On Joint Information Embedding and Lossy Compression

Alina Maor* and Neri Merhav

July 16, 2003

Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
{alinam@tx, merhav@ee}.technion.ac.il

Abstract

We consider the problem of optimum joint information embedding and lossy compression with respect to a fidelity criterion. The goal is to find the minimum achievable compression (composite) rate R_c as a function of the embedding rate R_e and the average distortion level Δ allowed, such that the average probability of error in decoding of the embedded message can be made arbitrarily small for sufficiently large block length. We characterize the minimum achievable composite rate and demonstrate how this minimum can be approached in principle.

1 Introduction

In the last few years, along with increasing awareness regarding the data protection, there is observed an increased interest in watermarking codes in their various applications. Watermarking is a form of hiding information in a host data set (covertext), usually an image, audio signal or video, creating a distorted version of the host data (stegotext, composite data). Successful retrieving of the watermark from the examined data indicates ownership, while on the other hand, damaging of the watermark beyond retrieving or its fabrication allows stealing the data or its forgery. There exists a variety of applications for data hiding ranging from classical steganography [1], [2], data authentication, copyright protection and copy control information [3], [4].

The requirements in watermarking scheme design are quite conflicting: In most applications, the watermark should be *perceptually transparent*, that is, invisible to the naked eye,

*This work is part of A. Maor's M.Sc. dissertation.

or, when audio signals are concerned, inaudible to the innocent listener. Watermarks must be *robust* to the distortion of watermarked data, caused by either conventional data processing (e.g., lossy compression, up/down-scaling, filtering, halftoning) or malicious attacks of the parties who wish to invalidate the watermark. As the reconstruction of the watermark is usually performed on a distorted version of stegotext (forgery), various distortion criteria measure the robustness of the embedding, while others ensure its initial transparency. Most watermarking schemes try to achieve the highest possible *information rate*, i.e., the amount of information that the embedded message should convey. In addition, it is usually assumed that if there exist two or more parties knowing the complete watermarking scheme, even if there exists a third party that is familiar with the watermarking technique but lacks some piece of information (e.g., secret key), it is very difficult if not impossible to 'crack' the watermarking scheme.

While most existing practical watermarking applications were designed and tested empirically (see, e.g., [1]-[4], [5]), the information-theoretic research activity in the problem area of watermarking is relatively new, evolving primarily around issues of system modeling, performance criteria, watermarking code design, and theoretical performance bounds. As covertext serves only as a 'carrier' of the watermark, it may be considered as a side information, as was first proposed in [6]. Thus, from the information-theoretic point of view, the watermarking problem is usually regarded as an instance of the problem of channel coding with side information, as originally treated in [19], [20] and [21]. The case where the side information is available to the encoder only is named public watermarking, and the case where it is available to the decoder as well is named private watermarking. A more general model of watermarking assumes that instead of full knowledge of covertext at the decoder, some partial information, that is statistically dependent on the covertext, is available. There exist various models for the watermarking schemes, where two main flows may be classified: One can either be interested in complete reconstruction of the watermark or only a partial reconstruction of the watermark, requiring the identification of its existence [7]. A variety of works [9]-[11] treat the problem of watermarking reconstruction from forgery as a hiding game between the information hider and the attacker.

Another aspect of the watermarking problem is that of joint information embedding and lossy compression, where quantization as well as entropy coding of the stegotext is treated as an integral part of the watermarking scheme. The general problem is given as follows:

There is a set of messages to be embedded in the covertext, subject to some distortion constraint. The composite sequence resulting from this embedding is compressed losslessly and the embedded message must be decoded reliably with or without access to the original host data. Karakos and Papamarcou [13, 14] and Willems and Kalker [15] study the tradeoffs between the distortion, the embedding rate and the composite rate for lossless compression. In [13], the attack-free version of private watermarking (fingerprinting) problem is treated, considering the case of a zero-mean Gaussian white noise process: the watermark, which is assumed to maintain some power and entropy constraints, is hidden in the covertext image, subject to a mean square distortion constraint, resulting in a Gaussian composite sequence, which is then compressed losslessly. The achievable rate region is established in terms of the relations between the composite rate, the embedding rate and the prescribed distortion constraint. In [14], a system similar to [13], which embeds watermarks in Gaussian covertext and distributes them in compressed form is studied. The performance of the system in the presence of an additive Gaussian attack (on decompressed covertext) is considered, and the achievable rate region is given. Willems and Kalker [15] study the attack-free case of the public joint watermarking-compression problem, where the covertext, the watermark and the composite sequence are drawn from finite alphabets. The model assumes that the composite sequence is subjected to some lossless symbol-by-symbol compression, the watermark is retrieved from reconstructed stegotext and, in addition, the covertext is estimated from the stegotext. The achievable region of composite rates, embedding rates, and distortion levels is characterized and a random coding algorithm is proposed for achieving any given point in the achievable region.

In this paper, we treat the attack-free version of the public problem of joint information embedding and entropy coding, where the covertext, the watermark and the stegotext are drawn from finite alphabets. As in [13] and [15], the data hiding and compression are cooperative and therefore are optimized jointly. The purpose of this paper is to characterize, in a more general way, the best achievable tradeoffs between the embedding rate R_e, the allowable average distortion Δ, and the composite rate R_c. Unlike in [15], in this paper, the lossless compression is performed per block rather than symbol-by-symbol. A single letter expression of the minimum achievable composite rate R_c^* is obtained as a function of R_e and Δ. The attainment conditions are established on R_e and Δ, beyond which there exists no reliable watermarking scheme. While in [15] it was speculated that for blockwise lossless compression of the composite sequence, the composite rate would be lower bounded by $H(Y)$, the entropy of the corresponding single-letter composite variable, here we show that this is true only for embedding rates that are above some threshold, whereas for lower embedding rates one can do better. In particular, the composite rate, in this lower region of embedding rates, can be made as small as $R_e + R(\Delta) < H(Y)$ (where $R(\Delta)$ is the rate distortion function of the covertext source) but cannot be reduced any further. The direct part of the coding theorem is based on showing that as long as R_e is not too large, one can construct 2^{NR_e} *disjoint* rate-distortion codebooks for the covertext source, and so, the watermark can be uniquely identified simply according to the codebook to which the composite sequence belongs. It should be pointed out that by this construction, we extend the Type Covering Lemma by Csiszár and Körner [16] to argue, that not only does a good reproduction codebook exist, but moreover, exponentially many different codebooks can be found. The continuous case is surprisingly different from the finite-alphabet case. Specifically, as we argue in the sequel, the composite rate in this case is *always* given by $R_e + R(\Delta)$ (and not only for low embedding rates) and that the embedding rate is unlimited. As for the Gaussian-quadratic case, the results of the public and the private watermarking problems, considering joint embedding and compression, coincide, since the result obtained in [13] is similar to ours.

The paper is organized as follows: in Section 2, we give notation conventions used through out the paper. Section 3 contains the system description and the problem definition. The main result is presented in Section 4. Sections 5 and 6 contain the proofs of the converse and the direct parts of the main coding theorem, respectively, while Sections 7 and 8 present an alternative expression of the formula of the minimum achievable composite rate, along with two examples of this formula for specific sources and distortion measures.

2 Notation Conventions and Preliminaries

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets, as well as most of the other sets, will be denoted by calligraphic letters. Similarly, random vectors, their realizations, and their alphabets will be denoted, respectively, by boldface capital letters, the corresponding boldface lower case letters, and calligraphic letters, superscripted by the dimensions. For example, the random vector $\boldsymbol{X} = (X_1, ..., X_N)$, ($N$-positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, ..., x_N)$ in $\mathcal{X}^N$, the Nth order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. The cardinality of a finite set $\mathcal{X}$ will be denoted by $|\mathcal{X}|$.

Let $\mathcal{Q}(\mathcal{X})$ denote the class of all discrete memoryless sources (DMSs) with a finite alphabet $\mathcal{X}$, and let Q denote a particular DMS in $\mathcal{Q}(\mathcal{X})$, i.e.,

$$\mathcal{Q}(\mathcal{X}) = \{Q : \forall x \in \mathcal{X}, \;\; Q(x) \geq 0, \;\; \sum_{x' \in \mathcal{X}} Q(x') = 1\}. \tag{1}$$

For a given positive integer N, let $\boldsymbol{X} = (X_1, ..., X_N)$, $X_i \in \mathcal{X}$, $i = 1, ..., N$, denote an N-vector drawn from a memoryless source Q, namely,

$$\begin{aligned}&\Pr\{X_i = x_i, \; i = 1, ..., N\} = \prod_{i=1}^{N} Q(x_i) \stackrel{\triangle}{=} Q(\boldsymbol{x}), \\ &\forall (x_1, ..., x_N), \; x_i \in \mathcal{X}, \; i = 1, 2, ..., N.\end{aligned} \tag{2}$$

Let $\mathcal{W}(\mathcal{Y}|\mathcal{X})$ denote the class of all discrete conditional probability mass functions (PMFs), henceforth referred to as channels from the finite alphabet $\mathcal{X}$ to a finite alphabet $\mathcal{Y}$, and let W denote a particular channel in $\mathcal{W}(\mathcal{Y}|\mathcal{X})$, i.e.,

$$\mathcal{W}(\mathcal{Y}|\mathcal{X}) = \{W \; : \; W(y|x) \geq 0, \; \sum_{y' \in \mathcal{Y}} W(y'|x) = 1, \; \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\}. \tag{3}$$

