# A Relationship Between Quantization and Watermarking Rates in the Presence of Gaussian Attacks

*by Damianos Karakos, Adrian Papamarcou*

**TR 2001-50**

# A Relationship Between Quantization and Watermarking Rates in the Presence of Gaussian Attacks

*Damianos Karakos* *Adrian Papamarcou*
karakos@eng.umd.edu adrian@eng.umd.edu

Department of Electrical and Computer Engineering
University of Maryland
College Park, MD 20742

**Abstract**

A system which embeds watermarks in $n$-dimensional i.i.d. Gaussian images and distributes them in compressed form is studied. The performance of the system in the presence of Gaussian attacks is considered, and the region of achievable watermarking and quantization rates is established under constraints on image distortion and watermark detectability. The performance of related schemes is also discussed.

## 1 Introduction

Over the last decade, considerable attention has been devoted to information hiding as a means of preserving ownership of intellectual property in multimedia data. Numerous articles (e.g. see [1, 2, 3]) and books (e.g. [4, 5]) explain the basics of information hiding (commonly referred to as watermarking), explore its many practical applications, and evaluate the performance of various watermarking schemes under a variety of attack scenarios.

Two key issues in the design of watermarking schemes are:

- **Transparency:** The hidden message should not interfere perceptually with the host signal (or *covertext* [6]). 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 in the sequel, 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). In the *private* detection scenario, the original image is available to the detector; in the *public* scenario, it is not.

Information hiding has also been studied from an information-theoretic perspective, notably in [7, 8, 9, 10, 11, 12, 6, 13, 14]. The model treated in this paper, which involves *joint* watermarking and image compression, has received less attention in the literature. A brief summary of our model 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.

Previous work involving this model [9, 14], focused on the case where the watermarked/compressed image was not subject to attacks (compression inherently introduces degradation, but cannot be construed as a malicious attack of the type studied in, e.g., [7, 13]). It was shown that, when the original image is i.i.d. Gaussian and an average quadratic distortion constraint is satisfied, 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 $R_Q$ is the quantization rate, $R_W$ is the watermarking rate, $P_I$ is the image variance (per dimension or pixel) and $D$ is the average quadratic distortion between the original image and the watermarked/compressed image. Since this result is subsumed in the analysis of this paper, no further discussion is in order here except for the following observation. The rates above are compatible with a naive encoding scheme whereby $nR_W$ bits are used to encode the watermark index and $n(R_Q - R_W)$ bits to represent the original image, where

$$R_Q - R_W > \frac{1}{2} \log\left(\frac{P_I}{D}\right) \; .$$

By standard rate-distortion theory for i.i.d. Gaussian sources, there are enough bits to represent the image with average distortion equal to $D$. Yet this scheme is entirely inadequate from a watermarking (or information hiding) perspective, since the image representation does not contain the watermark in any form whatsoever.

An interesting compression/watermarking scheme developed by Chen and Wornell [12] is *Quantization Index Modulation* (QIM), where an ensemble of quantizers—each corresponding to a particular watermark index—is used for compressing the image. The *regular* version of QIM, in which the watermarked image is communicated to the user as an index in a source codebook, is of relevance to our work and will be studied further in Section 4.

In summary, this paper contains final versions of results in [9, 14], together with extensions to the important case where the compressed images are subjected to additive memoryless Gaussian attacks. The main contribution is a coding theorem which establishes the region of all achievable rate pairs $(R_Q, R_W)$ such that the average per-symbol quadratic distortion between the original and the compressed image does not exceed a threshold $D$, and the watermark index is detectable with high probability in a *private* scenario, i.e., assuming that the original image is available to the detector. Achievability results are also presented for regular QIM in the *public* scenario, as well as for certain additive watermarking schemes.

The paper is organized as follows. The description and interpretation of the rate region $\mathcal{R}_{D,D_A}$ consisting of achievable $(R_Q, R_W)$ pairs is given in Section 2. The coding theorem that establishes $\mathcal{R}_{D,D_A}$ is proved in Section 3. Achievability results for other schemes that combine watermarking and compression are presented in Section 4. Finally, conclusions and directions for further research are given in Section 5.

