CCIT Report #464
January 2004

On Joint Information Embedding and Lossy Compression in the Presence of a Stationary Memoryless Attack Channel

Alina Maor* and Neri Merhav

January 13, 2004

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 decompressed composite sequence (stegotext) is distorted by a stationary memoryless attack, resulting in a forgery which in turn is fed into the decoder, whose task is to retrieve the embedded information. The goal of this paper is to characterize the maximum achievable embedding rate R_e (the embedding capacity C_e) as a function of the compression (composite) rate R_c and the allowed average distortion level Δ, 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 embedding capacity and demonstrate how it can be approached in principle. We also provide a single-letter expression of the minimum achievable composite rate as a function of R_e and Δ, below which there exists no reliable embedding scheme.

1 Introduction

The subject of watermarking and information embedding has been attracting a vast amount of attention of both the academic world and the industry, due to an increasing awareness for the need of data protection in its various forms: ownership identification, data forgery exposure, etc., as is extensively surveyed in e.g., [1]-[4] as well as in many other publications. Generally speaking, a good watermarking code should satisfy several conflicting requirements: One the one hand, the watermark should be *perceptually transparent*, that is,

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

invisible to the naked eye, or, when audio signals are concerned, inaudible to the innocent listener, while on the other hand, the watermark must also be *robust* to both conventional data processing (e.g., lossy compression, up/down-scaling, filtering, halftoning) and to potential malicious attacks by a party who may wish to invalidate the watermark by creating a forgery.

While most of the existing practical watermarking applications were designed and tested empirically (see, e.g., [1]-[5]), the information-theoretic research activity in the problem area of watermarking is relatively new, and it focuses primarily on issues of system modelling, performance criteria, watermarking code design, and theoretical performance bounds. From the information-theoretic point of view, the watermarking problem is usually regarded [6] as an instance of channel coding with side information [19]-[21], where the role of the side information is played by the covertext. The case where the side information is available to the encoder only is referred to as public watermarking, whereas the case where it is available to the decoder as well is termed private watermarking. In a variety of works (see, e.g., [7]-[10]) the watermarking problem is modelled as a game between the information hider and the attacker, where the former wishes to maximize a certain objective function, like the capacity or error exponent, while the latter strives to minimize this objective function.

Another aspect of the watermarking problem is that of joint information embedding and lossy compression, where quantization and entropy coding of the stegotext is carried out as an integral part of the watermarking scheme. The problem is 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 reliably decodable with or without access to the original host data, either directly from the stegotext or from its forgery. Although the compression of the composite sequence is lossless, the entire process is lossy since the reconstruction of the covertext from stegotext cannot be perfect after the watermark embedding. Karakos and Papamarcou [11, 12, 13], Willems and Kalker [14], and Merhav and Maor [22], study the tradeoffs between the distortion, the embedding rate and the composite rate, that is, the rate of lossless compression of the stegotext. In [11] and [12], the private watermarking (fingerprinting) problem is treated for the attack-free case and in the presence of the attack, respectively, assuming a Gaussian-quadratic model. In [11], the watermark is retrieved directly from the stegotext, while in [12] the stegotext is subjected to an additive Gaussian

attack resulting in a forgery from which the watermark is retrieved. For both cases, the achievable rate region is established in terms of the relations between the composite rate, the embedding rate and the prescribed distortion constraint. In [13], along with an extended analysis of the results of [11, 12], the achievable region is established for the finite alphabet case of private watermarking, and a general memoryless attack on the stegotext. Willems and Kalker [14] study the attack-free case of the public joint watermarking-compression problem for a finite alphabet covertext. The model in [14] assumes that the composite sequence is losslessly compressed symbol-by-symbol, 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 binning argument is proposed for achieving any given point in the achievable region. In [22], the attack-free public version of the problem is treated, both for the finite alphabet and the continuous alphabet cases. As in [11] and [14], the data hiding and compression are cooperative and therefore are optimized jointly, but unlike in [14], the lossless compression is performed per block rather than symbol-by-symbol. The main result of [22] is a single-letter expression of the minimum achievable composite rate as a function of the embedding rate and the allowable average distortion.

In this paper, we extend the model of [22] to include a stationary memoryless attack channel operating on the composite sequence. As in [22], the goal of this paper is to characterize the best achievable tradeoffs between the embedding rate R_e, the allowable average distortion Δ, and the composite rate R_c. The main result is a single-letter expression of the maximum achievable embedding rate R_e (embedding capacity C_e) as a function of R_c and Δ. We further argue that the achievable rate region of the continuous case is given by the same expression as in the finite-alphabet case.

The results of [22] are, of course, obtained as a special case for which the attack channel is the identity channel (i.e., no attack), but then (as in [22]), there is no longer need for the (Gel'fand-Pinsker) auxiliary random variable U since it simply coincides with the single-letter random variable Y that represents the stegotext. Indeed, the proof of achievability part in [22] is conceptually simpler and does not have the binning structure of the more general coding scheme presented here, which is in the spirit of the one of Gel'fand and Pinsker. In the presence of an attack, the choice of $U = Y$ is, in general, no longer optimal. This is in contrast to private watermarking [13], where the choice $U = Y$ is optimal for

all achievable embedding rates. It should be pointed out that for the continuous case, in the absence of the attack, the Gel'fand-Pinsker upper-bound on R_e can be safely omitted [22], due to the fact that not only is the watermark reliably recoverable from the composite sequence (there is a one-to-one mapping), but also, there is no limitation on the number of composite sequences (in contrast to the finite-alphabet case), except for the one imposed by the compressibility requirement.

The paper is organized as follows: In Section 2, we establish notation conventions used throughout the paper. Section 3 contains the system description and the problem definition. The coding theorem is presented in Section 4, and Sections 5 and 6 contain the proofs of the converse and the direct parts, respectively.

2 Notation Conventions and Preliminaries

Throughout the paper, scalar 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. The notations x_i^j and X_i^j, where i and j are integers and $i \leq j$, will designate segments $(x_i, ..., x_j)$ and $(X_i, ..., X_j)$, respectively, where for $i = 1$, the the subscript will be omitted. For example, the random vector $\boldsymbol{X} = X^N = X_1^N = (X_1, ..., X_N)$, ($N$-positive integer) may take a specific vector value $\boldsymbol{x} = x^N = x_1^N = (x_1, ..., x_N)$ 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 $|\mathcal{X}|$. For $i > j$, x_i^j (or X_i^j) will be understood as the null string.

