An Image Watermarking Scheme Based on Information Theoretic Principles

Avi Levy, Neri Merhav[1]
HP Laboratories Israel
HPL-2001-13
January 18th , 2001*

image watermarking, information hiding, side information

In this report we investigate information hiding systems with non-malicious attack channels in which the decoder does not have access to the original source data. Recent works show that such information hiding systems can be viewed as communication systems with side information at the encoder. Based on this observation, we propose a novel watermarking scheme, named scaled bin encoding, which is based on capacity achieving random encoding scheme for an additive Gaussian channel with side information. This scheme out-performs previous watermarking techniques when applied to synthetic and imagery data. Furthermore, a theoretical and computational framework for the implementation of a family of information hiding schemes is proposed. This family, named binary modulation schemes, provides a computable expression for the theoretical information embedding rate, as well as fast encoding and decoding procedures. The decoding procedure is based on the maximum likelihood principle that improves upon the standard correlation based decoding techniques. We describe experimental results of implementing various binary modulation schemes for the additive Gaussian information hiding system and for DCT domain image watermarking .

* Internal Accession Date Only Approved for External Publication

[1] This work was done while N. Merhav was also with HP Labs Israel

HP Labs Israel, Technion City, Haifa 32000, Israel

1 Introduction

An information hiding system consists of three ingredients - an encoder, an attack channel, and a decoder (see Figure 1). The encoder receives a "message" to be embedded in a source data sequence named the "covertext" sequence. It encodes the message in an encoding sequence named the "stegotext" sequence, under a distortion constraint that restricts it to be close to the original covertext sequence. The stegotext sequence is the input to an attack channel which is a stationary and memoryless communication channel. The attack channel output sequence is the input to the decoder which produces an estimate for the original encoded message. We assume that the covertext sequence is not available at the decoder.

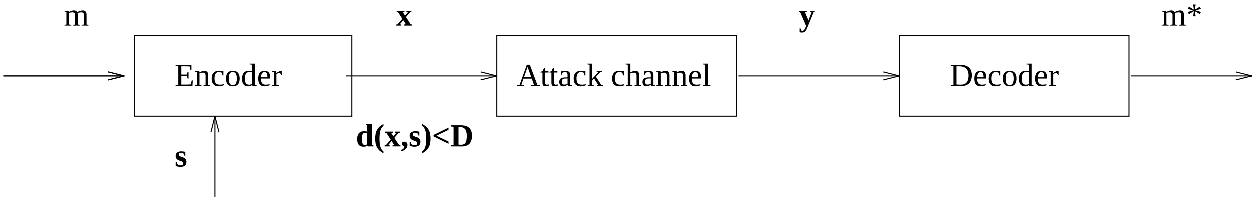

Figure 1: *Information hiding system: m - the message, $\mathbf{s}$ - the covertext sequence, $\mathbf{x}$ - the stegotext sequence, $\mathbf{y}$ - the attack channel output sequence, m^* - estimate message, $\mathbf{d}(\mathbf{x},\mathbf{y}) < D$ - the distortion constraint.*

Information hiding systems are modeled in either one of the following two scenarios:

1. An information game between the encoder, who needs to copyright-protect an original source sequence (the covertext) by embedding in it a signature (the message), and the attacker that attempts to remove this signature. In this scenario, both the encoder and the attacker are subjected to distortion constraints that prevent them from obscuring the content of the original sequence. The value of the game is the highest information rate that can reliably be decoded by a decoder which has no access to the original covertext sequence and for every attack channel that satisfies the distortion constraint [MO99, CL99].

2. A communication system where a message is to be embedded in a covertext sequence, with the restriction that the resulting stegotext sequence is subject to a distortion constraint. The stegotext sequence is subsequently being sent through a noisy communication channel that constitutes a non-malicious attack on the embedded message. By "non-malicious attack channel" we mean that the channel is not intentionally designed at removing the hidden information. The decoder decodes the embedded message from observing the attack channel output without resorting to the original covertext sequence [BCW00].

In this work, we adopt the second scenario and assume that the statistical properties of the attack channel are known both to the encoder and the decoder. This communication system model is useful for information hiding scenarios where the attacks on the stegotext signal are inflicted by non-malicious signal processing operations, such as filtering or compression.

Recently, the theoretical relationship between information hiding and channel coding with side information has been investigated [CMM99, MO99, Che00, SEG00]. Furthermore, it was shown that there exists a theoretical duality between information embedding and source coding with side information [BCW00, SEG00]. These relationships set the basis for information-theoretic analysis of information hiding systems and allow the derivation of explicit formulas for information hiding rates. In Section 2 of this report, we briefly present the theoretical results derived for information hiding systems in [MO99, Che00]. We note, however, that some of these results relate to the mean quadratic distortion measure, both for the encoder and for the attacker. This type of distortion measure has a great theoretical value and allows employing standard information-theoretic techniques. However, it does not model correctly some of the common applications of information hiding. As an example, under the mean quadratic distortion measure, the encoder and the attacker may severely damage the visual content of a covertext image by concentrating all the allowed distortion quota in a small region of its spatial or frequency domain representation. Standard models of the human visual system imply that a maximum distortion measure is more adequate for image watermarking applications [PZ98]. Motivated by this observation, we add to Section 2 some theoretical results pertaining to information hiding systems with the maximum distortion measure.

The theoretical results, presented in Section 2, provide insight into the nature of the information hiding problem. However, they do not provide practical means to calculate the capacity or to propose coding schemes for real information hiding systems (an important exception to this statement will be mentioned later). For this reason, we propose, in Section 3, a theoretical and computational framework named Binary Modulation Schemes (BMS's) for implementing efficient information hiding procedures. By limiting the scope to this class of schemes, we sacrifice optimality, to a certain extent, at the benefit of simplicity. We present computable expressions and bounds for the highest information embedding rate of a BMS and are able to construct a fast maximum likelihood decoder. The BMS framework allows for efficient implementations of common watermarking schemes such as those presented in [CKLS96, CW99] as well as the implementation of a new proposed scheme named Scaled Bin Encoding (SBE). This scheme is based on theoretical work of Costa [Cos83], where a capacity-achieving random coding scheme is proposed for the additive Gaussian channel with side information. The BMS implementation of the SBE scheme provides superior performance when applied to additive Gaussian information hiding systems and DCT domain image watermarking.

Efficient implementation of a BMS for image watermarking requires a stable statistical model for the covertext signal and the attack channel. The design parameters of the encoding and decoding procedures should be optimized to the statistical model in order to achieve a high information embedding rate and reliable decoding. In Section 4 we first describe a general framework for image watermarking statistical models, and then present a procedure for BMS transform domain watermarking. At the end of this section we provide a detailed description of a DCT domain statistical model.

In Section 5, we describe the implementations and compare the information embedding rates of three BMSs for the additive Gaussian channel. Then, we present the details of implementation and the results of SBE image watermarking in the DCT domain. Section 6 contains conclusions and discussion of future work.

2 Theoretical preliminaries

In this section of the report, we define an information hiding system with a non-malicious attack channel. We cite information-theoretic results pertaining to the related problem of coding for a communication channel with side information, and describe the equivalent results for information hiding systems with mean distortion constraint, established lately in the literature. Similar and somewhat stronger results for information hiding systems with distortion constraints based on maximum distortion measures, are then presented, while their proofs are deferred to the appendices. This body of theoretical results provides the foundation for the more practical analysis in the subsequent sections.

Throughout this report we use the following notation: Random variables are denoted by upper case letters (e.g., Z), and their realizations are denoted by lower case letters (e.g., z). Boldface letters denote sequences of length n (e.g., $\mathbf{Z} = (Z_1, ..., Z_n)$ is a random n-vector and $\mathbf{z} = (z_1, .., z_n)$ is a specific n-vector). Alphabets are denoted by caligraphic letters (e.g. $\mathcal{Z}$), the size of an alphabet is denoted by its absolute value (e.g. $\mid \mathcal{Z} \mid$) and the n-fold Cartesian product of these sets is denoted by the superscript n (e.g. $\mathcal{Z}^n$). The symbol m denotes the message to be transmitted in the information hiding system, $\mathbf{s}$ denotes the covertext sequence where the message is to be embedded, $\mathbf{x}$ denotes the encoder output - the stegotext sequence, $\mathbf{y}$ denotes the attack channel output sequence, and m^* denotes the estimated message at the decoder. A Gaussian probability distribution with mean μ and variance V is denoted $\mathcal{N}(\mu, V)$.

2.1 Definition of Information Hiding System (IHS)

In the definition of IHS, we make the following assumptions:

1. The message m is a sample from a uniformly distributed random variable M that takes values in the finite alphabet $\mathcal{M}$.
2. The covertext sequence $\mathbf{s}$ consists of n independent samples from a random variable S. The alphabet of S may be finite or continuous, and accordingly, its probability distribution or probability density is denoted by $p_S(s)$. This probability distribution/density is known to the encoder and the decoder, but the specific covertext sequence $\mathbf{s}$ is not available to the decoder.

3. The statistical characterization of the attack channel is fixed and known to both the encoder and the decoder.

An IHS consists of the following ingredients:

Constrained Encoder The encoder is a measurable mapping: $\mathbf{f_e} : \mathcal{M} \times \mathcal{S}^n \to \mathcal{X}^n$, which is subject to a distortion constraint. A distortion constraint is a pair (d, D) of a distortion measure $d : \mathcal{S} \times \mathcal{X} \to R^+$ and a distortion level $D \geq 0$. The encoding mapping should satisfy:

$$E[\frac{1}{n}\sum_{i=1}^{n} d(S_i, \mathbf{f_e}(m, \mathbf{S})_i)] \leq D,$$

where $\mathbf{f_e}(m, \mathbf{S})_i$ is the i-th coordinate of $\mathbf{f_e}(m, \mathbf{S})$.

Blind Attack Channel A stationary and memoryless communication channel with input variable X, output r.v. Y, and a transition probability matrix $p_{Y|X}(y|x)$.

Decoder The decoder is a measurable mapping $\mathbf{f_d} : \mathcal{Y}^n \to \mathcal{M}$. The output of the decoder, denoted m^*, is an estimate of the message m.

