A Relationship between Quantization and Distribution Rates of Digitally Fingerprinted Data[†]

Damianos Karakos
karakos@eng.umd.edu

Adrian Papamarcou
adrian@eng.umd.edu

Department of Electrical and Computer Engineering
and Institute for Systems Research
University of Maryland
College Park, MD 20742

Abstract

This paper considers a fingerprinting system where 2^{nR_W} distinct Gaussian fingerprints are embedded in respective copies of an n-dimensional i.i.d. Gaussian image. Copies are distributed to customers in digital form, using R_Q bits per image dimension. By means of a coding theorem, a rate region for the pair (R_Q, R_W) is established such that (i) the average quadratic distortion between the original image and each distributed copy does not exceed a specified level; and (ii) the error probability in decoding the embedded fingerprint in the distributed copy approaches zero asymptotically in n.

[†] This work was supported by the National Science Foundation under NSF Grant CDA-9700866. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Sorrento, Italy, June 25-30, 2000.

1 Introduction

The widespread use of digital data in commercial applications over the last few years has increased the need for copyright protection and authentication schemes. Especially for audio, image, video or multimedia data, information hiding has been suggested in the literature as the most effective means for protecting against unlawful use of the data (e.g. see [1, 2]). In a general framework, this is done by embedding a message into a host data set, such that (i) the hidden message does not perceptually interfere with the work being protected (a distortion constraint has to be satisfied), and (ii) the message must be difficult or impossible to remove without severely degrading the fidelity of the protected work. This hidden message can play the role either of a *watermark* or a *fingerprint*, depending on the application; a watermark carries copyright information related to the rightful owner of the protected work, and a fingerprint uniquely identifies each individual copy distributed, making it possible to trace any illegally distributed data back to the user [3, 4]. In some papers (e.g. [1, 5]) these two terms are used interchangeably. Also, note that in the steganography literature the original host image is called *covertext* and the resultant watermarked image is called *stegotext*.

Recently, there have been various approaches to information hiding, from an information-theoretic perspective. In [3, 6] O'Sullivan *et al.* give a general expression for the maximum rate of the set of messages that can be hidden within a host data set (hiding capacity) subject to an average distortion constraint, as well as the requirement that the message withstand *any* deliberate attack (subject to another average distortion constraint) aimed to destroy it. In this framework, the information hider and the attacker play a game, in which the hider selects the distribution on the watermarks such that it maximizes the mutual information between the hidden information and the output of the attack channel, while the attacker tries to minimize this mutual information. The authors assume that the attacker knows the hiding channel and that the decoder of the watermark knows the attack channel (additionally to the hiding channel). Also, the hider and the decoder share some common side information (e.g. a key or the host signal itself). For the case where the alphabets are continuous, it is shown that the optimal distributions used by the hider and the attacker have to be Gaussian (saddlepoint condition), provided the host data set is Gaussian distributed and the distortion measure is the mean square. Moreover, in this case, knowledge of the host signal at the decoder does not increase the hiding capacity. Additionally, in [7], Merhav considered a similar problem to [3], but from the point of view of computing the exponents of the probability of error.

Another version of the watermarking game has been investigated in [8, 9] by Cohen and Lapidoth, where the distribution on the data set is assumed Gaussian, and the watermark encoder and decoder are designed irrespective of the attacker model (that is, the watermark decoder does not use a maximum likelihood rule with respect to the attack). Also, [8, 9] considered both *peak* and *average* distortion constraints between the watermarked and the original image (as well as between the watermarked and the attacked image). It is proved that in the case of average distortion constraints, the coding capacity is zero. Moreover, knowledge of the host signal at the decoder does not increase the capacity.

Another interesting watermarking scheme is *quantization index modulation*, developed by Chen and Wornell [10, 11], in which the host signal is compressed by a quantizer that depends on the message to be hidden. An information-theoretic analysis of this system has also been developed [12]. In [13], Steinberg and Merhav consider the problem of watermark identification (i.e., detection whether a particular watermark resides in the covertext) and give bounds on the identification capacity for two alternative cases, where the covertext is known, or is not known, to the decoder.

In this paper, we study a problem that combines source and channel coding in a fingerprinting framework (see also [14]). This problem is motivated by the following scenario. A data distributor (e.g. a news agency) has to deliver an information sequence I^n (e.g., a digital image) to $M_n = 2^{nR_W}$ customers, such that each customer receives a different fingerprinted version of I^n. We call R_W the *distribution rate* of the fingerprints (or, equivalently, the *fingerprinting rate*). To that end, the agent creates M_n distinct signals $x^n(1), \dots, x^n(M_n)$ (the fingerprints) and uses them to generate M_n fingerprinted copies of I^n. In order for the fingerprints to be usable for a variety of data, they are created *independently* of the host signal I^n. Due to bandwidth limitations, the agent compresses the fingerprinted data at a rate of R_Q bits per image dimension subject to a fidelity criterion, prior to distribution.*

For security purposes, as well as for maximum usability, we assume that both the quantization and the reconstruction of the image are *independent* of the choice of the fingerprint set. This can be particularly useful in case the quantization is performed by an authority other than the information hider. In this way, the fingerprint set

*Note that in this paper we use the more precise term *fingerprint*, instead of *watermark* that we used in [14].

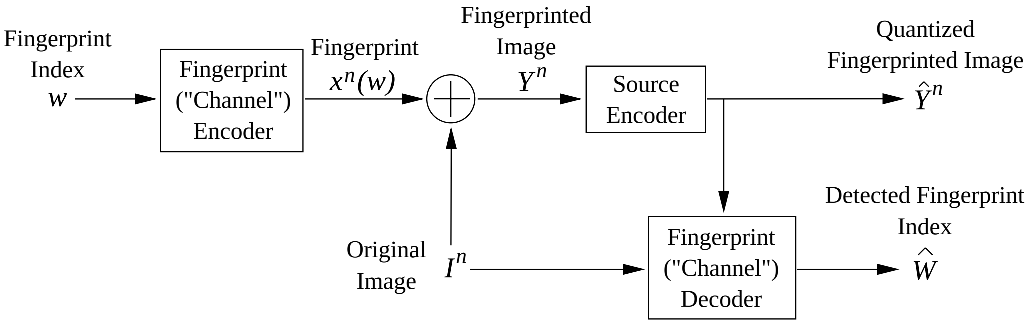

Figure 1: The fingerprinting/authentication system with quantization

does not need to be revealed to any intermediaries (otherwise, the security of the system would be at stake). In addition, the agent who generated the image should be able to discern which fingerprint is present in a fingerprinted (and subsequently compressed) image with a low probability of error (e.g., in case an authenticator needs to track down the initial owner of an illegally distributed image), using the original image as side information. In other words, fingerprints and source codewords have to be designed in such a way that knowledge of the fingerprint set and the original data suffices for detecting reliably the fingerprint in the compressed image.