Sources and channels will be denoted generically by the letter P subscripted by the name of the random variable and its conditioning, if applicable, e.g., $P_X(x)$ is the probability of $X = x$, $P_{Y|X}(y|x)$ is the conditional probability of $Y = y$ given $X = x$, and so on. Whenever clear from the context, these subscripts will be omitted. The class of all discrete memoryless sources (DMSs) with a finite alphabet $\mathcal{X}$ will be denoted by $\mathcal{P}(\mathcal{X})$, with P_X denoting a particular DMS in $\mathcal{P}(\mathcal{X})$, i.e.,

$$\mathcal{P}(\mathcal{X}) = \{P_X : \sum_{x \in \mathcal{X}} P_X(x) = 1, \quad \forall x \in \mathcal{X}, \quad P_X(x) \geq 0\}. \tag{1}$$

For a given positive integer N, the probability of any N-vector $\boldsymbol{x} = (x_1, ..., x_N)$ drawn from a DMS P_X, is given by

$$\Pr\{X_i = x_i,\ i = 1, ..., N\} = \prod_{i=1}^{N} P_X(x_i) \stackrel{\triangle}{=} P_X(\boldsymbol{x}). \tag{2}$$

Information-theoretic quantities will be denoted using the conventional notations [15, 16, 17]: For a pair of discrete random variables (X, Y) with a joint distribution $P_{XY}(x, y) = P_X(x)P_{Y|X}(y|x)$, the *entropy* of X will be 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 will use the latter as a subscript, for example, the *entropy* of X, induced by the source P_X, will be denoted by $H_{P_X}(X)$. The binary entropy function will be denoted by

$$h(\alpha) \stackrel{\triangle}{=} -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha), \quad 0 \leq \alpha \leq 1. \tag{3}$$

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

$$d : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^+. \tag{4}$$

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

$$d_{max} \stackrel{\triangle}{=} \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}} d(x, y) < \infty. \tag{5}$$

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:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_i). \tag{6}$$

We next describe the generic notation related to the method of types, which is widely used throughout this paper. For a given generic random variable (RV) $A \in \mathcal{A}$ (or a vector of RV's taking on values in $\mathcal{A}$), and a vector $\boldsymbol{a} \in \mathcal{A}^N$, the empirical probability mass function (EPMF) is a vector $P_{\boldsymbol{a}} = \{P_{\boldsymbol{a}}(a), a \in \mathcal{A}\}$, where $P_{\boldsymbol{a}}(a)$ is the relative frequency of the letter $a \in \mathcal{A}$ in the vector $\boldsymbol{a}$. For a scalar $\delta > 0$, the set $T^{\delta}_{P_A}$ of all δ-typical sequences is the set of the sequences $\boldsymbol{a} \in \mathcal{A}^N$ such that

$$(1 - \delta)P_A(a) \leq P_{\boldsymbol{a}}(a) \leq (1 + \delta)P_A(a) \tag{7}$$

for every $a \in \mathcal{A}$. The size of $T^{\delta}_{P_A}$ is bounded [16] by :

$$2^{N[(1-\delta)^2 H(A)-\delta]} \leq |T^{\delta}_{P_A}| \leq 2^{N[(1+\delta)^2 H(A)]}. \tag{8}$$

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

$$\Pr\left\{\boldsymbol{A} \notin T^{\delta}_{P_A}\right\} \leq \delta \tag{9}$$

for all N sufficiently large.

For a given generic channel $P_{B|A}(b|a)$ and for each $\boldsymbol{a} \in T^{\delta}_{P_A}$, the set $T^{\delta}_{P_{B|A}}(\boldsymbol{a})$ of all sequences $\boldsymbol{b}$ that are jointly δ-typical with $\boldsymbol{a}$, is the set of all $\boldsymbol{b}$ such that:

$$(1-\delta)P_{\boldsymbol{a}}(a)P_{B|A}(b|a) \leq P_{\boldsymbol{ab}}(a,b) \leq (1+\delta)P_{\boldsymbol{a}}(a)P_{B|A}(b|a), \tag{10}$$

for all $a \in \mathcal{A}, b \in \mathcal{B}$, where $P_{\boldsymbol{ab}}(a,b)$ denotes the fraction of occurrences of the pair (a,b) in $(\boldsymbol{a},\boldsymbol{b})$. Similarly as in eq. (7) [16], for all $\boldsymbol{a} \in T^{\delta}_{P_A}$, the size of $T^{\delta}_{P_{B|A}}(\boldsymbol{a})$ is bounded as follows:

$$2^{N[(1-\delta)^2 H(B|A)-\delta]} \leq |T^{\delta}_{P_{B|A}}(\boldsymbol{a})| \leq 2^{N[(1+\delta)^2 H(B|A)]}. \tag{11}$$

Finally, observe that for all $\boldsymbol{x} \in T^{\delta}_{P_X}$ and $\boldsymbol{y} \in T^{\delta}_{P_{Y|X}}(\boldsymbol{x})$, $d(\boldsymbol{x},\boldsymbol{y})$ is upper bounded by:

$$d(\boldsymbol{x},\boldsymbol{y}) \leq (1+\delta)^2 \sum_{x,y} P_X P_{Y|X}(y|x) d(x,y) \stackrel{\triangle}{=} (1+\delta)^2 E d(X,Y). \tag{12}$$

3 System Description and Problem Definition

Consider a general block coding scheme for joint watermark embedding and compression depicted in Fig. 1: A DMS P_X 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 watermarking messages, $v \in \{0, 1, ..., M-1\}$, 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.,

$$\Pr\{V = v\} = \frac{1}{M} \quad \text{for all} \quad v \in \{0, 1, ..., M-1\}. \tag{13}$$

The *embedding rate* of the scheme, R_e is defined by

$$R_e \stackrel{\triangle}{=} \frac{1}{N} \log M. \tag{14}$$

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 $\mathcal{Y}$.

