PH.D. THESIS

Digital Watermarking, Fingerprinting and Compression: An Information-Theoretic Perspective

by Damianos Karakos
Advisor: Prof. Adrian Papamarcou

PhD 2002-5

ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.

ISR is a permanent institute of the University of Maryland, within the Glenn L. Martin Institute of Technology/A. James Clark School of Engineering. It is a National Science Foundation Engineering Research Center.

Web site http://www.isr.umd.edu

ABSTRACT

Title of Dissertation: DIGITAL WATERMARKING, FINGERPRINTING AND COMPRESSION:
AN INFORMATION-THEORETIC PERSPECTIVE

Damianos Karakos, Doctor of Philosophy, 2002

Dissertation directed by: Professor Adrian Papamarcou
Department of Electrical and Computer Engineering

The ease with which digital data can be duplicated and distributed over the media and the Internet has raised many concerns about copyright infringement. In many situations, multimedia data (e.g., images, music, movies, etc) are illegally circulated, thus violating intellectual property rights. In an attempt to overcome this problem, watermarking has been suggested in the literature as the most effective means for copyright protection and authentication. Watermarking is the procedure whereby information (pertaining to owner and/or copyright) is embedded into host data, such that it is: (i) hidden, i.e., not perceptually visible; and (ii) recoverable, even after a (possibly malicious) degradation of the protected work. In this thesis, we prove some theoretical results that establish the fundamental limits of a general class of watermarking schemes.

The main focus of this thesis is the problem of joint watermarking and compression of images, which can be briefly described as follows: due to bandwidth or storage constraints, a watermarked image is distributed in quantized form, using R_Q bits per image dimension, and is subject to some additional degradation (possibly due to malicious attacks). The hidden message carries R_W bits per image dimension. Our main result is the determination of the region of allowable rates (R_Q, R_W), such that: (i) an average distortion constraint between the original and the watermarked/compressed image is satisfied, and (ii) the hidden message is detected from the degraded image with very high probability. Using notions from information theory, we prove coding theorems that establish the rate region in the following cases: (a) general i.i.d. image distributions, distortion constraints and memoryless attacks, (b) memoryless attacks combined with collusion (for fingerprinting applications), and (c) general—not necessarily stationary or ergodic—Gaussian image distributions and attacks, and average quadratic distortion constraints. Moreover, we prove a multi-user version of a result by Costa on the capacity of a Gaussian channel with known interference at the encoder.

DIGITAL WATERMARKING, FINGERPRINTING AND COMPRESSION: AN INFORMATION-THEORETIC PERSPECTIVE

by

Damianos Karakos

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2002

Advisory Committee:

Professor Adrian Papamarcou, Chairman/Advisor
Professor Benjamin Kedem
Professor Prakash Narayan
Professor Haralabos Papadopoulos
Professor Min Wu

DEDICATION

To my parents

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Professor Adrian Papamarcou, for his guidance, encouragement and support during my graduate studies.

Next, I would like to thank the other members of my committee: Prof. Prakash Narayan, Prof. Haralabos Papadopoulos, Prof. Min Wu and Prof. Benjamin Kedem, for their insightful comments and suggestions. Moreover, helpful discussions with Aaron Cohen from MIT and Prof. Neri Merhav from Technion, Israel, are gratefully acknowledged.

There are a number of other people who have helped me at various stages during my studies at UMD, and I am thankful to all of them. Of course, it is impossible to mention everyone; I apologize to those whom I omitted. Alexandros Labrinidis (who has been my roommate since 1995) has helped me on various aspects of graduate life, and has commented on parts of my thesis. Yannis Sismanis has assisted me with computer-related matters. Kaushik Chakraborty and Chunxuan Ye have offered me an intellectual environment during our informal information theory seminars at UMD. Vinay Chande and Andres Kwasinski have given me helpful comments during the last stages of my PhD. Finally, the ECE Computer Support staff (especially Julie Napp, Cliff Russell and Axel Persaud) have been very helpful by providing their timely expertise on many occasions.

Last, but not least, I would like to thank my parents, George and Anastasia Karakos, for all their encouragement and love.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

1.1 General Background

Over the last decade, principally due to the development of the Internet and the World-Wide-Web (WWW), distribution of digital multimedia data to a large population of users can be done very easily. Moreover, digital data can be duplicated very fast and without any degradation in quality—consider, for example, how common the copying of musical CDs has become in the last few years. Naturally, this situation has raised many concerns about possible violations of intellectual property rights. Unauthorized duplication and distribution of copyrighted material (photographs, music, movies, etc.), without appropriate compensation to the copyright holders, are becoming increasingly problematic. In order to fight piracy, many companies (especially in the entertainment and news industries) have devoted considerable attention to the development of information hiding (or *watermarking*) techniques. In plain terms, a watermark is a signal which is *hidden* (i.e., is not perceptually visible) inside some multimedia data, and carries information about this data (e.g., owner, title, date of creation, etc). Thus a watermark

uniquely identifies the work being protected, and helps resolve disputes about the ownership of the data.

As an example of the usefulness of watermarking, let us consider a simple scenario: Newspaper X publishes a photograph, for which it claims exclusive rights. Newspaper Y, also claiming to be the exclusive owner, publishes the same photograph after copying it from X. Without any special protection mechanism, X cannot prove that it is the rightful owner of the photograph. However, if X watermarks the photograph before publication (that is, X embeds a hidden message that identifies it as its legitimate owner), and is able to detect the watermark later in the illegally distributed copy, it will be able to supply proof of ownership in a court of law. On the other hand, to prevent detection of the watermark, Y may try to *remove* it from the picture by distorting the picture. That is, Y may attempt to *attack* the watermark so as to render it undetectable, without significantly degrading the quality of the image or affecting its commercial value. Careful design of the watermarking system can prevent this from happening.

There have been many instances of disputes or litigations on the intellectual ownership of multimedia data. A copyright violations lawsuit that received extensive publicity in the early 2000's, was that against Napster [1]. Napster was essentially a centralized database which allowed millions of users to freely distribute music files in a peer-to-peer network. The music files were unwatermarked and compressed in such a way that the quality of the reproduced music was very close to that of a Compact Disc (CD recording). However, all copyright information that normally accompanies the music written on a CD was lost. As a result, it was not an easy task for the music companies to prove that unauthorized distribution was indeed taking place through Napster. A watermarking scheme robust

to compression would have provided additional ammunition to the music industry, as the copyright information would have been inseparable from the music itself.

Due to its significance, the watermarking field has grown tremendously over the last five years. There are numerous articles (e.g., see [2, 3, 4] and the references therein) and books (e.g., [5, 6]) that explain the basics of watermarking, explore its many practical applications, and evaluate the performance of various schemes under a variety of attacks.

Two key issues in the design of watermarking schemes are:

- **Transparency:** The hidden message should not interfere perceptually with the host signal (or *covertext* [7]). This requirement is perfectly justified by the fact that watermarking aims at protecting *multimedia* data, which are sensitive, in general, to changes. In other words, an image or a musical piece could become useless if the introduced artifacts (due to watermarking) exceeded some perceptual threshold. The quality of the watermarked data must thus be comparable to that of the covertext, a requirement which is often expressed in terms of a distortion constraint.

- **Robustness:** The message must be detectable in the watermarked image (the covertext is assumed to be an image throughout this thesis, though similar techniques can be applied to other types of multimedia data), even after degradation due to malicious attacks or other processing (quantization, D/A conversion, etc). Of course, detectability of the watermark is closely related to the maximum amount of distortion that can be introduced by an attacker (for example, an attack that completely destroys an image would render the watermark detection impossible). A watermarking scheme is robust if it allows the hidden message to be accurately decoded in a distorted image whose

quality is close to that of the watermarked image (this requirement is again expressed in terms of a distortion constraint).

There are two detection scenarios: *private* and *public*. In the private detection scenario, the original image is available to the detector; in the public scenario, it is not. Although public detection schemes can be more useful in practice (since it is not always possible to have the original image available during the detection), private schemes usually offer more robustness.

One important application of information hiding is *fingerprinting* (also known as transaction tracking [6]). The fingerprint is a signal hidden inside an image, which satisfies the aforementioned transparency and robustness requirements. As opposed to a watermark, a fingerprint uniquely identifies each individual copy distributed, making it possible to trace illegally distributed data back to the user [8]. In other words, a fingerprint plays the role of a user's serial number. When both embedded into an image, watermark and fingerprint uniquely identify an *(owner, user)* pair.

Fingerprinting applications create new possibilities for attacks, mainly *collusion* attacks. In this type of attack, two or more users who possess fingerprinted copies of the same image combine their copies to produce a forged document in which the individual fingerprints maybe harder to detect (than without collusion) [9, 10, 11].

1.2 Literature Review on Information-Theoretic Aspects of Watermarking

Information hiding has also been studied from an information-theoretic perspective. Simply put, information-theoretic approaches treat watermarking as a com-

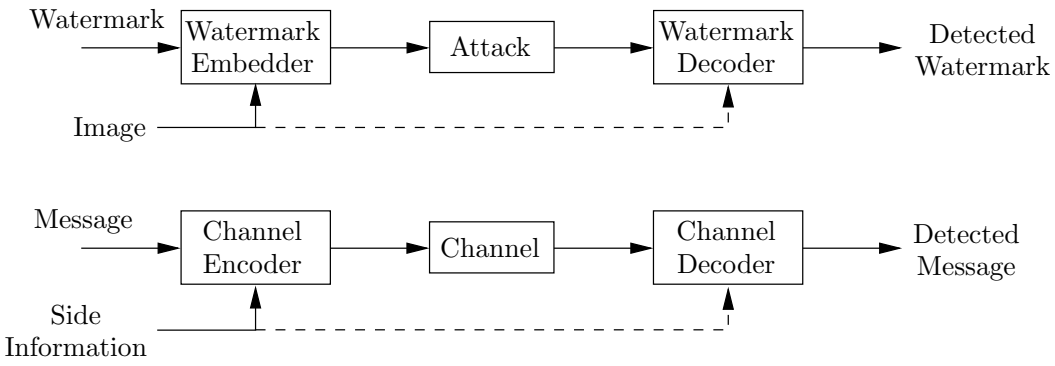


Figure 1.1: Watermarking system viewed as communication system with side information.

munication process, their main goal being to determine the maximum number of bits that can be hidden in (and reliably detected from) an image. In general, this number grows linearly with the image size, and the constant of proportionality is known as the *capacity* C. Practical watermarking implementations embed at a rate R bits/image dimension, where $R < C$.

1.2.1 Communication with Side Information

In the communication model for watermarking, the watermark embedder plays the role of the transmitter, the watermark detector plays the role of the receiver and the attack represents the communication channel (see Figure 1.1). The objective is to decode the hidden message reliably; the original image can be construed of as side information which is always available at the encoder, but is only available at the decoder in the private scenario (not the public one) [12, 4, 6].

Earlier results on communication with side information at the transmitter (such as those obtained by Gel'fand and Pinsker [13], Heegard and El Gamal [14] and Costa [15]) have been applied extensively to digital watermarking in the public

scenario. Costa's paper [15] in particular, has been instrumental in the development of many practical watermarking schemes [16, 6, 17]. Specifically, the author considers the following situation: a transmitter sends a signal X^n to a receiver (with n transmissions), through the following channel:

$$Y^n = X^n + S^n + Z^n$$

where S^n, Z^n are independent, i.i.d. Gaussian random vectors. Here, S^n is known to the encoder (but not the decoder), while Z^n acts as noise known to neither encoder or decoder. It is proved that the capacity of this channel is the same as if S^n were known to the decoder as well. Namely, $C = 1/2 \log(1 + P/N)$, where P is the average (per symbol) power of X^n and N is average power of Z^n. This result is quite surprising because one would expect S^n to act as interference at the receiver, thus hindering the detection of the message. The analogy to watermarking is obvious: in a public detection scenario, S^n would play the role of the covertext (known to the watermark embedder only), X^n would represent the watermark embedded, and Z^n would be the noise added by the attacker. Extensions of this result are developed by Yu *et al.* [18] and Cohen and Lapidoth [19].

1.2.2 Capacity of Watermarking Systems without Compression

One of the earliest information-theoretic studies of watermarking is by Moulin, O'Sullivan and Ettinger [20]. Extensions of that work are published in [21, 22, 9] by the first two authors. The main problem they consider is the following: a watermarker (or information hider) embeds a watermark into an image such that an average distortion constraint is met. An attacker modifies the watermarked

image using a memoryless attack to produce a forgery. The attacker is also limited by a distortion constraint, expressible either between the watermarked image and the forgery, or the original image and the forgery. Note that although the attacker does not have the original image in his possession, an average distortion constraint between the original image and the forgery can still be satisfied. The detector attempts to detect the watermark message from the forgery, possibly with the aid of some side information (e.g., the original image in a private scenario). Through a coding theorem, the authors compute the maximum achievable rate of the size of the watermark message set. They show that this maximum rate (i.e., capacity) is the value of a game played by the information hider and the attacker. In this game, the watermarker chooses his watermarking scheme so as to maximize the embedding rate while the attacker chooses an attack mechanism so as to minimize it. It is also assumed that the attacker knows the covert channel utilized by the watermarker, while the decoder knows the attack channel transition probabilities (more details about these assumptions will be given in Chapter 2). The authors also compute the capacity for the case where the covertext is i.i.d. Gaussian distributed, and show that the capacity remains the same in both the public and the private scenarios.

Another interesting information-theoretic treatment of digital watermarking is due to Cohen and Lapidoth [23, 7, 24]. In their work, they determine the capacity of a watermarking system, similar to the one considered by Moulin and O'Sullivan above, but with the following differences: (a) the attacker is not constrained to perform a memoryless attack; (b) the distortion constraints on the watermarker and the attacker are of the almost sure, rather than average, type; (c) neither encoder or decoder have any knowledge about the attack channel other than the

aforementioned distortion constraints; and (d) the attacker knows the details of the watermark embedder except for a secret key shared between encoder and decoder. In order to prove their result, the authors determine the value of the "mutual information game", which is the same as the value of the memoryless attack game (with average quadratic distortion constraints) that is studied in [9]. Moreover, the authors prove that, under average quadratic distortion constraints, there exists a (non-memoryless) attack that results in zero capacity.

A similar formulation for a vector Gaussian watermarking game is proposed by Cohen and Lapidoth in [7] and studied in detail in [25].

The capacity of a private watermarking scheme is studied by Somekh-Baruch and Merhav in [26, 27]. Here, the distortion constraint is expressed as follows: the probability that the distortion induced (by the watermarker or the attacker) exceeds a given threshold is upper bounded by an exponentially decaying function. This requirement (termed "large deviations distortion constraint") is more general than the almost sure constraint considered by Cohen [24]. In addition to determining the capacity, the authors investigate the exponent of the probability of decoding error; similar studies are conducted in [28]. Finally, the capacity of a public watermarking scheme is investigated in [29].

Steinberg and Merhav [30] study the identification capacity of a watermarking system, subject to a memoryless attack channel. For that model, they derive the maximum number of watermarks that can be embedded into an image, assuming that the decoder is interested only in determining (with vanishing probability of error) whether a particular watermark in present in the watermarked image or not.

1.2.3 Joint Watermarking and Compression

The problem of joint watermarking and compression of images—which is the main focus of this thesis—has received less attention in the literature. This problem can be formally described as follows: due to bandwidth or storage constraints, a watermarked image is quantized to R_Q bits per image dimension, corresponding to a source codebook index. The information is then delivered to the customer, who is assumed to have access to the source codebook. The compression scheme complies with the aforementioned transparency and robustness requirements, in that a distortion (fidelity) constraint is met, and the watermark is detectable from the reproduced (quantized), and possibly degraded, version of the image.

In [16], Chen and Wornell develop an interesting watermarking/compression scheme termed *Quantization Index Modulation* (QIM), where an ensemble of quantizers—each corresponding to a particular watermark index—is used for compressing the image. The authors provide an information-theoretic analysis of QIM based on Costa's work [15], and developed a practical scheme which makes use of dithered quantizers. Also, they distinguish between two versions of QIM: *regular* and *distortion-compensated* QIM. In regular QIM, the watermarked image is communicated to the user as an index in a source codebook, while in distortion-compensated QIM, the output of the encoder is a linear combination of the cover-text and the output of a quantizer. Regular QIM is of relevance to our work and will be studied further in Chapter 2.

The main goal of this thesis is the determination of regions of allowable rates (R_Q, R_W), where R_W is the rate of the watermark index set, under certain transparency and robustness requirements. In [31, 8], Karakos and Papamarcou examine the case where the watermarked/compressed image is not subject to attacks (com-

pression inherently introduces degradation, but cannot be construed as a malicious attack of the type studied in, e.g., [9, 24]). They show that, when the original image is i.i.d. Gaussian and an average quadratic distortion constraint is imposed, the region of allowable rates (R_Q, R_W) (for the no-attack case) is given by

$$\begin{aligned} R_Q &\geq \frac{1}{2}\log\left(\frac{P_I}{D}\right) \\ R_W &\leq R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right) \end{aligned}$$

where P_I is the image variance (per dimension or pixel) and D is the average quadratic distortion between the original image and the watermarked/compressed image. In this thesis, we extend this result to the case where the quantized, watermarked image is subject to malicious attacks, as well as to the case where fingerprinted images are distributed to different customers and are subject to collusion attacks. More details about these extensions can be found in the next section, where an overview of the thesis is presented.

1.3 Organization of the Thesis

The thesis is organized as follows. In Chapter 2 we establish the rate region of achievable rate pairs (R_Q, R_W) under memoryless attacks such that (i) an average quadratic constraint between the original image and the watermarked, compressed image is satisfied, and (ii) the probability of correct detection of the watermark from a distorted image approaches unity as the number of image dimensions approaches infinity. The following cases are considered:

- The memoryless attack is chosen independently of the embedding strategy and is known to both encoder and decoder. Results are obtained for two

statistical models: the general discrete alphabet case for arbitrary image distributions and distortion constraints, as well as the Gaussian case where the original image and the attack channel are i.i.d. Gaussian, and the distortion metric is quadratic.

- The information hider and the attacker play a game (similar to [9]). Specifically, the attacker knows the encoding function used by the watermarker, while the decoder knows the attack distribution.

Moreover, we give achievability results for the rate region of regular QIM (public scenario), as well as the rate region of additive watermarking schemes [31, 8, 32].

In Chapter 3 we establish the region of achievable rates for fingerprinting systems under collusion attacks. Similar statistical models (discrete and Gaussian) are considered here. The formulation is different than the one in Chapter 2, in that we allow a number of users to collude by producing a forgery which depends on more than one fingerprinted version of the same image. We demonstrate that collusion can be very effective in reducing the size of the rate region. We conclude Chapter 3 with a multi-user version of Costa's paper; we prove that in the public version of a fingerprinting system without compression, the maximum rate achievable under Gaussian symmetric collusion attacks is the same as in a private scenario.

In Chapter 4 we extend the theory derived by Yu *et al.* in [18] for non-stationary Gaussian models, to the case where watermarked images are quantized before distribution. We obtain a general rate region formula, and we give examples in which the general formula yields simpler expressions. Finally, conclusions and directions for further research are given in Chapter 5.

1.4 Information-Theoretic Contributions

In this section, we briefly summarize the most important contributions of this thesis, from an information-theoretic viewpoint.

As mentioned in the previous section, the main problem treated in this thesis is that of joint quantization and watermarking. Using terminology from communication theory, this problem can be stated as follows: a transmitter wishes to convey two kinds of information through a channel: a quantized form of a random vector (i.e., the image), and a message (the watermark index) taken from a particular set of messages. Each information is intended for two different "receivers", respectively: the watermark index has to be decoded only by one receiver (i.e., the watermark decoder) and not the other (i.e., the customer), while the quantized information is decodable by both. Moreover, the customer receives a noise-free copy of the quantized image, while the watermark decoder has access to side information (the original image), which is not available to the customer.

A naive way of performing the encoding would be to concatenate the two bitstreams that describe the two entities (image and watermark) and send the result through the channel. However, this particular encoding is inefficient in terms of protection, because an attacker can simply discard the watermark message without affecting the image quality. Instead, our goal is to embed the watermark information *inside* the compressed image representation. This is accomplished by designing appropriate source codewords, which can be used for conveying information through an attack channel. In other words, we study to what extent a source codebook can be used as a channel codebook, and we derive the relationship between the rates of these two codebooks. This relationship reveals two interesting (and quite surprising) properties: (i) at low quantization rates, Gaussian attack

noise does not affect the detectability of the watermark, and the only limitation imposed on the watermarking rate is due to the quantization itself; and (ii) at high (but finite) quantization rates the source code can achieve the watermarking rate of an *optimal* channel code.

The relationship between quantization and watermarking rates is derived in Chapter 2 for the single-watermark (or, single user) case. Chapter 3 is an extension of the result of Chapter 2 to the multi-user case, in which the decoder decodes reliably more than one (fingerprint) messages. The result derived in this case is relevant to the expression for the rate region of a multi-access channel [33]. Moreover, we prove that a Gaussian multi-access channel with side information available *only* at the transmitter has the same capacity (assuming that the rates of all users are the same), as if the side information were known at the receiver as well. This result extends Costa's single-user result. Finally, it is interesting that the analysis of Chapter 2 can be extended to the case of non-stationary and non-ergodic Gaussian images and attacks, as explained in Chapter 4.

1.5 Notation

The following symbols and conventions will be used consistently throughout the thesis.

Capital letters are used to denote random quantities, while small letters denote deterministic quantities (or realizations of the respective random variables). Also, a random variable can be a deterministic function of a random quantity (e.g., $\hat{Y} = \hat{y}(Q)$). All variables take values in sets that are represented by the corresponding script letters, e.g., $X \in \mathcal{X}$. A sequence of n variables $X_1, \ldots, X_n$ is denoted by X^n and belongs to the Cartesian product $\mathcal{X}^n$.

We denote the original image (covertext) by I^n. The integer n can be interpreted in many ways; it could represent the number of pixels of the image or the number of its most significant DCT coefficients. The exact meaning of n pertains to the particular implementation of the watermarking algorithm. For the purposes of this thesis, we assume that n represents the number of image values that are altered by means of the watermarking algorithm. Moreover, we assume that the watermark index W is uniformly distributed on the set $\{1, \ldots, 2^{nR_W}\}$, where R_W is the *watermarking rate*. The watermark decoder, which attempts to detect W from a watermarked, compressed and possibly distorted image, outputs its estimate $\hat{W}$ of W. Finally, a source codebook (quantizer) consists of 2^{nR_Q} elements $\{\hat{y}^n(1), \ldots, \hat{y}^n(2^{nR_Q})\}$, where R_Q is the *quantization rate*.

Chapter 2

Relationship between Quantization and Watermarking Rates in the Presence of Memoryless Attacks

In this chapter, we establish the region of achievable rates (R_Q, R_W) under memoryless attacks. We assume that a distortion constraint on the watermarker (and, possibly, on the attacker) is met and the probability of correct watermark detection goes to unity when n goes to infinity. The chapter is organized as follows: in Section 2.1 we give a summary of the results; proofs of the theorems can be found in Sections 2.2, 2.3, 2.4 and 2.5. Finally, we conclude the chapter with Section 2.6, where we discuss achievable rate regions of various schemes. The results of this chapter are extensions of results published in [31, 8, 34, 32].

2.1 Summary of Results

This section summarizes the results for: (i) the case in which the memoryless attack is fixed (in terms of the conditional probability distribution of the attack channel), for both discrete and continuous alphabets; and (ii) the case where the attacker

chooses its distribution with respect to the distribution of $\hat{Y}^n$ given I^n used by the watermarker, and with respect to some distortion constraint on the attack. In the latter case, we assume that the attacker knows the information hiding strategy utilized by the watermarker, and tries to minimize the achievable rate region by appropriate choice of the attack (while the watermarker tries to maximize the same region). Thus, watermarker and attacker play a game, in which the watermarker plays first (by choosing the information-hiding strategy) and the attacker plays next by choosing its attack strategy with respect to the strategy chosen by the watermarker. Note that the watermark encoder knows $p_{Z^n|\hat{Y}^n} = (p_{Z|\hat{Y}})^n$ only in case (i), while the watermark decoder knows (or can reliably estimate) $p_{Z|\hat{Y}}$ in both cases (i) and (ii).

2.1.1 Fixed Attack Channel

We first present results for discrete alphabets, and then for continuous alphabets with Gaussian distributions.

Discrete Alphabets

The general form of the watermarking/quantization system under consideration is shown in Figure 2.1. The watermark index W is uniformly distributed over a set of size 2^{nR_W}; I^n is the n-dimensional i.i.d. image, with distribution $p_{I^n}(i^n) = \prod_{j=1}^{n} p_I(i_j)$; and $\hat{Y}^n$ is the watermarked/quantized image which can be found in a source codebook of size 2^{nR_Q}. The attack channel is memoryless; its conditional probability distribution is given by

$$p_{Z^n|\hat{Y}^n}(z^n|\hat{y}^n) = \prod_{j=1}^{n} p_{Z|\hat{Y}}(z_j|\hat{y}_j)$$

We assume $p_{Z^n|\hat{Y}^n}$ to be known to both encoder and decoder.

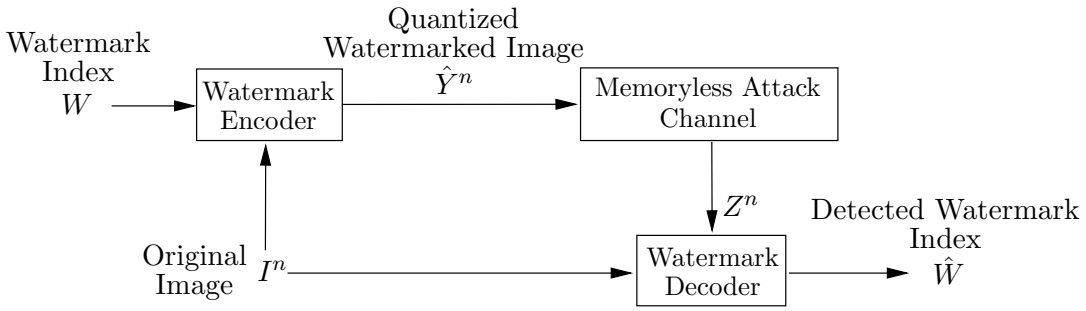


Figure 2.1: The general watermarking/quantization system with memoryless attacks.

The watermark decoder outputs $\hat{W}$, its estimate of W. We consider a private scenario here, so we assume that I^n is known at the decoder. The transparency and robustness requirements are expressed via the following constraints:

$$\textbf{Transparency:} \quad n^{-1}Ed(I^n, \hat{Y}^n) = n^{-1}\sum_{j=1}^{n} Ed(I_j, \hat{Y}_j) \leq D \tag{2.1}$$

$$\textbf{Robustness:} \quad \Pr\{\hat{W} \neq W\} \to 0 \text{ as } n \to \infty \tag{2.2}$$

where $d : \mathcal{I} \times \hat{\mathcal{Y}} \to \mathbf{R}^+$ is a given distortion function and D is a non-negative number.

Formally stated, we have the following definition of a private quantization/watermarking code.

Definition 2.1 *A $(2^{nR_Q}, 2^{nR_W}, n)$ private quantization/watermarking code consists of the following:*

- *A watermark set $\mathcal{M}_n = \{1, \ldots, 2^{nR_W}\}$.*
- *An encoding function $f : \mathcal{M}_n \times \mathcal{I}^n \to \hat{\mathcal{Y}}^n$ which maps a watermark index w and an image sequence i^n to a representation sequence $\hat{y}^n$ taken from the set $\{\hat{y}^n(1), \ldots, \hat{y}^n(2^{nR_Q})\}$.*

- *A decoding function* $g : \mathcal{Z}^n \times \mathcal{I}^n \rightarrow \mathcal{M}_n$ *which maps the output of the channel* z^n *and the original image* i^n *to an estimate* $\hat{w}$ *of* w.

For random W and I^n, we have the random quantities $\hat{Y}^n = f(W, I^n)$ and $\hat{W} = g(Z^n, I^n)$. A definition of a public quantization/watermarking code would be similar to the above, except that the decoder g would take as input only z^n.
We now state the following definitions:

Definition 2.2 *The probability of error in detecting watermark* w *is given by*

$$\mathcal{P}_e(w) = \Pr\{g(Z^n, I^n) \neq w | \hat{Y}^n = f(w, I^n)\}$$

Furthermore, the average probability of error for 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 watermark index* W *is uniformly distributed in* $\{1, \ldots, 2^{nR_W}\}$.

Definition 2.3 *For a* $(2^{nR_Q}, 2^{nR_W}, n)$ *quantization/watermarking code, the average (per-symbol) distortion is given by*

$$\bar{\mathcal{D}} = E[n^{-1} \sum_{j=1}^{n} d(I_j, f(W, I^n)_j)]$$

assuming that W *is uniformly distributed in* $\{1, \ldots, 2^{nR_W}\}$.

Definition 2.4 *A rate pair* (R_Q, R_W) *is achievable for distortion constraint* D, *if there exists a sequence of quantization/watermarking codes* $(2^{nR_Q}, 2^{nR_W}, n)$ *such that* $\max_w \mathcal{P}_e(w)$ *tends to 0 as* $n \rightarrow \infty$ *and* $\bar{\mathcal{D}} \leq D$. *Moreover, a rate region* $\mathcal{R}$ *of pairs* (R_Q, R_W) *is achievable if every element of* $\mathcal{R}$ *is achievable.*

Definition 2.5 *For a quantization/watermarking system operating at quantization rate R_Q, the watermarking capacity $C(R_Q)$ is defined as the supremum of all rates R_W such that (R_Q, R_W) is achievable.*

Our first result is stated as follows:

Theorem 2.1 *A private quantization/watermarking code $(2^{nR_Q}, 2^{nR_W}, n)$ satisfies the transparency and robustness requirements (2.1) and (2.2) respectively, if and only if $(R_Q, R_W) \in \mathcal{R}_D^{\text{dsc}}$, where*

$$\mathcal{R}_D^{\text{dsc}} = \left\{ (R_Q, R_W) : \right.$$
$$R_Q \geq \min_{p_{\hat{Y}|I}: Ed(I,\hat{Y}) \leq D} I(\hat{Y}; I)$$
$$\left. R_W \leq \max_{p_{\hat{Y}|I}: Ed(I,\hat{Y}) \leq D} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \right\}$$

Here,

$$\Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \stackrel{\Delta}{=} \min\{R_Q - I(\hat{Y}; I),\ I(Z; \hat{Y}|I)\} \tag{2.3}$$

and the mutual information quantities on the right-hand side of (2.3) are computed with respect to $p_I,\ p_{\hat{Y}|I}$ and $p_{Z|\hat{Y}}$. Theorem 2.1 holds for arbitrary discrete alphabets $\mathcal{I}^n, \hat{\mathcal{Y}}^n, \mathcal{Z}^n$. The superscript "dsc" in $\mathcal{R}_D^{\text{dsc}}$ is used to distinguish this rate region from the one obtained in the case of continuous alphabets. The proof of Theorem 2.1 can be found in Section 2.2.

Continuous Alphabets, Gaussian Distributions

A variant of Theorem 2.1 can be obtained in the case of continuous alphabets (all equal to $\mathbf{R}$). We consider the system shown in Figure 2.2. Here, I^n is i.i.d. Gaussian with variance P_I; the memoryless attack is described by the expression $Z^n = \beta_A \hat{Y}^n + V^n$ where β_A is a real scalar and V^n is i.i.d. Gaussian with variance

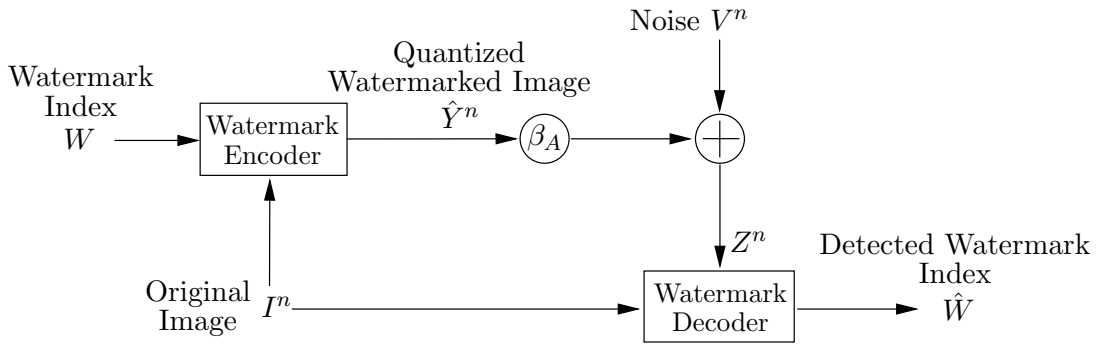


Figure 2.2: The watermarking/quantization system with Gaussian attacks combined with scaling.

P_V; and the distortion function d is the squared difference:

$$d(x, y) = (x - y)^2 \tag{2.4}$$

The distortion constraint (2.1) then becomes

$$n^{-1} E||I^n - \hat{Y}^n||^2 \leq D \tag{2.5}$$

The rate region $\mathcal{R}_D^{\text{gauss}}$ of achievable rates in the continuous case is then established by the following theorem.

Theorem 2.2 *A private, continuous alphabet quantization/watermarking code* $(2^{nR_Q}, 2^{nR_W}, n)$ *satisfies requirements (2.2) and (2.5), if and only if* $(R_Q, R_W) \in \mathcal{R}_D^{\text{gauss}}$*, where*

$$\mathcal{R}_D^{\text{gauss}} = \Bigg\{ (R_Q, R_W) : \\ R_Q \geq \left[\frac{1}{2} \log \left(\frac{P_I}{D} \right) \right]^+ \\ R_W \leq \max_{\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q} \right]} \min \left\{ R_Q - \frac{1}{2} \log(\gamma), \frac{1}{2} \log \left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V} \right) \right\} \Bigg\}$$

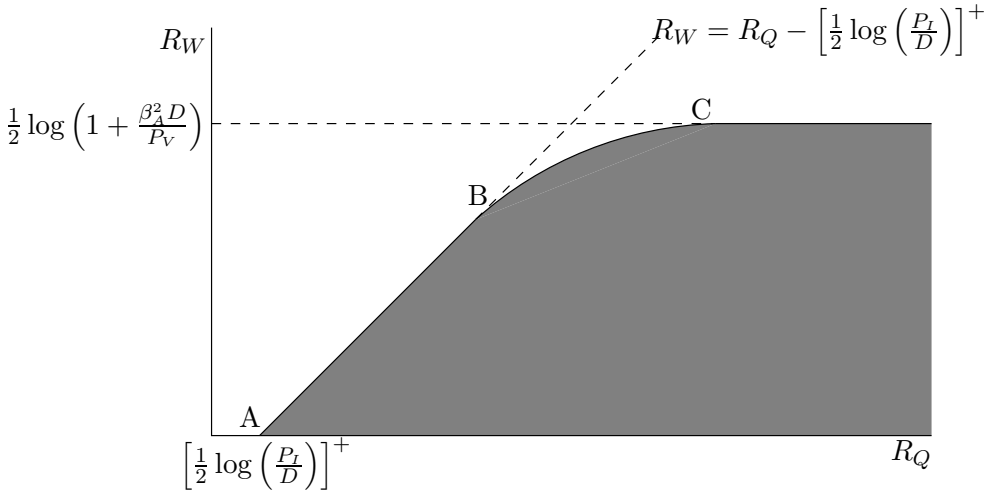


Figure 2.3: The rate region $\mathcal{R}_D^{\text{gauss}}$ of achievable rate pairs (R_Q, R_W).

