Practical Data-Hiding: Additive Attacks Performance Analysis

J. E. Vila-Forcén[1], S. Voloshynovskiy[1], O. Koval[1], F. Pérez-González[2] and T. Pun[1]

[1] University of Geneva, Department of Computer Science. 24 rue Général-Dufour, CH 1211, Geneva, Switzerland, {vila,svolos,koval,pun}@cui.unige.ch
[2] University of Vigo, Signal Theory and Communications Department. E-36200 Vigo, Spain, fperez@gts.tsc.uvigo.es

Abstract. The main goal of this tutorial is to review the theory and design the worst case additive attack (WCAA) for $|\mathcal{M}|$-ary quantization-based data-hiding methods using as performance criteria the error probability and the maximum achievable rate of reliable communications. Our analysis focuses on the practical scheme known as distortion compensation dither modulation (DC-DM). From the mathematical point of view, the problem of the worst case attack (WCA) design using probability of error as a cost function is formulated as the maximization of the average probability of error subject to the introduced distortion for a given decoding rule. When mutual information is selected as a cost function, a solution to the minimization problem should provide such an attacking noise probability density function (pdf) that will maximally decrease the rate of reliable communications for an arbitrary decoder structure. The obtained results demonstrate that, within the class of additive attacks, the developed attack leads to a stronger performance decrease for the considered class of embedding techniques than the additive white Gaussian or uniform noise attacks.

1 Introduction

Data-hiding techniques aim at reliably communicating the largest possible amount of information under given distortion constraints. Their resistance against different attacks determine the possible application scenarios. An extensive review of various application of digital data-hiding techniques is given in [21]. The knowledge of the WCA allows to create a fair benchmark for data-hiding techniques and makes it possible to provide reliable communications with the use of appropriate error correction codes.

In general, the digital data-hiding can be considered as a game between the data-hider and the attacker. This three-party two-players game was already investigated by Moulin and O'Sullivan [12] where two set-ups are analyzed. In the first one, the host is assumed to be available at both encoder and decoder prior to the transmission, the so-called *private game*. In the second one, the host is only available at the encoder as in Fig. 1, i.e., the *public game*. The performance

is analyzed with respect to the maximum achievable rate when the decoder is aware of the attacking channel and therefore *maximum likelihood* (ML) decoding is applied. A similar game-theoretic analysis of the $|\mathcal{M}|$-ary information detection problem, the so-called zero-rate spread spectrum watermarking problem, is performed in [11]. As in the previous case, it is assumed that the detector has the possibility to learn the statistics of the attacking channel.

In both cases [11, 12], the results were obtained under the assumption of continuous input alphabets. They lead to the conclusion that the optimal attacker strategy in the class of additive blockwise memoryless attacks corresponds to the application of the Gaussian test channel from the rate-distortion theory.

The knowledge of the attacking channel at the decoder is not a realistic case for most practical applications. Somekh-Baruch and Merhav considered the data-hiding problem in terms of maximum achievable rates and error exponents. They assumed that the host data is available either at both encoder and decoder [1] or only at the encoder [16] and supposed that neither encoder nor decoder is aware of the attacker strategy. In their consideration, the class of potentially applied attacks is significantly broader than in the previous study case [12] and includes any conditional pdf that satisfies a certain energy constraint. Although the solution of the problem is classically presented in terms of the achievable rate establishing the maximum number of messages $|\mathcal{M}|$ that can be reliably communicated, the error exponents solution is interesting in many practical applications where the objective is to minimize the probability of error at a given communications rate.

Quantization-based data-hiding methods have attracted attention in the watermarking community. They are a practical implementation of a binning technique for channels whose state is non-causally available at the encoder considered by Gel'fand-Pinsker [8]. Recently it has been also demonstrated [13] that quantization-based data-hiding performance coincides with the spread-spectrum (SS) data-hiding at the low-WNR by taking into account the host statistics and by abandoning the assumption of an infinite image to watermark ratio.

The quantization-based methods have been widely tested against a fixed channel and assuming that the channel transition pdf is available at the decoder. A *minimum Euclidean distance* (MD) decoder is implemented as a low-complexity equivalent of the ML decoder under the assumption of a channel pdf created by the symmetric extension of a monotonically non-increasing function [2].

It is a common practice in the data-hiding community to measure the performance in terms of the error rate for a given decoding rule as well as the maximum achievable rate of reliable communications. In this tutorial we will analyze the WCAA using both criteria. We restrict the encoding to the quantization-based one and the channel to the class of additive attacks only. We assume that the attacker might be informed of the encoding strategy and also of the decoding one for the error exponent analysis, while both encoder and decoder are uninformed of the channel. Furthermore, the encoder is aware of the host image but not of the attacking strategy.

It is important to note that the optimality of the attack critically relies on the input alphabet even under power-limited attacks. McKellips and Verdu showed that the additive white Gaussian noise (AWGN) is not the WCAA for discrete input alphabets such as pulse amplitude modulation in digital communications [10]. Similar conclusion for data-hiding was obtained by Pérez-González *et al.* [14], who demonstrated that the uniform noise attack performs worse than the AWGN attack for some watermark-to-noise ratios (WNRs). In [15], Pérez-González demonstrated that the AWGN cannot indeed be the WCAA because of its infinite support. Vila-Forcén *et al.* [19] and Goteti and Moulin [9] solved independently the min-max problem for distortion-compensated dither modulation (DC-DM) [3] in terms of probability of error for the fixed decoder, binary signaling, the subclass of additive attacks in data-hiding and detection-formulation, respectively. The additional difference between the two approaches consists in the definition of the cost function. While in the former case explicit computation of the probability of error is performed for the selected class of embedding strategies, in the latter one the Bhatacharyya bound is exploited in order to reduce the complexity of the considered game optimization problem. Simultaneously, Vila-Forcén *et al.* [20] and Tzschoppe *et al.* [18] derived the WCAA for DC-DM using the mutual information as objective function for additive attacks and binary signaling.