Let us also denote the class $\mathcal{P}(\mathcal{Y})$ of PMFs over a finite alphabet $\mathcal{Y}$ induced by PMFs $\{Q\}$ and channels W:

$$\mathcal{P}(\mathcal{Y}) = \{P : \; P(y) \stackrel{\triangle}{=} \sum_{x \in \mathcal{X}} Q(x) W(y|x), \; \sum_{y' \in \mathcal{Y}} P(y') = 1, \; \forall y \in \mathcal{Y}\}. \tag{4}$$

Information-theoretic quantities are denoted using the conventional notations [16, 17, 18]: For a pair of discrete random variables (X, Y) with a joint distribution $P(x, y) = Q(x)W(y|x)$, the *entropy* of X is denoted by $H(X)$, the *joint entropy* - by $H(X, Y)$, the *conditional entropy* of Y given X - by $H(Y|X)$, and the *mutual information* by $I(X;Y)$, where logarithms are defined to the base 2. When we wish to emphasize the dependence of an information-theoretic quantity on the underlying distribution, we use the latter as a subscript, for example, the *entropy* of X, induced by the source Q, will be denoted by $H_Q(X)$. The *relative entropy*, or *Kullback Leibler distance* of the pair of sources Q_1 and Q_2 is denoted by $D(Q_1||Q_2)$. The binary entropy function of a source $Q \sim \text{Bernoulli}(\alpha)$, $0 \leq \alpha \leq 1$, will be defined by

$$h(\alpha) \stackrel{\triangle}{=} -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha). \tag{5}$$

A *distortion measure* (or *distortion function*) is a mapping from the set $\mathcal{X} \times \mathcal{Y}$ into the set of non-negative reals:

$$d : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^+. \tag{6}$$

The distortion functions, considered in the paper, are bounded, i.e.,

$$d_{max} \stackrel{\triangle}{=} \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} d(x,y) < \infty. \tag{7}$$

The additive distortion $d(\boldsymbol{x}, \boldsymbol{y})$ between two vectors $\boldsymbol{x} \in \mathcal{X}^N$ and $\boldsymbol{y} \in \mathcal{Y}^N$ is given by:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} d(x_i, y_i). \tag{8}$$

For a set $\mathcal{B} \in \mathcal{Y}^N$, the minimum distortion between the elements of $\mathcal{B}$ and a vector $\boldsymbol{x} \in \mathcal{X}^N$ is denoted by:

$$d(\boldsymbol{x}, \mathcal{B}) \stackrel{\triangle}{=} \min_{\boldsymbol{y}\in\mathcal{B}} d(\boldsymbol{x}, \boldsymbol{y}). \tag{9}$$

The *rate-distortion function* $R(\Delta)$ of a memoryless source Q with respect to $d(\cdot, \cdot)$ is given by:

$$R(\Delta) = \min_{W:\sum_{x\in\mathcal{X},y\in\mathcal{Y}} Q(x)W(y|x)d(x,y)\leq\Delta} I(X;Y). \tag{10}$$

We next describe the notation related to the method of types, which is widely used throughout this paper. For a given memoryless source Q and a vector $\boldsymbol{x} \in \mathcal{X}^N$, the empirical probability mass function (EPMF) is a vector $P_{\boldsymbol{x}} = \{P_{\boldsymbol{x}}(a), a \in \mathcal{X}\}$, where $P_{\boldsymbol{x}}(a)$ is the relative frequency of the letter $a \in \mathcal{X}$ in the vector $\boldsymbol{x}$. For a scalar $\delta > 0$, the set T_Q^δ of all δ-typical sequences is the set of the sequences $\boldsymbol{x} \in \mathcal{X}^N$ such that

$$(1-\delta)Q(a) \leq P_{\boldsymbol{x}}(a) \leq (1+\delta)Q(a) \tag{11}$$

for every $a \in \mathcal{X}$. The size of T_Q^δ is bounded by [17]:

$$2^{N[(1-\delta)^2 H(X)-\delta]} \leq |T_Q^\delta| \leq 2^{N[(1+\delta)^2 H(X)]}. \tag{12}$$

It is also well-known (by the weak law of large numbers) that:

$$\Pr\left\{\boldsymbol{X} \notin T_Q^\delta\right\} \leq \delta \tag{13}$$

for all N sufficiently large.

For a given channel W and for each $\boldsymbol{x} \in T_Q^\delta$, the set $T_W^\delta(\boldsymbol{x})$ of all sequences $\boldsymbol{y}$ that are jointly δ-typical with $\boldsymbol{x}$, is the set of all $\boldsymbol{y}$ such that:

$$(1-\delta)P_{\boldsymbol{x}}(a)W(b|a) \leq P_{\boldsymbol{xy}}(a,b) \leq (1+\delta)P_{\boldsymbol{x}}(a)W(b|a), \tag{14}$$

for all $a \in \mathcal{X}, b \in \mathcal{Y}$, where $P_{\boldsymbol{xy}}(a,b)$ denotes the fraction of occurrences of the pair (a,b) in $(\boldsymbol{x},\boldsymbol{y})$.

Similarly as in eq. (11) [17], for all $\boldsymbol{x} \in T_Q^\delta$, the size of $T_W^\delta(\boldsymbol{x})$ is bounded as follows:

$$2^{N[(1-\delta)^2 H(Y|X)-\delta]} \leq |T_W^\delta(\boldsymbol{x})| \leq 2^{N[(1+\delta)^2 H(Y|X)]}. \tag{15}$$

Also, note that for all $\boldsymbol{x} \in T_Q^\delta$ and $\boldsymbol{y} \in T_W^\delta(\boldsymbol{x})$, $d(\boldsymbol{x},\boldsymbol{y})$ is upper bounded by:

$$d(\boldsymbol{x},\boldsymbol{y}) \leq (1+\delta)^2 \sum_{x,y} Q(x)W(y|x)d(x,y) = (1+\delta)^2 Ed(X,Y). \tag{16}$$

Finally, observe that for $\boldsymbol{x} \in T_Q^\delta$, $T_W^\delta(\boldsymbol{x}) \subseteq T_P^{\delta'}$, where $\delta' \stackrel{\triangle}{=} 2\delta + \delta^2$, since

$$(1-\delta)^2 P(b) \leq P_{\boldsymbol{y}}(b) \leq (1+\delta)^2 P(b), \tag{17}$$

where $P_{\boldsymbol{y}}(b)$ denotes the relative frequency of a letter $b \in \mathcal{Y}$ in the vector $\boldsymbol{y}$. The size of $T_P^{\delta'}$ is denoted similarly as in eq. (15) and

$$2^{N[(1-\delta)^2 H(Y)-\delta]} \leq |T_P^{\delta'}| \leq 2^{N[(1+\delta)^2 H(Y)]}. \tag{18}$$

3 System Description and Problem Definition

A general coding scheme for joint watemark embedding and compression is given in Fig. 1.

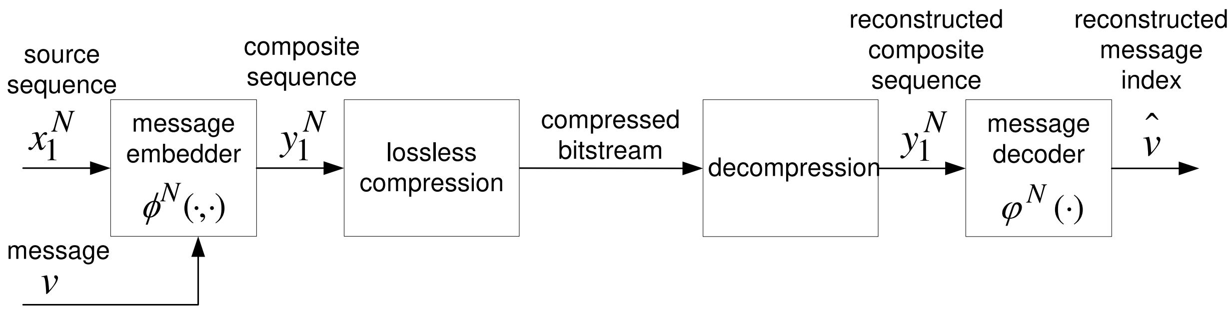

Figure 1: Block diagram of the system.

Let us consider a DMS Q that produces a sequence $\boldsymbol{X} = (X_1, ..., X_N)$ according to (2). This sequence will be referred to as the covertext sequence. One of M possible messages v, $v \in \{0, 1, ..., M-1\}$, is embedded into the covertext $\boldsymbol{x}$.

It is assumed that the message v is uniformly distributed across $\{0, 1, ..., M-1\}$, independently of $\boldsymbol{x}$, i.e.,

$$\Pr\{V = v\} = \frac{1}{M} \quad \text{for all} \quad v \in \{0, 1, ..., M-1\}. \tag{19}$$

The encoder (embedder) maps each pair $(\boldsymbol{x}, v)$ into a composite sequence, henceforth denoted as $\boldsymbol{y} = (y_1, y_2, ..., y_N)$, whose components take on values in a finite alphabet $\mathcal{Y}$.