This fingerprinting/compression system is depicted in Figure 1. Note that although there is no transmission medium involved, the quantizer acts as a deterministic channel that degrades the fingerprinted image. We will thus refer the fingerprint encoding/decoding as *channel* encoding/decoding. The main result of this paper is the determination of the allowable rates R_Q and R_W for the above system, under some weak assumptions described in the next section.

In comparison to the scenarios studied in [3, 9], our model considers a single fidelity criterion, namely the resultant distortion between the original data sequence and the fingerprinted/quantized data. And while quantization degrades the fingerprinted image, it cannot be construed as a malicious attack of the type modeled in [3, 9]. In our case, data compression and fingerprinting are cooperative (not competing) schemes, and must be optimized jointly. Moreover, we assume that the fingerprinted/quantized image is not further corrupted by any attack; therefore our result on the rate region can be considered as an *outer bound* on the achievable rate region obtained when an attack channel is present.

The paper is organized as follows: in Section 2 we give basic definitions and assumptions used in our model. The coding theorem is presented in Section 3, along with a sketch of the proof of the achievability and converse theorems. The complete proofs of the converse and the forward parts are given in Sections 4 and 5 respectively. Conclusions and directions for further research are given in Section 6.

2 Model and Assumptions

We first proceed to give some definitions pertaining to the system shown in Figure 1. In the sequel, all random quantities (scalars or vectors) appear in upper case (e.g. W, X^n, Y^n). Lower case is used to denote non-random quantities (e.g. w, x^n, n). For example, for a specific w, $x^n(w)$ is a deterministic vector, while $X^n = x^n(W)$ is a random vector—specifically, a deterministic mapping of a random variable.

Definition 1 *An $(2^{nR_Q}, n)$ source code consists of a codebook of n-dimensional vectors $\{\hat{y}^n(1), \ldots, \hat{y}^n(2^{nR_Q})\}$ and an encoder f which maps the image space $\mathbf{R}^n$ into that codebook.*

The mapping f does not depend on the particular fingerprint set used (i.e. the compression is done in a *fingerprint-independent* fashion). This can be particularly useful in cases where the compression is performed by an authority other than the information hider; then the hider need not reveal the fingerprints to any intermediaries and the security of the fingerprinting system is not compromised. The codebook is obviously available to the users, thus transmission of the fingerprinted image to each user requires no more than nR_Q bits.

Since the fingerprints have to be recoverable from the quantized image, we have a channel coding counterpart in our model, and thus the following definition.

Definition 2 *A $(2^{nR_W}, n)$ fingerprint code consists of a codebook of n-dimensional fingerprints $\{x^n(1), \ldots, x^n(2^{nR_W})\}$ and a decoder*

$$g : \mathbf{R}^n \times \{\hat{y}^n(1), \ldots, \hat{y}^n(2^{nR_Q})\} \rightarrow \{1, \ldots, 2^{nR_W}\}$$

The output of the decoder is denoted by $\hat{w}$. The notation $M_n \overset{def}{=} 2^{nR_W}$ will also be used.

The decoder g is known only to the agent/authenticator.

We make the following basic assumptions about the image and fingerprint model:

- The original image (or other multimedia data) I^n is i.i.d. $\mathcal{N}(0, P_I)$.
- Fingerprinting is additive, i.e., the fingerprinted image can be represented as

  $$Y^n = I^n + x^n(W)$$

  where W is the (random, in general) fingerprint index.
- The fingerprints $x^n(1), \ldots, x^n(M_n)$ satisfy the following:

  $$\lim_{n\to\infty} \max_{1\leq i\leq n} \left| \frac{1}{M_n} \sum_{w=1}^{M_n} x_i^2(w) - P_X \right| = 0 \tag{1}$$

  $$\lim_{n\to\infty} \left| \frac{1}{n} h(I^n + x^n(W)) - \frac{1}{2}\log(2\pi e)(P_I + P_X) \right| = 0 \tag{2}$$

  where W is assumed uniformly distributed over $\{1, \ldots, M_n\}$, $h(\cdot)$ is the differential entropy function, and $0 < P_X < P_I$.

The above assumptions are made mostly for the sake of tractability. Although the Gaussian model is not appropriate for most images of interest, it is possible (as argued in [7]) to model the components of the image as uncorrelated if whitening is performed prior to (additive) fingerprinting. The third assumption stipulates Gaussian-like features for the fingerprint set: conditions (1) and (2) are satisfied with high probability if the fingerprints are randomly generated using nM_n i.i.d. $\mathcal{N}(0, P_X)$ components. This assumption is further advocated in [1], where additive i.i.d. Gaussian fingerprints are claimed to have strong resilience to common signal processing operations (e.g., low-pass filtering), common geometric transformations and collusional attacks.

We now define the following performance metrics:

Definition 3 *The probability of error in decoding fingerprint $x^n(w)$ is given by*

$$\mathcal{P}_e(w) = \Pr\Big\{g(I^n, f(I^n + x^n(w))) \neq w\Big\}$$

Furthermore, the average probability of error for the decoder g is given by

$$\mathcal{P}_e = \frac{1}{2^{nR_W}} \sum_w \mathcal{P}_e(w)$$

and is equal to $\Pr\{W \neq \hat{W}\}$ when the fingerprint index W is uniformly distributed in $\{1, \ldots, 2^{nR_W}\}$.

Definition 4 *The average (per-symbol) quadratic distortion for fingerprint $x^n(w)$ is given by*

$$\bar{\mathcal{D}}(w) = E[n^{-1}||I^n - f(I^n + x^n(w)||^2]$$

The average quadratic distortion when the fingerprint index W is uniformly distributed in $\{1, \ldots, 2^{nR_W}\}$ is given by

$$\bar{\mathcal{D}} = E[n^{-1}||I^n - f(I^n + x^n(W)||^2] = \frac{1}{2^{nR_W}} \sum_w \bar{\mathcal{D}}(w)$$

Our objective in this paper is to compute:

1. the minimum value of R_Q such that $\bar{\mathcal{D}}$ does not exceed some value D, and

2. the maximum value of R_W such that the average probability of error $\mathcal{P}_e$ approaches zero, when the rate of the quantizer is R_Q and $\bar{\mathcal{D}}$ does not exceed D.

