

Informed data-hiding with partial side information at the decoder

S. Voloshynovskiy^a, O. Koval^a, F. Perez-Gonzalez^c,
M. Kivanc Mihcak^{b 1}

^a*CUI-University of Geneva,
7, route de Drize,
CH-1227 Geneva, Switzerland
<http://sip.unige.ch/>*

^b*Electrical Engineering and Computer Science Department,
Bosfor University,
34342 - Bebek/Istanbul, Turkey*

^c*Signal Processing in Communications Group,
Signal Theory and Communications Department,
University of Vigo, 36310 Vigo, Spain*

Abstract

In this paper, we extend a traditional robust data-hiding set-up with host state at the encoder to a case when a partial side information about host statistics is also available at the decoder. We demonstrate that the knowledge of host statistics at the decoder can relax the critical requirements of random binning-based methods concerning attack channel statistics at the encoder. We also analyze performance improvement of some known data-hiding methods showing that they are particular cases of the generalized set-up. Finally, we consider some related open research issues and future possible extensions.

Keywords: multimedia robust data-hiding, Gel'fand-Pinsker set-up, side information, stochastic image modeling, channel capacity.

¹ Further information: Send correspondence to S. Voloshynovskiy
Email addresses: svolos@unige.ch (S. Voloshynovskiy),
Oleksiy.Koval@unige.ch (O. Koval), fperez@gts.tsc.uvigo.es (F. Perez-Gonzalez), kivanc.mihcak@boun.edu.tr (M. Kivanc Mihcak).

1 Introduction

Digital data-hiding appeared as an emerging tool for multimedia security, processing and management. A tremendous amount of possible applications have been recently reported that include copyright protection, tamper proofing, content integrity verification and authentication, secret communications (steganography) and watermark-assisted media processing such as multimedia indexing, retrieval and quality enhancement [1].

Robust data-hiding is one of the key technologies for the above applications that require reliable watermark decoding after different intentional and unintentional attacks [2]. At the same time, it is facing two important problems related to the host interference and the lack of information about the attacker strategy at the encoder.

The design of host interference cancellation critically relies on the knowledge of the host realization at the encoder. The main assumption of almost all current data-hiding techniques consists in the availability of the host state (realization, an original multimedia file) at the encoder assuming some fixed attacker strategy. The related communications problem is the one considered by Gel'fand and Pinsker [3]. The Gel'fand-Pinsker set-up is based on a random binning argument contrarily to the classical random coding argument that does not take into account the channel state information. Costa considered the Gel'fand-Pinsker problem in a Gaussian set-up and mean squared distortion criteria and demonstrated that the capacity of the Gaussian channel with the Gaussian interfering host can be equal to the capacity of interference-free communications [4]. It is important to note that the generation of the data-hider codebook is based on a *compensation parameter* (parameter α in original Costa paper) that takes into account attack statistics that correspond to the variance of an additive white Gaussian noise (AWGN) in the Costa problem. If the compensation parameter is optimally selected assuming the knowledge of the attack channel variance, the Costa rate equals the channel capacity of the AWGN channel. However, if this parameter is not optimal, the host plays a crucial role in the system performance leading to a considerable rate loss.

Practical low-complexity implementations of the Costa set-up are based on structured codebooks that use scalar (1-D)/vector (multidimensional) quantizers/lattices and are known as *Distortion-Compensated Dither Modulation* (DC-DM) [5] and *Scalar Costa Scheme* (SCS) [6]. It should also be pointed out that both SCS and DC-DM completely disregard host statistics corresponding to host probability density function (pdf) for watermark design using the argument that the variance of multimedia data is much larger than watermark and noise variances. This is equivalent to a high-rate assumption in rate distortion theory that corresponds to the small distortion regime. In this case,

the host pdf is considered to be flat (uniform) within quantization interval which significantly simplifies the design and analysis of these methods. It also corresponds to the assumption of very low watermark-to-image ratio (WIR). This imposes some specific constraints on the system design that determine its performance in the broad range of watermark-to-noise ratios (WNR).

The related problem of host-interference cancellation based on the above principles refers to the knowledge of the noise variance at the encoder prior to the transmission. That is not a case for the robust data-hiding where the attack parameters are not available at the encoder and can be only estimated at the decoder. To relax the lack of this information, a particular version of the DC-DM known as a *dither modulation* (DM) completely disregards the actual noise variance in the attack channel and selects the constant value of the compensation parameter equal to 1, which is only optimal for asymptotically high-WNR regime.

Contrarily, the methods based on the *spread spectrum* (SS) principle sacrifice from the host interference since they do not take into account the host state at the encoder. However, at the same time they demonstrate superior performance at the low-WNR regime in contrast to the host pdf-independent quantization-based methods [6]. Recently, Perez-Freire *et. al.* [7] demonstrated that additional enhancement can be achieved by proper modeling of host pdf. This group of methods can be also considered based on the random binning argument assuming that the compensation parameter tends to zero.

The goal of this paper is to extend the Gel'fand-Pinsker set-up, where the host realization is non-causally available only at the encoder, to the communications with extra side information about the host statistics at the decoder. The overall objective is to relax the critical dependence of the Costa set-up on the knowledge of the attack channel variance and to achieve good performance at low- and high-WNR regimes simultaneously. In fact, as a consequence, we want to theoretically confirm the possibility to develop a hybrid scheme that can combine the best from quantization- and SS-based methods to perform optimally under channel ambiguity.

In the related publications of Moulin and O'Sullivan [8,9], the authors use similar set-up to consider data-hiding problem as information-theoretic game between the data-hider and the attacker in the scope of the Gel'fand-Pinsker problem with the maximum allowable embedding and attacking distortions. The side information available at the encoder and decoder is considered in the form of secret key and has a twofold role. First, it may be a cryptographic key that is independent of the host. Second, the side information can be in some dependence with the host signal forming a joint distribution that is of a particular interest in our analysis. Moreover, Moulin and Mihcak [10] have practically applied this framework to the analysis of data-hiding capacity of real

images using the so-called parallel Gaussian decomposition where the image is supposed to consist of a number of independent Gaussian channels. Thus, the optimal data-hiding system performs the power allocation among these channels to satisfy the imposed distortion constraints. The generic assumptions made there are, however, different from those used in our work, and so the final methodology that we have followed, also differs. The main differences consist in part of codebook design, generic character of side information and applied decomposition. The generic side information possibly correlated with the host data is considered to be symmetrically available at the encoder and the decoder in the Moulin and O’Sullivan set-up. This enables a possibility of optimal watermark power allocation among different channels. In our set-up, we assume the availability of asymmetric or so-called partial side information at the decoder that, being sub-optimal in terms of achievable rate, simplifies the encoder structure and relaxes the critical dependence of the optimal codebook design on the ambiguity concerning attack channel parameters that can be of interest for numerous practical applications. Finally, Moulin and Mihcak applied the generic parallel channel decomposition without providing any relationship to the statistics of the host data besides the assumption that samples in the parallel channels are Gaussian. In our analysis, we follow a similar decomposition of host data providing the so-called *source splitting* with Gaussian mixture distribution [11]. However, the analysis of our set-up refers to the link between global and local image statistics demonstrated on a particular example of Laplacian distribution. This makes possible to have a concrete implementation of the Moulin and O’Sullivan idea, where the hypothetical side information is represented in the form of local data variances, which are modeled using exponential distribution.

It should be also mentioned the set-up analyzed by Cover and Chiang [12] who have considered a generic channel coding problem with asymmetrically available side information. The main difference of this work with the set-up of Cover and Chiang consists in the definition of channel where the side information available at the encoder and decoder simultaneously defines the channel state taking into account the physical origin of the available side information. While only general information-theoretic argument was used in [12], the available side information is linked with the statistical properties of real images in this work.

Finally, the extension of the Gel’fand-Pinker set-up to real channels was performed by Cohen and Lapidoth [13] who have indicated that for the AWGN channels the statistics of the host have no influence on the channel capacity. This conclusion was deduced based on the assumption that the attack channel pdf parameters (variance of the AWGN channel) is also perfectly available at the encoder for the optimal selection of optimization parameter. Contrarily, if there is a small deviation of the channel variance from those available at the encoder, one might expect a considerable deviation of achievable rate from

the channel capacity. We intend to show in this paper that the presence of sequence correlated with host state on the side of decoder might relax this critical impact. At the same time, it should be pointed out that if the attack channel state is perfectly known at the encoder, the considered set-up performance coincides with the results of Cohen and Lapidoth.

The paper has the following structure. The basic set-up of side information-assisted data-hiding is considered in Section 2. The data-hiding concept with the host state at the encoder based on the random binning argument is briefly reviewed in Section 3 in the scope of generalized Gel'fand-Pinsker problem, Costa Gaussian set-up and discrete approximation of the Costa problem. Section 4 presents the source splitting principle and the corresponding relationship between local and global stochastic image models. A new data-hiding communications set-up with host state at the encoder and host statistics at the decoder is presented in Section 5 and some known data-hiding methods are considered as particular cases of this set-up. The experimental results demonstrating the efficiency of the proposed approach in terms of achievable rates are given in Section 6. Finally, Section 7 concludes the paper and presents some future research perspectives.

Notations We use capital letters to denote scalar random variables X , bold capital letters to denote vector random variables \mathbf{X} , corresponding small letters x and \mathbf{x} to denote the realizations of scalar and vector random variables, respectively. The superscript N is used to designate length- N vectors $\mathbf{x} = x^N = [x[1], x[2], \dots, x[N]]^T$ with k^{th} element $x[k]$. We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to indicate that a random variable X is distributed according to $p_X(x)$. The mathematical expectation of a random variable $X \sim p_X(x)$ is denoted by $E_{p_X}[X]$ or simply by $E[X]$. Calligraphic fonts \mathcal{X} denote sets $X \in \mathcal{X}$ and $|\mathcal{X}|$ denotes the cardinality of set \mathcal{X} . \mathbf{I}_N denotes the $N \times N$ identity matrix. We also define the WIR as $\text{WIR} = 10 \log_{10} \frac{\sigma_W^2}{\sigma_X^2}$ and the WNR as $\text{WNR} = 10 \log_{10} \frac{\sigma_W^2}{\sigma_Z^2}$ where σ_X^2 , σ_W^2 , σ_Z^2 represent the variances of host data, watermark and noise, respectively.

