming at giving easily interpretable results, for the remaining of this section we make the simplification $\boldsymbol{H} = \mathrm{diag}(h_{1,1}, \cdots, h_{M,M})$, so from the attacking/decoding point of view, the objective will be to minimize/maximize the arguments of each Q-function adding up in (9). The exact derivation of results without this simplification is a current issue of work [7]. Then, we can resort to Theorem 1 where $\mathcal{J} = \mathcal{S}_i$; $p_j = \sigma_{n_j}$; $q_j = \alpha_j$ and $t_j = x_{f_j}/h_{j,j}$ to show that the attack at the equilibrium of the game with distortion constraints as in (2), for each $j \in \mathcal{S}_i$, $i = 1, \cdots, N$, is given by

$$\sigma^{*\,2}_{n_j} = \left(\xi_i \alpha_j - \frac{x^2_{f_j}}{h^2_{j,j}} \right)^+, \tag{10}$$

where ξ_i is a suitably chosen parameter so that $\frac{1}{L} \sum_{j \in \mathcal{S}_i} (\xi_i \alpha_j - \frac{x^2_{f_j}}{h^2_{j,j}})^+ = D_c^2(i)$ for all $i = 1, \cdots, N$. If $\mathcal{S}_{0,i} \subset \mathcal{S}_i$, $i = 1, \cdots, N$, denotes the set of indices for which the right hand side of Eq. (10) is zero, then the optimal decoding strategy is

$$\beta_j^* = \begin{cases} \frac{K \alpha_j h^2_{j,j}}{x^2_{f_j}}, & \text{if } j \in \mathcal{S}_{0,i} \\ \frac{K}{\xi_i}, & \text{if } j \in \mathcal{S}_i \backslash \mathcal{S}_{0,i} \end{cases} \tag{11}$$

4. STRATEGIES FOR QUANTIZED PROJECTION

In the Quantized Projection (QP) method [6], the set of samples $\mathcal{S}_i$ assigned to one bit b_i, is projected by the embedder onto one dimension obtaining a variable r_{x_i}, which is later quantized with a uniform scalar quantizer with step $2\Delta_i$ so the centroids of the decision cells associated to $\hat{b}_i = 1$ and $\hat{b}_i = -1$ are respectively given by the unidimensional lattices $\Lambda_{+1} \triangleq 2\Delta_i \mathbb{Z} - \Delta_i/2$ and $\Lambda_{-1} \triangleq 2\Delta_i \mathbb{Z} + \Delta_i/2$. The linear projection function presented in [6] can be generalized so as to take into account the possibility of weighting the various dimensions. Thus, in this more general way, the projection can be constructed as

$$r_{y_i} = \sum_{j \in \mathcal{S}_i} y_j s_j \beta_j, \;\; i \in \{1, \cdots, N\} \tag{12}$$

with s_j having identical characteristics as in the previous section. In fact, a similar definition to (12) applies to r_{x_i} and r_{w_i}, i.e., the respective host image and projected watermark. Moreover, $r_{y_i} = r_{x_i} + r_{w_i}$. The embedder must select the watermark samples w_j, $j \in \mathcal{S}_i$, so that r_{y_i} in (12) effectively belongs to the desired lattice. As we discussed in the Introduction, it is reasonable to choose $w_j, j \in \mathcal{S}_i$, proportional to α_j so $w_j = \rho_i \alpha_j s_j$, for all $j \in \mathcal{S}_i$, and $\rho_i = r_{w_i} / (\sum_{j \in \mathcal{S}_i} \alpha_j \beta_j)$. In the decoder, the image $\boldsymbol{z}$ is projected similarly to (12) to obtain r_{z_i}, which is then quantized to yield $\hat{b}_i$, $i = 1, \cdots, N$.

A performance analysis for this data hiding scheme can be adapted from that in [6] to show that the probability of error $P_e(i)$ for the i-th bit can be approximated by

$$P_e(i) \approx 2Q(\Delta_i / 2\sigma_{r_{n_i}}), \;\; i \in \{1, \cdots, N\} \tag{13}$$

with $\sigma^2_{r_{n_i}}$ the variance of the projected noise.

Then, considering the monotonicity of the Q-function, we have that the functional that the decoder (attacker) should maximize (minimize) is

$$\frac{\Delta_i^2}{4\sigma^2_{r_{ni}}} = \frac{\tau_i^2 \left(\sum_{j \in \mathcal{S}_i} \alpha_j \beta_j \right)^2}{4 \sum_{j \in \mathcal{S}_i} \sigma^2_{n_j} \beta_j^2} \tag{14}$$

where $\tau_i \in [\sqrt{3}, 2]$ a parameter that weakly depends on $\boldsymbol{\beta}_i$, so it can be disregarded in the optimization.

Now, it is possible to apply Theorem 1 to (14) with $\mathcal{J} = \mathcal{S}_i$; $p_j = \sigma_{n_j}$; $q_j = \alpha_j$; $t_j = 0$, to conclude that the equilibrium of the game is achieved when $\sigma^{*\,2}_{n_j} = K_2 \alpha_j$, for any nonnegative constant K_2, and $\beta_j^* = \text{constant}$.