Note that the above definition of measured distortion takes into account the *combined* effect of fingerprinting and quantization. As we mentioned in Section 1, this is one of the distinctive differences between our model and the one considered in [3].

3 The Fingerprinting/Compression Coding Theorem

The coding theorem that establishes the bounds on R_Q, R_W consists of two parts, a direct and a converse part. Throughout, we use the notation $\hat{Y}^n = f(Y^n)$. In both parts, the following Markov conditions are used:

$$I^n, X^n \to Y^n \to \hat{Y}^n \tag{3}$$

$$Y^n \to I^n, X^n \to \hat{Y}^n \tag{4}$$

which hold because (i) $\hat{Y}^n$ is a function of Y^n; and (ii) $Y^n = I^n + X^n$ (where $X^n = x^n(W)$). Also, from (3), (4) and the data processing inequality [15], we get

$$I(I^n, X^n; \hat{Y}^n) = I(Y^n; \hat{Y}^n) \tag{5}$$

We begin by stating the converse theorem.

Theorem 1 *(Converse) For any $(2^{nR_Q}, n)$ source code and any $(2^{nR_W}, n)$ fingerprint set that satisfies conditions (1) and (2) above with $\bar{\mathcal{D}} \leq D$ and $\mathcal{P}_e \to 0$, the following must be true:*

$$\begin{aligned} R_Q &\geq r_q(D) \overset{def}{=} \frac{1}{2}\log\left(\frac{P_I^2}{(P_I + P_X)D - P_I P_X}\right) \\ R_W &\leq r_w(R_Q, D) \overset{def}{=} R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right) \end{aligned}$$

The proof of the converse is given in Section 4, and is composed of two arguments: a source coding and a channel coding argument. The source coding argument establishes the lower bound on the rate R_Q, while the channel coding argument establishes the upper bound on R_W, for any pair of source and fingerprint codes. It should be noted that R_Q approaches its lower bound $r_q(D)$ when $I^n - \hat{Y}^n$ and $\hat{Y}^n$ become approximately uncorrelated with the average per-symbol distortion between I^n and $\hat{Y}^n$ approaching D.

The forward (achievability) theorem is as follows.

Theorem 2 *(Forward) For any $\epsilon > 0$ and for any rate pair (R_Q, R_W) such that*

$$R_Q > r_q(D) \qquad \text{and} \qquad R_W < r_w(R_Q, D)$$

there exists a $(2^{nR_Q}, n)$ source code and a $(2^{nR_W}, n)$ code for the fingerprints such that (1) and (2) are satisfied, and such that for every fingerprint index w, $\bar{\mathcal{D}}(w) \leq D + \epsilon$ and $\mathcal{P}_e(w) < \epsilon$ as $n \to \infty$.

The rate region $\mathcal{R}_D$ of allowable rates (R_Q, R_W) is shown in Figure 2 (for a fixed D).

The proof of the forward theorem is given in Section 5. Briefly, a $(2^{nR_Q}, n)$ random source codebook is generated, where each codeword $\hat{Y}^n(q)$ consists of n i.i.d. $\mathcal{N}(0, P_I - d)$ components, such that $d \leq D \leq P_I$. Joint typicality encoding is used, with the proviso that distortion is measured between I^n and $\hat{Y}^n$. For the channel code, a random fingerprint code $X^n(1), \ldots, X^n(2^{nR_W})$ is generated, where each component X_i is i.i.d. Gaussian distributed with variance P_X. The decoder/authenticator, who has knowledge of I^n, uses the random fingerprint code together with the aforementioned random source code to form triplets $(I^n, X^n(k), \hat{Y}^n)$ for all possible $1 \leq k \leq 2^{nR_W}$. It declares fingerprint $\hat{W}$ present in the image if $(I^n, X^n(\hat{W}), \hat{Y}^n)$ is jointly typical with respect to a suitably chosen distribution. The probability of error is shown to vanish asymptotically as long as

$$R_W < \frac{1}{2}\log\left(\frac{P_I + P_X 2^{2R_Q}}{P_I + P_X}\right)$$

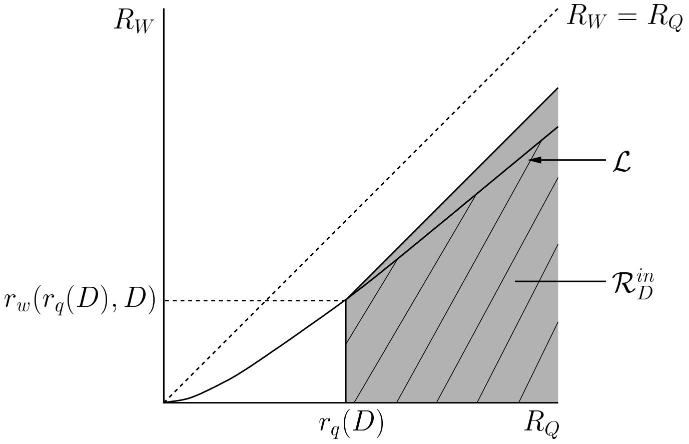

Figure 2: For a given distortion constraint D, the shaded area represents the region $\mathcal{R}_D$ of achievable pairs (R_Q, R_W). As D varies, the minimum source coding rate $r_q(D)$ and the maximum corresponding fingerprinting rate $r_w(r_q(D), D)$ parametrically define curve $\mathcal{L}$. The inner bound $\mathcal{R}_D^{in}$ is represented by the striped region.

This inequality, together with the constraint $R_Q > r_q(D)$, yields a non-convex inner bound $\mathcal{R}_D^{in}$ on the achievable rate-region $\mathcal{R}_D$ (see Figure 2). The entire region $\mathcal{R}_D$ can be achieved by timesharing. Finally, we extract and expurgate deterministic source and channel codes so that (1) and (2) are satisfied, together with $\mathcal{P}_e(w) \to 0$ and $\bar{\mathcal{D}}(w) \leq D + \epsilon$ for every fingerprint index w.

There are a number of observations that can be made with respect to Figure 2.

- The upper boundary of $\mathcal{R}_D$ is parallel to the diagonal $R_W = R_Q$. This is because $\mathcal{R}_D$ is the convex hull of $\mathcal{R}_D^{in}$, and the asymptotic slope of $\mathcal{L}$ equals unity.

- The entire region $\mathcal{R}_D$ lies below the diagonal $R_W = R_Q$. This is because for a given image I^n, all 2^{nR_W} fingerprinted copies have to be distinguishable through *different* quantization indices, i.e., $R_Q \geq R_W$.