2 Side information-aided data-hiding

This section presents the basic set-up of side information-aided digital data-hiding. This set-up was first proposed by Voyatzis and Pitas in 1999 [14] and was further extended and analyzed by Cannons and Moulin [15] in terms of statistical detection performance evaluation.

The block-diagram of side information-aided data-hiding is shown in Figure 1. This set-up corresponds to the classical data-hiding scenario where the data-

hider has the access to the uniquely assigned secret key K^N and to the non-causal interference $x^N \in \mathcal{X}^N$. We also assume that X^N is distributed according to $p_{X^N}(x^N)$. A message m , key k^N and non-causal host realization x^N are used by the encoder to produce a watermark w^N . The watermark w^N is embedded into the host data x^N , thus resulting in the watermarked data y'^N . The attacker attempting to deteriorate the reliable communications of the hidden message applies certain attack to the watermarked data by passing y'^N through the attack channel $p(y^N|y'^N)$.

The only difference with the classical set-up consists in the availability of the side information $S^N = \psi(X^N, K^N)$ at the decoder representing some key-dependent simplified representation of the host data. A function $\psi(\cdot)$ represents some hash or statistics that can be extracted from the host data X^N using a secret key K^N . The hashing can represent some extraction and key-dependent quantization of host statistics or features. It can for example be scalar or vector subtractive dither quantization [16]. The key K^N is chosen from the corresponding alphabet \mathcal{K}^N where all keys are distributed uniformly. The role of the key is to provide a source of randomness and to give the data-hider an information advantage over the attacker.

The secret key and the side information $S^N \in \mathcal{S}^N$ are communicated to the decoder via some private channel. The decoder combines this information with the channel output Y^N and produces the estimate of the original message \hat{m} . We will assume that the message $m \in \mathcal{M}$ is uniformly distributed over $\mathcal{M} = \{1, 2, \dots, |\mathcal{M}|\}$, with $|\mathcal{M}| = 2^{NR}$, where R is the data-hiding rate. It is assumed that the stego and attacked data are defined on $y'^N \in \mathcal{Y}'^N$ and $y^N \in \mathcal{Y}^N$, respectively. The distortion function is defined as:

$$d^N(x^N, y'^N) = \frac{1}{N} \sum_{i=1}^N d(x_i, y'_i), \quad (1)$$

where $d(x_i, y'_i) : \mathcal{X} \times \mathcal{Y}' \rightarrow \mathcal{R}^+$ denotes the element-wise distortion between x_i and y'_i .

Definition 1: A *discrete memoryless data-hiding channel* consists of five alphabets $\mathcal{X}, \mathcal{S}, \mathcal{W}, \mathcal{Y}', \mathcal{Y}$; probability transition matrix $p_{Y^N|W^N, X^N}(y^N|w^N, x^N)$ corresponding to the covert channel communications of the watermark W^N through the host image X^N , $p_{Y'^N|W^N, X^N}(y'^N|w^N, x^N)$, and attack channel $p_{Y^N|Y'^N}(y^N|y'^N)$ such that:

$$p_{Y^N|W^N, X^N}(y^N|w^N, x^N) = \sum_{y'^N} p_{Y'^N|W^N, X^N}(y'^N|w^N, x^N) p_{Y^N|Y'^N}(y^N|y'^N).$$

The attack channel is subject to the distortion constraint D^A :

$$\sum_{y'^N \in \mathcal{Y}'^N} \sum_{y^N \in \mathcal{Y}^N} d^N(y'^N, y^N) p_{Y^N|Y'^N}(y^N|y'^N) p_{Y'^N}(y'^N) \leq D^A, \quad (2)$$

where $p_{Y^N|Y'^N}(y^N|y'^N) = \prod_{i=1}^N p_{Y|Y'}(y_i|y'_i)$.

Definition 2: A $(2^{NR}, N)$ code for data-hiding channel consists of *message set* $\mathcal{M} = \{1, 2, \dots, 2^{NR}\}$, *encoding function*:

$$\phi^N : \mathcal{M} \times \mathcal{X}^N \times \mathcal{K}^N \rightarrow \mathcal{W}^N, \quad (3)$$

embedding function:

$$\varphi^N : \mathcal{W}^N \times \mathcal{X}^N \rightarrow \mathcal{Y}^N, \quad (4)$$

subject to the embedding distortion constraint D^E :

$$\frac{1}{|\mathcal{K}^N| |\mathcal{M}|} \sum_{k^N \in \mathcal{K}^N} \sum_{m \in \mathcal{M}} \sum_{x^N \in \mathcal{X}^N} d^N(x^N, \varphi^N(\phi^N(m, x^N, k^N))) p_{X^N}(x^N) \leq D^E \quad (5)$$

and *decoding function*:

$$g^N : \mathcal{Y}^N \times \mathcal{K}^N \times \mathcal{S}^N \rightarrow \mathcal{M}. \quad (6)$$

It is important to note that the corresponding covert channel satisfies memoryless properties:

$$p_{W^N, U^N | X^N, K^N}(w^N, u^N | x^N, k^N) = \prod_{i=1}^N p_{W, U | X, K}(w_i, u_i | x_i, k_i). \quad (7)$$

We define the *average probability of error* for $(2^{NR}, N)$ code as:

$$P_e^{(N)} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \Pr[g^N(Y^N, K^N, S^N) \neq m | M = m]. \quad (8)$$

Definition 3: A rate $R = \frac{1}{N} \log_2 |\mathcal{M}|$ is achievable for distortions (D^E, D^A) , if there exists a sequence of $(2^{NR}, N)$ codes with $P_e^{(N)} \rightarrow 0$ as $N \rightarrow \infty$.

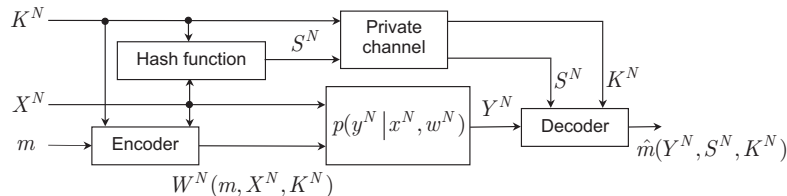


Fig. 1. Generalized block-diagram of side information-aided data-hiding.

We will refer to the case $S^N = 0$ as *blind data-hiding* and to the case $S^N = X^N$ as *non-blind data-hiding*. The intermediate case when $S^N = \psi(X^N, K^N)$ is known as *semi-blind data-hiding* [15]. The main idea behind the semi-blind data-hiding consists in the necessity to have image dependent keys and a decoder that is either operated by the content provider as in fingerprinting,

or at least requires some communications with the content provider. In the following, to simplify the notations we will consider only the communication aspect of the problem and skip the dependence of the encoding, decoding and hashing on the secret key implicitly assuming that the proper key management is established among all these entities.

3 Host state at the encoder: host realization

3.1 Gel'fand-Pinsker problem

This problem can be formulated as a reliable communication of a message $m \in \{1, 2, \dots, |\mathcal{M}|\}$, with $|\mathcal{M}| = 2^{NR}$, encoded into a sequence W^N , over the channel $p_{Y|W,X}(y|w, x)$ and interference X^N being known at the encoder but not at the decoder. The corresponding discrete memoryless channel is described by $\{\mathcal{W}, \mathcal{X}, p_{Y|W,X}(y|w, x), \mathcal{Y}\}$ where the side information is assumed to have a distribution $p_{\mathbf{X}}(\mathbf{x}) = \prod p_X(x_i)$ (Figure 2). The problem is to find the maximum rate of reliable communications $R = \frac{1}{N} \log_2 |\mathcal{M}|$.

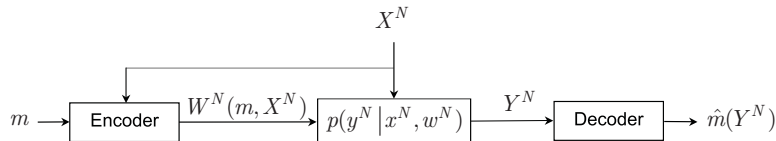


Fig. 2. Gel'fand-Pinsker channel coding with side information at the encoder: $Pr[m \neq \hat{m}(Y^N)] \rightarrow 0$ as $N \rightarrow \infty$.

Gel'fand and Pinsker [3] have shown using binning-based codebook construction that the capacity of this channel is:

$$C_X^{10} = \max_{p(u,w|x)} [I(U; Y) - I(U; X)], \quad (9)$$

where the maximization is over all auxiliary random variables U , such that $U \rightarrow (W, X) \rightarrow Y$ form a Markov chain. $I(U; Y)$ and $I(U; X)$ correspond to mutual information between the observed data Y and U , the host data X and U , respectively. It was shown in [3] that the maximization is attained with $p_{U,W|X}(u, w|x)$ such that W is deterministic given (U, X) .

Here and in the following, we use the superscripts to denote the availability (1 stands for "available" and 0 for "not available") of corresponding states or statistics used in the subscripts at the encoder and the decoder, respectively. Therefore, in the above case, "10" denotes the availability of the host state X at the encoder but not at the decoder.