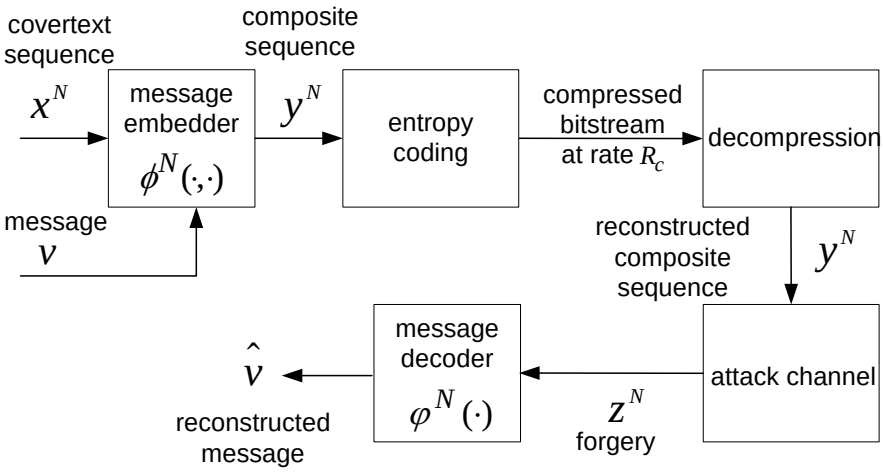

Figure 1: Block diagram of the system.

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

$$\boldsymbol{y} = \phi^N(\boldsymbol{x}, v) \stackrel{\triangle}{=} \big(\phi_1(\boldsymbol{x}, v), \phi_2(\boldsymbol{x}, v), ..., \phi_N(\boldsymbol{x}, v)\big) \tag{15}$$

where $\phi_n(\cdot,\cdot)$, $n = \{1, ..., N\}$ is the projection of $\phi^N(\cdot,\cdot)$, corresponding to the n-th coordinate. In order to maintain reasonable quality of the composite sequence, the following constraint is imposed: The expected distortion between the composite sequence $\boldsymbol{y}$ and the source sequence $\boldsymbol{x}$, defined by

$$Ed(\boldsymbol{X},\, \boldsymbol{Y}) \stackrel{\triangle}{=} Ed(\boldsymbol{X}, \phi^N(\boldsymbol{X}, V)) = \sum_{\boldsymbol{x}} \sum_{v} \frac{1}{M} P_X(\boldsymbol{x}) \frac{1}{N} \sum_{n=1}^{N} d(x_n, \phi_n(\boldsymbol{x}, v)) \tag{16}$$

should not exceed a prescribed level Δ.

The composite sequence $\boldsymbol{y}$ is entropy-coded, and the corresponding *composite rate* is defined by

$$\frac{H\big(\phi^N(\boldsymbol{X}, V)\big)}{N}, \tag{17}$$

and should not exceed a prescribed value, R_c.

The compressed composite sequence is sent to the decoder. After the decompression and before the watermarking decoding, $\boldsymbol{y}$ is distorted by an attacker modelled as a discrete stationary memoryless channel $P_{Z|Y}(z|y)$, which produces a forgery $\boldsymbol{z} = (z_1, z_2, ..., z_N)$, whose components take on values in a finite alphabet $\mathcal{Z}$. The decoder, that estimates the embedded message from $\boldsymbol{z}$, is given by:

$$\widehat{v} = \varphi^N(\boldsymbol{z}), \tag{18}$$

where

$$\varphi^N : \mathcal{Z}^N \to \{0, 1, ..., M-1\}. \tag{19}$$

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

$$P_e \stackrel{\triangle}{=} \Pr\{\varphi^N(\boldsymbol{Z}) \neq V\}. \tag{20}$$

An *achievable embedding rate* R_e for a pair (R_c, Δ) is an embedding rate, defined as in (14), such that for every $\epsilon > 0$, there exists a sufficient large N, an encoder ϕ^N and a decoder φ^N, that satisfy $P_e \leq \epsilon$, $Ed(\boldsymbol{X}, \boldsymbol{Y}) \leq \Delta$ and $\frac{H\left(\phi^N(\boldsymbol{X},V)\right)}{N} \leq R_c$. Our goal, in this paper, is to characterize the best achievable tradeoffs among Δ, R_c and R_e that maintain reliable estimation of V. In particular, we will be interested in the embedding capacity, $C_e(R_c, \Delta)$, which is the supremum of all achievable embedding rates for (R_c, Δ).

4 Main Result

Let $\mathcal{A}$ denote the set of all triples (U, X, Y) of random variables taking values in finite sets $\mathcal{U}$, $\mathcal{X}$, $\mathcal{Y}$, where $\mathcal{X}$ is the alphabet of the covertext, $\mathcal{Y}$ is the alphabet of the stegotext, and $\mathcal{U}$ is an arbitrary finite alphabet of size $|\mathcal{U}| \leq |\mathcal{X}| \cdot |\mathcal{Y}| + 1$, and the joint probability distribution of (U, X, Y), $P_{UXY}(u, x, y)$, is such that the marginal distribution of X is P_X, and $Ed(X, Y) \leq \Delta$. For any triple (U, X, Y), there exists a related quadruple (U, X, Y, Z), with Z taking values in $\mathcal{Z}$, such that

$$P_{U,X,Y,Z}(u, x, y, z) = P_{U,X,Y}(u, x, y) P_{Z|Y}(z|y), \tag{21}$$

where $P_{Z|Y}(z|y)$ is a transition probability of the discrete stationary memoryless attack channel.

We now present the main result of this paper, which is a single-letter characterization of the achievable region of (R_c, R_e, Δ).

Theorem 1. *Given a DMS P_X, an embedding rate R_e is achievable for a pair (R_c, Δ) if and only if there exists a triple of random variables $(U, X, Y) \in \mathcal{A}$ satisfying*

$$R_e \leq \min\{I(U; Z) - I(U; X), R_c - I(X; U, Y)\}. \tag{22}$$

The proof of the converse part of Theorem 1 is given in Section 5, and the proof of the direct part is provided in Section 6.

Obviously, the maximum achievable embedding rate $C_e(R_c, \Delta)$ is obtained by taking maximum among all the triples $(U, X, Y) \in \mathcal{A}$ maintaining the conditions of Theorem 1.