Here,

$$P_W(\gamma) \triangleq \frac{\gamma(P_I + D) - 2P_I + 2\sqrt{P_I(\gamma D - P_I)(\gamma - 1)}}{\gamma^2} \tag{2.6}$$

and $[\cdot]^+ \triangleq \max\{\cdot, 0\}$. The proof of the Theorem 2.2 can be found in Section 2.3. $\mathcal{R}_D^{\text{gauss}}$ is the shaded region in Figure 2.3. Its upper boundary is composed of:

- The segment AB on the straight line $R_W = R_Q - \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+$.
- The curved segment BC defined by the equation

  $$R_W = \max_{\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]} \min\left\{R_Q - \frac{1}{2}\log(\gamma), \frac{1}{2}\log\left(1 + \frac{\beta_A^2 P_W(\gamma)}{D_A}\right)\right\}$$

  for R_Q in the interval $\left[\frac{1}{2}\log\left(\max\left\{1, \frac{P_I}{D}\right\} + \frac{\beta_A^2 |P_I - D|}{P_V}\right), \frac{1}{2}\log\left(1 + \frac{P_I}{D} + \frac{\beta_A^2 (P_I + D)}{P_V}\right)\right]$, i.e., the projection of BC on the R_Q-axis.
- The half-line $\mathcal{C}_\infty$ which is parallel to the R_Q-axis and has vertex C. The R_W-ordinate on $\mathcal{C}_\infty$ is given by $\frac{1}{2}\log\left(1 + \frac{\beta_A^2 D}{P_V}\right)$.

Two key conclusions can be drawn from Figure 2.3:

- For rates $R_Q \in \left[\left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+,\ \frac{1}{2}\log\left(\max\left\{1, \frac{P_I}{D}\right\} + \frac{\beta_A^2|P_I - D|}{P_V}\right)\right]$, the watermarking rate R_W can be as high as $R_Q - \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+$, which is the maximum watermarking rate for the case of no attack (D_A=0). In other words, at low quantization rates, Gaussian attack noise does not degrade the performance of the system.

- When $R_Q \geq \frac{1}{2}\log\left(1 + \frac{P_I}{D} + \frac{\beta_A^2(P_I+D)}{P_V}\right)$, the maximum watermarking rate is constant and equal to $\frac{1}{2}\log\left(1 + \frac{\beta_A^2 D}{P_V}\right)$. This expression makes sense in the case $R_Q = \infty$, where the distortion in the original image is solely due to watermarking, and where $\beta_A^2 D$ represents the "signal" power in the AWGN Gaussian attack channel of variance P_V—hence the familiar expression for the capacity of that channel. It is surprising that in the case $R_Q < \infty$, there exists a quantization rate threshold above which quantization does not hinder the detection of the watermark, i.e., the watermarking rate can be as high as in the case of no compression.

2.1.2 The Watermarker vs. Attacker Game

Here, we assume that the attacker knows (or can estimate reliably) the joint distribution of $I^n, \hat{Y}^n$ (a similar assumption was made in [9, 22]). Thus, depending on the encoding algorithm and the rate R_Q chosen by the watermarker, the attacker chooses his conditional distribution function $p_{Z|\hat{Y}}$ (we use single letters here because we assume memoryless attacks), so as to minimize the maximum achievable rate $C(R_Q)$, subject to a fidelity criterion. This fidelity criterion can be expressed in the form of a distortion constraint between the watermarked image $\hat{Y}^n$ and the forgery Z^n. Similarly to the distortion constraint on the watermarker, we assume

that the distortion constraint on the attacker is given by

$$n^{-1}\sum_{j=1}^{n} Ed_A(\hat{Y}_j, Z_j) \leq D_A \tag{2.7}$$

for some non-negative D_A.

On the other hand, the watermarker (who designs the encoder/decoder pair) has to ensure that the detection of the watermark is reliable for *any* attack that the attacker is allowed to use. So, watermarker and attacker play a game, in which the former tries to maximize the achievable rate region while the latter tries to minimize it. Note that we call a rate region *maximized*, when $C(R_Q)$ is maximum for all R_Q (analogously for minimum). The "rules" of this game are expressed as follows:

- The encoder (watermark embedder) is designed without any knowledge of the attack conditional distribution $p_{Z|\hat{Y}}$. Therefore, in terms of the game evolution, the watermarker plays first, and designs an encoding function f (according to Definition 2.1) that remains fixed for the rest of the game. This function f induces a fixed conditional probability $p_{\hat{Y}^n|I^n}$. Moreover, the rate pair (R_Q, R_W) of the quantization/watermarking code is chosen without any knowledge of the attack and remains fixed. The watermarker must ensure that the rates are chosen such that the watermark is detected reliably for *any* attack distribution $p_{Z|\hat{Y}}$ chosen by the attacker.

- The attacker, who knows $p_{\hat{Y}^n|I^n}$, plays second and chooses a conditional distribution $p_{Z|\hat{Y}}$ for the attack such that the distortion constraint (2.7) is met.

- The watermarker plays next, and designs his watermark decoding algorithm g with respect to the distribution $p_{Z|\hat{Y}}$ chosen by the attacker.

Given the above game formulation, we now define an achievable rate pair (R_Q, R_W) as follows:

Definition 2.6 *A rate pair (R_Q, R_W) is achievable for distortion constraint D if there exists a sequence of quantization/watermarking codes $(2^{nR_Q}, 2^{nR_W}, n)$ such that for any admissible attack distribution $p_{Z|\hat{Y}}$, the maximal error probability $\max_w \mathcal{P}_e(w)$ tends to 0 as $n \to \infty$ and $\bar{\mathcal{D}} \leq D$. Moreover, a rate region $\mathcal{R}$ of pairs (R_Q, R_W) is achievable if every element of $\mathcal{R}$ is achievable.*

For the rest of our results in this sub-section, we assume that Definition 2.6 is in effect when we refer to achievable rates.

We now define the following sets of distributions, which we use in the sequel:

$$\mathcal{M}(p_I, D) \stackrel{\Delta}{=} \{p_{\hat{Y}^n|I^n} : n^{-1} \sum_{j=1}^{n} Ed(I_j, \hat{Y}_j) \leq D\} \tag{2.8}$$

$$\mathcal{M}^{ml}(p_I, D) \stackrel{\Delta}{=} \{p_{\hat{Y}|I} : Ed(I, \hat{Y}) \leq D\} \tag{2.9}$$

which contains all the *memoryless* distributions that belong to $\mathcal{M}(p_I, D)$. Also,

$$\mathcal{M}_A(p_I, f, D_A) \stackrel{\Delta}{=} \{p_{Z|\hat{Y}} : n^{-1} \sum_{j=1}^{n} Ed_A(f(W, I^n)_j, Z_j) \leq D_A\} \tag{2.10}$$

where $d_A : \hat{\mathcal{Y}} \times \mathcal{Z} \to \mathbf{R}^+$ is a distortion function. Observe that the set $\mathcal{M}_A(p_I, f, D_A)$ depends on f only through the induced conditional probability $p_{\hat{Y}^n|I^n} = p_{f(W,I^n)|I^n}$. Thus, instead of writing the attacker's set of admissible distributions as $\mathcal{M}_A(p_I, f, D_A)$, we use the notation

$$\mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A) \stackrel{\Delta}{=} \{p_{Z|\hat{Y}} : n^{-1} \sum_{j=1}^{n} Ed_A(\hat{Y}_j, Z_j) \leq D_A\} \tag{2.11}$$

Note that by the Markov condition

$$I^n \to \hat{Y}^n \to Z^n \tag{2.12}$$

we have $p_{Z^n,\hat{Y}^n,I^n} = (p_{Z|\hat{Y}})^n \, p_{\hat{Y}^n|I^n} \, (p_I)^n$.

The admissible set of distributions is defined by

$$\mathcal{A}(p_I, D, D_A) = \{(p_{\hat{Y}^n|I^n}, p_{Z|\hat{Y}}) : p_{\hat{Y}^n|I^n} \in \mathcal{M}(p_I, D),\ p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)\}$$

and is, in general, non-rectangular.

We now state the following theorem (for discrete alphabets):

Theorem 2.3 *Assume that, for all $n \geq 1$, the attacker knows $p_{\hat{Y}^n|I^n}$, the watermark decoder knows $p_{Z|\hat{Y}}$, and the attack distortion constraint (2.7) is satisfied. Then, a rate pair (R_Q, R_W) is achievable (i.e., it satisfies Definition 2.6) if and only if it belongs to the set*

$$\mathcal{R}^{\text{dsc}}_{D,D_A} = \Bigg\{ (R_Q, R_W) : \\ R_Q \geq \min_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I, D)} I(\hat{Y}; I) \\ R_W \leq \max_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I, D)} \ \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \Bigg\}$$

The proof of Theorem 2.3 can be found in Section 2.4.

Our final result in this section is the determination of the region of achievable rate pairs (R_Q, R_W) when all alphabets are continuous (and equal to $\mathbf{R}$), I^n is i.i.d. Gaussian with variance P_I, and when both distortion functions $d(\cdot,\cdot), d_A(\cdot,\cdot)$ are equal to the squared-error distortion function (2.4). In this case, Theorem 2.3 becomes:

Theorem 2.4 *Under the same assumptions as in Theorem 2.3, a rate pair*

(R_Q, R_W) is achievable if and only if it belongs to the set

$$\mathcal{R}_{D,D_A}^{\text{gauss}} = \Bigg\{ (R_Q, R_W) : \\ R_Q \geq \left[\frac{1}{2} \log \left(\frac{P_I}{D} \right) \right]^+ \\ R_W \leq \max_{\gamma \in \Gamma(R_Q, D, D_A)} \min \left\{ R_Q - \frac{1}{2} \log(\gamma),\ \frac{1}{2} \log \left(1 + \frac{P_W(\gamma)}{D_A} - \frac{1}{\gamma} \right) \right\} \Bigg\}$$

Moreover, if

$$\Gamma(R_Q, D, D_A) \stackrel{\Delta}{=} \left[\max\{1, P_I/D\},\ 2^{2R_Q} \right] \cap \{\gamma : \gamma P_W(\gamma) > D_A\} \tag{2.13}$$

is empty, then no positive R_W can be achieved for that particular R_Q.

The proof of Theorem 2.4 can be found in Section 2.5.

2.1.3 Other Schemes

Finally, Section 2.6 contains achievable rate regions of other schemes; namely, the public scenario of the regular QIM, as well the private scenario of additive schemes [8, 32] in the presence of additive Gaussian noise. In both of these cases, the achievable regions are subsets of the region derived in Theorem 2.2 (for $\beta_A = 1$ and $P_V = D_A$).

2.2 Proof of Theorem 2.1

The coding theorem which establishes the region $\mathcal{R}_D^{\text{dsc}}$, consists of a converse and a direct (achievability) part.

Converse Theorem

The converse theorem states that any $(2^{nR_Q}, 2^{nR_W}, n)$ code which satisfies constraints (2.1) and (2.2) must satisfy $(R_Q, R_W) \in \mathcal{R}_D^{\text{dsc}}$.

Proof: Let $\epsilon > 0$. We assume that the watermark index W is uniformly distributed in $\{1, \ldots, 2^{nR_W}\}$, that $\Pr\{W \neq \hat{W}\} < \epsilon$, and that the distortion constraint is met with equality:

$$\frac{1}{n}\sum_{j=1}^{n} Ed(I_j, \hat{Y}_j) = D \tag{2.14}$$

By virtue of the monotonicity of the region $\mathcal{R}_D^{\text{dsc}}$ in D, the constraint can then be relaxed to an inequality, as in (2.1).

A standard converse rate-distortion theorem (e.g., [33]) yields

$$R_Q \geq \frac{1}{n} I(I^n; \hat{Y}^n) \geq r_q(D) \tag{2.15}$$

where $r_q(D)$ is the rate-distortion function for the memoryless source I^n, with distortion D. Specifically,

$$r_q(D) = \min_{p_{\hat{Y}|I}: Ed(I,\hat{Y}) \leq D} I(I; \hat{Y}) \tag{2.16}$$

This establishes the lower bound on R_Q in the definition of $\mathcal{R}_D^{\text{dsc}}$.

The upper bound on R_W is obtained using two chains of inequalities. The first chain is as follows:

$$\begin{aligned}
R_W &= n^{-1}H(W) \\
&= n^{-1}H(W|I^n) && (2.17)\\
&= n^{-1}I(W; \hat{Y}^n|I^n) + n^{-1}H(W|\hat{Y}^n, I^n) && (2.18)\\
&\leq n^{-1}I(W; \hat{Y}^n|I^n) + n^{-1}H(W|Z^n, I^n) && (2.19)\\
&\leq n^{-1}I(W; \hat{Y}^n|I^n) + \epsilon && (2.20)\\
&= n^{-1}H(\hat{Y}^n|I^n) - n^{-1}H(\hat{Y}^n|I^n, W) + \epsilon \\
&= n^{-1}H(\hat{Y}^n|I^n) + \epsilon && (2.21)\\
&= n^{-1}H(\hat{Y}^n) - n^{-1}(H(\hat{Y}^n) - H(\hat{Y}^n|I^n)) + \epsilon
\end{aligned}$$

$$
\begin{aligned}
&\leq R_Q - n^{-1} I(\hat{Y}^n; I^n) + \epsilon && (2.22)\\
&= R_Q - H(I) + n^{-1} H(I^n|\hat{Y}^n) + \epsilon \\
&\leq R_Q - H(I) + n^{-1} \sum_{j=1}^{n} H(I_j|\hat{Y}_j) + \epsilon && (2.23)\\
&= R_Q - n^{-1} \sum_{j=1}^{n} I(I_j; \hat{Y}_j) + \epsilon && (2.24)
\end{aligned}
$$

where (2.17) holds because I^n is independent of W; (2.19) follows from the data processing inequality applied to the Markov chain $W \rightarrow (\hat{Y}^n, I^n) \rightarrow (Z^n, I^n)$; (2.20) is a consequence of Fano's inequality; (2.21) holds because $\hat{Y}^n$ is a deterministic function of I^n, W, (2.22) follows from $R_Q \geq n^{-1} H(\hat{Y}^n)$ and (2.23) is due to the inequalities $H(I^n|\hat{Y}^n) \leq \sum_{i=1}^{n} H(I_j|\hat{Y}^n) \leq \sum_{j=1}^{n} H(I_j|\hat{Y}_j)$.

The second chain of inequalities is as follows:

$$
\begin{aligned}
R_W &= n^{-1} H(W|I^n) && (2.25)\\
&= n^{-1} I(W; Z^n|I^n) + n^{-1} H(W|I^n, Z^n) \\
&\leq n^{-1} I(W; Z^n|I^n) + \epsilon && (2.26)\\
&= n^{-1} H(Z^n|I^n) - n^{-1} H(Z^n|I^n, W) + \epsilon \\
&= n^{-1} H(Z^n|I^n) - n^{-1} H(Z^n|\hat{Y}^n) + \epsilon && (2.27)\\
&\leq n^{-1} \sum_{j=1}^{n} H(Z_j|I_j) - n^{-1} \sum_{j=1}^{n} H(Z_j|Z_1, \ldots, Z_{j-1}, \hat{Y}^n) + \epsilon && (2.28)\\
&= n^{-1} \sum_{j=1}^{n} H(Z_j|I_j) - n^{-1} \sum_{j=1}^{n} H(Z_j|\hat{Y}_j) + \epsilon && (2.29)\\
&\leq n^{-1} \sum_{j=1}^{n} (H(Z_j|I_j) - H(Z_j|\hat{Y}_j, I_j) + \epsilon && (2.30)\\
&= n^{-1} \sum_{j=1}^{n} I(Z_j; \hat{Y}_j|I_j) + \epsilon && (2.31)
\end{aligned}
$$

where (2.25) is due to the independence of I^n and W; (2.26) follows from Fano's inequality; (2.27) holds because of the Markov chains $(I^n, W) \rightarrow \hat{Y}^n \rightarrow Z^n$ and

$\hat{Y}^n \to (I^n, W) \to Z^n$; (2.28) follows from the chain rule for the entropy; (2.29) holds because the attack channel is memoryless and therefore given $\hat{Y}_j$, Z_j is conditionally independent of everything else and (2.30) follows because conditioning reduces entropy.

Combining (2.24) and (2.31) we have:

$$R_W \leq \min \left\{ R_Q - n^{-1} \sum_{j=1}^{n} I(I_j; \hat{Y}_j), n^{-1} \sum_{j=1}^{n} I(Z_j; \hat{Y}_j | I_j) \right\} + \epsilon \tag{2.32}$$

We now observe that (2.32) depends only on the pmf's $p_{Z_j, \hat{Y}_j, I_j}(z, \hat{y}, i) = p_{Z|\hat{Y}}(z|\hat{y})\, p_{\hat{Y}_j|I_j}(\hat{y}|i)\, p_I(i)$, for each time instant $j = 1, \ldots, n$ ($p_{Z|\hat{Y}}$ and p_I do not depend on j due to memorylessness). Furthermore, $R_Q - I(I_j; \hat{Y}_j)$ and $I(Z_j; \hat{Y}_j | I_j)$ are both concave with respect to $p_{\hat{Y}_j|I_j}$. Hence, applying Jensen's inequality, we obtain:

$$n^{-1} \sum_{j=1}^{n} (R_Q - I(I_j; \hat{Y}_j)) \leq R_Q - I_a(I; \hat{Y}) \tag{2.33}$$

$$n^{-1} \sum_{j=1}^{n} I(Z_j; \hat{Y}_j | I_j) \leq I_a(Z; \hat{Y} | I) \tag{2.34}$$

where, for the computation of the mutual information in the right-hand side of (2.33) and (2.34), the "averaged" pmf

$$p^a_{\hat{Y}|I}(\hat{y}|i) \stackrel{\Delta}{=} n^{-1} \sum_{j=1}^{n} p_{\hat{Y}_j|I_j}(\hat{y}|i) \tag{2.35}$$

was used. It is easy to establish that $p^a_{\hat{Y}|I}$ satisfies the one-dimensional distortion constraint

$$Ed(I, \hat{Y}) = D \tag{2.36}$$

Then, combining (2.33), (2.34) and (2.36), from (2.32) we obtain

$$R_W \;\; \leq \;\; \Xi(R_Q,\; p_I,\; p^a_{\hat{Y}|I},\; p_{Z|\hat{Y}}) + \epsilon$$

$$
\begin{aligned}
&\leq \max_{\left\{ \begin{array}{c} p_{\hat{Y}|I}\colon \exists\ p_{\hat{Y}_1|I_1}\cdots p_{\hat{Y}_n|I_n}\ \text{s.t.} \\ p_{\hat{Y}|I}=n^{-1}\sum_{j=1}^{n} p_{\hat{Y}_j|I_j} \\ n^{-1}\sum_{j=1}^{n} Ed(I_j,\hat{Y}_j)=D \end{array} \right\}} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) + \epsilon \\
&\leq \max_{p_{\hat{Y}|I}:Ed(I,\hat{Y})=D} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) + \epsilon \qquad (2.37) \\
&\leq \max_{p_{\hat{Y}|I}:Ed(I,\hat{Y})\leq D} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) + \epsilon
\end{aligned}
$$

where inequality (2.37) is due to (2.36). By letting $\epsilon \to 0$, we conclude the proof of the converse. ■

Direct Theorem

We now show that $\mathcal{R}_D^{\text{dsc}}$ is achievable.

Proof: As required for $\mathcal{R}_D^{\text{dsc}}$, we limit the quantization rate to $R_Q \geq r_q(D)$ (where $r_q(D)$ was defined in (2.16)).

We use a random coding argument, where the watermark index W is assumed uniformly distributed in $\{1,\ldots,2^{nR_W}\}$. The technique is similar to the private version of regular QIM [16], in that 2^{nR_W} quantizers, each one indexed by a different watermark, are employed.

Codebook Generation: A set of 2^{nR_Q} i.i.d. sequences $\tilde{Y}^n$, is generated, such that each dimension is distributed according to some pmf $p_{\hat{Y}}$. The set is then partitioned into 2^{nR_W} subsets of 2^{nR_1} sequences each, i.e.,

$$R_Q = R_W + R_1$$

The w^{th} subset, consisting of sequences $\tilde{Y}^n(w,1),\ldots,\tilde{Y}^n(w,2^{nR_1})$, becomes the codebook for the w^{th} watermark.

Watermark Embedding: Given I^n and a deterministic w, the embedder identifies within the w^{th} codebook the first codeword $\tilde{Y}^n(w,q)$ such that the pair

$(I^n, \tilde{Y}^n(w,q))$ lies in the set $T_{I,\hat{Y}}(\epsilon)$ of typical pairs with respect to a bivariate $p_{I,\hat{Y}}$, such that the distortion constraint (2.1) is satisfied. The output of the embedder (encoder) is denoted by $\hat{Y}^n(w) = \tilde{Y}^n(w,q)$. If none of the codewords in the w^{th} codebook is jointly typical with I^n, then the embedder outputs $\hat{Y}^n(w) = 0$. In this manner, 2^{nR_W} watermarked versions of the image I^n are obtained: $\hat{Y}^n(1), \ldots, \hat{Y}^n(2^{nR_W})$. Clearly, for random W, the embedder output is $\hat{Y}^n(W)$.

Decoding: Again, the decoder has access to the original image I^n. Upon receiving Z^n, the decoder seeks among all watermarked versions $\hat{Y}^n(1), \ldots, \hat{Y}^n(2^{nR_W})$ of I^n a single $\hat{Y}^n(\hat{w})$ such that the triplet $(I^n, \hat{Y}^n(\hat{w}), Z^n)$ lies in $T^n_{I,\hat{Y},Z}(\epsilon)$, the set of typical triplets with respect to the trivariate distribution $p_{I,\hat{Y},Z}$, such that $p_{I,\hat{Y},Z} = p_{Z|\hat{Y}} p_{\hat{Y},I}$. If a unique such sequence $\hat{Y}^n(\hat{w})$ exists, then the decoder outputs $\hat{W} = \hat{w}$; otherwise, the decoder declares an error.

Error Events: Without loss of generality, we assume $W = 1$. We then have the following error events:

- E_1: $\hat{Y}^n(1) = 0$, i.e., there exists no $q \in \{1, \ldots, 2^{nR_1}\}$ such that $(I^n, \tilde{Y}^n(1,q)) \in T_{I,\hat{Y}}$.
- E_2: There exists a $\tilde{Y}^n(1,q) = \hat{Y}^n(1)$ such that $(I^n, \hat{Y}^n(1)) \in T_{I,\hat{Y}}$, but $(I^n, \hat{Y}^n(1), Z^n) \notin T_{I,\hat{Y},Z}$.
- E_3: $(I^n, \hat{Y}^n(1), Z^n) \in T_{I,\hat{Y},Z}$ but there also exists a $k > 1$ such that $(I^n, \hat{Y}^n(k), Z^n) \in T_{I,\hat{Y},Z}$.

The probability of error is then

$$\Pr\{\hat{W} \neq 1\} = \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$$

Behavior of $Pr(E_1)$: From standard rate-distortion theorems [33], we know that if $R_1 = R_Q - R_W > I(I;\hat{Y})$ (the mutual information of the bivariate $p_{I,\hat{Y}}$ defined above), then $\Pr(E_1) \to 0$ as $n \to \infty$. Equivalently, if

$$R_W \leq R_Q - I(I;\hat{Y}) - \epsilon \tag{2.38}$$

then $\Pr(E_1) \to 0$ as $n \to \infty$.

Behavior of $Pr(E_2)$: To show that $\Pr(E_2) \to 0$, it suffices to show that the triplet $(I^n, \hat{Y}^n(1), Z^n)$ lies in $T_{I,\hat{Y},Z}$ with probability approaching unity asymptotically. In the previous paragraph, we showed that $\Pr\{(I^n, \hat{Y}^n(1)) \in T_{I,\hat{Y}}\} \to 1$. Since Z^n is the output of a memoryless channel with conditional probability distribution $p_{Z|\hat{Y}}$ (that was used for the generation of the typical set $T_{I,\hat{Y},Z}$), it can be easily verified that Z^n is typical with $I^n, \hat{Y}^n(1)$ as well, with probability that approaches unity.

Behavior of $Pr(E_3)$:

$$\begin{aligned}
\Pr(E_3) &= \Pr\{\exists\, w \neq 1 : (I^n, \hat{Y}^n(w), Z^n) \in T_{I,\hat{Y},Z}\} \\
&\leq \sum_{w=2}^{2^{nR_W}} \Pr\{(I^n, \hat{Y}^n(w), Z^n) \in T_{I,\hat{Y},Z}\} \\
&= (2^{nR_W} - 1)\Pr\{(I^n, \hat{Y}^n(2), Z^n) \in T_{I,\hat{Y},Z}\}
\end{aligned} \tag{2.39}$$

where the last equality is due to the symmetry of the random code statistics. Since

$$\Pr\{(I^n, \hat{Y}^n(2)) \in T_{I,\hat{Y}}\} \to 1$$

and by construction, Z^n is independent of $\hat{Y}^n(2)$ given I^n, a standard argument (cf. the proof of Theorem 8.6.1 in [33]) yields

$$\Pr\{(I^n, \hat{Y}^n(2), Z^n) \in T_{I,\hat{Y},Z}\} \leq 2^{-n(I(Z;\hat{Y}|I)-(\epsilon/2))}$$

where the conditional mutual information is computed with respect to the trivariate $p_{I,\hat{Y},Z}$ defined earlier. Hence, if

$$R_W \leq I(Z;\hat{Y}|I) - \epsilon \tag{2.40}$$

it follows that the upper bound on $\Pr(E_3)$ in (2.39) vanishes asymptotically.

Thus, combining (2.38) and (2.40) and letting $\epsilon \to 0$, we obtain the achievable rate

$$R_W \;\leq\; \Xi(R_Q,\; p_I,\; p_{\hat{Y}|I},\; p_{Z|\hat{Y}}) \tag{2.41}$$

Then by maximizing (2.41) with respect to $p_{\hat{Y}|I}$ such that the distortion constraint (2.1) is met, we obtain the required result.

We have thus proved that if $(R_Q, R_W) \in \mathcal{R}_D^{\text{dsc}}$, then the average probability of error, over the ensemble of random codes, vanishes asymptotically with n. By a standard argument, there exists a deterministic code that achieves $\mathcal{R}_D^{\text{dsc}}$ with arbitrarily small probability of error (averaged over all the messages); and the codebook can be then expurgated to make the maximal probability of error arbitrarily small.

■

2.3 Proof of Theorem 2.2

As in the discrete case, here too we have a direct and a converse part.

Converse Theorem

The converse part states that if constraints (2.2) and (2.5) are satisfied by some $(2^{nR_Q}, 2^{nR_W}, n)$ code, then the rates (R_Q, R_W) must lie in $\mathcal{R}_D^{\text{gauss}}$, as defined in the statement of the theorem.

Proof: Let $\epsilon > 0$. We assume that the watermark index W is uniformly distributed in $\{1, \ldots, 2^{nR_W}\}$, that $\Pr\{W \neq \hat{W}\} < \epsilon$, and that the distortion constraint is met with equality (similarly to the discrete case):

$$\frac{1}{n}\sum_{j=1}^{n} E(I_j - \hat{Y}_j)^2 = D \tag{2.42}$$

Since I^n is i.i.d. Gaussian, a standard converse rate-distortion theorem (e.g., [33]) yields

$$R_Q \geq \frac{1}{n} I(I^n; \hat{Y}^n) \geq \left[\frac{1}{2} \log \left(\frac{P_I}{D} \right) \right]^+ \tag{2.43}$$

This establishes the lower bound on R_Q in the definition of $\mathcal{R}_D^{\text{gauss}}$.

The derivation of the upper bound on R_W can be simplified by considering the L_2-space spanned by vectors I^n and $\hat{Y}^n$, with inner product defined by

$$< U^n, V^n > \stackrel{\Delta}{=} \frac{1}{n} \sum_{j=1}^{n} E[U_j V_j].$$

for any random vectors U^n, V^n. The geometry of this space is shown in Figure 2.4, where the circle $\mathcal{C}$ has radius $\sqrt{D}$ corresponding to the distortion constraint (2.42). The lengths of I^n and $\hat{Y}^n$ are given by $\sqrt{P_I}$ and $\sqrt{P_{\hat{Y}}}$, respectively, where $P_{\hat{Y}} \stackrel{\Delta}{=} n^{-1} \sum_{j=1}^{n} E(\hat{Y}_j^2)$; while the angle between the two vectors is denoted by ϕ. As can be seen from Figure 2.4, when $P_I \geq D$, the maximum $\phi_{\max}$ of ϕ is obtained when $\hat{Y}^n$ is tangent to $\mathcal{C}$, in which case

$$\sin^2(\phi_{\max}) = \frac{D}{P_I}$$

Otherwise, if $P_I < D$, ϕ can take any value in $[0, \pi]$ (due to symmetry, we ignore negative angles). Note that for every ϕ (except when $P_I \geq D$ and $\phi = \phi_{\max}$), the line on which $\hat{Y}^n$ lies, meets $\mathcal{C}$ at two points. Equivalently, for every $\gamma \stackrel{\Delta}{=} \sin^{-2}(\phi)$ with $\gamma \geq \max\{1, P_I/D\}$, there are two positions of $\hat{Y}^n$ on $\mathcal{C}$ (the position farther from the origin is shown in every case in Figure 2.4).

Let $\lambda_0 I^n$ be the projection of $\hat{Y}^n$ on I^n, or equivalently, the MMSE estimator of $\hat{Y}^n$ among all scalar multiples on I^n:

$$\lambda_0 \stackrel{\Delta}{=} \arg \min_{\lambda \in \mathbf{R}} \frac{1}{n} \sum_{i=1}^{n} E(\hat{Y}_i - \lambda I_i)^2 \tag{2.44}$$

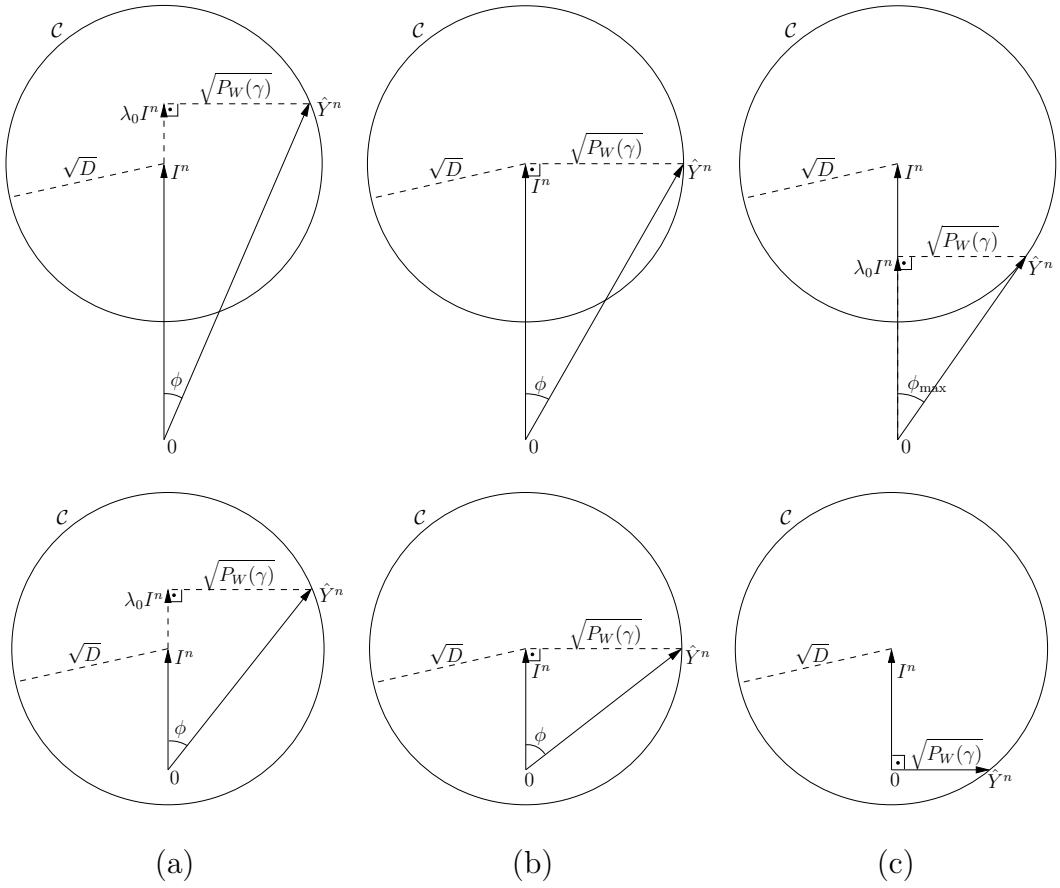


Figure 2.4: The 2nd moment space L_2 spanned by vectors I^n and $\hat{Y}^n$, shown for three different values of ϕ. The top figures correspond to the case $P_I \geq D$, while the bottom figures correspond to $P_I < D$. The circle $\mathcal{C}$ is the locus of all $\hat{Y}^n$ such that $n^{-1}E||I^n - \hat{Y}^n||^2 = D$. As ϕ increases from 0, $P_W(\gamma)$ increases monotonically (case (a)) until it reaches its maximum value D (case (b)), then decreases monotonically until $\phi = \phi_{\max}$ (case (c)). We do not consider the case $\phi > \pi/2$ when $P_I < D$, since it gives the same value for γ and $P_W(\gamma)$ as the angle $\pi - \phi$.

Let $P_{\hat{Y}|I}$ denote the resulting MMSE error, and note that

$$\sin^2(\phi) = \frac{P_{\hat{Y}|I}}{P_{\hat{Y}}}$$

From the geometry of Figure 2.4, using the Pythagorean theorem, it easily follows that

$$P_{\hat{Y}|I}(\gamma - 1) = \frac{(P_I + P_{\hat{Y}} - D)^2}{4P_I} \tag{2.45}$$

$P_{\hat{Y}}$ $(= \gamma P_{\hat{Y}|I})$ can then be eliminated to yield a quadratic equation for $P_{\hat{Y}|I}$ in terms of P_I, D and γ, with roots

$$P_{\hat{Y}|I} = \frac{\gamma(P_I + D) - 2P_I \pm 2\sqrt{P_I(\gamma D - P_I)(\gamma - 1)}}{\gamma^2} \tag{2.46}$$

Consistent with our earlier observation, there are two possible values of $P_{\hat{Y}|I}$ for every ϕ ($< \phi_{\max}$ if $P_I \geq D$) or equivalently, for every $\gamma \geq \max\{1, P_I/D\}$. The larger value is precisely $P_W(\gamma)$, as defined in (2.6).