- Setting $\hat{Y}^n$ identically equal to zero results in an average distortion equal to P_I and also makes it impossible to detect the fingerprint. Thus for $D \geq P_I$, both $r_q(D)$ and $r_w(r_q(D), D)$ equal zero, and $\mathcal{R}_D$ becomes the entire subdiagonal region $R_Q \geq R_W$.

- The other extreme (i.e., minimum) value of D is $\frac{P_I P_X}{P_I + P_X}$. This distortion is achieved when $\hat{Y}^n$ is a scaled version of Y^n, which requires infinite precision

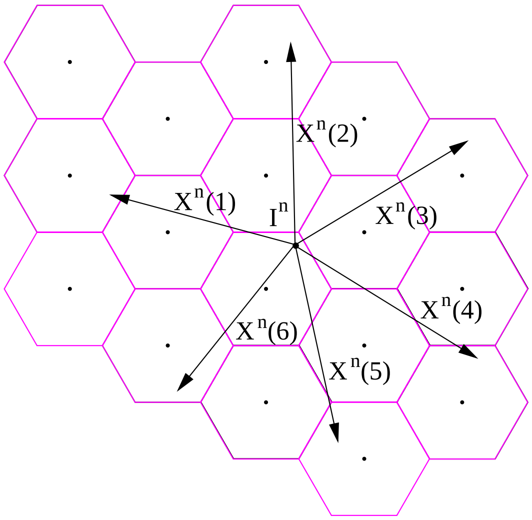

Figure 3: Example for $n = 2$ (plane). The hexagonal cells represent the encoding regions of the quantizer, and their centers are the representation vectors $\hat{Y}^n$. In order to decode the fingerprint X^n with low probability of error, it is necessary for $Y^n = I^n + X^n$ to fall into different encoding regions for different X^n with high probability.

and hence also an infinite rate $r_q(D)$. Obviously, the fingerprint can be perfectly reconstructed from I^n and $\hat{Y}^n$, thus the corresponding fingerprinting rate $r_w(r_q(D), D)$ is infinite.

- The region $\mathcal{R}_D$ grows monotonically with $D \in [P_I P_X/(P_I + P_X),\ P_I]$.

- Figure 2 depicts the rate region $\mathcal{R}_D$ obtained when *both* the distortion bound D *and* the fingerprint variance P_X are fixed. The effect of varying P_X (with D kept fixed) is to change the position of the left-hand boundary of $\mathcal{R}_D$: as P_X increases, so does the minimum quantization rate $r_q(D)$ given in Theorem 1. There is no change in the upper boundary of $\mathcal{R}_D$, since the expression for $r_w(R_Q, D)$ does not involve P_X. Thus for an application where the only constraints are upper bounds D and R on the average distortion and quantization rates, respectively, the maximum fingerprinting rate R_W would be given by

$$R - \frac{1}{2}\log\left(\frac{P_I}{D}\right) \ ,$$

and could be achieved (within ϵ) using *any* value of P_X such that

$$0 < P_X \leq \frac{P_I(P_I 2^{-2R_Q} - D)}{D - P_I} \ .$$

> The actual choice of P_X would depend on the possible attack scenarios (not studied in this paper); in general, higher values of P_X would be preferable.

Figure 3 depicts a quantizer in the case $n = 2$, where $2^{nR_W} = 6$ possible fingerprints are used. Clearly, the introduction of the fingerprints increases the distortion between the original image I^n and its representation $\hat{Y}^n$. Setting R_W above a certain limit would result in poor fingerprint detection, since versions of the image carrying different fingerprints could fall, with high probability, in the same encoding region.

4 Converse Theorem

The proof of Theorem 1 consists of two parts, the source coding part which establishes a lower bound on R_Q and the channel coding part which establishes an upper bound on R_W.

Let $\epsilon > 0$. We assume that the fingerprint index W is uniformly distributed in $\{1, \ldots, 2^{nR_W}\}$, $\mathcal{P}_e < \epsilon$, and

$$\bar{\mathcal{D}} = \frac{1}{n} \sum_{i=1}^{n} E(I_i - \hat{Y}_i)^2 \le D \ . \tag{6}$$

Let θ be the angle formed between the vectors Y^n and $\hat{Y}^n$ in L_2, where inner product is defined as the (componentwise) average correlation. In other words,

$$\cos(\theta) = \frac{n^{-1} \sum_{i=1}^{n} E(Y_i \hat{Y}_i)}{\left(n^{-1} \sum_{i=1}^{n} E(Y_i^2)\right)^{1/2} \left(n^{-1} \sum_{i=1}^{n} E(\hat{Y}_i^2)\right)^{1/2}} \tag{7}$$

Again, $X^n \overset{def}{=} x^n(W)$.

We now begin with the source coding part.

Source Coding Part: We have the usual chain of inequalities:

$$\begin{aligned} R_Q &\ge \frac{1}{n} H(\hat{Y}^n) \\ &= \frac{1}{n} H(\hat{Y}^n) - \frac{1}{n} H(\hat{Y}^n | Y^n) \\ &= \frac{1}{n} I(\hat{Y}^n; Y^n) && (8) \\ &\ge \inf_{p(\hat{y}^n | y^n)} \frac{1}{n} I(\hat{Y}^n; Y^n) && (9) \end{aligned}$$

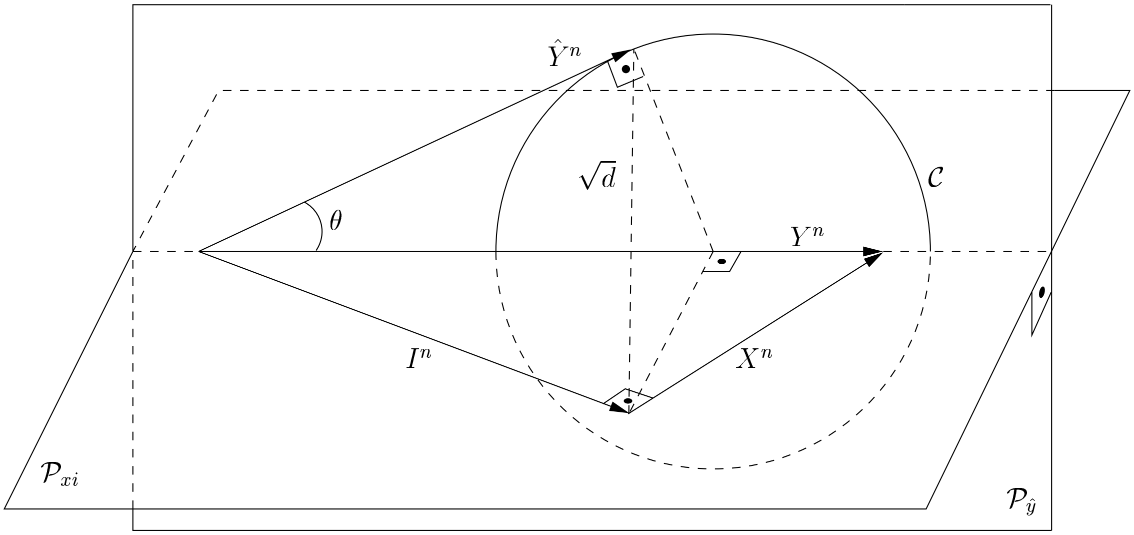

Figure 4: The 2nd moment space L_2 spanned by the vectors $(X^n, I^n, \hat{Y}^n)$.