The goal of this paper is to provide an overview of the WCAA against quantization-based data-hiding techniques, focusing on the core principles and basic performance measures used in the data-hiding community. We did not attempt to provide a comprehensive overview of all possible attacking strategies that could be applied against quantization-based methods. All these classes of attacks are rather broad for this review and include various geometrical transformation and signal processing attacks as well as attacks that combine prior information about scheme design with security leakages revealed by the attacker. The last group is the most dangerous one besides the fact that it requires some specific information about the data-hiding technique. The geometrical attacks are quite generic and can be applied to any data-hiding method disregarding any prior information about the codebook design. Signal processing attacks are generally based on the statistical priors about the host data and the watermark. The group of WCAA conforms to the signal processing attacks and directly exploits the knowledge of the watermark statistics caused by the structured codebook design. We refer the interested readers to [23, 24] for more information about attacks classification. More recent studies [4, 17, 22] address the impact of security leakages in the scope of information-theoretic analysis for geometrically structured and quantization-based codebooks and general reversibility of watermark embedding.

This paper aims at establishing the information-theoretic limits of $|\mathcal{M}|$-ary quantization-based data-hiding techniques and developing a benchmark that can be used for the fair comparison of different quantization-based methods.

The selection of the distortion compensation parameter α' (see Section 2.2) fixes the encoder structure for the quantization-based methods. Although the optimal α' can easily be determined when the power of the noise is available at

the encoder prior to the transmission [6], this is not always feasible for various practical scenarios. Nevertheless, the availability of the attacking power and of the attacking pdf is a very common assumption in most data-hiding schemes. We will demonstrate that for a specific decoder (MD decoder) it is possible to calculate the optimal α' independently of the attack variance and pdf for the block error probability as a cost function.

Notations We use capital letters to denote scalar random variables X, bold capital letters to denote vector random variables $\mathbf{X}$ and corresponding small letters x and $\mathbf{x}$ to denote the realizations of scalar and vector random variables, respectively. An information message and a set of messages with cardinality $|\mathcal{M}|$ is designated as $m \in \mathcal{M}, \mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\}$, respectively. A host signal distributed according to the pdf $f_{\mathbf{X}}(\mathbf{x})$ is denoted by $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x})$; $\mathbf{Z} \sim f_{\mathbf{Z}}(\mathbf{z})$, $\mathbf{W} \sim f_{\mathbf{W}}(\mathbf{w})$ and $\mathbf{V} \sim f_{\mathbf{V}}(\mathbf{v})$ represents the attack, the watermark and the received signal, respectively. The step of quantization is equal to Δ and the distortion-compensation factor is denoted as α'. The variance of the watermark is σ_W^2 and the variance of the attack is σ_Z^2. The watermark-to-noise ratio (WNR) is given by $\mathrm{WNR} = 10 \log_{10} \xi$, where $\xi = \frac{\sigma_W^2}{\sigma_Z^2}$. The set of natural numbers is denoted as $\mathbb{N}$ and $\mathbb{I}_N$ denotes the $N \times N$ identity matrix.

2 Problem Formulation

2.1 Data-Hiding Formulation of the Gel'fand-Pinsker Problem

The Gel'fand-Pinsker problem [8] has been recently revealed as the appropriate theoretical framework of data-hiding communications with side information. The Gel'fand-Pinsker data-hiding set-up is presented in Fig. 1. The random variable $\mathbf{X}$ stands for the host signal, which is independent and identically distributed (i.i.d.) according to $p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i)$ and available non-causally at the encoder. The encoder is a mapping $\phi : \mathcal{M} \times \mathcal{X}^N \times \mathcal{K} \to \mathcal{W}^N$, where the key $K \in \mathcal{K}, \mathcal{K} = \{1, 2, \ldots, |\mathcal{K}|\}$. The stego data $\mathbf{Y}$ is obtained using the embedding mapping: $\varphi : \mathcal{W}^N \times \mathcal{X}^N \to \mathcal{Y}^N$. The decoder estimates the embedded message as $\psi : \mathcal{V}^N \times \mathcal{K} \to \mathcal{M}$. According to this scheme, a key is available at both encoder and decoder. Nevertheless, key management is outside of the scope of this paper and will not be considered further.

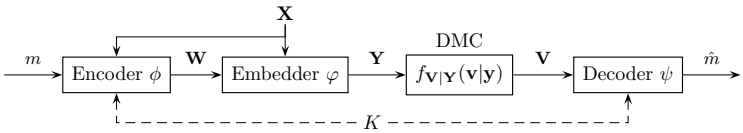

Fig. 1. Gel'fand-Pinsker data-hiding set-up

Two constraints apply to the Gel'fand-Pinsker in the data-hiding scenario: the embedding and the channel distortion constraints [12]. Let $d(\cdot, \cdot)$ be a non-

negative function and σ_W^2, σ_Z^2 be two positive numbers, the embedder and the channel are said to satisfy the embedding and channel distortion constraints if:

$$\sum_{\mathbf{x}\in\mathcal{X}^N}\sum_{\mathbf{y}\in\mathcal{Y}^N} d(\mathbf{x},\mathbf{y}) f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \leq \sigma_W^2; \quad \sum_{\mathbf{y}\in\mathcal{Y}^N}\sum_{\mathbf{v}\in\mathcal{V}^N} d(\mathbf{y},\mathbf{v}) f_{\mathbf{Y},\mathbf{V}}(\mathbf{y},\mathbf{v}) \leq \sigma_Z^2, \quad (1)$$

where $d(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} d(x_i, y_i)$.