The mutual information between I^n and $\hat{Y}^n$ is also related to the geometry of Figure 2.4. Specifically, if $\mu_0 \hat{Y}^n$ is the projection of I^n onto $\hat{Y}^n$, we have

$$\begin{aligned} I(I^n; \hat{Y}^n) &= h(I^n) - h(I^n|\hat{Y}^n) \\ &= h(I^n) - h(I^n - \mu_0 \hat{Y}^n|\hat{Y}^n) \\ &\geq h(I^n) - h(I^n - \mu_0 \hat{Y}^n) \end{aligned}$$

The differential entropy of $I^n - \mu_0 \hat{Y}^n$ is upper-bounded by that of an i.i.d. Gaussian vector having components of the same variance. By concavity of the logarithm, we then have

$$h(I^n - \mu_0 \hat{Y}^n) \leq \frac{n}{2} \log \left(\frac{2\pi e}{n} \sum_{j=1}^{n} E(I_j - \mu_0 \hat{Y}_j)^2 \right) = \frac{n}{2} \log(2\pi e P_{I|\hat{Y}})$$

where $P_{I|\hat{Y}}$ is defined similarly to $P_{\hat{Y}|I}$. Therefore

$$\frac{1}{n} I(I^n; \hat{Y}^n) \geq \frac{1}{2} \log \left(\frac{P_I}{P_{I|\hat{Y}}} \right) ,$$

and since

$$\frac{P_{I|\hat{Y}}}{P_I} = \frac{P_{\hat{Y}|I}}{P_{\hat{Y}}} = \sin^2(\phi) \ ,$$

we conclude that

$$\frac{1}{n} I(I^n; \hat{Y}^n) \geq \frac{1}{2} \log \left(\frac{1}{\sin^2(\phi)} \right) = \frac{1}{2} \log(\gamma) \tag{2.47}$$

As in the discrete case, the upper bound on R_W is obtained using two parallel chains of inequalities. The first chain is identical to inequalities (2.17)-(2.22). Thus, from (2.22) we obtain

$$R_W \leq R_Q - n^{-1} I(\hat{Y}^n; I^n) + \epsilon \tag{2.48}$$

and together with (2.47), we finally obtain (for any value of γ corresponding to the geometry of I^n and $\hat{Y}^n$)

$$R_W \leq R_Q - \frac{1}{2} \log(\gamma) + \epsilon \tag{2.49}$$

The second chain of inequalities is as follows (where λ_0 was defined in (2.44)):

$$\begin{aligned}
R_W &= n^{-1} H(W|I^n) && (2.50)\\
&= n^{-1} I(W; Z^n|I^n) + n^{-1} H(W|I^n, Z^n) \\
&\leq n^{-1} I(W; Z^n|I^n) + \epsilon && (2.51)\\
&= n^{-1} h(Z^n|I^n) - n^{-1} h(Z^n|I^n, W) + \epsilon \\
&= n^{-1} h(Z^n|I^n) - n^{-1} h(V^n|I^n, W) + \epsilon && (2.52)\\
&= n^{-1} h(\beta_A \hat{Y}^n - \beta_A \lambda_0 I^n + V^n|I^n) - n^{-1} h(V^n) + \epsilon && (2.53)\\
&\leq n^{-1} h(\beta_A(\hat{Y}^n - \lambda_0 I^n) + V^n) - \frac{1}{2} \log(2\pi e) P_V + \epsilon \\
&\leq \frac{1}{2} \log(2\pi e)(\beta_A^2 P_{\hat{Y}|I} + P_V) - \frac{1}{2} \log(2\pi e) P_V + \epsilon && (2.54)\\
&= \frac{1}{2} \log \left(1 + \frac{\beta_A^2 P_{\hat{Y}|I}}{P_V} \right) + \epsilon && (2.55)
\end{aligned}$$

where (2.50) holds because I^n is independent of W; (2.51) follows from Fano's inequality; (2.52) holds because $\hat{Y}^n$ is a function of I^n and W; (2.53) follows from the independence of V^n and (I^n, W); and (2.54) holds by the usual Gaussian bound on differential entropy.

From (2.55) and (2.46) we obtain

$$R_W \leq \frac{1}{2}\log\left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V}\right) + \epsilon$$

An upper bound on the range of γ can be deduced from (2.43), (2.47), resulting in

$$\max\left\{1, \frac{P_I}{D}\right\} \leq \gamma \leq 2^{2R_Q} \tag{2.56}$$

Thus

$$R_W \leq \max_{\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V}\right)\right\} + \epsilon \tag{2.57}$$

and taking $\epsilon \to 0$, we obtain the upper bound on R_W in the definition of $\mathcal{R}_D^{\text{gauss}}$. ■

The Upper Boundary of the Rate Region

Before proceeding to the proof of the direct theorem, it is instructive to examine the behavior of the upper bound on R_W as a function of R_Q:

$$r_W(R_Q) \triangleq \max_{\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V}\right)\right\} \tag{2.58}$$

Note that since R_Q is variable, the range of interest for γ is $[\max\{1, P_I/D\}, \infty)$.

The second argument of $\min\{\cdot,\cdot\}$ in (2.58) is independent of R_Q and monotone in $P_W(\gamma)$. From the proof of the converse theorem above, we know that $\sqrt{P_W(\gamma)}$ is the length of the error vector $\hat{Y}^n - \lambda_0 I^n$ when I^n and Y^n are as shown in Figure 2.4, with $\sin^{-2}(\phi) = \gamma$. Clearly, $\sqrt{P_W(\gamma)}$ increases monotonically as ϕ increases from $\phi = 0$ to $\phi = \arctan(\sqrt{D/P_I}) = \arcsin(\sqrt{D/(P_I + D)})$; then decreases

monotonically as ϕ increases to $\phi_{\max} = \arcsin(\min\{1, \sqrt{D/P_I}\})$. Equivalently (but in the reverse direction), as γ increases from $\gamma = \max\left\{1, \frac{P_I}{D}\right\}$ to $\gamma = 1 + \frac{P_I}{D}$ and then on to infinity, $P_W(\gamma)$ increases from $|P_I - D| \min\left\{1, \frac{D}{P_I}\right\}$ to D (its maximum value), and then decreases to 0. The function $\frac{1}{2}\log(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V})$ has similar behavior, and is plotted in Figure 2.5 against $\frac{1}{2}\log(\gamma)$. The initial (leftmost) and maximum R_W-ordinates on the curve are $\frac{1}{2}\log(1 + \frac{|P_I - D|}{P_V}\min\{1, \frac{D}{P_I}\})$ and $\frac{1}{2}\log(1 + \frac{D}{P_V})$, respectively.

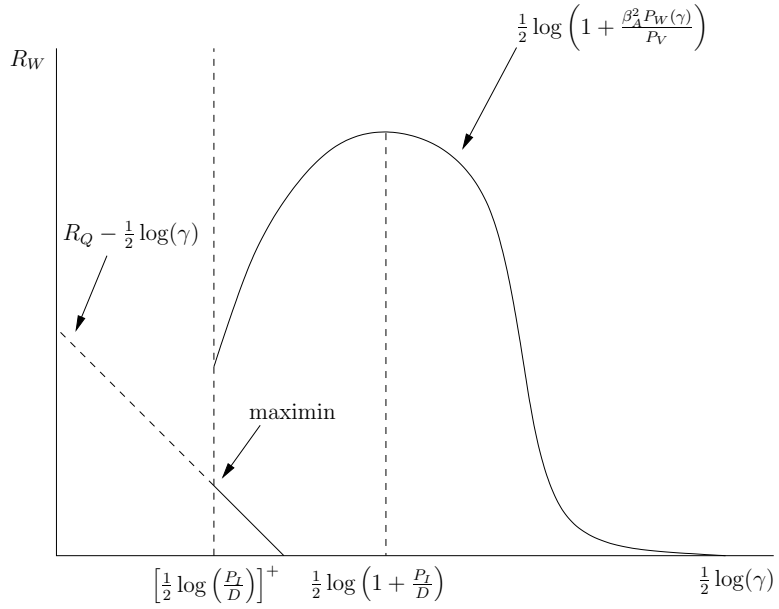
The first argument of $\min\{\cdot, \cdot\}$ in (2.58) involves R_Q and decreases monotonically from $R_Q - \left[\frac{1}{2}\log(\frac{P_I}{D})\right]^+$ to zero as γ ranges over the interval $[\max\{1, \frac{P_I}{D}\}, 2^{2R_Q}]$ of maximization in (2.58). Plotted against $\frac{1}{2}\log(\gamma)$, it yields a line segment of slope -1 (in Figure 2.5), whose position on the graph depends on the value of R_Q.

The behavior of $r_W(R_Q)$ as R_Q varies can be examined with the aid of Figure 2.5. There are three regimes of interest:

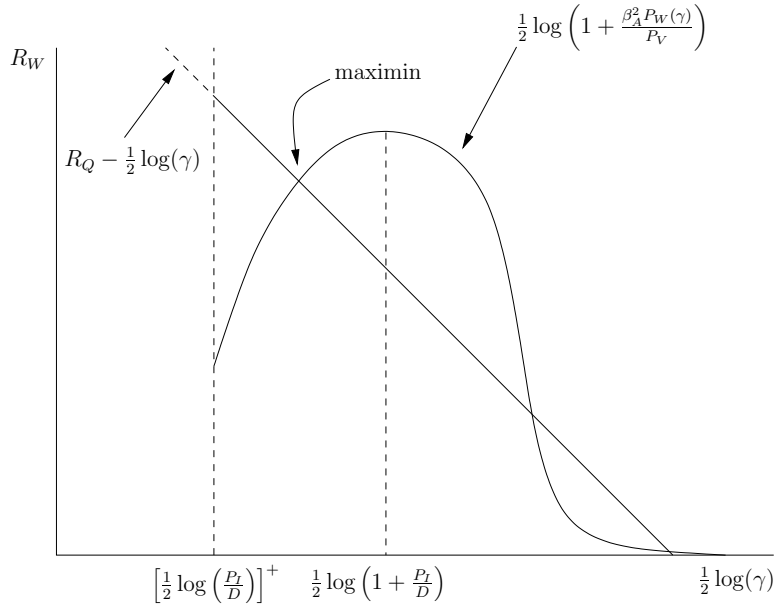
(a) In the first regime, the straight line segment lies entirely below the curve (Figure 2.5(a)). The maximin in (2.58) is then given by the maximum ordinate on the line segment, i.e., $r_W(R_Q) = R_Q - \frac{1}{2}\log(\max\{1, \frac{P_I}{D}\})$. This occurs for $R_Q \in \left[\left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+, \frac{1}{2}\log\left(\max\left\{1, \frac{P_I}{D}\right\} + \frac{\beta_A^2|P_I - D|}{P_V}\right)\right)$.

(b) In the second regime, the straight line segment intersects the rising portion of the curve (Figure 2.5(b)). The maximin in (2.58) is then given by the ordinate at the point of intersection (this value is given by the root of a cubic equation). This occurs for $R_Q \in \left[\frac{1}{2}\log\left(\max\left\{1, \frac{P_I}{D}\right\} + \frac{\beta_A^2|P_I - D|}{P_V}\right),\ \frac{1}{2}\log\left(1 + \frac{P_I}{D} + \frac{\beta_A^2(P_I + D)}{P_V}\right)\right]$.

(c) The third regime corresponds to all other values of R_Q, namely $R_Q > \frac{1}{2}\log(1 + \frac{P_I}{D} + \frac{\beta_A^2(P_I + D)}{P_V})$. In this case, the straight line segment either intersects the curve on its falling portion only (as in Figure 2.5(c)), or does not intersect it at all.

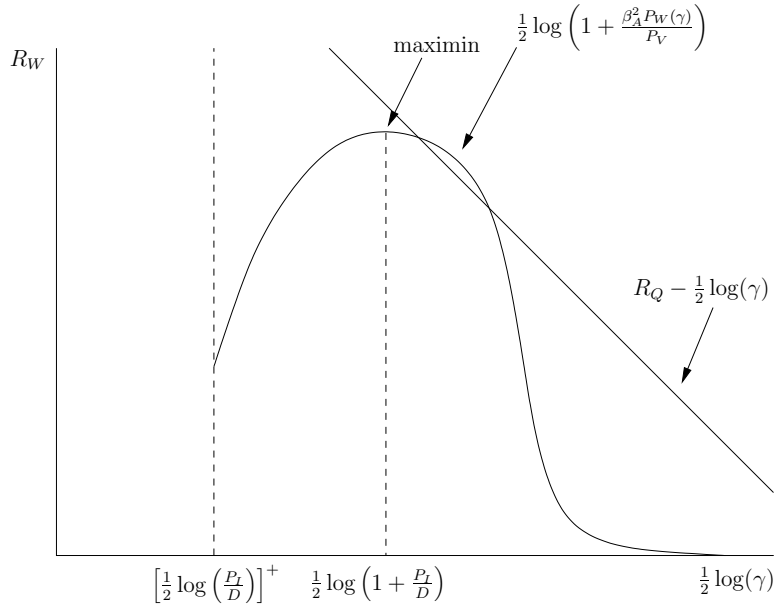


(a)



(b)

Figure 2.5: (a), (b): Plots of $R_Q-\frac{1}{2}\log(\gamma)$ and $\frac{1}{2}\log(1+\frac{\beta_A^2 P_W(\gamma)}{P_V})$ and determination of the maximin point for various values of R_Q (continued on next page).



(c)

Figure 2.5: (c): Determination of the maximin value when R_Q belongs to the third regime (continued from previous page).

The maximin value in (2.58) is then given by the maximum ordinate on the curve, i.e., $r_W(R_Q) = \frac{1}{2}\log(1 + \frac{\beta_A^2 D}{P_V})$. Note that this upper bound on R_W also follows by a simpler argument, namely that R_W can be no higher than the capacity of an AWGN channel with signal (i.e., watermark) power $\beta_A^2 D$ and noise power P_V (when no quantization noise is present, i.e., $R_Q = \infty$).

The three regimes obtained above correspond to the three segments AB, BC and $\mathcal{C}_\infty$ of the upper boundary of $\mathcal{R}_D^{\text{gauss}}$ described in Section 2.1.

Note: In the special case $P_V = 0$ (no attack), the curve in Figure 2.5 is displaced to $+\infty$ and only the first regime obtains, i.e., the bound on R_W is simply $R_W \leq R_Q - \left[\frac{1}{2}\log(\frac{P_I}{D})\right]^+$. The converse theorem then reduces to the channel coding part of the converse theorem in [35], and also the converse theorem of [8] for $R_F = 0$.

Direct Theorem

We now show that $\mathcal{R}_D^{\text{gauss}}$ is achievable.

Proof: First, we limit the quantization rate to $R_Q \geq \left[\frac{1}{2}\log(\frac{P_I}{D})\right]^+$. We use a random coding argument, similar to the discrete case (proved in Section 2.2).

Codebook Generation: Let $\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]$. A set of 2^{nR_Q} i.i.d. $\sim \mathcal{N}(0, \gamma P_W(\gamma))$ Gaussian sequences $\tilde{Y}^n$, is generated and partitioned into 2^{nR_W} subsets of 2^{nR_1} sequences each, i.e.,

$$R_Q = R_W + R_1 \tag{2.59}$$

The w^{th} subset, consisting of sequences $\tilde{Y}^n(w,1), \ldots, \tilde{Y}^n(w, 2^{nR_1})$, becomes the codebook for the w^{th} watermark.

Watermark Embedding: The procedure here is identical to the one given in the proof of Theorem 2.1. That is, from I^n and watermark index w the embedder identifies within the w^{th} codebook the first codeword $\tilde{Y}^n(w,q)$ such that the pair $(I^n, \tilde{Y}^n(w,q))$ lies in a typical set $T_{I,\hat{Y}}(\epsilon)$. The set $T_{I,\hat{Y}}(\epsilon)$ of typical pairs is constructed with respect to a bivariate Gaussian distribution $p_{I,\hat{Y}}$ having mean zero and covariance

$$K_{I,\hat{Y}} = \begin{bmatrix} P_I & \sqrt{(\gamma-1)P_I P_W(\gamma)} \\ \sqrt{(\gamma-1)P_I P_W(\gamma)} & \gamma P_W(\gamma) \end{bmatrix}$$

Note that the second moments in $K_{I,\hat{Y}}$ are consistent with the geometry of Figure 2.4, with $\gamma = \sin^{-2}(\phi)$. In particular, if the pair $(I^n, \hat{Y}^n)$ lies in $T_{I,\hat{Y}}(\epsilon)$, then the empirical second moments:

$$\frac{1}{n}\sum_{i=1}^n I_i^2, \qquad \frac{1}{n}\sum_{i=1}^n \hat{Y}_i^2 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n I_i\hat{Y}_i$$

are within ϵ (or a factor thereof) of the average values shown implicitly in Figure 2.4. This also means that the distortion constraint (2.42) is essentially met (since ϵ-differences can be safely ignored).

Decoding: Upon receiving $Z^n = \hat{Y}^n(W) + V^n$, the decoder seeks among all watermarked versions $\hat{Y}^n(1), \ldots, \hat{Y}^n(2^{nR_W})$ of I^n a single $\hat{Y}^n(\hat{w})$ such that the triplet $(I^n, \hat{Y}^n(\hat{w}), Z^n)$ lies in $T^n_{I,\hat{Y},Z}(\epsilon)$, the set of typical triplets with respect to the trivariate Gaussian distribution $p_{I,\hat{Y},Z}$ having zero mean and covariance matrix

$$K_{I,\hat{Y},Z} = \begin{bmatrix} P_I & \sqrt{(\gamma-1)P_I P_W(\gamma)} & \beta_A\sqrt{(\gamma-1)P_I P_W(\gamma)} \\ \sqrt{(\gamma-1)P_I P_W(\gamma)} & \gamma P_W(\gamma) & \beta_A \gamma P_W(\gamma) \\ \beta_A\sqrt{(\gamma-1)P_I P_W(\gamma)} & \beta_A\gamma P_W(\gamma) & \beta_A^2\gamma P_W(\gamma) + P_V \end{bmatrix}$$

If a unique such sequence $\hat{Y}^n(\hat{w})$ exists, then the decoder outputs $\hat{W} = \hat{w}$; otherwise, the decoder declares an error.

Note that $p_{I,\hat{Y},Z}(i,\hat{y},z) = p_{I,\hat{Y}}(i,\hat{y})p_V(z-\hat{y})$, where p_V is the marginal of the attack noise V^n.

Error Events: Carrying out the same analysis as in the proof of the discrete-case theorem, we find that the probability of error $\Pr\{W \neq \hat{W}\}$ can be arbitrarily small as long

$$R_W < \min\{R_Q - I(I;\hat{Y}), I(Z;\hat{Y}|I)\} \tag{2.60}$$

and since (from the definition of the typical sets $T_{I,\hat{Y}}, T_{I,\hat{Y},Z}$) we have $I(I;\hat{Y}) = \frac{1}{2}\log(\gamma)$ and $I(Z;\hat{Y}|I) = \frac{1}{2}\log\left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V}\right)$, we finally obtain

$$R_W < \min\left\{R_Q - \frac{1}{2}\log(\gamma), \frac{1}{2}\log\left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V}\right)\right\} \tag{2.61}$$

Choosing $\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]$ so as to maximize the right-hand side of (2.61), we can achieve the whole region $\mathcal{R}_D^{\text{gauss}}$.

This proof holds for random codes; however, using a standard expurgation argument we can argue that there exists a deterministic code such that the maximal probability of error is made arbitrarily small, for sufficiently large codelength n. ■

2.4 Proof of Theorem 2.3

We begin with the converse part of Theorem 2.3.

Converse Part

This part shows that every achievable rate (R_Q, R_W) (under the assumption that the attacker knows $p_{\hat{Y}^n|I^n}$, the decoder knows $p_{Z|\hat{Y}}$ and (2.7) is satisfied) must lie in $\mathcal{R}^{\text{dsc}}_{D,D_A}$.

Proof: Let $\epsilon > 0$. The watermarker chooses an encoding function f such that $p_{\hat{Y}^n|I^n} \in \mathcal{M}(p_I, D)$. By a standard rate-distortion theorem [33], we have

$$R_Q \geq \min_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I, D)} I(I; \hat{Y})$$

thus establishing the lower bound on R_Q. Now, since (R_Q, R_W) is achievable, we know that for every $p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)$, we have $\mathcal{P}_e < \epsilon$. Therefore, the converse of Theorem 2.1 (and hence inequality (2.32)) applies for every $p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)$. Thus

$$R_W \leq \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)} \min \left\{ R_Q - n^{-1} \sum_{j=1}^{n} I(I_j; \hat{Y}_j), n^{-1} \sum_{j=1}^{n} I(Z_j; \hat{Y}_j | I_j) \right\} + \epsilon \tag{2.62}$$

We can now use inequalities (2.33) and (2.34) to upper bound (2.62) and obtain

$$R_W \leq \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)} \Xi(R_Q,\ p_I,\ p^a_{\hat{Y}|I},\ p_{Z|\hat{Y}}) + \epsilon \tag{2.63}$$

where the single-letter pmf $p^a_{\hat{Y}|I}$, was defined in (2.35).

We now prove that $\mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A) = \mathcal{M}_A(p_I, (p^a_{\hat{Y}|I})^n, D_A)$ as follows:

$$\begin{aligned}
\mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A) &= \left\{ p_{Z|\hat{Y}} : n^{-1} \sum_{j=1}^{n} E d_A(\hat{Y}_j, Z_j) \leq D_A \right\} \\
&= \left\{ p_{Z|\hat{Y}} : n^{-1} \sum_{j=1}^{n} \sum_{z_j, \hat{y}_j, i_j} p_{Z|\hat{Y}}(z_j|\hat{y}_j)\, p_{\hat{Y}_j|I_j}(\hat{y}_j|i_j)\, p_I(i_j) d_A(\hat{y}_j, z_j) \leq D_A \right\} \\
&= \left\{ p_{Z|\hat{Y}} : \sum_{z,\hat{y},i} p_{Z|\hat{Y}}(z|\hat{y})\, p_I(i)\, d_A(\hat{y}, z) \left(n^{-1} \sum_{j=1}^{n} p_{\hat{Y}_j|I_j}(\hat{y}|i) \right) \leq D_A \right\} \quad (2.64) \\
&= \left\{ p_{Z|\hat{Y}} : \sum_{z,\hat{y},i} p_{Z|\hat{Y}}(z|\hat{y})\, p_I(i)\, p^a_{\hat{Y}|I}(\hat{y}|i)\, d_A(\hat{y}, z) \leq D_A \right\} \\
&= \mathcal{M}_A(p_I, (p^a_{\hat{Y}|I})^n, D_A) \quad (2.65)
\end{aligned}$$

where the equality in (2.64) is due to that fact that all variables $\{\hat{Y}_j, Z_j, I_j,\ j = 1, \ldots, n\}$ have the same support set as $\hat{Y}_1, Z_1, I_1$ respectively. Hence, from (2.63) and (2.65) we obtain

$$\begin{aligned}
R_W &\leq \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p^a_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p^a_{\hat{Y}|I},\ p_{Z|\hat{Y}}) + \epsilon \\
&\leq \max_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I, D)} \ \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) + \epsilon
\end{aligned} \quad (2.66)$$

where the last inequality stems from the fact that, as we saw in Section 2.2, $p^a_{\hat{Y}|I}$ satisfies the one-dimensional distortion constraint (2.36). Finally, by letting $\epsilon \to 0$, we conclude the proof. ■

Direct Part

The achievability of $\mathcal{R}^{\text{dsc}}_{D,D_A}$ is proved next.

Proof: Here, we use the proof of the direct part of Theorem 2.1. More specifically, we generate a random code exactly as in that proof, for a rate pair $(R_Q, R_W) \in \mathcal{R}^{\text{dsc}}_{D,D_A}$. Observe that the design of the encoder is oblivious of the attack channel

(as required by the rules of the game), and remains fixed for any distribution $p_{Z|\hat{Y}}$. The conditional distribution $p_{\hat{Y}^n|I^n}$ chosen by the encoder is $p^*_{\hat{Y}^n|I^n} = (p^*_{\hat{Y}|I})^n$, such that the quantity

$$\min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p^*_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}})$$

is maximized under the condition that $p^*_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I, D)$.

Assume now that the attacker uses some distribution $p'_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p^*_{\hat{Y}|I})^n, D_A)$. By the proof of Theorem 2.1 we know that the decoder (which presumably knows $p'_{Z|\hat{Y}}$), is able to detect the watermark with vanishing probability of error as long $R_W \leq \Xi(R_Q,\ p_I,\ p^*_{\hat{Y}|I},\ p'_{Z|\hat{Y}})$. Since $(R_Q, R_W) \in \mathcal{R}^{\text{dsc}}_{D,D_A}$, it follows that:

$$\begin{aligned} R_W &\leq \max_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I, D)} \ \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \\ &= \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, (p^*_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p^*_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \\ &\leq \Xi(R_Q,\ p_I,\ p^*_{\hat{Y}|I},\ p'_{Z|\hat{Y}}) \end{aligned} \tag{2.67}$$

as required.

In order to complete the proof, we need to prove that there exists a deterministic code with arbitrarily small probability of error for all possible attack channels. As is explained in [36], using random coding arguments for proving achievability results for a family of channels is not as straighforward as in the case of fixed channels. In our case, the family of channels is determined by the set $\mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)$. As we proved, for each $p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)$, the average probability of error over the ensemble of random codes is small. However, this does not immediately guarantee the existence of a single deterministic code which is simultaneously good *for all* $p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)$. In principle, for each $p_{Z|\hat{Y}}$, a *different* deterministic code could achieve small probability of error.

Existence of a deterministic code that achieves $\mathcal{R}_{D,D_A}^{\text{dsc}}$ can indeed be established using the same technique as in proving the existence of a deterministic code for compound channels [37]. This technique consists of: (a) first approximating $\mathcal{M}_A(p_I, (p^*_{\hat{Y}|I})^n, D_A)$ by a *discrete* set that contains sufficiently many channel distributions; (b) showing existence of a good code for a single (not compound) channel whose distribution function is an "average" of the distribution functions of the discrete set; and (c) proving that the code obtained in (b) can be used for the entire compound channel, with the original set of channel distributions. ■

2.5 Proof of Theorem 2.4

Theorem 2.3 can be applied to the continuous case as well. Replacing sums with integrals and pmf's with pdf's (where applicable), we can obtain a proof for the maximal achievable rate region for the case where

- $\mathcal{I} = \hat{\mathcal{Y}} = \mathcal{Z} = \mathbf{R}$.
- I^n is i.i.d. zero-mean Gaussian with variance P_I. The distribution of I is denoted p_I^G for simpicity.
- $d(x, y) = d_A(x, y) = (x - y)^2$.

Direct application of Theorem 2.3 gives:

$$\mathcal{R}_{D,D_A}^{\text{gauss}} = \Bigg\{ (R_Q, R_W) : \\ R_Q \geq \left[\frac{1}{2} \log \left(\frac{P_I}{D} \right) \right]^+ \\ R_W \leq \max_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I^G, D)} \; \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I^G, (p_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \Bigg\}$$

since we know that the rate-distortion function of an i.i.d. Gaussian source of variance P_I and distortion constraint D is given by

$$\min_{p_{\hat{Y}|I}:E(I-\hat{Y})^2\leq D} I(I;\hat{Y}) \;=\; \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+$$

Thus, it suffices to prove that

$$\max_{p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I^G,D)} \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(p_{\hat{Y}|I})^n,D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) =$$
$$\max_{\gamma\in\Gamma(R_Q,D,D_A)} \min\left\{R_Q-\frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1+\frac{P_W(\gamma)}{D_A}-\frac{1}{\gamma}\right)\right\} \quad (2.68)$$

where $\Gamma(R_Q,D,D_A)$ was defined in (2.13).

In order to prove (2.68), we show that the left-hand side of (2.68) is upper- and lower-bounded by the quantity on the right-hand side, i.e., by

$$\max_{\gamma\in\Gamma(R_Q,D,D_A)} \min\left\{R_Q-\frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1+\frac{P_W(\gamma)}{D_A}-\frac{1}{\gamma}\right)\right\}.$$

Observe that for any $p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I^G,D)$ and any $p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(p_{\hat{Y}|I})^n,D_A)$, the n-variate i.i.d. extensions of the scalar random variables $I,\hat{Y},Z$, whose joint distribution is $p_I^G\,p_{\hat{Y}|I}\,p_{Z|\hat{Y}}$, are consistent with the L_2-space geometry of Section 2.3. The quantities $\gamma, P_W(\gamma), P_{\hat{Y}}$ are defined similarly here.

Upper bound

We now define a class of conditional probabilities $\{p'_{Z|\hat{Y}}(p_{\hat{Y}|I})\}$ such that

$$(\forall\ p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I^G,D))\ \ p'_{Z|\hat{Y}}(p_{\hat{Y}|I})\in\mathcal{M}_A(p_I^G,(p_{\hat{Y}|I})^n,D_A)$$

By the above, it obviously holds that

$$\max_{p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I^G,D)} \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(p_{\hat{Y}|I})^n,D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \leq$$
$$\max_{p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I^G,D)} \Xi(R_Q,\ p_I^G,\ p_{\hat{Y}|I},\ p'_{Z|\hat{Y}}(p_{\hat{Y}|I})) \quad (2.69)$$

The construction of the class $p'_{Z|\hat{Y}}(p_{\hat{Y}|I})$ is as follows:

Case 1

If $P_{\hat{Y}} = E(\hat{Y}^2) \leq D_A$ then $p'_{Z|\hat{Y}}(p_{\hat{Y}|I})(z|\hat{y}) = \delta(z)$, i.e., $Z = 0$. Observe that $p'_{Z|\hat{Y}} \in \mathcal{M}_A(p_I^G, (p_{\hat{Y}|I})^n, D_A)$ since $E[(\hat{Y} - Z)^2] = E(\hat{Y}^2) \leq D_A$. Thus, the upper bound of (2.69) is zero in this case. In terms of the notation of Section 2.3 (proof of Theorem 2.2), the condition $P_{\hat{Y}} \leq D_A$ is equivalent to $\gamma P_W(\gamma) \leq D_A$, and hence $\Gamma(R_Q, D, D_A)$ is empty.

Case 2

If $P_{\hat{Y}} > D_A$ or, equivalently,

$$\gamma P_W(\gamma) > D_A \tag{2.70}$$

then $p'_{Z|\hat{Y}}(p_{\hat{Y}|I}) = \mathcal{N}(\beta_A \hat{Y}, P_V)$ where

$$\beta_A = 1 - \frac{D_A}{E(\hat{Y}^2)}, \quad P_V = \beta_A D_A$$

In other words, $Z = \beta_A \hat{Y} + V$, where V is zero-mean Gaussian with variance P_V, independent from $I, \hat{Y}$. It is straightforward to show that $E[(\hat{Y} - Z)^2] = D_A$ in this case.

We now have

$$I(Z; \hat{Y}|I) = h(Z|I) - h(V)$$

and applying the chain of inequalities (2.53)-(2.55), we obtain

$$\begin{aligned} I(Z; \hat{Y}|I) &\leq \frac{1}{2} \log \left(1 + \frac{\beta_A^2 P_W(\gamma)}{P_V} \right) && (2.71) \\ &= \frac{1}{2} \log \left(1 + \frac{P_W(\gamma)}{D_A} - \frac{1}{\gamma} \right) && (2.72) \end{aligned}$$

where the actual values of β_A and P_V were used in the last equality. Also, from (2.47), we obtain

$$I(I;\hat{Y}) \geq \frac{1}{2}\log(\gamma) \tag{2.73}$$

Hence, from (2.69), (2.72) and (2.73) we have

$$\begin{aligned}
&\max_{p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I,D)} \ \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I,(p_{\hat{Y}|I})^n,D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \\
&\quad\leq \max_{p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I,D)} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1+\frac{P_W(\gamma)}{D_A}-\frac{1}{\gamma}\right)\right\} \quad (2.74)\\
&\quad= \max_{\gamma\in\Gamma(R_Q,D,D_A)} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1+\frac{P_W(\gamma)}{D_A}-\frac{1}{\gamma}\right)\right\} \quad (2.75)
\end{aligned}$$

where (2.75) holds because the right-hand side of (2.74) depends only on γ, whose range is determined by (2.56) and by (2.70).

We now proceed to prove the lower bound.

Lower Bound

Firstly, if $\Gamma(R_Q, D, D_A)$ is empty, we set the lower bound to be trivially equal to zero, as required by the theorem.

If $\Gamma(R_Q, D, D_A)$ is non empty, we consider a conditional probability $\tilde{p}_{\hat{Y}|I}$ such that $(I, \hat{Y})$ are jointly zero-mean Gaussian with respect to the covariance matrix $K_{I,\hat{Y}}$ that we saw in the proof of the direct part of Theorem 2.2. That is,

$$K_{I,\hat{Y}} = \begin{bmatrix} P_I & \sqrt{(\gamma-1)P_I P_W(\gamma)} \\ \sqrt{(\gamma-1)P_I P_W(\gamma)} & \gamma P_W(\gamma) \end{bmatrix}$$

where $\gamma, P_W(\gamma)$ are defined similarly to Section 2.3. Moreover, we set

$$\gamma = \arg\max_{\gamma\in\Gamma(R_Q,D,D_A)} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1+\frac{P_W(\gamma)}{D_A}-\frac{1}{\gamma}\right)\right\} \tag{2.76}$$

Also, the matrix $K_{I,\hat{Y}}$ is consistent with the L_2-space geometry of Section 2.3, and hence the distortion constraint (2.42) is satisfied. Therefore,

$$\tilde{p}_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I^G, D).$$

Hence, we have

$$\begin{aligned}
&\max_{p_{\hat{Y}|I}\in\mathcal{M}^{ml}(p_I^G,D)} \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(p_{\hat{Y}|I})^n,D_A)} \Xi(R_Q,\ p_I^G,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \\
&\geq \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(\tilde{p}_{\hat{Y}|I})^n,D_A)} \Xi(R_Q,\ p_I^G,\ \tilde{p}_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \\
&\geq \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(\tilde{p}_{\hat{Y}|I})^n,D_A)} I(Z;\hat{Y}|I)\right\} \qquad (2.77)
\end{aligned}$$

where (2.77) was obtained by substituting $I(I;\hat{Y})$ with $\frac{1}{2}\log(\gamma)$ (from the Gaussianity of $\tilde{p}_{\hat{Y}|I}$).

We now have the following chain of inequalities:

$$\begin{aligned}
I(Z;\hat{Y}|I) &= I(\hat{Y}-\lambda_0 I; Z-\lambda_0\kappa I|I) && (2.78)\\
&= h(\hat{Y}-\lambda_0 I) - h(\hat{Y}-\lambda_0 I|Z-\lambda_0\kappa I, I) && (2.79)\\
&\geq h(\hat{Y}-\lambda_0 I) - h(\hat{Y}-\lambda_0 I|Z-\lambda_0\kappa I) && (2.80)\\
&= I(\hat{Y}-\lambda_0 I; Z-\lambda_0\kappa I) && (2.81)
\end{aligned}$$

where λ_0 is the linear MMSE coefficient for estimating $\hat{Y}$ given I (defined in (2.44)) and $\kappa \neq 0$ is an arbitrary constant; (2.79) was obtained since $\hat{Y}-\lambda_0 I$ is orthogonal to I (and hence they are independent); and (2.80) is due to the fact that conditioning reduces entropy.