The encoder is defined by the embedding function $\phi^N(\cdot, \cdot)$:

$$\boldsymbol{y} = \phi^N(\boldsymbol{x}, v) \stackrel{\triangle}{=} \big(\phi_1(\boldsymbol{x}, v), \phi_2(\boldsymbol{x}, v), ..., \phi_N(\boldsymbol{x}, v)\big) \tag{20}$$

where $\phi_n(\cdot, \cdot)$, $n = \{1, ..., N\}$ is the projection of $\phi^N(\cdot, \cdot)$, corresponding to the n-th coordinate. The decoder, that estimates the embedded message, is given by:

$$\widehat{v} = \varphi^N(\boldsymbol{y}), \tag{21}$$

where

$$\varphi^N : \mathcal{Y}^N \rightarrow \{0, 1, ..., M-1\}. \tag{22}$$

In order to maintain reasonable quality of the composite sequence, the following constraint is imposed on the system: The expected distortion between the composite sequence $\boldsymbol{y}$ and the source sequence $\boldsymbol{x}$, defined by

$$Ed(\boldsymbol{X}, \boldsymbol{Y}) \stackrel{\triangle}{=} Ed(\boldsymbol{X}, \phi^N(\boldsymbol{X}, V)) = \sum_{\boldsymbol{x}} \sum_{v} \frac{1}{M} Q(\boldsymbol{x}) \frac{1}{N} \sum_{n=1}^{N} d(x_n, \phi_n(\boldsymbol{x}, v)) \tag{23}$$

should not exceed a prescribed level Δ.

The composite sequence $\boldsymbol{y}$ is entropy-coded, i.e., the codeword length of $\boldsymbol{y}$ is defined as $l(\boldsymbol{y}) = \lceil -\log\big(\Pr\{\boldsymbol{y}\}\big)\rceil$, $\forall \boldsymbol{y} \in \mathcal{Y}^N$. The corresponding *composite rate* R_c, defined by

$$R_c \stackrel{\triangle}{=} \frac{El\big(\phi^N(\boldsymbol{X}, V)\big)}{N} \tag{24}$$

should be as small as possible. The *embedding rate*, R_e, defined by

$$R_e \stackrel{\triangle}{=} \frac{1}{N} \log M, \tag{25}$$

should be as large as possible.

The quality of estimation of V is judged according to the *average probability of error*, P_e, defined by:

$$P_e \stackrel{\triangle}{=} \Pr\{V \neq \varphi^N(\phi^N(\boldsymbol{X}, V))\}. \tag{26}$$

The average probability of error should, of course, be made as small as possible.

The objectives of simultaneously minimizing P_e and R_c, while maximizing R_e and maintaining distortion Δ, might be conflicting. It is, therefore, desirable to characterize the best achievable tradeoff between Δ, R_c and R_e that still enables reliable estimation of V.

An *achievable composite rate* R_c for a pair (R_e, Δ) is a composite rate such that for every $\epsilon > 0$, there exists a sufficient large N, an encoder ϕ^N and a decoder φ^N, that satisfy $P_e \leq \epsilon$ and $Ed(\boldsymbol{X}, \boldsymbol{Y}) \leq \Delta$ for R_c and R_e defined as in (24) and (25).

4 Main Result

We now present the main result of this paper, which is a single-letter expression for the minimum achievable composite rate.

Theorem 1. *For a DMS Q, defined as in (2), and a given pair (R_e, Δ), the minimum achievable composite rate R_c is given by*

$$R_c^*(R_e, \Delta) \stackrel{\triangle}{=} R_e + f(R_e, \Delta), \tag{27}$$

where

$$f(R_e, \Delta) \stackrel{\triangle}{=} \min_{\mathcal{S}(R_e,\Delta)} I(X;Y), \tag{28}$$

X being a random variable governed by Q, and

$$\mathcal{S}(R_e, \Delta) = \{W : Ed(X,Y) \leq \Delta, R_e \leq H(Y|X)\}. \tag{29}$$

Discussion:

The outline of the proof of the direct part of Theorem 1 is as follows: Given a fixed DMS Q and a channel W such that $Ed(X,Y) \leq \Delta$, we show that it is possible to find up to $2^{NH(Y|X)}$ disjoint codebooks of size $2^{NI(X;Y)}$, each maintaining the Type Covering Lemma by Csiszár and Körner [16]. Now, as long as $R_e \leq H(Y|X)$, it is possible to attribute a different watermark to each codebook, and thus, the watermark can be correctly retrieved according to the codebook to which the composite sequence belongs. The composite rate of this coding scheme is, therefore, $R_e + I(X;Y) \leq H(Y)$, and so, R_c^* is obtained by minimizing $I(X;Y)$ over a set of all channels maintaining the constraints of eq. (29).

In [15], the tradeoffs existing in the joint information embedding and compression problem are evaluated, focusing on the case of symbol-by-symbol compression of the composite sequences. Willems and Kalker obtain that the minimum achievable composite rate equals to $EL_H(Y)$, the expected codeword-length of a binary Huffman code of the composite variable, and conjecture that for blockwise lossless compression of the composite sequence the composite rate is lower bounded by $H(Y)$. As is shown in Section 7, for fixed Δ, $f(R_e, \Delta)$ is a monotonically non-decreasing and convex function of R_e, and therefore, an alternative expression of (27) is given by

$$R_c^*(R_e, \Delta) \triangleq \begin{cases} R_e + R(\Delta), & 0 \le R_e \le R_e^* \\ \min_{\mathcal{S}(R_e,\Delta)} H(Y), & R_e > R_e^*, \end{cases} \tag{30}$$

where

$$R_e^* \triangleq \max\{R_e : \ f(R_e, \Delta) = R(\Delta)\}, \tag{31}$$

proving that it is possible to achieve the composite rate as small as $R_e + R(\Delta) < H(Y)$ for values of R_e below R_e^*.

An easy extension of this work can be done for the case of continuous alphabets. The result of this extension turns out to be rather interesting: The minimum achievable composite rate is always given by the first expression of (30) and the embedding rate is unlimited. The proof of the converse part is identical to the one in Section 5 (except for dropping the constraint on R_e). In the proof of the direct part, the following property of the continuous sequences is used: For channels that achieve the rate distortion function, it is possible to generate infinitely many distinct nearly optimal rate-distortion codebooks that differ from each other only by arbitrarily small perturbations of one (representative) codebook, each one representing a different watermark message. Therefore, without any additional requirements on the system (like robustness to the attacks) the maximum achievable embedding rate is infinite. In [13], the embedding and the compression steps of the coding scheme are separated, but yet, we obtain that in case of the Gaussian alphabets, the achievable rate regions coincide for the public and the private joint watermarking-compression problems. Though, in general, for the same value of the composite rate, the embedding rate of the private watermarking is expected to be higher than this of the public watermarking. It is also interesting to notice that the coding scheme proposed in the proof of the direct coding theorem of [13] is substantially different from ours, using the random coding technique to create a single codebook of size 2^{NR_c} and then, evaluating the watermark by the joint typicality property of the decompressed composite sequence with the source sequence and the watermark. This coding scheme is implementable by the random binning technique [15], and while modified to be based on the joint typicality of the source and the composite sequences, achieves the results identical to ours for the case of the public version of the joint watermarking and compression problem. The idea of indexing of the disjoint codebooks by different watermarks is exploited also in [8], where the class of embedding methods (QIM, DC-QIM) is introduced, and the disjoint codebooks are created by different quantizers, each containing a unique set of the reconstruction vectors. Various case studies of implementation of QIM and DC-QIM, and its functionality and efficiency are provided in the paper, for both the public and the private versions of watermarking, in the presence of the attack, along with the proof of the information-theoretic optimality of the proposed methods, and estimations of the reachability of the embedding capacity. It is interesting to notice that the proof of the optimality of QIM (hidden QIM), which, in terms of quantization, gives the interpretation to the proof of the achievability of capacity by Gel'fand and Pinsker [20], is based on the technique of random binning. If analyzed from our point of view, in the attack-free case, QIM and DC-QIM obtain theoretical results which are identical to ours, and therefore, for certain practical problems, schemes achieving minimum R_c for prescribed (R_e, Δ) may be implementable by practical realizations of QIM.

5 Proof of the Converse Part

The proof of the converse is very similar to that of [15], but with some refinements. Let ϕ^N and φ^N be given, and assume that the distortion constraint Δ is satisfied. Consider a random variable I distributed uniformly over $\{1, 2, ..., N\}$, independently of all other random variables in the system, and let us denote

$$(X, Y) \stackrel{\triangle}{=} (X_I, Y_I). \tag{32}$$

Now, the probability distribution of (X, Y) is given by:

$$\Pr\{(X, Y) = (x, y)\} = \frac{1}{N} \sum_{n=1}^{N} \Pr\{(X_n, Y_n) = (x, y)\}. \tag{33}$$

Since, by hypothesis, the given system satisfies the expected distortion constraint,

$$\begin{aligned} \Delta &\geq Ed(\boldsymbol{X}, \boldsymbol{Y}) && (34) \\ &= \sum_{\boldsymbol{x},\boldsymbol{y}} Pr\{(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{x}, \boldsymbol{y})\} \frac{1}{N} \sum_{n=1}^{N} d(x_n, y_n) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{x,y} Pr\{(X_n,Y_n)=(x,y)\}d(x,y) \\
&= \sum_{x,y} Pr\{(X,Y)=(x,y)\}d(x,y) \\
&= Ed(X,Y).
\end{aligned}
$$

From the entropy coding of the composite sequence, we obtain:

$$
NR_c = El(\boldsymbol{Y}) \geq H(\boldsymbol{Y}) = H(\boldsymbol{Y},V) - H(V|\,\boldsymbol{Y}). \tag{35}
$$