Corollary 1. *The embedding capacity $C_e(R_c, \Delta)$ for a DMS Q is given by:*

$$C_e(R_c, \Delta) = \max_{(U,X,Y)\in\mathcal{A}} \min\{I(U;Z) - I(U;X), R_c - I(X;U,Y)\}. \tag{23}$$

Discussion:

The embedding capacity of the public watermarking scheme is, of course, smaller than or equal to the one obtained in private watermarking [13]. It should be pointed out that (22) and (23) are not obtained by a straightforward extension of the well-known analysis of Gel'fand and Pinsker [20], where the maximum achievable embedding rate was found without constraining the allowable distortion of the covertext and without requirements on the compressibility of the codewords. Neither the direct scheme proposed in [20] nor the proof of the converse part of [20] lend themselves to characterizing tradeoffs between the embedding rate and the composite rate. Here, an alternative coding scheme is proposed, which not only achieves the embedding capacity of [20], but also allows a characterization of a tradeoff between the embedding and the composite rates. The two schemes differ in their ways of creating of composite sequences. In [20], after choosing an auxiliary codeword $\boldsymbol{U} = (U_1, ..., U_N)$ for a pair $(\boldsymbol{X}, V)$, a composite sequence $\boldsymbol{Y} = (Y_1, ..., Y_N)$ is created by $Y_i = f(U_i, X_i)$, for some function f, and the only quantitative characterization on the possible compression rate is that it is upper-bounded by $H(Y)$. The mechanism of generating the composite sequences, proposed in this paper, is different and more complex than that of Gel'fand and Pinsker [20] (and these proposed in [7]-[10]) and this is in order to provide an enumerable composite set and to maintain the distortion constraint.

The proof of the converse part is strongly based on that of [20]. It should be noted that both the proposed coding scheme and the converse proof camouflage the fact that although the schemes were originally planned to provide reliable retrieval of the watermark from the distorted version of the composite sequence, reliable retrieval of the watermark is possible also directly from the composite sequence.

Corollary 1 presents the result of Theorem 1 in terms of the maximum achievable embedding rate - $C_e(R_c, \Delta)$. Another way to present Theorem 1 is in terms of the minimum

achievable composite rate:

Corollary 2. *The minimum achievable composite rate $R_c^*(R_e, \Delta)$ is given by:*

$$R_c^*(R_e, \Delta) = R_e + \min I(X; U, Y), \tag{24}$$

where the minimum is over $\mathcal{A} \bigcap \{(U, X, Y) : R_e \leq I(U; Z) - I(U; X)\}$.

In an attack-free case [22], where $Z = Y$ and the optimal choice of the auxiliary is $U = Y$, the result of (22) coincides with that of [22].

An extension of this work can be done for the case of continuous alphabets ([18], ch. 7), by considering the supremum of $\min\{I(U_d; Z_p) - I(U_d; X_p), R_c - I(X_p; U_d, Y_p)\}$ over all finite-alphabet auxiliary variables U_d and all partitions X_d, Y_d and Z_d of the source, channel input and channel output alphabets, respectively. The achievability scheme can also be presented directly, following the lines of the scheme provided in Section 6, where the considered sequences should satisfy weak typicality (see, e.g., [16], pp. $225 - 227$) rather than strong typicality used for the finite-alphabets case. This scheme demonstrates well a great difference between the continuous and the finite-alphabets cases: the number of composite sequences is finite for the finite-alphabet case, while for the case of continuous alphabets, there are infinitely many usable auxiliary and stegotext sequences. So, we can generate arbitrarily many distinct auxiliary and composite codebooks, that differ from each other only by arbitrarily small perturbations of one (representative) auxiliary and one corresponding stegotext codebook, with each auxiliary codebook representing a different watermarking message. The minimum sizes of these auxiliary and composite codes are dictated by properties of typical sequences, and a variation of the Rate-Distortion Theorem [17], respectively, establishing the first upper-bound to the embedding rate, in terms of the composite rate allowed. Now, in the presence of the attack, the number of usable auxiliary sequences is limited by the standard channel-coding argument, i.e., we cannot use more auxiliary codewords than we can distinguish at the output of the attack channel, and as a consequence, an additional upper-bound to R_e is determined. But, in an attack-free case, no channel coding is performed, the watermark is retrievable directly from the sent composite sequence, and therefore, the upper-bound $I(U; Z) - I(U; X)$ to R_e can be omitted, bringing us back to the result of [22], since the choice of $U = Y$ provides us with the highest achievable embedding rate for a given R_c.

5 Proof of the Converse Part of Theorem 1

Let $\left(\phi^N, \varphi^N\right)$ be a given encoder-decoder pair for which $Ed(\boldsymbol{X}, \boldsymbol{Y}) \leq \Delta$, $\frac{1}{N} H(\phi^N(\boldsymbol{X}, V)) \leq R_c$ and $P_e \leq \epsilon$. We start with Fano's inequality:

$$H(V|\boldsymbol{Z}) \leq h(P_e) + P_e \log(M-1) \leq 1 + P_e N R_e, \tag{25}$$

where $h(\cdot)$ is the binary entropy function. Since $V \rightarrow \boldsymbol{Y} \rightarrow \boldsymbol{Z}$ is a Markov chain, (25) implies:

$$H(V|\,\boldsymbol{Y}) \leq H(V|\boldsymbol{Z}) \leq 1 + P_e N R_e. \tag{26}$$

The embedding rate can therefore be upper-bounded as follows:

$$\begin{aligned}
NR_e &\overset{(a)}{=} H(V) \\
&= H(V|\boldsymbol{X}) \\
&= H(V|\boldsymbol{X}, \boldsymbol{Y}) + I(V; \boldsymbol{Y}|\boldsymbol{X}) \\
&= H(V|\,\boldsymbol{Y}) + I(V; \boldsymbol{Y}|\boldsymbol{X}) - I(V; \boldsymbol{X}|\,\boldsymbol{Y}) \\
&\overset{(b)}{\leq} 1 + P_e N R_e + H(\,\boldsymbol{Y}|\boldsymbol{X}) - H(\,\boldsymbol{Y}|V, \boldsymbol{X}) - I(V; \boldsymbol{X}|\,\boldsymbol{Y}) \\
&\overset{(c)}{=} 1 + P_e N R_e + H(\,\boldsymbol{Y}) - I(\boldsymbol{X}; \boldsymbol{Y}) - I(V; \boldsymbol{X}|\,\boldsymbol{Y}) \\
&= 1 + P_e N R_e + H(\,\boldsymbol{Y}) - I(\boldsymbol{X}; V, \boldsymbol{Y}) \\
&\leq 1 + P_e N R_e + N R_c - I(\boldsymbol{X}; V, \boldsymbol{Y}),
\end{aligned} \tag{27}$$

where:

- (a) follows from the assumption V has a uniform distribution,
- (b) from (26),
- (c) from the fact that $\boldsymbol{Y}$ is a function of $(\boldsymbol{X}, V)$.