Let $\tilde{Z}$ be a zero-mean Gaussian random variable such that $(I,\hat{Y},\tilde{Z})$ are jointly Gaussian and have the same second moments as $(I,\hat{Y},Z)$. A lower bound to (2.81) can be obtained as follows:

$$\begin{aligned}
&I(\hat{Y}-\lambda_0 I; Z-\lambda_0\kappa I) \\
&= h(\hat{Y}-\lambda_0 I) - h(\hat{Y}-\lambda_0 I|Z-\lambda_0\kappa I) \\
&= h(\hat{Y}-\lambda_0 I) - h(\hat{Y}-\lambda_0 I-\mu(Z-\lambda_0\kappa I)|Z-\lambda_0\kappa I),\quad \mu\neq 0 && (2.82)\\
&\geq h(\hat{Y}-\lambda_0 I) - h(\hat{Y}-\lambda_0 I-\mu(Z-\lambda_0\kappa I)) && (2.83)
\end{aligned}$$

$$
\begin{aligned}
&\geq h(\hat{Y}-\lambda_0 I)-\frac{1}{2}\log(E[(\hat{Y}-\lambda_0(1-\mu)I-\mu Z)^2]) && (2.84)\\
&= h(\hat{Y}-\lambda_0 I)-\frac{1}{2}\log(E[(\hat{Y}-\lambda_0(1-\mu)I-\mu\tilde{Z})^2]) && (2.85)\\
&= h(\hat{Y}-\lambda_0 I)-h(\hat{Y}-\lambda_0 I-\mu(\tilde{Z}-\lambda_0\kappa I)|Z^*-\lambda_0\kappa I) && (2.86)\\
&= I(\hat{Y}-\lambda_0 I;\tilde{Z}-\lambda_0\kappa I) && (2.87)
\end{aligned}
$$

where μ in (2.82) is the linear MMSE coefficient for estimating $\hat{Y}-\lambda_0 I$ given $Z-\lambda_0\kappa I$ (hence, $\hat{Y}-\lambda_0 I-\mu(Z-\lambda_0\kappa I)$ is orthogonal to $Z-\lambda_0\kappa I$); (2.83) was obtained because conditioning reduces entropy; (2.84) is due to the Gaussian entropy upper bound; (2.85) is from the definition of $\tilde{Z}$; and (2.86) holds because $\hat{Y}-\lambda_0 I-\mu(\tilde{Z}-\lambda_0\kappa I)$ is independent of $\tilde{Z}-\lambda_0\kappa I$ (since they are Gaussian and uncorrelated).

From (2.77), (2.81) and (2.87) we obtain

$$
\begin{aligned}
&\min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(\tilde{p}_{\hat{Y}|I})^n,D_A)} I(Z;\hat{Y}|I)\\
&\quad\geq \min_{p_{Z|\hat{Y}}\in\mathcal{M}_A(p_I^G,(\tilde{p}_{\hat{Y}|I})^n,D_A)} I(\hat{Y}-\lambda_0 I;\tilde{Z}-\lambda_0\kappa I)\\
&\quad= \min_{E(\tilde{Z}\hat{Y}),E(\tilde{Z}^2):\ E[(\tilde{Z}-\hat{Y})^2]\leq D_A} I(\hat{Y}-\lambda_0 I;\tilde{Z}-\lambda_0\kappa I) \qquad (2.88)
\end{aligned}
$$

where the last equality is because the quantity $I(\hat{Y}-\lambda_0 I;\tilde{Z}-\lambda_0\kappa I)$ depends on $p_{Z|\hat{Y}}$ only through the second moments $E(\tilde{Z}\hat{Y})=E(Z\hat{Y})$ and $E(\tilde{Z}^2)=E(Z^2)$ (since $\tilde{Z}$ is Gaussian).

Since $\hat{Y},\tilde{Z}$ are jointly Gaussian, we can express them in the form:

$$\tilde{Z}=\kappa\hat{Y}+U$$

where κ is the same as used in (2.88), and U is a zero-mean Gaussian variable independent of $\hat{Y}$ and I (because of the Markov condition (2.12)). Therefore, $E(\tilde{Z}\hat{Y})=\kappa\gamma P_W(\gamma)$ and $E(\tilde{Z}^2)=\kappa^2\gamma P_W(\gamma)+P_U$, where P_U is the variance of U.

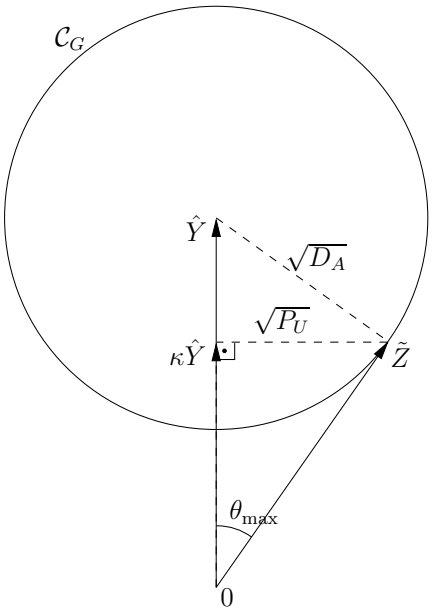


Figure 2.6: The L_2 space spanned by variables $\hat{Y}$ and $\tilde{Z}$, shown for the maximum possible $\theta = \theta_{\max}$ (when $P_{\hat{Y}} > D_A$).

Hence, we have

$$\begin{aligned} & I(\hat{Y} - \lambda_0 I; \tilde{Z} - \lambda_0 \kappa I) \\ &= h(\tilde{Z} - \lambda_0 \kappa I) - h(\tilde{Z} - \lambda_0 \kappa I | \hat{Y} - \lambda_0 I) \\ &= h(\kappa(\hat{Y} - \lambda_0 I) + U) - h(\kappa(\hat{Y} - \lambda_0 I) + U | \hat{Y} - \lambda_0 I) \\ &= \frac{1}{2}\log(2\pi e)(\kappa^2 P_W(\gamma) + P_U) - \frac{1}{2}\log(2\pi e) P_U \\ &= \frac{1}{2}\log\left(1 + \frac{\kappa^2 P_W(\gamma)}{P_U}\right) \end{aligned} \tag{2.89}$$

Therefore, (2.88) is equal to

$$\min_{(\kappa,\ P_U):\ E[(\hat{Y}-\tilde{Z})^2] \le D_A} \frac{1}{2}\log\left(1 + \frac{\kappa^2 P_W(\gamma)}{P_U}\right) \tag{2.90}$$

We observe now that minimizing the ratio κ^2/P_U in (2.90) is equivalent to minimizing the ratio $\kappa^2 P_{\hat{Y}}/P_U$. This ratio is equal to $\cot^2(\theta)$, where θ is shown in the L_2-space of Figure 2.6. Therefore, for computing the minimum in (2.90), it suffices

to maximize θ. We consider the following two cases: (i) if $P_{\hat{Y}} = \gamma P_W(\gamma) \leq D_A$ then the maximum possible θ is $\pi/2$ and hence (2.90) is equal to zero (note that $\Gamma(R_Q, D, D_A)$ is empty in this case); (ii) if $\gamma P_W(\gamma) > D_A$, then θ is maximized when $\tilde{Z}$ is tangent to the circle $\mathcal{C}_G$ and $E[(\hat{Y}-\tilde{Z})^2] = D_A$; in this case $\cot^2(\theta) = \frac{\gamma P_W(\gamma)}{D_A} - 1$ and (2.90) is equal to

$$\frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A} - \frac{1}{\gamma}\right) \tag{2.91}$$

By combining (2.76), (2.77), (2.88) and (2.91), we obtain

$$\begin{aligned} &\max_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I^G, D)} \ \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I^G, (p_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I^G,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \\ &\geq \max_{\gamma \in \Gamma(R_Q, D, D_A)} \ \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A} - \frac{1}{\gamma}\right)\right\} \end{aligned} \tag{2.92}$$

as required.

Combining (2.75) and (2.92), the theorem is proved. ∎

Remarks: In proving (2.68), we had to prove that

$$\max_{p_{\hat{Y}|I} \in \mathcal{M}^{ml}(p_I^G, D)} \ \min_{p_{Z|\hat{Y}} \in \mathcal{M}_A(p_I^G, (p_{\hat{Y}|I})^n, D_A)} \Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}}) \tag{2.93}$$

is upper- and lower-bounded by a suitable quantity. In the course of the proof, it became apparent that, in most cases of interest (that is, $P_{\hat{Y}} > D_A$), the most malicious attack is of the form

$$Z^n = \left(1 - \frac{D_A}{P_{\hat{Y}}}\right)\hat{Y}^n + V^n \tag{2.94}$$

where V^n is i.i.d. Gaussian, independent of $\hat{Y}^n$, with variance $(1 - \frac{D_A}{P_{\hat{Y}}})D_A$. As can be seen in Figure 2.6, this particular attack is equivalent to optimally quantizing an i.i.d. Gaussian source at distortion D_A. This attack has been proved to be optimal (from the attacker's point of view) in other watermarking schemes, as well [22, 24].

As we pointed out in Section 2.1, the set of admissible encoding and attack distributions $\mathcal{A}(p_I, D, D_A)$ is non-rectangular, since $\mathcal{M}_A(p_I, p_{\hat{Y}^n|I^n}, D_A)$ depends on the choice of $p_{\hat{Y}^n|I^n}$. Therefore, one cannot use any concavity and convexity properties of $\Xi(R_Q,\ p_I,\ p_{\hat{Y}|I},\ p_{Z|\hat{Y}})$ in order to establish the existence of a saddle-point that would give the value of (2.93). The approach we followed in this section (i.e., finding upper and lower bounds to (2.93)) overcomes this difficulty. Similar techniques were followed by [22, 24].

2.6 Performance of Other Schemes

In this section we present achievability results for certain schemes that combine watermarking and compression. Specifically, we investigate the relationship between watermarking and quantization rates in the presence of additive memoryless Gaussian noise of variance D_A, for the following systems:

- Regular Quantization Index Modulation (QIM) [16], where no knowledge of the original image is available at the decoder (public scenario).
- Additive watermarking, where the embedder computes the weighted sum of the original image and a watermark-dependent signal and then compresses the resulting vector using a universal (watermark non-specific) quantizer. A private detection scenario is assumed in this case.

Although our focus is on achievability results, the rate region $\mathcal{R}_D^{\text{gauss}}$ derived in Section 2.3 (for $\beta_A = 1$ and $P_V = D_A$,) can be taken as an outer bound on the achievable rate region of both schemes considered in this section.

A. Regular Quantization Index Modulation, Public Scenario

We consider the *regular* version of QIM [16] (distinct from *distortion-compensated*

QIM), since we require the output of the embedding process to be a quantized image (corresponding to an index in a source codebook).

Essentially, here we have an ensemble of 2^{nR_W} quantizers and their codebooks. Each quantizer corresponds to a different watermark index, and covers the entire image space with $2^{n(R_Q-R_W)}$ representation vectors (codewords). The watermark W is embedded into an original image I^n by quantizing I^n using the W^{th} quantizer, yielding a representation vector $\hat{Y}^n$. Detection of the watermark W in a (possibly corrupted) image Z^n entails mapping Z^n to a representation vector taken from the *union* of the 2^{nR_W} codebooks; the index of the codebook which contains that vector becomes the estimate $\hat{W}$ of the watermark W. (By contrast, the private detection scenario used in the proof of the direct theorem of Section 2.3 mapped Z^n to one of 2^{nR_W} representation vectors, each taken from a *different* codebook.)

As discussed in [16], achievable pairs (R_Q, R_W) for regular QIM (also called "hidden" QIM) under constraints (2.1) and (2.2) can be found using a well-known formula due to Gel'fand and Pinsker [13]:

$$R_Q = I(\hat{Y}; Z) = I(\hat{Y}; \hat{Y} + V) \tag{2.95}$$

$$R_W = [I(\hat{Y}; Z) - I(\hat{Y}; I)]^+ \tag{2.96}$$

The trivariate distribution $p_{I,\hat{Y},Z}(i, \hat{y}, z)$, can be taken as the Gaussian in the proof of the direct theorem in Section 2.3. Thus $p_{I,\hat{Y},Z}(i, \hat{y}, z) = p_{I,\hat{Y}}(i, \hat{y}) p_V(z - \hat{y})$, where I and $V = Z - \hat{Y}$ are independent with mean zero and variances P_I and D_A respectively; and $\hat{Y}$ also has mean zero and satisfies $E(\hat{Y} - I)^2 = D$. It should be noted again that the second moments of $p_{I,\hat{Y},Z}(i, \hat{y}, z)$ are consistent with the geometry of Figure 2.4.

We briefly investigate the behavior of (2.96) as R_Q (given by (2.95)) varies. Letting $P_{\hat{Y}} = \gamma P_W(\gamma) = E(\hat{Y}^2)$, we have from (2.95)

$$R_Q = \frac{1}{2}\log\left(1 + \frac{P_{\hat{Y}}}{D_A}\right)$$

and thus

$$P_{\hat{Y}} = D_A(2^{2R_Q} - 1) \tag{2.97}$$

Also, (2.96) gives

$$R_W = \left[R_Q - \frac{1}{2}\log(\gamma)\right]^+ \tag{2.98}$$

Setting $P_{\hat{Y}|I} = P_W(\gamma) = P_{\hat{Y}}/\gamma$ in (2.45) and expressing γ in terms of P_I, $P_{\hat{Y}}$ and D, we obtain (with the aid of (2.97))

$$R_W = \left[R_Q - \frac{1}{2}\log\left(\frac{P_I D_A(2^{2R_Q} - 1)}{P_I D_A(2^{2R_Q} - 1) - \frac{1}{4}(P_I + D_A(2^{2R_Q} - 1) - D)^2}\right)\right]^+ \tag{2.99}$$

The range of values of R_Q for which R_W in (2.99) is nonzero is a subinterval of

$$\left[\frac{1}{2}\log\left(1 + \frac{(\sqrt{P_I} - \sqrt{D})^2}{D_A}\right), \frac{1}{2}\log\left(1 + \frac{(\sqrt{P_I} + \sqrt{D})^2}{D_A}\right)\right]$$

whose exact endpoints are given by the roots of a cubic. Expression (2.99) is shown in Figure 2.7 as the dashed-dotted curved line. One can trivially achieve the rest of the region (below the horizontal, dashed-dotted line), by appending extra "dummy" bits to the output of the quantizer (thus increasing the rate R_Q). As can be seen from Figure 2.7, the watermarking rate R_W obtained using i.i.d. Gaussian codebooks is positive only for a finite range of values of R_Q (without appending the trivial bits). This is explained by the fact that as the quantization rate increases, the quantization cells shrink and thus it becomes increasingly likely that a corrupted image will be mistaken for an image generated by another quantizer (resulting in a different watermark index at the decoder). This difficulty does

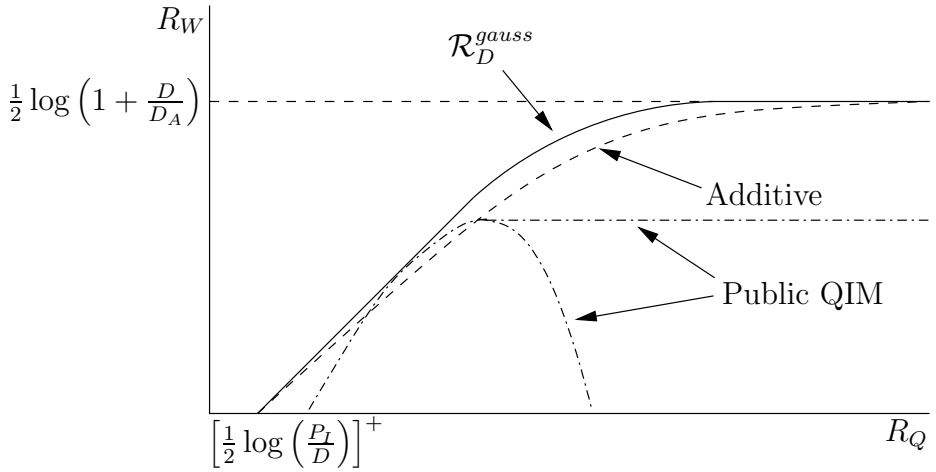


Figure 2.7: Inner bounds on the achievable rate regions for public QIM and private additive schemes. $\mathcal{R}_D^{\text{gauss}}$, for $\beta_A = 1$ and $P_V = D_A$, is an outer bound on the achievable rate regions of both schemes.

not arise when additive watermarking schemes (see, e.g., [31, 8]) are used. The analysis of such an additive scheme follows.

B. Additive Watermarking, Private Scenario

Additive watermarking schemes (see, e.g., [31, 8]) are immune to the problem discussed above, as they use a single quantizer which is not dependent on the embedded watermark. From a complexity/cost viewpoint, they are particularly attractive in applications where the same image is distributed to different customers (i.e., the embedded watermark is a fingerprint identifying the customer), as customers can use the same codebook in order to reconstruct their image.

In general, additive watermarking reduces to the computation of

$$Y^n = \alpha I^n + \beta x^n(W) \tag{2.100}$$

where W is the index of the watermark and $x^n(\cdot)$ is a n-dimensional signal that does not depend on the original image I^n. α, β are non-zero scalars. To further

compress Y^n, a universal quantizer (i.e., one that does not depend on the watermark embedded in Y^n) can be used:

$$\hat{Y}^n = f(Y^n)$$

subject to an appropriate distortion constraint ((2.1) in this case). The decoder attempts to detect W given $\hat{Y}^n$ and I^n with vanishing probability of error.

We obtain an inner bound on the achievable (R_Q, R_W) region using a random coding argument. First, we note that compressing Y^n is equivalent to compressing $\alpha^{-1}Y^n$, which effectively eliminates the parameter α in (2.100); and that the parameter β can be absorbed in the power of the watermark. Thus we use the simpler form

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

The watermarker generates a random channel codebook $\{X^n(1), \ldots, X^n(2^{nR_W})\}$, all components of which are i.i.d. Gaussian with variance P_X; and a random source codebook $\{\tilde{Y}^n(1), \ldots, \tilde{Y}^n(2^{nR_Q})\}$, also i.i.d. Gaussian with variance $P_{\hat{Y}}$, where both P_X and $P_{\hat{Y}}$ are free parameters in the model.

Y^n is encoded as $\hat{Y}^n = \tilde{Y}^n(q)$, where q is the smallest index such that the pair $(Y^n, \tilde{Y}^n(q))$ is jointly typical with respect to a bivariate Gaussian $p_{Y,\hat{Y}}$ having mean zero and covariance

$$K_{Y,\hat{Y}} = \begin{bmatrix} P_I + P_X & \frac{P_I+P_X}{2P_I}(P_I + P_{\hat{Y}} - D) \\ \frac{P_I+P_X}{2P_I}(P_I + P_{\hat{Y}} - D) & P_{\hat{Y}} \end{bmatrix}$$

Without going into detail, it is not difficult to show that joint typicality of Y^n and $\hat{Y}^n = \tilde{Y}^n(q)$ implies that the per-letter distortion between I^n and $\hat{Y}^n$ is, with probability approaching unity, no larger than $D + \epsilon$, which in turn implies that the distortion constraint (2.1) is essentially satisfied. By the usual rate-distortion

argument, taking

$$R_Q = I(Y;\hat{Y}) + \epsilon \tag{2.101}$$

ensures that, with probability approaching unity, a jointly typical pair $(Y^n, \tilde{Y}^n(q))$ can be found. (As expected from rate-distortion theory, $I(Y;\hat{Y}) \geq \frac{1}{2}\log(\frac{P_I}{D})$ with equality iff P_X=0 and $P_I = P_{\hat{Y}} + D$.)

Upon receiving $Z^n = \hat{Y}^n + V^n$, the watermark detector attempts to find a unique w such that the triplet $(I^n, X^n(w), Z^n)$ is jointly typical with respect to a trivariate Gaussian $p_{I,X,Z}$ having mean zero and covariance

$$K_{I,X,Z} = \begin{bmatrix} P_I & 0 & \frac{P_I+P_{\hat{Y}}-D}{2} \\ 0 & P_X & \frac{P_X(P_I+P_{\hat{Y}}-D)}{2P_I} \\ \frac{P_I+P_{\hat{Y}}-D}{2} & \frac{P_X(P_I+P_{\hat{Y}}-D)}{2P_I} & P_{\hat{Y}} + D_A \end{bmatrix}$$

(This distribution is consistent with $p_{Y,\hat{Y}}$ and the additive noise distribution p_V in the sense that $p_{Z|I,X}(z|i,x) = \int_{\hat{y}} p_{Y,\hat{Y}}(i+x,\hat{y})p_V(z-\hat{y})d\hat{y}/p_Y(i+x)$.) Again, without going into detail, it can be shown that if

$$R_W = I(X;I,Z) - \epsilon \tag{2.102}$$

then the probability of decoding error vanishes as $n \to \infty$.

Solving (2.101) for P_X, then substituting into (2.102) and letting $\epsilon \to 0$, we obtain the following achievable watermarking rate:

$$R_W = \frac{1}{2}\log\left(\frac{2^{2R_Q}(2D(P_I+P_{\hat{Y}}) - D^2 - (P_I - P_{\hat{Y}})^2 + 4D_A P_I)}{4P_I(2^{2R_Q}D_A + P_{\hat{Y}})}\right) \tag{2.103}$$

which is positive for $R_Q \geq \frac{1}{2}\log\left(\frac{P_I}{D}\right)$. (2.103) is maximized for

$$P_{\hat{Y}} = -2^{2R_Q}D_A + \sqrt{(2^{2R_Q}D_A + D)^2 + P_I(P_I + 2D_A(2^{2R_Q}-2) - 2D)}$$

yielding the final expression

$$R_W = \frac{1}{2}\log\left(\frac{2^{2R_Q}\left(4P_I(D+D_A)-\left(D+P_I+2^{2R_Q}D_A-\sqrt{(2^{2R_Q}D_A+D)^2+P_I(P_I+2D_A(2^{2R_Q}-2)-2D)}\right)^2\right)}{4P_I\sqrt{(2^{2R_Q}D_A+D)^2+P_I(P_I+2D_A(2^{2R_Q}-2)-2D)}}\right) \tag{2.104}$$

The corresponding curve is also shown in Figure 2.7 (the region below it being an inner bound on the achievable region for this additive scheme). As expected, when $R_Q \to \infty$, $\hat{Y}^n$ is negligibly different from $Y^n = I^n + X^n$ and thus R_W approaches the capacity of an AWGN channel.

Chapter 3

Fingerprinting and Collusion Attacks

In this chapter, we extend the results of Chapter 2 to the case of fingerprinting. As we mentioned in Chapter 1, fingerprinting is used for tracking illegal distributors of a protected image. More precisely, the information hider creates different fingerprinted versions of an image and distributes each to a respective customer. Each customer thus receives an image containing a fingerprint which uniquely identifies him. If a fingerprint is detected in an illegally distributed copy, then it is likely that the customer, to whom the fingerprint was assigned, is responsible for the illegal distribution.

Fingerprinting, like watermarking, has to adhere to some transparency and robustness requirements. That is, each fingerprinted copy should be of the same (or comparable) quality as the original image; and the hidden fingerprint should be recoverable even after degradation (possibly due to a malicious attack) of the protected work. An important consideration in fingerprinting is that attacks can be made more effective through collusion: two or more users who possess different fingerprinted copies of the *same* image can collude to produce a forgery. The fingerprint detector will then attempt to detect all the colluders (that is, all the

fingerprint indices) from the forgery.

Fingerprinting has attracted considerable attention during the recent years. Most work has focused on designing practical codes which are resistant to collusion attacks [38, 39, 40, 11]; little work has been done on information-theoretic aspects in terms of achievable rates [8, 41]. This chapter gives results on achievable rate regions in the presence of quantization.

The chapter is organized as follows: in Section 3.1 we summarize our results; Sections 3.2, 3.3 and 3.4 contain proofs of the theorems.

3.1 Summary of the Results

3.1.1 Discrete and Continuous Alphabets

The general form of the system under consideration is shown in Figure 3.1. The information hider creates 2^{nR_F} fingerprinted copies $\hat{Y}^n(1, I^n), \ldots, \hat{Y}^n(2^{nR_F}, I^n)$ of I^n, and distributes them to an equal number of customers using nR_Q bits per customer. We now assume that k (out of 2^{nR_F}) customers collude by combining their copies $\{\hat{Y}^n(W_1, I^n), \ldots, \hat{Y}^n(W_k, I^n)\}$ and produce a forgery Z^n. Then Z^n, together with I^n (in a private scenario), are provided to the fingerprint decoder, which outputs an estimate $\{\hat{W}_1, \ldots, \hat{W}_k\}$ of $\{W_1, \ldots, W_k\}$. Note that we have successful detection if all the fingerprint indices are correctly detected; this does not necessarily require $\hat{W}_l = W_l$ for all $1 \leq l \leq k$. In other words, the decoder tries to estimate the set $\{W_1, \ldots, W_k\}$, or, equivalently, any permutation of $(W_1, \ldots, W_k)$. The indices $W_1, \ldots, W_k$ are all distinct and the vector $(W_1, \ldots, W_k)$ is uniformly distributed in the set

$$\mathcal{F}(n,k) \stackrel{\Delta}{=} \{(w_1, \ldots, w_k) \in \{1, \ldots, 2^{nR_F}\}^k : \ (\forall\ l \neq m)\ w_l \neq w_m\}$$

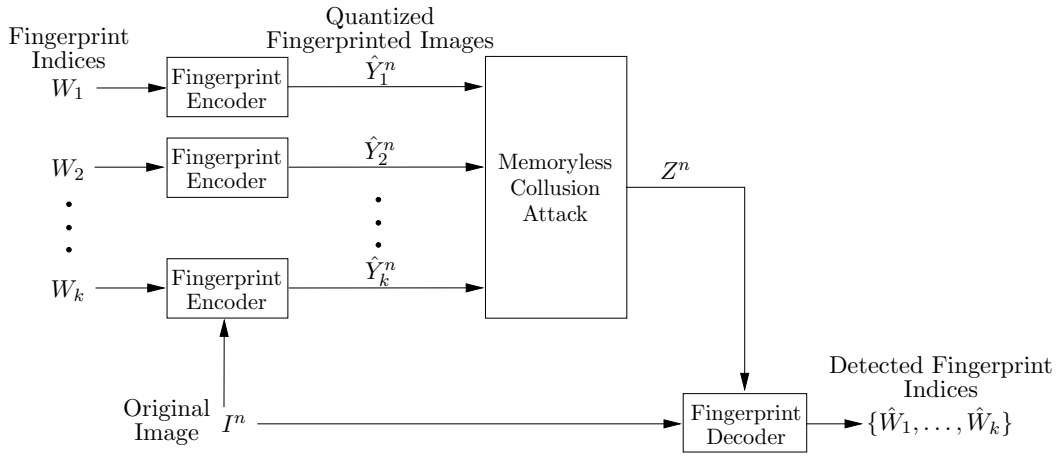


Figure 3.1: The general fingerprinting/quantization system with memoryless collusion attacks.

Note that the size of $\mathcal{F}(n,k)$ is $J(n,k) \stackrel{\Delta}{=} |\mathcal{F}(n,k)| = \frac{J_n!}{(J_n-k)!}$, where $J_n = 2^{nR_F}$. Using Stirling's approximation formula [42], we have

$$J(n,k) \doteq (J_n)^k \tag{3.1}$$

where we use the notation $a_n \doteq 2^{n(b\pm\epsilon)}$ to mean

$$\left| \frac{1}{n} \log a_n - b \right| < \epsilon$$

for n sufficiently large [33]. Furthermore,

$$\Pr\{W_l = w_l\} = \Pr\{W_1 = w_1\} = \frac{(J_n-1)!}{(J_n-1-(k-1))!} \cdot \frac{1}{J(n,k)} = \frac{1}{J_n}$$

We further assume that k, as well as the attack channel conditional distribution, are known to the fingerprint encoder and decoder. Similarly to Chapter 2, here too we assume that the attack is time-independent and memoryless; its conditional probability distribution is

$$p^{(k)}_{Z^n|\hat{Y}_1^n,\ldots,\hat{Y}_k^n} = (p^{(k)}_{Z|\hat{Y}_1,\ldots,\hat{Y}_k})^n$$

We assume that $k \geq 2$ and that it is fixed in n. The case $k = 1$ is the case of no collusion, treated in Chapter 2. For simplicity, we use the notation $\hat{Y}_l^n = \hat{Y}^n(W_l, I^n)$.

We now have the following definitions:

Definition 3.1 *A $(2^{nR_Q}, 2^{nR_F}, n)$ private quantization/fingerprinting code consists of the following:*

- *A fingerprint set $\mathcal{F}_n = \{1, \ldots, 2^{nR_F}\}$.*
- *An encoding function $f : \mathcal{F}_n \times \mathcal{I}^n \to \hat{\mathcal{Y}}^n$ which maps a fingerprint index w and an image sequence i^n to a representation sequence $\hat{y}^n$ taken from the set $\{\hat{y}^n(1), \ldots, \hat{y}^n(2^{nR_Q})\}$.*
- *A decoding function $g^{(k)} : \mathcal{Z}^n \times \mathcal{I}^n \to \mathcal{F}(n,k)$, which maps the output of the channel z^n and the original image i^n to an estimate $(\hat{w}_1, \ldots, \hat{w}_k)$ of a permutation of $(w_1, \ldots, w_k)$.*

Definition 3.2 *The probability of error in detecting fingerprints $(w_1, \ldots, w_k)$ is given by*

$$\mathcal{P}_e(w_1, \ldots, w_k) = \Pr\{g^{(k)}(Z^n, I^n) \notin \mathcal{P}(w_1, \ldots, w_k) | \hat{Y}_1^n = f(w_1, I^n), \ldots, \hat{Y}_k^n = f(w_k, I^n)\}$$

where $\mathcal{P}(S)$ is the set of all permutations of the ordered set S. Furthermore, the average probability of error for decoder $g^{(k)}$ is given by

$$\mathcal{P}_e = \frac{1}{J(n,k)} \sum_{(w_1,\ldots,w_k) \in \mathcal{F}(n,k)} \mathcal{P}_e(w_1, \ldots, w_k)$$

and is equal to $\Pr\{(\hat{W}_1, \ldots, \hat{W}_k) \notin \mathcal{P}(W_1, \ldots, W_k)\}$ when the fingerprint index vector $(W_1, \ldots, W_k)$ is uniformly distributed in $\mathcal{F}(n,k)$.

Definition 3.3 *For a* $(2^{nR_Q}, 2^{nR_F}, n)$ *quantization/fingerprinting code, the average (per-symbol) distortion is given by*

$$\bar{\mathcal{D}} = E[n^{-1} \sum_{j=1}^{n} d(I_j, f(W, I^n)_j)]$$

assuming that W *is uniformly distributed in* $\{1, \ldots, 2^{nR_F}\}$.

Definition 3.4 *A rate pair* (R_Q, R_F) *is achievable for distortion constraint* D, *if there exists a sequence of quantization/fingerprinting codes* $(2^{nR_Q}, 2^{nR_F}, n)$ *such that*

$$\bar{\mathcal{D}} \leq D \tag{3.2}$$

$$\max_{w_1,\ldots,w_k} \mathcal{P}_e(w_1, \ldots, w_k) \to 0 \quad as \quad n \to \infty \tag{3.3}$$

Moreover, a rate region $\mathcal{R}$ *of pairs* (R_Q, R_F) *is achievable if every element of* $\mathcal{R}$ *is achievable.*

We can now state the following theorem.

Theorem 3.1 *A private quantization/fingerprinting code* $(2^{nR_Q}, 2^{nR_F}, n)$ *satisfies the transparency and robustness requirements (3.2) and (3.3), respectively, if and only if* $(R_Q, R_F) \in \mathcal{R}_D^{\text{dsc, F}}$, *where*

$$\mathcal{R}_D^{\text{dsc, F}} = \Bigg\{ (R_Q, R_F) : \\ R_Q \geq \min_{p_{\hat{Y}|I}: Ed(I,\hat{Y}) \leq D} I(\hat{Y}; I) \\ R_F \leq \max_{p_{\hat{Y}|I}: Ed(I,\hat{Y}) \leq D} \min \Big\{ R_Q - I(I; \hat{Y}), \\ \min_{1 \leq l \leq k} \min_{S_l \subset \{1,\ldots,k\}} \frac{1}{l} I(Z; \{\hat{Y}_s\}_{s \in S_l} | \{\hat{Y}_t\}_{t \in \bar{S}_l}) \Big\} \Bigg\}$$

where S_l is any subset of $\{1,\ldots,k\}$ with l elements, and $\bar{S}_l = \{1,\ldots,k\} - S_l$. For the computation of $I(Z;\{\hat{Y}_s\}_{s\in S_l}|\{\hat{Y}_t\}_{t\in\bar{S}_l})$, the conditional probabilities $p_{\hat{Y}_1,\ldots,\hat{Y}_k|I} = (p_{\hat{Y}|I})^n$ and $p^{(k)}_{Z|\hat{Y}_1,\ldots,\hat{Y}_k}$ are used.

The proof of Theorem 3.1 can be found in Section 3.2.

We have also considered a continuous-alphabet, Gaussian analogue of Theorem 3.1. In particular, if we assume that (i) all (single-letter) alphabets are equal to $\mathbf{R}$, (ii) the image I^n is i.i.d. Gaussian with variance P_I, (iii) the distortion function is the squared-error (i.e., (2.4)) and (iv) the output of the attack channel is $Z^n = \sum_{l=1}^{k} \lambda_{k,l}\hat{Y}_l^n + V^n$ where $\lambda_{k,l}$ are scalar, fixed quantities and V^n is i.i.d. Gaussian with variance P_V, then Theorem 3.1 becomes:

Theorem 3.2 *(Gaussian case) A private, continuous alphabet quantization/ fingerprinting code $(2^{nR_Q}, 2^{nR_F}, n)$ satisfies requirements (3.2) and (3.3), if and only if $(R_Q, R_F) \in \mathcal{R}_D^{\text{gauss, F}}$, where*

$$\mathcal{R}_D^{\text{gauss, F}} = \Bigg\{ (R_Q, R_F) : \\ R_Q \geq \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+ \\ R_F \leq \max_{\gamma\in\left[\max\left\{1,\frac{P_I}{D}\right\},2^{2R_Q}\right]} \min\Bigg\{ R_Q - \frac{1}{2}\log(\gamma), \\ \min_{1\leq l\leq k} \min_{S_l\subset\{1,\ldots,k\}} \frac{1}{2l}\log\left(1 + \frac{\sum_{s\in S_l}\lambda_{k,s}^2 P_W(\gamma)}{P_V}\right)\Bigg\}\Bigg\}$$

where $P_W(\gamma)$ was defined in (2.6), and S_l in Theorem 3.1.