Costa considered the Gel'fand-Pinsker problem for the i.i.d. Gaussian case and mean square error distance [5]. The embedder φ produces $\mathbf{Y} = \mathbf{W} + \mathbf{X}$, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbb{I}_N)$ and the channel output is obtained as: $\mathbf{V} = \mathbf{X} + \mathbf{W} + \mathbf{Z}$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbb{I}_N)$. The estimate of the message $\hat{m}$ is obtained at the decoder as in the Gel'fand-Pinsker set-up.

2.2 Quantization-Based Data-Hiding Techniques:

Aiming at reducing the Costa codebook exponential complexity, a number of practical data-hiding algorithms exploit *structured codebooks* instead of random ones. The most famous discrete approximations of Costa problem are known as DC-DM [3] and scalar Costa scheme (SCS) [6]. The structured codebooks are designed using quantizers (or lattices [7]) in order to achieve host interference cancellation. In the case of DC-DM, the stego data is obtained as follows:

$$\phi_{\text{DC-DM}}(m, x, \alpha') = y = x + \alpha'(Q_m(x) - x), \quad (2)$$

where $Q_m(\cdot)$ denotes a vector or scalar quantizer for the message m and $0 < \alpha' \leq 1$ is the analogue of the Costa optimization parameter α. If $\alpha' = 1$, the DC-DM simplifies to the DM: $\phi_{\text{DM}}(m, x) = \phi_{\text{DC-DM}}(m, x, 1)$.

3 Error Probability as a Cost Function

When the average error probability is selected as a cost function, we formulate the problem of Fig. 1 as:

$$P_B^{*(N)} = \min_{\phi,\psi} \max_{f_{V|Y}(\cdot|\cdot)} P_B(\phi, \psi, f_{V|Y}(\cdot|\cdot)). \quad (3)$$

The error probability depends on the particular encoder/decoder pair (ϕ, ψ) and the attacking channel $f_{\mathbf{V}|\mathbf{Y}}(\mathbf{v}|\mathbf{y})$, i.e., $P_B(\phi, \psi, f_{V|Y}(v|y)) = \Pr[\hat{m} \neq m | M = m]$. Here, we assume that the attacker knows both encoder and decoder strategies and selects its attacking strategy accordingly. Both encoder and decoder choose their strategy without knowing the attack in advance. Although this is a very conservative set-up, it is also important for various practical scenarios.

The more advantageous set-up for the data-hider is based on the assumption that the decoder selects its strategy knowing the attacker choice:

$$\min_{\phi} \max_{f_{V|Y}(\cdot|\cdot)} \min_{\psi} P_B(\phi, \psi, f_{V|Y}(\cdot|\cdot)). \quad (4)$$

Here, the attacker knows only the encoding function, which is fixed prior to the attack, and the decoder is assumed to be aware of the attack pdf.

In the general case, Somekh-Baruch and Merhav [1] have shown that the following inequalities apply for the above scenarios :

$$\min_{\phi,\psi} \max_{f_{V|Y}(\cdot|\cdot)} P_B(\phi, \psi, f_{V|Y}(\cdot|\cdot)) \geq \min_{\phi} \max_{f_{V|Y}(\cdot|\cdot)} \min_{\psi} P_B(\phi, \psi, f_{V|Y}(\cdot|\cdot)) \quad (5)$$

$$= \min_{\phi} \max_{f_{V|Y}(\cdot|\cdot)} P_B(\phi, \psi^{\mathrm{ML}}, f_{V|Y}(\cdot|\cdot)), \quad (6)$$

where (6) assumes that the decoder is aware of the attacking pdf and therefore the minimization at the decoder results in the optimal ML decoding strategy ψ^{ML}. Using (6) one can write:

$$\min_{\phi} \max_{f_{V|Y}(\cdot|\cdot)} P_B(\phi, \psi^{\mathrm{MD}}, f_{V|Y}(\cdot|\cdot)) \geq \min_{\phi} \max_{f_{V|Y}(\cdot|\cdot)} P_B(\phi, \psi^{\mathrm{ML}}, f_{V|Y}(\cdot|\cdot)), \quad (7)$$

with equality if, and only if, the MD decoder coincides with the optimal ML decoder. In the class of additive attacks, the attacking channel transition pdf is only determined by the pdf of the additive noise $f_Z(z)$. Finally, in this analysis we assume independence of the error probability on the quantization bin where the received signal v lies (because the error decision region $\overline{\mathcal{R}}_m$ has periodical structure and the host pdf $f_X(x)$ is assumed to be asymptotically constant within each quantization bin).

The problem (3) implies that the attacker might know both encoding and decoding strategy. Here, we target finding the WCAA pdf and the optimum fixed encoding strategy independently of the particular attacking case which guarantees reliable communications and provides an upper bound on the error probability.

3.1 Additive White Gaussian Noise Attack

The probability of error is determined using the equivalent noise pdf given by the convolution of the self-noise (a delta in the DM case and a uniform in the DC-DM one) with the attacking noise. The analytical expression for the error probability does not exist, and it is evaluated numerically. The error probability for the DM and the DC-DM under the AWGN attack is depicted in Fig. 2.

3.2 Uniform Noise Attack

It was shown [14] that the uniform noise attack produces higher error probability than the AWGN attack for some particular WNR in the binary signaling case. This fact contradicts the common belief that the AWGN is the WCAA for all data-hiding methods since it has the highest differential entropy among all pdfs with bounded variance.