Following [20], let us define N auxiliary random variables $\widetilde{U}(1), ...\widetilde{U}(N)$:

$$\widetilde{U}(i) = (V, Z_1^{i-1}, X_{i+1}^N). \tag{28}$$

Then,

$$\begin{aligned}
I(\boldsymbol{X}; V, \boldsymbol{Y}) &\overset{(a)}{=} I(\boldsymbol{X}; V, \boldsymbol{Y}, \boldsymbol{Z}) \\
&= \sum_{i=1}^{N} I(X_i; V, \boldsymbol{Y}, \boldsymbol{Z}|X_{i+1}^N)
\end{aligned} \tag{29}$$

$$
\begin{aligned}
&\stackrel{(b)}{=} \sum_{i=1}^{N} I(X_i; V, \boldsymbol{Y}, \boldsymbol{Z}|X_{i+1}^N) + \sum_{i=1}^{N} I(X_i; X_{i+1}^N) \\
&= \sum_{i=1}^{N} I(X_i; V, \boldsymbol{Y}, \boldsymbol{Z}, X_{i+1}^N) \\
&= \sum_{i=1}^{N} I(X_i; V, Z_1^{i-1}, X_{i+1}^N, Y_i, Y_1^{i-1}, Y_{i+1}^N, Z_i^N) \\
&\stackrel{(c)}{\geq} \sum_{i=1}^{N} I(X_i; \widetilde{U}(i), Y_i), \qquad (30)
\end{aligned}
$$

where:

(a) follows from the Markov chain $\boldsymbol{X} \rightarrow (V, \boldsymbol{Y}) \rightarrow \boldsymbol{Z}$. (31)

(b) from the fact that the covertext source is memoryless, and

(c) from the data processing theorem and (28).

On substituting (29) into (27), we obtain

$$NR_e \leq 1 + P_e NR_e + NR_c - \sum_{i=1}^{N} I(X_i; \widetilde{U}(i), Y_i). \qquad (32)$$

Also, from the proof of the converse part of [20], it is known that

$$NR_e \leq 1 + P_e NR_e + \sum_{i=1}^{N} [I(\widetilde{U}(i); Z_i) - I(\widetilde{U}(i); X_i)]. \qquad (33)$$

Combining (32) with (33) and dividing the resulting inequality by N, gives:

$$R_e(1 - P_e) - \frac{1}{N} \leq \min \left\{ \frac{1}{N} \sum_{i=1}^{N} [I(\widetilde{U}(i); Z_i) - I(\widetilde{U}(i); X_i)], R_c - \frac{1}{N} \sum_{i=1}^{N} I(X_i; \widetilde{U}(i), Y_i) \right\}. \qquad (34)$$

Now, consider a time-sharing random variable T distributed uniformly over $\{1, 2, ..., N\}$, independently of all other random variables in the system, and let us denote a quadruple of random variables

$$(\widetilde{U}, X, Y, Z) \stackrel{\triangle}{=} (\widetilde{U}_T, X_T, Y_T, Z_T). \qquad (35)$$

The probability distribution of $(\widetilde{U}, X, Y, Z)$ is given by:

$$\Pr\{(\widetilde{U}, X, Y, Z) = (\widetilde{u}, x, y, z)\} = \frac{1}{N} \sum_{n=1}^{N} \Pr\{(\widetilde{U}_n, X_n, Y_n, Z_n) = (\widetilde{u}, x, y, z)\}. \qquad (36)$$

Therefore, by definition of T:

$$\frac{1}{N} \sum_{i=1}^{N} [I(\widetilde{U}(i); Z_i) - I(\widetilde{U}(i); X_i)] \quad = \quad I(\widetilde{U}; Z|T) - I(\widetilde{U}; X|T) \qquad (37)$$

$$
\begin{aligned}
&= I(\widetilde{U},T;Z) - I(Z;T) - I(\widetilde{U},T;X) + I(X;T) \\
&\leq I(\widetilde{U},T;Z) - I(\widetilde{U},T;X),
\end{aligned}
$$

where the last step is due to the fact that $\boldsymbol{X}$ is stationary and memoryless and hence, $I(X;T) = 0$, and the fact that $I(Z;T)$ is non-negative. Also,

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} I(X_i;\widetilde{U}(i),Y_i) &= I(X;\widetilde{U},Y|T) \\
&= I(X;\widetilde{U},Y|T) + I(X;T) \\
&= I(X;\widetilde{U},T,Y),
\end{aligned} \tag{38}
$$

where the second equality is again due to the fact that $\boldsymbol{X}$ is stationary and memoryless.

Let us define now a new random variable $U \stackrel{\triangle}{=} (\widetilde{U},T)$. Exchanging variables in (37) and (38) and substituting the obtained result into (34), provides us with the following expression:

$$
R_e(1-P_e) - \frac{1}{N} \leq \min\big\{I(U;Z) - I(U;X), R_c - I(X;U,Y)\big\}. \tag{39}
$$

By hypothesis, the given system satisfies $P_e \leq \epsilon$, and hence, by taking the limit $\epsilon \to 0$ as $N \to \infty$ in (39), we obtain:

$$
R_e \leq \min\big\{I(U;Z) - I(U;X), R_c - I(X;U,Y)\big\}. \tag{40}
$$

Next, the expected distortion constraint is satisfied by the system, and so,

$$
\begin{aligned}
\Delta &\geq Ed(\boldsymbol{X},\boldsymbol{Y}) \\
&= \sum_{\boldsymbol{x},\boldsymbol{y}} Pr\{(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{x},\boldsymbol{y})\}\frac{1}{N}\sum_{i=1}^{N} d(x_i,y_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{x,y} Pr\{(X_i,Y_i) = (x,y)\}d(x,y) \\
&= \sum_{x,y} Pr\{(X,Y) = (x,y)\}d(x,y) \\
&= Ed(X,Y).
\end{aligned} \tag{41}
$$

