Partially reversible data hiding with pure message communications over state-dependent channels*

S. Voloshynovskiy [a] , O. Koval [a] , E. Topak [b] , J.E. Vila Forcen [c]
P. Comesana Alfaro [d] [1]

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

[b] *Motorola Turkey, Istanbul Turkey*

[c] *Carlos III University of Madrid,*
C/ Madrid 126, 28903 - Getafe (Madrid), Spain

[d] *Signal Processing in Communications Group,*
Signal Theory and Communications Department,
University of Vigo, 36310 Vigo, Spain

Abstract

In this paper, we analyze the reversibility of data hiding techniques based on random binning as a by-product of pure message communications. We demonstrate the capabilities of unauthorized users to perform hidden data removal using solely a signal processing approach based on optimal estimation as well as consider reversibility on the side of authorized users who have knowledge of the key used for the message hiding. In fact, we show that knowledge of the auxiliary random variable, used in the codebook construction of random binning techniques, is sufficient to perform the optimal reversibility procedure. We compare the optimal rate-distortion region results obtained using more involved coding strategies based on hybrid random binning with those utilizing uncoded transmission. This analysis is performed for the generalized Gel'fand-Pinsker formulation, Gaussian Costa setup and particular practical schemes based on structured codebooks. Finally, we consider some related open issues and possible future extensions.

[1] Further information: Send correspondence to S. Voloshynovskiy

* Part of this paper appeared as S. Voloshynovskiy, O. Koval, E. Topak, J.E. Vila-Forcen, P. Comesana, and T. Pun. On reversibility of random binning techniques: multimedia perspectives. In 9th Conference on Communications and Multimedia Security, CMS 2005, Lecture Notes in Computer Science. Springer-Verlag Heidelberg, September 2005.

Keywords: robust data hiding, Gel'fand-Pinsker setup, side information, reversibility, worst case attacks.

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].

Most of these applications are facing an important problem related to the host interference. The design of host interference cancellation critically relies on knowledge of the channel state at the encoder. The main assumption behind almost all current data hiding techniques consists in the availability of the state realization (host) at the encoder assuming some fixed attacker strategy. The related issue in communications under the assumption of a fixed channel with random parameters was considered by Gel'fand and Pinsker [2]. The Gel'fand-Pinsker setup is based on a *random binning* argument contrary to the classical *random coding* argument that does not take into account the channel state information for the codebook generation. Costa considered the Gel'fand-Pinsker problem in a Gaussian formulation and mean squared distortion criterion and demonstrated using *random binning*-based codebook design that the capacity of the Gaussian channel with the Gaussian interfering host can be equal to the capacity of interference-free communications [3]. Recent advantages in the design of practical capacity achieving codes makes this technique even more attractive for various purposes [4,5].

The wide practical use of the Gel'fand-Pinsker setup has raised a number of interesting problems related to its security and robustness in multimedia security applications as well as its reversibility in multimedia management and communications. Although these aspects seem to be unrelated from the first point of view, they have a lot of common issues that can be used for the optimal design of binning-based techniques.

Email addresses: svolos@unige.ch (S. Voloshynovskiy), Oleksiy.Koval@unige.ch (O. Koval), www.motorola.com/tr (E. Topak), jemilio@tsc.uc3m.es (J.E. Vila Forcen), pcomesan@gts.tsc.uvigo.es (P. Comesana Alfaro).

In particular, the growing demands of multimedia security require not only the robust marking of digital content by assigning and embedding some copyright related index (a.k.a. a message), but also a possibility to recover the original data after embedding. Such a need for the reversibility has emerged in those applications requiring the authentication where the embedded watermark might be interpreted as a content modification or distortion leading to some faulty conclusion. For example, the medical content watermarking is a very useful option for the content and patient privacy protection. However, once embedded, the watermark might cause certain visual degradation that can be incorrectly interpreted. Moreover, sometimes the information carried by a previous watermark might be outdated and the content owner or manager needs to replace it by a new one by removing the old one without content quality degradation. That is why it is highly desirable to have a reversibility option where the authorized user can remove his/her old watermark. These applications also include forensics of documents, art work authentication, remote sensing and imaging. Obviously, this option can be only executed by authorized parties who posses the correct secret key and it is commonly known as *reversible watermarking* or *data hiding*. Moreover, the reversibility or original content recovery can be demanded from both the watermarked version and its degraded counterpart. Therefore, the question of recovery accuracy requires additional careful investigation.

The traditional Gel'fand-Pinsker based data hiding methods do not address this issue due to the inherently different assumptions and requirements behind their design. It is worth mentioning that the main objective of these methods is the maximization of the data hiding rate, i.e., the maximum number of messages that can be embedded and reliably extracted from the degraded watermarked data under some embedding and channel distortion constraints, but not the construction of the best recovered copy of the original data. The first practical methods that addressed the reversible watermarking were probably the algorithms proposed by Honsinger *et. al.* [6] and Fridrich *et. al.* [7]. In these methods, the part of image represented by the least significant bits has been losslessly compressed and embedded back to the original image. Obviously, such an approach is not robust to various modifications and is only useful for authentication applications. Nevertheless, it represents a quite interesting concept that stimulated numerous publications [8–11], which tried to extend this approach mostly focusing on the practical issues of robustness to the distortion and increase of the data hiding rate. However, their deep theoretical investigation remained mostly uncovered.

To our best knowledge, the information-theoretic analysis of joint data hiding and host recovery problem was originally addressed in the publications of Martinian *et. al.* [12] and Willems and Kalker [13]. Sutivong *et. al.* [14] considered a similar formulation for state-dependent channels in digital communications, where the channel state needs to be estimated at the decoder jointly with

the communicated message. The extension of these coding strategies as well as error-exponent analysis in data hiding applications were performed in our recent work [15]. To achieve this joint coding-estimation goal, the original Gel'fand-Pinsker coding scheme is essentially changed in all the above publications. The generalized coding approach behind these techniques consists of the usage of two separate coding methods where one of which is addressing *pure watermark embedding* (message communications) and the other one is responsible for the *host recovery*.

More particularly, the main objective behind the approach of Martinian *et. al.* [12] is to demonstrate the existence of the embedding rate-restoration distortion pair for two critical cases. In following, we will refer to it as a rate-distortion pair. The first case targets the optimization problem where the distortion is minimized without any constraint on the achievable rate. Contrarily, the second case addresses the optimization problem where the rate is maximized without constraint on the restoration distortion. The other rate-distortion pairs have not been investigated in this work. A similar in spirit strategy was addressed by Willems and Kalker [13] who succeeded to demonstrate the extreme cases of: *reversible noise-free embedding*, i.e., reversible data hiding only for the perfect idealized noise-free channels as an extension of the work of Fridrich *et. al.* [7], and *reversible and robust embedding*, i.e., the perfect reversibility with zero-distortion and any possible data hiding rate that also corresponds to the first case of Martinian *et. al.*, and finally *partially reversible noise-free embedding*, i.e., only partial restoration of host is requested that allowed to increase the data hiding rate. However, the general coding strategy was not presented for joint message embedding and host recovery due to difficulties with the converse part of the coding theorem. Finally, Sutivong *et. al.* [14] were the first who presented this generalized setup for all achievable rate-distortion pairs based on the discussed hybrid coding.

Although the above theoretical findings are of significant interest for numerous applications, the used coding techniques assume essential deviation from the Gel'fand-Pinsker framework. It has an important consequence for a number of already existing and widely used techniques based on the Gel'fand-Pinsker coding, where no additional information is used to assist the host recovery and the coding can not be modified. We will refer to this feature as backward compatibility. Therefore, the reversibility of the Gel'fand-Pinsker based data hiding schemes still remains an unsolved and emerging problem. Moreover, the investigation of the theoretical limits of the Gel'fand-Pinsker based data hiding reversibility represents a huge interest for the security analysis revealing the worst watermark removal strategy in the scope of a watermark removal attack, which was practically investigated in our previous work [16]. Thus, knowledge of the reversibility limits might shed more light on the development of efficient countermeasures. Such a formulation would naturally harmonize with the noise-free embedding studied by Willems and Kalker, which however as-

sumed a different coding strategy. Moreover, such an analysis would be also extremely useful for imperfect channels that introduce certain causal distortions to the watermarked data and where one attempts at their compensation based on the hidden data [17]. Finally, both the Gel'fand-Pinsker coding and all the above theoretical reversibility techniques assume perfect knowledge of the channel distortions in advance that is obviously not the case for virtually all practical scenarios.

That is why we formulate the goal of this paper as the theoretical investigation of the Gel'fand-Pinsker based data hiding methods, i.e., the methods originally designed for pure message embedding but not for the reversibility, in the reversibility formulation. The solution to this problem should reveal the answers to a number of above problems, including:

(a) the backward compatibility;
(b) the authorized removal of a watermark with the purpose of recovery of the original content for authentication, forensic analysis or another watermark embedding;
(c) the security analysis with respect to the best unauthorized watermark removal;
(d) performance under prior ambiguity;
(e) the investigation of the gap between the achievable rate-distortion region of classical Gel'fand-Pinsker reversibility and complex hybrid schemes.

The problems we formulate in this paper are obviously different to those considered by Martinian *et. al.* [12], Willems and Kalker [13] and Sutivong *et. al.* [14] in both formulation and used coding strategies. The main difference consists in the part of the problem formulation where the above authors use the modified coding setup while the setup under our analysis is based on the random binning Gel'fand-Pinsker coding and reversibility is considered as a by-product of pure message communications.