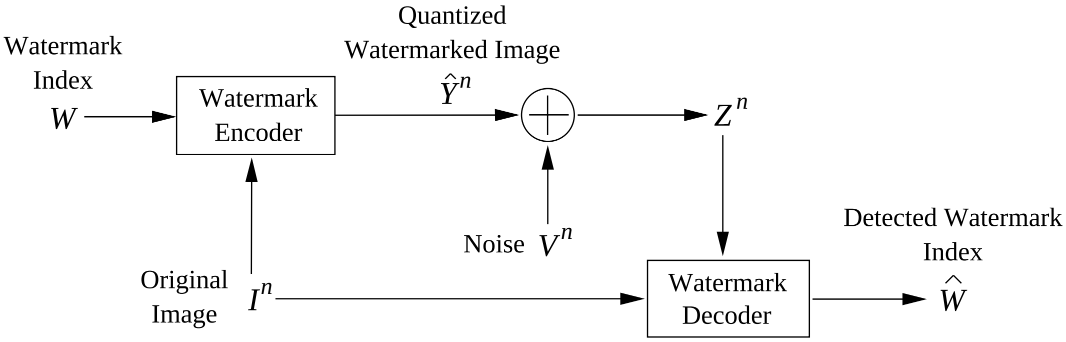


Figure 1: The watermarking/authentication system with quantization

## 2 The Rate Region

The watermarking/quantization system under consideration is shown in Figure 1. In the embedding process, $W$ is the watermark index which is uniformly distributed over a set of size $2^{nR_W}$; $I^n$ is the i.i.d. $n$-dimensional Gaussian image of (per-symbol) variance $P_I$; and $\hat{Y}^n$ is the watermarked/quantized image which can be found in an source codebook of size $2^{nR_Q}$. The attack is modeled as additive i.i.d. Gaussian noise $V^n$ of (per-symbol) variance $D_A$, and is assumed independent of $\hat{Y}^n$. The watermark decoder outputs $\hat{W}$, its estimate of $W$. The transparency and robustness requirements are expressed via the following constraints:

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

$$\Pr\{\hat{W} \neq W\} \to 0, \text{ as } n \to \infty \tag{2}$$

The converse and achievability results of Section 3 establish the following region $\mathcal{R}_{D,D_A}$ of achievable rates $(R_Q, R_W)$:

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

where

$$P_W(\gamma) \triangleq \frac{\gamma(P_I + D) - 2P_I + 2\sqrt{P_I(\gamma D - P_I)(\gamma - 1)}}{\gamma^2} \tag{3}$$

$\mathcal{R}_{D,D_A}$ is the shaded region in Figure 2. Its upper boundary is composed of:

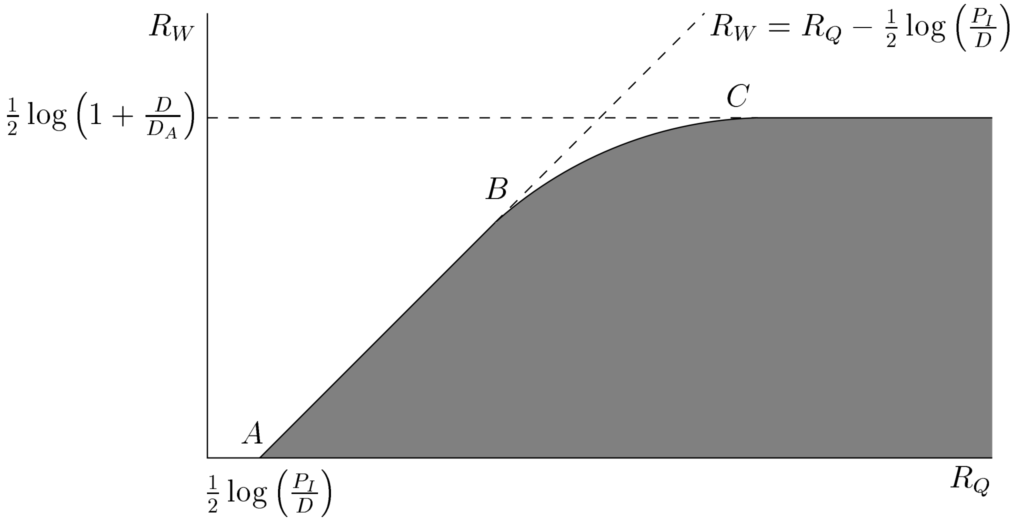


Figure 2: The rate region $\mathcal{R}_{D,D_A}$ of achievable rate pairs $(R_Q, R_W)$.

- The segment $AB$ on the straight line $R_W = R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right)$.

- The curved segment $BC$ defined by the equation

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

  for $R_Q$ in the interval $[\frac{1}{2}\log\left(\frac{P_I}{D}+\frac{P_I-D}{D_A}\right), \frac{1}{2}\log\left(1+\frac{P_I+D}{D_A}+\frac{P_I}{D}\right)]$, i.e., the projection of BC on the $R_Q$-axis. As we shall see later, $R_W$ is also given by the root of a cubic equation.