Without going into deep details of the proof of (9) presented in [3], we demonstrate the main considerations that define the achievable rate as the trade-off in the probability of error at the encoder and decoder. Codebook consists of $J|\mathcal{M}|$ codewords $u^N(m, j)$, $m \in \{1, 2, \dots, |\mathcal{M}|\}$, $j \in \{1, 2, \dots, J\}$ generated independently at random according to the marginal distribution $p_U(\cdot)$ and allocated into $|\mathcal{M}| = 2^{NR}$ bins with $J = 2^{NR'}$ codewords in each bin for every message m . $R' = \frac{1}{N} \log_2 J$ bits are used to describe the host at the encoder.

Encoder: Given the message $m \in \mathcal{M}$ and the host data x^N , the encoder seeks a codeword $u^N(m, j)$ such that $(u^N(m, j), x^N) \in A_\epsilon^{*(N)}(U, X)$ (see Appendix A for the definition of typicality). If such a codeword $u^N(m, j)$ is found, the encoder produces the watermark $w^N = \phi^N(u^N(m, j), x^N)$.

The probability that there is no $u^N(m, j)$ that is strongly jointly typical with x^N should be bounded to guarantee encoder errorless performance:

$$P_{e_{enc}} \leq e^{-2^{NR'} 2^{-N[I(U;X)+\epsilon]}}, \quad (10)$$

where ϵ to be a small non-negative number and $P_{e_{enc}} \rightarrow 0$ as $N \rightarrow \infty$ and $\epsilon \rightarrow 0$, if $R' > I(U; X)$ or if $J > 2^{N[I(U;X)+\epsilon]}$. This result suggests to have a relatively large J in order to assure errorless performance of the encoder.

Decoder: Given the output of the channel y^N , the decoder seeks a codeword $u^N(\tilde{m}, j)$, such that $(u^N(\tilde{m}, j), y^N) \in A_\delta^{*(N)}(U, Y)$ in the set of all codewords $1 \leq j \leq J$, $1 \leq m \leq |\mathcal{M}|$. If the decoder finds a unique jointly typical pair, it declares that the sent message was $\hat{m} = \tilde{m}$. Otherwise, an error is declared.

The probability of error at the decoder can be bounded by:

$$P_{e_{dec}} < 2^{N[R+R']} 2^{-N[I(U;K)-\delta]}. \quad (11)$$

Thus, we require $\delta \rightarrow 0$, $N \rightarrow \infty$ and $R + R' < I(U; Y)$ to have $P_{e_{dec}} \rightarrow 0$.

Combining the rate constraints obtained for the errorless performance of the encoder and decoder, one can provide the condition for the complete system that guarantees errorless message communications, i.e., $R < (I(U; Y) - I(U; X))$, which corresponds to the result (9).

3.2 Costa problem

Costa considered the Gel'fand-Pinsker problem for the Gaussian context and mean-square error distance [4]. The corresponding fixed channel $p_{Y|W,X}(y|w, x)$ is the Gaussian one with $X \sim \mathcal{N}(0, \sigma_X^2)$ and additive $Z \sim \mathcal{N}(0, \sigma_Z^2)$ (Figure 3). The embedding distortion constraint is imposed in the form of $E[W^2] \leq \sigma_W^2$ and the attack distortion corresponds to the variance of the additive Gaussian

noise σ_Z^2 . The auxiliary random variable was chosen as $U = W + \alpha X$ with compensation parameter α that defines the achievable communications rate:

$$R(\alpha, \sigma_X^2) = \frac{1}{2} \log_2 \frac{\sigma_W^2 (\sigma_W^2 + \sigma_X^2 + \sigma_Z^2)}{\sigma_W^2 \sigma_X^2 (1 - \alpha)^2 + \sigma_Z^2 (\sigma_W^2 + \alpha^2 \sigma_X^2)}. \quad (12)$$

Costa has shown that the optimal compensation parameter is $\alpha_{opt} = \frac{\sigma_W^2}{\sigma_W^2 + \sigma_Z^2}$ that requires the knowledge of σ_Z^2 at the encoder. In this case, the rate does not depend on the host variance and:

$$R(\alpha_{opt}) = C^{AWGN} = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_W^2}{\sigma_Z^2} \right) \quad (13)$$

that corresponds to the capacity of the AWGN channel without interference.

It is important to note that the number of codewords in each bin of the message of the Gel'fand-Pinsker set-up is approximately equal to $2^{NI(U;X)}$. In the Costa set-up, $I(U; X) = \frac{1}{2} \log_2 \left(1 + \alpha^2 \frac{\sigma_X^2}{\sigma_W^2} \right)$. This implies that the larger variance of the host σ_X^2 , the larger number of codewords are needed at the encoder in each bin. This means that for the capacity achieving scheme the codebook should be extended to take into account larger number of host states.

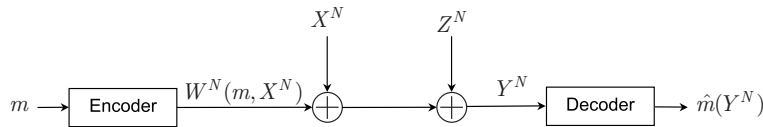


Fig. 3. Costa channel coding with the host state information at the encoder.

3.3 Scalar Costa Scheme: discrete approximation of Costa problem

The Costa set-up operates with the Gaussian random codebook that has exponential complexity. To reduce the complexity a number of practical data-hiding algorithms use structured codebooks instead of random ones based on the above considered binning argument [5,6]. These structured codebooks are designed based on the quantizers (lattices) that finally should provide the independence of the watermark (considered to be the quantization noise) and the host at the high-rate quantization regime (low-distortions).

The group of quantization-based methods follows an analogy between binning strategy in the Gel'fand-Pinsker problem and the same principle of quantization attempting to approximate host-dependent selection of codewords [6]. The auxiliary random variable U in this set-up is approximated by:

$$U = W + \alpha' X = \alpha' Q_m(X), \quad (14)$$

leading to the jointly typical sequences $(u^N(m, j), x^N) = (\alpha' Q_m(x^N), x^N)$, where $Q_m(\cdot)$ denotes a vector or scalar quantizer for the message m . In the

simplified version of the SCS (or DC-DM) the quantizer is chosen to be the uniform scalar one working at the high-rate assumption where the pdf of the host signal X is assumed to tend towards a uniform distribution [5,6]. This produces a uniformly distributed watermark $W = U - \alpha'X = \alpha'Q_m(X) - \alpha'X$. The resulting stego data is obtained as:

$$y' = x + w = x + \alpha'(Q_m(x) - x). \quad (15)$$

In the above case of uniform quantizer, the watermark will be uniform, as a consequence of the high-rate quantization assumption, with variance $\sigma_W^2 = \alpha'^2 \frac{\Delta^2}{12}$, where Δ is the quantization step. Therefore, the selection of the rate maximizing α designed for the Gaussian watermark in the Costa set-up is not any more optimal in the above case (for this reason we use α').

4 Stochastic host modeling

The quantization-based data-hiding methods completely disregard the host statistics due to the high-rate assumption. Another “extrema” of data-hiding host image modeling is a zero-mean i.i.d. Gaussian model. Although the Gaussian model leads to nice closed-form solutions in many cases, the practical application of this model to real images is very questionable. The development of accurate and tractable stochastic image models is a very important and challenging problem. The most simple and widely used class of stochastic image models that represents the stochastic behavior of image coefficients after decorrelation in transform domains (such as discrete cosine transform (DCT) domain or discrete wavelet transform (DWT) domain) is an i.i.d. Generalized Gaussian (GG) pdf. The GG model captures a global behavior of coefficients. A particular case of this model is the Laplacian pdf which is obtained when the shape parameter in the GG pdf is equal to 1 [17]. A lot of practical image coders and denoisers are designed based on the Laplacian model [17,18].

Contrarily to global statistical models, local modeling is targeting accurate description of statistical properties of data at a more fine level. In wavelet domain, zero-trees, Hidden Markov Trees, Gaussian Scale Mixture are examples of such local image models. It was demonstrated that such models outperform global ones in various image processing applications [19–21]. Remarkably, there is a link that can be established between global and local statistical models by means of hierarchical or doubly stochastic modeling [11,22].

According to a doubly stochastic model coefficients in the wavelet domain are distributed independently according to a conditional distribution $p(\cdot|\zeta)$, where parameter $\zeta \sim p(\zeta)$. Assuming that $\zeta = \sigma_X^2$, one can establish such a link between local and global stochastic models in the wavelet domain via parallel

source splitting [11] based on the analysis of joint distribution $p(x, \sigma_X^2)$. Using chain rule for probability one obtains:

$$p_{X, \Sigma_X^2}(x, \sigma_X^2) = p_{\Sigma_X^2}(\sigma_X^2)p_{X|\Sigma_X^2}(x|\sigma_X^2), \quad (16)$$

where $p_{\Sigma_X^2}(\sigma_X^2)$ represents the marginal variance distribution and the conditional distribution $p_{X|\Sigma_X^2}(x|\sigma_X^2)$ is supposed to capture local data statistical behavior. The global data statistics correspond to the marginal distribution:

$$p_X(x) = \int_0^\infty p_{X, \Sigma_X^2}(x, \sigma_X^2)d\sigma_X^2 = \int_0^\infty p_{X|\Sigma_X^2}(x|\sigma_X^2)p_{\Sigma_X^2}(\sigma_X^2)d\sigma_X^2. \quad (17)$$

A particular case of interest is given by the *infinite Gaussian mixture model*² [23]. According to this model, in our case $p_{X|\Sigma_X^2}(x|\sigma_X^2)$ takes a form of zero-mean conditional Gaussian distribution, i.e., $p_{X|\Sigma_X^2}(x|\sigma_X^2) = \frac{1}{\sqrt{2\pi\sigma_X^2}}e^{-\frac{x^2}{2\sigma_X^2}}$. Therefore, local image statistics are modeled as Gaussian while the global distribution $p_X(x)$ is obtained specifying the variance distribution $p_{\Sigma_X^2}(\sigma_X^2)$. For the case that is of interest for various transform domains, the global Laplacian pdf is obtained as a weighted mixture of zero-mean conditionally Gaussian pdfs given exponentially distributed local variance $p_{\Sigma_X^2}(\sigma_X^2) = \lambda_1 e^{-\lambda_1 \sigma_X^2}$, where λ_1 is a scale parameter of the exponential distribution and:

$$p_X(x) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{x^2}{2\sigma_X^2}} \lambda_1 e^{-\lambda_1 \sigma_X^2} d\sigma_X^2 = \sqrt{\frac{\lambda_1}{2}} e^{-\sqrt{2\lambda_1}|x|}. \quad (18)$$

This simple relationship provides a fundamental link between the global and local statistics of image coefficients. Therefore, the same data can be considered to be locally zero-mean Gaussian with the variance distributed according to the exponential pdf and, simultaneously, having the Laplacian global statistics. The Gaussian mixture model is also the basis for the *parallel channel decomposition* of stochastic image sources that makes it possible to use the simple relationship for the Gaussian statistics. Moreover, properly selecting the variance distribution, one can obtain a general class of Generalized Gaussian distributions $p_X(x)$. However, in the following we will concentrate only on the Laplacian case for demonstration purposes.