At the same time, the closest to our formulation can be found in the work of Eggers *et. al.* [10], who considered the reversibility of quantization-based data hiding as a structured codebook approximation of the random binning scheme. Our work extends previous results in part of a generalized consideration of digital data hiding for random binning schemes based on the Gel'fand-Pinsker setup and Costa formulation [3]. Moreover, we also link our work with the authentication problem and state information transmission problem analyzed in [12] and [14], respectively. In fact, we will demonstrate that our setup designed for pure information hiding closely achieves the results based on more complex joint pure information and host data coding using power sharing considered in [14] under specific circumstances.

The paper has the following structure. The basic information-theoretic setup

of side information-assisted data hiding is considered in Section 2. Here we briefly review the necessary fundamentals of the data hiding concept with the host state at the encoder based on the random binning argument in the scope of the generalized Gel'fand-Pinsker problem, Costa's Gaussian setup and a discrete approximation of the Costa problem. Section 3 presents the analysis of reversibility problem for both unauthorized and authorized users. The results of computer simulation are presented in Section 4. Finally, Section 5 concludes the paper and presents some future research perspectives.

Notations We use capital letters to denote random variables X, small letters x to denote their realizations. The superscript N is used to designate length-N vectors $x^N = [x[1], x[2], ..., x[N]]^T$ with k^{th} element $x[k]$. We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to indicate that a discrete 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 $X \in \mathcal{X}$ with the cardinality $|\mathcal{X}|$. $\mathbb{R}^+$ denotes the set of non-negative real numbers. For the vectors, $U^N \in \mathcal{U}^N$ denotes all U^N from $\mathcal{U}^N$ and $U^N \in \mathcal{U}^N(K = k)$ denotes all U^N from the subclass of $\mathcal{U}^N$ defined by K. We use $H(X)$, $h(X)$ and $I(X;Y)$ to denote the entropy of X, differential entropy of X and the mutual information between X and Y respectively. $\mathbf{0}$ denotes a vector of zeros and $\mathbf{I}_N$ designates a $N \times N$ identity matrix. We also define the watermark-to-image ratio (WIR) as WIR=$10 \log_{10} \frac{\sigma_W^2}{\sigma_X^2}$ and the watermark-to-noise ratio (WNR) as 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 Gel'fand-Pinsker setup: random binning in data hiding

In this section, we briefly introduce the Gel'fand-Pinsker coding framework needed for the reversibility analysis. We also review its extension to the Gaussian setup performed by Costa [3] and introduce the known low-complexity approximation based on quantization based methods. We will need these results to explain the difference in the available prior information on the side of authorized user contrary to those in the possession of the opponent. In particular, the fundamental role of the auxiliary random variable and the corresponding codebook construction will be considered. This section has the reviewing and generalization character and can be omitted by the readers familiar with this framework.

The block-diagram of the Gel'fand-Pinsker problem in data hiding formulation similar to the setup analyzed in [18] is presented in Fig. 1.

According to this setup, the data hider has the access to the uniquely assigned

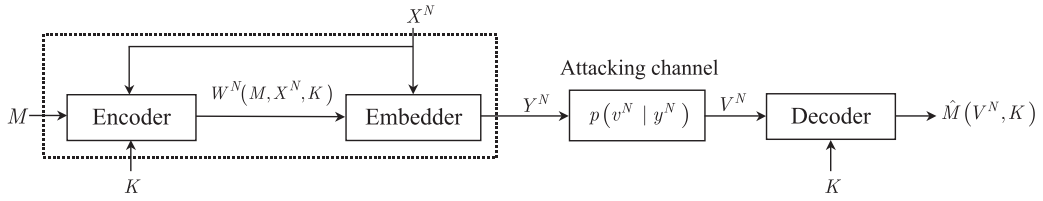


Fig. 1. Generalized Gel'fand-Pinsker channel coding with side information at the encoder: data hiding formulation.

secret key K that is uniformly distributed over the set $\mathcal{K} = \{1, 2, ..., |\mathcal{K}|\}$ and to the non-causally known interference $X^N \in \mathcal{X}^N$, $X^N \sim p_{X^N}(x^N)$. The message M, the key K and the non-causally known host realization X^N are used by the encoder to produce the 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 a certain attack to the watermarked data by passing Y^N through the attack channel $p_{V^N|Y^N}(v^N|y^N)$ that is assumed to be discrete memoryless one, i.e., $p_{V^N|Y^N}(v^N|y^N) = \prod_{i=1}^{N} p_{V|Y}(v_i|y_i)$. Finally, the decoder estimates the message $\hat{M}$ based on the attacked data V^N and the available key K. We will assume that the message $M \in \mathcal{M}$ is uniformly distributed over $\mathcal{M} = \{1, 2, ..., |\mathcal{M}|\}$, with $|\mathcal{M}| = 2^{NR}$, where R is the data hiding rate and N is the length of all involved vectors X^N, W^N, Y^N and V^N. It is assumed that the stego and attacked data are defined on $Y^N \in \mathcal{Y}^N$ and $V^N \in \mathcal{V}^N$, respectively. The distortion function is defined as:

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

where $d(x_i, y_i) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ denotes the element-wise distortion between x_i and y_i.

Definition 1: A *fixed discrete memoryless data hiding channel in Gel'fand-Pinsker formulation* consists of four alphabets $\mathcal{X}$, $\mathcal{W}$, $\mathcal{Y}$, $\mathcal{V}$ and a fixed transition probability matrix $p_{V^N|W^N,X^N}(v^N|w^N, x^N)$ that corresponds to the covert channel communications of the watermark W^N via $p_{Y^N|W^N,X^N}(y^N|w^N, x^N)$ with the host data X^N as the random parameter and the discrete memoryless attack channel $p_{V^N|Y^N}(v^N|y^N)$ such that $p_{V^N|W^N,X^N}(v^N|w^N, x^N) = \Sigma_{y^N} p_{Y^N|W^N,X^N}(y^N|w^N, x^N) p_{V^N|Y^N}(v^N|y^N)$. The attack channel is subject to the distortion constraint D^A:

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

where $p_{V^N|Y^N}(v^N|y^N) = \prod_{i=1}^{N} p_{V|Y} p(v_i|y_i)$.

Definition 2: A $(2^{NR}, N)$ code for the given data hiding channel consists of *message set* $\mathcal{M} = \{1, 2, ..., 2^{NR}\}$, *key set* $\mathcal{K} = \{1, 2, \ldots, |\mathcal{K}|\}$, *encoding*

function:

$$\phi^N : \mathcal{M} \times \mathcal{X}^N \times \mathcal{K} \to \mathcal{W}^N, \tag{3}$$

embedding function:

$$\varphi^N : \mathcal{W}^N \times \mathcal{X}^N \to \mathcal{Y}^N, \tag{4}$$

subject to the embedding distortion constraint D^E for a given user with $K = k$:

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

and *decoding function*:

$$g^N : \mathcal{V}^N \times \mathcal{K} \to \mathcal{M}. \tag{6}$$

We define the *average probability of error* for this code as:

$$P_e^{(N)} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \Pr[g^N(V^N, k) \neq m | M = m]. \tag{7}$$

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

Definition 4: The capacity of data hiding channel is the supremum of all achievable rates for a given distortion pair (D^E, D^A).

Theorem 1 (Data hiding capacity for the fixed channel) [18]: A rate R is achievable for a given distortion pair (D^E, D^A) and the discrete memoryless attack channel $p_{V^N|Y^N}(v^N|y^N) = \prod_{i=1}^N p_{V|Y} p(v_i|y_i)$, iff $R < C$ where:

$$C = \max_{p(u,w|x)} R, \tag{8}$$

where $R = I(U;V) - I(U;X)$ and U to be an auxiliary random variable $U \in \mathcal{U}(K = k)$, with $|\mathcal{U}(K = k)| \leq |\mathcal{X}||\mathcal{W}| + 1$ for all $K \in \{1, 2, \ldots, |\mathcal{K}|\}$. We assume $p(k, x^N, u^N, w^N, y^N, v^N) = p(k)p(x^N)p(u^N|x^N)\underline{\mathbf{1}}\{u^N \in \mathcal{U}^N(K = k)\}p(w^N|u^N, x^N)p(y^N|x^N, w^N)p(v^N|y^N)$ to reflect the technicality behind the codebook and watermark generations as well as channel degradations, where $\underline{\mathbf{1}}\{.\}$ denotes the indicator function.

We should also mention here that in a more general case we need to consider minimization with respect to $p(v^N|y^N)$ defined on the class of all channels with the bounded distortion [18]. However, the fixed channel model is considered in this paper to introduce the basic concept of reversibility due to the following reason. It is assumed that the data hiding game has a saddle point and the protocol is designed to achieve the optimal solution.

The proof of theorem 1 in the general case of active attacker is provided by Moulin and O'Sullivan [18] and the details can be found in the referred paper. However, it is important to emphasize that the main difference with our setup is the codebook construction and the corresponding interpretation of the user key. In the scope of this paper, the key K is considered uniquely as the index that defines the codebook of a particular user. When K is known, i.e., $K = k$, the entropy of a codeword U^N reduces from $H(U^N)|_{U^N \in \mathcal{U}^N}$ to $H(U^N)|_{U^N \in \mathcal{U}^N (K=k)} = 2^{N[R+R']}$ that corresponds to the total number of codewords in the binning technique considered by Gel'fand and Pinsker, where R' is the rate of host data X^N representation. Therefore, when K is unknown, $H(U^N)|_{U^N \in \mathcal{U}^N} = H(U^N)$. Contrary, Moulin and O'Sullivan have a broader understanding of the key as a sort of side information that can be in some relationship with X^N and is shared between the encoder and the decoder. Therefore, we assume that K is solely a cryptographic key that is independent of X^N. To make this difference more clear and to explain the considered communications setup, in the following, we detail the code construction.