As for the AWGN attacking case, we assume that the MD decoder is used and the probability of error is calculated as the integral of the equivalent noise pdf over the error region. The corresponding performance of the DC-DM under the uniform noise attack is presented in Fig. 3. Since we are assuming fixed decoder, the error probability for the binary case can be higher than 0.5.

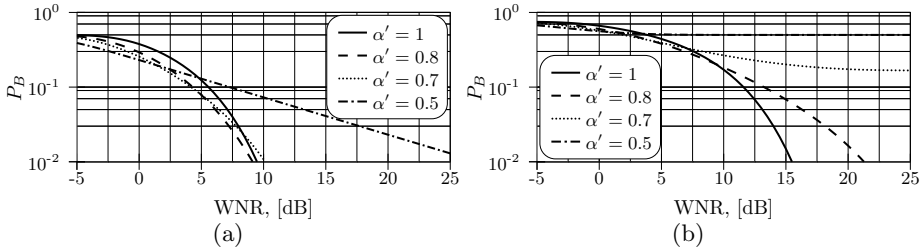

Fig. 2. Error probability analysis results for the AWGN attack case: (a) binary signaling and (b) quaternary signaling

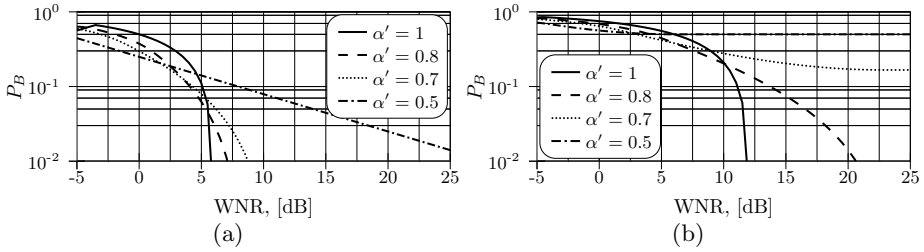

Fig. 3. Error probability for the uniform noise attack case: (a) binary signaling and (b) quaternary signaling

3.3 The Worst Case Additive Attack

The problem of the WCAA for digital communications based on binary pulse amplitude modulation (PAM) was considered in [10] using the error probability under attack power constraint. In this paper, the problem of the WCAA is addressed for the quantization-based data-hiding methods.

The problem (4) for the DM with the fixed MD decoder is given by:

$$\min_{\alpha'} \max_{f_Z(\cdot)} P_B(\alpha', \psi^{\mathrm{MD}}, f_Z(\cdot)), \tag{8}$$

where the encoder is optimized over all α' such that $0 < \alpha' \leq 1$, and the attacker selects the attack pdf $f_Z(\cdot)$ maximizing the error probability P_B. Since the encoder must be fixed in advance in the practical set-ups, we will first solve the above min-max problem as an internal maximization problem for a given encoder/decoder pair:

$$\max_{f_Z(\cdot)} P_B(\alpha', \psi^{\mathrm{MD}}, f_Z(\cdot)) = \max_{f_Z(\cdot)} \int_{\overline{\mathcal{R}}_m} f_V(v|M=m)dv, \tag{9}$$

where $0 < \alpha' \leq 1$, subject to the constraints:

$$\int_{-\infty}^{\infty} f_Z(z)dz = 1, \qquad \int_{-\infty}^{\infty} z^2 f_Z(z)dz \leq \sigma_Z^2, \tag{10}$$

where the first constraint follows from the pdf definition and σ_Z^2 constrains the attack power. The obtained error probabilities are depicted in Fig. 4, where the

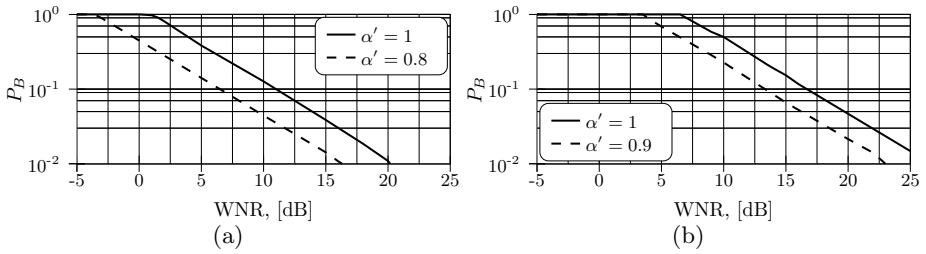

Fig. 4. WCAA error probability optimization results: (a) binary signaling and (b) quaternary signaling

maximum is equal to 1 since we are assuming that the decoder is fixed (MD decoder) and it is completely known to the attacker. In a different decoding case when it is possible to invert the bit values, the maximum error probability will be equal to 0.5.

We approximate the performance of the WCAA by a so-called $3-\delta$ attack whose pdf is presented in Fig. 5. The $3-\delta$ attack provides a simple and powerful attacking strategy, which approximates the performance of the WCAA and might be used for testing different data-hiding algorithms. In order to demonstrate how accurate this approximation is, one needs to compare the average error probability caused by this attack versus the numerically obtained results.

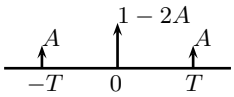

Fig. 5. $3-\delta$ attack, $0 \leq A \leq 0.5$

The corresponding performance for the DM and the DC-DM under the 3-δ attack is presented in Fig. 6. The comparison between Fig. 4 and Fig. 6 demonstrates that the 3-δ attack produces asymptotically the same error probability as the optimization results. The presented results (Fig. 2, Fig. 3 and Fig. 4) demonstrate that the gap between the AWGN attack and the real worst case attack can be larger than 5dB in terms of the WNR.