It is important to note that one of the most efficient lossy image compression algorithms known as *estimation-quantization* codec [24] is based on this model. In fact, omitting the practical details of side information communications between the encoder and the decoder, Hjorungnes, Lervik and Ramstad [11] were the first who theoretically demonstrated that the rate gain between Laplacian and infinite Gaussian mixture models can be as much as 0.312 bits/sample for high-rate regime.

² In general zero-mean assumption is not necessarily required.

5 Partial side information at the decoder: host statistics

In this section, we extend the results of Section 3 to the case of side information available at the decoder. Our set-up can be positioned between the above considered Gel'fand-Pinsker set-up (9), where the side information in the form of host realization is available non-causally at the encoder, and the Wolfowitz set-up [25], where the host realization is available at both encoder and decoder that results in the capacity:

$$C_X^{11} = \max_{p(w|x)} I(W; Y|X). \quad (19)$$

In our ‘‘asymmetric’’ set-up, similarly to the formulation considered in [12], the host realization is available at the encoder but only the realization of host statistics is presented at the decoder as $S^N = \sigma_X^{2N}$ according to Figure 1. It is important to underline that the realization of the host parameters describing the Laplacian distribution is the N -length vector of local variances that determines the statistics of the parallel Gaussian channels in the source splitting model. Our motivation to select this side information as well as expectation of the gain in data-hiding application is based on the fact that $I(X; \Sigma_X^2) \leq 0$ [12,22].

The particular problem of interest for our analysis is the Costa version of the Gel'fand-Pinsker set-up. We consider here an extended version of the Costa problem with the Laplacian host realization available at the encoder and the corresponding host statistics in the form of variances of infinite Gaussian Mixture model defined according to the source splitting (17) as partial side information about the host available at the decoder (Figure 4).

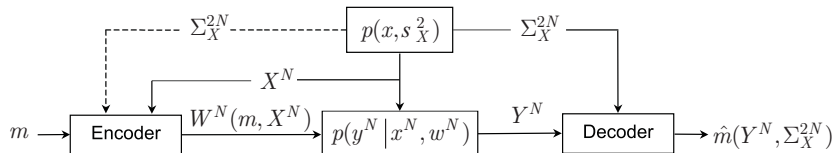


Fig. 4. Channel coding with host state at the encoder and host statistics at the decoder.

Therefore, besides the channel state X^N availability, the encoder potentially might have an access to Σ_X^{2N} (shown in dashed line). However, we will concentrate our analysis on the case when X^N is available at the encoder and Σ_X^{2N} is given at the decoder only. It is assumed that X^N and Σ_X^{2N} have the following joint distribution $p(x^N, \Sigma_X^{2N}) = \prod_{i=1}^N p(x_i, \sigma_{X_i}^2)$. The encoder produces a sequence $W^N(m, X^N)$ where $m \in \{1, 2, \dots, 2^{NR}\}$. The decoding of message \hat{m} is performed based on Y^N and Σ_X^{2N} as $\hat{m}(Y^N, \Sigma_X^{2N})$. The resulting averaged probability of error is $\frac{1}{2^{NR}} \sum_{m \in \mathcal{M}} \Pr[m \neq \hat{m}(Y^N, \Sigma_X^{2N}) | M = m]$ where we assume that all messages m are drawn uniformly over $\{1, 2, \dots, 2^{NR}\}$.

To derive a necessary intuition we first consider the above set-up for the DMC without taking into account the host realization at the encoder as shown in Figure 5. This implies that the watermark sequence $W^N(m)$ is generated without taking into account host state X^N similarly to the spread-spectrum methods. Thus, W^N is also independent with Σ_X^2 . In this case:

$$I(W; Y, \Sigma_X^2) = I(W; \Sigma_X^2) + I(W; Y | \Sigma_X^2) =$$

$$I(W; Y | \Sigma_X^2) = E_{\Sigma_X^2} \left[I(W; Y | \Sigma_X^2 = \sigma_X^2) \right], \quad (20)$$

where the second equality follows from the independence of W and Σ_X^2 .

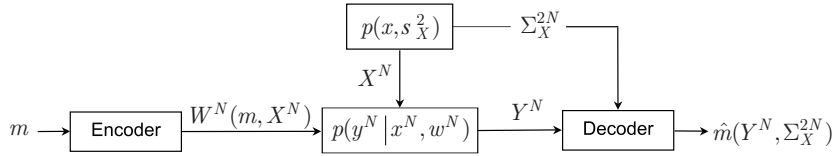


Fig. 5. Channel coding with host statistics at the decoder.

In this case, the codebook generation is based on the random coding technique. In particular, the encoder generates 2^{NR} i.i.d. codewords from distribution $p_W(w^N) = \prod_{i=1}^N p_W(w_i)$ and sends the corresponding codeword $W^N(m)$ to the channel to communicate the message $m = \{1, 2, \dots, 2^{NR}\}$. The decoder seeks a jointly typical triplet $(w^N(\hat{m}), y^N, \sigma_X^{2N}) \in A_\epsilon^{*(N)}(W, Y, \Sigma_X^2)$, ($A_\epsilon^{*(N)}(W, Y, \Sigma_X^2)$ denotes a set of jointly typical triplets (W, Y, Σ_X^2)) based on the received vector y^N and available side information σ_X^{2N} over all $w^N(\hat{m})$. Here, we use the definition of strong joint typicality (see Appendix A). If no such an \hat{m} exists or if there is more than one, then an error is declared. The analysis of the probability of error, i.e., an event when $\hat{m} \neq m$, averaged over all codes includes several possible cases. The first type of errors occurs when sent and received sequences are not strongly jointly typical and the second type of errors corresponds to the situation when a wrong codeword is strongly jointly typical with the received signal. The probability of the first event tends to zero according to the asymptotic equipartition property (AEP) [26]. To constrain the second probability of error it is possible to show that the rate should be bounded by $R < I(W; Y | \Sigma_X^2)$ for sufficiently large N . This concludes the analysis of the case when the host state is not available at the encoder.

In the extended set-up (Figure 4) with the host state available at the encoder, one can use random binning to incorporate the host state into the codebook design. Assuming that the host statistics Σ_X^{2N} are available at both encoder and decoder one has:

$$I(U; Y, \Sigma_X^2) - I(U; X, \Sigma_X^2) = I(U; \Sigma_X^2) + I(U; Y | \Sigma_X^2) - I(U; \Sigma_X^2)$$

$$- I(U; X | \Sigma_X^2) = I(U; Y | \Sigma_X^2) - I(U; X | \Sigma_X^2). \quad (21)$$

Following the above analysis, we will extend the Gel'fand-Pinsker set-up (9) to the case when the host statistics are available at the decoder and formulate the result in a form of Conjecture 1.

Conjecture 1: *If the host realization is non-causally available at the encoder according to the Gel'fand-Pinsker problem for the fixed channel $p_{Y|W,X}(y|w,x)$, and if the host statistics that govern this particular host realization are known at the decoder with $(X_i, \Sigma_{X_i}^2)$ to be pairwise i.i.d. $p(x_i, \sigma_{X_i}^2)$, then the capacity of this scheme is:*

$$C_{X, \Sigma_X^2}^{10,01} = \max_{p(u,w|x)} [I(U; Y, \Sigma_X^2) - I(U; X)], \quad (22)$$

where $C_{X, \Sigma_X^2}^{10,01}$ denotes the capacity with the host X available at the encoder and the statistics Σ_X^2 available at the decoder. A brief sketch of the achievability part of the conjecture 1 is given in Appendix A. In our analysis we assume the joint distribution $p(x, \sigma_X^2, u, w, y)$ describing relationship between all random variables to be $p(x, \sigma_X^2)p(u, w|x)p(y|w, x) = p(\sigma_X^2)p(x|\sigma_X^2)p(u|x)p(w|u, x) \times p(y|w, x)$ according to the source splitting model $p(x, \sigma_X^2)$ considered in Section 4, encoder design $p(u, w|x)$ considered by Gel'fand-Pinsker [3] and the channel model $p(y|w, x)$.

Before starting the consideration of particularities of the set-up considered in Conjecture 1, it is important to discover its applicability to the analysis to image data-hiding. In fact, it is widely realised in various image processing applications that for instance in discrete wavelet transform domain local variances of image coefficients are not independently distributed but are realisations of a slowly varying random field [24]. Therefore, one needs to generalize the above mentioned set-up to the case of channels with memory. In particular, assuming local correlation structure of the variance field, the problem under the analysis will have a blockwise memoryless structure. By other words, it is supposed that image is divided on a set of N/L independent non-overlapping blocks with correlated data inside each block. Similarly to [8], the following definitions are required to perform the generalizations.