- A half-line parallel to the $R_Q$-axis with vertex $C$. The $R_W$-ordinate is given by $\frac{1}{2}\log\left(1+\frac{D}{D_A}\right)$.

Two key conclusions can be drawn from Figure 2:

- For quantization rates $R_Q \in \left[\frac{1}{2}\log\left(\frac{P_I}{D}\right), \frac{1}{2}\log\left(\frac{P_I}{D}+\frac{P_I-D}{D_A}\right)\right]$, the watermarking rate $R_W$ can be as high as $R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right)$, 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 $R_Q \geq \frac{1}{2}\log\left(1+\frac{P_I+D}{D_A}+\frac{P_I}{D}\right)$, the maximum watermarking rate is constant and equal to $\frac{1}{2}\log\left(1+\frac{D}{D_A}\right)$. This expression makes sense in the case $R_Q = \infty$, where the distortion in the original image is solely due to watermarking, and where $D$ represents the "signal" power in the AWGN Gaussian attack channel of variance $D_A$—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.

# 3 The Coding Theorem

The coding theorem which establishes the region of all achievable rate pairs $(R_Q, R_W)$, consists of a converse and a direct (achievability) part.

The converse part states that no rate pairs $(R_Q, R_W) \notin \mathcal{R}_{D,D_A}$ are achievable.

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

$$\frac{1}{n}\sum_{i=1}^{n} E(I_i - \hat{Y}_i)^2 \leq D \tag{4}$$

We know from rate-distortion theory [15] that (since $I^n$ is Gaussian distributed) $R_Q$ should be at least as high as the rate-distortion function of a Gaussian source with average distortion $D$. Hence,

$$R_Q \geq \frac{1}{2}\log\left(\frac{P_I}{D}\right) \tag{5}$$

First, we have the following chain of inequalities:

$$\begin{aligned}
R_W &= n^{-1}H(W|I^n, V^n) && (6)\\
&= n^{-1}I(W; \hat{Y}^n|I^n, V^n) + n^{-1}H(W|I^n, \hat{Y}^n, V^n) &&\\
&\leq n^{-1}I(W; \hat{Y}^n|I^n, V^n) + n^{-1}H(W|I^n, Z^n) && (7)\\
&\leq n^{-1}I(W; \hat{Y}^n|I^n, V^n) + \epsilon && (8)\\
&= n^{-1}H(\hat{Y}^n|I^n, V^n) - n^{-1}H(\hat{Y}^n|W, I^n, V^n) + \epsilon &&\\
&\leq n^{-1}H(\hat{Y}^n|I^n) + \epsilon && (9)\\
&= n^{-1}H(\hat{Y}^n) - n^{-1}(H(\hat{Y}^n) - H(\hat{Y}^n|I^n)) + \epsilon &&\\
&\leq R_Q - n^{-1}I(\hat{Y}^n; I^n) + \epsilon && (10)\\
&\leq R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right) + \epsilon && (11)
\end{aligned}$$

where (6) holds because $I^n, V^n$ are independent of $W$, (7) follows from $H(W|I^n, \hat{Y}^n, V^n) = H(W|I^n, Z^n, \hat{Y}^n, V^n) \leq H(W|I^n, Z^n)$, (8) is a consequence of

Fano's inequality, (9) holds because $H(\hat{Y}^n|W, I^n, V^n) = 0$ (since $\hat{Y}^n$ is a function of $W, I^n$), (10) follows from $R_Q \geq n^{-1}H(\hat{Y}^n)$ and (11) holds because $n^{-1}I(\hat{Y}^n; I^n)$ is always greater than or equal to, the rate-distortion function of a Gaussian source of power $P_I$ and average distortion $D$.

We next have:

$$\begin{aligned} R_W &= n^{-1}H(W|I^n) && (12) \\ &= n^{-1}I(W; Z^n|I^n) + n^{-1}H(W|I^n, Z^n) \\ &\leq n^{-1}I(W; Z^n|I^n) + \epsilon && (13) \\ &= n^{-1}h(Z^n|I^n) - n^{-1}h(Z^n|I^n, W) + \epsilon \\ &= n^{-1}h(Z^n|I^n) - n^{-1}h(V^n|I^n, W) + \epsilon && (14) \\ &= n^{-1}h(\hat{Y}^n - \lambda I^n + V^n|I^n) - n^{-1}h(V^n) + \epsilon && (15) \\ &\leq n^{-1}h(\hat{Y}^n - \lambda I^n + V^n) - \frac{1}{2}\log(2\pi e)D_A + \epsilon \\ &\leq \frac{1}{2}\log(2\pi e)(G + D_A) - \frac{1}{2}\log(2\pi e)D_A + \epsilon && (16) \\ &= \frac{1}{2}\log\left(1 + \frac{G}{D_A}\right) + \epsilon && (17) \end{aligned}$$