An *Encoding-decoding scheme* for an IHS is an ordered quadruple $(\mathcal{M}, \mathbf{f_e}, \mathbf{f_d}, n)$ of a message alphabet, an encoding mapping, a decoder and covertext sequence length. The *rate* of an encoding-decoding scheme $\mathcal{E} = (\mathcal{M}, \mathbf{f_e}, \mathbf{f_d}, n)$ is defined as $R(\mathcal{E}) = \frac{1}{n}\log_2(\mid \mathcal{M} \mid)$. An encoding-decoding scheme $\mathcal{E}$ defines a conditional probability distribution $p_{\mathbf{Y}|M}(\mathbf{y}|m)$ on the random sequence $\mathbf{Y}$ given a message m:

$$p_{\mathbf{Y}|M}(\mathbf{y}|m) = \sum_{\mathbf{s}\in\mathcal{S}^n} \prod_i p_{Y|X}(y_i|\mathbf{f_e}(\mathbf{s}, m)_i) p_S(s_i).$$

Based on the assumption that M is distributed uniformly, the *average error probability* of $\mathcal{E}$ is defined as:

$$p_e(\mathcal{E}) = \frac{1}{\mid \mathcal{M} \mid} \sum_{m\in\mathcal{M}} Pr\{\mathbf{f_d}(\mathbf{Y}) \neq m | M = m\}.$$

A number $R > 0$ is called an achievable rate if for evey $\epsilon > 0$ and all sufficiently large n, there exists an encoding-decoding scheme $\mathcal{E}$ s.t. $R(\mathcal{E}) > R - \epsilon$ and $p_e(\mathcal{E}) < \epsilon$. The *capacity* is defined as the supremum of all the achievable rates and is denoted by C.

A remark on distortion constraints is due at this point: The distortion constraint defined above is a *mean distortion constraint*, since it is defined by the expected value of an average scalar distortion measure. A special case of the mean distortion constraint is the *maximum distortion constraint:* $\max_i d(s_i, \mathbf{f_e}(m, \mathbf{s})_i) \leq D$ a.s.. In order to see that this is indeed a special case, one can observe that

$$\max_i d(s_i, x_i) \leq D \; a.s. \Leftrightarrow E[\frac{1}{n}\sum_{i=1}^{n} 1_{\{d(s_i,x_i)\leq D\}}] = 0,$$

where 1_A denotes the indicator of an event A. Since the maximum distortion constraint is more adequate to certain information hiding application, as explained in the Introduction, we refer specifically to this constraint by the notation $\mathbf{d_S}$, while the general mean distortion constraint is denoted $\mathbf{d_M}$

2.2 Information Hiding Systems and Communication Channels with Side information

In this subsection, we show that the information hiding problem with mean distortion constraint is intimately related to the problem of Channel coding with Side Information - (SIC). The covertext sequence in the IHS plays a similar roll to the side information sequence in the SIC. The SIC problem, depicted at Figure (2), was introduced by Shannon [Sha58], for causal side information, and was later investigated in its general form, i.e. non-causal side information, by many authors (e.g. [GP80]). This relationship enables the derivation of a formula for the capacity of an information hiding problem, based on the analogous formula available for SIC.

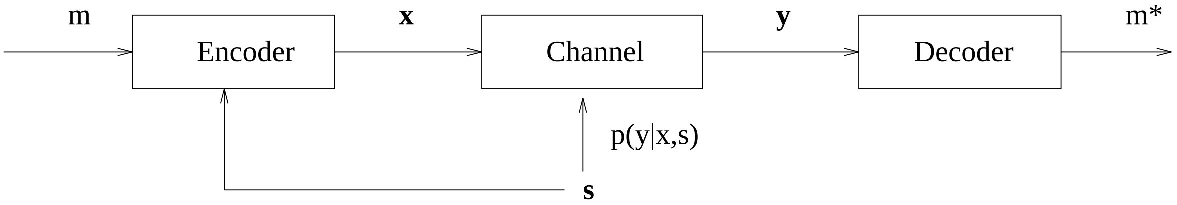

Figure 2: *Channel with side informatio: m - the message,* $\mathbf{s}$ *- the side information sequence,* $\mathbf{x}$ *- the input sequence,* $\mathbf{y}$ *- the channel output sequence,* m^* *- estimate message,* $p(y|x,s)$ *- transition probaility*

Definition 2.1 (A Channel with Side Information - SIC) *A channel with side information is a memoryless channel with input variable* X *and output random variable* Y*, whose transition probability matrix* $\{p_{Y|XS}(y|x,s)\}$*, depends on a side information variable* S*, which is distributed according to a p.d.f.* $p_S(s)$*.*

In this section, we assume that the side information sequence $\mathbf{s}$, consisting of independent samples from the r.v. S, is known to the encoder but is not available to the decoder.

The definition of encoding and decoding schemes, rate, average decoding error probability and channel capacity for a SIC can be found in [GP80]. These definitions are essentially similar to the definitions given in Section 2.1, except for the fact that the channel transition probabilities are dependent on the state r.v. S and the absence of a distortion constraint at the encoder.

The following theorem, by Gelfand and Pinsker, provides a formula for the capacity of a channel with side information.

Theorem 2.1 *[GP80] Let K denote a channel with side information with transition probability matrix $\{p_{Y|XS}(y|x,s)\}$. Denote by $\mathcal{A}$ the set of all triplets (U,S,X) of random variables taking values in $\mathcal{U} \times \mathcal{S} \times \mathcal{X}$ respectively, where ($\mathcal{U}$ is any arbitrarily large finite set) s.t.*

$$\sum_{u,x} p_{USX}(u,s,x) = p_S(s).$$

For every $A \in \mathcal{A}$, the joint distribution of (U,S,X,Y) is defined by:

$$p_{USXY}(u,s,x,y) = p_{USX}(u,s,x)p_{Y|XS}(y|x,s).$$

For every $A \in \mathcal{A}$ define $R(A) = I(U;Y) - I(U;S)$. Then, the capacity of K is given by $\max_{A\in\mathcal{A}} R(A)$.

Remark: As explained in [GP80], it turns out that the capacity-achieving conditional p.d.f. $p_{UX|S}(ux|s) = p_{U|S}(u|s)p_{X|US}(x|u,s)$ is such that $p_{X|US}(x|u,s)$ puts all its mass on one letter $x = f(u,s)$ for some deterministic function f.

The following theorem provides a formula for the capacity of an information hiding system with a maximum distortion constraint. It's similarity to Theorem (2.1) is apparent and its proof is omitted since it is identical in essence to the proof given in [GP80]. A similar theorem for the more general case of IHS, with mean distortion constraint is proved in [BCW00].

Theorem 2.2 *Let H be an information hiding system, with a maximum distortion constraint $\mathbf{d}_S$ and and distortion level D. Denote by $\mathcal{A}$ the set of all triplets (U,S,X) of random variables taking values in $\mathcal{U} \times \mathcal{S} \times \mathcal{X}$ ($\mathcal{U}$ is an arbitrarily large finite set) s.t.*

$$\sum_{u,x} p_{USX}(u,s,x) = p_S(s) \text{ and } \mathbf{d}_S < D \text{ a.s..}$$

For every $A \in \mathcal{A}$, the joint distribution of (U,S,X,Y) is defined by:

$$p_{USXY}(u,s,x,y) = p_{USX}(u,s,x)p_{Y|X}(y|x).$$

For every $A \in \mathcal{A}$ define the rate $R(A) = I(U;Y) - I(U;S)$. Then the capacity of the information hiding system H is equal to $\max_{A \in \mathcal{A}} R(A)$.

The remark that follows Theorem 2.1 continues to be applicable here.

We show now that the information hiding problem with the maximum distortion constraint is, in fact, a special case of the channel coding with side information problem. This is evident from the next lemma in which we construct, for each information hiding problem, an equivalent problem of channel coding with side information. The proof of the lemma is given in Appendix A.

Lemma 2.3 *Let H be an IHS with a maximum distortion constraint, $\mathbf{d_S}$, based on a distortion measure $d(x,s)$, and distortion level D. Denote by $p_{Y|X}(y|x)$ the transition probability matrix of its attack channel. Define a channel with side information K in the following way: The input, output and side information alphabet sets denoted respectively - $\mathcal{X}, \mathcal{Y}, \mathcal{S}$, are identical to those of H. The p.d.f. $p_S(s)$ of the side information r.v. S is equal to the p.d.f of the covertext r.v. of the information hiding system H. Define the transition probability matrix of K as:*

$$p_{Y|XS}(y|x,s) = \begin{cases} p_{Y|X}(y|x) & for \quad d(x,s) < D \\ \tilde{p}_{Y|S}(y|s) & for \quad d(x,s) \geq D \end{cases} \tag{1}$$

where $\tilde{p}_{Y|S}(y|s) = \frac{1}{N(s)} \sum_{x \in X_s} p_{Y|X}(y|x)$, $X_s = \{x \mid d(x,s) < D\}$, $N(s) = \sum_y \sum_{x \in X_s} p_{Y|X}(y|x)$.

Then, $C_H = C_K$.

The formula, provided by Theorems 2.1 and 2.2, for the capacity of a SIC and an IHS is ,in general, very difficult to compute. However, there are certain cases where this computation becomes simpler. These are the cases where the capacity of the system with the side information unknown to the decoder is equal to the capacity of the same system but with the side information available to the decoder.

We cite here a result due to Costa, showing that such an equality exists in the case of additive Gaussian SIC, and then provide a related example of an IHS with maximum distortion constraint.

Theorem 2.4 *[Cos83] Let K be a memoryless, stationary channel with side information, whose output r.v. Y is given by: $Y = X + S + Z$. The side information sequence is a realization of n i.i.d. r.v. , distributed $\mathcal{N}(0, Q)$, and is known only to the encoder. The noise r.v. $Z \sim \mathcal{N}(0, N)$ is independent of S and X. The input alphabet is $\mathcal{X} = \mathbf{R}^1$ and the input sequence $\mathbf{x}$ is constrained by $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$. Then the capacity of K is equal to $\frac{1}{2}\log_2(1+\frac{P}{N})$.*

This theorem implies that the additive Gaussian SIC capacity is independent of Q and is equal to the capacity of the additive Gaussian channel with signal to noise ratio $\frac{P}{N}$. The next corollary to this theorem implies that the same capacity equality holds for the additive Gaussian IHS with quadratic mean distortion constraint.

Corollary 2.5 *Define an IHS in the following way: The covertext sequence is a realization of n i.i.d. r.v. distributed $\mathcal{N}(0, Q)$. The mean distortion constraint is based on the quadratic distortion measure: $d(x, y) = (x - s)^2$. The attack channel is memoryless and stationary, with output $Y = X + Z$, where $Z \sim \mathcal{N}(0, N)$ is independent of S and of X. Then, the capacity of this IHS is equal to the capacity of the additive Gaussian channel with signal to noise ratio of $\frac{P}{N}$ (without distortion constraint), i.e. $\frac{1}{2}\log_2(1 + \frac{P}{N})$.*

The next theorem provides an example of an IHS with a maximum distortion constraint, where the capacity of the system with the covertext signal known to the encoder and the decoder is equal to the capacity of the system where it is known to the encoder only. The proof of the theorem can be found in Appendix B.