It is trivial to show that in the special case $\lambda_{k,l} = \lambda_k$ for all l, the region $\mathcal{R}_D^{\text{gauss, F}}$ becomes

$$\mathcal{R}_D^{\text{gauss, F}} = \Bigg\{ (R_Q, R_F) :$$

$$R_Q \geq \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+$$

$$R_F \leq \max_{\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2k}\log\left(1 + \frac{k\lambda_k^2 P_W(\gamma)}{P_V}\right)\right\}\Bigg\}$$

The proof of Theorem 3.2 can be found in Section 3.3.

3.1.2 A Simple Optimization of the Gaussian Collusion Attack

Adopting a conservative approach, and consistent with the game formulation of Chapter 2, we can assume that the colluders know the statistics of the embedding strategy. Namely, they know the joint distribution $p_{I^n,\hat{Y}^n}$. Therefore, they might wish to optimize the attack in terms of the parameters $\{\lambda_{k,l}\}_{l=1}^k, P_V$, such that a distortion constraint is satisfied. One such distortion constraint is the following:

$$(\forall\ l \leq k) \quad \frac{1}{n}E||\hat{Y}_l^n - Z^n||^2 \leq D_A \tag{3.4}$$

In other words, the forgery should look similar to *every* fingerprinted copy in the collusion. Such a requirement is quite reasonable, assuming that the colluders want to be fair to each other. The ultimate goal of the attackers is to minimize the achievable region $\mathcal{R}_D^{\text{gauss, F}}$ subject to the above distortion constraint. This can be done by minimizing

$$r_F(\gamma, k, \boldsymbol{\lambda}_k, P_V) \stackrel{\Delta}{=} \min_{1\leq l\leq k} \min_{S_l \subset \{1,\ldots,k\}} \frac{1}{2l}\log\left(1 + \frac{\sum_{s\in S_l}\lambda_{k,s}^2 P_W(\gamma)}{P_V}\right) \tag{3.5}$$

with respect to $\boldsymbol{\lambda}_k \stackrel{\Delta}{=} (\lambda_{k,1}, \ldots, \lambda_{k,k})$ and P_V, such that (3.4) is satisfied.

Assuming that k and P_V are fixed, we make the following observations:

1. For every $l \leq k$, the region $\boldsymbol{\Lambda}_{k,l}$ of allowable $\boldsymbol{\lambda}_k$ such that the constraint $n^{-1}E||\hat{Y}_l^n - Z^n||^2 \leq D_A$ is satisfied, is a convex set. Therefore, the region

$\boldsymbol{\Lambda}_k = \cap_{l=1}^{k} \boldsymbol{\Lambda}_{k,l}$ (which contains all $\boldsymbol{\lambda}_k$ that satisfy (3.4)) is convex. Hence, for any $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_m \in \boldsymbol{\Lambda}_k$, we have $\frac{\boldsymbol{\kappa}_1 + \cdots + \boldsymbol{\kappa}_m}{m} \in \boldsymbol{\Lambda}_k$.

2. The set of constraints in (3.4) is symmetric with respect to the $\lambda_{k,l}$'s. Therefore, if $\boldsymbol{\lambda}_k \in \boldsymbol{\Lambda}_k$, then any permutation of the elements of $\boldsymbol{\lambda}_k$ will also lie in $\boldsymbol{\Lambda}_k$. That is, $\mathcal{P}(\boldsymbol{\lambda}_k) \in \boldsymbol{\Lambda}_k$, where $\mathcal{P}(\boldsymbol{\lambda}_k)$ is the set that contains all distinct permutations of $\boldsymbol{\lambda}_k$.

3. For every $m \geq 1$, the function $r(\lambda_1, \ldots, \lambda_m) \stackrel{\Delta}{=} \lambda_1^2 + \cdots + \lambda_m^2$ is convex. Therefore,

$$r(\lambda_1, \ldots, \lambda_m) \geq r\left(\frac{\lambda_1 + \cdots + \lambda_m}{m}, \ldots, \frac{\lambda_1 + \cdots + \lambda_m}{m}\right) \tag{3.6}$$

with equality if and only if $\lambda_1 = \cdots = \lambda_m$.

Based on the observations above, we have the following theorem.

Theorem 3.3 *The value of $\boldsymbol{\lambda}_k$ which minimizes $r_F(\gamma, k, \boldsymbol{\lambda}_k, P_V)$ subject to the constraint (3.4), satisfies $\boldsymbol{\lambda}_k^* = \lambda_k^*(1, \ldots, 1)$ for some scalar λ_k^*.*

Proof: Let us assume that $\boldsymbol{\lambda}_k^* = (\lambda_{k,1}^*, \ldots, \lambda_{k,k}^*) \in \boldsymbol{\Lambda}_k$, and let P_V^* minimize r_F. We distinguish between the following cases:

- The minimum of r_F equals

$$\frac{1}{2l} \log\left(1 + \frac{\sum_{s \in S_l} \lambda_{k,s}^2 P_W(\gamma)}{P_V}\right)$$

for some S_l, and coefficients $\lambda_{k,s} = \lambda_k^*$ for all $s \in S_l$. Without loss of generality, we can take $\boldsymbol{\lambda}_k^* = \lambda_k^*(1, \ldots, 1)$.

- Suppose now that for $t, s \in S_l$ (where S_l attains the minimum in (3.5)) we have $\lambda_{k,s}^* \neq \lambda_{k,t}^*$. By switching the s^{th} and t^{th} elements of $\boldsymbol{\lambda}_k^*$, we obtain $\boldsymbol{\lambda}_k'$

which, by Observation 2, lies in $\mathbf{\Lambda}_k$. From Observation 1, $\boldsymbol{\kappa}_k^* \stackrel{\Delta}{=} \frac{\boldsymbol{\lambda}_k^* + \boldsymbol{\lambda}_k'}{2} \in \mathbf{\Lambda}_k$, and from Observation 3 we have:

$$\sum_{m \in S_l} (\lambda_{k,m}^*)^2 > \sum_{m \in S_l} (\kappa_{k,m}^*)^2$$

where we have strict inequality because $\lambda_{k,s}^* \neq \lambda_{k,t}^*$. Hence, $r_F(\gamma, k, \boldsymbol{\kappa}_k^*, P_V^*) < r_F(\gamma, k, \boldsymbol{\lambda}_k^*, P_V^*)$ (contradiction). Therefore, all elements of $\boldsymbol{\lambda}_k^*$ should be equal.

The proof of the theorem is thus concluded. ■

From the above, we obtain that the optimal choice (with respect to the colluders' point of view) is $\lambda_{k,l} = \lambda_k, \ \forall\ l \leq k$. Then, (3.4) becomes

$$\begin{aligned}(k-1)(k-2)\lambda_k^2(\gamma-1)P_W(\gamma) + (k-1)\lambda_k^2\gamma P_W(\gamma) \\ + 2(k-1)\lambda_k(\lambda_k-1)(\gamma-1)P_W(\gamma) + (\lambda_k-1)^2\gamma P_W(\gamma) + P_V \leq D_A \quad (3.7)\end{aligned}$$

where $D_A \geq P_W(\gamma)\left(1 - \frac{1}{k}\right)$. Also, the rate region $\mathcal{R}_D^{\text{gauss, F}}$, optimized for (3.7), becomes:

$$\begin{aligned}\mathcal{R}_{D,D_A}^{\text{gauss, F}} = \Bigg\{ & (R_Q, R_F) : \\ & R_Q \geq \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right)\right]^+ \\ & R_F \leq \max_{\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]} \min\Bigg\{ R_Q - \frac{1}{2}\log(\gamma), \\ & \qquad \min_{\{\lambda_k, P_V\}:\ (3.7) \text{ is satisfied}} \frac{1}{2k}\log\left(1 + \frac{k\lambda_k^2 P_W(\gamma)}{P_V}\right)\Bigg\}\Bigg\}\end{aligned}$$

For computing the optimal values of λ_k, P_V, an analysis similar to the one in Section 2.5 can be carried out here. It can then be shown that the optimal values

are

$$\lambda_k^* = \begin{cases} 0 & \text{if } D_A > \gamma P_W(\gamma) \\ \frac{\gamma P_W(\gamma) - D_A}{P_W(\gamma)(1+(\gamma-1)k)} & \text{otherwise} \end{cases}$$

and

$$P_V^* = ((D_A - P_W(\gamma))k + P_W(\gamma))\lambda_k^*$$

Therefore,

$$\mathcal{R}_{D,D_A}^{\text{gauss, F}} = \Bigg\{ (R_Q, R_F) : R_Q \geq \left[\frac{1}{2} \log \left(\frac{P_I}{D} \right) \right]^+$$
$$R_F \leq \max_{\gamma \in \Gamma(R_Q, D, D_A)} \min \Bigg\{ R_Q - \frac{1}{2}\log(\gamma),$$
$$\frac{1}{2k} \log \left(1 + \frac{k(\gamma P_W(\gamma) - D_A)}{(1+(\gamma-1)k)((D_A - P_W(\gamma))k + P_W(\gamma))} \right) \Bigg\} \Bigg\}$$

where $\Gamma(R_Q, D, D_A)$ was defined in (2.13).

Figure 3.2 shows $\mathcal{R}_{D,D_A}^{\text{gauss, F}}$ for different values of k, when $P_I = 150$ and $D = D_A = 50$. We can immediately see that, as the number of colluders increases, the rate region shrinks. Therefore, collusion can be a very effective means of attack. Note that for $k = 1$, the rate region shown is the one achieved under the optimum Gaussian attack (2.94), which is consistent with the values of λ_1^* and P_V^* derived above.

3.1.3 A Multi-User Costa Scheme

Thus far, all fingerprinting systems considered in this chapter operate in a private scenario. We also present a public scenario which is analogous to Costa's formulation [15] for the multi-access case. More precisely, assume that the output of the

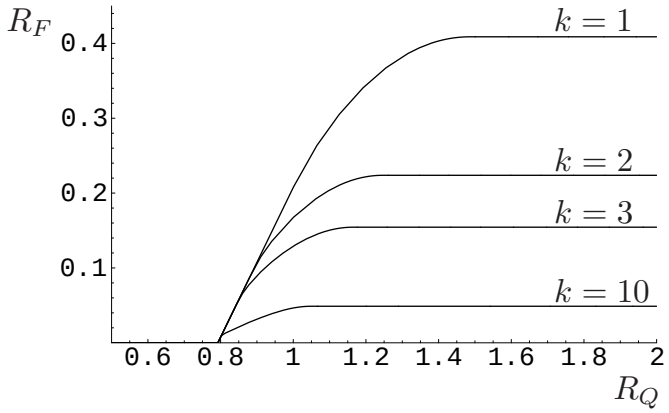


Figure 3.2: The rate region achieved under an optimized Gaussian attack, for different values of k (number of colluders).

encoder is $Y_l^n = f(W_l, I^n) = I^n + X^n(W_l, I^n)$ (without quantization) such that a squared-error distortion constraint is met. That is,

$$(\forall\ l)\ \ n^{-1}E||I^n - Y_l^n||^2 = n^{-1}E||X_l^n||^2 \leq D \tag{3.8}$$

The k colluders combine their fingerprinted images linearly and adding i.i.d. Gaussian noise, i.e.,

$$Z^n = \sum_{l=1}^{k} \lambda_k Y_l^n + V^n$$

The detector receives Z^n and attempts to detect all $W_1, \ldots, W_k$ without any knowledge of I^n. As it turns out, using an encoding technique similar to Costa's in [15], we can achieve any rate R_F such that

$$R_F \leq \frac{1}{2k} \log\left(1 + \frac{\lambda_k^2 k D}{P_V}\right)$$

which is precisely the maximum rate achievable in $\mathcal{R}_D^{\text{gauss, F}}$ when $R_Q = \infty$. The proof of this result can be found in Section 3.4.

3.2 Proof of Theorem 3.1

As we saw in Section 3.1, the W_l's are uniformly distributed in $\{1, \ldots, J_n\}$, but not independent. From (3.1), we have

$$\begin{aligned}
&\frac{1}{J(n,k)} \sum_{(w_1,\ldots,w_k)\in\mathcal{F}(n,k)} \mathcal{P}_e(w_1,\ldots,w_k) \\
&\quad \doteq \frac{1}{(J_n)^k} \sum_{(w_1,\ldots,w_k)\in\mathcal{F}(n,k)} \mathcal{P}_e(w_1,\ldots,w_k) \\
&\quad \leq \frac{1}{(J_n)^k} \sum_{(w_1,\ldots,w_k)\in(\mathcal{F}_n)^k} \mathcal{P}_e(w_1,\ldots,w_k)
\end{aligned}$$

The last expression is the average probability of error resulting by choosing the W_l's in an i.i.d. fashion (in which case they will not necessarily be distinct). From the above inequality it suffices to show that this probability of error is asymptotically vanishing.

By virtue of the above bound, we assume from now on, and in all the proofs of this chapter, that $(W_1, \ldots, W_k)$ are independent and uniformly distributed in $\{1, \ldots, J_n\}$. Hence, given I^n, the random variables $\hat{Y}_l^n = f(W_l, I^n)$ are independent (since the W_l's are independent and $\hat{Y}^n$ is a function of W and I^n). Thus, we have the Markov condition

$$(\forall\ l \neq m) \quad \hat{Y}_l^n \to I^n \to \hat{Y}_m^n \tag{3.9}$$

and therefore:

$$p_{\hat{Y}_1^n,\ldots,\hat{Y}_k^n|I^n} = \prod_{l=1}^{k} p_{\hat{Y}_l^n|I^n} = (p_{\hat{Y}^n|I^n})^k$$

where the last equality is due to the fact that all $\hat{Y}_l^n$'s are outputs of the *same* function f.

Converse Theorem

We first prove the following lemma, which is a variant of Fano's inequality [33].

Lemma 3.1 *Let $\hat{W}_l \in \{1, \ldots, J_n\}$ for all $1 \leq l \leq k$. Then*

$$H(W_1|\hat{W}_1, \ldots, \hat{W}_k) \leq \log(k+1) + \mathcal{P}_e n R_F$$

where $\mathcal{P}_e = \Pr\{(\hat{W}_1, \ldots, \hat{W}_k) \notin \mathcal{P}(W_1, \ldots, W_k)\}$, and $\mathcal{P}(S)$ is the set of all permutations of an ordered set S.

Proof: First, we define an "error" variable T, such that

$$T = \begin{cases} 1 & \text{if } \hat{W}_1 = W_1 \\ 2 & \text{if } \hat{W}_2 = W_1 \\ \vdots & \vdots \\ k & \text{if } \hat{W}_k = W_1 \\ 0 & \text{otherwise} \end{cases}$$

Then, we have

$$\begin{aligned} H(W_1, T|\hat{W}_1, \ldots, \hat{W}_k) &= H(W_1|\hat{W}_1, \ldots, \hat{W}_k, T) + H(T|\hat{W}_1, \ldots, \hat{W}_k) \\ &= H(W_1|\hat{W}_1, \ldots, \hat{W}_k) + H(T|\hat{W}_1, \ldots, \hat{W}_k, W_1) \end{aligned}$$

Observe that $H(T|\hat{W}_1, \ldots, \hat{W}_k, W_1) = 0$. Also, $H(T|\hat{W}_1, \ldots, \hat{W}_k) \leq H(T) \leq \log(k+1)$. Hence, we have

$$H(W_1|\hat{W}_1, \ldots, \hat{W}_k) \leq H(W_1|\hat{W}_1, \ldots, \hat{W}_k, T) + \log(k+1) \tag{3.10}$$

We now have

$$\begin{aligned} H(W_1|\hat{W}_1, \ldots, \hat{W}_k, T) &= \sum_{l=0}^{k} \Pr\{T = l\} H(W_1|\hat{W}_1, \ldots, \hat{W}_k, T = l) \\ &= \Pr\{T = 0\} H(W_1|\hat{W}_1, \ldots, \hat{W}_k, T = 0) \qquad (3.11) \\ &\leq \mathcal{P}_e n R_F \qquad (3.12) \end{aligned}$$

where (3.11) holds because $H(W_1|\hat{W}_1, \ldots, \hat{W}_k, T = l) = 0$ for all $l \neq 0$. Thus, combining (3.10) and (3.12), the lemma is proved. ■

The converse theorem states that any rate pair (R_Q, R_F) that satisfies constraints (3.2) and (3.3) must lie in $\mathcal{R}_D^{\text{dsc, F}}$.

Proof: Let $\epsilon > 0$. We assume that $\Pr\{(\hat{W}_1, \ldots, \hat{W}_k) \notin \mathcal{P}(W_1, \ldots, W_k)\} < \epsilon$ and that the distortion constraint is met with equality:

$$\frac{1}{n} \sum_{j=1}^{n} Ed(I_j, \hat{Y}_{l,j}) = D \tag{3.13}$$

where $\hat{Y}_l^n = (\hat{Y}_{l,1}, \ldots, \hat{Y}_{l,n})$ is the fingerprinted image of the l-th user. Since all fingerprinted copies are generated through the same encoder f, it follows that

$$(\forall\ l)\ \ p_{\hat{Y}_l^n | I^n} = p_{\hat{Y}^n | I^n}$$

and the Markov conditions (3.9) are satisfied.

As we saw in Section 2.2, the lower bound on R_Q in the definition of $\mathcal{R}_D^{\text{dsc, F}}$ can be established using a standard argument from rate-distortion theory [33].

For establishing the upper bound on R_F, we have:

$$\begin{aligned}
R_F &= n^{-1} H(W_1) \\
&= n^{-1} H(W_1 | I^n) && (3.14) \\
&= n^{-1} I(W_1; \hat{Y}_1^n | I^n) + n^{-1} H(W_1 | \hat{Y}_1^n, I^n) && (3.15) \\
&\leq n^{-1} I(W_1; \hat{Y}_1^n | I^n) + n^{-1} H(W_1 | Z^n, I^n) && (3.16) \\
&\leq n^{-1} I(W_1; \hat{Y}_1^n | I^n) + n^{-1} H(W_1 | \hat{W}_1, \ldots, \hat{W}_k) && (3.17) \\
&\leq n^{-1} I(W_1; \hat{Y}_1^n | I^n) + \epsilon && (3.18) \\
&= n^{-1} H(\hat{Y}_1^n | I^n) - n^{-1} H(\hat{Y}_1^n | I^n, W_1) + \epsilon \\
&= n^{-1} H(\hat{Y}_1^n | I^n) + \epsilon && (3.19) \\
&= n^{-1} H(\hat{Y}_1^n) - n^{-1} (H(\hat{Y}_1^n) - H(\hat{Y}_1^n | I^n)) + \epsilon \\
&\leq R_Q - n^{-1} I(\hat{Y}_1^n; I^n) + \epsilon && (3.20)
\end{aligned}$$

$$
\begin{aligned}
&= R_Q - H(I) + n^{-1} H(I^n|\hat{Y}_1^n) + \epsilon \\
&\leq R_Q - H(I) + n^{-1} \sum_{j=1}^{n} H(I_j|\hat{Y}_{1,j}) + \epsilon \qquad (3.21) \\
&= R_Q - n^{-1} \sum_{j=1}^{n} I(I_j; \hat{Y}_{1,j}) + \epsilon \\
&= R_Q - n^{-1} \sum_{j=1}^{n} I(I_j; \hat{Y}_j) + \epsilon \qquad (3.22)
\end{aligned}
$$

where (3.14) is true because I^n is independent of W_1; (3.16) follows from the Markov chain $W_1 \rightarrow (\hat{Y}_1^n, I^n) \rightarrow (Z^n, I^n)$; (3.17) holds because $H(W_1|Z^n, I^n) = H(W_1|g^{(k)}(Z^n, I^n), Z^n, I^n) \leq H(W_1|g^{(k)}(Z^n, I^n)) = H(W_1|\hat{W}_1, \ldots, \hat{W}_k)$; (3.18) follows from Lemma 3.1; (3.19) holds because $\hat{Y}_1^n$ is a deterministic function of I^n, W_1, (3.20) follows from $R_Q \geq n^{-1} H(\hat{Y}_1^n)$ and (3.21) is due to the inequalities $H(I^n|\hat{Y}_1^n) \leq \sum_{i=1}^{n} H(I_j|\hat{Y}_1^n) \leq \sum_{j=1}^{n} H(I_j|\hat{Y}_{1,j})$.

For l with $1 \leq l \leq k$, we denote by S_l any subset of $\{1, \ldots, k\}$ such that $|S_l| = l$. Let $S_l = \{s_1, \ldots, s_l\}$. We use the following notation:

$$
W_{S_l} \stackrel{\Delta}{=} (W_{s_1}, \ldots, W_{s_l}), \qquad \hat{Y}_{S_l}^n \stackrel{\Delta}{=} (\hat{Y}_{s_1}^n, \ldots, \hat{Y}_{s_l}^n)
$$

and

$$
\hat{Y}_{S_l,j} \stackrel{\Delta}{=} (\hat{Y}_{s_1,j}, \ldots, \hat{Y}_{s_l,j})
$$

where $\hat{Y}_{s,j}$ is the j-th element of $\hat{Y}_s^n$. Also, we denote $\bar{S}_l = \{1, \ldots, k\} - S_l$.

For each $1 \leq l \leq k$, and for each S_l (as defined above), we obtain the following chain of inequalities:

$$
\begin{aligned}
lR_F &= n^{-1} H(W_{S_l}|W_{\bar{S}_l}, I^n) \qquad (3.23) \\
&= n^{-1} I(W_{S_l}; Z^n|W_{\bar{S}_l}, I^n) + n^{-1} H(W_{S_l}|W_{\bar{S}_l}, I^n, Z^n) \\
&\leq n^{-1} I(W_{S_l}; Z^n|W_{\bar{S}_l}, I^n) + n^{-1} \sum_{m=1}^{l} H(W_{s_m}|Z^n, I^n)
\end{aligned}
$$

$$
\begin{aligned}
&\leq n^{-1} I(W_{S_l}; Z^n | W_{\bar{S}_l}, I^n) + l\epsilon && (3.24)\\
&= n^{-1} H(Z^n | W_{\bar{S}_l}, I^n) - n^{-1} H(Z^n | W_1, \ldots, W_k, I^n) + l\epsilon \\
&= n^{-1} H(Z^n | \hat{Y}_{\bar{S}_l}, I^n) - n^{-1} H(Z^n | \hat{Y}_1^n, \ldots, \hat{Y}_k^n) + l\epsilon && (3.25)\\
&\leq n^{-1} \sum_{j=1}^{n} H(Z_j | \hat{Y}_{\bar{S}_l, j}, I_j) \\
&\quad - n^{-1} \sum_{j=1}^{n} H(Z_j | Z_1, \ldots, Z_{j-1}, \hat{Y}_1^n, \ldots, \hat{Y}_k^n) + l\epsilon && (3.26)\\
&= n^{-1} \sum_{j=1}^{n} H(Z_j | \hat{Y}_{\bar{S}_l, j}, I_j) - n^{-1} \sum_{j=1}^{n} H(Z_j | \hat{Y}_{1,j}, \ldots, \hat{Y}_{k,j}) + l\epsilon && (3.27)\\
&\leq n^{-1} \sum_{j=1}^{n} (H(Z_j | \hat{Y}_{\bar{S}_l, j}, I_j) - H(Z_j | \hat{Y}_{1,j}, \ldots, \hat{Y}_{k,j}, I_j) + l\epsilon && (3.28)\\
&= n^{-1} \sum_{j=1}^{n} I(Z_j; \hat{Y}_{S_l, j} | \hat{Y}_{\bar{S}_l, j}, I_j) + l\epsilon && (3.29)
\end{aligned}
$$

where (3.23) is due to the independence of I^n and $W_1, \ldots, W_k$; (3.24) follows from Lemma 3.1; (3.25) holds because of the Markov chains $(I^n, W_1, \ldots, W_k) \to (\hat{Y}_1^n, \ldots, \hat{Y}_k^n) \to Z^n$ and $(\hat{Y}_1^n, \ldots, \hat{Y}_k^n) \to (I^n, W_1, \ldots, W_k) \to Z^n$; (3.26) follows from the chain rule for the entropy; (3.27) holds because the attack channel is memoryless and therefore given $(\hat{Y}_{1,j}, \ldots, \hat{Y}_{k,j})$, the variable Z_j is conditionally independent of everything else and (3.28) follows because conditioning reduces entropy.

Note that (3.29) is true for all $1 \leq l \leq k$ and all $S_l \subset \{1, \ldots, k\}$. Hence, together with (3.22), we obtain

$$
\begin{aligned}
R_F \leq \min \Bigg\{ & R_Q - \frac{1}{n} \sum_{j=1}^{n} I(I_j; \hat{Y}_j), \\
& \min_{1 \leq l \leq k} \min_{S_l \subset \{1, \ldots, k\}} \frac{1}{ln} \sum_{j=1}^{n} I(Z_j; \hat{Y}_{S_l, j} | \hat{Y}_{\bar{S}_l, j}, I_j) \Bigg\} + \epsilon
\end{aligned} \tag{3.30}
$$

Using a similar approach as in Section 2.2, we can argue that, since $I(I_j; \hat{Y}_j)$ and

$I(Z_j; \hat{Y}_{S_l,j} | \hat{Y}_{\bar{S}_l,j}, I_j)$ are concave with respect to $p_{\hat{Y}_j|I_j}$, the following are true:

$$n^{-1} \sum_{j=1}^{n} (R_Q - I(I_j; \hat{Y}_j)) \leq R_Q - I_a(I; \hat{Y}) \tag{3.31}$$

$$n^{-1} \sum_{j=1}^{n} I(Z_j; \hat{Y}_{S_l,j} | \hat{Y}_{\bar{S}_l,j}, I_j) \leq I_a(Z; \hat{Y}_{S_l} | \hat{Y}_{\bar{S}_l}, I) \tag{3.32}$$

where the mutual information expressions on the right-hand side of (3.31) and (3.32) are computed with respect to the pmf $p^a_{\hat{Y}|I} = n^{-1} \sum_{j=1}^{n} p_{\hat{Y}_j|I_j}$ that was defined in (2.35). The remainder of the proof follows in the same way as in the converse in Section 2.2, and will be omitted.

Note: The same result can be obtained if, instead of using $p^a_{\hat{Y}|I}$, we introduce a time-sharing variable Q uniformly distributed over $\{1, \ldots, n\}$ on the conditions side of each mutual information functional. In that case, the final expression involves maximization with respect to the joint distribution $p_{\hat{Y}|I,Q}\, p_{Q|I}\, p_I$. The rate region obtained is the same as the region $\mathcal{R}_D^{\text{dsc, F}}$ described by the expression in the statement of the theorem. ■

Direct Theorem

We now show that $\mathcal{R}_D^{\text{dsc, F}}$ is achievable.

Proof: As usual (cf. proofs of Theorems 2.1, 2.3), we limit the quantization rate to $R_Q \geq r_q(D)$.

We present an outline of the proof here. Many of the details are quite straightforward, or come directly from proofs presented elsewhere.

We assume that the indices $W_1, \ldots, W_k$ are uniformly distributed in $\{1, \ldots, J_n\}$. Furthermore, we use a random coding argument and we finally establish that there exists a deterministic code that achieves arbitrarily small probability of error.

Codebook Generation: The codebook generation is identical to the one given in the direct part of Theorem 2.1. A set of 2^{nR_Q} sequences $\tilde{Y}^n$ is generated i.i.d. according to a pmf $p_{\hat{Y}}$, and then the set is partitioned uniformly into 2^{nR_F} subsets.

Fingerprint Embedding: The embedding is again identical to the procedure described in the proof of Theorem 2.1. Given I^n and a fingerprint index w, the encoder outputs $\hat{Y}^n(w)$, as determined by joint typicality with respect to some probability distribution $p_{\hat{Y}|I}$ which satisfies

$$n^{-1}Ed(I^n, \hat{Y}^n) \leq D$$

Decoding: The decoder receives Z^n (generated from $\hat{Y}^n(W_1), \ldots, \hat{Y}^n(W_k)$). In the sequel, we will refer to $\hat{Y}^n(W_l)$ as $\hat{Y}_l^n$. The decoder then seeks a k-tuple $(\hat{Y}^n(\hat{w}_1), \ldots, \hat{Y}^n(\hat{w}_k))$ such that $(I^n, \hat{Y}^n(\hat{w}_1), \ldots, \hat{Y}^n(\hat{w}_k), Z^n)$ belongs to a set $T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)$, the set of typical k-tuples with respect to the distribution $p_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}$ $= p_{Z|\hat{Y}_1,\ldots,\hat{Y}_k}(p_{\hat{Y}|I})^k p_I$ (observe that the last equality is due to the Markov conditions (3.9)). If a unique set of indices $\{\hat{w}_1, \ldots, \hat{w}_k\}$ exists (their ordering is immaterial here) then the decoder outputs it, otherwise it declares an error.

Probability of Error: Without loss of generality, we assume that $W_1 = 1, \ldots, W_k = k$. Consistent with the proof of Theorem 2.1, we again have three kinds of error events:

(i) E_1: I^n is not represented well (i.e., in terms of the distortion constraint) by at least one of $\hat{Y}_1^n, \ldots, \hat{Y}_k^n$.

(ii) E_2: Assuming E_1^c (i.e., that E_1 did not occur), $(I^n, \hat{Y}^n(1), \ldots, \hat{Y}^n(k), Z^n) \notin T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)$.

(iii) E_3: Assuming $(E_1 \cup E_2)^c$, there exists a k-tuple $(w_1, \ldots, w_k) \notin \mathcal{P}(1, \ldots, k)$ such that $(I^n, \hat{Y}^n(w_1), \ldots, \hat{Y}^n(w_k), Z^n) \in T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)$.

As we proved in Section 2.2, the probability of event E_1 approaches zero as long

as

$$R_F \leq R_Q - I(I;\hat{Y}) - \epsilon \tag{3.33}$$

Moreover, it can be easily proved that the probability of E_2 goes to zero.

We can upper-bound the probability of the error event E_3 as follows:

$$\begin{aligned}
\Pr(E_3) &\leq \Pr\{\exists\ (w_1,\ldots,w_k) \neq (1,\ldots,k) : \\
&\quad (I^n, \hat{Y}^n(w_1),\ldots,\hat{Y}^n(w_k), Z^n) \in T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)\} \qquad (3.34)\\
&= \Pr\{(\exists\ S_l \subset \{1,\ldots,k\},\ 1 \leq l \leq k)\ \wedge\ (\exists\ (w_1,\ldots,w_k)\ : \\
&\quad (w_s \neq s,\ \forall\ s \in S_l)\ \wedge\ (w_t = t,\ \forall\ t \in \bar{S}_l)) : \\
&\quad (I^n, \hat{Y}^n(w_1),\ldots,\hat{Y}^n(w_k), Z^n) \in T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)\} \\
&\leq \sum_{l=1}^{k} \sum_{S_l \subset \{1,\ldots,k\}} \sum_{\substack{(w_1,\ldots,w_k):\\ w_s \neq s,\ \forall\, s \in S_l\\ w_t = t,\ \forall\, t \in \bar{S}_l}} 1 \times \\
&\quad \times \Pr\{(I^n, \hat{Y}^n(w_1),\ldots,\hat{Y}^n(w_k), Z^n) \in T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)\} \qquad (3.35)
\end{aligned}$$

where the right-hand side of (3.34) is clearly an upper bound on $\Pr(E_3)$, since there are fingerprint index combinations (e.g., when $(w_1,\ldots,w_k)$ is a permutation of $(1,\ldots,k)$) that do not lead to error. Also, all probabilities are computed under the condition that Z^n is the output of the channel whose input is $\hat{Y}^n(1),\ldots,\hat{Y}^n(k)$. Moreover, S_l and $\bar{S}_l$ are defined similarly as in the converse part; S_l denotes any subset of $\{1,\ldots,k\}$ that has l elements and $\bar{S}_l = \{1,\ldots,k\} - S_l$.

Because of the symmetry of the random code and because Z^n is independent of $\hat{Y}^n(w), \forall\ w > k$ given I^n, a standard argument similar to the one used in the proof of the direct part of Theorem 14.3.1 in [33], gives:

$$\Pr\{(I^n, \hat{Y}^n(w_1),\ldots,\hat{Y}^n(w_k), Z^n) \in T^n_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}(\epsilon)\} \leq 2^{-n(I(Z;\{\hat{Y}_s\}_{s\in S_l}|\{\hat{Y}_t\}_{t\in\bar{S}_l})-o(1))}$$

for all $w_s \neq s,\ w_t = t$ with $s \in S_l$ and $t \in \bar{S}_l$. Note that $o(1)$ approaches zero with

ϵ. Hence, (3.35) gives:

$$\Pr(E_3) \leq \sum_{l=1}^{k} \sum_{S_l \subset \{1,\ldots,k\}} (2^{nR_F})^l \; 2^{-n(I(Z;\{\hat{Y}_s\}_{s\in S_l}|\{\hat{Y}_t\}_{t\in \bar{S}_l})-o(1))} \tag{3.36}$$

Hence, (3.36) approaches zero as n goes to infinity provided

$$(\forall\; 1 \leq l \leq k)\,(\forall\; S_l \subset \{1,\ldots,k\}) \quad R_F \leq \frac{1}{l} I(Z;\{\hat{Y}_s\}_{s\in S_l}|\{\hat{Y}_t\}_{t\in \bar{S}_l}) - o(1)$$

or, equivalently, when

$$R_F \leq \min_{1\leq l\leq k} \; \min_{S_l \subset \{1,\ldots,k\}} \; \frac{1}{l} I(Z;\{\hat{Y}_s\}_{s\in S_l}|\{\hat{Y}_t\}_{t\in \bar{S}_l}) - o(1) \tag{3.37}$$

Finally, we combine (3.33) with (3.37) and maximizing with respect to all $p_{\hat{Y}|I}$ such that $Ed(I,\hat{Y}) \leq D$, we obtain the achivability of $\mathcal{R}_D^{\text{dsc, F}}$. The existence of a deterministic code can be proved using a standard expurgation argument. ■

3.3 Proof of Theorem 3.2

Converse Theorem

We begin with the converse part, which establishes that all $(2^{nR_Q}, 2^{nR_F}, n)$ codes which satisfy conditions (3.2) and (3.3) have rates $(R_Q, R_W) \in \mathcal{R}_D^{\text{gauss, F}}$.

Proof: Let $\epsilon > 0$. Similarly to the converse of the discrete case (Section (3.2)), here too we assume that the fingerprint indices $W_1, \ldots, W_k$ are independent, each one being uniformly distributed in $\{1, \ldots, 2^{nR_F}\}$, that $\mathcal{P}_e < \epsilon$, and that the distortion constraint is met with equality:

$$(\forall\; l) \quad \frac{1}{n} \sum_{j=1}^{n} E||I^n - \hat{Y}_l^n||^2 = D \tag{3.38}$$

The lower bound on R_Q is the standard rate-distortion function, and is trivially established.

For establishing the upper bound on R_F, we need to consider the L_2-space of Section 2.3. The quantities $\phi, \gamma, P_W(\gamma), \lambda_0$ have exactly the same meaning. Moreover, the upper bound

$$R_F \leq R_Q - \frac{1}{2}\log(\gamma) \tag{3.39}$$

follows from (3.22) and (2.47), so, no further discussion is needed here.