From Fano's inequality, we obtain:

$$
H(V|\,\boldsymbol{Y}) \leq h(P_e) + P_e \log(M-1) \leq 1 + P_e N R_e. \tag{36}
$$

where $h(\cdot)$ is the binary entropy function. Thus,

$$
\begin{aligned}
NR_c &\overset{(a)}{\geq} H(\boldsymbol{Y}) && (37)\\
&\overset{(b)}{\geq} H(\boldsymbol{Y},V) - 1 - P_e N R_e \\
&= H(V) + H(\boldsymbol{Y}|V) - 1 - P_e N R_e \\
&\geq H(V) + I(\boldsymbol{Y};\boldsymbol{X}|V) - 1 - P_e N R_e \\
&\overset{(c)}{=} NR_e + H(\boldsymbol{X}|V) - H(\boldsymbol{X}|\,\boldsymbol{Y},V) - 1 - P_e N R_e \\
&\overset{(d)}{\geq} (1-P_e)NR_e - 1 + H(\boldsymbol{X}) - H(\boldsymbol{X}|\,\boldsymbol{Y}) \\
&\overset{(e)}{=} (1-P_e)NR_e - 1 + \sum_{n=1}^{N} H(X_n) - \sum_{n=1}^{N} H(X_n|\,\boldsymbol{Y},X_1^{n-1}) \\
&\overset{(f)}{\geq} (1-P_e)NR_e - 1 + \sum_{n=1}^{N} H(X_n) - \sum_{n=1}^{N} H(X_n|Y_n) \\
&\overset{(g)}{=} (1-P_e)NR_e - 1 + NH(X|I) - NH(X|Y,I) \\
&\overset{(h)}{\geq} (1-P_e)NR_e - 1 + NH(X) - NH(X|Y) \\
&= (1-P_e)NR_e - 1 + NI(X;Y) && (38)\\
&= (1-P_e)NR_e - 1 + NH(Y) - NH(Y|X) \\
&\overset{(i)}{\geq} (1-P_e)NR_e - 1 + H(\boldsymbol{Y}) - NH(Y|X) && (39)
\end{aligned}
$$

where:

- (a) follows from the entropy coding,
- (b) from (35) and (36)
- (c) from the assumption V has a uniform distribution,

(*d*) from the fact that $\boldsymbol{X}$ and V are independent and the fact that conditioning reduces entropy,

(*e*) from the chain rule of entropy and the assumption that the source is memoryless,

(*f*) from the fact that conditioning reduces entropy,

(*g*) from the definition of (X, Y) through I,

(*h*) from the fact that (X, I) are independent, and the fact that conditioning reduces entropy, and

(*i*) from the chain rule of entropy.

From (37) and (38), we obtain:

$$NR_c \geq (1 - P_e)NR_e - 1 + NI(X; Y) \tag{40}$$

and from (37) and (39) we obtain:

$$H(\boldsymbol{Y}) \geq (1 - P_e)NR_e - 1 + H(\boldsymbol{Y}) - NH(Y|X). \tag{41}$$

Subtracting $H(\boldsymbol{Y})$ from both sides of (41) and dividing both equations (40) and (41) by N, we obtain:

$$R_c \geq (1 - P_e)R_e - \frac{1}{N} + I(X; Y) \tag{42}$$

and

$$H(Y|X) \geq (1 - P_e)R_e - \frac{1}{N}. \tag{43}$$

By hypothesis, the given system satisfies $P_e \leq \epsilon$, and hence, by taking the limit $\epsilon \to 0$ as $N \to \infty$ in (43) and (42), we obtain both:

$$R_c \geq R_e + I(X; Y) \tag{44}$$

and

$$R_e \leq H(Y|X). \tag{45}$$

Obviously, the existence of a channel W that satisfies $Ed(X, Y) \leq \Delta$, $R_e \leq H(Y|X)$ and $R_c \geq R_e + I(X; Y)$ is equivalent to

$$R_c \geq R_c^*(R_e, \Delta) = R_e + \min_{\mathcal{S}(R_e, \Delta)} I(X; Y) = R_e + f(R_e, \Delta). \tag{46}$$

which completes the proof of the converse part.

6 Proof of the Direct Part

Before getting into the technical details, we present the outline of the proof. First, we show that for the given DMS Q and a channel W such that $Ed(X,Y) \leq \Delta$, it is possible to partition $T_P^{\delta'}$ into approximately $2^{NH(Y|X)}$ disjoint subsets, each of which with cardinality of approximately $2^{NI(X;Y)}$, such that for every sequence $\boldsymbol{x} \in T_Q^{\delta}$, each subset contains at least one element that is jointly typical with $\boldsymbol{x}$, i.e., belongs to $T_W^{\delta}(\boldsymbol{x})$. The proof has two stages: In Lemma 1 below, we prove that there exists a subset $\mathcal{B}$ with the above mentioned properties within a large enough subset $\mathcal{F} \subseteq T_P^{\delta'}$. Specifically, it is possible to find such $\mathcal{B}$ if $|\mathcal{F}| \geq 2^{-N\delta}|T_P^{\delta'}|$ and $|\mathcal{F} \bigcap T_W^{\delta}(\boldsymbol{x})| \geq 2^{-N\delta}|T_W^{\delta}(\boldsymbol{x})|$ for every $\boldsymbol{x} \in T_Q^{\delta}$. We also show that for each $\boldsymbol{x} \in T_Q^{\delta}$, $|\mathcal{B} \bigcap T_W^{\delta}(\boldsymbol{x})| \leq 2^{5N\delta}$. Afterwards, we apply Lemma 1 recursively on $T_P^{\delta'}$: In each recursion, we remove a subset of approximately $2^{NI(X;Y)}$ from $T_P^{\delta'}$, which serves as a rate-distortion codebook corresponding to a certain watermark message, then another such subset is removed from the remaining set of approximately $|T_P^{\delta'}| - 2^{NI(X;Y)}$ sequences, and so on. The removing procedure is proceeded as long as there are enough typical sequences left to apply Lemma 1. Finally, we obtain approximately $2^{NH(Y|X)}$ disjoint subsets of $T_P^{\delta'}$. The fact that these codebooks are all disjoint enables reliable decoding of the watermark.

Now, for a channel W that achieves the minimum in (28), we present a coding scheme that attributes each message v to a different subset, creating a different code-book for each message. We prove that this scheme achieves P_e as small as desired, for a sufficiently large N, that it maintains the distortion constraint for $R_e \leq H(Y|X)$, and that it achieves $R_c^*(R_e, \Delta)$ as given in (27).

We begin with the following Lemma.

Lemma 1. *Given $Q \in \mathcal{Q}(\mathcal{X})$, $W \in \mathcal{W}(\mathcal{Y}|\mathcal{X})$ and scalars $\Delta \geq \sum_{x,y} Q(x)W(y|x)d(x,y)$ and $\delta > 0$, there exists sufficiently large N_0 such that for all $N \geq N_0$, the following holds true: If a set $\mathcal{F} \subseteq T_P^{\delta'}$ satisfies $|\mathcal{F}| \geq 2^{-N\delta}|T_P^{\delta'}|$ and $|\mathcal{F} \bigcap T_W^{\delta}(\mathbf{x})| \geq 2^{-N\delta}|T_W^{\delta}(\mathbf{x})|$ for all $\mathbf{x} \in T_Q^{\delta}$, then there exists a set $\mathcal{B} \subset \mathcal{F}$ such that for all $\mathbf{x} \in T_Q^{\delta}$:*

$$2^{2N\delta}\frac{|T_P^{\delta'}|}{|T_W^{\delta}(\mathbf{x})|} \leq |\mathcal{B}| \leq 2^{3N\delta}\frac{|T_P^{\delta'}|}{|T_W^{\delta}(\mathbf{x})|}, \tag{47}$$

$$d(\mathbf{x}, \mathcal{B}) \leq (1+\delta)^2\Delta, \quad \forall \mathbf{x} \in T_Q^{\delta}, \tag{48}$$

$$|\mathcal{B} \bigcap T_W^{\delta}(\mathbf{x})| \leq 2^{5N\delta}, \quad \forall \mathbf{x} \in T_Q^{\delta}. \tag{49}$$

Lemma 1 is proved in Appendix A.1.

Proof of the Direct Part of Theorem 1

For a fixed DMS Q, a channel W and scalars $\Delta \geq Ed(X,Y)$ and $\delta > 0$, we apply Lemma 1 recursively on the set $T_P^{\delta'}$. In each recursion, it is examined whether the conditions of Lemma 1 are satisfied. If satisfied, there exists a subset $\mathcal{B} \in T_P^{\delta'}$ that meets (47)-(49). This subset is found and removed from the set of the typical sequences and the recursion is repeated. In each step, the recursion is operated on a smaller set of the typical sequences, until the typical sequences are exhausted. The maximum number of the steps of the recursion determines the maximum number of the distinct subsets of $T_P^{\delta'}$ that satisfy the claims of the Lemma. Let us notice that if within each subset exist identical elements, (which is possible for $I(X;Y) > \frac{1}{2}H(Y)$), then less then $|B|$ elements of $T_P^{\delta'}$ are removed in that step. We next restrict ourselves to the "worst" case in which all elements of each subset are distinct. This is a worst case in the sense of minimizing the number of steps of the recursion. Yet, we prove that even in the worst case, the proposed coding scheme achieves the minimum composite rate for a fixed pair of the embedding rate and allowed distortion.