It remains to show that the alphabet of the random variable U can be limited by $|U| \leq |\mathcal{X}|\cdot|\mathcal{Y}|+1$. To this end, we will use the support lemma (cf. [15]), which is based on Carathéodory's theorem, according to which, given J real valued continuous functionals f_j, $j = 1,...,J$ on the set $\mathcal{P}(\mathcal{X})$ of probability distributions over the alphabets $\mathcal{X}$, and given any

probability measure μ on the Borel σ-algebra of $\mathcal{P}(\mathcal{X})$, there exist J elements $Q_1, ...Q_J$ of $\mathcal{P}(\mathcal{X})$ and J non-negative reals, $\alpha_1, ..., \alpha_J$, such that $\sum_{j=1}^{J} \alpha_j = 1$ and for every $j = 1, ..., J$

$$\int_{\mathcal{P}(\mathcal{X})} f_j(Q)\mu(dQ) = \sum_{i=1}^{J} \alpha_i f_j(Q_i). \tag{42}$$

Before we actually apply the support lemma, we first rewrite the relevant mutual informations of (40) in a more convenient form for the use of this lemma. As for the first upper bound to R_e, we have:

$$I(U;Z) - I(U;X) \quad = \quad H(Z) - H(Z|U) - H(X) + H(X|U), \tag{43}$$

and for the second upper bound to R_e, we have

$$\begin{aligned} R_c - I(X;U,Y) &= R_c - I(X;U) - I(X;Y|U) && (44)\\ &= R_c - H(X) + H(X|U) - H(X|U) + H(X|U,Y) \\ &= R_c - H(X) + H(X|U,Y) \\ &= R_c - H(X) + H(X,Y|U) - H(Y|U). && (45) \end{aligned}$$

For a given joint distribution of (X, Y, Z), $H(Z)$ and $H(X)$ are both given and unaffected by U. Therefore, in order to preserve prescribed values of $I(U;Z) - I(U;X)$ and $R_c - I(X;U,Y)$, it is sufficient the preserve the associated values $H(X|U) - H(Z|U)$ and $H(X,Y|U) - H(Y|U)$.

Let us define the the following functionals of a generic distribution Q over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \times \mathcal{Y}$ is assumed, without loss of generality, to be $\{1, 2, ..., m\}$, $m \stackrel{\triangle}{=} |\mathcal{X}| \cdot |\mathcal{Y}|$:

$$f_i(Q) = Q(x,y), \ i \stackrel{\triangle}{=} (x,y) = 1, ..., m-1 \tag{46}$$

$$f_m(Q) = \sum_{x,y} Q(x,y) \sum_{z} P_{Z|Y}(z|y) \log \frac{\sum_{x,y} Q(x,y) P_{Z|Y}(z|y)}{\sum_{y,z} Q(x,y) P_{Z|Y}(z|y)}. \tag{47}$$

Next define

$$f_{m+1}(Q) = \sum_{x,y} Q(x,y) \log \frac{\sum_x Q(x,y)}{Q(x,y)}. \tag{48}$$

Applying now the support lemma, we find that there exists a random variable U (jointly distributed with (X,Y)), whose alphabet size is $|U| = m + 1 = |\mathcal{X}| \cdot |\mathcal{Y}| + 1$ and it satisfies simultaneously:

$$\sum_{u} \Pr\{U = u\} f_i(P(\cdot|u)) = P_{XY}(x,y), \ i = 1, ..., m-1, \tag{49}$$

$$\sum_u \Pr\{U=u\} f_m(P(\cdot|u)) = H(X|U) - H(Z|U), \tag{50}$$

and

$$\sum_u \Pr\{U=u\} f_{m+1}(P(\cdot|u)) = H(X,Y|U) - H(Y|U). \tag{51}$$

It should be pointed out that this random variable maintains the prescribed distortion level $Ed(\boldsymbol{X}, \boldsymbol{Y})$ of the system, since the $P_{XY}(x,y)$ is preserved. This completes the proof of the converse part.

6 Proof of the Direct Part of Theorem 1

In this section, we show that given a triple of random variables $(U,X,Y) \in \mathcal{A}$ and positive numbers R_e, R_c and Δ such that $R_e \leq \min\{I(U;Z) - I(U;X), R_c - I(X;U,Y)\}$ and $Ed(X,Y) \leq \Delta$, then for any $\epsilon > 0$ and sufficiently large N, there exists a code of embedding rate R_e, for the attack channel $P_{Z|Y}$, with composite rate below R_c, error probability $P_e \leq \epsilon$, and $Ed(\boldsymbol{X}, \boldsymbol{Y}) \leq (1+\epsilon)\Delta$.

Let us denote three functions of a scalar $\delta > 0$, which will be used later on:

$$\epsilon_b \stackrel{\triangle}{=} (\delta^2 - 2\delta)H(U) - (\delta^2 + 2\delta)H(U|X) - \delta, \tag{52}$$

$$\epsilon_y \stackrel{\triangle}{=} (\delta^2 - 2\delta)H(Y|U) - (\delta^2 + 2\delta)H(Y|X,U) - \delta, \tag{53}$$

and

$$\epsilon_u \stackrel{\triangle}{=} (\delta^2 - 2\delta)H(Z) - (\delta^2 + 2\delta)H(Z|U) - \delta. \tag{54}$$

We next describe the mechanisms of random code selection and the encoding and decoding operations. Fix δ such that $2\delta + \max\{2 \cdot \exp^{-2^{N\delta}} + 2^{-N\delta}, \delta^2\} \leq \epsilon$.

Auxiliary Code Generation:

We first construct an auxiliary code capable of embedding 2^{NR_e} watermarks by a random selection technique. First, 2^{NR_u}, $R_u \leq I(U;Z) - \epsilon_u - \delta$, sequences $\{\boldsymbol{U}_i\}$, $i \in [1, ..., 2^{NR_u}]$, are drawn independently from $T^{\delta}_{P_U}$. Let us denote the set of these sequences by $\mathcal{C}$. The elements of $\mathcal{C}$ are equally distributed between $M \stackrel{\triangle}{=} 2^{NR_e}$ bins, each bin of size $m = 2^{NR}$, $R \geq I(X;U) + \epsilon_b + \delta$. A different watermark index is attached to each bin, identifying a

sub-code representing the watermark. We denote the codewords of bin v, $v \in [1, 2, ..., M]$, by $\boldsymbol{U}(v, k)$, $k \in [1, 2, ..., m]$.