5. 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 [8]. First, in Figure 1 the P_e's resulting when different strategies are considered for spread-spectrum (Section 3) are shown. Wiener filtering prior to decoding and 50 pixels per bit (i.e., $L = 50$) have been used. Three cases are analyzed: a) the noise variance $\sigma^2_{n_j}$ at each sample is made proportional to α_j^2, and $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\alpha}$, that is, the classical cross-correlating decoder; b) an attack as in (a) but the optimal decoding weights $\boldsymbol{\beta}$ for this attack are employed (shown in [7]); c) the plot labeled as "saddle-point" corresponds to the equilibrium solution. In all cases,

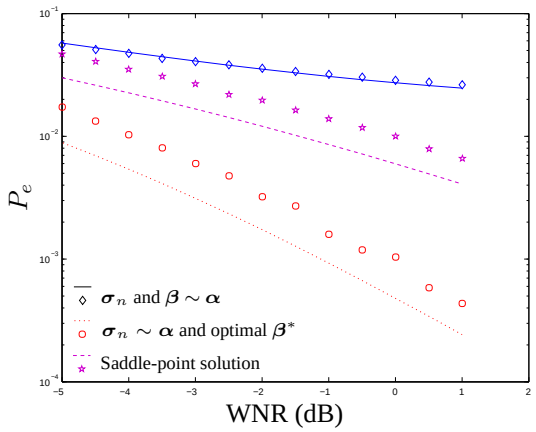

Fig. 1. BER versus WNR for spread-spectrum ($L = 50$) showing three different attacking/decoding strategies.

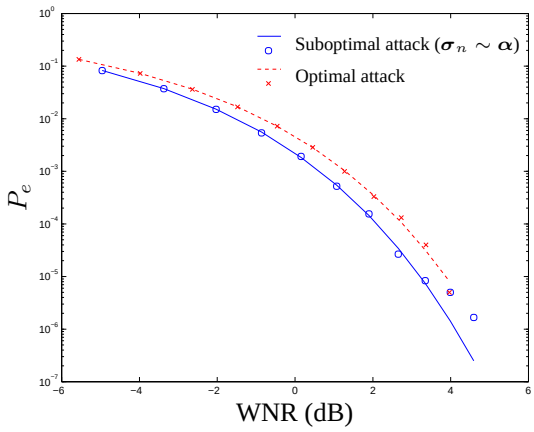

Fig. 2. BER versus WNR corresponding to the optimal and suboptimal attacks for QP ($L = 10$) when the decoder follows the optimum strategy.

the theoretical results lie close to the empirical ones, although for those where the optimal $\boldsymbol{\beta}^*$ is used the difference is larger. This can be explained by the fact that $\boldsymbol{\beta}^*$ depends on the Wiener filter coefficients, which in turn vary after having hidden the information. It is very interesting to note that the classical spread-spectrum solution can be significantly improved by using an optimal decoding strategy. The performance at the game-equilibrium lies somewhere in the middle between the two former cases. Recall that this is a satisfactory strategy for both parties due to the existence of a saddle-point.

Regarding QP, Figure 2 shows the results of comparing the optimal strategy given by the saddle-point solution against a suboptimal attack for the QP scheme with 10 pixels per information bit (watermarking is performed in the spatial domain). This suboptimal attack consists in perceptually-shaped noise, for which the noise variance $\sigma_{n_j}^2$ is proportional to the squared perceptual mask α_j^2. We can see that the difference between both strategies is not too large due to the fact that in this scenario all values of the perceptual mask are very close to each other.

6. CONCLUSIONS

Two different data hiding methods have been analyzed from a game-theoretic point of view, using the BER as a cost function. We have provided a theorem which proves useful not only for establishing the optimal joint decoding and attacking strategies, but also for showing that these in fact constitute a saddle-point from which neither the attacker nor the decoder are interested in deviating. For spread-spectrum, the game equilibrium is achieved by a water-filling attack and a decoding strategy that is richer than the classical cross-correlating receiver: where the filtered host image samples are small (i.e., small host-interference), the optimal weights are constant; while for those samples where host interference is large, the optimal weights somehow resemble those used in the cross-correlating receiver. In fact, QP can be considered a limiting case, where host interference is almost eliminated, so no weighting is needed for decoding in the saddle-point solution. However, this reasoning should not be taken too far: for dither modulations (DM), where host-interference is also nonexisting, the saddle-point solution turns out to be more involved than those given here [7].

7. REFERENCES

[1] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. on Information Theory*, vol. 49, pp. 563–593, March 2003.

[2] P. Moulin and M. Mıhak, "A framework for evaluating the data-hiding capacity of image sources," *IEEE Trans. on Image Processing*, vol. 11, pp. 1029–1042, September 2002.

[3] A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 48, pp. 1639–1667, June 2002.

[4] J. J. Eggers and B. Girod, *Informed Watermarking*. Kluwer Academic Publishers, 2002.

[5] P. Moulin and A. Ivanovic, "The zero-rate spread-spectrum watermarking game," *IEEE Trans. on Signal Processing*, vol. 51, pp. 1098–1117, April 2003.

[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*, vol. 51, pp. 960–980, April 2003. Special Issue "Signal Processing for Data Hiding in Digital Media & Secure Content Delivery".

[7] P. Comesaña, F. Pérez-González, and F. Balado, "Attacking and decoding strategies for data hiding games with ber payoffs," *In preparation.*

[8] 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.