Let us define the variables for the k-th step of the recursion, $k \in \{1, 2, ..., K\}$, where K will be defined later on. Let $\mathcal{B}(k)$ be a set that satisfies the claims of Lemma 1, for the k-th step and $m \stackrel{\triangle}{=} |\mathcal{B}(k)|$ as defined in (47). Let $\mathcal{F}(k)$ denote the set of typical sequences meeting the conditions of Lemma 1 in the k-th step:

$$\text{Initialization } (k=1): \qquad \mathcal{F}(1) = T_P^{\delta'}, \tag{50}$$

$$k-\text{th step}: \qquad \mathcal{F}(k) = T_P^{\delta'} - \bigcup_{i=1}^{k-1} \mathcal{B}(i), \tag{51}$$

$$|\mathcal{F}(k)| = |T_P^{\delta'}| - (k-1)m, \tag{52}$$

where (52) holds since $\{\mathcal{B}(i)\}$ are all disjoint.

Let us also define the cardinality of $\mathcal{B}(k) \bigcap T_W^{\delta}(\boldsymbol{x})$ by $N(\boldsymbol{x}, k)$, and let $\mathcal{F}(k) \bigcap T_W^{\delta}(\boldsymbol{x})$ be denoted by $\mathcal{F}_{\boldsymbol{x}}(k)$. Then,

$$\text{Initialization } (k=1): \qquad \mathcal{F}_{\boldsymbol{x}}(1) = T_W^{\delta}(\boldsymbol{x}), \tag{53}$$

$$k-\text{th step}: \qquad \mathcal{F}_{\boldsymbol{x}}(k) = T_W^{\delta}(\boldsymbol{x}) - \bigcup_{i=1}^{k-1} \Big(\mathcal{B}(i) \bigcap T_W^{\delta}(\boldsymbol{x})\Big), \tag{54}$$

$$|\mathcal{F}_{\boldsymbol{x}}(k)| = |T_W^{\delta}(\boldsymbol{x})| - \sum_{i=1}^{k-1} N(\boldsymbol{x}, i), \tag{55}$$

where by (49), $N(\boldsymbol{x}, i) \leq 2^{5N\delta}$ for all $i \in \{1, 2, ..., k-1\}$ and $\boldsymbol{x} \in T_Q^\delta$. Thus,

$$|\mathcal{F}_{\boldsymbol{x}}(k)| \geq |T_W^\delta(\boldsymbol{x})| - (k-1)2^{5N\delta}. \tag{56}$$

Evidently, Lemma 1 is implementable as long as

$$|\mathcal{F}(k)| \geq 2^{-N\delta}|T_P^{\delta'}| \tag{57}$$

and

$$|\mathcal{F}_{\boldsymbol{x}}(k)| \geq 2^{-N\delta}|T_W^\delta(\boldsymbol{x})|. \tag{58}$$

The constraints (57) and (58) dictate the maximum number of steps of the recursion and therefore,

$$K = \min\left\{\frac{2^{-N\delta}|T_P^{\delta'}|}{m}, \frac{2^{-N\delta}|T_W^\delta(\boldsymbol{x})|}{2^{5N\delta}}\right\}. \tag{59}$$

From (47), we obtain

$$2^{-4N\delta}|T_W^\delta(\boldsymbol{x})| \leq \frac{2^{-N\delta}|T_P^{\delta'}|}{m} \leq 2^{-3N\delta}|T_W^\delta(\boldsymbol{x})|, \tag{60}$$

and hence,

$$K = 2^{-6N\delta}|T_W^\delta(\boldsymbol{x})|. \tag{61}$$

We now summarize our coding scheme: For a fixed pair (R_e, Δ), let us consider the DMC W that achieves (28), hence $Ed(X,Y) \leq \Delta$, $R_e \leq H(Y|X)$ and $I(X;Y) = f(R_e, \Delta)$. Let us divide $T_P^{\delta'}$ into $2^{-6N\delta}|T_W^\delta(\boldsymbol{x})|$ disjoint subsets, each of size m, $2^{2N\delta}\frac{|T_P^{\delta'}|}{|T_W^\delta(\boldsymbol{x})|} \leq m \leq 2^{3N\delta}\frac{|T_P^{\delta'}|}{|T_W^\delta(\boldsymbol{x})|}$, as described above. We propose a coding scheme that attributes to each message, $v \in \{0, 1, ..., 2^{NR_e}-1\}$, a different subset, that is, a different codebook for this message.

The embedding scheme works as follows: Upon receiving a pair $(\boldsymbol{x}, v)$, where $\boldsymbol{x} \in T_Q^\delta$, the encoder chooses a composite sequence $\boldsymbol{y}$ from the codebook of v such that $\boldsymbol{y} \in T_W^\delta(\boldsymbol{x})$. The mapping from $T_Q^\delta \times \{0, 1, ..., 2^{NR_e} - 1\}$ to $T_P^{\delta'}$ is one-to-one with respect to the watermark, since all codebooks are disjoint. If $\boldsymbol{x} \notin T_Q^\delta$, a pre-defined error sequence is transmitted. Upon receiving a $\boldsymbol{y}$, the decoder checks whether the sequence belongs to one of the codebooks, or if it is an error-sequence. If the $\boldsymbol{y}$ is a codeword, the decoder retrieves the index v of the codebook to which it belongs and this is the decoded message. Otherwise, an error is announced.

The decoding error occurs when $\boldsymbol{x} \notin T_Q^\delta$. By (13), the probability of such an event vanishes as $N \to \infty$. Since the distortion function is assumed bounded by d_{max} (see (7)), then:

$$Ed(\boldsymbol{X}, \boldsymbol{Y}) \leq (1 - P_e)(1 + \delta)^2 \Delta + P_e d_{max}, \tag{62}$$

where the right-hand side becomes arbitrarily close to Δ for large N and small δ. Finally, the composite rate R_c is given by:

$$\begin{aligned} R_c &= \frac{1}{N} \log(2^{NR_e} m) \qquad (63) \\ &\leq \frac{1}{N} \log \left[2^{NR_e} 2^{3N\delta} \frac{|T_P^{\delta'}|}{|T_W^\delta(\boldsymbol{x})|} \right] \\ &\leq \frac{1}{N} \log \left[2^{NR_e} 2^{N[(1+\delta)^2 H(Y) - (1-\delta)^2 H(Y|X) + 4\delta]} \right] \\ &= R_e + (1+\delta)^2 H(Y) - (1-\delta)^2 H(Y|X) + 4\delta \\ &= R_e + (1+\delta)^2 I(X;Y) + 4\delta H(Y|X) + 4\delta \\ &= R_c^*(R_e, \Delta) + \epsilon(\delta), \qquad (64) \end{aligned}$$

where the second inequality follows from (15) and (18) and where $\epsilon(\delta) \to 0$ as $\delta \to 0$. This completes the proof of the direct part, as $\delta > 0$ is arbitrarily small.

7 An Alternative Expression of $R_c^*(R_e, \Delta)$

In this section, we prove that eqs. (27) and (30) are equivalent, as mentioned in the Discussion that follows Theorem 1. For $R_e \leq R_e^*$, the proof is trivial, by definition of R_e^*. Before we move on to the case $R_e > R_e^*$, let us first establish some important properties of $R_c^*(R_e, \Delta)$.

Lemma 2. *The function $R_c^*(R_e, \Delta)$, as defined in (27), is monotonically non-decreasing and convex in R_e, for fixed Δ.*

Lemma 2 is proved in Appendix A.2.

We now return to the proof of the equivalence between (27) and (30), proceeding with the case $R_e > R_e^*$. The function $f(R_e, \Delta)$ takes a constant value (plateau), $R(\Delta)$, for any $R_e \leq R_e^*$, and by definition, increases for $R_e > R_e^*$. Clearly, by convexity, $f(R_e, \Delta)$ cannot have an additional plateau for $R_e > R_e^*$. Therefore, for $R_e > R_e^*$, the channel W^* that achieves $f(R_e, \Delta)$, satisfies $R_e = H^*(Y|X)$, where $H^*(Y|X)$ is a conditional entropy induced by Q and W^*.

Let us define by $H^*(Y)$ and $I^*(X;Y)$ the entropy of Y and the mutual information between X and Y induced by Q and W^*, respectively. Then,

$$\begin{aligned} R_c^*(R_e,\Delta) &\stackrel{\triangle}{=} R_e + \min_{\mathcal{S}(R_e,\Delta)} I(X;Y) \\ &= H^*(Y|X) + I^*(X;Y) \\ &= H^*(Y) \\ &\geq \min_{\mathcal{S}(R_e,\Delta)} H(Y). \end{aligned} \tag{65}$$

On the other hand,

$$\begin{aligned} R_c^*(R_e,\Delta) &\stackrel{\triangle}{=} \min_{\mathcal{S}(R_e,\Delta)} [R_e + I(X;Y)] \\ &\leq \min_{\mathcal{S}(R_e,\Delta)} [H(Y|X) + I(X;Y)] \\ &= \min_{\mathcal{S}(R_e,\Delta)} H(Y), \end{aligned} \tag{66}$$

and therefore

$$R_c^*(R_e,\Delta) = \min_{\mathcal{S}(R_e,\Delta)} H(Y), \tag{67}$$

which completes the proof.

8 Examples

In this section, we provide two examples of calculation of $R_c^*(R_e,\Delta)$. The first example demonstrates a binary source with a one-sided distortion measure. For a given pair (R_e,Δ), $R_c^*(R_e,\Delta)$ is calculated following (27), and the achieving channels and the maximum achievable embedding rate are also found. The second example treats a case of a binary symmetric source with the Hamming distortion measure. Also there exists a value of the maximum achievable R_e beyond which $R_c^*(R_e,\Delta) = \infty$. The obtained minimizing channel is unique for a fixed Δ for all achievable embedding rates.