The error probability as a function of the distortion compensation parameter for a given WNR demonstrates that the $3-\delta$ attacking scheme is worse than either the uniform or Gaussian ones (Fig. 7). If the noise attack is known, it is possible to select such an α' that minimizes the error probability for the given WNR in Fig. 7. For example, if WNR = 0dB and Gaussian noise is applied, the optimal distortion compensation factor is $\alpha' = 0.53$, resulting in $P_B = 0.23$. Nevertheless, the encoder and the decoder are in general uninformed of the attacking strategy in advance and a mismatch in the attacking scheme may cause a bit error probability[3] of 1, while for $\alpha' = 0.66$ the maximum bit error probability is $P_B = 0.33$.

[3] In general the maximum bit error probability is equal to 1 for the fixed MD decoder.

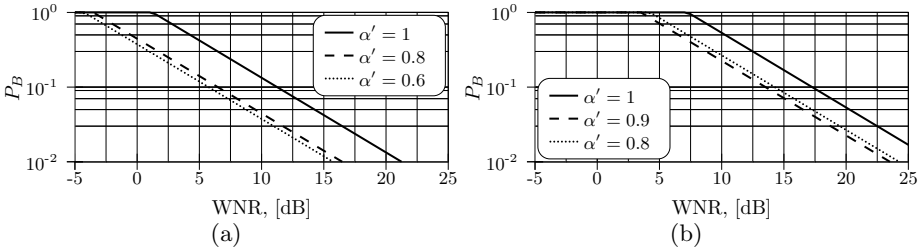

Fig. 6. Error probability analysis results for the $3-\delta$ attack case: (a) binary signaling and (b) quaternary signaling

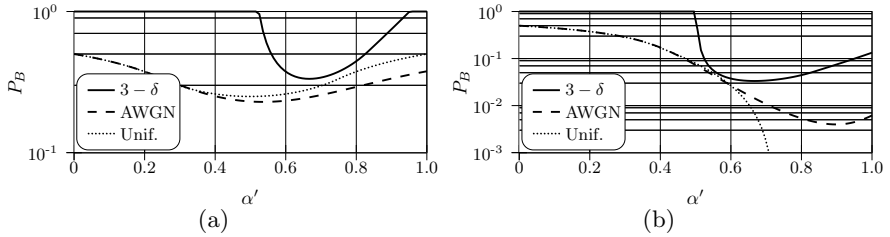

Fig. 7. Error probability comparison as a function of the distortion compensation parameter for the $3-\delta$, Gaussian and uniform attacks and binary signaling: (a) WNR = 0dB, (b) WNR = 10dB

In order to find the optimal compensation parameter value that will allow the data-hider to upper bound the error probability introduced by the WCAA, we analyzed the error probability given by the 3-δ attack. Surprisingly, it was found that, independently of the operational WNR, $\alpha' = \alpha'_{\text{opt}} = \frac{2(|\mathcal{M}|-1)}{2|\mathcal{M}|-1}$ guarantees the lowest error probability of the analyzed data-hiding techniques under the WCAA (Fig. 8). Having this bound on the error probability, it is possible to guarantee reliable communications using proper error correction codes. Therefore, one can select such a fixed distortion compensation parameter $\alpha' = \alpha'_{\text{opt}}$ at the uninformed encoder and the MD decoder, which guarantees a bounded error probability. Substituting $\alpha' = \alpha'_{\text{opt}}$ into the error probability, one obtains the upper bound on the error probability:

$$P_B(\alpha'_{\text{opt}}) = \frac{1}{6}|\mathcal{M}|(|\mathcal{M}|-1)\xi^{-1}. \tag{11}$$

4 Mutual Information as a Cost Function

The analysis of the WCAA with mutual information as a cost function is crucial for the fair evaluation of quantization-based data-hiding techniques. It provides the information-theoretic performance limit (in terms of achievable rate of reliable communications) that can be used for benchmarking of different practical robust data-hiding algorithms. Moulin *et al.* [12] considered the maximum achievable rate in the Gel'fand-Pinsker set-up as a max-min problem:

$$C = \max_{\phi} \min_{f_{V|Y}(\cdot|\cdot)} \left[I(U;V) - I(U;X) \right], \tag{12}$$

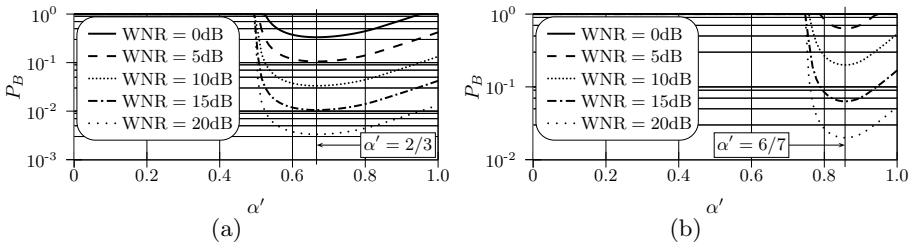

Fig. 8. Error probability analysis results as a function of the distortion compensation parameter α' for the $3-\delta$ attack: (a) binary signaling and (b) quaternary signaling

for a blockwise memoryless attack, the embedder distortion constraint σ_W^2 and the attacker distortion constraint σ_Z^2. In the case of quantization-based methods the mutual information is measured between the communicated message M and the channel output V [15] and the above problem is given by:

$$\max_{\phi} \min_{f_{V|Y}(\cdot|\cdot)} I_{\phi, f_{V|Y}(\cdot|\cdot)}(M; V'). \tag{13}$$

where $V' = Q_{\Delta}(V) - V$, since it was shown in [15] that modulo operation does not reduce the mutual information between V and M if the host is assumed to be flat within the quantization bins.