Theorem 2.6 *Define an IHS, denoted H, in the following way: The covertext sequence is a realization of n i.i.d. r.v. taking values in $\{0, \ldots, Q - 1\}$, and the stegotext variable X takes values in the same set. H is subject to a maximum distortion constraint $\mathbf{d}_S(\mathbf{x}, \mathbf{s}) = max_i \mid x_i - s_i \mid mod(P)$, where P is a positive integer that divides Q. The memoryless and stationary attack channel has output $Y = (X + Z)mod(Q)$, where the noise Z, independent of S and of X, is distributed uniformly in $\{0, \ldots, Q - 1\}$. Then, the capacity of H is equal to the capacity of a memoryless and stationary channel with input r.v. X that takes values in $\{0, \ldots, P - 1\}$, and output r.v. $Y = (X + Z)mod(P)$, where the noise r.v. Z, independent of X, is distributed uniformly in $\{0, \ldots, Q - 1\}$.*

Theorem 2.4 is the basis for the scaled bin encoding scheme developed in Section 3.3. It is applicable to image transform-domain watermarking since transform domain coefficient distributions are continuous and can be approximated in some important cases by the Gaussian distribution - see Section 5 for more details. Theorem 2.6 can serve equivalently as a basis for spatial domain image watermarking schemes, where the pixel gray level values are in the set $\{0, \ldots, 255\}$, and the change in pixel values can stretch only for few gray levels. The spatial domain based watermark approach was not developed in this research.

3 Binary Modulation Schemes

In this section, we investigate a family of block encoding schemes for information hiding applications, named Binary Modulation Schemes (BMSs). In restricting our attention to this family, we ignore more general encoding schemes that may provide higher embedding rates. However, we gain simplicity and we are able to construct an efficient maximum likelihood decoding algorithm, without significant loss of capacity. Using the notation of Section 2, a BMS can be described as two-stage procedure:

1. The message m is encoded as a binary sequence $\mathbf{b}$, using a binary error correcting code.
2. The sequence $\mathbf{b}$ is modulated, using a scalar binary modulating mapping f, in a stegotext sequence: $\mathbf{x} = (f(b_1, s_1), ..., f(b_n, s_n))$. The modulating mapping $f : \{0,1\} \times \mathcal{S} \to \mathcal{X}$ satisfies the distortion constraint: $d(s, f(b,s)) < D$ for every $b \in \{0,1\}$ and $s \in \mathcal{S}$.

The advantage of the BMSs stems from the fact that they transform the complicated information hiding problem, into a simpler problem of binary channel coding. Indeed, upon selecting a specific modulating mapping f the information hiding problem is reduced to the coding problem of a communication channel K_f with a binary input variable B, an output Y and a transition probability:

$$P^f_{Y|B}(y|b) = \sum_{s \in \mathcal{S}} P_S(s) P_{Y|X}(y|f(b,s)). \tag{2}$$

The capacity of this channel is given by $C_f = \max_{P_B} I_f(B,Y)$ where the maximum is taken over all distributions P_B of the binary input variable B, and I_f denotes mutual information induced by f. One can appreciate that this formula is much simpler than the one provided by Theorem (2.2), although there is no guarantee that it achieves the same optimal value. It follows that given an information hiding problem one can construct a suitable BMS by:

1. Selecting an modulating mapping f^{opt} that maximizes, or nearly maximizes, the capacity C_f.
2. Choosing a binary error correcting code for the channel $K_{f^{opt}}$.

We derive now a numerically computable expression for the capacity C_f. For the sake of brevity, we denote $P^f_b(y) = P^f_{Y|B}(y|B=b)$, and $p = P_B(0)$, $q = 1 - p = P_B(1)$. The explicit dependence of P^f_b on the mapping f is given by equation (2). The mutual information of B and Y is given by:

$$I_f(B,Y) = \sum_{b,y} P^f_{BY}(b,y) \log \frac{P^f_{BY}(b,y)}{P^f_B(b) P^f_Y(y)},$$

where logarithm here and throughout the sequel are taken to the base 2. Substituting $P^f_{BY}(b,y) = P^f_B(b)P^f_{Y|B}(y|b)$, the mutual information can be written as:

$$I_f(B,Y) = \sum_y pP^f_0(y)(\log P^f_0(y) - \log(P^f_Y(y))) + qP^f_1(y)(\log P^f_1(y) - \log(P^f_Y(y))).$$

Substituting $P^f_Y(y) = pP^f_0(y) + qP^f_1(y)$ and performing some algebraic manipulations one obtains:

$$I_f(B,Y) = H(p) - p\sum_y P^f_0(y)\log\left(1 + \frac{qP^f_1(y)}{pP^f_0(y)}\right) - q\sum_y P^f_1(y)\log\left(1 + \frac{pP^f_0(y)}{qP^f_1(y)}\right), \quad (3)$$

Using this expression for the mutual information, the capacity $C_f = \max_p I_f(B,Y)$ can be calculated numerically. In cases where an analytic approach is required, the following upper and lower bounds, derived in Appendix C, are easier to compute:

$$\frac{1}{8\ln 2} \parallel P^f_0 - P^f_1 \parallel^2 \leq C_f \leq \frac{D^f_0 D^f_1}{D^f_0 + D^f_1}, \quad (4)$$

where $\parallel \cdot \parallel$ is the Euclidean norm, $D^f_0 = D(P^f_0 \parallel P^f_1)$ and $D^f_1 = D(P^f_1 \parallel P^f_0)$. Formula (3) and the bounds in (4) provide the computational means for comparing the theoretical performance of different modulating functions. See Appendix D for an example of such a comparison for a certain type of BMSs.

3.1 Modulating mappings for BMS

A modulating mapping $f(b,s)$ can be represented as a pair of modulating functions $f_0(s) = f(0,s)$ and $f_1(s) = f(1,s)$. In designing such a pair of modulating functions, one should take into account the following two (conflicting) considerations:

1. Both f_0 and f_1 should satisfy the maximum distortion constraint $d(f_b(s),s) \leq D$ for every $b \in \{0,1\}$ and $s \in \mathcal{S}$.
2. For each value of s, the encoding values $f_0(s)$ and $f_1(s)$ should be as separated as possible, in order to minimize the probability of decoding error.

Two types of modulating mappings can be derived from information hiding schemes suggested in the literature:

1. Cox *et. al.* [CKLS96] suggested a pair of skewed perturbation functions (PF's) such as

   $$f_b(x) = x + (-1)^b \epsilon \text{ or } f_b(x) = x[1 + (-1)^b \epsilon],$$

   where ϵ is a small constant. Other versions of perturbation functions were proposed later, see for example [PBBC97].

2. Chen and Wornel [CW99] suggested a scheme named Dither Modulation (DM) that utilizes a pair of skewed quantizers q_0 and q_1, e.g.

   $$q_b(s) = q(s + d_b) - d_b,$$

   where $q(s)$ is a scalar quantizer and d_0, d_1 are constants. The quantizers are constructed in such a way that the distortion constraint is satisfied.

The main drawback of PF and DM schemes is that they do not explicitly use the knowledge of the attack channel statistics. Hence they are not adaptive to the features of the noise introduced to the stegotext signal. We will show, in section 5, that the DM scheme performs well, as long as the ratio of noise power to allowed distortion level is low, but collapses otherwise. The PF schemes are more stable but have inferior performance in the low noise case. In Section 3.3, we propose a novel BMS named scaled bin encoding which is adaptive to the statistics of the attack channel and is superior to both the DM and PF schemes. In spite of the performance deficiency of the PF schemes, when compared to the SBE scheme, they are simpler to implement, and are more robust to inaccuracies of the statistical models. Hence, we investigated few variants of Cox's original schemes, using the theoretical tools developed at the beginning of Section 3. A short account of this investigation is available at Appendix D.

Regardless the choice of the specific modulating functions f_0 and f_1 a BMS can be decoded using a Maximum Likelihood (ML) decoder. The ML decoding has theoretical and practical advantages over the correlation decoding procedures previously suggested in [CKLS96, PBBC97]. In the next section, we show, based on familiar properties of the Hadamard transform [Lem79], that when the set of error correcting codewords constitute a linear space, a ML decoding procedure can be efficiently computed, using the Hadamard transform.

3.2 Maximum Likelihood Computations with Fast Hadamard Transform

In this section, we exploit the fact that the Hadamard transform can be utilized for fast calculations over linear spaces [Lem79]. We comment that the computational procedure described in this section does not uses the stationarity of the covertext sequence and the attack channel. Following the ML paradigm, we write the probability of observing the output sequence $\mathbf{y}$ assuming that the encoded message $\mathbf{b}$ has been transmitted as:

$$P(\mathbf{y}|\mathbf{b}) = \prod_{i=1}^{n} P_i(y_i|b_i). \tag{5}$$

Hence, for a fixed output sequence $\mathbf{y}$, a ML decoding procedure amounts to computing an expression of the form:

$$\mathbf{b}^* = \arg\max_{\mathbf{b}\in\mathcal{B}} \prod_{i=1}^{n} P_i(y_i|b_i), \tag{6}$$

where $\mathcal{B}$ is the set of codewords. Note that $P_i(y_i|b_i)$ depends only on b_i, y_i, and on the statistical model of the IHS. We show now that if the code is linear, namely the set $\mathcal{B}$ is a linear subspace of $\{0,1\}^n$, then one can use the fast Hadamard transform in order to efficiently compute b^*.

It is clear that in expression (6) the product over $P_i(y_i|b_i)$ can be replaced by a sum over $\log P_i(y_i|b_i)$. Writing $\log P_i(y_i|b_i) = \frac{1}{2}(\log P_i(y_i|0) + \log P_i(y_i|1)) + (-1)^{b_i}\frac{1}{2}(\log P_i(y_i|0) - \log P_i(y_i|1))$ and noticing that the first expression on the right hand side does not depend on the sequence $\mathbf{b}$, one concludes that:

$$\mathbf{b}^* = \arg\max_{\mathbf{b}\in\mathcal{B}} \sum_{i=1}^{n} (-1)^{b_i} w_i, \tag{7}$$

where $w_i = \frac{1}{2}(\log P_i(y_i|0) - \log P_i(y_i|1))$.