Code construction: Introduce an auxiliary random variable U^N with an alphabet $\mathcal{U}^N$ via $p_{U^N|X^N}(.|.)$. Generate $|\mathcal{M}||\mathcal{J}||\mathcal{K}|$ distinct codewords $u^N(m, j, k)$, $m = \{1, 2, \cdots, |\mathcal{M}|\}$, $j = \{1, 2, \cdots, |\mathcal{J}|\}$, $k = \{1, 2, \cdots, |\mathcal{K}|\}$, with $|\mathcal{M}| = 2^{NR}$ and $|\mathcal{J}| = 2^{NR'}$ independently at random according to the marginal distribution $p_{U^N}(.)$ and distribute them into corresponding codebooks and bins. It is assumed for the sake of simplicity that each codeword uniquely appears only in one codebook. However, communications scenarios where the same codeword might appear in different codebooks can be considered as the alternative realization of permutational codebook construction according to randomized codes [19]. The existence of such a codebook can be shown according to sphere packing lemma [20]. Provided that the number of possible sequences that can be obtained from the distribution of codewords is much larger than the total number of codewords to be generated, i.e., $2^{H(U^N)} \gg |\mathcal{M}||\mathcal{J}||\mathcal{K}| = |\mathcal{K}|2^{N[R+R']}$, such a codebook construction is realizable. The resulting set of randomly designed codebooks composed of codewords $u^N(m, j, k)$ for $|\mathcal{K}|$ users is organized as shown in Fig. 2.

Since for the fixed $K = k$ the proof of the theorem follows a standard argument of the Gel'fand and Pinsker that can be found in [2], it will skipped here and the interested reader is encouraged to consult the referred paper.

2.1 Costa setup: Gaussian assumption

Costa considered the Gel'fand-Pinsker problem for the Gaussian context and mean-square error distance [3]. In our user-based codebook construction dependent on K, the corresponding fixed channel $p_{V^N|W^N,X^N}(v^N|w^N, x^N)$ is the

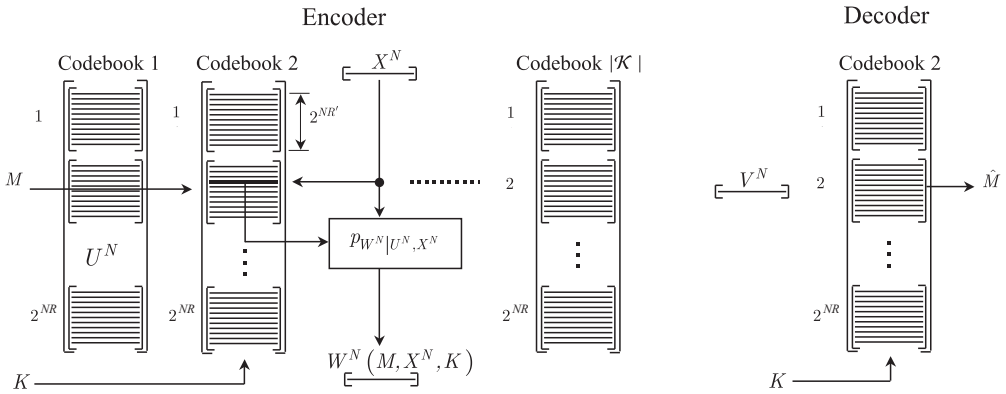


Fig. 2. Generalized Gel'fand-Pinsker data hiding codebook construction: encoding and decoding.

Gaussian one with $X^N \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$ and additive $Z^N \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$ (Fig. 3). The auxiliary random variable was chosen in the form $W^N = U^N - \alpha X^N$, $U^N \in \mathcal{U}^N(K = k)$, with optimization parameter α maximizing the rate:

$$R(\alpha) = \frac{1}{2} \log_2 \frac{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_Z^2)}{\sigma_W^2 \sigma_X^2 (1-\alpha)^2 + \sigma_Z^2(\sigma_W^2 + \alpha^2 \sigma_X^2)}. \tag{9}$$

Costa has shown that the optimal α is given by:

$$\alpha_{opt} = \frac{\sigma_W^2}{\sigma_W^2 + \sigma_Z^2}, \tag{10}$$

which requires knowledge of σ_Z^2 at the encoder. If $\alpha = \alpha_{opt}$, $R(\alpha)$ does not depend on the host variance and:

$$R(\alpha_{opt}) = C^{AWGN} = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_W^2}{\sigma_Z^2}\right). \tag{11}$$

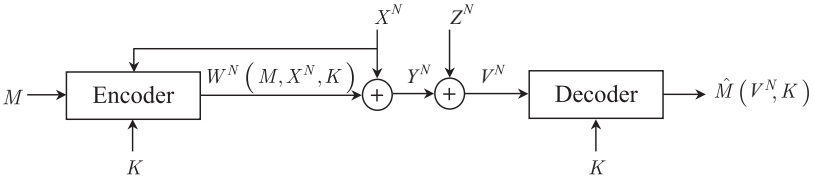


Fig. 3. Generalized Costa setup.

2.2 Scalar Costa Scheme: discrete approximation of Costa problem

To reduce the complexity of Costa set-up operating with the Gaussian random codebook that has the exponential complexity, a number of practical watermarking algorithms exploit structured codebooks based on the above

presented binning technique [21,22]. These structured codebooks are designed using dither quantizers (lattices) that should provide the independence of watermark (considered to be the quantization noise) and host data [23].

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 the host-dependent selection of codewords. The auxiliary random variable U^N, $U^N \in \mathcal{U}^N(K = k)$, in this setup can be interpreted as:

$$U^N = W^N + \alpha' X^N = \alpha' Q^N_{M,K}(X^N), \tag{12}$$

where $Q^N_{m,k}(.)$ denotes a vector or scalar quantizer for the message m and the key k and α' designates the so-called compensation factor, an analog of Costa α parameter. In the simplified version of the so-called Scalar Costa Scheme (SCS) the quantizer is chosen to be the uniform scalar one working at the high-rate assumption where the pdf of host signal X^N is assumed to be flat within a quantization interval [23] (we do not use here the word "bin" to avoid misinterpretation with the message bin in the Gel'fand-Pinsker problem) [21,22]. This leads to the uniformly distributed watermark $W^N = U^N - \alpha' X^N = \alpha' Q^N_{M,K}(X^N) - \alpha' X^N$. The resulting stego data are obtained as:

$$y^N = x^N + w^N = x^N + \alpha'(Q^N_{m,k}(x^N) - x^N). \tag{13}$$

The watermark in this case will be uniform and the embedding distortion is equal to the variance of watermark given by $\sigma^2_W = \alpha'^2 \frac{\Delta^2}{3}$ (as a consequence of high-rate quantization watermark design), where Δ denotes the quantization step. Therefore, the selection of rate maximizing α designed for the Gaussian watermark in the Costa setup is not any more optimal in the above case (for this reason we use α').

3 Partially reversible data hiding

In this section, we consider the reversibility of the Gel'fand-Pinsker based data hiding methods. This analysis will be performed for two cases of unauthorized and authorized users possessing different prior information. The unauthorized users perform the reversibility solely based on the watermarked version that can be either noisy or noise-free depending on the application. The authorized users have additionally access to the secret key besides the availability of the watermarked data. In turn, knowledge of the secret key provides the access to the particular codebook construction considered in Section 2 (Fig. 2) that is powerful prior information for the construction of reversible scheme.

In both cases, the setup under our analysis is based on the Gel'fand-Pinsker

framework for the condition of the achievable rate $R \leq C$, which originally addresses the problem of pure message m embedding with the maximum achievable rate R. Thus the Gel'fand-Pinsker problem in the data hiding formulation (8) is to maximize the rate R to achieve the channel capacity under the embedding distortion constraint. This formulation of classical robust data hiding does not assume any need for the host x^N recovery at the decoder.

The only known information-theoretic setups addressing the reversible data hiding are based on the essential modification of the binning strategy of the Gel'fand and Pinsker [12,13]. These methods use hybrid coding strategies where the additional coding technique is used for the encoding of the host state at the encoder to guarantee its best estimation at the decoder. Thus, the power for the message embedding is shared with the power for the host state encoding within the embedding constraint D^E. Moreover, the host recovery at the decoder does not benefit from knowledge of the communicated message once correctly decoded. Obviously, being useful for the authentication, such coding strategies are of little interest for robust data hiding used for content copyright protection, tracking, tracing and the interaction with the watermarked printed materials and analog audio. Nevertheless, to develop the useful intuition behind these hybrid coding strategies using the same mathematical notations and uniform terminology as well as to have fair benchmarking with the coding strategies directly addressing optimal joint message embedding and host estimation, we present the problem formulation, the used coding and main results for the state-of-the-art methods in Appendices A and B. In particular, the problem of joint message message communications and channel state estimation considered by Sutivong *et. al.* [14] is summarized in Appendix A. The problems considered by Martinian *et. al.* [12] and Willems and Kalker [13] are presented in Appendix B. An important conclusion needed for our further consideration consists in the statement that the solution of the joint problem is the trade-off between the amount of reliably embedded/communicated pure information with the rate R and the accuracy of the host/channel state estimation. The optimal coding strategy achieving this trade-off is based on the hybrid uncoded transmission, consisting in the host scaling at the encoder and the corresponding estimation at the decoder, and the random binning with the power allocation based on some trade-off power sharing factor $0 \leq \gamma \leq 1$. It is important to note that the coding strategies are working completely independently and do not assist each other at the decoding/estimation stage. The uncoded transmission aims solely at the host/channel state estimation and the random binning based message encoding is used for the pure message embedding/communications. Moreover, the random binning coding part is always designed to guarantee the capacity achieving rate $R = C$ within the allowed fraction of the shared power. Finally, the benefit of information carried by the message embedding part is completely disregarded at the host recovery. This aspect remains uncovered and little studied. That is why in this paper, contrary to the prior work summa-

rized in Appendices A and B, we will consider an entirely different mechanism of reversibility that is based on the usage of side information extracted from the message embedding part to assist the host recovery. However, at the same time we do not impose any strict capacity achieving constraint, thus allowing for the rate $R \leq C$. Additionally, contrary to the existing methods of host estimation in the state dependent channels that are developed assuming perfect availability of the channel statistics (channel variance in the Gaussian setup) at the encoder prior to the transmission and considering the estimation accuracy as an additional constraint for the optimal communication protocol design, in this paper the problem is formulated in a different way. It is assumed that the protocol is optimized for information communication under the relaxed requirements on channel statistics availability at the encoder while the reversibility (accuracy of the host estimation) is considered as a granted option, i.e., as a by-product of optimal message communications.