where (12) holds because $I^n$ is independent of $W$, (13) follows from Fano's inequality, (14) holds because $\hat{Y}^n$ is a function of $I^n, W$, (15) follows from the independence of $V^n$ and $I^n, W$ and (16) holds because the Gaussian distribution provides an upper bound on the entropy of a continuous variable with variance $G + D_A$, where

$$\begin{aligned} G &\triangleq \min_{\lambda \in \mathbf{R}} \frac{1}{n}\sum_{i=1}^{n} E(\hat{Y}_i - \lambda I_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} E(\hat{Y}_i - \lambda_0 I_i)^2 \end{aligned}$$

Figure 3 shows the space $L_2$ of second moments. We can easily compute that

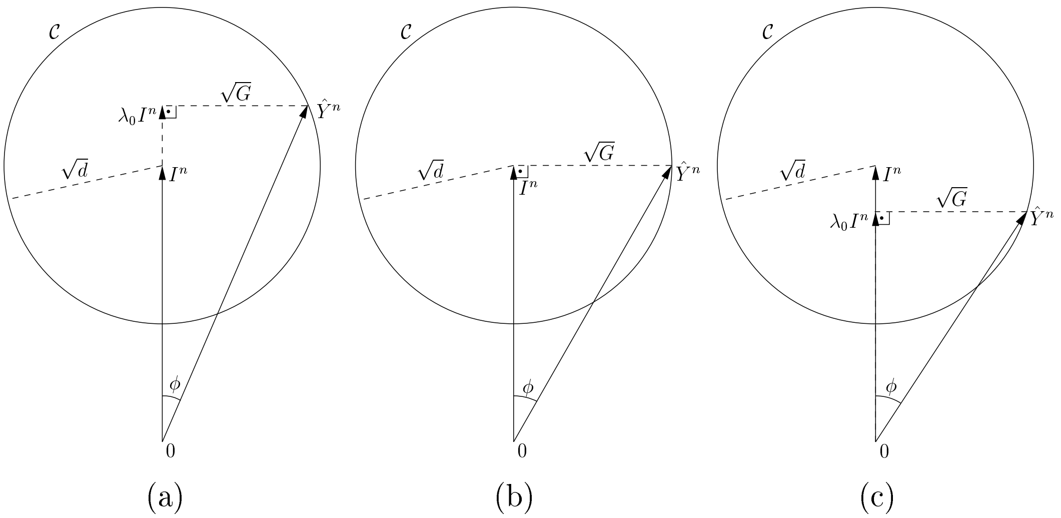
$$\lambda_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{I_i\hat{Y}_i}{P_I}$$

and together with (4) we get

$$\lambda_0 \geq \frac{P_I + P_{\hat{Y}} - D}{2P_I} > 0 \tag{18}$$

where $P_{\hat{Y}} \triangleq \frac{1}{n}\sum_{i=1}^{n} E(\hat{Y}_i^2)$. Also, from the Pythageorean theorem we have

$$\lambda_0^2 P_I + G = P_{\hat{Y}} \tag{19}$$



(a) (b) (c)

Figure 3: The 2nd moment space $\mathcal{L}_2$ spanned by vectors $I^n, \hat{Y}^n$, shown for three different values of $\phi$. The circle $\mathcal{C}$ is the locus of all $\hat{Y}^n$ such that $n^{-1}E||I^n - \hat{Y}^n||^2 = d \leq D$. As $\phi$ increases from 0, $G$ monotonically increases (case (a)) until it reaches its highest value $d$ (case (b)) and then starts decreasing monotonically (case (c)).