Since $\mathcal{B}$ is a linear subspace, each sequence $\mathbf{b} \in \mathcal{B}$ can be written as a linear combination of some basis vectors $\{\mathbf{e}^j\}_{j=1}^m$, where $m = \dim(\mathcal{B})$. Using this basis, $\mathcal{B}$ can be represented as the set of all linear combinations of its basis vectors: $\mathcal{B} = \{\sum_{j=1}^m c_j \mathbf{e}^j\}_{\mathbf{c}\in\{0,1\}^m}$. Substituting this representation in (7) one obtains:

$$\mathbf{c}^* = \arg\max_{\mathbf{c}\in\{0,1\}^m} \sum_{i=1}^{n} (-1)^{\sum_{j=1}^m c_j e_i^j} w_i,$$

where $\mathbf{c}^*$ is the representation of $\mathbf{b}^*$ in the basis $\{\mathbf{e}^j\}_{j=1}^m$. The sum over the index i can be re-organized in such a way that indices with common $\{e_i^j\}_{j=1}^m$ configuration are grouped together, yielding: $\arg\max_{\mathbf{c}\in\{0,1\}^m}\sum_{\mathbf{a}\in\{0,1\}^m}\sum_{\{i:\{e_i^j\}_{j=1}^m=\mathbf{a}\}}(-1)^{\mathbf{ac}}w_i$. Denoting $\bar{w}_{\mathbf{a}} = \sum_{\{i:\{e_i^j\}_{j=1}^m=\mathbf{a}\}} w_i$ the previous expression can be written as:

$$\mathbf{c}^* = \arg\max_{\mathbf{c}\in\{0,1\}^m} \Big(\sum_{\mathbf{a}\in\{0,1\}^m} (-1)^{\mathbf{ac}}\bar{w}_{\mathbf{a}} \Big). \tag{8}$$

It is easily verified that the expression $\sum_{\mathbf{a}\in\{0,1\}^m}(-1)^{\mathbf{ac}}\bar{w}_{\mathbf{a}}$ is the m-order Hadamard transform of the vector $\mathbf{v} \in \mathbf{R}^{2^m}$ defined by $v_{i(\mathbf{a})} = \bar{w}_{\mathbf{a}}$, where $i(\mathbf{a})$ is the integer whose binary representation is $\mathbf{a}$. The optimal sequence $\mathbf{c}^*$ of (8) is found by computing the fast Hadamard transform of the vector $\mathbf{v}$, defined above, and then choosing the entry that yields the maximum value. The optimal sequence $\mathbf{b}^*$ of the original ML expression is now obtained by substituting $\mathbf{c}^*$ in the expansion of $\mathbf{b}^*$ which yields: $\mathbf{b}^* = \sum_{j=1}^m c_j^* \mathbf{e}^j$. Utilizing the fast Hadamard transform reduces the complexity of the ML decoding to $O((2^m + n)m)$, while direct calculation has an $O((2^m)^2 + nm)$ complexity.

3.3 Scaled Bin Encoding

In this section, we propose a BMS, named Scaled Bin Encoding (SBE), which is adaptive to the statistics of the attack channel. This scheme is motivated by [Cos83], where a capacity-achieving random coding scheme for the additive Gaussian side information channel is suggested. The SBE scheme, though provably optimal only for the additive Gaussian information hiding channel with mean quadratic distortion constraint, yields superior results when applied to synthetic data and real imagery. Independently of our work [LM00], a modified version of the DM scheme named "distortion compensation", which has some resemblance to the SBE scheme, was suggested in [Che00].

The additive Gaussian SIC, investigated in [Cos83], has output $Y = X + S + Z$, where S, Z are independent r.v.'s, distributed according to $\mathcal{N}(0, Q)$ and $\mathcal{N}(0, N)$, respectively. The input sequences, in this case, are subjected to the power constraint: $\sum_{i=1}^n x_i^2 \leq P$. It is shown in [Cos83] that the capacity of this SIC is: $C = \frac{1}{2}\log(1 + \frac{P}{N})$. The proof is based on a random coding scheme where codewords are chosen as typical sequences corresponding to a normal r.v. $U = \alpha S + W$, where $W \sim \mathcal{N}(0, P)$ is independent of S and the optimal value of the parameter α is shown to be $\frac{P}{P+N}$. The codewords are then equally distributed into approximately 2^{nC} bins s.t. each bin represents a message. Given a side information sequence $\mathbf{s}$ and a message m, the corresponding codeword $\mathbf{u}(m, \mathbf{s})$ is chosen from the m related bin in such a way that the sequence $\mathbf{u}(m, \mathbf{s}) - \alpha\mathbf{s}$ is approximately orthogonal to $\mathbf{s}$. The sequence transmitted through the channel is $\mathbf{x} = \mathbf{u}(m, \mathbf{s}) - \alpha\mathbf{s}$.

Consider now the additive Gaussian IHS with output r.v. $Y = X + Z$, where S, Z are defined as above. We assume a mean square distortion constraint $\mathbf{d_m} = E(\frac{1}{n}\sum_{i=1}^n (s_i - x_i)^2) \leq P$. Note that the input variable X in the IHS case is equivalent to the sum $X + S$ in the SIC case. Therefore, a capacity-achieving coding scheme is derived from the SIC scheme described above by choosing a similar set of codewords. Choosing the codeword $\mathbf{u}(m, \mathbf{s})$ is done as above, but the transmitted sequence in this case is $\mathbf{x} = \mathbf{u}(m, \mathbf{s}) + (1 - \alpha)\mathbf{s}$.

A deterministic version of this random coding scheme can be constructed by choosing a well designed set of $\mid \mathcal{M} \mid$ vector-quantizers $\mathbf{q}(m, \mathbf{s})$, where $\mathbf{q} : \mathcal{M} \times \mathcal{S}^n \to \mathcal{X}^n$, and then setting $\mathbf{u}(m, \mathbf{s}) = q(m, \alpha\mathbf{s})$. This type of scheme is named Scaled Bin Encoding (SBE). Given a covertext sequence $\mathbf{s}$ and a message m the SBE scheme is defined by the stegotext sequence:

$$\mathbf{x} = \mathbf{s} + (\mathbf{q}((m, \alpha\mathbf{s}) - \alpha\mathbf{s}), \tag{9}$$

where α is the scaling parameter.

A BMS implementation of this SBE scheme is defined in the following way:

$$x_i = s_i + (q^{b_i}(\alpha s_i) - \alpha s_i),$$

where (q^0, q^1) is a pair of scalar quantizers and $\mathbf{b} = (b_1, \ldots, b_n)$ is the binary encoding sequence of the message m. The scalar quantizers are chosen according to the considerations mentioned in the beginning of this section, e.g. a pair of shifted scalar quantizers. If all the assumptions made in [Cos83] were satisfied, the scaling parameter should be taken as $\alpha = \frac{P}{P+N}$. However, certain independence assumptions may not be valid and hence a more careful examination, presented in Appendix E, yields the corrected formula:

$$\alpha = \frac{P - \frac{M(M+N)}{Q}}{P + N + \frac{M^2}{Q}}, \tag{10}$$

where $M = E(S(X - S))$. Note that for $M = 0$, equation (10) reduces to $\alpha = \frac{P}{P+N}$.

As mentioned in Section 3.2 the BMS implementation of the SBE information hiding scheme can be decoded by an efficient decoder. To this end, the probabilities $P(\mathbf{y}|\mathbf{b})$ for the SBE modulating mapping should be computed. These probabilities can be approximated by:

$$P(\mathbf{y}|\mathbf{b}) = \prod_{i=1}^{n} P(y_i|b_i) = \prod_{i=1}^{n} \sum_{u \in Q^{b_i}} P_U(u) P_{Y|U}(y|u), \tag{11}$$

where $\mathcal{Q}^b$ is the set of representation levels that the quantizer q^b admits. By substituting the probabilities $P_U(u)$ and $P_{Y|U}(y|u)$, computed in Appendix E (see Equations (21) and (20)), in (11) the following explicit expression for $P(\mathbf{y}|\mathbf{b})$ is obtained:

$$P(\mathbf{y}|\mathbf{b}) = \prod_{i=1}^{n} \sum_{u \in Q^{b_i}} P_{\mathcal{N}(0,V_U)}(u) P_{\mathcal{N}(0,V_T)}(y_i - \gamma u), \tag{12}$$

where, $P_{\mathcal{N}(0,V)}(\cdot)$ denotes the Gaussian probability density function with zero mean and variance V. The parameters γ, V_U, V_T appearing in (12) are calculated in Appendix E: $\gamma = \frac{Q+N+P}{\alpha Q+N+P}$, $V_T = \frac{Q+P+N}{\alpha Q+P+N} N$ and $V_U = \alpha^2 Q + P$, see (19), (20), (21) respectively.

4 Application of BMS to Image Watermarking

The BMS information hiding schemes described in Section 3 can be applied to transform domain image watermarking, where a selected subset of transform coefficients constitutes the covertext sequence. The main burden of such an application is due to the fact that the features of the transform domain coefficient statistics and the features of the noise are not fully known. Furthermore, these features vary in the image transform domain and may differ from one image to another. Hence, an image watermarking implementation of BMS requires the construction of a reliable statistical model, whose parameters are either stable across a wide range of images or can be estimated from the image data. In this section, we present a general framework for the required statistical model, general procedures for encoding and decoding, and a statistical model for a BMS watermarking implementation in the DCT domain.

A statistical model for a BMS watermarking application in a transform domain should include the following ingredients:

1. A statistical model for the transform coefficients, that is, a probability law of the covertext sequence.
2. A model for the human visual system sensitivity, from which the distortion measure $d(x, s)$ and the distortion level D are derived.
3. A statistical model of the noise, namely, a conditional probability distribution $P_{Y|X}$ of the attack channel.

4.1 A Procedure for BMS Watermarking

Based on a statistical model with the previously described ingredients, a BMS encoding and decoding procedures can be described in the following way:

Encoding - The encoder, depicted in Figure 3, receives the covertext image **i** and the message m and outputs the watermarked image $\mathbf{i}'$. The watermarking encoding procedure comprises the following stages:

1. Compute an invertible transform domain representation of the image (e.g. DCT, FFT, DWT) denoted by $t = F(\mathbf{i})$. Choose a subset of n transform coefficients to be watermarked. This coefficient sequence is denoted $\mathbf{s} = (s_1, \ldots, s_n)$.

2. Encode the message m as a sequence of n bits denoted $\mathbf{b} = (b_1, \ldots, b_n)$, using an error correcting code that maps each block of l message bits into a block of k bits where $k > l$. For the sake of simplicity we assume that k divides n.
3. Modulate the binary sequence $\mathbf{b}$ within the coefficient sequence $\mathbf{s}$ using the BMS procedure described below and obtain the watermarked coefficients sequence $\mathbf{x}$. The BMS procedure is performed repeatedly for each block of k coefficients until the sequence $\mathbf{s}$ is exhausted.
4. Replace the coefficient sequence $\mathbf{s}$ in the transformed image $\mathbf{t}$ with the watermarked coefficient sequence $\mathbf{x}$ and denote the resulting watermarked transform representation $\mathbf{t}'$.
5. Invert the watermarked transform domain representation and obtain the watermarked image $\mathbf{i}' = F^{-1}(\mathbf{t}')$.

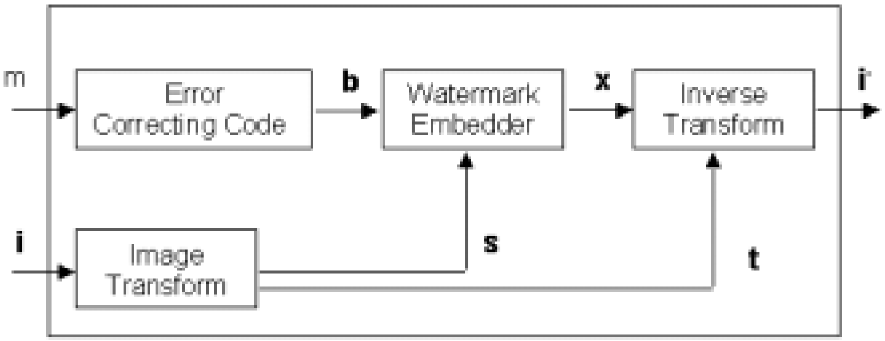

Figure 3: *Block diagram of the encoder*

Binary Modulation Scheme procedure - The procedure receives a sequence $\mathbf{b} = (b_1, \ldots, b_k)$ of k bits and a covertext sequence $\mathbf{s} = (s_1, \ldots, s_k)$ of k transform coefficients. The procedure constructs, for each coefficient s_i, a pair of SBE functions:

$$f_i^b(s_i) = s_i(1-\alpha) + (q_i^b(\alpha s_i)),$$

where $b \in \{0, 1\}$. The quantizers q_i^b and the parameter α are determined, as in Appendices E and F, based on estimations of the relevant statistical model parameters. The index i indicates that these estimations may depend on the index of the coefficient s_i. (In Section 4.2 we demonstrate an estimation technique for these parameters, applicable to DCT coefficients). The procedure outputs the stagotext sequence $\mathbf{x} = (f_1^{b_1}(s_1), \ldots, f_n^{b_n}(s_n))$.