Definition 6. A blockwise memoryless attack channel subject to attacking distortion constraint D^A is a conditional p.m.f. $p_{Y^N|Y'^N}(y^N|y'^N)$ such that:

$$\sum_{y'^L \in \mathcal{Y}'^L} \sum_{y^L \in \mathcal{Y}^L} d^L(y'^L, y^L) p_{Y^L|Y'^L}(y^L|y'^L) p_{Y'^L}(y'^L) \leq D^A. \quad (23)$$

Moreover, for blocks $y'^L_i = \{y'_{Li}, y'_{L(i+1)}, \dots, y'_{L(i+L-1)}\}$, $1 \leq i \leq N/L$, such that

$y'^N = \{y'_1, y'_2, \dots, y'_{N/L}\}$ one has

$$p_{Y^N|Y'^N}(y^N|y'^N) = \prod_{i=1}^{N/L} p_{Y^L|Y'^L}(y_i^L|y'_i{}^L). \quad (24)$$

Definition 7. A blockwise memoryless covert channel subject to embedding distortion constraint D^E is a conditional p.m.f. $p_{W^L, U^L|X^L, K^L}(w^L, u^L|x^L, k^L)$ such that:

$$\sum_{w^L, u^L, x^L, k^L} p_{W^L, U^L|X^L, K^L}(w^L, u^L|x^L, k^L) p_{X^L, K^L}(x^L, k^L) \leq D^E \quad (25)$$

and

$$p_{W^N, U^N|X^N, K^N}(w^N, u^N|x^N, k^N) = \prod_{i=1}^{N/L} p_{W^L, U^L|X^L, K^L}(w_i^L, u_i^L|x_i^L, k_i^L). \quad (26)$$

Under these specified conditions, the maximum achievable rate will be defined similarly to [8] according to the following Conjecture.

Conjecture 2: A maximum achievable rate over the defined blockwise memoryless channel is given by

$$C_{X, \Sigma_X^2}^{10,01} = \frac{1}{L} \max_{p(u^L, w^L|x^L)} \left[I(U^L; Y^L, \Sigma_X^{2L}) - I(U^L; X^L) \right]. \quad (27)$$

Since the proof of this conjecture closely follows the strategy presented in [8], we refer the interested reader to this reference for more details.

It is important to note that according to the made assumptions of asymmetrically available side information about the channel state (host realization is available at the encoder and only local variances are provided to the decoder) the following statement is true. According to the link provided by parallel source splitting between global i.i.d. statistical data models and conditionally i.i.d. local models, the side information X^L available to the encoder is following a distribution $p(x^L) = \prod_{i=1}^L p(x_i)$. Thus, without loss of optimality it is assumed that the encoder in the blockwise memoryless case is equivalent to the one designed for a memoryless case. Thus, one can conclude that $I(U^L; X^L) = LI(U; X)$. Therefore, (27) can be rewritten as follows:

$$C_{X, \Sigma_X^2}^{10,01} = \frac{1}{L} \max_{p(u, w|x)} \left[I(U^L; Y^L, \Sigma_X^{2L}) - LI(U; X) \right], \quad (28)$$

Moreover, further considering $I(U^L; Y^L, \Sigma_X^{2L})$ one obtains:

$$\begin{aligned} I(U^L; Y^L, \Sigma_X^{2L}) &= H(U^L) - H(U^L|Y^L, \Sigma_X^{2L}) \\ &= \sum_{i=1}^L H(U_i) - \sum_{i=1}^L H(U_i|U^{i-1}, Y^L, \Sigma_X^{2L}) \end{aligned} \quad (29)$$

$$= \sum_{i=1}^L H(U_i) - \sum_{i=1}^L H(U_i|Y^L, \Sigma_X^{2L}) \quad (30)$$

$$= \sum_{i=1}^L H(U_i) - \sum_{i=1}^L H(U_i|Y_1, \dots, Y_i, \dots, Y_L, \Sigma_{X_1}^2, \dots, \Sigma_{X_i}^2, \dots, \Sigma_{X_L}^2) \quad (31)$$

$$\geq \sum_{i=1}^L H(U_i) - \sum_{i=1}^L H(U_i|Y_i, \Sigma_{X_i}^2) \quad (32)$$

$$= LH(U) - LH(U|Y, \Sigma_X^2), \quad (33)$$

where (29) is obtained based on the encoder structure and chain rule for entropy [26], where $U^{i-1} = \{U_1, U_2, \dots, U_{i-1}\}$; (30) follows from the i.i.d. generation of the codewords at the encoder; (32) is obtained since conditioning reduces entropy.

Therefore, it is possible to bound (27) as follows:

$$C_{X, \Sigma_X^2}^{\prime 10, 01} \geq \max_{p(u, w|x)} [I(U; Y, \Sigma_X^2) - I(U; X)]. \quad (34)$$

Thus, one can claim that purely memoryless consideration of the problem of hidden information transfer maximization over a blockwise memoryless channel with state information asymmetrically available to encoder and decoder is lower bounded by the capacity of such a memoryless channel.

In the remaining part of the paper we will investigate some particular aspects of the obtained lower bound.

First, it should be noticed that $\Sigma_X^2 \rightarrow X \rightarrow U$ form a Markov chain. Thus Σ_X^2 and U are independent given X and consequently $I(U; \Sigma_X^2|X) = 0$. Therefore, one can further develop (22) as:

$$\begin{aligned} I(U; Y, \Sigma_X^2) - I(U; X) &= I(U; Y, \Sigma_X^2) - I(U; X) - I(U; \Sigma_X^2|X) \\ &= I(U; Y, \Sigma_X^2) - I(U; X, \Sigma_X^2) = I(U; Y|\Sigma_X^2) - I(U; X|\Sigma_X^2), \end{aligned} \quad (35)$$

where the last equation follows from (21).

Substituting (35) into (22), one obtains:

$$\begin{aligned}
C_{X,\Sigma_X}^{10,01} &= \max_{p(u,w|x)} [I(U; Y|\Sigma_X^2) - I(U; X|\Sigma_X^2)] \\
&= E_{p_{\Sigma_X^2}} \left[\max_{p(u,w|x)} [I(U; Y|\Sigma_X^2 = \sigma_X^2) - I(U; X|\Sigma_X^2 = \sigma_X^2)] \right] \\
&= \int_0^\infty p_{\Sigma_X^2}(\sigma_X^2) \left[\max_{p(u,w|x)} [I(U; Y|\Sigma_X^2 = \sigma_X^2) - I(U; X|\Sigma_X^2 = \sigma_X^2)] \right] d\sigma_X^2,
\end{aligned} \tag{36}$$

where the expectation is performed with respect to the distribution of host statistics $p_{\Sigma_X^2}(\sigma_X^2)$.

Contrarily, in the case when the host statistics are available at the encoder (dashed line in Figure 4), the above set-up should also incorporate this information in the design of the encoder defined by $p(u, w|x, \sigma_X^2)$ leading to the capacity $C_{X,\Sigma_X}^{10,11}$:

$$C_{X,\Sigma_X}^{10,11} = \max_{p(u,w|x,\sigma_X^2)} [I(U; Y|\Sigma_X^2) - I(U; X|\Sigma_X^2)], \tag{37}$$

the analysis of which is out of the scope of this paper. The particularities of this set-up have been considered by Moulin and O'Sullivan [8]. In our formulation, this result follows from the distribution $p(x, \sigma_X^2)p(u, w|x, \sigma_X^2)p(y|w, x) = p(\sigma_X^2)p(x|\sigma_X^2)p(u|x, \sigma_X^2)p(w|u, x, \sigma_X^2)p(y|w, x)$. It differs from the set-up defined in conjecture 1 by the design of auxiliary random variable $p(u|x, \sigma_X^2)$ and watermark generation part $p(w|u, x, \sigma_X^2)$ that potentially might require to provide an optimal power allocation according to the available statistics Σ_X^2 .

It should be also pointed out that:

$$\begin{aligned}
&I(U; Y|\Sigma_X^2) - I(U; X|\Sigma_X^2) \\
&= H(U|\Sigma_X^2) - H(U|Y, \Sigma_X^2) - H(U|\Sigma_X^2) + H(U|X, \Sigma_X^2) \\
&= H(U|X, \Sigma_X^2) - H(U|Y, \Sigma_X^2) \\
&\leq H(U|X, \Sigma_X^2) - H(U|Y, X, \Sigma_X^2) \\
&= I(U; Y|X, \Sigma_X^2) \leq I(W; Y|X, \Sigma_X^2),
\end{aligned} \tag{38}$$

where the first inequality is due to the reduction of entropy by conditioning and the last one is the consequence of data processing inequality [26].

Thus, $C_{X,\Sigma_X}^{10,01}$ is less than the capacity if both encoder and decoder have access to X^N and the decoder to Σ_X^{2N} , i.e.:

$$C_{X,\Sigma_X}^{11,01} = \max_{p(w|x)} I(W; Y|X, \Sigma_X^2), \tag{39}$$

and if both encoder and decoder have access to X^N and Σ_X^{2N} , i.e.:

$$C_{X, \Sigma_X^2}^{11,11} = \max_{p(w|x, \sigma_X^2)} I(W; Y|X, \Sigma_X^2), \quad (40)$$

that is an equivalent of the Wolfowitz problem (19).

Consider the expectation term in (36) for the fixed statistics $\Sigma_X^2 = \sigma_X^2$ under the AWGN attack that corresponds to the Gaussian set-up investigated by Costa. In this case, the internal maximization problem can be expressed as the rate $R(\alpha, \sigma_X^2)$ in (12).

Obviously, if the noise variance is perfectly known at the encoder, the Costa set-up reaches the channel capacity and there is no real necessity to use the host statistics at the decoder. However, if the attacking channel state is unknown at the encoder, the selection of optimal α for all possible attacks even in the scope of the AWGN scenario is an ambiguous problem. In this paper, we assume that the data-hider knows the attack distribution but the attack variance is an unknown parameter that can vary from one application to another. In this paper, we will address the Gaussian attack that is shown to be the worst case attack against Costa set-up within the class of additive noise attacks according to the information-theoretic game [9].