8.1 A binary source with a one-sided distortion measure

Consider a binary source $X \sim \text{Bernoulli}(p)$, with a distortion measure,

$$d(X,Y) \stackrel{\triangle}{=} \begin{cases} 1, & X=0, Y=1, \\ 0, & X=Y, \\ \infty, & \text{otherwise}. \end{cases} \tag{68}$$

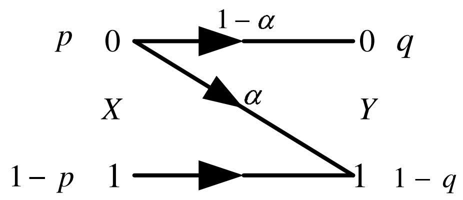

Figure 2: Z-Channel.

The test channel in this case, is obviously a Z-channel as depicted in Fig. 2.
The channel output Y is binary, $Y \sim$ Bernoulli(q), with $q = p(1-\alpha)$. Thus,

$$H(Y) = h(p(1-\alpha)), \tag{69}$$

and

$$H(Y|X) = ph(\alpha), \tag{70}$$

where $h(\cdot)$ is a binary entropy. The mutual information between X and Y, given by

$$I(X;Y) = h(p(1-\alpha)) - ph(\alpha), \tag{71}$$

is monotonically decreasing with α, since

$$\frac{\partial I(X;Y)}{\partial \alpha} = p\log(\frac{\alpha p}{1-p+\alpha p}) < 0, \tag{72}$$

for $\alpha > 0$.
We wish to determine the minimum achievable composite rate $R_c^*(R_e, \Delta)$, as denoted in (27), for a given pair (R_e, Δ). The average distortion measure is given by

$$Ed(X,Y) = p\alpha. \tag{73}$$

The channel achieving $R_c^*(R_e, \Delta)$ maintains $\min_{R_e \le H(Y|X), Ed(X,Y) \le \Delta} I(X;Y)$, and therefore, from (72), it is determined by the maximum α, satisfying the two following inequalities:

$$p\alpha \le \Delta \quad \text{and} \quad R_e \le ph(\alpha), \tag{74}$$

or equivalently,

$$\alpha \le \frac{\Delta}{p} \quad \text{and} \quad h^{-1}\left(\frac{R_e}{p}\right) \le \alpha \le 1 - h^{-1}\left(\frac{R_e}{p}\right), \tag{75}$$

where $h^{-1}(\cdot)$ represents the inverse entropy function with the support of $[0, \frac{1}{2}]$. For the case where $\frac{\Delta}{p} < h^{-1}(\frac{R_e}{p})$, the allowed range of α is empty. For $h^{-1}(\frac{R_e}{p}) \le \frac{\Delta}{p} \le 1 - h^{-1}(\frac{R_e}{p})$, the minimizing channel is with $\alpha = \frac{\Delta}{p}$, and for $1 - h^{-1}(\frac{R_e}{p}) < \frac{\Delta}{p}$, the distortion constraint no longer affects the composite rate, and so, the optimum channel is with $\alpha = 1 - h^{-1}(\frac{R_e}{p})$, provided that $R_e \le p$. Therefore, $R_c^*(R_e, \Delta)$ is given by

$$R_c^*(R_e, \Delta) = \begin{cases} R_e + h\big(p - \Delta\big) - ph\Big(\frac{\Delta}{p}\Big), & ph^{-1}\big(\frac{R_e}{p}\big) \le \Delta \le p\Big(1 - h^{-1}\big(\frac{R_e}{p}\big)\Big), \\ & 0 \le R_e \le p, \\ R_e + h\Big(ph^{-1}\big(\frac{R_e}{p}\big)\Big) - ph\Big(1 - h^{-1}\big(\frac{R_e}{p}\big)\Big), & p\Big(1 - h^{-1}\big(\frac{R_e}{p}\big)\Big) < \Delta, \\ & 0 \le R_e \le p, \\ \infty, & \Delta < ph^{-1}\big(\frac{R_e}{p}\big), \quad 0 \le R_e \le p, \\ & \text{or} \quad R_e > p. \end{cases} \tag{76}$$

8.2 The Binary Symmetric Source and Hamming Distortion measure

Consider the binary symmetric source and the Hamming distortion measure along with the class of channels depicted in Fig. 3.

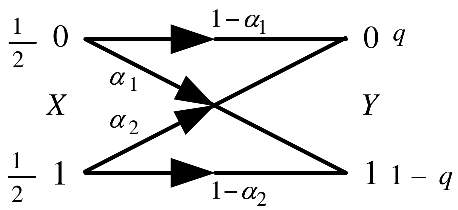

Figure 3: Binary Symmetric Source and Hamming distortion measure.

The channel output Y is binary, $Y \sim \text{Bernoulli}(q)$, with $q = \frac{1}{2}(1 - \alpha_1 + \alpha_2)$.

$$H(Y) = h(q) = h\Big(\frac{1}{2}(1 - \alpha_1 + \alpha_2)\Big), \tag{77}$$

and

$$H(Y|X) = \frac{1}{2}(h(\alpha_1) + h(\alpha_2)). \tag{78}$$

The mutual information between X and Y is given by

$$I(X;Y) = h\Big(\frac{1}{2}(1 - \alpha_1 + \alpha_2)\Big) - \frac{1}{2}(h(\alpha_1) + h(\alpha_2)). \tag{79}$$

We wish to determine the minimum achievable composite rate $R_c^*(R_e, \Delta)$ for a given pair (R_e, Δ). The distortion constraint $Ed(X, Y) \leq \Delta$ $\left(\Delta \leq \frac{1}{2}\right)$, interpreted as

$$\frac{1}{2}(\alpha_1 + \alpha_2) \leq \Delta, \tag{80}$$

dictates the family of channels (α_1, α_2) maintaining the distortion constraint.

We first argue that the test channel that achieves $f(R_e, \Delta)$ is symmetric, i.e., $\alpha_1 = \alpha_2 = \alpha$. Consider two channels $p_1(y|x)$ and $p_2(y|x)$ with (α_1, α_2) and (α_2, α_1), respectively. Those channels, henceforth referred to as original channels, share the same average distortion ((eq. 80)), and induce the same conditional entropy (eq. (78)) and mutual information (eq. (79)), denoted by $H_{1,2}(Y|X)$ and by $I_{1,2}(X;Y)$, respectively. Now, consider the mixture distribution $p_3(y|x) = \frac{1}{2}p_1(y|x) + \frac{1}{2}p_2(y|x)$, which is the symmetric channel $(\frac{\alpha_1+\alpha_2}{2}, \frac{\alpha_1+\alpha_2}{2})$, and let us denote by $H_3(Y|X)$ and by $I_3(X;Y)$ the conditional entropy and the mutual information induced by $p_3(y|x)$, respectively. Obviously, the average distortion of the mixture channel remains as this of the original channels, $H_3(Y|X) \geq H_{1,2}(Y|X)$ and $I_3(X;Y) \leq I_{1,2}(X;Y)$ due to the concavity of the conditional entropy and the convexity of the mutual information as a functionals of the conditional distribution. Hence, the mixture channel satisfies the same distortion and R_e constraints (eq. (29)) as the original channels, it may even satisfy larger R_e constraint, and it may have smaller mutual information. Thus, we obtain that given a pair (R_e, Δ), R_c^* is always achievable by a symmetric channel. The family of symmetric channels (α, α), induces (eq. (79))

$$I(X;Y) = 1 - h(\alpha), \tag{81}$$

and therefore, in order to determine R_c^*, we look for α which minimizes (81) under the constraints of eq. (29). For $\Delta \leq \frac{1}{2}$ and $R_e \leq h(\Delta)$, the minimum in (81) is achieved with $\alpha = \Delta$, and for $\Delta > \frac{1}{2}$ and $R_e \leq 1$ it is achieved with $\alpha = \frac{1}{2}$, giving:

$$R_c^*(R_e, \Delta) \stackrel{\triangle}{=} \begin{cases} R_e + 1 - h\Big(\min\left\{\Delta, \frac{1}{2}\right\}\Big), & 0 \leq \Delta, \quad 0 \leq R_e \leq h\Big(\min\left\{\Delta, \frac{1}{2}\right\}\Big) \\ \infty, & \text{otherwise.} \end{cases} \tag{82}$$

Appendix

A.1 Proof of Lemma 1

In this subsection we prove that within a large enough subset $\mathcal{F} \subseteq T_P^{\delta'}$ satisfying the conditions of Lemma 1, there exists a subset $\mathcal{B}$ that meets (47)-(49). The first two claims of

Lemma 1, (48) and (47), are proved in the spirit of the proof of the Type Covering Lemma by Csisźar and Körner [16] while proof of (49) is based on large deviations theory. For every sequence $\boldsymbol{x} \in T_Q^\delta$, let us denote:

$$\mathcal{F}_{\boldsymbol{x}} \triangleq \mathcal{F} \bigcap T_W^\delta(\boldsymbol{x}). \tag{A.1}$$

For a set $\mathcal{B} \subset \mathcal{F}$, let $m \triangleq |\mathcal{B}|$. Let us also denote by $\mathcal{U}(\mathcal{B})$ the set of those $\boldsymbol{x} \in T_Q^\delta$, for which $d(\boldsymbol{x}, \mathcal{B}) > (1+\delta)^2\Delta$ and by $\mathcal{B}_m$ the family of all subsets consisting of m, not necessarily distinct, elements of $\mathcal{F}$. By $\mathcal{B}_m^{1,2}$ we denote the family of all subsets of size m that satisfy (47) and (48), and by $\mathcal{B}_m^3$ the family of all subsets of size m that satisfy (49). We apply the method of random selection to $\mathcal{B}_m$ and show that there exists $\mathcal{B} \in \mathcal{B}_m$, such that m is bounded as in (47), $d(\boldsymbol{x}, \mathcal{B}) \leq (1+\delta)^2\Delta$, and (49) holds, i.e., $\mathcal{B} \in \mathcal{B}_m^{1,2} \bigcap \mathcal{B}_m^3$.