For each $1 \leq l \leq k$ and for each S_l (as defined in Section 3.2), we have from (3.25):

$$lR_F \leq n^{-1}h(Z^n|\hat{Y}_{\bar{S}_l}, I^n) - n^{-1}h(Z^n|\hat{Y}_1^n, \ldots, \hat{Y}_k^n) + l\epsilon \tag{3.40}$$

where we replaced the discrete entropy $H(\cdot)$ with the differential entropy $h(\cdot)$, since Z^n is a continuous random vector. Recall that $Z^n = \sum_{l=1}^{k} \lambda_{k,l}\hat{Y}_l^n + V^n$, where V^n is i.i.d. Gaussian of variance P_V.

From (3.40), we continue the chain of inequalities as follows:

$$\begin{aligned}
lR_F \leq\ & \\
& n^{-1}h\left(Z^n - \sum_{s\in\bar{S}_l}\lambda_{k,s}\hat{Y}_s^n \middle| \hat{Y}_{\bar{S}_l}, I^n\right) \\
& -n^{-1}h\left(Z^n - \sum_{l=1}^{k}\lambda_{k,l}\hat{Y}_l^n \middle| \hat{Y}_1^n, \ldots, \hat{Y}_k^n\right) + l\epsilon \\
=\ & n^{-1}h\left(\sum_{s\in S_l}\lambda_{k,s}\hat{Y}_s^n + V^n \middle| \hat{Y}_{\bar{S}_l}, I^n\right) - n^{-1}h(V^n) + l\epsilon && (3.41)\\
=\ & n^{-1}h\left(\sum_{s\in S_l}\lambda_{k,s}\hat{Y}_s^n + V^n \middle| I^n\right) - \frac{1}{2}\log(2\pi e)P_V + l\epsilon && (3.42)\\
=\ & n^{-1}h\left(\sum_{s\in S_l}\lambda_{k,s}(\hat{Y}_s^n - \Psi I^n) + V^n \middle| I^n\right) - \frac{1}{2}\log(2\pi e)P_V + l\epsilon && (3.43)\\
\leq\ & n^{-1}h\left(\sum_{s\in S_l}\lambda_{k,s}(\hat{Y}_s^n - \Psi I^n) + V^n\right) - \frac{1}{2}\log(2\pi e)P_V + l\epsilon && (3.44)
\end{aligned}$$

$$
\begin{aligned}
&\leq n^{-1}\sum_{j=1}^{n} h\left(\sum_{s\in S_l}\lambda_{k,s}(\hat{Y}_{s,j}-\Psi^{(j)}I^n)+V_j\right)-\frac{1}{2}\log(2\pi e)P_V+l\epsilon \\
&\leq n^{-1}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\left(E\left[\left(\sum_{s\in S_l}\lambda_{k,s}(\hat{Y}_{s,j}-\Psi^{(j)}I^n)\right)^2\right]+P_V\right) \\
&\quad -\frac{1}{2}\log(2\pi e)P_V+l\epsilon \qquad (3.45)\\
&\leq \frac{1}{2}\log(2\pi e)\left(\frac{1}{n}\sum_{j=1}^{n}E\left[\left(\sum_{s\in S_l}\lambda_{k,s}(\hat{Y}_{s,j}-\Psi^{(j)}I^n)\right)^2\right]+P_V\right) \\
&\quad -\frac{1}{2}\log(2\pi e)P_V+l\epsilon \qquad (3.46)\\
&= \frac{1}{2}\log\left(1+\frac{1}{P_V}\frac{1}{n}\sum_{j=1}^{n}E\left[\left(\sum_{s\in S_l}\lambda_{k,s}(\breve{Y}_{s,j}-\Psi^{(j)}I^n)\right)^2\right]\right)+l\epsilon \qquad (3.47)
\end{aligned}
$$

where (3.41) holds because V^n is independent of all other variables; (3.42) follows from the fact that, given I^n, $\hat{Y}_{S_l}$ is independent from $\hat{Y}_{\bar{S}_l}$. The Ψ used in (3.43) is a $n\times n$ matrix; the j-th row of Ψ is denoted $\Psi^{(j)}$ while $\Psi^{(j)}I^n$ is the best linear estimator of $\hat{Y}_{s,j}$ given I^n, and is the same for every s (since we know that all $\{\hat{Y}_s^n\}_{s=1}^k$ have the same statistics); (3.44) holds because conditioning reduces entropy; (3.45) is the Gaussian entropy upper bound and (3.46) is a consequence of Jensen's inequality. Finally, in (3.47), we assume that $(\{\breve{Y}_s^n\}_{s\in S_l}, I^n)$ are jointly zero-mean Gaussian, they have the same second moments as $(\{\hat{Y}_s^n\}_{s\in S_l}, I^n)$ and they satisfy the same Markov conditions (3.9):

$$
(\forall\ l\neq m)\quad \breve{Y}_l^n\rightarrow I^n\rightarrow\breve{Y}_m^n \qquad (3.48)
$$

It is easy to establish that $E[(\breve{Y}_{l,j}-\Psi^{(j)}I^n)(\breve{Y}_{m,j}-\Psi^{(j)}I^n)]=0$ for all $l\neq m$. Indeed,

$$
\begin{aligned}
&E[(\breve{Y}_{l,j}-\Psi^{(j)}I^n)(\breve{Y}_{m,j}-\Psi^{(j)}I^n)] \\
&\quad = E[E[(\breve{Y}_{l,j}-\Psi^{(j)}I^n)(\breve{Y}_{m,j}-\Psi^{(j)}I^n)|I^n]]
\end{aligned}
$$

$$
\begin{aligned}
&= E[E[\breve{Y}_{l,j}\ \breve{Y}_{m,j}|I^n] - \Psi^{(j)}I^n E[\breve{Y}_{l,j} + \breve{Y}_{m,j}|I^n] + (\Psi^{(j)}I^n)^2] \\
&= E[E[\breve{Y}_{l,j}|I^n]E[\breve{Y}_{m,j}|I^n] - \Psi^{(j)}I^n E[\breve{Y}_{l,j}|I^n] \\
&\quad -\Psi^{(j)}I^n E[\breve{Y}_{m,j}|I^n] + (\Psi^{(j)}I^n)^2] \qquad (3.49)
\end{aligned}
$$

because $E[\breve{Y}_{l,j}\ \breve{Y}_{m,j}|I^n] = E[\breve{Y}_{l,j}|I^n]E[\breve{Y}_{m,j}|I^n]$ from (3.48). Since $(\{\breve{Y}_s^n\}_{s\in S_l}, I^n)$ is Gaussian, we have that $E[\breve{Y}_{l,j}|I^n] = E[\breve{Y}_{m,j}|I^n] = \Psi^{(j)}I^n$. Substituting in (3.49), we obtain a value of zero.

Thus, (3.47) equals

$$
\begin{aligned}
&\frac{1}{2}\log\left(1 + \frac{1}{P_V}\sum_{s\in S_l}\lambda_{k,s}^2\left(\frac{1}{n}\sum_{j=1}^{n}E[(\breve{Y}_{s,j} - \Psi^{(j)}I^n)^2]\right)\right) + l\epsilon \\
&\leq \frac{1}{2}\log\left(1 + \frac{1}{P_V}\sum_{s\in S_l}\lambda_{k,s}^2\left(\frac{1}{n}\sum_{j=1}^{n}E[(\breve{Y}_{s,j} - \lambda_0 I_j)^2]\right)\right) + l\epsilon \qquad (3.50) \\
&= \frac{1}{2}\log\left(1 + \frac{\sum_{s\in S_l}\lambda_{k,s}^2 P_W(\gamma)}{P_V}\right) + l\epsilon \qquad (3.51)
\end{aligned}
$$

where λ_0 in (3.50) is as defined in Section 2.3; the inequality stems from the fact that $\lambda_0 I_j$ cannot be a better estimator of $\breve{Y}_{s,j}$ than $\Psi^{(j)}I^n$, hence the mean-square-error $E[(\breve{Y}_{s,j} - \lambda_0 I_j)^2]$ can only be higher; and (3.51) follows from the definition of $P_W(\gamma)$.

Thus, from (3.47) and (3.51) we have

$$
R_F \leq \frac{1}{2l}\log\left(1 + \frac{\sum_{s\in S_l}\lambda_{k,s}^2 P_W(\gamma)}{P_V}\right) + \epsilon \qquad (3.52)
$$

Finally, since (3.52) holds for all $1 \leq l \leq k$ and all $S_l \subset \{1, \ldots, k\}$, we have:

$$
R_F \leq \min_{1\leq l\leq k}\ \min_{S_l\subset\{1,\ldots,k\}}\ \frac{1}{2l}\log\left(1 + \frac{\sum_{s\in S_l}\lambda_{k,s}^2 P_W(\gamma)}{P_V}\right) + \epsilon
$$

Thus, (3.39) and maximization with respect to $\gamma \in \left[\max\left\{1, \frac{P_I}{D}\right\}, 2^{2R_Q}\right]$, yields the required result. ■

Direct Theorem

Proof: The proof of the direct part follows immediately from the direct part of Theorem 2.2 (Section 2.3) and the direct part of Theorem 3.1 (Section 3.2). More precisely: the fingerprint generation and embedding procedures are identical to those in Section 2.3, and the fingerprint detection as well as the computation of the probability of error follow from Section 3.2. What has to be determined is the joint probability distribution $p_{I,\hat{Y}_1,\ldots,\hat{Y}_k,Z}$ used by the joint typicality detector. As expected, this distribution is jointly zero-mean Gaussian, with covariance matrix

$$K_{I,\hat{Y}_1,\ldots,Z} =$$

$$\begin{bmatrix} P_I & \sqrt{(\gamma-1)P_I P_W(\gamma)} & \cdots & \sum_{l=1}^{k} \lambda_{k,l}\sqrt{(\gamma-1)P_I P_W(\gamma)} \\ \sqrt{(\gamma-1)P_I P_W(\gamma)} & \gamma P_W(\gamma) & \cdots & \lambda_{k,1}\gamma P_W(\gamma)+\sum_{l=2}^{k}\lambda_{k,l}(\gamma-1)P_W(\gamma) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{l=1}^{k} \lambda_{k,l}\sqrt{(\gamma-1)P_I P_W(\gamma)} & \lambda_{k,1}\gamma P_W(\gamma)+\sum_{l=2}^{k}\lambda_{k,l}(\gamma-1)P_W(\gamma) & \cdots & P_Z \end{bmatrix}$$

where $P_Z = \sum_{l=1}^{k} \lambda_{k,l}^2 \gamma P_W(\gamma) + \sum_{l \neq m} \lambda_{k,l}\lambda_{m,l}(\gamma-1)P_W(\gamma) + P_V$.

Arguments similar to the ones in Section 3.2 for the existence of deterministic codes can be given here, thus concluding the proof. ■

3.4 A Multi-User Costa Scheme

In this Section, we consider collusion attacks on fingerprinted images, under a *public* detection scenario. Furthermore, we assume no quantization.

As usual, I^n is an i.i.d. Gaussian image of per-symbol variance P_I. There are 2^{nR_F} fingerprint indices, each one corresponding to a particular customer. The information hider generates 2^{nR_F} fingerprinted copies of I^n, say $Y^n(1, I^n), \ldots, Y^n(2^{nR_F}, I^n)$, such that the following distortion constraint is satisfied:

$$(\forall\ w) \quad \frac{1}{n} E||I^n - Y^n(w, I^n)||^2 \leq D \tag{3.53}$$

Equivalently, we can assume that $Y^n(w, I^n) = I^n + X^n(w, I^n)$, and therefore (3.53) becomes

$$(\forall\ w) \quad n^{-1}E||X^n(w, I^n)||^2 \leq D \tag{3.54}$$

Let $W_1, \ldots, W_k$ be the indices of the colluders, each uniformly distributed in $\{1, \ldots, 2^{nR_F}\}$ and independent. The colluders produce the forgery Z^n as

$$Z^n = \sum_{l=1}^{k} \lambda_k Y_l^n + V^n$$

where, by definition, $Y_l^n = Y^n(W_l, I^n)$ and V^n is i.i.d. Gaussian noise with variance P_V (per dimension). The decoder produces estimates $\hat{W}_1, \ldots, \hat{W}_k$ of $W_1, \ldots, W_k$ without knowledge of I^n.

We use a random coding argument for our achievability proof. The approach is the multi-user extension of Costa's proof [15]. We trace the following steps:

Codebook Generation: First, 2^{nR_U} sequences U^n are generated i.i.d. Gaussian with variance $D + \alpha^2 P_I$ per dimension. Next, these sequences are distributed uniformly into 2^{nR_F} bins. Therefore, each bin w contains $U^n(w, 1), \ldots, U^n(w, 2^{n(R_U - R_F)})$.

Fingerprint Embedding: Given I^n and fingerprint index w, the embedder seeks within bin w a $U^n(w, q)$ which is jointly typical with I^n (that is, $(I^n, U^n(w, q)) \in T^n_{I,\hat{U}}(\epsilon)$). Joint typicality is with respect to some joint Gaussian distribution with covariance matrix

$$K_{I,\hat{U}} = \begin{bmatrix} P_I & \alpha P_I \\ \alpha P_I & D + \alpha^2 P_I \end{bmatrix}$$

An error is declared if no such U^n can be found. Otherwise, the encoder sets $\hat{U}^n(w, I^n) = U^n(w, q)$ and outputs $Y^n = \hat{U}^n(w, I^n) + (1 - \alpha)I^n$. The selected sequence $\hat{U}^n(w, I^n)$ is also distortion-typical, in the sense that $n^{-1}E||\hat{U}^n(w, I^n) - \alpha I^n||^2 = n^{-1}E||X^n(w, I^n)||^2 \leq D + \epsilon$.

Decoding: The decoder, given Z^n, seeks $U^n(\hat{w}_1, q_1), \ldots, U^n(\hat{w}_k, q_k)$ (belonging to bins $\hat{w}_1, \ldots, \hat{w}_k$, respectively) such that $(U^n(\hat{w}_1, q_1), \ldots, U^n(\hat{w}_k, q_k), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)$. Here, $T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)$ is the set of jointly typical tuples with respect to a joint Gaussian distribution with covariance matrix

$$K_{\hat{U}_1,\ldots,\hat{U}_k,Z} = \begin{bmatrix} D+\alpha^2 P_I & \cdots & \alpha^2 P_I & \lambda_k(D+k\alpha P_I) \\ \alpha^2 P_I & \ddots & \alpha^2 P_I & \lambda_k(D+k\alpha P_I) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_k(D+k\alpha P_I) & \cdots & \lambda_k(D+k\alpha P_I) & \lambda_k^2 k(D+kP_I)+P_V \end{bmatrix}$$

If *exactly* k typical $U^n(\hat{w}_1, q_1), \ldots, U^n(\hat{w}_k, q_k)$ are found, then the decoder outputs $\hat{W}_l = \hat{w}_l$ for all $1 \leq l \leq k$. Otherwise, an error is declared.

Probability of Error: Without loss of generality, we assume that $W_1 = 1, \ldots, W_k = k$ (this is the worst-case scenario in which all the fingerprint indices are different). We now have the following error events:

(i) E_1: No $U^n(l, q)$ can be found for some $l \leq k$, such that $(I^n, U^n(l, q)) \in T^n_{I,\hat{U}}$.

(ii) E_2: Assuming E_1^c (i.e., that all bins $1, \ldots, k$ contain U^n's which are typical with I^n), not all bins $1, \ldots, k$ contain $U^n(1, q_1), \ldots, U^n(k, q_k)$ respectively, such that $(U^n(1, q_1), \ldots, U^n(k, q_k), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}$.

(iii) E_3: Assuming $(E_1 \cup E_2)^c$, there exists a tuple $(w_1, \ldots, w_k) \notin \mathcal{P}(1, \ldots, k)$ and there exist $U^n(w_1, q_1), \ldots, U^n(w_k, q_k)$ in bins $w_1, \ldots, w_k$ respectively, such that $(U^n(w_1, q_1), \ldots, U^n(w_k, q_k), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)$. Note that $\mathcal{P}(1, \ldots, k)$ is the set that contains all ordered permutations of the tuple $(1, \ldots, k)$.

Behavior of E_1: We know, from rate-distortion theory, that if the number of elements in each bin is at least $2^{n(I(I;\hat{U})+\epsilon)}$, then $\Pr\{E_1\} \to 0$ as n approaches infinity. This is equivalent to

$$R_U - R_F \geq I(I; \hat{U}) + \epsilon \tag{3.55}$$

where the mutual information is computed with respect to the Gaussian joint

distribution with covariance matrix $K_{I,\hat{U}}$. Hence, by substitution, (3.55) becomes

$$R_U - R_F \geq \frac{1}{2}\log\left(1 + \frac{\alpha^2 P_I}{D}\right) + \epsilon \tag{3.56}$$

Behavior of E_2: To show that $\Pr\{E_2\} \to 0$, it suffices to show that $(\hat{U}^n(1, I^n), \ldots, \hat{U}^n(k, I^n), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}$ with probability that approaches 1. From the previous paragraph, we know that $\Pr\{(I^n, \hat{U}^n(l, I^n)) \in T^n_{I,\hat{U}}(\epsilon)\} \to 1$ for all $1 \leq l \leq k$. Since $Z^n = \sum_{l=1}^k \lambda_k Y_l^n + V^n = \sum_{l=1}^k \lambda_k \hat{U}^n(l, I^n) + k\lambda_k(1-\alpha)I^n + V^n$ and V^n is independent of $(I^n, \hat{U}^n(1, I^n), \ldots, \hat{U}^n(k, I^n))$, it follows easily that the empirical correlations obtained from $(\hat{U}^n(1, I^n), \ldots, \hat{U}^n(k, I^n), Z^n)$ are within a factor of ϵ of the corresponding entries of $K_{\hat{U}_1,\ldots,\hat{U}_k,Z}$ with probability approaching unity asymptotically. Hence, typicality is established with probability that approaches 1 (therefore $\Pr\{E_2\} \to 0$).

Behavior of E_3: We upper-bound the probability of the error event E_3 as follows:

$$\begin{aligned}
\Pr(E_3) \;\leq\; & \Pr\{(\exists\ (w_1, \ldots, w_k) \neq (1, \ldots, k)) \wedge ((\forall\ 1 \leq l \leq k)\ \exists\ U^n(w_l, q_l)) : \\
& (U^n(w_1, q_1), \ldots, U^n(w_k, q_k), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)\} \qquad (3.57) \\
= \;& \Pr\{(\exists\ S_l \subset \{1, \ldots, k\},\ 1 \leq l \leq k)\ \wedge\ (\exists\ (w_1, \ldots, w_k)\ : \\
& (w_s \neq s,\ \forall\ s \in S_l)\ \wedge\ (w_t = t,\ \forall\ t \in \bar{S}_l))\ \wedge \\
& ((\forall\ 1 \leq l \leq k)\ \exists\ U^n(w_l, q_l)) : \\
& (U^n(w_1, q_1), \ldots, U^n(w_k, q_k), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)\} \\
\leq \;& \sum_{l=1}^{k} \sum_{S_l \subset \{1,\ldots,k\}} \sum_{\substack{(w_1,\ldots,w_k): \\ w_s \neq s,\ \forall\, s \in S_l \\ w_t = t,\ \forall\, t \in \bar{S}_l}} \sum_{\{q_m\}, 1 \leq m \leq k} 1 \times \\
& \times \Pr\{(U^n(w_1, q_1), \ldots, U^n(w_k, q_k), Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)\} \qquad (3.58)
\end{aligned}$$

Assume that there exists a set $WQ'(r)$ that contains r pairs $(w'_1, q'_1), \ldots, (w'_r, q'_r)$, which belong to the set $WQ = \{(w_1, q_1), \ldots, (w_k, q_k)\}$ such that $\hat{U}^n(w'_s, I^n) =$

$U^n(w'_s, q'_s)$ for all $1 \leq s \leq r$. Then, because of the symmetry in the construction of the U^n's and their symmetric contribution to Z^n, we have

$$\Pr\{(U^n(w_1,q_1),\ldots,U^n(w_k,q_k),Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)\} =$$
$$\Pr\{(\{\hat{U}^n(w'_s,I^n)\}_{s=1}^r, \{U^n(w_m,q_m)\}_{(w_m,q_m)\in WQ-WQ'(r)}, Z^n) \in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)\} \doteq$$
$$2^{-n((k-r)h(\hat{U})-h(\hat{U}_{r+1},\ldots,\hat{U}_k|\hat{U}_1,\ldots,\hat{U}_r,Z))}$$

where the differential entropies are computed with respect to the joint Gaussian with covariance matrix $K_{\hat{U}_1,\ldots,\hat{U}_k,Z}$.

Then, continuing from (3.58) we have

$$\begin{aligned}
\Pr\{E_3\} &\leq \sum_{l=1}^{k} \sum_{S_l\subset\{1,\ldots,k\}} \sum_{\substack{(w_1,\ldots,w_k):\\ w_s\neq s,\ \forall\, s\in S_l\\ w_t=t,\ \forall\, t\in \bar{S}_l}} \sum_{\{q_m\},m\in S_l} \sum_{\{q_\nu\},\nu\in\bar{S}_l} 1 \times \\
&\quad \times \Pr\{(U^n(w_1,q_1),\ldots,U^n(w_k,q_k),Z^n)\in T^n_{\hat{U}_1,\ldots,\hat{U}_k,Z}(\epsilon)\} \\
&\doteq \sum_{l=1}^{k} \sum_{S_l\subset\{1,\ldots,k\}} \sum_{w_s\neq s,\ \forall\, s\in S_l} 2^{n(R_U-R_F)l} \sum_{r=0}^{k-l} 2^{n(R_U-R_F)(k-l-r)} \times \\
&\quad \times 2^{-n((k-r)h(\hat{U})-h(\hat{U}_{r+1},\ldots,\hat{U}_k|\hat{U}_1,\ldots,\hat{U}_r,Z))} \\
&= \sum_{l=1}^{k} \sum_{S_l\subset\{1,\ldots,k\}} \sum_{w_s\neq s,\ \forall\, s\in S_l} \sum_{r=0}^{k-l} 2^{n(R_U-R_F)(k-r)} \times \\
&\quad \times 2^{-n((k-r)h(\hat{U})-h(\hat{U}_{r+1},\ldots,\hat{U}_k|\hat{U}_1,\ldots,\hat{U}_r,Z))} \\
&\leq \sum_{l=1}^{k} \binom{k}{l} \sum_{r=0}^{k-l} 2^{nR_U(k-r)-nR_F(k-l-r)}\, 2^{-n((k-r)h(\hat{U})-h(\hat{U}_{r+1},\ldots,\hat{U}_k|\hat{U}_1,\ldots,\hat{U}_r,Z))}
\end{aligned} \tag{3.59}$$

The last expression in (3.59) approaches zero if for every $1 \leq l \leq k$ and $0 \leq r \leq k-l$ we have

$$R_U(k-r) - R_F(k-l-r) < (k-r)h(\hat{U}) - h(\hat{U}_{r+1},\ldots,\hat{U}_k|\hat{U}_1,\ldots,\hat{U}_r,Z) \tag{3.60}$$

Observe that when $k=1$ (no collusion) then (3.60) becomes $R_U < I(\hat{U};Z)$, as expected from [15].

It can be proved by induction that

$$(k-r)h(\hat{U}) - h(\hat{U}_{r+1},\ldots,\hat{U}_k|\hat{U}_1,\ldots,\hat{U}_r,Z) = (k-r)\frac{1}{2}\log\left(1+\frac{\alpha^2 P_I}{D}\right) + \frac{1}{2}\log\left(\frac{D^2\lambda_k^2(k-r)+r\alpha^2P_IP_V+D(\lambda_k^2kP_I(k+r\alpha^2-2r\alpha)+P_V)}{k\alpha^2P_IP_V+D(k^2\lambda_k^2P_I(1-\alpha)^2+P_V)}\right)$$

Thus, from (3.60) we obtain that for every $1 \le l \le k$ and $0 \le r \le k-l$

$$R_U < R_F\left(1-\frac{l}{k-r}\right)+\frac{1}{2}\log\left(1+\frac{\alpha^2P_I}{D}\right) + \frac{1}{2(k-r)}\log\left(\frac{D^2\lambda_k^2(k-r)+r\alpha^2P_IP_V+D(\lambda_k^2kP_I(k+r\alpha^2-2r\alpha)+P_V)}{k\alpha^2P_IP_V+D(k^2\lambda_k^2P_I(1-\alpha)^2+P_V)}\right) \tag{3.61}$$

Hence, from (3.56) and (3.61) we obtain that for every $1 \le l \le k$ and $0 \le r \le k-l$

$$R_F \le \frac{1}{2l}\log\left(\frac{D^2\lambda_k^2(k-r)+r\alpha^2P_IP_V+D(\lambda_k^2kP_I(k+r\alpha^2-2r\alpha)+P_V)}{k\alpha^2P_IP_V+D(k^2\lambda_k^2P_I(1-\alpha)^2+P_V)}\right) \tag{3.62}$$

It is easy to check that the right-hand side of (3.62) is maximized when $\alpha = \frac{Dk\lambda_k^2}{Dk\lambda_k^2+P_V}$. By substituting in (3.62), we obtain

$$R_F < \min_{1\le l\le k}\ \min_{0\le r\le k-l}\ \frac{1}{2l}\log\left(1+\frac{D\lambda_k^2(k-r)}{P_V}\right) \tag{3.63}$$

or, equivalently,

$$R_F < \frac{1}{2k}\log\left(1+\frac{D\lambda_k^2k}{P_V}\right) \tag{3.64}$$

which is the required result. Observe that R_F cannot be higher because this is the maximum rate achieved in $\mathcal{R}_D^{\text{gauss, F}}$ when $R_Q = \infty$ (and I^n is known at the decoder).

Chapter 4

General Gaussian Images and Attacks

All results obtained so far are based on two main assumptions: that the attacks are memoryless, and the original image I^n is i.i.d. In this chapter, we consider again the problem of quantization of watermarked data, but under the following assumptions: (i) the attack noise is additive and Gaussian but not necessarily i.i.d. (or even stationary), and (ii) the original image I^n is Gaussian, but not necessarily stationary, either. We derive achievable quantization and watermarking rates whose values depend on the image size n. Although these rates may not have a limit as $n \to \infty$ (like it happens in the i.i.d. case), probabilities of error do approach zero for very large n.

In our analysis, we use the theory developed in [18, 43, 44]. The problem that was studied in [18, 43] is the "colored" paper version of [15]. Specifically, the authors consider a single block of n transmissions, in which the received signal is given by

$$Y^n = X^n + S^n + Z^n$$

where X^n is the transmitted signal and Z^n, S^n are independent Gaussian processes with arbitrary finite-dimensional covariance matrices (thus are not necessarily sta-

tionary or ergodic). S^n (which plays the role of the "colored" paper) is available non-causally to the transmitter only. Moreover, the transmission is power limited in the usual sense:

$$\frac{1}{n}\sum_{j=1}^{n} E(X_j^2) \leq P$$

The main result of [18, 43] is that there exists a $(e^{n(C_n-4\epsilon)}, n)$ code over a one-shot use of n transmissions such that the probability of error is upper bounded by $e^{-n\alpha(\epsilon)}$, where $\alpha(\epsilon)$ does not depend on the statistics of Z^n. C_n is called the capacity of the Gaussian channel over one shot of n transmissions, and is given by

$$C_n = \max_{K_{X^n}:\ n^{-1}\mathrm{tr}(K_{X^n})\leq P(1-\epsilon)} \frac{1}{2n}\log\left(\frac{|K_{X^n}+K_{Z^n}|}{|K_{Z^n}|}\right)$$

where K_{X^n}, K_{Z^n} are the covariance matrices of X^n and Z^n, respectively, the maximum is attained for a non-singular K_{X^n}, and ϵ is an arbitrary positive number. Although C_n may fluctuate arbitrarily with n, the probability of error approaches zero for large n, as long as the rate of the code is upper-bounded by C_n.

In this chapter, we consider a private watermarking scheme, where the watermarked image is distributed in quantized form. In Section 4.1, we give an overview of the general Gaussian watermarking/quantization model and we state the main result, i.e., the theorem which establishes the region of achievable rate pairs (R_Q, R_W). Section 4.2 contains the proof of the main theorem; finally, in Section 4.3 we consider special cases of the general result.

4.1 Main Result

We consider a system similar to the one shown in Figure 2.2. We assume that the watermark index W is uniformly distributed over a set of size $2^{nR_W^{(n)}}$; the original image I^n is zero-mean Gaussian with covariance matrix K_{I^n}, and the additive noise

V^n is zero-mean Gaussian with covariance matrix K_{V^n}, independent of (I^n, W). Note that $R_W^{(n)}$ depends on n. The output of the encoder is a sequence $\hat{Y}^n$ which belongs to a source codebook of size $2^{nR_Q^{(n)}}$ (again, note the dependence of $R_Q^{(n)}$ on n). The output of the attack is $Z^n = B_A\hat{Y}^n + V^n$, where B_A is a fixed $n \times n$ matrix. For each n, both encoder and decoder know B_A and the statistics of the noise, i.e., K_{V^n}. For simplicity of notation, we will drop the superscript n when we refer to the correlation matrices; for example, K_I denotes K_{I^n}. Moreover, $K_{I\hat{Y}}$ is the cross-correlation matrix $E[I^n(\hat{Y}^n)^t]$, where the superscript t denotes the matrix transpose.

The definitions of a private quantization/watermarking code are similar to those in Section 2.1, except that the rates (R_Q, R_W) are no longer constant in n. Moreover, we say that a rate pair $(R_Q^{(n)}, R_W^{(n)})$ is achievable over one shot of n transmissions if there exists a sequence of codes $(2^{nR_Q^{(n)}}, 2^{nR_W^{(n)}}, n)$ such that

$$\frac{1}{n}\sum_{j=1}^{n} E[(I_j - \hat{Y}_j)^2] \leq D \tag{4.1}$$

and

$$(\forall\ w) \quad P_e^{(n)}(w) \stackrel{\Delta}{=} \Pr\{g(Z^n, I^n) \neq w | X^n = f(w, I^n)\} \to 0, \text{ as } n \to \infty \tag{4.2}$$

Also, as usual,

$$P_e^{(n)} = \frac{1}{2^{nR_W^{(n)}}} \sum_{w=1}^{2^{nR_W^{(n)}}} P_e^{(n)}(w)$$

is the average probability of error, assuming that the watermark indices are uniformly distributed in $\{1, \ldots, 2^{nR_W^{(n)}}\}$.

We now state the main result of this chapter:

Theorem 4.1 *Let $\epsilon > 0$ and consider the rate region*

$$\mathcal{R}_D^{\text{general, gauss}} = \Bigg\{ (R_Q^{(n)}, R_W^{(n)}) :$$

$$R_Q^{(n)} \geq \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2}\log\left(\frac{\sigma_j^2}{\min\{\xi, \sigma_j^2\}}\right), \text{ where } \xi \text{ satisfies } \sum_{j=1}^{n}\min\{\xi, \sigma_j^2\} = nD;$$

$$R_W^{(n)} \leq \max_{\substack{K_{I\hat{Y}}, K_{\hat{Y}}: \\ n^{-1}\text{tr}(K_I + K_{\hat{Y}} - 2K_{I\hat{Y}}) \leq D}} \min \Bigg\{ R_Q^{(n)} - \frac{1}{2n}\log\left(\frac{|K_I|}{|K_I - K_{I\hat{Y}}K_{\hat{Y}}^{-1}K_{I\hat{Y}}^t|}\right),$$

$$\frac{1}{2n}\log\left(\frac{|B_A(K_{\hat{Y}} - K_{\hat{Y}I}K_I^{-1}K_{\hat{Y}I}^t)B_A^t + K_V|}{|K_V|}\right)\Bigg\}\Bigg\}$$

where $\sigma_1^2, \ldots, \sigma_n^2$ are the eigenvalues of I^n. If the maximum in the above expression is achieved for non-singular matrices $K_{\hat{Y}}, K_{I\hat{Y}}$, then there exists a sequence of $(2^{nR_Q^{(n)}}, 2^{nR_W^{(n)}}, n)$ quantization/watermarking codes with $(R_Q^{(n)} - \epsilon, R_W^{(n)} + \epsilon) \in \mathcal{R}_D^{\text{general, gauss}}$ such that conditions (4.1) and (4.2) are satisfied. Conversely, any sequence of $(2^{nR_Q^{(n)}}, 2^{nR_W^{(n)}}, n)$ codes with $(R_Q^{(n)}, R_W^{(n)}) \notin \mathcal{R}_D^{\text{general, gauss}}$ violates either conditions (4.1) or (4.2).

The proof of this theorem can be found in Section 4.2.

Note: Assuming that the matrices B_A and K_V are non-singular, we can write the second upper bound on $R_W^{(n)}$ as follows:

$$\frac{1}{2n}\log\left(\frac{|\Theta(K_{\hat{Y}} - K_{\hat{Y}I}K_I^{-1}K_{\hat{Y}I}^t)\Theta^t + \Delta|}{|\Delta|}\right) \leq$$

$$\frac{1}{2n}\sum_{j=1}^{n}\log\left(\frac{(\Theta(K_{\hat{Y}} - K_{\hat{Y}I}K_I^{-1}K_{\hat{Y}I}^t)\Theta^t)_{jj} + \Delta_{jj}}{\Delta_{jj}}\right)$$

where $\Delta = \text{diag}(\delta_1^2, \ldots, \delta_n^2)$ satisfies

$$B_A^{-1}K_V(B_A^{-1})^t = \Theta\Delta\Theta^t, \quad \Theta\Theta^t = \mathbf{I}$$

(whitening transformation). The right-hand side of the inequality becomes tight only when $\Theta(K_{\hat{Y}} - K_{\hat{Y}I}K_I^{-1}K_{\hat{Y}I}^t)\Theta^t$ is diagonal (cf. the argument for the capacity

of an additive colored Gaussian noise channel in [33]). However, this does not necessarily give a tight upper bound for $R_W^{(n)}$, because the first upper bound

$$R_W^{(n)} \leq R_Q^{(n)} - \frac{1}{2n} \log \left(\frac{|K_I|}{|K_I - K_{I\hat{Y}} K_{\hat{Y}}^{-1} K_{I\hat{Y}}^t|} \right)$$

may become suboptimally small.

4.2 Proof of Theorem 4.1

Before we proceed with the proof, it is useful to consider some important definitions and lemmas.

The first lemma proves the asymptotic equipartition property (AEP) for arbitrary Gaussian stochastic processes. Note that, in general, the AEP holds only for stationary and ergodic processes [33]; that is, if $\{X_j\}$ is a stationary and ergodic process with entropy rate h, then

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \to h$$

with probability one; here p is the joint distribution of $(X_1, \ldots, X_n)$, and h is defined as

$$h = \lim_{n \to \infty} H(X_n | X_{n-1}, \ldots, X_1) = \lim_{n \to \infty} \frac{H(X_1, \ldots, X_n)}{n}$$

However, Gaussian processes are special because they obey the AEP without any assumption on stationarity or ergodicity. Although the entropy rate may not exist, the per time unit differential entropy h_n of $(X_1, \ldots, X_n)$ plays the same role, where

$$h_n(X) \triangleq \frac{h(X_1, \ldots, X_n)}{n}$$

We thus have the following lemma (proved in [44]):

Lemma 4.1 *If $\{X_j\}$ is an arbitrary Gaussian stochastic process, then*

$$-\frac{1}{n}\log p(X_1,\ldots,X_n) - h_n(X) \to 0$$

with probability one.