Therefore, the main goal of the foregoing study consists in the analysis of host reversibility versus achievable rate in state dependent channels for the different assumptions and priors. More particularly, the problem is formulated in the following way that consists of two parts and includes the classical Gel'fand-Pinsker part and a reversibility part:

- *Gel'fand-Pinsker part*: the encoder is aiming at sending a message $M \in \{1, 2, ..., 2^{NR}\}$ to the receiver with some defined rate R via the host dependent channel by generating a sequence $W^N(M, X^N, K)$ (3) that is embedded into the host X^N according to (4); the decoder (6) aims at estimating the message $\hat{M}$ based on the degraded version V^N of Y^N and the secret key K.
- *Reversibility part*: is considered for two parties:
  - *the unauthorized user*, who has the access to the degraded version of V^N in the general case or just a watermarked version Y^N in a particular one, and aims at the best estimate of $\hat{X}^N$ without any knowledge of the secret key K and the corresponding side information as a mapping:

$$\psi^N : \mathcal{V}^N \to \hat{\mathcal{X}}^N; \tag{14}$$

  - *the authorized user*, who, besides V^N or Y^N, has access to the secret key K and thus to the corresponding auxiliary random variable estimate $\hat{U}^N = U^N$ as long as the data hiding rate $R \leq C$ that should assist the host estimation as a mapping:

$$\psi^N : \mathcal{V}^N \times \mathcal{U}^N \to \hat{\mathcal{X}}^N. \tag{15}$$

Therefore, we will assume a typical robust digital watermarking setup where the data hider selects the rate of message embedding such that $0 \leq R \leq C$ and investigate the best possible estimates of the host as the solution to the

optimization problem:

$$\hat{x}^N = \underset{x^N \in \mathcal{X}^N : R \leq C}{\arg\max}\ p(x^N|v^N), \tag{16}$$

for the unauthorized user and:

$$\hat{x}^N = \underset{x^N \in \mathcal{X}^N : R \leq C}{\arg\max}\ p(x^N|v^N, u^N), \tag{17}$$

for the authorized one.

3.1 Unauthorized user reversibility

In many multimedia applications, the unauthorized users are considered as those who do not have access to the secret key used for the datahiding. Nevertheless, these users might be motivated in certain circumstances to estimate the original host X^N based on the available noisy version of stego data V^N (Fig. 4). As a possible application one can mention printing protocol where both document fidelity and information authenticity should be provided. Possible application list is quite extensive and we refer all interested readers to [17] for more details. Moreover, as it was mentioned in the beginning of this paper, reversibility can be motivated by certain constraints in medical and military applications as well as considered as a version of the worst case attack against informed data hiding [16,24,25].

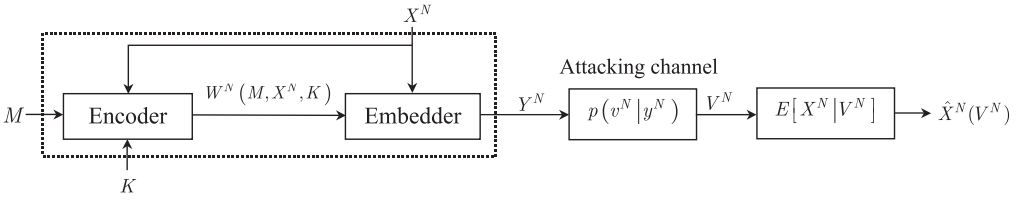


Fig. 4. Reversibility setup for the unauthorized user.

To estimate X^N, one can use either maximum a posteriori probability (MAP) estimator (16) or minimum mean square error (MMSE) estimator that coincide in the case of Gaussian setup. Therefore, the MMSE estimate of unauthorized user is obtained as:

$$\hat{X}^N = E[X^N|V^N]. \tag{18}$$

Assume that all random vectors in this setup are mutually independent zero-mean Gaussian, i.e., $X^N \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$, $W^N \sim \mathcal{N}(\mathbf{0}, \sigma_W^2 \mathbf{I}_N)$ and $Z^N \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$, contrary to the above case of uncoded transmission where the watermark W_X^N is colinear with X^N. In this case, the embedding distortion is $D^E = \sigma_W^2$ and the attacker distortion corresponds to the variance of AWGN, i.e., $D^A = \sigma_Z^2$.

Assuming $v^N = x^N + w^N + z^N$, one has the following MMSE estimate [26]:

$$\hat{X}^N = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2 + \sigma_Z^2} V^N. \tag{19}$$

The variance of this estimate corresponds to the variance of data (host) restoration by the unauthorized user and will be denoted as D^r_{MMSE}:

$$D^r_{MMSE} = E[d^N(\hat{X}^N, X^N)] = \frac{\sigma_X^2(\sigma_W^2 + \sigma_Z^2)}{\sigma_X^2 + \sigma_W^2 + \sigma_Z^2}. \tag{20}$$

It is important to note that $\hat{X}^N$ depends on the variances of original host, watermark and noise. In the asymptotic case of infinitively large host variance ($\sigma_X^2 \to \infty$) no reliable estimate is possible.

Moreover, perfect host restoration is not feasible in this setup even in the noiseless case ($\sigma_Z^2 = 0$), i.e., based on Y^N, due to watermark presence since the original host is considered as a completely random process for this kind of users. This level of accuracy does not obviously correspond to the demands of numerous multimedia and security applications requiring perfect watermark reversibility. It naturally reflects the price of information lack for the unauthorized users.

It is also interesting to note that this result coincides with those based on the optimal hybrid joint coding of Sutivong *et. al.* [14] for the power sharing parameter $\gamma = 1$, i.e., the pure message communication, presented in Appendix A (43). This also demonstrates that the considered estimate does not benefit from any side information and the strategy of Sutivong *et. al.* [14] is equivalent to those of unauthorized users.

3.2 Authorized user reversibility

In the case of authorized user, the secret key is available at the decoder side. The knowledge of key considerably extends the possibilities of host restoration at the decoder side in comparison with the unauthorized user. The block-diagram of this setup is shown in Fig. 5. The main idea behind is to use the so-called *genie-aided* or *multistage decoding* similarly to multiple access channel decoding [27]. In the scope of the advocated approach, this idea consists first in the reliable decoding of message m that assumes the availability of $\hat{U}^N = u^N(m, j, k)$, $\hat{U}^N \in \mathcal{U}^N(K = k)$, and then in the sequential $\hat{U}^N$-assisted estimation of $\hat{X}^N$. Our main results is based on the property that $\hat{U}^N = U^N$ as long as $R \leq C$ that can be used for the reversibility.

Depending on the reversibility setup, several scenarios are possible.

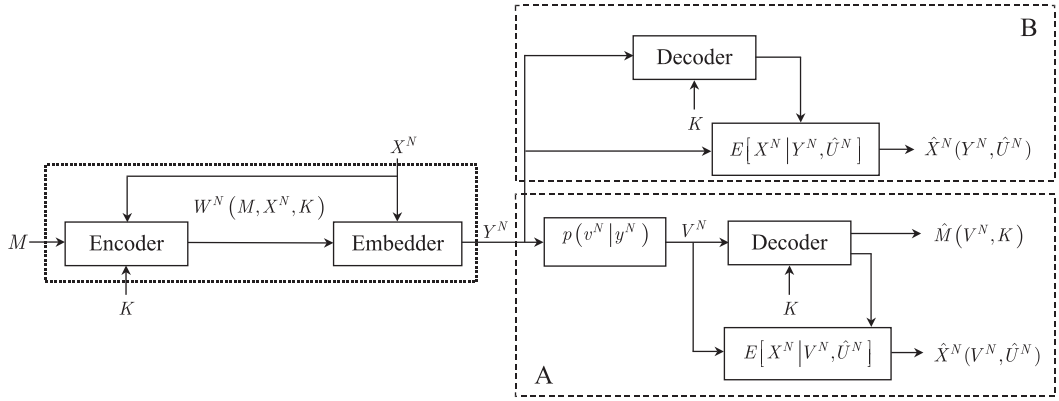


Fig. 5. Reversibility setup for the authorized user.