Decoding - The decoder, depicted in Figure 4, receives the attacked version of the image, denoted $\mathbf{i}^*$, and estimates the encoded message using the following procedure:

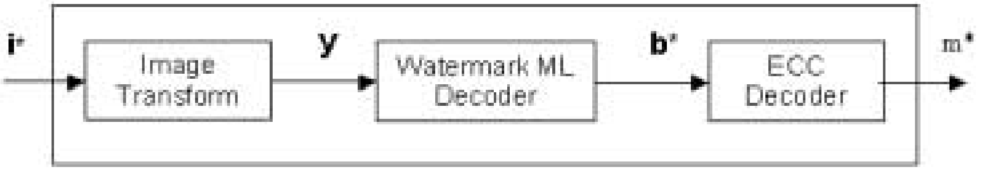

Figure 4: *Block diagram of the decoder*

1. Compute the same transform domain representation of the image $\mathbf{i}^*$ and choose the same subset of transform coefficients used in the encoding scheme. This subset is denoted $\mathbf{y} = (y_1, \ldots, y_n)$.
2. Extract the binary sequence $\mathbf{b}^*$ from $\mathbf{y}$ using the ML decoding procedure described below. The ML decoding procedure is performed repeatedly for each block of k coefficients until the sequence $\mathbf{y}$ is exhausted.
3. Decode the binary sequence $\mathbf{b}^*$ and obtain the estimate message m^*.

ML decoding procedure - The ML decoding procedure receives a coefficient sequence $\mathbf{y} = (y_1, \ldots, y_k)$ and outputs the decoded bit sequence $\mathbf{b}^* = (b_1^*, \ldots, b_k^*)$ using the procedures described in Sections 3.3 and 3.2. For each possible codeword $\mathbf{b}$ (where $\mathbf{b}$ ranges over the 2^l possible codewords) the procedure computes the ML score:

$$Sc(\mathbf{b}) = P\{\mathbf{y}|\mathbf{b}\} = \prod_{i=1}^{k} P_i(y_i|b_i)$$

and chooses the code word $\mathbf{b}$, that maximizes it, as the estimate sequence $\mathbf{b}^*$. The probability $P_i(y_i|b_i)$ are computed according to Formula (12) in Section 3.3. The parameters of the ML formula are derived in the same way described in the BMS procedure above, except for the variance Q_i which is estimated from the variance of y_i and the given statistical model. See Section 4.2 for an example of such an estimation for DCT coefficients.

4.2 A Statistical Model for DCT Domain Image Watermarking

In this section, we describe a statistical model of a DCT domain watermarking which is used in the implementation of the SBE scheme described in Section 5.2. We have chosen to work with the DCT, since the distribution of the DCT coefficients can be approximated by the Gaussian p.d.f.. However, the same techniques can be applied to other transforms such as the DFT or the DWT with appropriate statistical models. We start with the statistical model for the DCT domain, outlined according to the framework detailed in the beginning of Section 4.

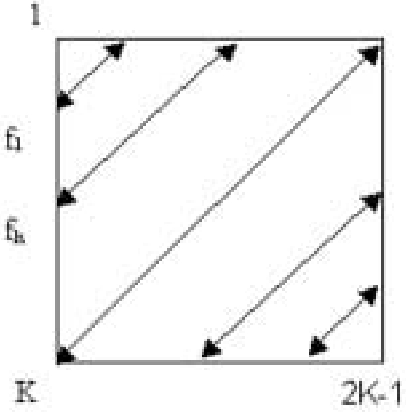

Figure 5: *DCT domain diagram*

DCT coefficients statistical model Given a $K \times K$ gray-level image, its DCT transform is a coefficient array of dimension $K \times K$, as well. The DCT coefficients are indexed by their spatial frequency indices (ν_x, ν_y), where ν_x is the horizontal index and ν_y is the vertical index, both taking on values from 0 to $K-1$. In the DCT domain, the upper-left corner constitutes the lowest frequency $(0,0)$ and the lower-right corner constitute the highest frequency $(K-1, K-1)$. We approximate the DCT coefficients distribution by a Gaussian distribution with zero mean and variance that depends on the radial frequency $\nu_r = \sqrt{\nu_x^2 + \nu_y^2}$. For the sake of simplicity we substitute the non integer radial frequencies indices with corresponding off-diagonal indices in the DCT domain and hence the frequency index ν will actually denote an off-diagonal index. Note that the off-diagonal indices ranges from 1 - the upper-left corner, through K - the main diagonal, to $2K-1$ - the lower-right corner (see Figure (5) for graphical illustration). The DCT coefficients, denoted c_i, are ordered according to their frequency, in such a way that coefficients which correspond to the same off-diagonal have consecutive indices. The set of coefficients corresponding to the off-diagonal indexed ν are denoted $\mathbf{S}_\nu$ and their number denoted $\mid \mathbf{s}_\nu \mid$. Their variance is estimated by:

$$Q_\nu = \frac{1}{\mid \mathbf{s}_\nu \mid} \sum_{c_i \in \mathbf{s}_\nu} c_i^2.$$

Sensitivity of the human visual system A simple model for the sensitivity of the human visual system assumes an absolute difference distortion measure in the frequency domain $d_u(s,x) =\mid s-x \mid$. However, the allowed distortion level may change with the frequency, so that for a DCT coefficient of frequency ν, the absolute difference distortion measure is given by:

$$d_u(c,\tilde{c}) < \epsilon\sqrt{Q_\nu},$$

where ϵ is a constant with value between 0.1 and 0.2 for the middle and high frequency coefficients. For low frequency coefficients ϵ is smaller, since the visual system is more sensitive to those frequencies.

In a more accurate model (see e.g. [PZ98]), the distortion measure is the absolute difference measure only for small coefficient values, but for larger values it is inversely scaled by the absolute value of the coefficient. This "adaptive" distortion measure can be described by:

$$d_a(s,x) = \frac{\mid s-x \mid}{\max(\mid s \mid, \delta)}$$

where the parameter δ, as well as the distortion level, may depend on the frequency.

The statistical model of the noise We have measured the effect of standard image processing operations, such as filtering, scaling and compression on the statistical properties of the DCT coefficients. We have found out that the most complicated statistical effect is generated by compression and that a statistical model for the distortion due to compression in the DCT domain is general enough to describe also the other forms of noise. We have, therefore, concentrated on developing a noise model for JPEG compression. Denoting $\mathbf{y}_\nu$ the set of coefficients of the JPEG-compressed image corresponding to the set $\mathbf{s}_\nu$ of the uncompressed image, we assumed a linear model:

$$\mathbf{y}_\nu = a(\nu)\mathbf{s}_\nu + Z_\nu,$$

where $a(\nu)$ is a deterministic coefficient, and Z_ν, is a zero-mean noise r.v., independent of $\mathbf{s}_\nu$. The coefficients $a(\nu)$ and the noise variances $Var(Z_\nu) = N_\nu$ were estimated, for certain values of the JPEG quality parameter, using a linear regression procedure. A typical example is presented in Figure 6. It turns out that for each JPEG quality level parameter, there is a range of medium frequencies for which $a(\nu)$ is close to 1, and hence the corresponding coefficients are the preferred candidates for watermarking, as explained below.

The above described statistical model suggests that only medium frequency coefficients should be watermarked. The reason is the following: The number of low frequency coefficients is relatively small and the human visual system is very sensitive to perturbation therein, hence they hardly contribute to the overall capacity. The high frequency coefficients

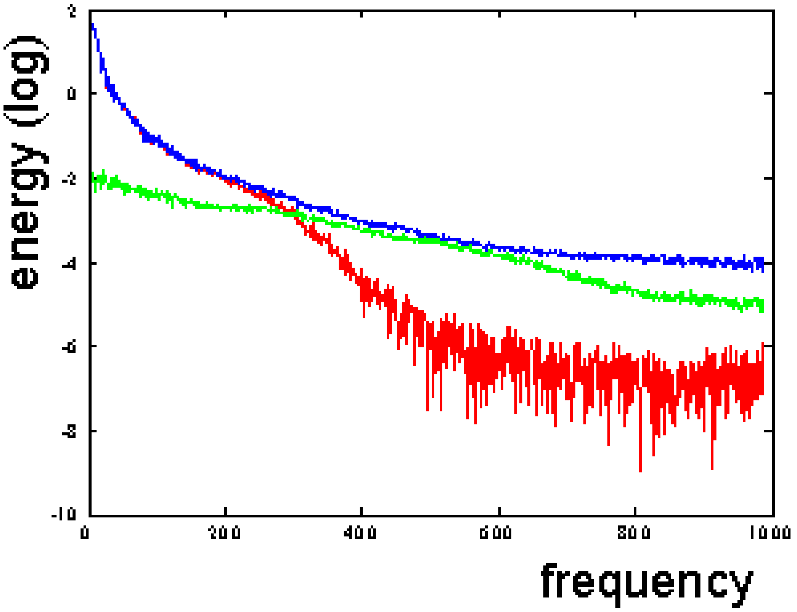

Figure 6: *Statistical model of JPEG compression noise in the DCT domain. The blue graph describes the energy of the original signal* ($\mathbf{s}_\nu$)*, the red graph describes the energy of the attenuated signal* ($a(\nu)\mathbf{s}_\nu$) *and the green graph describes energy of the noise signal* ($\mathbf{z}_\nu$)*.*

are practically filtered out by the JPEG compression and therefore, a watermark signal introduced into these coefficients is sharply attenuated. In other words, the signal to noise ratio for this coefficients is so small that they have only negligible contribution to the overall capacity. Consequently, the medium frequency DCT coefficients, for which the factor $a(\nu)$ is close to one are the only candidates for watermarking. Note that the average power of this set of coefficients is significantly higher than the power of the JPEG compression noise and the power of the allowed watermarking signal. Therefore, the decoder can estimate the statistical parameters of the original image DCT coefficients from the statistics of the watermarked and compressed image.

5 Experimental Results

The BMS information hiding technique has been tested on synthetic data and on real imagery data. In Section 5.1, we describe the experimental results for the additive Gaussian information hiding system. Applications to DCT domain image watermarking are described in Section 5.2.