Composite Sequence Generation:
For each auxiliary sequence $\boldsymbol{U}(v, k) = \boldsymbol{u}$, a set of $m_y \stackrel{\triangle}{=} 2^{NR_y}$, $R_y \geq I(X;Y|U) + \epsilon_y + \delta$, composite sequences $\{\boldsymbol{Y}_j\}$, $j \in [1, ..., m_y]$, are independently drawn from $T^{\delta}_{P_{Y|U}}(\boldsymbol{u})$. We denote this set by $\mathcal{C}(\boldsymbol{U}(v, k))$ and its elements by $\boldsymbol{Y}(v, \mathcal{C}(\boldsymbol{U}(v, k)), j)$. Note that the 2^{NR_u} sets $\{\mathcal{C}(\boldsymbol{U}(v, k))\}$ may not be all mutually exclusive.

Encoding/Embedding:
Upon receiving a pair $(\boldsymbol{x}, v)$, the encoder acts as follows:

1. If $\boldsymbol{x} \in T^{\delta}_{P_X}$ and bin number v contains a sequence $\boldsymbol{U}(v, k) = \boldsymbol{u}$ such that (s.t.) the pair $(\boldsymbol{x}, \boldsymbol{u}) \in T^{\delta}_{P_{XY}}$, the first $\boldsymbol{Y}(v, \mathcal{C}(\boldsymbol{U}(v, k)), j) = \boldsymbol{y}$ found in $\mathcal{C}(\boldsymbol{U}(v, k))$, such that $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}) \in T^{\delta}_{P_{XUY}}$, is chosen for transmission. If there exist more than one jointly δ-typical with $\boldsymbol{x}$ sequences, the described above process is applied to the the first matching $\boldsymbol{U}(v, k)$ found in a bin's list.

2. If $\boldsymbol{x} \notin T^{\delta}_{P_X}$, or $\nexists \boldsymbol{U}(v, k) = \boldsymbol{u}$ s.t. $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}) \in T^{\delta}_{P_{XUY}}$, an arbitrary error message is transmitted.

Decoding:
Upon receiving $\boldsymbol{Z} = \boldsymbol{z}$, the decoder finds all sequences $\{\boldsymbol{U}(v, k) = \boldsymbol{u}\}$, so that the pairs $(\boldsymbol{u}, \boldsymbol{z}) \in T^{\delta}_{P_{UZ}}$. If all found $\{\boldsymbol{U}(v, k)\}$ belong to a single bin, the index of this bin is decoded as the watermark $\widetilde{v}$. Otherwise (if $\nexists \boldsymbol{U}(v, k) = \boldsymbol{u}$ s.t. $(\boldsymbol{u}, \boldsymbol{z}) \in T^{\delta}_{P_{UZ}}$ or there exist more than one bin containing such a sequence), an error is declared.

We now turn to the analysis of the error probability, the distortion, and the compressibility of the composite sequence. For each pair $(v, \boldsymbol{x})$, a particular choice of a code $\mathcal{C}$ and related choices of $\{\mathcal{C}(\boldsymbol{U}(v, k))\}$, the possible causes for incorrect watermark decoding are the following:

1. $\boldsymbol{x} \notin T^{\delta}_{P_X}$. Let the probability of this event be defined as P_{e_1}.

2. $\boldsymbol{x} \in T^{\delta}_{P_X}$, but in bin no. v $\nexists \boldsymbol{u}$ s.t. $(\boldsymbol{x}, \boldsymbol{u}) \in T^{\delta}_{P_{XU}}$. Let the probability of this event be defined as P_{e_2}.

3. $\boldsymbol{x} \in T^{\delta}_{P_X}$, and bin no. v contains $\boldsymbol{U}(v,k) = \boldsymbol{u}$ s.t. $(\boldsymbol{x}, \boldsymbol{u}) \in T^{\delta}_{P_{XU}}$, but $\nexists \boldsymbol{y} \in \mathcal{C}(\boldsymbol{U}(v,k))$ s.t. $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}) \in T^{\delta}_{P_{XUY}}$. Let the probability of this event be defined as P_{e_3}.

4. $\boldsymbol{x} \in T^{\delta}_{P_X}$, and bin no. v contains $\boldsymbol{u}$ s.t. $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}) \in T^{\delta}_{P_{XUY}}$, but $(\boldsymbol{u}, \boldsymbol{z}) \notin T^{\delta}_{P_{UZ}}$. Let the probability of this event be defined as P_{e_4}.

5. $\boldsymbol{x} \in T^{\delta}_{P_X}$, in bin no. v $\exists \boldsymbol{u}$ s.t. $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}) \in T^{\delta}_{P_{XUY}}$ and $(\boldsymbol{u}, \boldsymbol{z}) \in T^{\delta}_{P_{UZ}}$, but there exists another bin no. $\widetilde{v}$ that contains $\widetilde{\boldsymbol{u}}$ s.t. $(\boldsymbol{z}, \widetilde{\boldsymbol{u}}) \in T^{\delta}_{P_{UZ}}$. Let the probability of this event be defined as P_{e_5}.

If none of those events occur, the message v is retrieved correctly from $\boldsymbol{z}$, and the distortion constraint between $\boldsymbol{x}$ and $\boldsymbol{y}$ is satisfied, as follows from (12).

The average probability of error P_e is bounded by

$$P_e \leq P_{e_1} + P_{e_2} + P_{e_3} + P_{e_4} + P_{e_5}. \tag{55}$$

The fact that $P_{e_1} \to 0$ follows from (9). As for P_{e_2}, we have:

$$P_{e_2} \stackrel{\triangle}{=} \prod_{k=1}^{m} \Pr\{(\boldsymbol{x}, \boldsymbol{U}(v,k)) \notin T^{\delta}_{P_{XU}}\}. \tag{56}$$

Now, by (8), for every k:

$$\begin{aligned} \Pr\{(\boldsymbol{x}, \boldsymbol{U}(v,k)) \notin T^{\delta}_{P_{XU}}\} &= 1 - \Pr\{(\boldsymbol{x}, \boldsymbol{U}(v,k)) \in T^{\delta}_{P_{XU}}\} \\ &= 1 - \frac{|T^{\delta}_{P_{U|X}}(\boldsymbol{x})|}{|T^{\delta}_{P_U}|} \\ &\leq 1 - \frac{2^{N[(1+\delta)^2 H(U|X)]}}{2^{N[(1-\delta)^2 H(U) - \delta]}} \\ &= 1 - 2^{-N[I(X;U) + \epsilon_b]}, \end{aligned} \tag{57}$$

where ϵ_b is given by (52). Substitution of (57) into (56) provides us with the following upper-bound:

$$P_{e_2} \leq \left[1 - 2^{-N[I(X;U)+\epsilon_b]}\right]^m \leq \exp\left\{ - 2^{NR} \cdot 2^{-N[I(X;U)+\epsilon_b]} \right\} \to 0, \tag{58}$$

double-exponentially rapidly since $R \geq I(X;U) + \epsilon_b + \delta$.