Note that if the Gaussian $\{X_j\}$ has an entropy rate h, then the above lemma implies that $-\frac{1}{n}\log p(X_1,\ldots,X_n) \to h$, as expected. In proving Lemma 4.1, it is argued in [44] that $(X^n)^t K_{X^n}^{-1} X^n$ has a chi-square distribution with n degrees of freedom, i.e., it has the same distribution as $\sum_{j=1}^n Z_j^2$, where Z_j are i.i.d. $\sim \mathcal{N}(0,1)$. By utilizing the Chernoff bound, it is then proved that, if $|K_{X^n}| > 0$ for all n (where K_{X^n} is the covariance matrix of $(X_1,\ldots,X_n)$),

$$\Pr\left\{\left|-\frac{1}{n}\log p(X_1,\ldots,X_n) - h_n(X)\right| > \epsilon\right\} < 2^{-n\epsilon'}, \text{ for all } \epsilon > 0 \tag{4.3}$$

where $\epsilon' = \epsilon - 1/2\log(1+2\epsilon)$ is a positive quantity which approaches zero as ϵ approaches zero. Hence, the rate of the convergence depends only on ϵ, and not on $|K_{X^n}|$.

We now provide the following definitions, consistent with [44, 43].

Definition 4.1 *Let (X^n, Y^n) be jointly distributed with density $p(x^n, y^n)$. Let*

$$h_n(X) = \frac{1}{n}h(X^n), \quad h_n(Y) = \frac{1}{n}h(Y^n), \quad h_n(X,Y) = \frac{1}{n}h(X^n,Y^n)$$

Then, the set $T_{X,Y}^n(\epsilon)$ of jointly ϵ-typical (X^n, Y^n) is defined by

$$T_{X,Y}^n(\epsilon) = \left\{ (x^n,y^n) \in \mathbf{R}^n \times \mathbf{R}^n : \right.$$
$$\left|-\frac{1}{n}\log p(x^n) - h_n(X)\right| < \epsilon$$
$$\left|-\frac{1}{n}\log p(y^n) - h_n(Y)\right| < \epsilon$$
$$\left.\left|-\frac{1}{n}\log p(x^n,y^n) - h_n(X,Y)\right| < \epsilon\right\}$$

where $p(x^n), p(y^n)$ are the marginals of X^n, Y^n respectively.

From the definition above and Lemma 4.1, it is trivial to prove the following:

Lemma 4.2 *Let X^n be an arbitrary Gaussian stochastic process, with $|K_{X^n}| > 0$. Then,*

$$\Pr\{X^n \in T_X^n(\epsilon)\} > 1 - \epsilon$$

for n sufficiently large.

We now have the following Lemma, proved in [43].

Lemma 4.3 *Let X^n, Y^n be jointly Gaussian. The volume of the typical set $T_{X,Y}^n(\epsilon)$, denoted by $|T_{X,Y}^n(\epsilon)|$, satisfies:*

$$(1 - 2^{-n\epsilon'})2^{n(h_n(X,Y)-\epsilon)} \leq |T_{X,Y}^n(\epsilon)| \leq 2^{n(h_n(X,Y)+\epsilon)} \tag{4.4}$$

Also, let U^n and V^n be two independent Gaussian sequences with the same marginals as X^n and Y^n respectively. Then

$$(1 - 2^{-n\epsilon'})2^{-n(I_n(X;Y)-3\epsilon)} \leq \Pr\{(U^n, V^n) \in T_{X,Y}^n(\epsilon)\} \leq 2^{-n(I_n(X;Y)+3\epsilon)} \tag{4.5}$$

where, by definition, $I_n(X;Y) = \frac{1}{n} I(X^n; Y^n)$.

The proof of Lemma 4.3 is very similar to the proof of the corresponding result for i.i.d. sources, found in [33]. Moreover, we have two more lemmas:

Lemma 4.4 *For all $(x^n, y^n) \in T_{X,Y}^n(\epsilon)$*

$$p_{Y^n}(y^n) \geq p_{Y^n|X^n}(y^n|x^n)\, 2^{-n(I_n(X;Y)+3\epsilon)}$$

where the distributions $p_{Y^n}, p_{Y^n|X^n}$ are the ones used in the definition of $T_{X,Y}^n(\epsilon)$.

Lemma 4.5 *For $0 \leq x, y \leq 1, n > 0$,*

$$(1 - xy)^n \leq 1 - x + e^{-yn}.$$

The proof of Lemma 4.4 is very similar to the proof of Lemma 13.5.2 in [33], and Lemma 4.5 is identical to Lemma 13.5.3 in [33].

We will now begin the proof of Theorem 4.1, starting with the converse part.

Converse Theorem

Proof: Let $\epsilon > 0$. We assume that the watermark index W is uniformly distributed in $\{1, \ldots, 2^{nR_W^{(n)}}\}$, that $\Pr\{W \neq \hat{W}\} < \epsilon$, and that the distortion constraint is met with equality:

$$\frac{1}{n}\sum_{j=1}^{n} E[(I_j - \hat{Y}_j)^2] = D \tag{4.6}$$

which is equivalent to

$$\frac{1}{n}\mathrm{tr}(K_I + K_{\hat{Y}} - 2K_{I\hat{Y}}) = D$$

Without loss of generality, we assume that K_I is non-singular, therefore, $|K_I| > 0$ (if not, we can linearly transform I^n into a vector of lower dimension which has a non-singular covariance matrix.)

First, we will derive the lower bound for $R_Q^{(n)}$. The derivation is similar to the case of parallel Gaussian sources [33]:

$$\begin{aligned}
R_Q^{(n)} &\geq n^{-1}H(\hat{Y}^n) \\
&\geq n^{-1}(H(\hat{Y}^n) - H(\hat{Y}^n|I^n)) \\
&= n^{-1}I(\hat{Y}^n; I^n) \\
&= n^{-1}(h(I^n) - h(I^n|\hat{Y}^n)) \\
&= n^{-1}(h(I^n) - h(I^n - \hat{Y}^n|\hat{Y}^n)) \\
&\geq n^{-1}(h(I^n) - h(I^n - \hat{Y}^n)) && (4.7) \\
&= n^{-1}(h(\tilde{I}^n) - h(I^n - \hat{Y}^n)) && (4.8) \\
&= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\sigma_j^2 - n^{-1}h(I^n - \hat{Y}^n) && (4.9)
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\sigma_j^2 - \frac{1}{n}\sum_{j=1}^{n}h(I_j-\hat{Y}_j) \\
&\geq \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\sigma_j^2 - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)D_j \qquad (4.10)\\
&= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log\left(\frac{\sigma_j^2}{D_j}\right) \qquad (4.11)
\end{aligned}
$$

where (4.7) is true because conditioning reduces entropy; $\tilde{I}^n$ in (4.8) is the Karhunen-Loève transformation (KLT) of I^n, i.e.,

$$\tilde{I}^n = Q^t I^n, \quad \text{s.t. } QQ^t = \mathbf{I}$$

($\mathbf{I} = \text{diag}(1,\ldots,1)$) with $K_{\tilde{I}} = Q^t K_I Q$ and hence $h(\tilde{I}^n) = h(I^n)$; $\sigma_1^2,\ldots,\sigma_n^2$ in (4.9) are the eigenvalues of K_I; and, finally, we used the Gaussian upper bound on the entropy in (4.10), where $D_j \stackrel{\Delta}{=} E[(I_j-\hat{Y}_j)^2]$.

Our goal now is to minimize (4.11) with respect to D_j, subject to the constraint (4.6), or equivalently,

$$\frac{1}{n}\sum_{j=1}^{n}D_j = D$$

This can be done using Lagrange multipliers and the reverse "water-filling" method [33]. It can thus be proved that (4.11) is minimized when

$$D_j = \begin{cases} \xi, & \text{if } \xi < \sigma_j^2 \\ \sigma_j, & \text{if } \xi \geq \sigma_j^2 \end{cases}$$

where ξ is chosen such that $\sum_{j=1}^{n} D_j = nD$, thus establishing the lower bound for $R_Q^{(n)}$.

We now establish the first upper bound on $R_W^{(n)}$. By following the same chain of inequalities as in the converse in Section 2.2, we obtain from (2.22):

$$R_W^{(n)} \leq R_Q^{(n)} - n^{-1}I(\hat{Y}^n;I^n) + \epsilon \qquad (4.12)$$

We have

$$\begin{aligned} I(\hat{Y}^n; I^n) &= h(I^n) - h(I^n|\hat{Y}^n) \\ &= h(I^n) - h(I^n - M\hat{Y}^n|\hat{Y}^n) && (4.13) \\ &\geq h(I^n) - h(I^n - M\hat{Y}^n) && (4.14) \\ &\geq \frac{1}{2}\log(2\pi e)^n|K_I| - \frac{1}{2}\log(2\pi e)^n|K_{I-M\hat{Y}}| && (4.15) \\ &= \frac{1}{2}\log\left(\frac{|K_I|}{|K_{I-M\hat{Y}}|}\right) && (4.16) \end{aligned}$$

where (4.13) is true for any arbitrary $n \times n$ matrix M; (4.14) holds because conditioning reduces entropy; and the Gaussian upper bound on the entropy was used in (4.15).

We can now set $M = K_{I\hat{Y}}K_{\hat{Y}}^{-1}$. Then (see, for example, [45]) $M\hat{Y}^n$ is the MMSE linear estimator of I^n given $\hat{Y}^n$. The covariance matrix of the error is $K_{I-M\hat{Y}} = K_I - K_{I\hat{Y}}K_{\hat{Y}}^{-1}K_{I\hat{Y}}^t$, so, by substitution in (4.16) and together with (4.12), we finally obtain:

$$R_W^{(n)} \leq R_Q^{(n)} - \frac{1}{2n}\log\left(\frac{|K_I|}{|K_I - K_{I\hat{Y}}K_{\hat{Y}}^{-1}K_{I\hat{Y}}^t|}\right) + \epsilon \qquad (4.17)$$

The second set of inequalities establishes the second upper bound on $R_W^{(n)}$ as follows:

$$\begin{aligned} R_W^{(n)} &= n^{-1}H(W|I^n) \\ &= n^{-1}I(W; Z^n|I^n) + n^{-1}H(W|I^n, Z^n) \\ &\leq n^{-1}I(W; Z^n|I^n) + \epsilon && (4.18) \\ &= n^{-1}h(Z^n|I^n) - n^{-1}h(Z^n|W, I^n) + \epsilon \\ &= n^{-1}h(B_A(\hat{Y}^n - \Lambda I^n) + V^n|I^n) - n^{-1}h(V^n) + \epsilon && (4.19) \\ &\leq n^{-1}h(B_A(\hat{Y}^n - \Lambda I^n) + V^n) - n^{-1}h(V^n) + \epsilon && (4.20) \end{aligned}$$

$$\leq \frac{1}{2n}\log(2\pi e)^n(|B_A K_{\hat{Y}-\Lambda I} B_A^t + K_V|) - \frac{1}{2n}\log(2\pi e)^n|K_V| + \epsilon \quad (4.21)$$

$$= \frac{1}{2n}\log\left(\frac{|B_A K_{\hat{Y}-\Lambda I} B_A^t + K_V|}{|K_V|}\right) + \epsilon \quad (4.22)$$

where (4.18) is a consequence of Fano's inequality; (4.19) is true for any $n \times n$ matrix Λ; conditioning reduces entropy in (4.20); and we used the Gaussian upper bound on the entopy of $B_A(\hat{Y}^n - \Lambda I^n) + V^n$ in (4.21).

By setting $\Lambda = K_{\hat{Y}I}K_I^{-1}$, we have $K_{\hat{Y}-\Lambda I} = K_{\hat{Y}} - K_{\hat{Y}I}K_I^{-1}K_{\hat{Y}I}^t$. Then (4.22) becomes

$$R_W^{(n)} \leq \frac{1}{2n}\log\left(\frac{|B_A(K_{\hat{Y}} - K_{\hat{Y}I}K_I^{-1}K_{\hat{Y}I}^t)B_A^t + K_V|}{|K_V|}\right) + \epsilon \quad (4.23)$$

Hence, combining (4.17) and (4.23) and maximizing with respect to $K_{I\hat{Y}}$ and $K_{\hat{Y}}$ such that (4.6) is satisfied, we obtain the required result (letting $\epsilon \to 0$, as usual).

■

We now proceed with the proof of the direct part.

Direct Theorem

Proof: We use a random coding argument. Let $\epsilon > 0$ and let W be uniformly distributed in $\{1, \ldots, 2^{nR_W^{(n)}}\}$. As required for $\mathcal{R}_D^{\text{general, gauss}}$, we limit the quantization rate to $R_Q^{(n)} \geq r_q^{(n)}(D)$, where

$$r_q^{(n)}(D) \triangleq \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log\left(\frac{\sigma_j^2}{\min\{\xi, \sigma_j^2\}}\right), \quad \text{s.t.} \quad \sum_{j=1}^{n}\min\{\xi, \sigma_j^2\} \leq nD$$

The encoding/decoding and analysis of the probability of error follow. Notice the similarities with the achievability proof of Section 2.2.

Codebook Generation: A set of $2^{nR_Q^{(n)}}$ sequences $\tilde{Y}^n$ is generated, such that each sequence is generated independently of every other sequence, according to the joint

Gaussian $\mathcal{N}(0, K_{\hat{Y}})$, for some non-singular matrix $K_{\hat{Y}}$. The set is then partitioned into $2^{nR_W^{(n)}}$ subsets of $2^{nR_1^{(n)}}$ sequences each, i.e.,

$$R_Q^{(n)} = R_W^{(n)} + R_1^{(n)}$$

The w^{th} subset, consisting of sequences $\tilde{Y}^n(w,1), \ldots, \tilde{Y}^n(w, 2^{nR_1^{(n)}})$, becomes the codebook for the w^{th} watermark.

Watermark Embedding: Given I^n and a deterministic w, the embedder identifies within the w^{th} codebook the first codeword $\tilde{Y}^n(w,q)$ such that the pair $(I^n, \tilde{Y}^n(w,q))$ lies in the set $T^n_{I,\hat{Y}}(\epsilon)$ of typical pairs with respect to a joint Gaussian distribution of $p_{I^n,\hat{Y}^n}$, whose covariance matrix is

$$K_{I,\hat{Y}} = \begin{bmatrix} K_I & K_{I\hat{Y}} \\ K_{\hat{Y}I} & K_{\hat{Y}} \end{bmatrix}$$

for covariance matrices $K_{I\hat{Y}}, K_{\hat{Y}}$ such that the distortion constraint (4.6) is satisfied. The output of the embedder (encoder) is denoted by $\hat{Y}^n(w) = \tilde{Y}^n(w,q)$. If none of the codewords in the w^{th} codebook is jointly typical with I^n, then the embedder outputs $\hat{Y}^n(w) = 0$. This way, $2^{nR_W^{(n)}}$ watermarked versions of the image I^n can be obtained: $\hat{Y}^n(1), \ldots, \hat{Y}^n(2^{nR_W^{(n)}})$. For random W, the embedder output is $\hat{Y}^n(W)$.

Decoding: The decoder has access to the original image I^n, and together with Z^n, it seeks among all watermarked versions $\hat{Y}^n(1), \ldots, \hat{Y}^n(2^{nR_W^{(n)}})$ of I^n a single $\hat{Y}^n(\hat{w})$ such that the triplet $(I^n, \hat{Y}^n(\hat{w}), Z^n)$ lies in $T^n_{I,\hat{Y},Z}(\epsilon)$, the set of typical triplets with respect to a joint Gaussian distribution $p_{I^n,\hat{Y}^n,Z^n}$ with covariance matrix

$$K_{I,\hat{Y},Z} = \begin{bmatrix} K_I & K_{I\hat{Y}} & K_{I\hat{Y}}B_A^t \\ K_{\hat{Y}I} & K_{\hat{Y}} & K_{\hat{Y}}B_A^t \\ B_A K_{\hat{Y}I} & B_A K_{\hat{Y}} & B_A K_{\hat{Y}} B_A^t + K_V \end{bmatrix}$$

If a unique such sequence $\hat{Y}^n(\hat{w})$ exists, then the decoder outputs $\hat{W} = \hat{w}$; otherwise, the decoder declares an error.

Error Events: Without loss of generality, we assume $W = 1$. Similarly to the proof in Section 2.2, we have the following error events:

- E_1: $\hat{Y}^n(1) = 0$, i.e., there exists no $q \in \{1, \ldots, 2^{nR_1^{(n)}}\}$ such that $(I^n, \tilde{Y}^n(1,q)) \in T^n_{I,\hat{Y}}(\epsilon)$.
- E_2: There exists a $\tilde{Y}^n(1,q) = \hat{Y}^n(1)$ such that $(I^n, \hat{Y}^n(1)) \in T^n_{I,\hat{Y}}(\epsilon)$, but $(I^n, \hat{Y}^n(1), Z^n) \notin T^n_{I,\hat{Y},Z}(\epsilon)$.
- E_3: $(I^n, \hat{Y}^n(1), Z^n) \in T^n_{I,\hat{Y},Z}(\epsilon)$ but there also exists a $k > 1$ such that $(I^n, \hat{Y}^n(k), Z^n) \in T^n_{I,\hat{Y},Z}(\epsilon)$.

The probability of error is then

$$\Pr\{\hat{W} \neq 1\} = \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$$

Behavior of $Pr(E_1)$: We prove that, if the rate of the size of each codebook (i.e., $R_1^{(n)}$) is at least equal to to $I_n(I;\hat{Y}) + \epsilon$ (as defined in Lemma 4.3), then $\Pr(E_1) \to 0$ as $n \to \infty$. The proof is similar to the achievability proof of the rate-distortion theorem for i.i.d. sources in [33]. It proceeds as follows:

$$\begin{aligned}
&\Pr(E_1) = \\
&\qquad \Pr\{(I^n, \tilde{Y}^n(1,1)) \notin T^n_{I,\hat{Y}}(\epsilon) \wedge \ldots \wedge (I^n, \tilde{Y}^n(1, 2^{nR_1^{(n)}})) \notin T^n_{I,\hat{Y}}(\epsilon)\} \\
&\leq \int_{i^n:\ i^n \in T^n_I(\epsilon)} p_{I^n}(i^n) \Pr\left\{ \bigcap_{q=1}^{2^{nR_1^{(n)}}} \left((i^n, \tilde{Y}^n(1,q)) \notin T^n_{I,\hat{Y}}(\epsilon) \right) \middle| I^n = i^n \right\} di^n \\
&\qquad + \int_{i^n:\ i^n \notin T^n_I(\epsilon)} p_{I^n}(i^n) di^n
\end{aligned}$$

$$\leq \int_{i^n:\ i^n \in T_I^n(\epsilon)} p_{I^n}(i^n) \Pr\left\{ \bigcap_{q=1}^{2^{nR_1^{(n)}}} \left((i^n, \tilde{Y}^n(1,q)) \notin T_{I,\hat{Y}}^n(\epsilon) \right) \right\} di^n + \epsilon \quad (4.24)$$

$$= \int_{i^n:\ i^n \in T_I^n(\epsilon)} p_{I^n}(i^n) (1 - \Pr\{(i^n, \tilde{Y}^n(1,1)) \in T_{I,\hat{Y}}^n(\epsilon)\})^{2^{nR_1^{(n)}}} di^n + \epsilon \quad (4.25)$$

$$= \int_{i^n:\ i^n \in T_I^n(\epsilon)} p_{I^n}(i^n) \left(1 - \int_{\hat{y}^n:(i^n,\hat{y}^n) \in T_{I,\hat{Y}}^n(\epsilon)} p_{\hat{Y}^n}(\hat{y}^n)\, d\hat{y}^n \right)^{2^{nR_1^{(n)}}} di^n + \epsilon \quad (4.26)$$

$$\leq \int_{i^n:\ i^n \in T_I^n(\epsilon)} p_{I^n}(i^n) \Bigg(1 - \int_{\hat{y}^n:(i^n,\hat{y}^n) \in T_{I,\hat{Y}}^n(\epsilon)} p_{\hat{Y}^n|I^n}(\hat{y}^n|i^n) 2^{-n(I_n(\hat{Y};I)+3\epsilon)}\, d\hat{y}^n \Bigg)^{2^{nR_1^{(n)}}} di^n + \epsilon \quad (4.27)$$

$$\leq \int_{i^n:\ i^n \in T_I^n(\epsilon)} p_{I^n}(i^n) \Bigg(1 - \int_{\hat{y}^n:(i^n,\hat{y}^n) \in T_{I,\hat{Y}}^n(\epsilon)} p_{\hat{Y}^n|I^n}(\hat{y}^n|i^n)\, d\hat{y}^n + e^{-2^{n(R_1^{(n)} - I_n(\hat{Y};I)-3\epsilon)}} \Bigg) di^n + \epsilon \quad (4.28)$$

$$\leq 1 - \int_{i^n,\hat{y}^n:(i^n,\hat{y}^n) \in T_{I,\hat{Y}}^n(\epsilon)} p_{I^n,\hat{Y}^n}(i^n,\hat{y}^n)\, di^n d\hat{y}^n + e^{-2^{n(R_1^{(n)} - I_n(\hat{Y};I)-3\epsilon)}} + \epsilon$$

$$= 1 - \Pr\{(I^n, \hat{Y}^n) \in T_{I,\hat{Y}}^n(\epsilon)\} + e^{-2^{n(R_1^{(n)} - I_n(\hat{Y};I)-3\epsilon)}} + \epsilon$$

$$\leq e^{-2^{n(R_1^{(n)} - I_n(\hat{Y};I)-3\epsilon)}} + 2\epsilon \quad (4.29)$$

where (4.24) holds because $\tilde{Y}^n(1,q)$ is independent of I^n for all q and the second integral in the preceding expression is equal to $\Pr\{I^n \notin T_I^n(\epsilon)\}$ which is less than ϵ (by Lemma 4.2). Also, (4.25) is true because all $\tilde{Y}^n(1,q)$ are independently generated for all q. The distribution $p_{\hat{Y}^n}$ was used in the inner integral of (4.26), since it is equal to $p_{\tilde{Y}^n}$ by construction. Also, Lemma 4.4 was used in (4.27), Lemma 4.5 was used in (4.28), and (4.29) follows because $\Pr\{(I^n, \hat{Y}^n) \in T_{I,\hat{Y}}^n(\epsilon)\} > 1 - \epsilon$

(from Lemma 4.2).

In order for (4.29) to be made arbitrarily small, it suffices that

$$R_1^{(n)} \geq I_n(\hat{Y}; I) + 3\epsilon$$

or, equivalently,

$$R_W^{(n)} \leq R_Q^{(n)} - I_n(\hat{Y}; I) - 3\epsilon \tag{4.30}$$

Thus, for arbitrarily small ϵ, choosing a sufficiently large n, we can have $\Pr(E_1) < 4\epsilon$, irrespectively of the covariance matrices K_I, K_V, as long as (4.30) is satisfied. Moreover, the distortion constraint (4.6) is satisfied (within ϵ), by virtue of the choice of $K_{I\hat{Y}}, K_{\hat{Y}}$ in the definition of the typical set $T^n_{I,\hat{Y}}(\epsilon)$.

Behavior of $Pr(E_2)$: To show that $\Pr(E_2) \to 0$, it suffices to show that the triplet $(I^n, \hat{Y}^n(1), Z^n)$ lies in $T^n_{I,\hat{Y},Z}(\epsilon)$ with probability approaching unity asymptotically. In the previous paragraph, we showed that $\Pr\{(I^n, \hat{Y}^n(1)) \in T^n_{I,\hat{Y}}(\epsilon)\} \to 1$. However, we cannot claim that $(I^n, \hat{Y}^n(1), Z^n)$ is jointly Gaussian and thereby immediately show that $\Pr(E_2) \to 0$ through the use of Lemma 4.2. We observe the following:

$$\begin{aligned}
&\Pr\{(I^n, \hat{Y}^n(1), Z^n) \in T^n_{I,\hat{Y},Z}(\epsilon)\} \\
&= \int_{\substack{i^n,\hat{y}^n,z^n: \\ (i^n,\hat{y}^n,z^n)\in T^n_{I,\hat{Y},Z}(\epsilon)}} p_{I^n,\hat{Y}^n(1),Z^n}(i^n, \hat{y}^n, z^n)\, dz^n di^n d\hat{y}^n \\
&= \int_{\substack{i^n,\hat{y}^n: \\ (i^n,\hat{y}^n)\in T^n_{I,\hat{Y}}(\epsilon)}} p_{I^n,\hat{Y}^n(1)}(i^n, \hat{y}^n) \int_{\substack{z^n: \\ (i^n,\hat{y}^n,z^n)\in T^n_{I,\hat{Y},Z}(\epsilon)}} p_{Z^n|I^n,\hat{Y}^n(1)}(z^n|i^n, \hat{y}^n)\, dz^n di^n d\hat{y}^n
\end{aligned} \tag{4.31}$$

We now note that the conditional distribution of Z^n given $I^n, \hat{Y}^n(1)$ is equal to the unconditional distribution of V^n (which is independent of $I^n, \hat{Y}^n(1)$). That is,

$$p_{Z^n|I^n,\hat{Y}^n(1)}(z^n|i^n, \hat{y}^n) = p_{V^n}(z^n - B_A\hat{y}^n)$$

On the other hand, $p_{V^n}(z^n - B_A\hat{y}^n) = p_{\tilde{Z}^n|I^n,\hat{Y}}(z^n|i^n,\hat{y})$, where $(I^n,\hat{Y}^n,\tilde{Z}^n)$ are jointly Gaussian with the same covariance matrix as in the definition of $T^n_{I,\hat{Y},Z}(\epsilon)$. Moreover, for $(i^n,\hat{y}^n) \in T^n_{I,\hat{Y}}(\epsilon)$, the set $\{z^n : (i^n,\hat{y}^n,z^n) \in T^n_{I,\hat{Y},Z}(\epsilon)\}$ is the same as the set $\{z^n : z^n \in T^n_Z(\epsilon) \ \wedge \ z^n \in T^n_{Z|\hat{Y}^n=\hat{y}^n,I^n=i^n}(\epsilon)\}$. This is true because when $(i^n,\hat{y}^n,z^n) \in T^n_{I,\hat{Y},Z}(\epsilon)$, then $p_{\tilde{Z}^n|I^n,\hat{Y}^n}(z^n|i^n,\hat{y}^n)$ is approximately (within ϵ) equal to $2^{-h(Z^n|I^n,\hat{Y}^n)}$, and also

$$\begin{aligned}
& h(Z^n|I^n,\hat{Y}^n) \\
&= \int_{i^n,\hat{y}^n} p_{I^n,\hat{Y}^n}(i^n,\hat{y}^n) h(Z^n|I^n=i^n,\hat{Y}^n=\hat{y}^n) \\
&= \int_{i^n,\hat{y}^n} p_{I^n,\hat{Y}^n}(i^n,\hat{y}^n) h(V^n) \qquad (4.32) \\
&= h(V^n) \\
&= h(V^n|I^n=i^n,\hat{Y}^n=\hat{y}^n) \\
&= h(Z^n|I^n=i^n,\hat{Y}^n=\hat{y}^n)
\end{aligned}$$

where (4.32) and the remaining equalities hold because of the independence between V^n and $(I^n,\hat{Y}^n)$.

Hence, (4.31) equals

$$\begin{aligned}
&\int_{\substack{i^n,\hat{y}^n: \\ (i^n,\hat{y}^n)\in T^n_{I,\hat{Y}}(\epsilon)}} p_{I^n,\hat{Y}^n(1)}(i^n,\hat{y}^n) \int_{\substack{z^n: z^n\in T^n_Z(\epsilon) \ \wedge \\ z^n\in T_{Z|I^n=i^n,\hat{Y}^n=\hat{y}^n}}} p_{\tilde{Z}^n|I^n,\hat{Y}^n}(z^n|i^n,\hat{y}^n)\, dz^n\, di^n\, d\hat{y}^n = \\
&\int_{\substack{i^n,\hat{y}^n: \\ (i^n,\hat{y}^n)\in T^n_{I,\hat{Y}}(\epsilon)}} p_{I^n,\hat{Y}^n(1)}(i^n,\hat{y}^n) \times \\
&\times \Pr\{\tilde{Z}^n \in (T^n_Z(\epsilon) \cap T^n_{Z|I^n=i^n,\hat{Y}^n=\hat{y}^n}(\epsilon))|I^n=i^n,\hat{Y}^n=\hat{y}^n\}\, di^n\, d\hat{y}^n \qquad (4.33)
\end{aligned}$$

Since $\tilde{Z}^n$ given $(I^n=i^n,\hat{Y}^n=\hat{y}^n)$ is Gaussian, Lemma 4.1 holds. Therefore, for all $(i^n,\hat{y}^n) \in T^n_{I,\hat{Y}}(\epsilon)$, $\Pr\{\tilde{Z}^n \in T^n_{Z|I^n=i^n,\hat{Y}^n=\hat{y}^n}(\epsilon)|I^n=i^n,\hat{Y}^n=\hat{y}^n\} > 1-\epsilon$. Also,

$\Pr\{\tilde{Z}^n \in T_Z^n(\epsilon)\} > 1-\epsilon$ and $\Pr\{(I^n, \hat{Y}^n(1)) \in T_{I,\hat{Y}}^n(\epsilon)\} \to 1$. Then, it can be easily proved that $\Pr\{\tilde{Z}^n \in (T_Z^n(\epsilon) \cap T_{Z|I^n=i^n,\hat{Y}^n=\hat{y}^n}^n(\epsilon))|I^n = i^n, \hat{Y}^n = \hat{y}^n\} > 1-\epsilon$ for all $(i^n, \hat{y}^n) \in T' \subset T_{I,\hat{Y}}^n(\epsilon)$, such that $\Pr\{(I^n, \hat{Y}^n(1)) \in T'\} > 1-\epsilon$. Hence (4.33) equals at least $(1-\epsilon)\Pr\{(I^n, \hat{Y}^n(1)) \in T_{I,\hat{Y}}^n(\epsilon)\} > (1-\epsilon)^2$. Thus, $\Pr(E_2) < 1-(1-\epsilon)^2 < \epsilon$ for sufficiently large n and for any non-singular matrices $K_I, K_{\hat{Y}}, K_{I\hat{Y}}, K_V$.

Behavior of $Pr(E_3)$:

$$\begin{aligned}
\Pr(E_3) &= \Pr\{\exists\, w \neq 1 : (I^n, \hat{Y}^n(w), Z^n) \in T_{I,\hat{Y},Z}^n(\epsilon)\} \\
&\leq \sum_{w=2}^{2^{nR_W^{(n)}}} \Pr\{(I^n, \hat{Y}^n(w), Z^n) \in T_{I,\hat{Y},Z}^n(\epsilon)\} \\
&= (2^{nR_W^{(n)}} - 1)\Pr\{(I^n, \hat{Y}^n(2), Z^n) \in T_{I,\hat{Y},Z}^n(\epsilon)\} \qquad (4.34)
\end{aligned}$$

where the last equality is due to the symmetry in the random code generation. Since $\Pr\{(I^n, \hat{Y}^n(2)) \in T_{I,\hat{Y}}^n(\epsilon)\} \to 1$ and by construction, Z^n is independent of $\hat{Y}^n(2)$ given I^n, a standard argument (cf. the proof of Theorem 8.6.1 in [33]) yields

$$\Pr\{(I^n, \hat{Y}^n(2), Z^n) \in T_{I,\hat{Y},Z}^n(\epsilon)\} \leq 2^{-n(I_n(Z;\hat{Y}|I)-(\epsilon/2))}$$

where the conditional mutual information is computed with respect to the Gaussian joint distribution defined earlier. Thus, if

$$R_W^{(n)} \leq I_n(Z;\hat{Y}|I) - \epsilon \qquad (4.35)$$

it follows that the upper bound on $\Pr(E_3)$ in (4.34) vanishes asymptotically.

Thus, combining (4.30) and (4.35) and letting $\epsilon \to 0$, we obtain the achievable rate

$$R_W^{(n)} \;\leq\; \min\,\{R_Q^{(n)} - I_n(\hat{Y};I),\; I_n(Z;\hat{Y}|I)\} \qquad (4.36)$$

Since all distributions are Gaussian in the computation of the mutual information

quantities in (4.36), using Schur's formula [43]

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| \cdot |A - BD^{-1}C|$$

it is easy to see that (4.36) takes the form

$$R_W^{(n)} \leq \min \left\{ R_Q^{(n)} - \frac{1}{2n} \log \left(\frac{|K_I|}{|K_I - K_{I\hat{Y}} K_{\hat{Y}}^{-1} K_{I\hat{Y}}^t|} \right), \right.$$
$$\left. \frac{1}{2n} \log \left(\frac{|B_A(K_{\hat{Y}} - K_{\hat{Y}I} K_I^{-1} K_{\hat{Y}I}^t) B_A^t + K_V|}{|K_V|} \right) \right\} \tag{4.37}$$

Then by maximizing (4.37) with respect to non-singular $K_{\hat{Y}}, K_{I\hat{Y}}$ such that the distortion constraint (4.6) is met, we obtain the required result.

By using a standard expurgation argument, we can now show the existence of a deterministic code which achieves $\mathcal{R}_D^{\text{general, gauss}}$ with vanishing probability of error.

■

4.3 Special Cases

In this section, we consider special cases of the general Gaussian watermarking channel. Specifically, we find simple expressions for $\mathcal{R}_D^{\text{general, gauss}}$, in the following situations:

- *Parallel Gaussian model:* V^n consists of independent, but not necessarily identically distributed components. That is, the j-th element of V^n is zero-mean Gaussian with variance τ_j^2. Also, we assume that I^n has independent components, i.e., the j-th element of I^n is zero-mean Gaussian with variance σ_j^2. Moreover, B_A is a diagonal matrix, i.e., $B_A = \text{diag}(\beta_1, \ldots, \beta_n)$.

- *Blockwise independent model:* I^n and V^n are blockwise memoryless. That is, they consist of n/L blocks of L elements each, and the joint distribution

of I^n, V^n within each block is zero-mean Gaussian with arbitrary covariance matrices $K_I^{(L)}, K_V^{(L)}$ respectively. Moreover, each block is independent of each other. Similarly, B_A has a blockwise-diagonal form; specifically,

$$B_A = \begin{bmatrix} B_A^{(L)} & 0 & \cdots & 0 \\ 0 & B_A^{(L)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_A^{(L)} \end{bmatrix}$$

where $B_A^{(L)}$ is a $L \times L$ matrix.

We now analyze each special case in more detail.