Let $Z^m \triangleq Z_1 Z_2 ... Z_m$, denote m independent copies of a random variable Z, uniformly distributed over $\mathcal{F}$. Let

$$1\{Z^m \notin \mathcal{B}_m^{1,2}\} \triangleq \begin{cases} 1 & \text{if } Z^m \notin \mathcal{B}_m^{1,2}, \\ 0 & \text{otherwise }. \end{cases} \tag{A.2}$$

and

$$1\{Z^m \notin \mathcal{B}_m^{3}\} \triangleq \begin{cases} 1 & \text{if } Z^m \notin \mathcal{B}_m^{3}, \\ 0 & \text{otherwise }, \end{cases} \tag{A.3}$$

and finally, define a new random variable:

$$\Gamma(Z^m) \triangleq 1\{Z^m \notin \mathcal{B}_m^{1,2}\} + 1\{Z^m \notin \mathcal{B}_m^{3}\}. \tag{A.4}$$

To prove the existences of $\mathcal{B} \in \mathcal{B}_m$, which satisfies (47), (48) and (49) at the same time, we will show that $E\Gamma(Z^m) < 1$.

$$\begin{aligned} E\Gamma(Z^m) &= E\Big\{1\{Z^m \notin \mathcal{B}_m^{1,2}\} + 1\{Z^m \notin \mathcal{B}_m^{3}\}\Big\} \\ &= E\Big\{1\{Z^m \notin \mathcal{B}_m^{1,2}\}\Big\} + E\Big\{1\{Z^m \notin \mathcal{B}_m^{3}\}\Big\} \\ &= \Pr\{Z^m \notin \mathcal{B}_m^{1,2}\} + \Pr\{Z^m \notin \mathcal{B}_m^{3}\}. \end{aligned} \tag{A.5}$$

Let us denote $\Pr\{Z^m \notin \mathcal{B}_m^{1,2}\}$ using the random set $\mathcal{U}(Z^m)$, which is a set of all $\boldsymbol{x} \in T_Q^\delta$ satisfying $d(\boldsymbol{x}, Z^m) > (1+\delta)^2\Delta$, namely, $d(\boldsymbol{x}, Z_i) > (1+\delta)^2\Delta$, for all $i \in \{1, 2, ..., m\}$. Using the union bound we obtain:

$$\Pr\{Z^m \notin \mathcal{B}_m^{1,2}\} = \Pr\left\{\bigcup_{\boldsymbol{x} \in T_Q^\delta} \boldsymbol{x} \in \mathcal{U}(Z^m)\right\} \leq \sum_{\boldsymbol{x} \in T_Q^\delta} \Pr\{\boldsymbol{x} \in \mathcal{U}(Z^m)\}. \tag{A.6}$$

By (16), $\boldsymbol{y} \in T_W^\delta(\boldsymbol{x})$ for $\boldsymbol{x} \in T_Q^\delta$ implies $d(\boldsymbol{x}, \boldsymbol{y}) \leq (1+\delta)^2\Delta$.

Therefore, for each $i \in \{1, 2, ...m\}$,

$$\Pr\left\{d(\boldsymbol{x}, Z_i) > (1+\delta)^2\Delta\right\} \leq \Pr\left\{Z_i \notin \mathcal{F}_{\boldsymbol{x}}\right\} = 1 - \frac{|\mathcal{F}_{\boldsymbol{x}}|}{|\mathcal{F}|} \leq 1 - 2^{-N\delta}\frac{|T_W^\delta(\boldsymbol{x})|}{|T_P^{\delta'}|}, \tag{A.7}$$

where the last inequality follows from the constraints on cardinality of $\mathcal{F}$ and $\mathcal{F}_{\boldsymbol{x}}$ for each $\boldsymbol{x} \in T_Q^\delta$. By the independence of $\{Z_i\}$,

$$\Pr\{\boldsymbol{x} \in \mathcal{U}(Z^m)\} = \prod_{i=1}^{m} \Pr\{d(\boldsymbol{x}, Z_i) > (1+\delta)^2\Delta\} \leq \left[1 - 2^{-N\delta}\frac{|T_W^\delta(\boldsymbol{x})|}{|T_P^{\delta'}|}\right]^m. \tag{A.8}$$

Applying the inequality $(1-a)^b \leq \exp(-ab)$ to (A.8) gives:

$$\Pr\{\boldsymbol{x} \in \mathcal{U}(Z^m)\} \leq \exp\left\{-m 2^{-N\delta}\frac{|T_W^\delta(\boldsymbol{x})|}{|T_P^{\delta'}|}\right\}, \tag{A.9}$$

which for

$$m \geq 2^{2N\delta}\frac{|T_P^{\delta'}|}{|T_W^\delta(\boldsymbol{x})|} \tag{A.10}$$

gives

$$\Pr\{\boldsymbol{x} \in \mathcal{U}(Z^m)\} \leq \exp\{-2^{N\delta}\}. \tag{A.11}$$

On substituting (A.11) into (A.6) we obtain

$$\Pr\left\{Z^m \notin \mathcal{B}_m^{1,2}\right\} \leq |T_Q^\delta| \exp\left\{-2^{N\delta}\right\} \leq |T_Q^\delta| 2^{-2^{N\delta}} \leq 2^{N\log|\mathcal{X}| - 2^{N\delta}}. \tag{A.12}$$

In view of (A.5), to complete the proof of Lemma 1, it is now enough to show that $\Pr\left\{Z^m \notin \mathcal{B}_m^3\right\} \to 0$ as $N \to \infty$. For a given sequence $\boldsymbol{x} \in T_Q^\delta$, let us denote by $N(\boldsymbol{x})$ the number of elements of Z^m in $T_W^\delta(\boldsymbol{x})$, i.e., the number of $Z_i \in \mathcal{F}_{\boldsymbol{x}}$:

$$N(\boldsymbol{x}) = \sum_{i=1}^{m} 1\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}, \tag{A.13}$$

where

$$1\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\} \triangleq \begin{cases} 1 & \text{if } Z_i \in \mathcal{F}_{\boldsymbol{x}}, \\ 0 & \text{otherwise .} \end{cases} \tag{A.14}$$

Then, $\Pr\left\{Z^m \notin \mathcal{B}_m^3\right\}$ is the probability of the existence of at least one $\boldsymbol{x} \in T_Q^\delta$ with $N(\boldsymbol{x}) > 2^{5N\delta}$, i.e.,

$$\Pr\left\{Z^m \notin \mathcal{B}_m^3\right\} = \Pr\{\exists \boldsymbol{x} \in T_Q^\delta : \; N(\boldsymbol{x}) > 2^{5N\delta}\}. \tag{A.15}$$

By the union bound, we obtain

$$\Pr\left\{Z^m \notin \mathcal{B}_m^3\right\} \leq \sum_{\boldsymbol{x} \in T_Q^\delta} \Pr\{N(\boldsymbol{x}) > 2^{5N\delta}\}. \tag{A.16}$$

Therefore, we are interested in upper-bounding $\Pr\{N(\boldsymbol{x}) > 2^{5N\delta}\}$, simultaneously for all $\boldsymbol{x} \in \mathcal{X}$. Obviously, $\Big\{1\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}\Big\}$ are Bernoulli i.i.d. random variables with

$$\Pr\Big\{1\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\} = 1\Big\} = \Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\} = \frac{E\{N(\boldsymbol{x})\}}{m}. \tag{A.17}$$

Note that $\Pr\{N(\boldsymbol{x}) > 2^{5N\delta}\} = \Pr\{\frac{1}{m}N(\boldsymbol{x}) > \frac{2^{5N\delta}}{m}\}$, and so, for

$$m \leq 2^{3N\delta}\frac{|T_P^{\delta'}|}{|T_W^{\delta}(\boldsymbol{x})|}, \tag{A.18}$$

we obtain

$$\begin{aligned}
\frac{2^{5N\delta}}{m} &\geq 2^{2N\delta}\frac{|T_W^{\delta}(\boldsymbol{x})|}{|T_P^{\delta'}|} && \text{(A.19)}\\
&\geq 2^{N\delta}\frac{|\mathcal{F}_{\boldsymbol{x}}|}{|\mathcal{F}|}\\
&= 2^{N\delta}\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\} && \text{(A.20)}\\
&> \Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}. && \text{(A.21)}
\end{aligned}$$

This means that $\Pr\{N(\boldsymbol{x}) > 2^{5N\delta}\}$ is the probability of a large deviations event associated with the empirical mean of $\Big\{1\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\} = 1\Big\}_i$ and hence can be upper-bounded as follows [16]:

$$\Pr\{N(\boldsymbol{x}) > 2^{5N\delta}\} \leq 2^{-mD\left(\frac{2^{5N\delta}}{m}||\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}\right)}, \tag{A.22}$$

where $D(\alpha\|\beta)$, for $\alpha, \beta \in [0,1]$ is defined as $\alpha\log\left(\frac{\alpha}{\beta}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-\beta}\right)$.
Now, $D\big(\frac{2^{N\epsilon}}{m}||\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}\big)$ is lower bounded as follows:

$$\begin{aligned}
D\Big(\frac{2^{5N\delta}}{m}||\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}\Big) &= \frac{2^{5N\delta}}{m}\log\frac{\frac{2^{5N\delta}}{m}}{\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}} && \text{(A.23)}\\
&+ \Big(1-\frac{2^{5N\delta}}{m}\Big)\log\frac{1-\frac{2^{5N\delta}}{m}}{1-\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}}\\
&\geq N\delta\frac{2^{5N\delta}}{m} && \text{(A.24)}\\
&+ \Big(1-\frac{2^{5N\delta}}{m}\Big)\log\frac{1-\frac{2^{5N\delta}}{m}}{1-\Pr\{Z_i \in \mathcal{F}_{\boldsymbol{x}}\}} && \text{(A.25)}\\
&\geq \Big(N\delta - \log(e)\Big)\frac{2^{5N\delta}}{m}, && \text{(A.26)}
\end{aligned}$$

where (A.24) follows from expressing the first term of (A.23) via (A.20), and (A.26) is obtained by applying of the inequality $\log(x) \geq (1-\frac{1}{x})\log(e)$ to (A.25) and reducing some positive terms of the obtained expression.