5.1 Additive Gaussian Information Hiding Channel

In these experiments, the random covertext sequences were generated from i.i.d. samples of a Gaussian r.v. with zero mean and unit variance. The absolute difference distortion measure with distortion level $D = 0.1$ was used for the distortion constraint. The following three BMS information hiding schemes were tested:

SBE - Scaled Bin Encoding scheme: The SBE scheme, described in Section 3.3, was implemented using a pair of uniform scalar quantizers (see Appendix F) and the parameter α calculated according to Formulas (16) in Appendix E.

DM - Dither Modulation scheme: The DM scheme, described in Section 3, was implemented using the same pair of uniform scalar quantizers used for the SBE scheme.

PF - Perturbation Function scheme: The Perturbation Function scheme with modulating functions, described by the item denoted "PF4" in Appendix D, was implemented.

Table 1: *Decoding error probabilities of the Gaussian IHS with low signal to noise ratio S/N=0.5 and capacity: 0.29 bits per channel use. (A '*' denotes a very high error probability.)*

Method/Rate	*7/64*	*8/128*	*9/256*	*10/512*	*11/1024*	*13/4096*
SBE	0.18232	0.0216	0	0	0	0
DM	*	*	*	*	*	0.69512
PF	*	*	0.046429	0.0005102	0	0

Table 2: *Decoding error probabilities of the Gaussian IHS with medium signal to noise ratio S/N=1.0 and capacity: 0.5 bits per channel use. (A '*' denotes a very high error probability.)*

Method/Rate	*6/32*	*7/64*	*8/128*	*9/256*	*10/512*	*11/1024*
SBE	0.070968	0.0036426	0	0	0	0
DM	*	*	*	*	0.0015337	0
PF	*	*	*	0.025255	0	0

Table 3: *Decoding error probabilities of the Gaussian IHS with high signal to noise ratio S/N=2.0 and capacity: 0.79 bits per channel use. (A '*' denotes a very high error probability.)*

Method/Rate	*6/32*	*7/64*	*8/128*	*9/256*	*10/512*
SBE	0.00063906	0	0	0	0
DM	*	0.0023006	0	0	0
PF	*	*	*	0.0005102	0

The performance of each information hiding scheme was tested under three levels of noise power (variance). Random noise sequences were generated from i.i.d. samples of a Gaussian r.v. with zero-mean and variance: $0.2, 0.1, 0.05$, corresponding to signal to noise ratios of $0.5, 1, 2$ respectively. For each noise level, 10^3 binary random sequences, comprising of 10^3 bits each, where embedded using the appropriate BMS and subsequently decoded using the fast ML procedure described in Section 3.2. The resulting error probabilities are summarized in Tables 1, 2 and 3.

The specific values of information embedding rates appearing in these tables are due to the ECC used in these tests. The error correcting codewords we have used were the rows of the Hadamard matrices of varoius orders, together with their binary complementary sequences. The rate - R - of the code is determined from the order - h - of the Hadamard matrix by the relation: $R = \frac{h+1}{2^h}$, where in our experiments $h = 5, \ldots, 12$.

It is evident that the SCE scheme outperforms the DM and PF schemes at all signal to noise (S/N) categories. The PF method is compatible only at low S/N ratios and the DM method performs well at high S/N ratios but collapses at low S/N ratios. Notice also, that the practical embedding rates (the rates where the measured error probability is 0) are far below the theoretical rates which are specified in the relevant tables.

5.2 DCT Domain Image Watermarking

In this section, we present the results of applying the BMS procedure described in Section 4.1, based on the statistical model presented in Section 4.2. The experiments consist of embedding 100 randomly generated sequences of 10^3 bits, in each one of a 10 gray scale images, using the SBE scheme. We used the encoding and decoding procedures described in Section 4.1 with the Hadamard ECC described in Section 5.1. The attack channel consists of $1 : 10$ rate JPEG compression. The decoding error probabilities were measured for various information embedding rates up to the rate that yields zero probability. The results are summarized in Table 4, where one can observe that the SBE scheme performance are slightly lower than its performance when applied to the additive Gaussian IHS with equal signal to noise ratios. The DM and PF schemes performs poorly on the same gray scale images, achieving zero error probability only at the rate of 11/1024 bits per channel use, which is less than $\frac{1}{3}$ of the equivalent SBE rate.

In these experiments, the modulating functions for the SBE scheme were based on a pair of adaptive quantizers. These quantizers preserve the visual quality of the image better than the uniform quantizers, and hence are preferable for image watermarking applications. The exact definitions of these quantizers appear in Appendix F. The parameters of the adaptive quantizers and other encoding and decoding parameters depends on the allowed distortion level and on the distortion level to noise power ratio. In order to determine these factors, we measured, for each image, the average power of the noise introduced to the selected set of coefficients, by the $1 : 10$ rate JPEG compression procedure. The allowed distortion level of the watermark signals was set to this measured noise power, so that in average the S/N ratio was equal to 1. For all the test images, this watermark signal power produced a

Table 4: *Decoding error probabilities for real imagery watermarking with signal to noise power S/N=1.0 and capacity 0.5 bits per channel use.*

Image/Rate	*7/64*	*8/128*	*9/256*
Bridge	0.023	0.0001	0
Flower1	0.0012	0	0
Flower2	0.001	0	0
Gondola	0.0048	0.0003	0
House	0.0077	0.0006	0
Lena	0.0006	0	0
Leopard	0.0049	0.0003	0
Tiffany	0.0012	0.0011	0
Tiger	0.0014	0.0004	0
Zebras	0.0088	0	0

watermarked image that was perceptually identical to the original image. We comment that this parameter tuning strategy is applicable to a broad range of watermarking applications, where the noise features are unknown, but one can expect a ”worst case” noise level. The features of the ”worst case” noise can be measured and the watermarking scheme parameters are tuned accordingly.

6 Discussion and Conclusion

We have presented the theoretical foundation for information hiding and its connection to channel coding with side information. Based on this foundation, a theoretical result by Costa, pertaining to the additive Gaussian channel with side information, inspired an information hiding scheme which has superior performance in comparison to previously proposed information hiding techniques.

We proposed the binary modulation schemes as an information hiding framework where error correcting codes are combined with scalar modulation. This framework allows for an easy application of the spread spectrum methods suggested by Cox *at. al.* [CKLS96] as well as more sophisticated watermarking schemes such as the dither modulation scheme suggested by Chen and Wornel [CW99] and the scaled bin encoding proposed in this work. We developed a computable expression together with upper and lower bounds for the binary modulation schemes information capacity. We showed that a binary modulation scheme can be decoded with efficient maximum likelihood procedure, based on the Hadamard transform.

We proposed the scaled bin encoding scheme, which is based on [Cos83], where a capacity-achieving random coding scheme, for the additive Gaussian side information channel, is presented. The scaled bin encoding scheme is adaptive to the features of the expected noise in the attack channel, and has lower decoding error probability at higher information embedding rates in comparison to previously proposed information hiding schemes.

We presented experimental results that compare binary modulation scheme implementations of three watermarking methods: perturbation function, dither modulation and scaled bin encoding. These approaches were tested for the additive Gaussian information hiding channel with various signal to noise parameters. In all cases the scaled bin encoding out performs the other techniques. On real imagery data we applied the scaled bin encoding scheme to medium frequency DCT coefficients and found its information embedding capacity. To achieve best results we tuned the parameters of the scaled bin encoding scheme to a statistical model of JPEG compression and to an analysis of the human visual system sensitivity.

We propose the following directions in which this work can be further developed:

1. Searching for a technique that, given the parameters of the information hiding problem, finds the binary modulating mapping that maximizes the information hiding capacity. This goal can be achieved by solving the optimization problem defined by equation (3).
2. Developing statistical models for image transform domains other than the DCT, e.g. DWT, FFT, sub-images DCT etc. The goal is to find the transform that provides robust statistical features and is less affected by common image processing operations.

7 Apendices

A Proof of Lemma 2.3

Proof: We show that for each encoding and decoding scheme for H there is a corresponding scheme of K with the same rate and the same (or smaller) decoding error probability and vice versa. Hence the capacities of these two channels are equal.

It is clear that every encoding and decoding scheme for H is an encoding and decoding scheme for K with the same average decoding error probability. To prove the converse, let $\mathcal{E}(\mathcal{M}, F, \tilde{F}, n)$ be an encoding and decoding scheme for K . If there exist code words $F(m, \mathbf{s})$ that do not satisfy the constraint $\mathbf{d}_{\mathbf{S}}(F(m, \mathbf{s}), \mathbf{s})$, then they are not acceptable for an encoding and decoding scheme for H. We show, however, that if this is the case, then these code words can be modified in such a way that they do satisfy the constraint and the resulting encoding and decoding scheme has a lower average error probability and the same rate. The modified scheme is an acceptable encoding and decoding scheme for H and thus the lemma is proved.

Let $m \in \mathcal{M}$ be a message encoded by the sequences $F(m, \mathbf{s})$. Let $A^m = \{\mathbf{y} \mid \tilde{F}(\mathbf{y}) = m\}$ be the decoding region of the message m. Assume that the sequence $F(m, \mathbf{s})$ is unique for the variable $\mathbf{s}$ i.e., $F(m, \mathbf{s}) \neq F(m, \mathbf{s}')$ for every $\mathbf{s}' \neq \mathbf{s}$. We will show that if there exists an index j s.t. $d(s_j, F(m,\mathbf{s})_j) \geq D$, then $F(m,\mathbf{s})_j$ can be replaced by another symbol α s.t. $d(s_j, \alpha) < D$ and the resulting code word has smaller decoding error probability.

We now compute the probability that an output sequence $\mathbf{y}$ of the channel K belongs to the decoding region A^m given that the input is the encoding sequence $\mathbf{x} = F(m,\mathbf{s})$:

$$P\{\mathbf{y} \in A^m | \mathbf{x}\} = \sum_{\mathbf{y} \in A^m} \prod_{i=1}^{n} p_{Y|XS}(y_i|x_i, s_i)$$

Collecting the terms of the form $p_{Y|XS}(y|x_j, s_j)$ this probability can be written as:

$$P\{\mathbf{y} \in A^m | \mathbf{x}\} = \sum_{y \in \mathcal{Y}} \lambda_y p_{Y|XS}(y|x_j, s_j)$$

where $\lambda_y = \sum_{\mathbf{y} \in A^m, y_j = y} \prod_{i=1, i \neq j}^{n} p_{Y|XS}(y_i|x_i, s_i)$ are nonnegative constants. We define $\lambda_y = 0$, when there is no $\mathbf{y} \in A^m$ with $y_j = y$.