The infimum in (9) is taken over all the distributions $p(\hat{y}^n|y^n)$ for which the conditions (1), (2), (3), (4) and (6) are satisfied. It is well known (e.g., [16, page 69]) that

$$\frac{1}{n} I(\hat{Y}^n; Y^n) \geq \frac{1}{2} \log \left(\frac{1}{\sin^2(\theta)} \right) \tag{10}$$

where θ was defined in (7).

In order to minimize (10) with respect to θ, we consider the L_2 space spanned by the vectors X^n, I^n and $\hat{Y}^n$ depicted in Figure 4. The following observations are in order:

1. Y^n lies on the plane $\mathcal{P}_{xi}$ spanned by X^n and I^n.

2. The projection of $\hat{Y}^n$ on the $\mathcal{P}_{xi}$ plane lies on Y^n, due to Markov condition (3). Equivalently, $\hat{Y}^n$ and Y^n belong to a plane $\mathcal{P}_{\hat{y}}$ which is orthogonal to $\mathcal{P}_{xi}$.

3. The circle $\mathcal{C}$ on the plane $\mathcal{P}_{\hat{y}}$ is the locus of all $\hat{Y}^n$ such that $\bar{\mathcal{D}} = d$.

We obtain a lower bound on $\frac{1}{n} I(\hat{Y}^n; Y^n)$ by minimizing (10), or equivalently, by maximizing θ. This happens when $\hat{Y}^n$ is tangent to $\mathcal{C}$ and d takes the maximum allowable value, namely D. It then follows from the geometry of Figure 4 that $\hat{Y}^n - I^n$ is orthogonal to $\hat{Y}^n$ and that $\sin(\theta) = \frac{\sqrt{D(P_I+P_X)-P_I P_X}}{P_I}$. Substituting in (10), we obtain

$$R_Q \geq \frac{1}{2} \log \left(\frac{P_I^2}{(P_I + P_X)D - P_I P_X} \right) \tag{11}$$

as required.

Channel Coding Part: Let $\epsilon > 0$ and $\mathcal{P}_e < \epsilon$. Let the rate of the quantizer be $R_Q \geq r_q(D)$ and the distortion constraint (6) be satisfied. Since the fingerprint index W is uniformly distributed, we have:

$$\begin{aligned} R_W &= \frac{1}{n}H(W) \\ &= \frac{1}{n}I(W; I^n, \hat{Y}^n) + \frac{1}{n}H(W|I^n, \hat{Y}^n) \\ &\leq \frac{1}{n}I(W; I^n, \hat{Y}^n) + \epsilon \qquad (12) \\ &\leq \frac{1}{n}I(X^n; I^n, \hat{Y}^n) + \epsilon \\ &= \frac{1}{n}I(\hat{Y}^n; Y^n) - \frac{1}{n}I(\hat{Y}^n; I^n) + \epsilon \qquad (13) \end{aligned}$$

where (12) is due to Fano's inequality [15] and (13) follows from (5). Since the rate of the quantizer is R_Q, we have from (8) that $n^{-1}I(\hat{Y}^n; Y^n) \leq R_Q$, and from (13) we obtain

$$\begin{aligned} R_W - \epsilon &\leq R_Q - \frac{1}{n}I(\hat{Y}^n; I^n) \\ &= R_Q - h(I) + \frac{1}{n}h(I^n|\hat{Y}^n) \\ &\leq R_Q - \frac{1}{2}\log\left(\frac{P_I}{n^{-1}\sum_{i=1}^n E(I_i - \hat{Y}_i)^2}\right) \\ &\leq R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right) \qquad (14) \end{aligned}$$

Taking ϵ arbitrarily small yields the desired bound on R_W and concludes the proof of the channel coding converse.

5 Forward Theorem

The proof of Theorem 2 uses a random source code $\mathcal{C}_{\hat{Y}}$ and a random channel code $\mathcal{C}_X$, generated independently of each other. The codebook for $\mathcal{C}_{\hat{Y}}$ consists of 2^{nR_Q} sequences $\tilde{Y}^n(1), \ldots, \tilde{Y}^n(2^{nR_Q})$, whose components are i.i.d. $\mathcal{N}(0, P_I - d)$; while the codebook for $\mathcal{C}_X$ consists of $M_n = 2^{nR_W}$ sequences $X^n(1), \ldots, X^n(M_n)$, whose components are i.i.d. $\sim \mathcal{N}(0, P_X)$. We first show that

$$\begin{aligned} \Pr\{|n^{-1}||I^n - \hat{Y}^n||^2 - D| < \epsilon\} &\rightarrow 1 \qquad (15) \\ \Pr\{\hat{W} = W\} &\rightarrow 1 \qquad (16) \end{aligned}$$

where the probabilities are computed with respect to the joint (product) distribution of I^n, W, $\mathcal{C}_{\hat{Y}}$ and $\mathcal{C}_X$. We then extract deterministic codes that satisfy the conditions in the statement of the theorem.

Source Coding: The fingerprinted image $Y^n = I^n + X^n(W)$ (where W is a random fingerprint index, uniformly distributed in $\{1, \ldots, 2^{nR_W}\}$) is represented by the codeword $\hat{Y}^n = f(Y^n) = \tilde{Y}^n(q)$, where q is the smallest index such that the pair $(Y^n, \tilde{Y}^n(q))$ is jointly typical with respect to a distribution $p_{Y\hat{Y}}$ defined below. If no such $q \in \{1, \ldots, 2^{nR_Q}\}$ can be found, then $\hat{Y}^n = \tilde{Y}^n(0) \overset{def}{=} 0$.

It should be emphasized that the function f used by the encoder is independent of the fingerprint set; and that the encoder only sees Y^n, and not the original image I^n. Thus the distribution $p_{Y\hat{Y}}$ used in the typicality criterion for determining $\hat{Y}^n$ must be such that the average distortion constraint between I^n and $\hat{Y}^n$ is (indirectly) met. Such $p_{Y\hat{Y}}$ can be chosen using parameters derived from the proof of the converse theorem, as demonstrated below.