Scenario (A) (noisy case). This scenario, bottom part of Fig. 5, refers to the situation when the data hider designs the scheme for a particular fixed channel $p(v^N|y^N)$ and the certain achievable rate [2] using a corresponding codebook construction. The decoder should properly estimate the sent message according to the mapping based on V^N and K (6). At the same time, the authorized user is interested to estimate the host based on the available possibly distorted data V^N and decoded U^N using a key-dependent mapping:

$$\psi^N : \mathcal{V}^N \times \mathcal{U}^N(\mathcal{K}) \to \hat{\mathcal{X}}^N. \tag{21}$$

The criterion that judges the performance of (21) is defined based on the mean-squared estimation error:

$$D^r = E[d^N(\hat{X}^N, X^N)] = \frac{1}{N} E\left[\left\|\hat{X}^N - X^N\right\|^2\right]. \tag{22}$$

Therefore, besides the errorless message decoding, the problem is to design the estimator ψ^N that produces the minimum mean-squared estimation error based on V^N and U^N.

Scenario (B) (noiseless case). The second scenario of interest, depicted in the upper part of Fig. 5, refers to the situation when the data hider designs the codebook for a particular fixed channel $p(v^N|y^N)$ (for instance, that corresponds to the optimal attacker strategy in data hiding game [18]), performs data hiding procedure and stores the data in the form of Y^N for himself and at the same time makes it available via the channel $p(v^N|y^N)$. After a certain time, the data hider finds it necessary to recover the original host data due to some reasons caused by the loss of original data, its unavailability due to the time or access restrictions, some network/storage failures or the requirements of legal procedures in medical or law enforcement institutions. In these circumstances, the authorized user knows the key and has the undistorted wa-

[2] We do not require here that the rate should coincide with the channel capacity that is a valid assumption for numerous practical applications.

termarked data Y^N. The problem now is to design of a proper mapper ψ^N that can produce the MMSE estimate of X^N based on Y^N. What is of particular interest is to establish a possibility to perfectly restore the original data X^N, i.e., to achieve restoration distortion equal to zero.

We split our analysis in two parts. First, we consider the generalized reversibility of the Gel'fand-Pinsker problem for the authorized user. Secondly, we analyze the Costa setup to have a fair comparison with the previously considered scenario of unauthorized user reversibility. Along the analysis of Costa reversibility, we will indicate the particularities of practical data hiding schemes reversibility based on the structured codebooks.

The problem formulation that will be a common basis for the setups below can be given as follows. In the case of authorized user it is supposed that the distorted version of the watermarked data V^N and the key K are available. The problem is to design the estimate $\hat{X}^N$ of the original data X^N based on V^N using all information about the data hiding scheme design and corresponding codebook of the user defined by the key K. The quality of the obtained estimate should be validated by the restoration distortion D^r.

3.2.1 Reversibility of the Gel'fand-Pinsker setup

In the analysis of the Gel'fand-Pinsker setup, we assume that conditions of reliable message communications provided by Theorem 1 are satisfied and $\hat{m} = m$ with $P_e^{(N)} \to 0$ as $N \to \infty$. This implies that given the distorted data v^N and the key k, the decoder can uniquely identify the codeword $\hat{u}^N(\hat{m}, \hat{j}, k) = u^N(m, j, k)$ used at the encoder for watermark generation under conditions on th errorless performance of the decoder (Appendix C).

Therefore, in the noisy case corresponding to the scenario A, one can design a proper estimator of $\hat{x}^N$ based on v^N for the fixed channel $p(v^N|y^N)$ and errorless knowledge of $u^N(m, j, k)$. The decoder forms the MMSE estimate $\hat{x}^N$, i.e., the best linear estimate $\hat{x}^N$ given v^N and $u^N(m, j, k)$:

$$\hat{X}^N = E[X^N|V^N, U^N(W^N(M, X^N, K), X^N)], \tag{23}$$

where we emphasize that $u^N(m, j, k)$ is a function of the known message m, key k and the host realization x^N itself.

In the noiseless case (scenario B), $v^N = y^N$ and $y^N = \varphi^N(x^N, w^N)$ according to the embedding rule (4). Since $w^N = \phi^N(m, x^N, k)$ and assuming that $\hat{m} = m$ is correctly decoded that is obviously the case for the noiseless transmission and known k, one can substitute w^N into y^N obtaining $y^N = \varphi^N(x^N, \phi^N(m, x^N, k))$ and find $\hat{x}^N$ assuming the existence of uniqueness of the solution in the above equation for the functions $\varphi^N(.)$ and $\phi^N(.)$. In this case, $\hat{x}^N = x^N$ and the

authorized user can obtain the perfect estimate of the original data.

3.2.2 Reversibility of the Costa setup

To exemplify the above framework, we consider reversibility of the Costa setup assuming $X^N \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$, $W^N \sim \mathcal{N}(\mathbf{0}, \sigma_W^2 \mathbf{I}_N)$ and $Z^N \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$. The distorted version of the watermarked data $v^N = x^N + w^N + z^N$ is available at the decoder as well as the authorized user key k. This makes possible to find $u^N(m, j, k) \in \mathcal{U}^N(K = k)$ according to the argument presented in the previous subsection. From the Costa assumption about the auxiliary random variable, one can express the watermark as:

$$W^N = U^N - \alpha X^N. \tag{24}$$

Substituting W^N into V^N, one obtains:

$$V^N = (1 - \alpha)X^N + U^N + Z^N. \tag{25}$$

The MMSE estimate of X^N, $\hat{X}^N = E[X^N | V^N, U^N]$, assuming Gaussian data statistics is given by:

$$\hat{X}^N = E[X^N | V^N, U^N] = aV^N + bU^N, \tag{26}$$

where $a = \sigma_X^2 \sigma_W^2 (1-\alpha)(-2\alpha\sigma_W^2\sigma_X^2 + \sigma_X^2\sigma_W^2 + \alpha^2\sigma_W^2\sigma_X^2 + \alpha^2\sigma_Z^2\sigma_X^2 + \sigma_Z^2\sigma_W^2)^{-1}$, $b = \sigma_X^2(\sigma_W^2\alpha + \alpha\sigma_Z^2 - \sigma_W^2)(-2\alpha\sigma_W^2\sigma_X^2 + \sigma_X^2\sigma_W^2 + \alpha^2\sigma_W^2\sigma_X^2 + \alpha^2\sigma_Z^2\sigma_X^2 + \sigma_Z^2\sigma_W^2)^{-1}$.

The variance of this estimator is:

$$D^r(\alpha) = E[d^N(\hat{X}^N, X^N)] = \frac{\sigma_X^2 \sigma_W^2 \sigma_Z^2}{\sigma_W^2 \sigma_X^2 (1-\alpha)^2 + \sigma_Z^2(\sigma_W^2 + \alpha^2 \sigma_X^2)}. \tag{27}$$

In the noiseless case (scenario B), $\sigma_Z^2 = 0$, using (24) and the facts that $\alpha \neq 1$ and $Y^N = X^N + W^N$, the (26) reduces to:

$$\begin{aligned} \hat{X}^N &= \frac{1}{1-\alpha}(V^N - U^N) = \frac{1}{1-\alpha}(Y^N - U^N) \\ &= \frac{1}{1-\alpha}(X^N + W^N - \alpha X^N - W^N) = X^N \end{aligned} \tag{28}$$

that leads to $D^r = 0$ and provides the perfect reversibility of the watermark.

In the above analysis we have referred to the generic selection of parameter α. However, actually it depends on the variance of the watermark and noise, i.e., the maximum allowed embedding and attacking distortions. Normally, in the practice of digital data hiding, the actual value of applied attack variance is rarely known in advance at the encoder. Thus, α is selected keeping in mind

some critical, the least favorable, or average conditions of system applications [28]. This definitely provides the mismatch between the optimal parameter and actual one that leads to some decrease in the system performance in terms of maximum achievable rate that will be shown by the results of our simulation.

Nevertheless, it is interesting to investigate the hypothetical system performance in terms of reversibility, if one assumes the perfect knowledge of the operational scenario at the encoder that makes possible to choose the optimal parameter according to the Costa result (10). In this case, substituting $\alpha = \alpha_{opt}$ into (27), one obtains:

$$D^r(\alpha_{opt}) = \frac{\sigma_X^2(\sigma_W^2 + \sigma_Z^2)}{\sigma_X^2 + \sigma_W^2 + \sigma_Z^2} \tag{29}$$

that coincides with the estimation variance of the unauthorized user (20). The coincidence of (29) with (20) under condition $\alpha = \alpha_{opt}$ indicates a very interesting fact that the data hider as the authorized user can not benefit from the knowledge of U^N at the estimator in terms of estimation accuracy of X^N under optimal selection of parameter α. Obviously, the Gel'fand-Pinsker and Costa setups are designed to achieve the maximum rate of reliable communications but not the minimum possible distortion of the host communicated via the noisy channel. This justifies that the side information-assisted host estimation accuracy in this setup cannot exceed those provided by the estimation without any side information. Therefore, this scheme is not the optimal one when two constraints are imposed simultaneously [14]. The option of reversibility was considered rather a granted one along the main line of reliable message communications. Nevertheless, in the case of mismatch, i.e., $\alpha \neq \alpha_{opt}$, this disadvantage provides the increase of restoration accuracy.

Therefore, for the generic selection of α in the Costa setup, the rate distortion pair $(R(\alpha), D^r(\alpha))$ is obtained as:

$$\begin{aligned}&(R(\alpha), D^r(\alpha))\\ &= \left(\frac{1}{2}\log_2 \frac{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_Z^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_Z^2(\sigma_W^2 + \alpha^2\sigma_X^2)}, \frac{\sigma_X^2\sigma_W^2\sigma_Z^2}{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_Z^2(\sigma_W^2 + \alpha^2\sigma_X^2)}\right).\end{aligned} \tag{30}$$

Obviously, when $\alpha = \alpha_{opt}$, the rate distortion pair (30) coincides with the results [14] presented in Appendix A (43) for $\gamma = 1$. However, in the specific cases when $\alpha \neq \alpha_{opt}$, one can obtain quite interesting for practice results.