Therefore, for the generic α and corresponding rate $R(\alpha, \sigma_X^2)$ (12), the equation (36) can be rewritten as:

$$R_{X, \Sigma_X^2}^{10,01}(\alpha) = \int_0^\infty R(\alpha, \sigma_X^2) p_{\Sigma_X^2}(\sigma_X^2) d\sigma_X^2, \quad (41)$$

where the defined rate can only approach channel capacity for the optimal selection of α since $R(\alpha = \alpha_{opt}, \sigma_X^2) = C^{AWGN}$ according to Costa results [4]. Finally, it should be pointed out that the following inequality holds:

$$R_{X, \Sigma_X^2}^{10,01}(\alpha) \leq R(\alpha_{opt}) \quad (42)$$

with the equality for $\alpha = \alpha_{opt}$. The equality follows from the fact that for the informed encoder, that is aware of the noise variance, $\alpha = \alpha_{opt}$ and $R(\alpha_{opt})$ coincides with C^{AWGN} according to (13) that does not depend on the host variance, i.e., $R(\alpha_{opt}, \sigma_X^2) = R(\alpha_{opt})$. Thus:

$$R_{X, \Sigma_X^2}^{10,01}(\alpha_{opt}) = R(\alpha_{opt}) \int_0^\infty p_{\Sigma_X^2}(\sigma_X^2) d\sigma_X^2 = C^{AWGN}. \quad (43)$$

In the following, we will consider some particular cases of different α selection to link our new set-up with several well-known data-hiding techniques.

5.1 SS data-hiding: host statistics at the decoder

In the following, we will consider the SS data-hiding as a particular case of the Costa set-up when $\alpha = 0$. In this case, the auxiliary random variable $U = W + \alpha X = W$ is host independent (no host state is taken into account for the design of the watermark at the encoder).

This choice of the compensation parameter corresponds to the asymptotic case of very low-WNR regime when $\sigma_Z^2 \rightarrow \infty$. For these specific conditions, the SS data-hiding is known to approach the capacity of the AWGN channel.

The corresponding rate (12) for $\alpha = 0$ is:

$$R(0, \sigma_X^2) = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_W^2}{\sigma_X^2 + \sigma_Z^2} \right) \quad (44)$$

that represents the well-known result for the capacity of SS systems.

Therefore, to approach the capacity of the Gel'fand-Pinsker problem for the Gaussian set-up at the low-WNR it is sufficient to have only one codeword in each message bin since $I(U; X) = 0$ and $2^{NI(U; X)} = 1$. This means that the selection of the codeword W is host-independent and the classical SS-type communications based on the random coding can approach capacity. However, at the high-WNR regime this scheme sacrifices from the considerable host interference that requires to increase the amount of codewords in each bin depending on the host statistics (variance).

Under this assumption, equation (41) represents the rate of spread-spectrum data-hiding with side information about host statistics at the decoder.

Conjecture 3: *If the Laplacian host realization is not taken into account at the encoder and the host statistics are used at the decoder according to the source splitting model, then the achievable rate of the scheme is:*

$$R_{X, \Sigma_X^2}^{10,01}(0) = \int_0^\infty \frac{1}{2} \log_2 \left(1 + \frac{\sigma_W^2}{\sigma_X^2 + \sigma_Z^2} \right) p_{\Sigma_X^2}(\sigma_X^2) d\sigma_X^2. \quad (45)$$

5.2 DM data-hiding: host statistics at the decoder

By analogy with the DM ($\alpha' = 1$), we will also recall the Costa set-up for $\alpha = 1$. This corresponds to the encoder adaptation to the asymptotic situation of very high-WNR regime when $\sigma_Z^2 \rightarrow 0$ and $\alpha \rightarrow 1$. For this conditions, the DM is known to approach the capacity of the AWGN channel using high-

dimensional lattices (we neglect here 1.53 dB shaping loss due to uniform pdf of the watermark contrarily to the Gaussian pdf for the scalar case) [27].

In this case, the Costa auxiliary random variable $U = W + \alpha X = W + X$. Thus, $I(U; X) = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_X^2}{\sigma_W^2} \right)$ which requires an infinite number of codewords for each message bin when $\sigma_X^2 \rightarrow \infty$. Obviously, the design of the watermark W is host-state-dependent and the capacity achieving scheme is based on the random binning argument.

The corresponding rate (12) for $\alpha = 1$ is:

$$R(1, \sigma_X^2) = \frac{1}{2} \log_2 \left(\frac{\sigma_W^2}{\sigma_Z^2} + \frac{\sigma_W^2}{\sigma_X^2 + \sigma_Z^2} \right). \quad (46)$$

Conjecture 4: *If the Laplacian host realization is taken into account at the encoder based on the random binning argument and the host statistics are used at the decoder according to the source splitting model, then the achievable rate of the scheme is:*

$$R_{X, \Sigma_X^2}^{10,01}(1) = \int_0^\infty \frac{1}{2} \log_2 \left(\frac{\sigma_W^2}{\sigma_Z^2} + \frac{\sigma_W^2}{\sigma_X^2 + \sigma_Z^2} \right) p_{\Sigma_X^2}(\sigma_X^2) d\sigma_X^2. \quad (47)$$

6 Experimental results

To perform a fair comparison of the proposed approach we will compare different methods under the AWGN attack. Figure 6 summarizes the known results for the Costa set-up with the optimal selection of the compensation parameter in order to approach the capacity of the AWGN channel (13). We present on the same plot the performance of practical discrete approximations of the Costa scheme based on the binary-SCS with correspondent optimally selected compensation parameter and the binary DM [6] as well as the performance of the SS-based methods for WIR=-6 dB and WIR=-16 dB for Gaussian and Laplacian hosts.

The difference in the achievable rates for the SS-based methods between Gaussian and Laplacian hosts is not significant. It manifests itself only in the high-WNR regime. Since the host signal is acting as the interference to the watermark, the achievable rate is higher for the Laplacian host since its interference influence is smaller in comparison to the Gaussian host with the same variance.

To investigate the impact of the partial side information at the decoder according to the proposed framework we performed the analysis of the uninformed

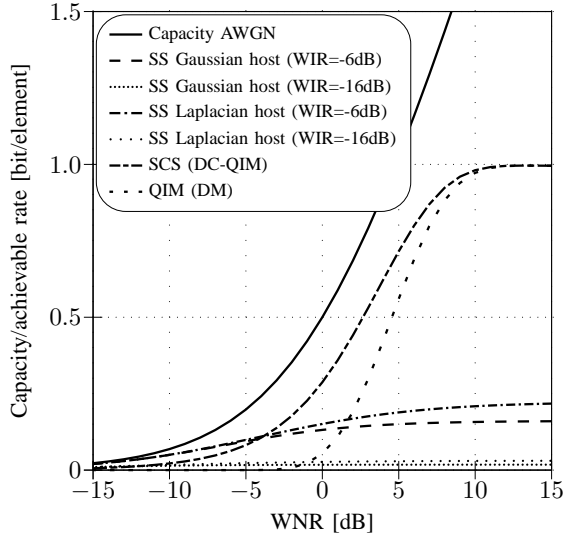


Fig. 6. Achievable rates of Costa set-up, SCS, DC-DM and SS for WIR=-6 dB and WIR=-16 dB in assumption of Gaussian and Laplacian hosts.

decoder in terms of host statistics according to the Costa rate (12) for various values of the compensation parameter α and two WIRs equal to -6 dB and -16 dB shown in Figures 7 and 9, respectively. The achievable rates of Costa set-up with partial side information at the decoder $R_{X,\Sigma_X^2}^{10,01}(\alpha)$ according to (41) for WIR=-6 dB and WIR=-16dB are obtained by numerical integration and are presented in Figures 8 and 10, respectively. The difference $R_{X,\Sigma_X^2}^{10,01}(\alpha) - R(\alpha)$ is shown in Figures 11 and 12 for WIR=-6 dB and WIR=-16 dB, respectively.

For $\alpha = 0$ that corresponds to the SS-based data-hiding, $R(0)$ approaches channel capacity at the low-WNR (Figures 7,9) for both WIRs. However, at the high-WNR regime, the host variance has a crucial impact on the system performance that is observed as a considerable rate decrease (especially for WIR=-16 dB). Using partial side information at the decoder, the rate $R_{X,\Sigma_X^2}^{10,01}(0)$ is significantly increased at the high-WNR regime for both WIRs with respect to the rate $R(0)$.

Considering another asymptotic case of $\alpha = 1$, which corresponds to the DC-DM-based selection of compensation parameter and scheme adaptation to the high-WNR, we observe that $R(1)$ approaches the AWGN channel capacity. Contrarily, at the low-WNR regime, its performance is considerably degraded due to the overestimated number of codewords in each message bin and a corresponding high probability of error at the decoder attempting to seek a jointly typical $u^N(m, j)$ and y^N . The proposed set-up $R_{X,\Sigma_X^2}^{10,01}(1)$ again outperforms $R(1)$ in this case.

Assuming the targeted range of operational WNRs to be $[-5; 10]$ dB one can select the compensation parameter in the range of $0.2 \leq \alpha \leq 0.4$ to resolve the trade-off between the host interference cancellation and system

robustness under attack channel ambiguity. This selection of α requires a limited number of codewords in each bin to cope with the host interference cancellation problem and an informed “adaptive” decoder that will perform the estimation of “channels goodness” prior to the decoding. The design of practical data-hiding schemes based on the proposed set-up is an important and challenging problem that will be a subject of our future research.