To estimate P_{e_3}, we repeat the technique of the previous step:

$$P_{e_3} \stackrel{\triangle}{=} \prod_{j=1}^{m_y} \Pr\{(\boldsymbol{x},\, \boldsymbol{Y}(v, \mathcal{C}(\,\boldsymbol{U}(v,k)), j)) \notin T^{\delta}_{P_{XY}}\}. \tag{59}$$

Again, by the property of the typical sequences, for every j:

$$\Pr\{(\boldsymbol{x},\, \boldsymbol{Y}(v, \mathcal{C}(\,\boldsymbol{U}(v,k)), j)) \notin T^{\delta}_{P_{XY}}\} \leq 1 - 2^{-N[I(X;Y|U)+\epsilon_y]}, \tag{60}$$

where ϵ_y is given by (53) and therefore, substitution of (60) into (59) gives

$$P_{e_3} \leq \left[1 - 2^{-N[I(X;Y|U)+\epsilon_y]}\right]^{m_y} \leq \exp\left\{ - 2^{NR_y} \cdot 2^{-N[I(X;Y|U)+\epsilon_y]} \right\} \to 0, \tag{61}$$

double-exponentially rapidly since $R_y \geq I(X;Y|U) + \epsilon_y + \delta$.

The estimation of P_{e_4} is again based on property of typical sequences. Since $\boldsymbol{Z}$ is an output of N successive uses of a memoryless attack channel $P_{Z|Y}$ with input $\boldsymbol{Y}(v, \mathcal{C}(\,\boldsymbol{U}(v,k)), j)$ and by the assumption of this step $(\boldsymbol{x},\, \boldsymbol{U}(v,k),\, \boldsymbol{Y}(v, \mathcal{C}(\,\boldsymbol{U}(v,k)), j)) \in T^{\delta}_{P_{XUY}}$, from (9) we obtain

$$P_{e_4} = \Pr\{(\boldsymbol{x},\, \boldsymbol{U}(v,k),\, \boldsymbol{Y}(v, \mathcal{C}(\,\boldsymbol{U}(v,k)), j),\, \boldsymbol{Z}) \notin T^{\delta}_{P_{XUYZ}}\} \leq \delta, \tag{62}$$

and similarly to P_{e_1} can be made as small as desired by an appropriate choice of δ.

Finally, we estimate P_{e_5} as follows:

$$\begin{aligned} P_{e_5} &= \Pr\{\exists \widetilde{v} \neq v : (\,\boldsymbol{U}(\widetilde{v},k),\, \boldsymbol{Z}) \in T^{\delta}_{P_{UZ}}\} && (63) \\ &\leq \sum_{\widetilde{v}\neq v, k\in[1,2,\ldots,m]} \Pr\{(\,\boldsymbol{U}(\widetilde{v},k),\, \boldsymbol{Z}) \in T^{\delta}_{P_{UZ}}\} \\ &\leq (2^{NR_e} - 1) 2^{NR} \Pr\{\,\boldsymbol{U}(\widetilde{v},k),\, \boldsymbol{Z} \in T^{\delta}_{P_{UZ}}\} \\ &\leq 2^{NR_u} 2^{-N[I(U;Z)-\epsilon_u]}, && (64) \end{aligned}$$

where ϵ_u is given by (54). Now, since $R_u \leq I(U;Z) - \epsilon_u - \delta$, $P_{e_5} \to 0$.

Since $P_{e_i} \to 0$ for i = 1,2,3,4,5, their sum tends to zero as well, implying that there exist at least one choice of an auxiliary code $\mathcal{C}$ and related choices of sets $\{\mathcal{C}(\,\boldsymbol{U}(v,k))\}$ that give rise to the reliable watermark decoding.

The embedding rate of the above described scheme is determined by the maximum possible number of auxiliary bins, i.e.,

$$\begin{aligned} R_e &= \frac{1}{N} \log\left(\frac{R_u}{m}\right) && (65) \\ &\leq I(U;Z) - I(U;X) - \epsilon_b - \epsilon_u - 2\delta. \end{aligned}$$

Now, let us denote by N_c the total number of composite sequences used in the described above scheme:

$$N_c = M \cdot m \cdot m_y = 2^{N[R_e + I(X;U,Y) + \epsilon_b + \epsilon_y]}. \tag{66}$$

For sufficiently small values of δ, ϵ_b and ϵ_y vanish and can be neglected, giving $N_c = 2^{N[R_e + I(X;U,Y)]}$. Now, since the compression procedure applied to the composite sequences is lossless, it satisfies

$$\frac{1}{N} H(\boldsymbol{Y}) \leq \frac{1}{N} \log(N_c) = R_e + I(X;U,Y) \leq R_c, \tag{67}$$

which completes the proof of the direct part. Finally, since ϵ_b, ϵ_u and δ are arbitrarily small, R_e can be made as close as desired to $\min\{I(U;Z) - I(U;X), R_c - I(X;U,Y)\}$.

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] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563–593, March 2003.

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

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

[10] 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.

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

[12] D. Karakos and A. Papamarcou, "A Relationship Between Quantization and Watermarking Rates in the Presence of Additive Gaussian Attacks," *IEEE Trans. Inform. Theory*, vol. 49, no. 8, pp. 1970-1982, August 2003.

[13] D. Karakos, "Digital Watermarking, Fingerprinting and Compression: An Information-Theoretic Perspective", Ph.D. Thesis, UMD, College Park, MD, 2002. Available on-line at **http://www.ece.umd.edu/ karakos/publications.html**.

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

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

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

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

[18] R.G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[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.

[22] A. Maor and N. Merhav, "On Joint Information Embedding and Lossy Compression," submitted to *IEEE Trans. Inform. Theory*, July 2003. Also, Technical Report, CCIT Pub. no. 450, EE Pub. no. 1391, Technion – I.I.T., November 2003. Available at **http://www.ee.technion.ac.il/people/merhav/**.