Rewriting the inequalities (5)–(6) for the mutual information we have:

$$\max_{\phi} \min_{f_{V|Y}(\cdot|\cdot)} I_{\phi, f_{V|Y}(\cdot|\cdot)}(M; V') \leq \max_{\phi} I_{\phi, \tilde{f}_{V|Y}(\cdot|\cdot)}(M; V'), \tag{14}$$

with equality if, and only if, the fixed attack $\tilde{f}_{V|Y}(\cdot|\cdot)$ coincides with the WCAA. Thus, the decoder in Fig. 1 is not fixed and we assume that the channel attack pdf $f_{V|Y}(\cdot|\cdot)$ is available at the decoder (informed decoder) and, consequently, ML decoding is performed. Under previous assumptions of quantization-based embedding and additive attack, it is possible to rewrite (13) as:

$$\max_{\alpha'} \min_{f_Z(\cdot)} I_{\alpha', f_Z(\cdot)}(M; V'). \tag{15}$$

Assuming equiprobable symbols, one obtains [15, 20]:

$$I_{\alpha', f_Z(\cdot)}(M; V') = D\left(f_{V'|M}(v'|M=1) || f_{V'}(v')\right), \tag{16}$$

where $D(\cdot||\cdot)$ denotes the Kullback-Leibler distance (KLD). The next section is dedicated to the analysis of the DM and the DC-DM under the AWGN attack, the uniform noise attack and the WCAA.

4.1 Additive White Gaussian Noise Attack

When the DM and the DC-DM undergo the AWGN, no closed analytical solution to the mutual information minimization problem exists; the minimization was therefore performed using numerical computations. The results of this analysis for the binary and quaternary cases are shown in Fig. 9.

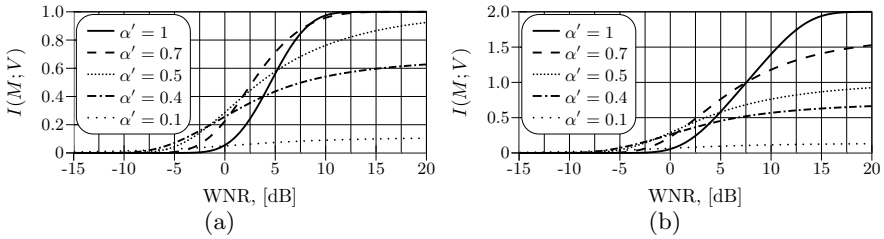

Fig. 9. Mutual information analysis results for the AWGN attack case and different α' and WNR values: (a) binary signaling and (b) quaternary signaling

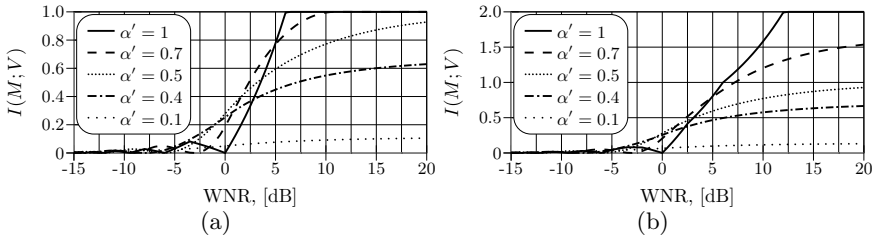

Fig. 10. Mutual information analysis results for the uniform noise attack case: (a) with binary signaling and (b) quaternary signaling

4.2 Uniform Noise Attack

It was shown [14] that the uniform noise attack is stronger than the AWGN attack for some WNRs when the error probability is used as a cost function. One of the properties of the KLD measure states that it is equal to zero if, and only if, the two pdfs are equal. In case the uniform noise attack is applied, this condition holds for some particular values of WNR for the mutual information given by (16). It can be demonstrated that $I(M;V') = 0$ when $\xi = \frac{\alpha'^2}{k^2}, k \in \mathbb{N}$ for the $|\mathcal{M}|$-ary signaling. The mutual information of quantization-based data-hiding techniques for the uniform noise attacking case with binary and quaternary signaling is depicted in Fig. 10.

The uniform noise attack guarantees that it is not possible to communicate using the DC-DM at $\xi \leq \alpha'^2$, and therefore distortion compensation parameter α' has a strong influence on the performance at the low-WNR. As a consequence, $\xi = \alpha'^2$ represents the WNR corresponding to zero rate communication, if the attacking variance satisfies $\sigma_Z^2 \geq \frac{D_w}{\alpha'^2}$.

For example (binary signaling, Fig. 10(a)), if the data-hider anticipates a WNR $= -6$dB, he/she could select $\alpha' = 0.7$ to maximize the mutual information. Nevertheless, at the WNR $= -3$dB the mutual information is zero for $\alpha' = 0.7$. Therefore, it is possible for the attacker to inhibit reliable communications by applying an attack 3dB lower in power than the data-hider prediction in this example. This forces the data-hider to decrease the value of α'. Therefore, the attacker can inhibit communications by making less efforts. In this example, to reduce the power of the attack on 3dB from the embedder prediction is favorable for the attacker.