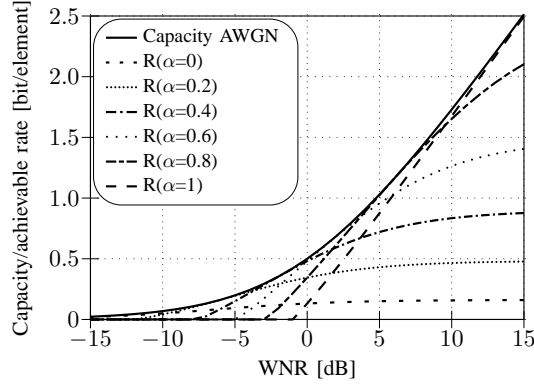


Fig. 7. Achievable rates of Costa set-up $R(\alpha, \sigma_X^2)$ for Gaussian host with different α and WIR=-6 dB.

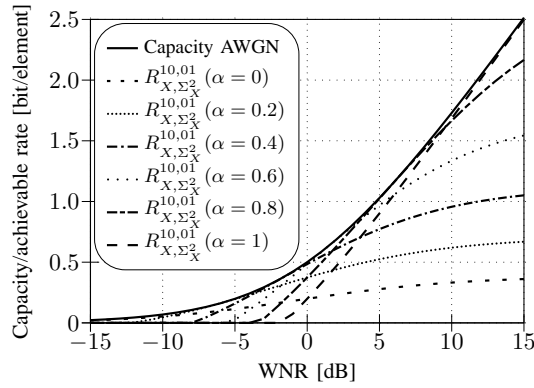


Fig. 8. Achievable rates of Costa set-up with partial side information at the decoder $R_{X, \Sigma_X^2}^{10,01}(\alpha)$ for WIR=-6 dB.

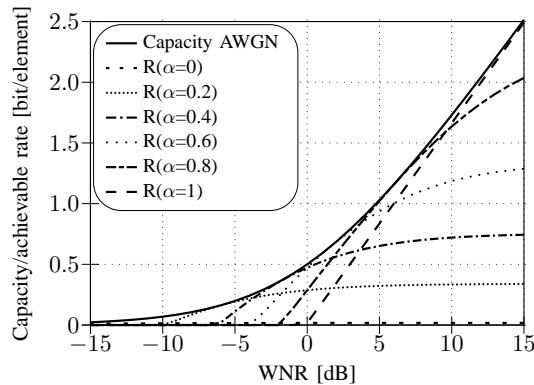


Fig. 9. Achievable rates of Costa set-up $R(\alpha, \sigma_X^2)$ for Gaussian host with different α and WIR=-16 dB.

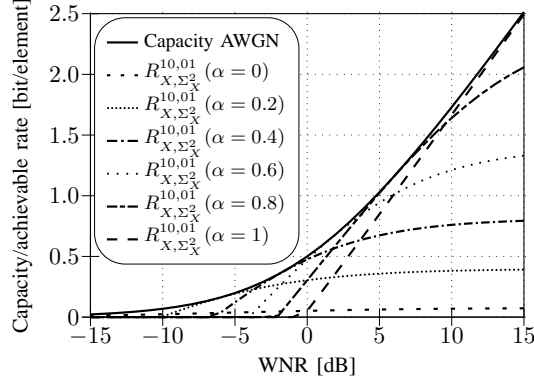


Fig. 10. Achievable rates of Costa set-up with partial side information at the decoder $R_{X, \Sigma_X^2}^{10,01}(\alpha)$ for WIR=-16 dB.

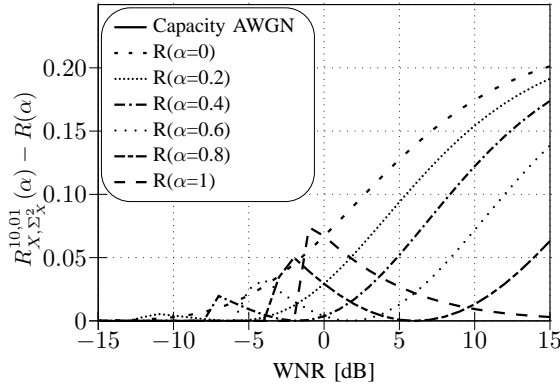


Fig. 11. Rate gain as the difference $R_{X, \Sigma_X^2}^{10,01}(\alpha) - R(\alpha, \sigma_X^2)$ for WIR=-6 dB.

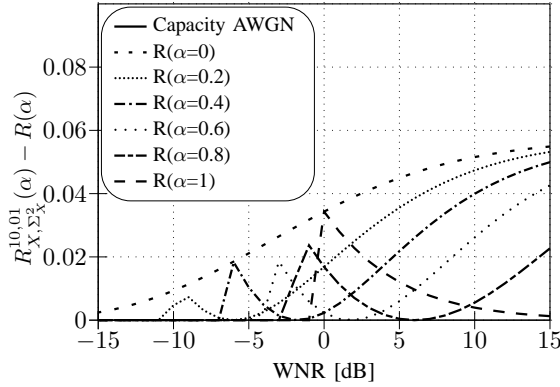


Fig. 12. Rate gain as the difference $R_{X, \Sigma_X^2}^{10,01}(\alpha) - R(\alpha, \sigma_X^2)$ for WIR=-16 dB.

7 Conclusions and future work

In this paper, we considered robust data-hiding with host state available at the encoder and partial side information at the decoder. We demonstrated that the knowledge of host statistics at the decoder can relax the critical requirements of quantization-based methods concerning attack channel statistics ambiguity

at the encoder. In our analysis we have considered two well-known set-ups based on the SS and DC-DM techniques. Traditionally, SS-based methods are considered as a particular case of the Costa set-up when α is fixed and equal to zero disregarding the actual WNR that corresponds to the encoder adaptation to the low-WNR regime. The DC-DM is considered as an asymptotic case when $\alpha = 1$.

In this paper, the mismatch in the assumption concerning α and operational WNR is compensated by the proper modeling of host at the decoder that considerably increases the performance of the SS-based methods at the high-WNR regime as well as the performance of quantization-based methods at the low-WNR regime.

Following the Gel'fand-Pinsker binning strategy, we have investigated the proposed asymmetric set-up for various values of α . When $\alpha = 0$, each message has only one codeword in each bin disregarding host state. For small values of $\alpha \leq 0.4$ and the side information at the decoder, one can find a good trade-off between the approaching capacity at the low-WNR regime and compensation of host interference at the high-WNR regime.

The future extensions include two main lines of research. New practical quantization-based methods can be designed taking into account host statistics and the work on this emerging issue is a subject of our ongoing research.

Amplitude scaling, or so-called value-metric attack, or equivalently fading is one more important problem of practical quantization-based algorithms. We believe that using the proposed approach one can also find a solution to this problem under proper watermark power control.

It would be also interesting to analyse the rate loss due to the mismatch between host statistics at the encoder and decoder, the so-called imperfect side information, that should be an important issue for practical data-hiding schemes.

Acknowledgment

This paper was partially supported by SNF PB grant 114613 and SNF projects 20020-121635 and 200021-119770 and Swiss IM2 project. The authors acknowledge the help and stimulating seminar on the subject of random binning and Gel'fand-Pinsker problem given by Prof. I.E. Telatar (EPFL, Switzerland) for the members of the Stochastic Information Processing (SIP) group. S. Voloshynovskiy and O. Koval thank to Prof. A. Lapidoth and Dr. G. Kramer for numerous stimulating discussions during an MIT course given at the ETH Zurich. The authors are thankful to the members of SIP group for many hours

spent during group seminars in clarifying the considered set-ups and also to the anonymous reviewers for their valuable comments.

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.

Appendix A

We briefly consider here the achievability part of Conjecture 1. The main idea is to resolve the trade-off between the number of codewords needed at the encoder at each bin of the message to cancel host interference and the number of uniquely distinguishable codewords at the decoder.

We assume that $m \in \{1, 2, \dots, |\mathcal{M}|\}$, $j \in \{1, 2, \dots, J\}$ with $|\mathcal{M}| = 2^{NR}$ and $J = 2^{NR'}$.

Code construction: Introduce an auxiliary random variable U with alphabet \mathcal{U} via $p_{U|X}(\cdot)$. Generate $J|\mathcal{M}|$ codewords $u^N(m, j)$ independently at random according to the marginal $p_U(\cdot)$ and allocate them into $|\mathcal{M}|$ bins with $J = 2^{NR'}$ codewords in each bin for every message m .

Encoder: Given the message to be sent m and host signal x^N , the encoder seeks a codeword $u^N(m, j)$ such that $(u^N(m, j), x^N) \in A_\epsilon^{*(N)}(U, X)$, i.e., the encoder seeks a jointly typical pair $(u^N(m, j), x^N)$ in the set of strongly jointly typical sequences $A_\epsilon^{*(N)}(U, X)$. Here, we use the definition of the *strongly typical set* [26], p. 358 $A_\epsilon^{*(N)}(X)$ with respect to $p_X(\cdot)$ that is the set of N -tuples x^N satisfying:

$$A_\epsilon^{*(N)}(X) = \begin{cases} x^N : \text{for all } a \in \mathcal{X} \\ \mathbf{N}(a|x^N) = 0, \text{ if } p_X(a) = 0, \\ \left| \frac{1}{N} \mathbf{N}(a|x^N) - p_X(a) \right| < \frac{\epsilon}{|\mathcal{X}|}, \end{cases} \quad (48)$$

where $\mathbf{N}(a|x^N)$ is the number of a occurrences in the sequence x^N and ϵ is an arbitrary small positive constant.

The *strongly jointly typical* sequences (x^N, y^N) with respect to the joint distribution $p_{XY}(\cdot)$ on $\mathcal{X} \times \mathcal{Y}$ satisfy:

$$A_\epsilon^{*(N)}(X, Y) = \begin{cases} (x^N, y^N) : \text{for all } a \in \mathcal{X}, b \in \mathcal{Y}, \\ \mathbf{N}(a, b|x^N, y^N) = 0, \text{ if } p_{XY}(a, b) = 0, \\ \left| \frac{1}{N} \mathbf{N}(a, b|x^N, y^N) - p_{XY}(ab) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}, \end{cases} \quad (49)$$

where $\mathbf{N}(a, b|x^N, y^N)$ is the number of the pair (a, b) occurrences in the pair of sequences (x^N, y^N) .