If $\alpha = 0$, that corresponds to the so-called *spread spectrum* data hiding, (30) reduces to:

$$(R(\alpha = 0), D^r(\alpha = 0)) = \left(\frac{1}{2}\log_2\left(1 + \frac{\sigma_W^2}{\sigma_X^2 + \sigma_Z^2}\right), \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}\right). \tag{31}$$

If $\alpha = 1$, that corresponds by analogy to the so-called *quantization index modulation* data hiding as the generalization of the SCS, (30) becomes:

$$(R(\alpha = 1), D^r(\alpha = 1)) = \left(\frac{1}{2} \log_2 \frac{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_Z^2)}{\sigma_Z^2(\sigma_W^2 + \sigma_X^2)}, \frac{\sigma_X^2 \sigma_W^2}{\sigma_X^2 + \sigma_W^2} \right), \quad (32)$$

where the achieved rate is higher in comparison to (31) but the distortion depends on the variance of the watermark σ_W^2 and remains constant with $\sigma_Z^2 \to 0$.

The same analysis can be extended to the discrete approximation of the Costa setup based on the SCS. The only difference consists in the statistics of the watermark that are defined by the selection of the quantizer. The reversibility of the SCS in a particular assumption when the host pdf is not taken into account, i.e., assuming $\sigma_X^2 \to \infty$, was peformed by Eggers *et. al.* [10].

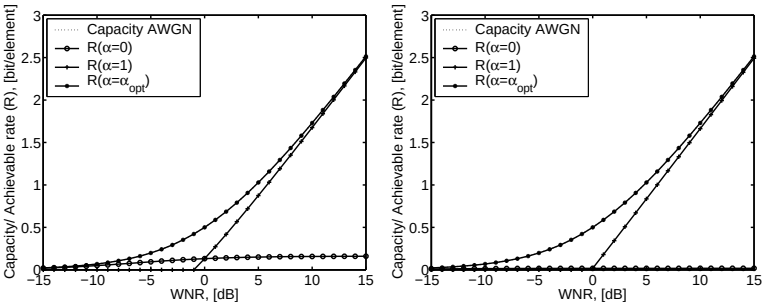
To be consistent with our analysis, one can show that in the noiseless case:

$$\hat{X}^N = \frac{1}{(1 - \alpha')}(Y^N - \alpha' Q_{m,k}^N(Y^N)) = X^N, \quad (33)$$

where we used the facts that $\alpha' \neq 1$ and $U^N = \alpha' Q_{m,k}^N(Y^N)$, $U^N \in \mathcal{U}^N(K = k)$, for the SCS and knowledge of quantizer for the decoded message m and the given key k. The equality follows from the observation that $y^N = x^N + \alpha'(Q_{m,k}^N(x^N) - x^N)$ and $Q_{m,k}^N(y) = Q_{m,k}^N(x)$.

4 Computer simulation

To confirm the theoretical findings, we have performed the experimental validation of different reversibility scenarios for the Gaussian setup and compare them with the known results. Fig. 6 summarizes the known results for the achievable rates of the Costa setup (9) with different values of optimization parameter α for the WIR values of -6 dB and -16 dB. In particular, for two asymptotic cases, when $\alpha = 0$ one obtains the performance of spread spectrum data hiding and when $\alpha = 1$ it corresponds to the SCS with $\alpha' = 1$ originating from dither modulation. These results are demonstrated to underline the critical dependence of the achieved rates on the selection of α. Obviously, the capacity of the AWGN channel is achieved for $\alpha = \alpha_{opt}$ (11) that provides interference-free communications. It should be again pointed out here that the Costa design of α_{opt} aims at maximizing the achievable rate and does not assume any constraints on the host restoration distortion in the case of reversible data hiding. To investigate the impact of α on the restoration distortion, we have performed a number of simulations for different types of



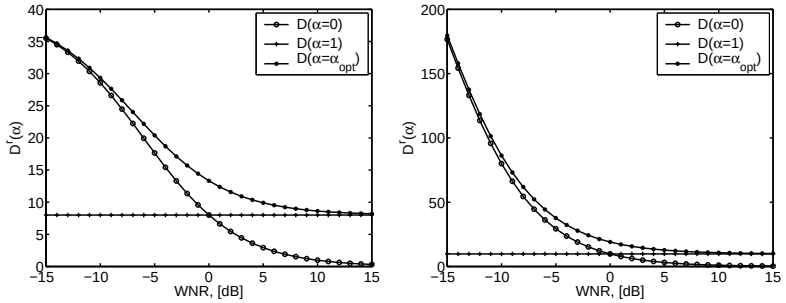
(a) (b)

Fig. 6. Achievable rate in the Costa setup: (a) WIR=-6 dB and (b) WIR=-16 dB.

users. First, assuming unauthorized user, who is aware only of the host, watermark and noise statistics, we have applied the MMSE estimation (19). The variance of this estimate D^r_{MMSE} is equal to the variance $D^r(\alpha = \alpha_{opt})$ of the corresponding estimate obtained by the authorized user (20) that is plotted in Fig. 7 for both WIRs. Secondly, assuming the authorized user with knowledge of the key (sequentially we suppose knowledge of U^N), we have computed the variance of the restored host $D^r(\alpha)$ according to (27) for the considered values of α (Fig. 7).

The obtained results illustrate the non-optimality of the Costa selection $\alpha = \alpha_{opt}$ for host recovery at the decoder based on the MMSE estimate. They allow to conclude about the estimation accuracy improvement at low WNRs in comparison with the unauthorized user or authorized user with $\alpha = \alpha_{opt}$ when α parameter increases. However, at high WNRs the situation is the opposite one. This behavior is justified by the fact that for $\alpha = 0$ (spread spectrum communications) $U^N = W^N$ and it represents additional interference source for host communications. Therefore, $D^r(\alpha = 0) = \frac{\sigma_X^2 \sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}$ and asymptotically perfect host recovery at high WNRs ($\sigma_Z^2 \to 0$) is possible that is shown in Fig. 7. Contrary, for $\alpha = 1$, the result is equal to $\frac{\sigma_X^2 \sigma_W^2}{\sigma_X^2 + \sigma_W^2}$ that is independent from σ_Z^2. The result for $\alpha = \alpha_{opt}$ asymptotically converges to that for $\alpha = 1$ as WNR$\to \infty$.

In order to finalize the consideration of reversibility of the Costa problem we computed the corresponding achievable rate-distortion pairs in this setup and compared them with the optimal results (41) and (42) when power and time(space)-sharing are used. The obtained results are presented in Fig. 8. These results demonstrate that depending on maximization of the communication rate or minimization of estimation distortion in certain applications, different scenarios are possible. Evidently, the maximum rate of reliable communications for the whole range of WNRs might be achieved only when $\alpha = \alpha_{opt}$. Oppositely, when more accurate host estimation is necessary for the fixed com-



(a) (b)

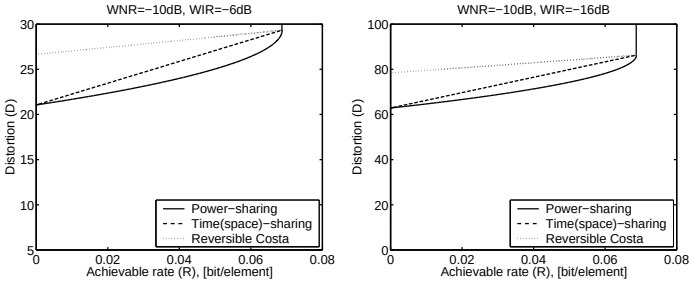
Fig. 7. Restoration distortion in the Costa setup: (a) WIR=-6 dB and (b) WIR=-16 dB.

munication rate, deviation from the rate maximization conditions are required.

It is interesting to note that the optimal rate-distortion pair (R, D), given by (41), (42), obtained using power-sharing can be asymptotically achieved based on the considered reversibility of the Costa scheme as WNR$\rightarrow +\infty$ (Figures 8,g and h). The gap also reduces as WIR$\rightarrow -\infty$. This result outperforms the time-sharing setup under considered conditions. Therefore, knowledge of U^N at the estimator of X^N can help reduce the estimation variance contrary to a particular case considered in [14] when the case $\alpha = \alpha_{opt}$ was only analyzed. It should be pointed out that the codebook construction in our consideration remains the same as for the Gel'fand-Pinsker or Costa cases and no sophisticated extra hybrid coding techniques were used as in [14].