Let $d \leq D$. Consider a bivariate Gaussian density $p_{Y\hat{Y}}$ having zero mean and covariance matrix

$$K_{Y\hat{Y}} = \begin{bmatrix} P_I + P_X & (P_I + P_X)(P_I - d)/P_I \\ (P_I + P_X)(P_I - d)/P_I & P_I - d \end{bmatrix} ,$$

and denote its marginals by p_Y and $p_{\hat{Y}}$. The typical set corresponding to $p_{Y\hat{Y}}$ is

$$T^n_{Y\hat{Y}}(\epsilon) = \Bigg\{ (y^n, \hat{y}^n) : \quad \begin{array}{l} \left| -\frac{1}{n} \log p_Y(y^n) - \frac{1}{2} \log(2\pi e)(P_I + P_X) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p_{\hat{Y}}(\hat{y}^n) - \frac{1}{2} \log(2\pi e)(P_I - d) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p_{Y\hat{Y}}(y^n, \hat{y}^n) - \frac{1}{2} \log(2\pi e)^2 |K_{Y\hat{Y}}| \right| < \epsilon \Bigg\} , \end{array}$$

where—with a slight abuse of notation—$p_Y(y^n)$, $p_{\hat{Y}}(\hat{y}^n)$ and $p_{Y\hat{Y}}(y^n, \hat{y}^n)$ are the n-fold i.i.d. products of $p_Y(y)$, $p_{\hat{Y}}(\hat{y})$ and $p_{Y\hat{Y}}(y, \hat{y})$, respectively.

Since each of the sequences $\tilde{Y}^n(r)$ is i.i.d. $(p_{\hat{Y}})$ and independent of Y^n, the probability that the pair $(Y^n, \tilde{Y}^n(r))$ belongs to $T^n_{Y\hat{Y}}(\epsilon)$ is lower-bounded by $2^{-n(I(\hat{Y};Y)+\epsilon)}$. Here $I(\hat{Y}; Y)$ denotes the mutual information of the bivariate distribution $p_{Y\hat{Y}}$:

$$I(\hat{Y}; Y) = \frac{1}{2} \log \left(\frac{P_I^2}{(P_I + P_X)d - P_I P_X} \right)$$

A standard argument (e.g., [15, page 356]) allows us to conclude that for

$$R_Q \geq I(\hat{Y}; Y) + 3\epsilon = \frac{1}{2} \log \left(\frac{P_I^2}{(P_I + P_X)d - P_I P_X} \right) + \epsilon \ , \tag{17}$$

there exists an index $q \geq 1$ such that $(Y^n, \tilde{Y}^n(q))$ lies in $T^n_{Y\hat{Y}}(\epsilon)$ (and thus $\hat{Y}^n = \tilde{Y}^n(q)$) with probability approaching unity.

The desired property (15) follows from a stronger result, namely that with probability approaching unity, the configuration of I^n, X^n, Y^n and $\hat{Y}^n$ in the L_2 space induced by *empirical* (not ensemble) correlations is approximately that shown in Figure 4. This result will also be of use in the channel coding argument given later.

Lemma 1 *With probability approaching unity, the triplet* $(I^n, X^n, \hat{Y}^n)$ *is typical with respect to the trivariate Gaussian distribution* $p_{IX\hat{Y}}$ *having zero mean and covariance matrix*

$$K_{IX\hat{Y}} = \begin{bmatrix} P_I & 0 & P_I - d \\ 0 & P_X & P_X(P_I - d)/P_I \\ P_I - d & P_X(P_I - d)/P_I & P_I - d \end{bmatrix}$$

Proof: Since $P_{IX\hat{Y}}$ is Gaussian, typicality is also expressed as follows: the empirical correlations obtained from $(I^n, X^n, \hat{Y}^n)$ should be within ϵ (or a factor thereof) of the corresponding entries of $K_{IX\hat{Y}}$. Since $\Pr\{(I^n, X^n) \in T^n_{IX}(\epsilon)\} \to 1$ and $\Pr\{\hat{Y}^n \in T^n_{\hat{Y}}(\epsilon)\} \to 1$, it remains to show that

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} I_i\hat{Y}_i - (P_I - d)\right| < \epsilon\right\} \to 1 \ . \tag{18}$$

and

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\hat{Y}_i - \frac{P_X(P_I - d)}{P_I}\right| < \epsilon\right\} \to 1 \tag{19}$$

The fact that $\Pr\{(Y^n, Y^n) \in T^n_{Y\hat{Y}}(\epsilon)\} \to 1$ implies that

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i + I_i)\hat{Y}_i - \frac{(P_I + P_X)(P_I - d)}{P_I}\right| < \epsilon\right\} \to 1 \ , \tag{20}$$

and thus it suffices to prove one of the two relationships, w.l.o.g. (18).

The vector I^n can be decomposed as

$$I^n = \alpha Y^n + Z^n \tag{21}$$

where Z^n is an i.i.d. Gaussian vector independent of Y^n and, by the Markov condition (3), also independent of $\hat{Y}^n$. It can be easily shown that $\alpha = \frac{P_I}{P_I + P_X}$, and that the variance of each Z_i equals $\frac{P_I P_X}{P_I + P_X}$.

From (20) and (21), we obtain

$$\Pr\left\{\left|\left(1+\frac{P_X}{P_I}\right)\frac{1}{n}\sum_{i=1}^{n} I_i\hat{Y}_i - \left(1+\frac{P_X}{P_I}\right)\frac{1}{n}\sum_{i=1}^{n} Z_i\hat{Y}_i - \frac{(P_I+P_X)(P_I-d)}{P_I}\right| < \epsilon\right\} \to 1 \tag{22}$$

It is easy to show that $\Pr\{|n^{-1}\sum_{i=1}^{n} Z_i\hat{Y}_i| < \epsilon\} \to 1$ by conditioning on the sequence $\hat{Y}^n$ and applying the weak law of large numbers to the i.i.d. sequence Z^n (which is independent of $\hat{Y}^n$). Thus (22) yields (18), as required. ■

Channel Coding: Again, the fingerprint index W is assumed to be uniformly distributed in $\{1, \dots, 2^{R_W}\}$. The rate of the random source code is set at $R_Q = r_q(d) + \epsilon$ (where $d \leq D$), which guarantees (15).