Therefore, the message m defines the bin and the host realization selects a particular $u^N(m, j)$ from this bin. If such a codeword $u^N(m, j)$ is found, the encoder produces the watermark $w^N = f^N(u^N(m, j), x^N)$.

We must bound the probability that there is no codeword that is strongly jointly typical with X^N . In fact, this probability should go to zero as $N \rightarrow \infty$ and $\epsilon \rightarrow 0$.

The probability that the given sequence x^N has not a jointly typical pair with an independently chosen codeword $u^N(m, j)$, $j = \{1, 2, \dots, 2^{NR'}\}$ for a given message m is $\prod_{j=1}^{2^{NR'}} \Pr[(U^N(m, j), x^N) \notin A_\epsilon^{*(N)}(U, X)]$. The average probability for all X^N coming from the distribution $p_{X^N}(x^N)$ and belonging to $x^N \in A_\epsilon^{*(N)}(X)$ is:

$$P_{e_1} = \sum_{x^N \in A_\epsilon^{*(N)}(X)} p_{X^N}(x^N) \prod_{j=1}^{2^{NR'}} \Pr[(U^N(m, j), x^N) \notin A_\epsilon^{*(N)}(U, X)] \quad (50)$$

$$= \sum_{x^N \in A_\epsilon^{*(N)}(X)} p_{X^N}(x^N) \left[1 - \Pr[(U^N(m, j), x^N) \in A_\epsilon^{*(N)}(U, X)] \right]^{2^{NR'}} \quad (51)$$

From Lemma 13.6.2 [26], p. 359, we have:

$$\Pr[(U^N(m, j), x^N) \in A_\epsilon^{*(N)}(U, X)] \geq 2^{-N[I(U;X)+\delta]}, \quad (52)$$

where δ goes to zero as $\epsilon \rightarrow 0$ and $N \rightarrow \infty$.

Substituting this in (51) and using inequality $(1-x)^n \leq e^{-nx}$, we have:

$$P_{e_1} \leq e^{-2^{NR'} 2^{-N[I(U;X)+\delta]}}, \quad (53)$$

which goes to 0 as $N \rightarrow \infty$ if $R' > I(U; X)$ or equivalently if $J > 2^{N[I(U;X)+\delta]}$. This suggests to have a relatively large number J .

Decoder: Given the signal on the output of the channel y^N and the side information σ_X^{2N} , the decoder seeks a codeword $u^N(m, j)$ satisfying

$$(u^N(m, j), y^N, \sigma_X^{2N}) \in A_\epsilon^{*(N)}(U, Y, \Sigma_X^2)$$

in the set of all codewords $1 \leq j \leq J$, $1 \leq m \leq |\mathcal{M}|$, i.e., among all $J|\mathcal{M}|$ codewords. If the decoder finds only one unique jointly typical triple, it declares that the sent message was $\hat{m} = m$. Otherwise, an error is declared.

Suppose $y^N \in A_\epsilon^{*(N)}(Y)$ and the decoder finds an $\tilde{m} = m$ and \tilde{j} such that $(u^N(\tilde{m}, \tilde{j}), y^N, \sigma_X^{2N}) \in A_\epsilon^{*(N)}(U, Y, \Sigma_X^2)$. Denote probability of this event as $\Pr[(U^N(\tilde{m}, \tilde{j}), y^N, \sigma_X^{2N}) \in A_\epsilon^{*(N)}(U, Y, \Sigma_X^2)]$. Excluding the correct $\hat{m} = m$, the total number of codewords $u^N(\tilde{m}, \tilde{j})$ for which $\hat{m} \neq m$ is $(2^{NR} - 1)2^{NR'}$ or $(|\mathcal{M}| - 1)J$. Therefore, the total probability of such kind of error is:

$$P_{e_2} = (2^{NR} - 1)2^{NR'} \Pr[(U^N(\tilde{m}, \tilde{j}), y^N, \sigma_X^{2N}) \in A_\epsilon^{*(N)}(U, Y, \Sigma_X^2)]. \quad (54)$$

According to [26], the following upper bound exists: $\Pr[(U^N(\tilde{m}, \tilde{j}), y^N, \sigma_X^{2N}) \in A_\epsilon^{*(N)}(U, Y, \Sigma_X^2)] \leq 2^{-N[I(U; Y, \Sigma_X^2) - \delta]}$. Therefore, the probability P_{e_2} can be bounded as:

$$P_{e_2} < 2^{N[R+R']} 2^{-N[I(U; Y, \Sigma_X^2) - \delta]}. \quad (55)$$

Thus, we require that ϵ is small, N is large and $R + R' < I(U; Y, \Sigma_X^2)$ to have $P_{e_2} \rightarrow 0$. Obviously, the higher $|\mathcal{M}|$ and J , the higher probability of incorrect decoding.

Encoder-decoder trade-off in selection of J : The selection of J should resolve the encoder/decoder trade-off for a given $|\mathcal{M}|$. From one side J should be sufficiently large to make the encoder failure probability low, i.e., to guarantee the existence of one jointly typical pair at the encoder $J2^{-NI(U; X)} > 1$ or equivalently $\frac{1}{N} \log_2 J > I(U; X)$. From the other side, J should be small enough to avoid the failure of the decoder, i.e., $\frac{1}{N} \log_2 |\mathcal{M}| + \frac{1}{N} \log_2 J < I(U; Y, \Sigma_X^2)$. If J is chosen correctly, $R < I(U; Y, \Sigma_X^2) - I(U; X)$ that corresponds to equation given in Conjecture 1.

References

- [1] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. Morgan Kaufmann Publishers, Inc., San Francisco, 2001.
- [2] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, "Attack modelling: Towards a second generation watermarking benchmark," *Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking*, vol. 81, pp. 1177–1214, June 2001.
- [3] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Probl. Control and Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
- [4] M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

- [5] B. Chen and G. W. Wornell, “Quantization index modulation: A class of provably good methods for digital watermarking and information embedding,” *IEEE Trans. on Information Theory*, vol. 47, pp. 1423–1443, May 2001.
- [6] J. Eggers, J. Su, and B. Girod, “A blind watermarking scheme based on structured codebooks,” in *Secure images and image authentication, IEE Colloquium*, London, UK, April 2000.
- [7] L. Perez-Freire, F. Perez-Gonzalez, 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, pp. 235–238.
- [8] P. Moulin and J. O’Sullivan, “Information-theoretic analysis of information hiding,” *IEEE Trans. on Information Theory*, vol. 49, pp. 563–593, March 2003.
- [9] P. Moulin, “The role of information theory in watermarking and its application to image watermarking,” *Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking*, vol. 81, no. 6, pp. 1121–1139, 2001.
- [10] P. Moulin and M. K. Mihcak, “A framework for evaluating the data-hiding capacity of image sources,” *IEEE Trans. on Image Processing*, vol. 11, no. 9, pp. 1029–1042, September 2002.
- [11] A. Hjorungnes, J. Lervik, and T. Ramstad, “Entropy coding of composite sources modeled by infinite gaussian mixture distributions,” in *IEEE Digital Signal Processing Workshop*, 20-24 January 1996, pp. 235–238.
- [12] T. Cover and M. Chiang, “Duality between channel capacity and rate distortion with two-sided state information,” *IEEE Transactions on Information Theory*, vol. 48, pp. 1629–1638, June 2002.
- [13] A. Cohen and A. Lapidoth, “The gaussian watermarking game,” *IEEE Transactions on Information Theory*, vol. 48, pp. 1639 – 1667, June 2002.
- [14] G. Voyatzis and I. Pitas, “The use of watermarks in the protection of digital multimedia products,” *Proceedings of IEEE*, vol. 87, pp. 1197–1207, July 1999.
- [15] J. L. Cannons and P. Moulin, “Design and statistical analysis of a hash-aided image watermarking system,” *IEEE Trans. on Image Processing*, vol. 13, pp. 1393–1406, October 2004.
- [16] N. Jayant and P. Noll, *Digital Coding of Waveforms*. London: Prentice-Hall Int., 1984.
- [17] Y. Yoo, A. Ortega, and B. Yu, “Image subband coding using context-based classification and adaptive quantization,” *IEEE Trans. on Image Processing*, vol. 8, pp. 1702–1215, August 1999.
- [18] P. Moulin and J. Liu, “Analysis of multiresolution image denoising schemes using generalized-gaussian and complexity priors,” in *IEEE Trans. on Information Theory*, vol. 45, no. 3, April 1999, pp. 909–919.

- [19] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Shift-invariant denoising using wavelet-domain hidden Markov trees," in *Conference Record of The Thirty-Third Asilomar Conference on Signals, Systems and Computers*, 1999.
- [20] C. Creusere, "A new method of robust image compression based on the embedded zerotree wavelet algorithm," *IEEE Transactions on Image Processing*, vol. 6, pp. 1436–1442, 1997.
- [21] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using gaussian scale mixtures in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, pp. 1338–1351, 2002.
- [22] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. on Image Processing*, vol. 10, pp. 1647–1658, March 2001.
- [23] R. Wilson, "Mgmm: Multiresolution gaussian mixture models for computer vision." in *ICPR*, 2000, pp. 1212–1215.
- [24] S. LoPresto, K. Ramchandran, and M. Orhard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Data Compression Conference 97*, Snowbird, Utah, USA, 1997, pp. 221–230.
- [25] J. Wolfowitz, *Coding Theorems in Information Theory*. Springer-Verlag, 3rd ed. New York, 1978.
- [26] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley and Sons, New York, 1991.
- [27] D. Forney and G. Ungerboeck, "Modulation and coding for linear gaussian channels," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2384–2415, 1998.