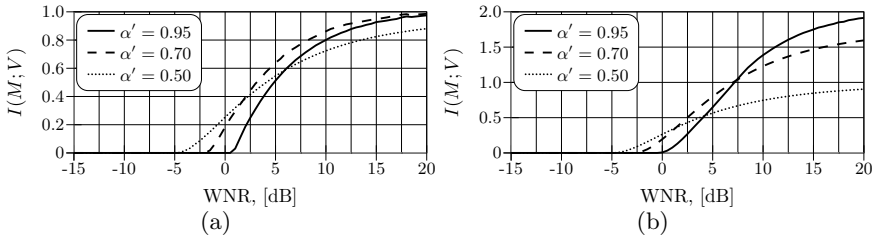

Fig. 11. Mutual information analysis results for the WCAA case: (a) binary signaling and (b) quaternary signaling

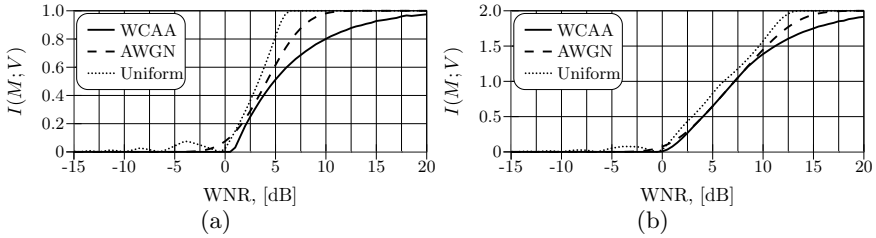

Fig. 12. Comparison of different attacks using mutual information as a cost function: (a) $\alpha' = 0.95$, binary signaling and (c) $\alpha' = 0.95$, quaternary signaling

4.3 The Worst Case Additive Attack

The problem of the WCAA using the mutual information as a cost function can be formulated using (15). Since the encoder must be fixed in advance as for the probability of error analysis case, we solve the max-min problem as a constrained minimization problem:

$$\min_{f_Z(\cdot)} I_{\alpha', f_Z(\cdot)}(M; V') = \min_{f_Z(\cdot)} D\left(f_{V'|M}(v'|M=1) || f_{V'}(v')\right), \tag{17}$$

where $0 < \alpha' \leq 1$. The constraints in (17) are the same as with the error probability oriented analysis case (10). Unfortunately, this problem has no closed form solution and it was solved numerically. The obtained results are presented for different α' values in Fig. 11. In comparison with the AWGN and the uniform noise attacks, they demonstrate that the developed attack produces the maximum possible loss in terms of the mutual information for all WNRs (Fig. 12).

In the analysis of the WCAA using the error probability as a cost function, the optimal α' parameter was found. Unfortunately, it is not the case in the mutual information oriented analysis, and its value varies with the WNR. In Fig. 13 the optimum α' values as a function of the WNR are presented for different input distributions in comparison with the optimum SCS parameter [6]. It demonstrates that SCS optimum distortion compensation parameter designed for the AWGN is also a good approximation for the WCAA case.

Using the optimum α' for each WNR, the resulting mutual information (17) is presented in Fig. 14(a) for different cardinality of the input alphabet compared to the performance of the AWGN using the optimized $\alpha = \alpha_{opt}$ parameter [12]. The obtained performance demonstrates that the developed WCAA is worse than the AWGN whenever the optimum α' is selected.

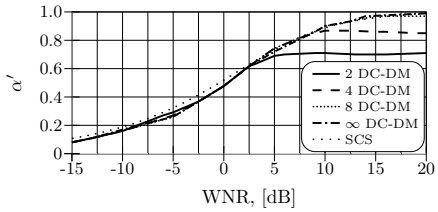

Fig. 13. Optimum distortion compensation parameter α' when the mutual information is selected as a cost function

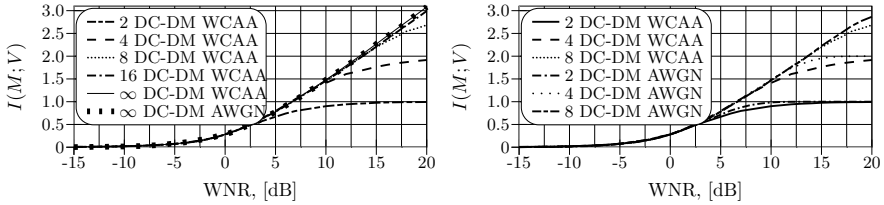

Fig. 14. Maximum achievable rate for different cardinality of the input alphabet under the WCAA compared to the AWGN (a) for $|\mathcal{M}| \to \infty$ and (b) for $|\mathcal{M}| < \infty$

It is possible to observe in Fig. 14(a) that the impact of the WCAA is very similar to the AWGN and that the difference in terms of the mutual information is negligible. Although the AWGN is not the WCAA, its performance is an accurate and practical approximation to the WCAA in the asymptotic case when $|\mathcal{M}| \to \infty$. For $|\mathcal{M}| < \infty$, the difference might be important for some WNRs and it is needed to consider the real WCAA as it is presented in Fig. 14(b).

5 Conclusions

In this tutorial we analyzed the performance of quantization-based data-hiding techniques from the probability of error and mutual information perspectives. The comparison between the analyzed cost functions demonstrated that in a rigid scenario with a fixed decoder, the attacker can decrease the rate of reliable communication more severely than by using either the AWGN or the uniform noise attacks. We showed that the AWGN attack is not the WCAA in general, and we obtained an accurate and practical analytical approximation to the WCAA, the so-called $3-\delta$ attack, when the cost function is the probability of error for the fixed MD decoder. For the $3-\delta$ attack, $\alpha' = \frac{2(|\mathcal{M}|-1)}{2|\mathcal{M}|-1}$ was found to be the optimal value for the MD decoder that allows to communicate with an upper bounded probability of error for a given WNR. This value could be fixed without prior knowledge of the attacking pdf.

The analysis results obtained by means of numerical optimization showed that there exists a worse attack than the AWGN when the mutual information was used as a cost function. Contrarily to the error probability analysis case, the optimal distortion compensation parameter (α') depends on the operational WNR for the mutual information analysis case. The particular behaviour of

the mutual information under uniform noise attack was considered, achieving zero-rate communication for attacking variances σ_Z^2 such that $\sigma_Z^2 \geq \frac{D_w}{\alpha'^2}$. The presented results should serve as a basis for the development of fair benchmarks for various data-hiding technologies under the assumptions of high rate and $\sigma_X^2 \gg \sigma_W^2$.