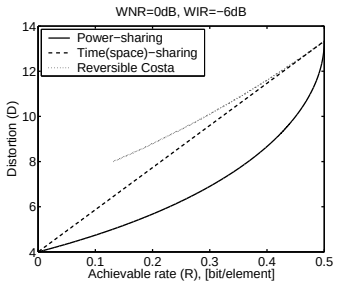
5 Conclusions

In this paper, the problem of reversibility of data hiding techniques based on random binning principle was analyzed as a by-product of pure message communications. Estimation-based reversibility was generally formulated within the Gel'fand-Pinsker framework and quantitatively analyzed in the Costa setup. We demonstrated that in the noisy case the unauthorized user is capable to remove the hidden data using optimal MMSE estimate with the same host data reconstruction distortion that the authorized one if the codebook is designed for $\alpha = \alpha_{opt}$. Contrary, non-optimal in the message communications sense selection of α together with the access to the proper codeword U^N provides significant estimation performance improvement for the authorized users. In the noiseless case ($\sigma_Z^2 \rightarrow 0$), knowledge of U^N allows the authorized user to completely recover the host data that is never possible for the unauthorized user. Similar analysis was performed for the quantization-based approximation of Costa setup. Moreover, the performance of the considered partially

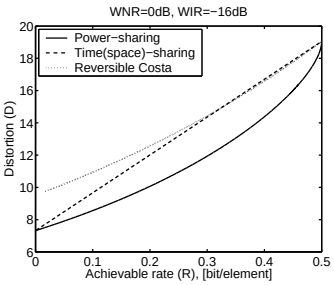


(a)

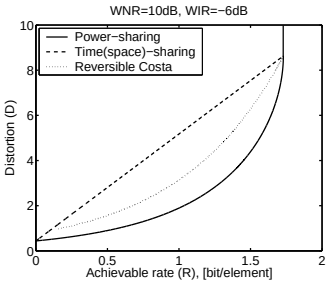
(b)



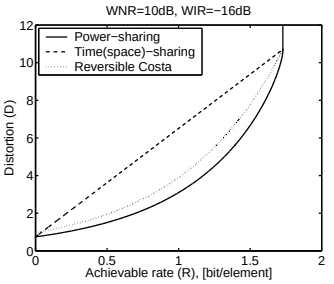
(c)



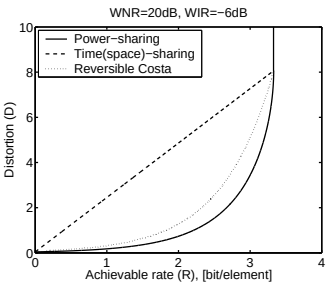
(d)



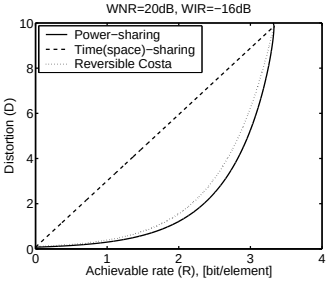
(e)



(f)



(g)



(h)

Fig. 8. Optimal (R, D) trade-off regions for the Gaussian setup.

reversible data hiding was compared with more involved coding strategies based on joint uncoded transmission and random binning. The conditions of asymptotic convergence of results are shown in the paper.

The future research will be focused on the extension of the above results to the codebooks with the repetition of codewords U^N for different users. We will also consider a case of multiple available copies Y_1^N, Y_2^N, ..., Y_L^N generated from the same host data with different message or keys.

Moreover, in the performed analysis we have considered the reversibility in the assumption of knowledge of sequence u^N at the decoder based on knowledge of key used for the data hiding. This provided useful bounds on the expected results of reversibility. Considering the reversibility as a sort of attacking strategy attempting at the complete removal of any trace of watermark from the stego data, it would be highly interesting to estimate the attacker efforts in learning u^N without knowledge of the secret key. We expect that this analysis can also provide useful insights on possible countermeasures on the codebook construction allowing to prevent perfect reversibility or increasing the complexity of the attacker in learning u^N.

6 Acknowledgements

This paper was partially supported by SNF projects 200021-111643 and 200021-1119770. We would like to thank to F. Pérez-González for his valuable feedback on the first version of this paper. The authors are also thankful to P. Moulin and T. Kalker for several discussions on the subject of reversibility and their valuable comments. The authors also appreciate communications with M. E. Haroutunian and enjoy the joint work on the error exponent consideration of reversibility.

7 Appendix A

The problem of joint message communications and channel state estimation was considered by Sutivong *at al.* [14] for the Gaussian case and the MMSE distortion measure. The problem was formulated as:

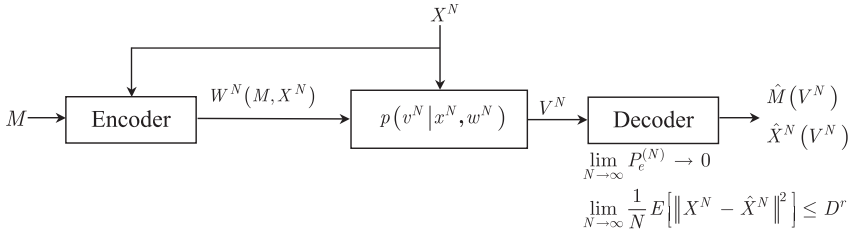
$$R = I(U;V) - I(U;X) \geq 0, \tag{34}$$

$$\frac{1}{N} E\left[\left\|W^N\right\|^2\right] \leq D^E, \tag{35}$$

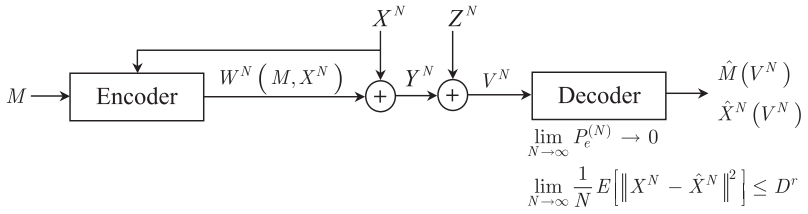
$$E\left[d^N(X^N, \hat{X}^N)\right] \leq D^r. \tag{36}$$

The solution to this problem is a trade-off between the amount of reliably communicated pure information with the rate R and the accuracy of channel (host) state estimation.

More particularly, the problem is formulated in the following way: the transmitter is aiming at sending a message $M \in \{1, 2, ..., 2^{NR}\}$ to the receiver as well as to provide the conditions for the accurate channel state X^N estimation (Fig. 9,a). For this purpose the encoder generates a sequence $W^N(M, X^N)$ and transmits it over the channel. From the channel output V^N, which is obtained according to the probabilistic mapping $p_{V^N|W^N,X^N}(v^N|w^N, x^N) = \prod_{i=1}^{N} p_{V|W,X}(v_i|w_i, x_i)$, the decoder is trying to estimate the sent message M as well as the host state X^N.



(a)



(b)

Fig. 9. Pure information and state information transmission over state dependent channels: (a) general formulation; (b) Gaussian setup.

In the case of Sutivong *et. al.* [14], this problem was considered for the Gaussian setup (Fig. 9,b), where the received signal is $V^N = W^N + X^N + Z^N$, $X^N \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$ and $Z^N \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$ are independent. As in the Costa setup, the host state is available non-causally at the encoder. A $(2^{NR}, N)$ code for this channel consists of an input message index set $\mathcal{M}$, encoder mapping:

$$\phi_S^N : \{1, 2, ..., 2^{NR}\} \times \mathcal{X}^N \to \mathcal{W}^N, \tag{37}$$

producing the channel input $W^N(M, X^N)$ subject to the power constraint $\frac{1}{N}\sum_{i=1}^{N} E\left[W_i^2\right] \leq \sigma_W^2$ and decoding mappings:

$$g_S^N : \mathcal{V}^N \to \{1, 2, ..., 2^{NR}\}, \tag{38}$$

$$\psi^N : \mathcal{V}^N \rightarrow \mathcal{X}^N. \tag{39}$$

The performance is measured by the error probability (7) and by the MMSE:

$$E\left[d^N(X^N, \hat{X}^N(V^N))\right] = \frac{1}{N} E\left[\left\|X^N - \hat{X}^N(V^N)\right\|^2\right], \tag{40}$$

where $d^N(.,.)$ is defined by (1), $\hat{X}^N$ is the estimate of X^N and the expectation is with respect to the joint pdf $p_{X^N,\hat{X}^N}(x^N, \hat{x}^N)$.

Definition 5: A rate-distortion pair (R, D) is said to be achievable if there exists a sequence of $(2^{NR}, N)$ codes such that $P_e^{(N)} \rightarrow 0$ and $E\left[d^N(X^N, \hat{X}^N)\right] \leq D$ for $N \rightarrow \infty$.

Theorem 2 [14]: The optimal (R, D) trade-off for the above considered additive channel $V^N = W^N + X^N + Z^N$ is a closure of the convex hull of (R, D) pairs satisfying:

$$R \leq \frac{1}{2} \log_2 \left(1 + \frac{\gamma \sigma_W^2}{\sigma_Z^2}\right), \tag{41}$$

$$D \geq \frac{\sigma_X^2(\gamma\sigma_W^2 + \sigma_Z^2)}{\left(\sigma_X + \sqrt{(1-\gamma)\sigma_W^2}\right)^2 + \gamma\sigma_W^2 + \sigma_Z^2}, \tag{42}$$

where $0 \leq \gamma \leq 1$ is the factor of power sharing.

The details behind the proof of Theorem 2 can be found in [14] and the interested reader is encouraged to review the referred paper.

It is important to mention the coding strategy that consists of hybrid uncoded transmission and random binning with the power allocation based on γ. The uncoded transmission is aiming at the channel state estimation and the random binning is used for the interference-free message communications.

Evidently, by varying γ, it is possible to compromise the rate of pure information transmission and host state estimation accuracy. In particular, two limiting cases were analyzed in [14]. The first one ($\gamma = 1$) corresponds to the pure information transmission similar to the Costa setup (Fig. 3), while in the second one ($\gamma = 0$) all the available power σ_W^2 is spent for the uncoded channel state communications via state dependent channel (Fig. 10).