The decoder/authenticator has possession of $I^n, \hat{Y}^n$, as well as $\mathcal{C}_{\hat{Y}}$ and $\mathcal{C}_X$. To detect the fingerprint, the decoder forms all triplets $(I^n, X^n(k), \hat{Y}^n)$ for $1 \leq k \leq 2^{nR_W}$, and tests each one for typicality with respect to the trivariate distribution $p_{IX\hat{Y}}$ introduced in Lemma 1.

- If there exists a unique index j such that $(I^n, X^n(j), \hat{Y}^n) \in T^n_{IX\hat{Y}}(\epsilon)$, then the decoder outputs $\hat{W} = j$.
- Otherwise, the decoder outputs $\hat{W} = 0$, thereby declaring an error.

To compute the probability of error $\Pr\{\hat{W} \neq W\}$, we assume w.l.o.g. that $W = 1$ (since $\Pr\{\hat{W} \neq W\} = \Pr\{\hat{W} \neq w | W = w\}$ for any w). An error will occur only if one of the following events occurs.

1. $\hat{Y}^n = 0$, i.e., there exists no $q \in \{1, \dots, 2^{nR_Q}\}$ such that $(I^n + X^n(1), \hat{Y}^n(q)) \in T^n_{Y,\hat{Y}}(\epsilon)$. As this is not typical with respect to the $P_{\hat{Y}}$-marginal of $p_{IX\hat{Y}}$, the decoder declares an error.
2. There exists $q \in \{1, \dots, 2^{nR_Q}\}$ such that $(I^n + X^n(1), \hat{Y}^n(q)) \in T^n_{Y,\hat{Y}}(\epsilon)$, but the triplet $(I^n, X^n(1), \hat{Y}^n(q))$ does not belong to $T^n_{I,X,\hat{Y}}(\epsilon)$.
3. There exists q (as in Event 2) satisfying $(I^n, X^n(1), \hat{Y}^n(q)) \in T^n_{I,X,\hat{Y}}(\epsilon)$, but there also exists $k > 1$ such that $(I^n, X^n(k), \hat{Y}^n(q)) \in T^n_{I,X,\hat{Y}}(\epsilon)$.

Under Event 1, there is no good representation $\hat{Y}^n$ of $I^n + X^n(1)$. From the source coding argument, the probability of this event is asymptotically vanishing provided

$R_Q > r_q(d)$. The same is true about the probability of Event 2, by virtue of Lemma 1.

The probability of Event 3 is upper bounded as follows, assuming that $\hat{Y}^n = f(I^n + X^n(1))$.

$$\begin{aligned} & \Pr\{\exists\ w \neq 1 : (I^n, X^n(w), \hat{Y}^n) \in T^n_{I,X,\hat{Y}}(\epsilon)\} \\ \leq & \sum_{w=2}^{2^{nR_W}} \Pr\{(I^n, X^n(w), \hat{Y}^n) \in T^n_{I,X,\hat{Y}}(\epsilon)\} \\ = & \ 2^{nR_W} \Pr\{(I^n, X^n(2), \hat{Y}^n) \in T^n_{I,X,\hat{Y}}(\epsilon)\} \end{aligned} \tag{23}$$

The quantity $\Pr\{(I^n, X^n(2), \hat{Y}^n) \in T^n_{IX\hat{Y}}(\epsilon)\}$ can be upper-bounded by $2^{-n(I(X;I,\hat{Y})-\epsilon)}$, since

- $(I^n, \hat{Y}^n)$ lies in $T^n_{I\hat{Y}}(\epsilon)$; and
- by construction, $\hat{Y}^n$ depends only on I^n and $X^n(1)$, and is therefore independent of $X^n(2)$.

It can be easily shown that the mutual information $I(X; I, \hat{Y})$ equals $\frac{1}{2}\log\left(\frac{P_I d}{(P_I+P_X)d - P_I P_X}\right)$. Therefore, in order for (23) to vanish asymptotically, it suffices that

$$R_W < \frac{1}{2}\log\left(\frac{P_I d}{(P_I + P_X)d - P_I P_X}\right) - o(1) \tag{24}$$

Since $R_Q = r_q(d) + \epsilon$, we have

$$d = \frac{P_I(P_I + P_X 2^{2(R_Q-\epsilon)})}{(P_I + P_X)2^{2(R_Q-\epsilon)}}$$

and by substitution, (24) becomes

$$R_W < \frac{1}{2}\log\left(\frac{P_I + P_X 2^{2(R_Q-\epsilon)}}{P_I + P_X}\right) - o(1) \tag{25}$$

Thus (25) guarantees that $\Pr\{\hat{W} \neq W\} \to 0$.

We note here that the rate pair (R_Q, R_W) can be chosen arbitrarily close to the point $\big(r_q(D),\, r_w(r_q(D), D)\big)$ on the curve $\mathcal{L}$ which forms the upper boundary of the region $\mathcal{R}_D^{in}$ (see Figure 2). Once we establish the existence of deterministic codes with the desired properties, we will argue by time-sharing that the entire region $\mathcal{R}_D$ is achievable.

Deterministic Codes: With (R_Q, R_W) as above, consider a deterministic source codebook $\hat{y}^n(1), \ldots, \hat{y}^n(2^{nR_Q})$ and a deterministic channel codebook $x^n(1), \ldots, x^n(M_n)$ (where $M_n = 2^{nR_W}$) satisfying the following conditions:

$$\frac{1}{M_n}\sum_{w=1}^{M_n} \Pr\{|n^{-1}||I^n - f(I^n + x^n(w))||^2 - D| > \epsilon\} < \epsilon \tag{26}$$

$$\frac{1}{M_n}\sum_{w=1}^{M_n} P_e(w) < \epsilon \tag{27}$$

$$\max_{1\le i\le n}\left|\frac{1}{M_n}\sum_{w=1}^{M_n} x_i^2(w) - P_X\right| < \epsilon \tag{28}$$

$$\left|\frac{1}{n}h(I^n + x^n(W)) - h_0\right| < \epsilon \tag{29}$$

where $h_0 = \frac{1}{2}\log(2\pi e)(P_I + P_X)$. The existence of such codes is guaranteed by the fact that the corresponding random codes satisfy each of the above conditions with probability approaching unity asymptotically.