Substituting the definition of $p_{Y|XS}$ from (1) and recalling that $d(x_j, s_j) \geq D$ one gets:

$$\begin{aligned} P\{\mathbf{y} \in A^m | \mathbf{x}\} &= \sum_{y \in \mathcal{Y}} \lambda_y \frac{1}{N(s_j)} \sum_{x \in X_{s_j}} p_{Y|X}(y|x) \\ &= \frac{1}{N(s_j)} \sum_{x \in X_{s_j}} \sum_{y \in \mathcal{Y}} \lambda_y p_{Y|X}(y|x) \end{aligned}$$

The last expression is an average over the set X_{s_j}. Therefore, there exists a symbol $\alpha \in X_{s_j}$ such that $\sum_{y \in \mathcal{Y}} \lambda_y p_{Y|X}(y|\alpha) \geq \frac{1}{N(s_j)} \sum_{x \in X_{s_j}} \sum_{y \in \mathcal{Y}} \lambda_y p_{Y|X}(y|x)$.

It follows that by replacing the j^{th} entry of the sequence $\mathbf{x} = F(m,\mathbf{s})$ with α the probability of decoding error can only decrease. Applying the same procedure for all the entries in $F(m,\mathbf{s})$ s.t. $d(s_i, F(m,\mathbf{s})_i) \geq D$ produces a new codeword $F'(m,\mathbf{s})$ with smaller decoding error probability. This new code word is acceptable as a code word for the channel H.

This argument proves the lemma for the case where for each message m the codewords $\{F(m,\mathbf{s})\}$ and $\{F(m,\mathbf{s}')\}$ are distinct when $s \neq s'$. Consider now the case where for some message m there are k side information sequences $\mathbf{s}_1, \ldots, \mathbf{s}_k$ which are encoded by the same codeword. We have showed, that for any codeword $F(m,\mathbf{s})$, there is a procedure that produces a different codeword $F'(m,\mathbf{s})$ which is acceptable for an encoding and decoding scheme for H and has lower decoding error probability. Note that this procedure depends on the sequence $\mathbf{s}$, and hence even if one starts with the same codeword $F(m,\mathbf{s}_l)$ for $l = 1, \ldots, k$, the procedure may produce different outputs. Hence, if one replaces the unique codeword, corresponding to all the k sequences $\mathbf{s}_1,\ldots,\mathbf{s}_k$, with the k (not necessarily distinct) output codewords of this procedure, the new encoding and decoding scheme has the same rate and has a smaller average probability of decoding error. $\Box$

B Proof of Theorem 2.6

We prove a theorem on the capacity of a SIC from which theorem 2.6 is easily derived.

Theorem B.1 *Let K be a memoryless communication channel with side information whose output Y is defined by: $Y = (X+S+Z)modQ$. The side information S and the noise Z are independent random variables taking values in $\{0,\ldots,Q-1\}$ where Q is a positive integer. The input variable X takes values in $\{0,\ldots,P-1\}$ where P is a positive integers that divides Q. If Z is distributed uniformly (with no restriction on the distribution of S) then K has the same capacity as the communication channel C whose output is defined by $Y = (X+Z)modP$, where the input variable X and the noise Z are defined as above.*

Proof: To show that the two capacities are equal we start by proving that the capacity C_C of C is greater or equal to the capacity C_K of K. Denote by K' the modification of the channel K, where the side information sequence $\mathbf{s}$ is made available to the decoder. It is clear that $C_{K'} \geq C_K$. In the channel K' the decoder, observes $Y' = (X+Z)modQ$ and in the channel C the decoder observes $Y = (X+Z)modP$. However, since Z is distributed uniformly $H(X|Y') = H(X|Y)$ and hence the capacity of these two channels are equal. It follows that $C_C = C_{K'} \geq C_K$.

To prove the converse, let $\mathcal{E}^C = (\mathcal{M}, f_e^C, f_d^C, n)$ be an encoding and decoding scheme for C. We show that there exists an encoding and decoding scheme $\mathcal{E}^K = (\mathcal{M}, f_e^K, f_d^K, n)$ for K with the same rate and the same average decoding error probability . Define the modulating function f_e^K using the modulating function f_e^C in the following way: $f_e^K(m,\mathbf{s}) = (f_e^C(m) - \mathbf{s})modP$ and define the decoding function $f_d^K(\mathbf{y}) = f_d^C((\mathbf{y})modP)$. Since $Q = \alpha P$, it follows that for every integer a, $((a)mod(Q))mod(P) = (a)mod(P)$. Therefore, for every noise sequence $\mathbf{z}$, the following equalities are satisfied:

$$\begin{aligned} f_d^K(f_e^K(m,\mathbf{s}) + \mathbf{s} + \mathbf{z}) &= f_d^C((((f_e^C(m) - \mathbf{s})modP + \mathbf{s} + \mathbf{z})modQ)modP) \\ &= f_d^C(((f_e^C(m) - \mathbf{s})modP + \mathbf{s} + \mathbf{z})modP) \\ &= f_d^C((f_e^C(m) + \mathbf{z})modP) \end{aligned}$$

It turns out that for each message m and each noise sequence $\mathbf{z}$ the decoded message m^* of the scheme $\mathcal{E}^K$ is equal, regardless the side information sequence $\mathbf{s}$ to the decoded message of the scheme $\mathcal{E}^C$ given the same message and noise sequence. Since the distribution of the noise sequences is identical for both channels it follows that for each message m:

$$P_K(f_d^K(\mathbf{y}) \neq m|m) \quad = \quad P_C(f_d^C(\mathbf{y}) \neq m|m)$$

Therefore, both the rate and the decoding error probability of the encoding and decoding schemes $\mathcal{E}^K$ and $\mathcal{E}^C$ are equal. □

C Deriving Upper and Lower Bounds for BMS Channel Capacity

In this appendix, we use the notation and formulas introduced in Section 3. By differentiating (3) with respect to p one receives:

$$\frac{\partial I(B,Y)}{\partial p} = \sum_y P_1(y) \log(q + p\frac{P_0(y)}{P_1(y)}) - \sum_y P_0(y) \log(p + q\frac{P_1(y)}{P_0(y)}). \tag{13}$$

At the point p_m where (13) equals zero, the function $I(B,Y)$ achieves a maximum, since the mutual information is a concave function of p. By substituting the condition $\frac{\partial I(B,Y)}{\partial p} = 0$ in (3), one finds the maximal value of $I(B,Y)$ to be:

$$I^{max}(B,Y) = -\sum_y P_0(y) \log(p_m + q_m\frac{P_1(y)}{P_0(y)}) = -\sum_y P_1(y) \log(q_m + p_m\frac{P_0(y)}{P_1(y)}).$$

Define two functions of the variable p:

$$I_0(p) = -\sum_y P_0(y) \log(p + q\frac{P_1(y)}{P_0(y)}) \; I_1(p) = -\sum_y P_1(y) \log(q + p\frac{P_0(y)}{P_1(y)}).$$

and then according to the previous equality, $I^{max}(B,Y) = I_0(p_m) = I_1(p_m)$. Note that $I_0(0) = D(P_0||P_1)$, $I_0(1) = 0$, $I_1(0) = 0$, $I_1(1) = D(P_1||P_0)$ and both $I_0(p)$ and $I_1(p)$ are convex functions since $\frac{\partial^2 I_0(p)}{\partial p^2} = \sum_y P_0(y)\frac{(P_0(y)-P_1(y))^2}{(pP_0(y)+qP_1(y))^2} \geq 0$ and $\frac{\partial^2 I_1(p)}{\partial p^2} = \sum_y P_1(y)\frac{(P_0(y)-P_1(y))^2}{(pP_0(y)+qP_1(y))^2} \geq 0$.

In the (I,p) plane the curves defined by $I_0(p)$ and $I_1(p)$ intersect at (I^{max}, p_m). From the convexity of $I_0(p)$ and $I_1(p)$ it follows that this point lies strictly below the intersection point (I', p_a) of the lines $I(p) = pD(P_1 \parallel P_0)$ and $I(p) = -pD(P_0 \parallel P_1) + D(P_0 \parallel P_1)$. Direct calculation shows that $I' = \frac{D_0 D_1}{D_0+D_1}$, where $D_0 = D(P_0 \parallel P_1)$ and $D_1 = D(P_1 \parallel P_0)$. Hence, an upper bound for I^{max} is given by:

$$I^{max}(B,Y) \leq \frac{D_0 D_1}{D_0 + D_1}.$$

To derive a lower bound for $I^{max}(B,Y)$ we use the inequality $D(P_0 \parallel P_1) \geq \frac{1}{2\ln 2} \parallel P_0 - P_1 \parallel^2$ [CT91]. From the equality $I^{max}(B,Y) = D(P_0 \parallel pP_0 + qP_1)$, it follows that:

$$I^{max}(B,Y) \geq \frac{q_m^2}{2\ln 2} \parallel P_0 - P_1 \parallel^2 .$$

Similarly $I^{max}(B,Y) = D(P_1 \parallel pP_0 + qP_1)$ implies that:

$$I^{max}(B,Y) \geq \frac{p_m^2}{2\ln 2} \parallel P_0 - P_1 \parallel^2 .$$

Combining the two inequalities yields:

$$I^{max}(B,Y) \geq \frac{1}{8\ln 2} \parallel P_0 - P_1 \parallel^2 .$$

D Investigating the Perturbation Function BMS

Perturbation Functions (PF's) encoding schemes have the general formula:

$$f_b(x) = x + (-1)^b g(x)\delta,$$

where the coefficient δ and the function $g(x)$ are chosen in accordance with the distortion constraint and the attack channel statistics. Due to their simplicity and robustness to statistical instability, the PF encoding schemes are an important subclass of BMS. Using Formula (3) for the capacity C_f and Formula (4) for the upper bound, developed at Section 3, the theoretical performance of few PF's encoding schemes were investigated in the context of the additive Gaussian IHS.

the following PF encoding schemes were compared:

1. PF1: $X = S + (-1)^b \delta std(S)$
2. PF2: $X = S(1 + (-1)^b \delta)$

3. PF3: $X = S(1 + (-1)^b \delta) + (-1)^b \delta std(S)$
4. PF4: $X = S + (-1)^b \delta \mid S \mid +(-1)^b \delta std(S)$

Where $std(S)$ stands for the standard deviation of the r.v. S. The results of these comparisons are presented in Figures 7 and 8, where $S \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{N}(0, 0.01)$. One can observe that the upper bounds are not tight, however, they preserve the relative performance order of the PF schemes. Scheme PF4 above, that has the best theoretical performance, was used in the experimental tests described at Section 5.