Substituting (A.26) into (A.22), we obtain

$$\begin{aligned}
\Pr\{N(\boldsymbol{x}) > 2^{5N\delta}\} &\leq 2^{-m\left(N\delta - \log(e)\right)\frac{2^{5N\delta}}{m}} && \text{(A.27)}\\
&= 2^{-2^{5N\delta}\left(N\delta - \log(e)\right)}, && \text{(A.28)}
\end{aligned}$$

and hence, substitution of (A.28) and the upper-bound on the size of T_Q^δ (12) into (A.16) gives:

$$\begin{aligned}
\Pr\left\{Z^m \notin \mathcal{B}_m^3\right\} &\leq |T_Q^\delta| 2^{-2^{5N\delta}\left(N\delta_2 - \log(e)\right)} && \text{(A.29)}\\
&\leq 2^{N(1+\delta)^2 H(X) - 2^{5N\delta}\left(N\delta - \log(e)\right)}.
\end{aligned}$$

For $N \to \infty$, $\Pr\left\{Z^m \notin \mathcal{B}_m^3\right\} \to 0$, which completes the proof of Lemma 1.

A.2 Proof of Lemma 2

In this subsection we prove that the function $R_c^*(R_e, \Delta)$, as defined in (27), is monotonically non-decreasing and convex in R_e, for fixed Δ.

The proof of monotonicity is simple: For increasing R_e, $f(R_e, \Delta)$, and therefore also $R_c^*(R_e, \Delta)$, are defined over decreasing minimization sets. We next establish the convexity with respect to R_e.

Let us consider the two points $(R_{e1}, R_c^*(R_{e1}, \Delta))$ and $(R_{e2}, R_c^*(R_{e2}, \Delta))$, which lie on the $R_c^*(R_e, \Delta)$ curve, where $R_{e1} < R_{e2}$. Let the joint distributions that achieve these pairs be $P_1(x, y) = Q(x)W_1(y|x)$ and $P_2(x, y) = Q(x)W_2(y|x)$. Let us also denote $R_{e\lambda} = \lambda R_{e1} + (1-\lambda)R_{e2}$. To prove convexity in R_e, we must show that

$$R_c^*(\lambda R_{e1} + (1-\lambda)R_{e2}, \Delta) \leq \lambda R_c^*(R_{e1}, \Delta) + (1-\lambda)R_c^*(R_{e2}, \Delta). \tag{A.30}$$

Equivalently, since both sides of (A.30) contain term R_e, which cancels, we must show:

$$f(R_{e\lambda}, \Delta) \leq \lambda f(R_{e1}, \Delta) + (1-\lambda) f(R_{e2}, \Delta) \tag{A.31}$$

for every $\lambda \in [0, 1]$.

Consider the distribution $W_\lambda = \lambda W_1 + (1-\lambda)W_2$. The mutual information is a convex functional of the conditional distribution, and the conditional entropy is a concave functional of the conditional distribution for a given distribution Q on $\mathcal{X}$, therefore:

$$I_{P_\lambda}(X;Y) \leq \lambda f(R_{e1}, \Delta) + (1-\lambda) f(R_{e2}, \Delta) \tag{A.32}$$

$$H_{P_\lambda}(Y|X) \geq \lambda H_{P_1}(Y|X) + (1-\lambda) H_{P_2}(Y|X) \tag{A.33}$$

where $I_{P_\lambda}(X;Y)$ and $H_{P_\lambda}(Y|X)$ are mutual information and conditional entropy of Y given X, induced by Q and W_λ; $H_{P_1}(Y|X)$ and $H_{P_2}(Y|X)$ are conditional entropies of Y given X, induced by Q and W_1 and W_2, respectively.

Since

$$f(R_{e\lambda}, \Delta) = \min_{\mathcal{S}(R_{e\lambda},\Delta)} I(X;Y), \tag{A.34}$$

where

$$\mathcal{S}(R_{e\lambda}, \Delta) = \{W(Y|X) : R_{e\lambda} \leq H(Y|X), Ed(X,Y) \leq \Delta\}, \tag{A.35}$$

then, by showing that

$$f(R_{e\lambda}, \Delta) \leq I_\lambda(X;Y), \tag{A.36}$$

the proof of convexity will be completed.

The average distortion $Ed(X,Y)$ is an affine functional of the distribution, and therefore, the distortion constraint is fulfilled by $W_\lambda = \lambda W_1 + (1-\lambda)W_2$.

From (A.33), using definitions of the sets of channels that achieve $f(R_{e1}, \Delta)$ and $f(R_{e2}, \Delta)$, we obtain that $R_{e1} \leq H_{P_1}(Y|X)$ and $R_{e2} \leq H_{P_2}(Y|X)$ and therefore $H_{P_\lambda} \geq R_{e\lambda}$.

Hence, we obtain that $W_\lambda \in \mathcal{S}(R_{e\lambda}, \Delta)$, and therefore (A.36) holds.

References

[1] F.A.P. Petitcolas, R.J. Anderson and M.G. Kuhn, "Information Hiding - A Survey," *Proc. IEEE*, vol. 87, no. 7. pp. 1062–1078, July 1999.

[2] R.J. Anderson and F.A.P. Petitcolas, "On the Limits of Stenography," *IEEE J. Comm*, vol. 16, no. 4. pp. 474–481, May 1998.

[3] M. Barni, C.I. Podilchuk F. Bartolini and E.J. Delp, "Watermark Embedding: Hiding a Signal Within a Cover Image," *IEEE Comm. Magazine*, pp. 102–108, August 2001.

[4] M.D. Swanson, M. Kobayashi and A.H. Tewfik, "Multimedia Data-Embedding and Watermarking Technologies," *Proc. IEEE Inform. Theory*, vol. 86, no. 6. pp. 1064–1087, June 1998.

[5] I.J. Cox, J. Kilian, F.T. Leighton and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," *IEEE Trans. Image Proc.*, vol. 6, no. 12. pp. 1673–1687, Dec. 1997.

[6] I.J. Cox, M.L. Miller and A.L. McKellips, "Watermarking as Communications with Side Information," *Proc. IEEE*, vol. 87, no. 7. pp. 1127–1141, July 1999.

[7] Y. Steinberg and N. Merhav, "Identification in the Presence of Side Information with Application to Watermarking," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1410–1422, May 2001.

[8] B. Chen and G.W. Wornell, "Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[9] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *submitted to the IEEE Trans. Inform. Theory*, preprint, October 1999.

[10] N. Merhav, "On Random Coding Error Exponents of Watermarking Systems," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 420–430, March 2000.

[11] A.S. Cohen and A. Lapidoth, "The Gaussian Watermarking Game," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1639–1667, June 2002.

[12] A. Somekh-Baruch and N. Merhav, "On the Error Exponent and Capacity Games of Private Watermarking Systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 537–562, March 2003.

[13] D. Karakos and A. Papamarcou, "A relationship between Quantization and Distortion Rates of Digitally Fingerprinted Data,". Also Institute for Systems Research Technical Report, TR 2000-51, UMD, December 2000, available at **http://www.isr.umd.edu/TechReports**.

[14] D. Karakos and A. Papamarcou, "A Relationship Between Quantization and Watermarking Rates in the Presence of Additive Gaussian Attacks, to appear in " *IEEE Trans. Inform. Theory*, August 2003. Also, Institute for Systems Research Technical Report TR 2001-50, UMD, available at **http://www.isr.umd.edu/TechReports**.

[15] F. Willems and T. Kalker, "Reversible Embedding Methods", *Proc. 40th Allerton Conference on Communications Control and Computing*, Monticello, IL, October 2002.

[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, New York, 1991.

[18] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression*, edited by T. Kailath, Prentice - Hall, 1971.

[19] C.E. Shannon, "Channels with Side Information at the Transmitter," *IBM J.*, pp. 289–293, October 1958.

[20] S.I. Gel'fand and M.S. Pinsker, "Coding for Channel with Random Parameters," *Prob. Control. Inform. Theory*, vol. 9(1), pp. 19–31, 1980.

[21] M.H.M. Costa, "Writing on dirty paper," *IEEE Trans. on Inform. Theory*, vol. IT-29, pp. 439–441, May 1983.