By a standard expurgation argument, a proportion δ of the watermaks can be removed from the channel code so that each remaining fingerprint $x^n(w)$ satisfies

$$\Pr\{|n^{-1}||I^n - f(I^n + x^n(w))||^2 - D| > \epsilon\} < \epsilon/\delta \tag{30}$$

and

$$\mathcal{P}_e(w) < \epsilon/\delta\ , \tag{31}$$

where the ratio ϵ/δ vanishes with ϵ (e.g., δ varies as $\sqrt{\epsilon}$). Using the fact that

$$||I^n - f(I^n + x^n(w))|| \le ||I^n||$$

and the dominated convergence theorem, it is straightforward to show that (30) implies

$$\bar{\mathcal{D}}(w) = n^{-1}E[||I^n - g(f(I^n + x^n(w))||^2] \le D + \epsilon \tag{32}$$

for every fingerprint $x^n(w)$ in the new (i.e., expurgated) channel code.

It remains to show that conditions (1) and (2) are also satisfied by the new channel code. (1) follows from (28) after a slight modification of the random channel code which does not affect the asymptotic relationships of interest. Specifically, we truncate each fingerprint component at $\pm\beta$, where $\beta = \beta(n)$ increases slowly with n. As a result, removal of δM_n fingerprints from the original deterministic code has asymptotically negligible effect on the sum in (28), which can then be renormalized to yield (1) for the new code.

To establish condition (2) for the new channel code, we consider the n-variate density of $I^n + x^n(W)$ resulting from choosing $x^n(W)$ uniformly over the following sets: (a) the original deterministic code; (b) the set of codewords that were removed from that code; and (c) the new code. Denoting these densities by $\mathbf{p}_a$, $\mathbf{p}_b$ and $\mathbf{p}_c$, respectively, we have

$$h(\mathbf{p}_a) \le -\log(1-\delta) + (1-\delta)h(\mathbf{p}_c) - \delta \int \mathbf{p}_b \log \mathbf{p}_c$$

The integral on the r.h.s. of the above inequality can be upper-bounded using explicit forms for $\mathbf{p}_b$ and $\mathbf{p}_c$ together with the power condition (1), which holds for both the original and the new code. Without going into detail, the resulting bound is of the form

$$\frac{1}{n}h(\mathbf{p}_a) \le o(1) + \frac{1}{n}h(\mathbf{p}_c)$$

Since $n^{-1}h(\mathbf{p}_a)$ is within ϵ of h_0 (by (29)) and $n^{-1}h(\mathbf{p}_c)$ can be no larger than $h_0 + \epsilon$ (i.e., the entropy of a Gaussian distribution with the same second moments), the required result follows.

Timesharing: Thus far, we have established the achievability of the region $\mathcal{R}_D^{in}$ depicted in Figure 2. To show that the entire region $\mathcal{R}_D$ is achievable for a particular distortion bound D, consider any point (R_Q, R_W) in $\mathcal{R}_D$. Since the asymptotic slope of $\mathcal{L}$ equals unity, the point in question will lie on a straight line segment joining $\Big(r_q(D)+\epsilon,\ r_w(r_q(D), D)-\epsilon\Big)$ with another point below the curve $\mathcal{L}$, corresponding to a lower distortion bound $D' < D$. If $\lambda \in (0,1)$ is the appropriate mixture coefficient, then partitioning the image I^n into blocks of λn and $(1-\lambda)n$ symbols and applying the corresponding source and fingerprint codes will yield an average distortion no larger than $\lambda D + (1-\lambda)D' + \epsilon$. It is straightforward to show that the mixture of fingerprinting codes will also satisfy conditions (1) and (2). Thus the point (R_Q, R_W) is achievable for distortion bound D.

This concludes the proof of the forward part of the coding theorem.

6 Concluding Remarks

There are a number of possible extensions to the fingerprinting model considered in this paper. One could combine our model with the one in [3], assuming that the fingerprinted images are further corrupted by attacks. Generalizations of our results for non-Gaussian images and fingerprints, non-quadratic distortion metrics, and quantizers that depend explicitly on the fingerprint used (as in [10]) would be also welcome. Another interesting model that could offer increased security, would involve a trusted authority via which agents and customers would communicate. In this case, the trusted authority combines watermarking and fingerprinting using a superposition of codewords; one for the agent who generated the original image and one for the customer who receives the watermarked/fingerprinted copy [17].

References

[1] I. Cox, J. Kilian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997.

[2] M.D.Swanson, M.Kobayashi, and A.H.Tewfik. Multimedia data-embedding and watermarking technologies. *Proceedings of the IEEE*, 86(6):1064–1087, June 1998.

[3] P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding. *submitted to the IEEE Transactions on Information Theory, preprint*, October 1999.

[4] C. Cachin. An information-theoretic model for steganography. In D. Aucsmith, editor, *Proc. of 2nd Workshop on Information Hiding*, Portland, Oregon, April 1998. Springer-Verlag.

[5] J. Kilian, F.T.Leighton, L.R.Matheson, T.G.Shamoon, R.E.Tarjan, and F. Zane. Resistance of digital watermarks to collusive attacks. In *Proc. IEEE Int. Symp. on Information Theory*, page 271, Boston, MA, August 1998. IEEE.

[6] J. O'Sullivan, P. Moulin, and J. M. Ettinger. Information theoretic analysis of steganography. In *Proc. IEEE Int. Symp. on Information Theory*, page 297, Boston, MA, August 1998.

[7] N. Merhav. On random coding error exponents of watermarking systems. *IEEE Transactions on Information Theory*, 46:420–430, March 2000.

[8] A. Cohen and A. Lapidoth. On the Gaussian watermarking game, Laboratory for Information and Decision Systems report, LIDS-P-2464, MIT. Nov 1999.

[9] A. Cohen and A. Lapidoth. On the Gaussian watermarking game. In *Proc. IEEE Int. Symp. on Information Theory*, page 48, Sorrento, Italy, June 2000.

[10] B. Chen and G. Wornell. An information-theoretic approach to the design of robust digital watermarking systems. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2061–2064, March 1999.

[11] B. Chen and G. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. In *Proc. IEEE Int. Symp. on Information Theory*, page 46, Sorrento, Italy, June 2000.

[12] B. Chen. Private communication. 2000.

[13] Y. Steinberg and N. Merhav. Identification in the presence of side information with application to watermarking. In *Proc. IEEE Int. Symp. on Information Theory*, page 45, Sorrento, Italy, June 2000.

[14] D. Karakos and A. Papamarcou. A relationship between quantization and distribution rates of digitally watermarked data. In *Proc. IEEE Int. Symp. on Information Theory*, page 47, Sorrento, Italy, June 2000.

[15] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[16] H. Dym and H. P. McKean. *Gaussian Processes, Function Theory, and the Inverse Spectral Problem.* Academic Press, 1976.

[17] D. Karakos and A. Papamarcou. Fingerprinting, watermarking and quantization of Gaussian data. *To be submitted to the IEEE Int. Symp. on Information Theory*, June 2001.