4.3.1 Parallel Gaussian Model

In order to find the matrices $K_{\hat{Y}}, K_{I\hat{Y}}$ that attain the maximum in the definition of $\mathcal{R}_D^{\text{general, gauss}}$, we consider the converse part of section 4.2. We will first find upper bounds on (4.12) and (4.20), and then show that these upper bounds are attainable by a particular selection of matrices $K_{\hat{Y}}, K_{I\hat{Y}}$.

From (4.12) and (4.14), we obtain the following (we let $\epsilon \to 0$ for simplicity, and we set M to be such that $M\hat{Y}^n$ is the MMSE linear estimator of I^n given $\hat{Y}^n$):

$$\begin{aligned} R_W^{(n)} &\leq R_Q^{(n)} - n^{-1}h(I^n) + n^{-1}h(I^n - M\hat{Y}^n) \\ &\leq R_Q^{(n)} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\sigma_j^2 + \frac{1}{n}\sum_{j=1}^{n} h(I_j - M^{(j)}\hat{Y}^n) \qquad (4.38) \\ &\leq R_Q^{(n)} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\sigma_j^2 \\ &\quad + \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)E[(I_j - M^{(j)}\hat{Y}^n)^2] \qquad (4.39) \\ &\leq R_Q^{(n)} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\sigma_j^2 + \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)E[(I_j - \mu_j\hat{Y}_j)^2] \qquad (4.40) \end{aligned}$$

$$= R_Q^{(n)} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log\left(\frac{\sigma_j^2}{\sigma_j^2\sin^2(\phi_j)}\right) \tag{4.41}$$

$$= R_Q^{(n)} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(\gamma_j) \tag{4.42}$$

where (4.38) holds because the sum of the individual entropies is greater than the entropy of the vector, and $M^{(j)}$ is the j-th row of the matrix M; (4.39) is the Gaussian upper bound to the entropy; (4.40) is true because $E[(I_j - \mu_j\hat{Y}_j)^2] \geq E[(I_j - M^{(j)}\hat{Y}^n)^2]$ for any scalar μ_j (recall that $M^{(j)}\hat{Y}^n$ is the MMSE linear estimator of I_j given $\hat{Y}^n$), ϕ_j in (4.41) is the angle between I_j and $\hat{Y}_j$ in the L_2 space of second moments (similar to Figure 2.4), and we define $\gamma_j = \sin^{-2}(\phi_j)$ in (4.42).

Thus, as it can be easily verified, the upper bound (4.42) can be achieved by the matrices $K_{\hat{Y}}^* = \mathrm{diag}(\gamma_1 P_W(\sigma_1^2, \gamma_1, D_1), \ldots, \gamma_n P_W(\sigma_n^2, \gamma_n, D_n))$ and $K_{I\hat{Y}}^* = \mathrm{diag}(\sqrt{(\gamma_1 - 1)\sigma_1^2 P_W(\sigma_1^2, \gamma_1, D_1)}, \ldots, \sqrt{(\gamma_n - 1)\sigma_n^2 P_W(\sigma_n^2, \gamma_n, D_n)})$, where $D_j \triangleq E[(I_j - \hat{Y}_j)^2]$ and

$$P_W(\sigma_j^2, \gamma_j, D_j) \triangleq \frac{\gamma_j(\sigma_j^2 + D_j) - 2\sigma_j^2 + 2\sqrt{\sigma_j^2(\gamma_j D_j - \sigma_j^2)(\gamma_j - 1)}}{\gamma_j^2}$$

As we shall now prove, $K_{\hat{Y}}^*$ and $K_{I\hat{Y}}^*$ maximize the second upper bound of $R_W^{(n)}$.

From (4.20), we obtain (for $\epsilon \to 0$ and linear MMSE matrix Λ):

$$\begin{aligned}
R_W^{(n)} &\leq n^{-1}h(B_A(\hat{Y}^n - \Lambda I^n) + V^n) - n^{-1}h(V^n) \\
&\leq \frac{1}{n}\sum_{j=1}^{n} h(\beta_j(\hat{Y}_j - \Lambda^{(j)}I^n) + V_j) - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(2\pi e)\tau_j^2 && (4.43)\\
&\leq \frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{2}\log(2\pi e)\beta_j^2(E[(\hat{Y}_j - \Lambda^{(j)}I^n)^2] + \tau_j^2) - \frac{1}{2}\log(2\pi e)\tau_j^2\right] && (4.44)\\
&\leq \frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{2}\log(2\pi e)\beta_j^2(E[(\hat{Y}_j - \lambda_j I_j)^2] + \tau_j^2) - \frac{1}{2}\log(2\pi e)\tau_j^2\right] && (4.45)\\
&= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log\left(1 + \frac{\beta_j^2 P_W(\sigma_j^2, \gamma_j, D_j)}{\tau_j^2}\right) && (4.46)
\end{aligned}$$

where (4.43) is obtained by the usual upper bound by the entropy of independent random variables; (4.44) is the Gaussian entropy upper bound (where $\Lambda^{(j)}$ is the j-th row of Λ); (4.45) holds because $\lambda_j I_j$ cannot be a better estimator of $\hat{Y}_j$ (given I^n) than ΛI^n; and finally, (4.46) is true because we choose λ_j such that $\lambda_j I_j$ is the MMSE linear estimator of $\hat{Y}_j$ given I_j.

Again, it can be easily verified that $K^*_{\hat{Y}}$ and $K^*_{I\hat{Y}}$ attain the upper bound (4.46). Hence, $\mathcal{R}_D^{\text{general, gauss}}$ takes the form:

$$\mathcal{R}_D^{\text{general, gauss (1)}} = \Bigg\{ (R_Q^{(n)}, R_W^{(n)}) :$$
$$R_Q^{(n)} \geq \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2}\log\left(\frac{\sigma_j^2}{\min\{\xi, \sigma_j^2\}}\right), \text{ where } \xi \text{ satisfies } \sum_{j=1}^{n}\min\{\xi,\sigma_j^2\} = nD$$
$$R_W^{(n)} \leq \max_{\substack{\{\gamma_j, D_j\}: \\ \sum_{j=1}^n D_j = nD \\ \gamma_j \geq \max\{1, \sigma_j^2/D_j\} \\ \sum_{j=1}^n \log(\gamma_j) \leq 2nR_Q^{(n)}}} \min\Bigg\{ R_Q^{(n)} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log(\gamma_j),$$
$$\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}\log\left(1 + \frac{\beta_j^2 P_W(\sigma_j^2, \gamma_j, D_j)}{\tau_j^2}\right)\Bigg\}\Bigg\}$$

Hence, under this memoryless attack scenario, the optimum hiding strategy is also memoryless (but not necessarily identically distributed). A similar conclusion (but for the case of no compression) was reached in [9] as well. Unfortunately, the optimal values for $\{\gamma_j, D_j\}$ (which attain the maximum in the definition of $\mathcal{R}_D^{\text{general, gauss (1)}}$) are difficult to determine analytically in the general case, mainly due to the complexity of $P_W(\cdot,\cdot,\cdot)$. However, we were able to perform a numerical optimization, for the following simple case:

$$(\forall\ 1 \leq j \leq n_1) \qquad \sigma_j^2 = \sigma_1^2, \quad \tau_j^2 = \tau_1^2, \quad \beta_j = \beta_1$$
$$(\forall\ n_1 + 1 \leq j \leq n) \qquad \sigma_j^2 = \sigma_2^2, \quad \tau_j^2 = \tau_2^2, \quad \beta_j = \beta_2$$

for some $\sigma_1^2, \sigma_2^2, \tau_1^2, \tau_2^2, \beta_1, \beta_2, n_1$. It is easy to establish that the values of $\{\gamma_j, D_j\}$ that maximize the upper bound on $R_W^{(n)}$ in the definition of $\mathcal{R}_D^{\text{general, gauss (1)}}$ are of the form:

$$\begin{aligned} (\forall\ 1 \le j \le n_1) \qquad & \gamma_j^* = \gamma_1^*, \qquad D_j^* = D_1^*, \\ (\forall\ n_1 + 1 \le j \le n) \qquad & \gamma_j^* = \gamma_2^*, \qquad D_j^* = D_2^*, \end{aligned}$$

due to the symmetry of the optimization equations with Lagrange multipliers. Thus, $\mathcal{R}_D^{\text{general, gauss (1)}}$ becomes:

$$\begin{aligned} \mathcal{R}_D^{\text{general, gauss (1)}} = \Bigg\{ & (R_Q, R_W) : \\ & R_Q \ge \frac{\nu}{2} \log\left(\frac{\sigma_1^2}{\min\{\xi, \sigma_1^2\}}\right) + \frac{1-\nu}{2} \log\left(\frac{\sigma_2^2}{\min\{\xi, \sigma_2^2\}}\right), \\ & \text{where } \xi \text{ satisfies: } \nu \min\{\xi, \sigma_1^2\} + (1-\nu)\min\{\xi, \sigma_2^2\} = D \\ & R_W \le \max_{\substack{\{\gamma_1, D_1, \gamma_2, D_2\}: \\ \nu D_1 + (1-\nu) D_2 = D \\ \gamma_j \ge \max\{1, \sigma_j^2/D_j\} \\ \nu \log(\gamma_1) + (1-\nu)\log(\gamma_2) \le 2R_Q^{(n)}}} \min\ \Bigg\{ R_Q - \frac{\nu}{2}\log(\gamma_1) - \frac{1-\nu}{2}\log(\gamma_2), \\ & \frac{\nu}{2}\log\left(1 + \frac{\beta_1^2 P_W(\sigma_1^2, \gamma_1, D_1)}{\tau_1^2}\right) + \frac{1-\nu}{2}\log\left(1 + \frac{\beta_2^2 P_W(\sigma_2^2, \gamma_2, D_2)}{\tau_2^2}\right) \Bigg\} \Bigg\} \end{aligned}$$

where $\nu = n_1/n \le 1$.

Figure 4.1 shows achievable rate regions and the optimal values of γ_1, γ_2, D_1 (as functions of R_Q), as determined from a numerical optimization. We considered the following cases:

(a) $\sigma_1^2 = 4, \sigma_2^2 = 3, \tau_1^2 = \tau_2^2 = 2, \beta_1 = \beta_2 = 1, D = 1, \nu = 0.5$.

Then, the minimum value for R_Q is $\frac{1}{4}\log(4) + \frac{1}{4}\log(3) = 0.8962$ ($\xi = 1$). The maximum value of R_W (when $R_Q = \infty$) is $\frac{1}{2}\log(1 + \frac{1}{2}) = 0.2925$. Note that the values of σ_1^2, σ_2^2 do not affect the maximum R_W (since at infinite R_Q, the

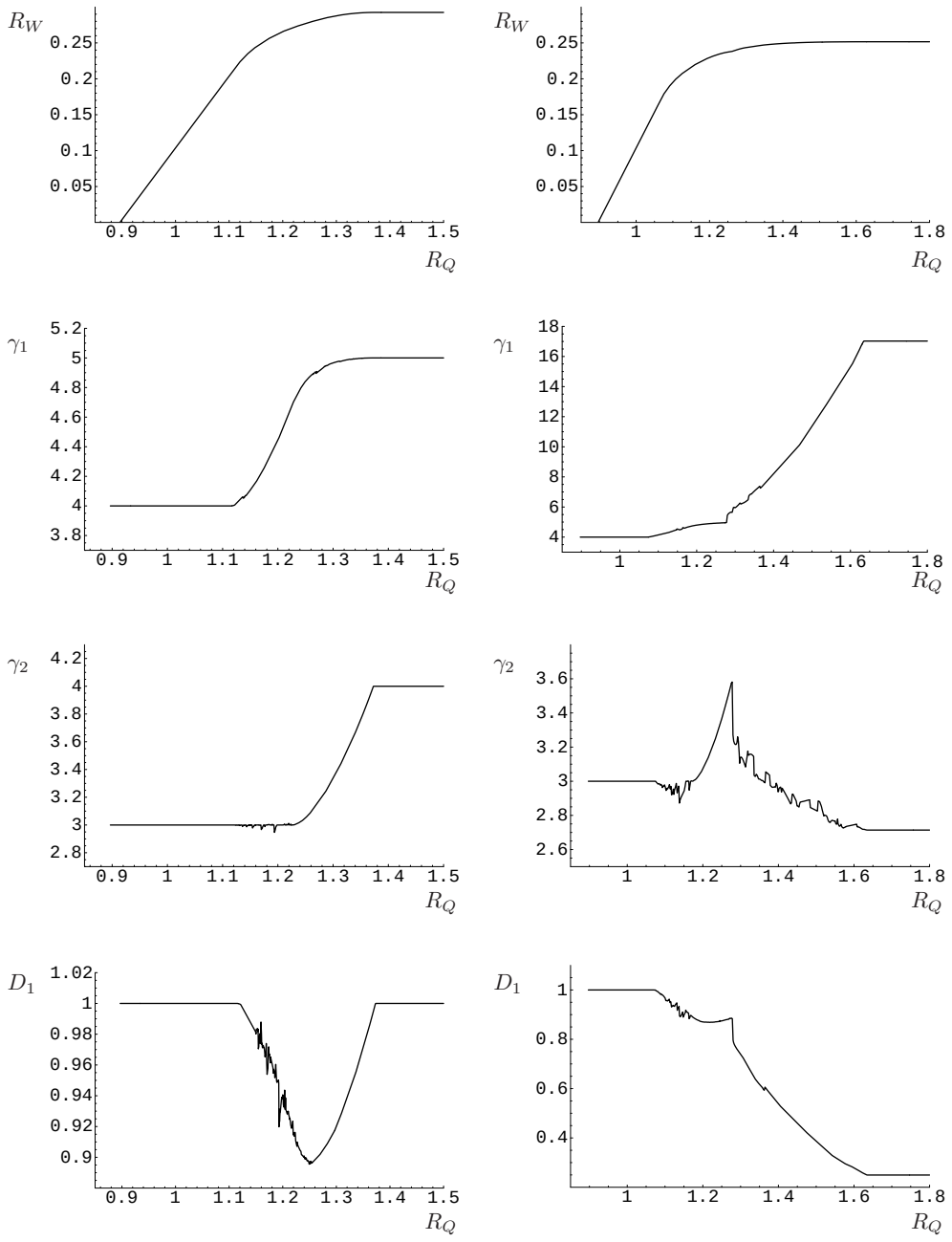


Figure 4.1: The rate region and the optimal values for γ_1, γ_2, D_1 as functions of R_Q for the two examples (a), (b) of parallel Gaussian channels.

watermark embedding is additive and the original image can be subtracted completely at the decoder).

(b) $\sigma_1^2 = 4, \sigma_2^2 = 3, \tau_1^2 = 3.5, \tau_2^2 = 2, \beta_1 = \beta_2 = 1, D = 1, \nu = 0.5$.
The minimum R_Q is again 0.8962 and the maximum R_W is $\frac{1}{4}\log(1 + \frac{0.25}{3.5}) + \frac{1}{4}\log(1 + \frac{1.75}{2}) = 0.2516$ (as determined by optimal water-filling in a parallel Gaussian channel with noise variances $3.5, 2$ and total signal power 2).

The various "squiggles" which appear on the plots are due to numerical artifacts in the optimization. Also, D_2 can be easily determined from D_1 (since $D_2 = \frac{D - \nu D_1}{1-\nu}$).

We now observe the following:

- Both rate regions have three rate regimes: (i) Low R_Q regime, in which the maximum R_W is a linear function of R_Q with slope 1; the first part of the $\min(\cdot, \cdot)$ in the expression for $\mathcal{R}_D^{\text{general, gauss}}$ is dominant, as was also observed in the memoryless case of Chapter 2; γ_1, γ_2, D_1 remain constant in this regime, taking the optimal reverse water-filling values $4, 3, 1$ respectively (i.e., the ones that minimize $n^{-1}I(I^n; \hat{Y}^n)$). (ii) Intermediate R_Q regime, where the upper boundary of the rate region is "curved", corresponding to the second part of the $\min(\cdot, \cdot)$ of $\mathcal{R}_D^{\text{general, gauss}}$; in this case, γ_1, γ_2 both change such that $P_W(\sigma_j^2, \gamma_j, D_j)$ increases until it reaches its maximum value, D_j. (iii) High R_Q regime, where γ_1, γ_2, D_1 and the maximum R_W remain constant; their values correspond to optimum water-filling in a Gaussian parallel channel when $R_Q = \infty$. For case (a) these values are: $D_1 = 1, \gamma_1 = 1 + \sigma_1^2/D_1 = 5, \gamma_2 = 1 + \sigma_2^2/D_2 = 4$ and for case (b) they are: $D_1 = 0.25, \gamma_1 = 17, \gamma_2 = 2.714$.

- As in the memoryless case of Chapter 2, the low and high R_Q regimes have

the same impact here as well; at low quantization rates the Gaussian noise does not degrade the performance of the system, and at high (but finite) quantization rates, the compression does not hinder the watermark detection.

4.3.2 Blockwise Independent Model

Various blockwise models were also considered in [7, 25, 46, 9]. Our formulation is analogous to the one in [9]. Similarly to the previous subsection, here too we obtain the expression for $\mathcal{R}_D^{\text{general, gauss}}$ by deriving achievable upper bounds to expressions (4.12) and (4.20).

Let

$$M = \begin{bmatrix} M^{(L,1)} & 0 & \cdots & 0 \\ 0 & M^{(L,2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M^{(L,n/L)} \end{bmatrix} \quad \Lambda = \begin{bmatrix} \Lambda^{(L,1)} & 0 & \cdots & 0 \\ 0 & \Lambda^{(L,2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Lambda^{(L,n/L)} \end{bmatrix}$$

where M, Λ are $n \times n$ matrices and $M^{(L,j)}, \Lambda^{(L,j)}$ are $L \times L$ matrices for all $1 \leq j \leq n/L$. Moreover, $M^{(L,j)}\hat{Y}^{(L,j)}$ is the MMSE linear estimator of $I^{(L,j)}$ given $\hat{Y}^{(L,j)}$, and $\Lambda^{(L,j)}I^{(L,j)}$ is the MMSE linear estimator of $\hat{Y}^{(L,j)}$ given $I^{(L,j)}$ (the notation $X^{(L,j)}$ means the j-th, L-size block of vector X^n). We have again the following chains:

$$\begin{aligned} R_W^{(n)} &\leq R_Q^{(n)} - n^{-1}h(I^n) + n^{-1}h(I^n - M\hat{Y}^n) \\ &\leq R_Q^{(n)} - \frac{1}{n}\frac{n}{L}h(I^L) + \frac{1}{n}\sum_{j=1}^{n/L} h(I^{(L,j)} - M^{(L,j)}\hat{Y}^{(L,j)}) \qquad (4.47) \\ &\leq R_Q^{(n)} - \frac{1}{2L}\log(2\pi e)^L |K_I^{(L)}| + \\ &\quad \frac{1}{n}\sum_{j=1}^{n/L} \frac{1}{2}\log(2\pi e)^L |K_I^{(L)} - K_{I\hat{Y}}^{(L,j)}(K_{\hat{Y}}^{(L,j)})^{-1}K_{\hat{Y}I}^{(L,j)}| \qquad (4.48) \end{aligned}$$

$$\leq \quad R_Q^{(n)} - \frac{1}{2L}\log\left(\frac{1}{n/L}\sum_{j=1}^{n/L}\frac{|K_I^{(L)} - K_{I\hat{Y}}^{(L,j)}(K_{\hat{Y}}^{(L,j)})^{-1}K_{\hat{Y}I}^{(L,j)}|}{|K_I^{(L)}|}\right) \quad (4.49)$$

and

$$\begin{aligned}
R_W^{(n)} &\leq n^{-1}h(B_A(\hat{Y}^n - \Lambda I^n) + V^n) - n^{-1}h(V^n) \\
&\leq \frac{1}{n}\sum_{j=1}^{n/L} h(B_A^{(L)}(\hat{Y}^{(L,j)} - \Lambda^{(L,j)}I^{(L,j)} + V^{(L,j)}) - \frac{1}{L}h(V^L) \\
&\leq \frac{1}{n}\sum_{j=1}^{n/L}\frac{1}{2}\log(2\pi e)^L(|B_A^{(L)}(K_{\hat{Y}}^{(L,j)} - K_{\hat{Y}I}^{(L,j)}(K_I^L)^{-1}K_{I\hat{Y}}^{(L,j)})(B_A^{(L)})^t + K_V^{(L)}|) \\
&\quad - \frac{1}{2L}\log(2\pi e)^L|K_V^{(L)}| \\
&\leq \frac{1}{2L}\log\left(\frac{1}{n/L}\sum_{j=1}^{n/L}\frac{|B_A^{(L)}(K_{\hat{Y}}^{(L,j)} - K_{\hat{Y}I}^{(L,j)}(K_I^L)^{-1}K_{I\hat{Y}}^{(L,j)})(B_A^{(L)})^t + K_V^{(L)}|}{|K_V^{(L)}|}\right)
\end{aligned} \quad (4.50)$$

We observe now that both inequalities (4.49) and (4.50) are consequences of Jensen's inequality [33], and they can be satisfied with equality if and only if

$$(\forall\ 1 \leq j \leq n/L) \qquad K_{\hat{Y}}^{(L,j)} = K_{\hat{Y}}^{(L)}, \qquad K_{\hat{Y}I}^{(L,j)} = K_{\hat{Y}I}^{(L)}$$

From the above it follows that the rate region $\mathcal{R}_D^{\text{general, gauss}}$ becomes

$$\begin{aligned}
&\mathcal{R}_D^{\text{general, gauss (2)}} = \Bigg\{(R_Q^{(n)}, R_W^{(n)}) : \\
&R_Q^{(n)} \geq \frac{1}{2L}\sum_{j=1}^{L}\log\left(\frac{\sigma_j^2}{\min\{\xi,\sigma_j^2\}}\right),\ \text{where } \xi \text{ satisfies } \sum_{j=1}^{L}\min\{\xi,\sigma_j^2\} = LD \\
&R_W^{(n)} \leq \\
&\max_{\substack{K_{I\hat{Y}}^{(L)}, K_{\hat{Y}}^{(L)}: \\ L^{-1}\text{tr}(K_I^{(L)} + K_{\hat{Y}}^{(L)} - 2K_{I\hat{Y}}^{(L)}) \leq D}} \min\ \Bigg\{R_Q^{(n)} - \frac{1}{2L}\log\left(\frac{|K_I^{(L)}|}{|K_I^{(L)} - K_{I\hat{Y}}^{(L)}(K_{\hat{Y}}^{(L)})^{-1}K_{\hat{Y}I}^{(L)}|}\right), \\
&\frac{1}{2L}\log\left(\frac{|B_A^{(L)}(K_{\hat{Y}}^{(L)} - K_{\hat{Y}I}^{(L)}(K_I^{(L)})^{-1}K_{I\hat{Y}}^{(L)})(B_A^{(L)})^t + K_V^{(L)}|}{|K_V^{(L)}|}\right)\Bigg\}\Bigg\}
\end{aligned}$$

where $\sigma_1^2, \ldots, \sigma_L^2$ are the eigenvalues of $K_I^{(L)}$.

In other words, $\mathcal{R}_D^{\text{general, gauss (2)}}$ has the same form as $\mathcal{R}_D^{\text{general, gauss}}$, where n is replaced by L. This is not surprising, since the blocks are independent from each other and therefore the rate region should only depend on the statistics of one block.

Chapter 5

Concluding Remarks

The main focus of this thesis has been the determination of the largest possible rate regions for information-hiding systems that combine watermarking with quantization. In this last chapter, we review the key issues that were involved in our study, and we present a unifying perspective. Finally, we conclude this dissertation with directions for future research.

Single-User Watermarking: In Chapter 2, we derived the relationship between quantization and watermarking rates under the following assumptions: (i) the original image is i.i.d. distributed; (ii) attacks are memoryless; (iii) average distortion constraints are imposed on the watermarker and the attacker; and (iv) the original image is available at the detector (private scenario). We studied the cases of discrete alphabets, as well as continuous alphabets with Gaussian images and attacks (under quadratic distortion measures). Moreover, we considered fixed attacks, as well as optimized attacks. In the latter case, we formulated the game played between the watermarker and the attacker, and we determined the resulting rate region. It is interesting that, in the Gaussian case, the optimal attack is equivalent to optimal compression of a Gaussian source.

Fingerprinting and Collusion Attacks: In Chapter 3, we studied a fingerprinting/quantization system subject to different types of collusion attacks. The assumptions (i)-(iv) listed above were also applicable here. We demonstrated that collusion can significantly reduce the rate region. Furthermore, for attacks involving linear combinations plus Gaussian noise and satisfying a symmetric distortion constraint, we showed that the optimal choice for the colluders is to perform a symmetric attack (i.e., one where all the multiplicative coefficients are the same). Finally, in a public scenario without compression, we proved a multi-user analogue of Costa's result [15]: that the maximum fingerprinting rate achievable is the same as in a private scenario.

General Gaussian Images and Attacks: In Chapter 4 we considered Gaussian images and attacks which are not necessarily stationary or ergodic. We derived a general formula for the rate region in a private scenario and under average quadratic distortion constraints. Moreover, we examined two special cases; namely, the parallel Gaussian model and the blockwise-independent model, and we obtained simpler expressions for the rate region.

5.1 A Common Theme

The results that we derived in this thesis share a common structure. Specifically, the maximum watermarking (or fingerprinting) rate achievable in a joint watermarking/quantization system is given by the general formula:

$$R_W = \max_{\mathcal{C}_{\hat{Y}}} \min \left\{ R_Q - n^{-1}\mathcal{I}_1(\mathcal{C}_{\hat{Y}}; I^n),\ n^{-1}\mathcal{I}_2(\mathcal{C}_{\hat{Y}}; \mathcal{A}|I^n) \right\} \tag{5.1}$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ are mutual information quantities, I^n corresponds to the original image statistics, $\mathcal{C}_{\hat{Y}}$ represents the code used by the information hider and $\mathcal{A}$

represents the attack channel. The two arguments of the minimum in (5.1) lead to the following interpretations:

- $nR_W \leq nR_Q - \mathcal{I}_1(\mathcal{C}_{\hat{Y}}; I^n)$, or, equivalently, $nR_Q \geq nR_W + \mathcal{I}_1(\mathcal{C}_{\hat{Y}}; I^n)$. Thus, the number of bits needed to describe a watermarked/compressed image (i.e., nR_Q) is at least the sum of the number of bits needed to describe the watermark index (i.e., nR_W) and the number of bits needed to describe the original image (i.e., $\mathcal{I}_1(\mathcal{C}_{\hat{Y}}; I^n)$) at some distortion level. One would expect this to be true considering the dual aim (watermarking and compression) of the embedding process. It is surprising that at sufficiently low quantization rates, (i.e., where the first argument of the minimum in (5.1) may prevail), no additional bits may be needed in order to provide immunity against attacks.

- $nR_W \leq \mathcal{I}_2(\mathcal{C}_{\hat{Y}}; \mathcal{A}|I^n)$. Thus, the number of bits needed to describe the watermark index cannot exceed the number of bits recoverable at the output of an attack channel $\mathcal{A}$ with side information I^n. This is consistent with viewing the watermark set as a code for information transmission through this channel. It is surprising that for sufficiently high quantization rates (i.e., where the second argument of the minimum in (5.1) may prevail), compression may not affect the detectability of the watermark.

5.2 Directions for Future Research

There are a number of extensions to the problem of joint watermarking and compression of images treated in this thesis. The most important ones, are as follows:

Unknown Collusion Attacks: In Chapter 3, we assumed that the collusion attack channel is fixed and known to the decoder. As a consequence, the information

hider knows the number of colluders. Such an assumption is not always justified; in practical situations, the decoder needs to estimate the number of colluders and the attack channel. There exist various channel estimation techniques [36] for single-user channels, based on training sequences. It would be interesting to obtain such techniques for multi-user channels, as well.

The Public Scenario: All results derived in this dissertation (except the achievable rates of public QIM of Chapter 2 and the multi-user Costa scheme of Chapter 3) assume a private scenario. We were not able to establish the region of achievable rates in a public scenario, even for the simple Gaussian case of Chapter 2. Such a region should be a subset of the region obtained in the private case; however, whether or not it is a proper subset is still an open problem. Note that in the case of no attacks, it was shown in [8] that public and private scenarios yield the same rate region.

Other Types of Distortion Constraints: It would be interesting to establish rate regions for joint watermarking and compression systems under distortion constraints that do not involve averaging of distortion measures. For example, motivated by the work in [24] and [27] respectively, two possible types of distortion constraints could be: (i) in the almost sure sense, i.e., the distortion between two vectors, does not exceed a threshold with probability one; and (ii) in the large-deviations sense, where the probability that the distortion between two vectors exceeds a threshold is upper-bounded by an exponentially-decaying function of n.

BIBLIOGRAPHY

[1] Robert W. Lucky. Music on hold. *IEEE Spectrum Magazine*, 37(7), July 2000.

[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] F. Petitcolas, R. Anderson, and M. Kuhn. Information hiding - a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999.

[4] M. Barni, F. Bartolini, I.J. Cox, J. Hernandez, and F. Perez-Gonzalez. Digital watermarking for copyright protection: A communications perspective. *IEEE Communications Magazine*, 39(8):90–133, August 2001.

[5] S. Katzenbeisser and F. Petitcolas. *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, 2000.

[6] I.Cox, J. Bloom, and M. Miller. *Digital Watermarking*. Morgan Kaufmann Publishers, 2001.

[7] A. Cohen and A. Lapidoth. The capacity of the vector Gaussian watermarking game. In *Proc. IEEE Int. Symp. on Information Theory*, page 5, Washington, DC, June 2001.

[8] D. Karakos and A. Papamarcou. Fingerprinting, watermarking and quantization of Gaussian data. In *Proc. 39th Allerton Conference on Communication, Control and Computing (Invited Talk)*, Monticello, Illinois, October 2001.

[9] P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding, preprint, available at http://www.ifp.uiuc.edu/∼moulin/paper.html. December 2001.

[10] 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.

[11] W. Trappe, M. Wu, and K.J.R. Liu. Collusion-resistant fingerprinting for multimedia. In *Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP02)*, May 2002.

[12] I.J. Cox, M.L. Miller, and A. McKellips. Watermarking as communications with side information. *Proceedings of the IEEE*, 87(7):1127–1141, 1999.

[13] S. Gel'fand and M. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

[14] C. Heegard and A.A. El Gamal. On the capacity of computer memory with defects. *IEEE Transactions on Information Theory*, 29:731–739, 1983.

[15] M. H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29:439–441, May 1983.

[16] B. Chen and G. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47(4):1423–1443, May 2001.

[17] J. Eggers, J. Su, and B. Girod. Robustness of a blind image watermarking scheme. In *Proc. of the IEEE International Conference on Image Processing*, pages 17–20, Sept. 2000.

[18] W. Yu, A. Sutivong, D. Julian, T. Cover, and M. Chiang. Writing on colored paper. In *Proc. IEEE Int. Symp. on Information Theory*, Washington, DC, June 2001.

[19] A. Cohen and A. Lapidoth. Generalized writing on dirty paper. In *Proc. IEEE Int. Symp. on Information Theory*, page 227, Lausanne, Switzerland, June 2002.

[20] 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.

[21] P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding. In *Proc. IEEE Int. Symp. on Information Theory*, page 19, Sorrento, Italy, June 2000.

[22] J. O'Sullivan and P. Moulin. Some optimal properties of optimal information hiding and information attacks. In *Proc. 39th Allerton Conference on Communication, Control and Computing (Invited Talk)*, Monticello, Illinois, October 2001.

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

[24] A. Cohen and A. Lapidoth. The Gaussian watermarking game. *IEEE Transactions on Information Theory*, 48(6):1639 –1667, June 2002.

[25] A. S. Cohen. *Information Theoretic Analysis of Watermarking Systems*. PhD Thesis, MIT, Cambridge, MA, 2001.

[26] Anelia Somekh-Baruch and N. Merhav. On the watermarking game of the random coding error exponent with large deviations distortion constraints. In *Proc. IEEE Int. Symp. on Information Theory*, page 7, Washington, DC, June 2001.

[27] Anelia Somekh-Baruch and N. Merhav. On the error exponent and capacity games of private watermarking systems. Submitted to the *IEEE Transactions on Information Theory*, available at http://tiger.technion.ac.il/users/merhav, June 2001.

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

[29] Anelia Somekh-Baruch and N. Merhav. On the capacity game of public watermarking systems. In *Proc. IEEE Int. Symp. on Information Theory*, page 223, Lausanne, Switzerland, June 2002. Also, submitted to the *IEEE Trans. on Information Theory*, available at http://tiger.technion.ac.il/users/merhav.

[30] Y. Steinberg and N. Merhav. Identification in the presence of side information with application to watermarking. *IEEE Transactions on Information Theory*, 47(4):1410–1422, May 2001.

[31] 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.

[32] D. Karakos and A. Papamarcou. A relationship between quantization and watermarking rates in the presence of gaussian attacks. In *Proc. 39th Conference on Information Sciences and Systems (CISS-2002)*, Princeton, New Jersey, March 2002.

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

[34] D. Karakos and A. Papamarcou. A relationship between quantization and watermarking rates in the presence of Gaussian attacks, Institute for Systems Research technical report, TR 2001-50, University of Maryland, available at http://www.isr.umd.edu/TechReports. Dec 2001.

[35] D. Karakos and A. Papamarcou. A relationship between quantization and distribution rates of digitally watermarked data, Institute for Systems Research technical report, TR 2000-51, UMD, available at http://www.isr.umd.edu/TechReports. Dec 2000.

[36] A. Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Transactions on Information Theory*, 44(6):2148–2177, October 1998.

[37] D. Blackwell, L. Breiman, and A.J. Thomasian. The capacity of a class of channels. *Ann. Math. Stat.*, 30(4):1229–1241, 1959.

[38] D. Boneh and J. Shaw. Collusion secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44:1897–1905, May 1998.

[39] P. Moulin and A. Briassouli. The Gaussian fingerprinting game. In *Proc. 39th Conference on Information Sciences and Systems (CISS-2002)*, Princeton, New Jersey, March 2002.

[40] W. Trappe, M. Wu, and K.J.R. Liu. Anti-collusion fingerprinting for multimedia. Submitted to the *IEEE Transactions on Signal Processing*, Dec. 2001.

[41] J. Su, J. Eggers, and B. Girod. Capacity of digital watermarks subjected to an optimal collusion attack. In *Proc. European Signal Proc. Conf. (EUSIPCO'00)*, Sept. 2000.

[42] G.R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford Science Publications, 1992.

[43] W. Yu, A. Sutivong, D. Julian, T. Cover, and M. Chiang. Writing on colored paper. To be submitted to the *IEEE Transactions on Information Theory*, available at http://www.stanford.edu/∼weiyu/, 2002.

[44] T. Cover and S. Pombra. Gaussian feedback capacity. *IEEE Transactions on Information Theory*, 35(1):37–43, January 1989.

[45] C. W. Helstrom. *Elements of Signal Detection and Estimation*. Prentice-Hall, 1995.

[46] A. Cohen and A. Lapidoth. Watermarking capacity for Gaussian sources. In *Proc. 39th Allerton Conference on Communication, Control and Computing (Invited Talk)*, Monticello, Illinois, October 2001.