Acknowledgment

This paper was partially supported by SNF Professorship grant No PP002-68653/1, Interactive Multimodal Information Management (IM2) project and by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT. The authors are thankful to the members of the Stochastic Image Processing group at University of Geneva and to Pedro Comesaña and Luis Pérez-Freire of the Signal Processing in Communications Group at University of Vigo for many helpful and interesting discussions. The information in this document reflects only the author's 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. A. Somekh-Baruch and N. Merhav. On the error exponent and capacity games of private watermarking systems. *IEEE Trans. on Information Theory*, 49(3):537–562, March 2003.
2. Mauro Barni and Franco Bartolini. *Watermarking Systems Engineering*. Marcel Dekker, Inc., New York, 2004.
3. 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*, 47:1423–1443, 2001.
4. P. Comesaña-Alfaro, L. Pérez-Freire, and F. Pérez-González. An information-theoretic framework for assessing security in practical watermarking and data hiding scenarios. In *WIAMIS 2005, 6th International Workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Switzerland, April 13-15 2005.
5. M. Costa. Writing on dirty paper. *IEEE Trans. on Information Theory*, 29(3):439–441, 1983.
6. Joachim J. Eggers, Robert Bäuml, Roman Tzschoppe, and Bernd Girod. Scalar costa scheme for information embedding. *IEEE Transactions on Signal Processing*, 51(4):1003–1019, April 2003.
7. U. Erez and R. Zamir. Lattice decoding can achieve 0.5 log(1+snr) over the additive white gaussian noise channel using nested codes. In *Proceedings of IEEE International Symposium on Information Theory*, page 125, Washington DC, USA, June 2001.
8. S.I. Gel'fand and M.S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.
9. A. K. Goteti and P. Moulin. Qim watermarking games. In *Proc. ICIP*, Oct. 2004.
10. A. McKellips and S. Verdu. Worst case additive noise for binary-input channels and zero-threshold detection under constraints of power and divergence. *IEEE Transactions on Information Theory*, 43(4):1256–1264, July 1997.

11. P. Moulin and A. Ivanovic. The zero-rate spread-spectrum watermarking game. *IEEE Transactions on Signal Processing*, 51(4):1098–1117, April 2003.
12. P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Trans. on Information Theory*, 49(3):563–593, October 2003.
13. L. Peréz-Freire, F. Peréz-González, and S. Voloshynovskiy. Revealing the true achievable rates of scalar costa scheme. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Siena, Italy, September 29 - October 1 2004.
14. F. Peréz-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, Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery*, 51(4), April 2003.
15. Fernando Pérez-González. The importance of aliasing in structured quantization modulation data hiding. In *International Workshop on Digital Watermarking*, Seoul, Korea, 2003.
16. A. Somekh-Baruch and N. Merhav. On the capacity game of public watermarking systems. *IEEE Trans. on Information Theory*, 49(3):511–524, March 2004.
17. E. Topak, S. Voloshynovskiy, O. Koval, M.K. Mihcak, and T. Pun. Towards geometrically robust data-hiding with structured codebooks. *ACM Multimedia Systems Journal, Special Issue on Multimedia and Security*, 2005. submitted.
18. R. Tzschoppe, R. Bäuml, R. Fischer, A. Kaup, and J. Huber. Additive Non-Gaussian Attacks on the Scalar Costa Scheme (SCS). In *Proceedings of SPIE Photonics West, Electronic Imaging 2005, Security, Steganography, and Watermarking of Multimedia Contents VII (EI120)*, volume 5681, San Jose, USA, January 16-20 2005.
19. J. E. Vila-Forcén, S. Voloshynovskiy, O. Koval, F. Peréz-González, and Thierry Pun. Worst case additive attack against quantization-based watermarking techniques. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Siena, Italy, September 29 - October 1 2004.
20. J. E. Vila-Forcén, S. Voloshynovskiy, O. Koval, F. Peréz-González, and Thierry Pun. Worst case additive attack against quantization-based data-hiding methods. In *Proceedings of SPIE Photonics West, Electronic Imaging 2005, Security, Steganography, and Watermarking of Multimedia Contents VII (EI120)*, San Jose, USA, January 16-20 2005.
21. S. Voloshynovskiy, F. Deguillaume, O. Koval, and T. Pun. Information-theoretic data-hiding: Recent achievements and open problems. *International Journal of Image and Graphics*, 5(1):1–31, 2005.
22. S. Voloshynovskiy, O. Koval, E. Topak, J.E. Vila-Forcén, P. Comesaña, and Thierry Pun. On reversibility of random binning techniques: multimedia perspectives. In *9th IFIP TC-6 TC-11 Conference on Communications and Multimedia Security (CMS 2005)*, Salzburg Austria, September 2005.
23. S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun. Attack modelling: Towards a second generation benchmark. *Signal Processing*, 81(6):1177–1214, June 2001. Special Issue: Information Theoretic Issues in Digital Watermarking, 2001. V. Cappellini, M. Barni, F. Bartolini, Eds.
24. S. Voloshynovskiy, S. Pereira, T. Pun, J. Eggers, and J. Su. Attacks on digital watermarks: Classification, estimation-based attacks and benchmarks. *IEEE Communications Magazine (Special Issue on Digital watermarking for copyright protection: a communications perspective)*, 39(8):118–127, 2001. M. Barni, F. Bartolini, I.J. Cox, J. Hernandez, F. Pérez-González, Guest Eds. Invited paper.