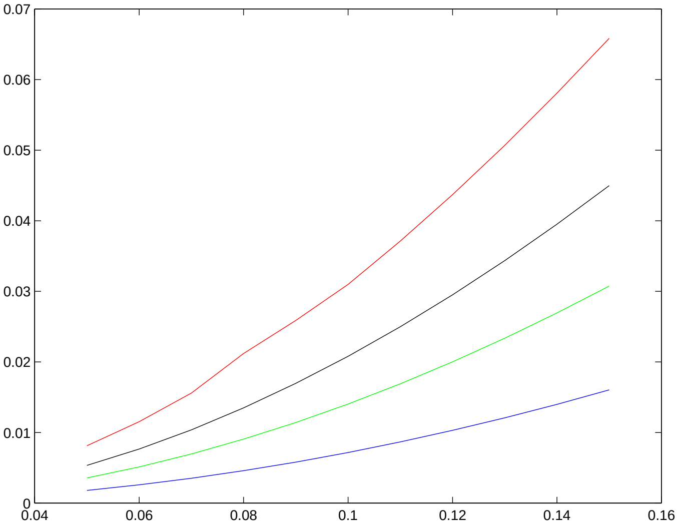

Figure 7: *A diagram of the capacity (in bits per channel use), as a function of the parameter δ, for perturbation function BMS: PF1 - blue, PF2 - green, PF3 - black, PF4 - red.*

E Deriving Sub-optimal Values for the SBE Scheme Parameters

Using the notation of Section 3.3, we recall that the codewords of an SBE scheme are typical sequences of the r.v. $U = \alpha S + W$. We assume that these codewords, when considered as random samples $\mathbf{u}(M, \mathbf{S})$ ($\mathbf{u}$ is the modulating mapping of the SBE scheme), can be well approximated by n i.i.d. realizations of U. Such an approximation will be faithful, in cases where the allowed distortion level P and the noise power N are significantly smaller than the covertext signal power Q. This situation is typical to certain transform domain image watermarking schemes where $\frac{P}{Q} \cong \frac{N}{Q} \ll 1$.

To decode the binary message $\mathbf{b}$ from the channel output one needs to estimate the encoding sequence $\mathbf{u}$ from the output sequence $\mathbf{y}$. As a basis for this estimation we consider a linear dependence $U = \beta Y + V$, where V is a Gaussian r.v. with zero mean, and the constant β has

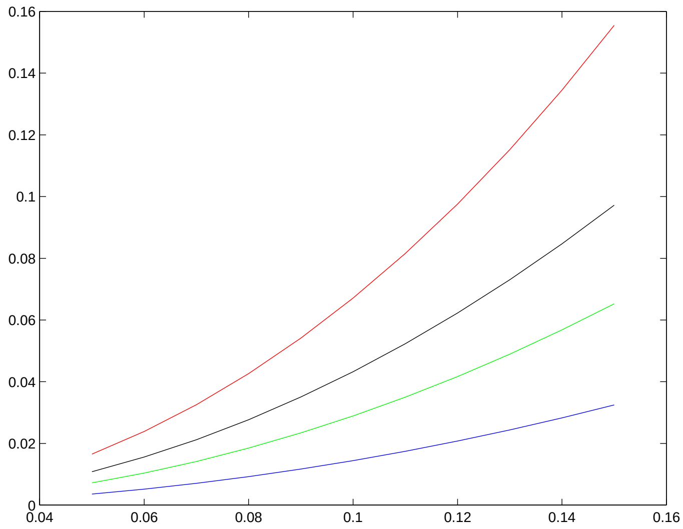

Figure 8: *A diagram of the upper bounds (in bits per channel use), as a function of the parameter δ, for perturbation function BMS: PF1 - blue, PF2 - green, PF3 - black, PF4 - red.*

the value that minimizes $Var(V)$. Such linear dependence exists in the additive Gaussian IHS, where U and Y are jointly Gaussian. We look now for the values of β, α that minimize $Var(U - \beta Y)$ and hence produce the best estimation in the mean quadratic distance sense. We need to solve the following optimization problem:

$$\alpha^{opt} = \arg\min_{\alpha} E(\beta(\alpha)Y - U(\alpha))^2. \tag{14}$$

To simplify the computation we assume that the noise variable Z is not correlated with S and with W, although the derivation can be extended to the case where $E(ZS)$ and $E(ZW)$ are non-zero. As a first step we determine the optimal value of β for fixed α. Direct differentiation yields: $\beta^{opt} = \frac{E(YU)}{E(Y^2)}$, and by substituting this expression in (14), one finds that solving (14) is equivalent to solving:

$$\alpha^{opt} = \arg\min_{\alpha}(E(Y^2)E(U^2) - E^2(YU)). \tag{15}$$

Substituting $Y = S + W + Z$ and $U = \alpha S + W$ in (15) the optimization problem reads:

$$\alpha^{opt} = \arg\min_{\alpha}(\alpha^2(QP + QN - M^2) + \alpha(2MN + 2M^2 - 2QP) + PQ + PN - M^2)$$

Differentiating with respect to α yields the optimal value:

$$\alpha = \frac{P - \frac{M(M+N)}{Q}}{P + N + \frac{M^2}{Q}}. \tag{16}$$

Returning to β^{opt}, we substitute $Y = S + W + Z$ and $U = \alpha S + W$ in $\beta^{opt} = \frac{E(YU)}{E(Y^2)}$ and obtain:

$$\beta^{opt} = \frac{\alpha Q + P + M(1+\alpha)}{Q + P + N + 2M}. \tag{17}$$

Substituting the optimal value of α from (16) into (17) gives the explicit expression:

$$\beta = \frac{P}{P+N} + M\frac{N}{(P+N)(Q+P+N+2M)}. \tag{18}$$

Note that for negligible M one can write $\beta^{opt} \cong \alpha^{opt} \cong \frac{P}{P+N}$.

In order to perform a ML decoding, one should compute the conditional probability $P_{Y|U}\{y|u\}$ and the apriory probability $P_U(u)$. For computing the first probability, we use again a linear dependence of the form: $Y = \gamma(\alpha)U + T$, where T is a Gaussian r.v. independent of U and $\gamma(\alpha)$ is a parameter that minimizes the variance of T, denoted V_T. Substituting the optimal value $\alpha = \frac{P}{P+N}$, and using the same technique as before, we arrive at:

$$\gamma^{opt} = \frac{Q+N+P}{\alpha^{opt}Q + N + P} \quad \text{and} \quad V_T = \frac{Q+P+N}{\alpha^{opt}Q + P + N}N. \tag{19}$$

It follows that

$$P\{y|u\} = P_{\mathcal{N}(0,V_T)}(y - \gamma^{opt}u), \tag{20}$$

where $P_{\mathcal{N}(0,V_T)}(\cdot)$ denotes the Gaussain probability density function with zero mean and variance V_T. For the second probability we substitute $\alpha = \frac{P}{P+N}$ in the formula $U = \alpha S + W$ and obtain:

$$P_U(u) = P_{\mathcal{N}(0,V_U)}(u), \tag{21}$$

where $V_U = (\alpha^{opt})^2 Q + P$ is the variance of U.

F Uniform and Adaptive Quantizers

In this appendix, we present the uniform and adaptive quantizers referred to in Section 5. It is evident that these quantizers are related to the absolute difference and the adaptive distortion measures described in Section 4.2. As the adaptive distortion measure is a better model for the human visual system sensitivity, the corresponding adaptive quantizers has advantage over the uniform quantizer in image watermarking applications.

Uniform quantizers The pair of uniform quantizers that constitute the modulating mapping are defined by the following two formulas:

$$q^0(s) = 2D\, round(\frac{s}{2D} - \frac{1}{4}) + \frac{D}{2}$$

and

$$q^1(s) = 2D\, round(\frac{s}{2D} + \frac{1}{4}) - \frac{D}{2},$$

where $round(\cdot)$ returns the closest integer value to its real input and D is a parameter. The uniform quantizers are designed to satisfy the absolute difference distortion constraint with distortion level D. One should note that the uniform quantizers satisfy the mean quadratic distortion constraint for distortion level that equal D^2.

Adaptive quantizers The pair of adaptive quantizers are defined by their set of representation levels. The modulating mapping maps each input value to the closest representation level of the selected quantizer. Although the covertext r.v. may have an infinite support, e.g. a Gaussian r.v., it is enough to consider finite quantization sets since the probability of observing very large values is negligible. We define two sets of positive representation levels by:

$$q^p(i) = \frac{\delta}{\epsilon}[(1 - \frac{\epsilon}{4})(1+\epsilon)^i - 1]$$

and

$$q^n(i) = \frac{\delta}{\epsilon}[(1 + \frac{\epsilon}{4})(1+\epsilon)^i - 1].$$

The representation levels of the adaptive quantizer pair are defined as: $\mathbf{q}^0 = [-inv(\mathbf{q}^p), \mathbf{q}^n]$ and $\mathbf{q}^1 = [-inv(\mathbf{q}^n), \mathbf{q}^p]$, where $inv(\mathbf{v})$ denotes the vector $\mathbf{v}$ in reverse order. The range of the index i is determined so that the two sets cover the range of the r.v. S up to negligible probability. The parameters ϵ and δ are designed so that the modulating mapping satisfies the adaptive distortion constraint, i.e. the representation levels are dense near zero and become more sparse as their absolute value grows. We found that by setting $\epsilon = \sqrt{\frac{P}{Q}}$ and $\delta = 3.3\sqrt{P}$ the quantizers satisfy also the mean quadratic distortion constraint with distortion level P.

G References

[BCW00] R.J. Barron, B. Chen, and G.W. Wornell. The duality between information embedding and source coding with side information and its implications//application. *Preprint*, 2000.

[Che00] B. Chen. *Design and analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems.* PhD thesis, June 2000.

[CKLS96] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for images, audio and video. In *Proc. IEEE Int. Conf. on Image Processing*, volume III, pages 243–246, September 1996.

[CL99] A. Cohen and A. Lapidoth. On the gaussian watermarking game. *Preprint*, 1999.

[CMM99] I.J. Cox, M.L. Miller, and A.L. McKellips. Watermarking as communications with side information. *Proceedings of the IEEE*, 87(7):1127–1141, July 1999.

[Cos83] M.H. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[CT91] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wily & Sons, 1991.

[CW99] B. Chen and G.W. Wornell. Dither modulation: a new approach to digital watermarking and information embedding. In *IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents*, volume 3657, pages 52–57, January 1999.

[GP80] S.I. Gelfand and M.S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

[Lem79] A. Lempel. Hadamard and m-sequence transforms are permutationally similar. *Applied Optics*, 18(24):4064–5, 1979.

[LM00] A. Levy and N. Merhav. A watermarking scheme based on scaled bin encoding. *HPL Invention Disclosure*, January 2000.

[MO99] P. Moulin and J.A. O'Sullivan. Information-theoretic analysis of information hiding. *Preprint*, 1999.

[PBBC97] A. Piva, M. Barni, F. Bartolini, and V. Cappellini. Dct-based watermark recovering without resorting to the uncorrupted original image. In *Proc. IEEE Int. Conf. on Image Processing*, pages 520–523, July 1997.

[PZ98] C.I. Podilchuk and W. Zeng. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 16(4):525–539, May 1998.

[SEG00] J.K. Su, J.J. Eggers, and B. Girod. Channel coding and rate distortion with side information: Geometric interpretation and illustration of duality. *Preprint*, 2000.

[Sha58] C.E. Shannon. Channels with side information at the transmitter. *IBM J. Res. Develop.*, 2:289–293, October 1958.