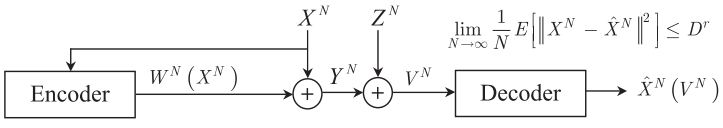


Fig. 10. Channel state communications via state dependent Gaussian channel.

The rate distortion pairs for these setups are obtained from (41) and (42) for

the corresponding values of γ:

$$\gamma = 1 : (R, D) = \left(\frac{1}{2} \log_2 \left(1 + \frac{\sigma_W^2}{\sigma_Z^2} \right), \frac{\sigma_X^2 (\sigma_W^2 + \sigma_Z^2)}{\sigma_X^2 + \sigma_W^2 + \sigma_Z^2} \right), \tag{43}$$

$$\gamma = 0 : (R, D) = \left(0, \frac{\sigma_X^2 \sigma_Z^2}{(\sigma_X + \sigma_W)^2 + \sigma_Z^2} \right). \tag{44}$$

It is important to note that for $\gamma = 1$ (Costa setup), the estimation of the host signal $\hat{X}^N$ is just the MMSE estimation solely based on the received signal V^N. No side information from the results of message decoding and particularly from U^N, which might be known at the decoder as well as at the encoder, is used to assist estimation. It should be also pointed out that optimal $\alpha = \alpha_{opt}$ was explicitly used in the derivations. Obviously, this selection of α maximizes the rate in Costa communications protocol. However, σ_Z^2 is not always known in advance at the encoder and thus the general result for any α similar to the pair (R, D) (41), (42) would be of interest for practical applications. Therefore, it is interesting to establish the impact of this side information about U^N and generic α on the estimation accuracy.

8 Appendix B

The data hiding counterpart of the Gel'fand-Pinsker problem has several main differences with the communications formulation presented in Appendix A. First, besides the application particularities, it includes the security aspect referred to the fact that the only authorized party can embed and decode information based on the secret key k considered in Section 2. Secondly, contrary to the fixed or "passive" communication channel with random parameters considered by Gel'fand and Pinsker, the data hiding includes "active" channel represented by the attacker that attempts to deteriorate the reliable communications modifying the stego data within the range of allowable distortions D^A (2). Thirdly, instead of input channel power constraint, one should satisfy the semantic similarity with the host data according to the distortion measure D^E (5). All these factors are considered in Section 2.

Martinian *et. al.* [12] and Willems and Kalker [13] were among the first who considered the problem of joint message communications (embedding) and channel state estimation (host recovery or restoration) in the information-theoretic formulation. Besides the fact that the problem was called differently, e.g., Martinian *et. al.* [12] called it *authentication* and Willems and Kalker [13] referred to it as *reversibile data hiding*, one can summarize both approaches similarly to the Sutivong *et. al.* formulation [14], where the main

difference comes from the "semantic" codebook construction that should guarantee distortion-typicality [27] according to the second condition. Contrary, Sutivong *et. al.* constrain only the input power of the sequence W^N.

Although the authors did not succeed to present the complete (R, D)-region similarly to Sutivong *et. al.* [14] in the communications setup, a number of interesting observations has been developed. In particular, Martinian *et. al.* [12] have shown that the inner bounds of (R, D)-region can be achieved using uncoded transmission similarly to Sutivong *et. al.* [14], while the outer bounds are achieved using binning strategy for the message communications adopted to various WNRs.

Willems and Kalker [13] considered coding theorems for reversible embedding and provided achievable rate-distortion pairs for the scenarios of standard noise-free embedding, reversible noise-free embedding, reversible and robust embedding and partially reversible noise-free embedding. Although they didn't provide results for the case of partially reversible and robust embedding, they conjectured that they would be similar to the ones given in [14].

9 Appendix C

The probability of error at the decoder in finding U^N used at the encoder under the condition of known key k, i.e., $U^N \in \mathcal{U}^N(K = k)$, can be upper bounded by:

$$P_{e_U} \leq (2^{N[R+R']} - 1)2^{-[I(U^N;V^N)-\delta]}, \tag{45}$$

$$< 2^{N[R+R']}2^{-[I(U^N;V^N)-\delta]} \tag{46}$$

that coincides with the condition of reliable decoding of message M under the constraints of small δ and large N and assuming $R + R' < \frac{1}{N}I(U^N; V^N)$ [2]. Therefore, the condition of errorless message M decoding also corresponds to the errorless knowledge of $\hat{U}^N$ at the decoder. Thus, one can state that once the message m is reliably decoded, one can also know u^N.

References

[1] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*, Morgan Kaufmann Publishers, Inc., San Francisco, 2001.

[2] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Probl. Control and Inf. Theory* **9**(1), pp. 19–31, 1980.

[3] M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory* **29**, pp. 439–441, May 1983.

[4] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. on Information Theory* **48**, pp. 1639–1667, June 2002.

[5] L. Pérez-Freire, F. Pérez-González, and S. Voloshynovskiy, "An accurate analysis of scalar quantization-based data hiding," *IEEE Transactions on Information Forensics and Security* **1**, pp. 80–86, March 2006.

[6] C. W. Honsinger, M. R. P. Jones, and J. C. Stoffel, "Lossless recovery of an original image containing embedded data," *U.S. Patent # 6278791* , 1999.

[7] J. .Fridrich, M. Goljan, and R. Du, "Lossless data embedding–new paradigm in digital watermarking," *EURASIP J. Appl. Signal Process.* **2002**(2), pp. 185–196, 2002.

[8] A. M. Alattar, "Reversible watermark using the difference expansion of a generalized integer transform," *IEEE Transactions on Image Processing* **13**, pp. 1147–1156, August 2004.

[9] M. U. Celik, G. Sharma, A. M. Tekalp, and E. Saber, "Reversible data hiding," in *International Conference on Image Processing (ICIP 2002)*, pp. 157–160, (NY, USA), Sept. 2002.

[10] J. Eggers, R. Buml, R. Tzschoppe, and B. Girod, "Inverse mapping of SCS-watermarked data," in *Eleventh European Signal Processing Conference (EUSIPCO'2002)*, (Toulouse, France), September 3-6 2002.

[11] J. Tian, "Wavelet-based reversible watermarking for authentication," in *SPIE Security and Watermarking of Multimedia Cont. IV*, **4675**, (San Jose, USA), Jan. 2002.

[12] E. Martinian, G. W. Wornell, and B. Chen, "Authentication with distortion criteria," *IEEE Trans. on Information Theory* , pp. 2523–2542, July 2005.

[13] F. M. J. Willems and T. Kalker, "Coding theorems for reversible embedding," in *Proc. DIMACS Series in Discrete Mathematics and Theoretical Computer Science; American Mathematical Society*, **66**, pp. 61–76, 2004.

[14] A. Sutivong, M. Chiang, T. Cover, and Y.-H. Kim, "Channel capacity and state estimation for state-dependent Gaussian channels," *IEEE Trans. on Information Theory* **51**, pp. 1486–1495, April 2005.

[15] M. Haroutunian, S. Tonoyan, O. Koval, and S. Voloshynovskiy, "On reversible information hiding system," in *Proceedings of the IEEE International Symposium on Information Theory*, (Toronto, Canada), July 6–11 2008.

[16] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgärtner, and T. Pun, "A generalized watermark attack based on stochastic watermark estimation and perceptual remodulation," in *Proceedings of the SPIE International Conference on Security and Watermarking of Multimedia Contents II*, **3971**, pp. 358–370, (San Jose, USA), 23–28 January 2000.

[17] S. Voloshynovskiy, O. Koval, F. Deguillaume, and T. Pun, "Visual communications with side information via distributed printing channels: extended multimedia and security perspectives," in *Proceedings of the SPIE International Conference on Security and Watermarking of Multimedia Contents III*, (San Jose, USA), January 2004.

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

[19] P. Moulin and R. Koetter, "Data-hiding codes (tutorial paper)," in *Proceedings IEEE*, **93**, pp. 2083–2127, Dec. 2005.

[20] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[21] B. Chen and G. W. Wornell, "Quantization Index Modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory* **47**, pp. 1423–1443, May 2001.

[22] J. Eggers, J. Su, and B. Girod, "A blind watermarking scheme based on structured codebooks," in *Secure Images and Image Authentication, IEE Colloquium*, pp. 4/1–4/6, (London, UK), April 2000.

[23] N. Jayant and P. Noll, *Digital coding of waveforms: principles and applications to speech and video*, Prentice-Hall, 1984.

[24] R. Tzschoppe, R. Bauml, R. Fischer, A. Kaup, and J. Huber, "Additive non-gaussian attacks on the scalar costa scheme (scs)," in *SPIE Electronic Imaging 2005. Security, Steganography, and Watermarking of Multimedia Contents VII*, **5681**, (San Jose, USA), October 2004.

[25] J. V. Forcen, S. Voloshynovskiy, O. Koval, F. Perez-Gonzalez, and T. Pun, "Worst case additive attack against quantization-based watermarking techniques," in *IEEE International Workshop on Multimedia Signal Processing*, (Siena, Italy), October 2004.

[26] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall Signal Processing Series, 1993.

[27] T. Cover and J. Thomas, *Elements of Information Theory.*, Wiley and Sons, New York, 1991.

[28] J. F. Vila, S. Voloshynovskiy, O. Koval, E. Topak, and T. Pun, "Asymmetric side data-hiding: optimization of achievable rate for Laplacian host," in *SPIE Electronic Imaging 2006, Security, Steganography, and Watermarking of Multimedia Contents VIII*, (San Jose, USA), January 15-19 2006.