Finally, it can be easily shown that

$$\begin{aligned} \frac{1}{n}I(I^n;\hat{Y}^n) &\geq \frac{1}{2}\log\left(\frac{1}{\sin^2(\phi)}\right) \\ &= \frac{1}{2}\log\left(\frac{P_{\hat{Y}}}{G}\right) \end{aligned} \tag{20}$$

Hence, from (10) and (20) we have

$$R_W \leq R_Q - \frac{1}{2}\log\left(\frac{P_{\hat{Y}}}{G}\right) + \epsilon \tag{21}$$

Let us define $\gamma \triangleq P_{\hat{Y}}/G \geq 1$. Then, by (21) and the fact that the minimum value $\frac{1}{n}I(I^n;\hat{Y}^n)$ can take is $\frac{1}{2}\log\left(\frac{P_I}{D}\right)$, we obtain

$$\frac{P_I}{D} \leq \gamma \leq 2^{2R_Q} \tag{22}$$

Moreover, (18) and (19) give

$$G(\gamma-1) = \lambda_0^2 P_I \geq \frac{(P_I+\gamma G - D)^2}{4P_I}$$

from which (solving for $G$) we get

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

and, therefore,

$$G \le P_W(\gamma) \tag{23}$$

where $P_W(\gamma)$ was defined in Section 2. So, combining (17), (21), (22) and (23) we obtain

$$R_W \le \max_{\gamma \in \left[\frac{P_I}{D}, 2^{2R_Q}\right]} \min\left\{R_Q - \frac{1}{2}\log(\gamma), \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)\right\} + \epsilon \tag{24}$$

By taking $\epsilon \to 0$ in (24), and together with (5) we obtain the required result.

*Behavior of $P_W(\gamma)$:* Let's assume that $d = D$. From (20), we get that $\gamma = \sin^{-2}(\phi)$. In Figure 3, it is shown that $G$ (which is equal to $P_W(\gamma)$ for $d = D$) is a continuous function of $\phi$ that increases monotonically as $\phi$ increases from 0 to $\arctan(\sqrt{D/P_I})$, and decreases monotonically when $\phi > \arctan(\sqrt{D/P_I})$. Observe that since $P_I \ge D$, $\phi$ is always between 0 and $\pi/2$ (hence $\sin^{-2}(\phi)$ is a monotonically decreasing function). Now, since sin and log are continuous and monotonous functions, it is obvious that $\frac{1}{2}\log(1 + \frac{P_W(\gamma)}{D_A})$ has the following behavior: (a) it increases monotonically and continuously for $\gamma \in [\frac{P_I}{D}, 1 + \frac{P_I}{D}]$, and (b) it decreases monotonically and continuously for $\gamma > 1 + \frac{P_I}{D}$. Obviously, the highest value of $P_W(\gamma)$ is $D$, achieved for $\gamma = 1 + \frac{P_I}{D}$. So, when $\gamma = \frac{P_I}{D}$ then $\frac{1}{2}\log(1 + \frac{P_W(\gamma)}{D_A}) = \frac{1}{2}\log(1 + \frac{D}{D_A}(1 - \frac{D}{P_I}))$ and when $\gamma = 1 + \frac{P_I}{D}$ then $\frac{1}{2}\log(1 + \frac{P_W(\gamma)}{D_A}) = \frac{1}{2}\log(1 + \frac{D}{D_A})$.

*Behavior of the upper bound (24):* By replacing $\gamma$ with $2^{2\zeta}$ (where $\zeta \in [\frac{1}{2}\log(\frac{P_I}{D}), R_Q]$), (24) can be written as

$$R_W \le \max_{\zeta \in \left[\frac{1}{2}\log(\frac{P_I}{D}), R_Q\right]} \min\left\{R_Q - \zeta,\ \frac{1}{2}\log\left(1 + \frac{P_W(2^{2\zeta})}{D_A}\right)\right\} \tag{25}$$

We consider the following ranges for $R_Q$: (i) $R_Q \in [\frac{1}{2}\log(\frac{P_I}{D}), \frac{1}{2}\log(\frac{P_I}{D} + \frac{P_I - D}{D_A})]$. In this case, $\zeta \in [\frac{1}{2}\log(\frac{P_I}{D}), \frac{1}{2}\log(\frac{P_I}{D} + \frac{P_I - D}{D_A})]$. This corresponds to the situation described in Figure 4(a), and it is obvious that the maximization in (24) is accomplished for $\zeta = \frac{1}{2}\log(\frac{P_I}{D})$. Therefore, (25) is equal to $R_Q - \frac{1}{2}\log(\frac{P_I}{D})$ in this case, and corresponds to the linear segment $(A, B)$ of Figure 2. (ii) $R_Q \in [\frac{1}{2}\log(\frac{P_I}{D} + \frac{P_I - D}{D_A}), \frac{1}{2}\log(1 + \frac{P_I}{D} + \frac{P_I + D}{D_A})]$. Then, as shown in Figure 4(b), the maximin of (25) is attained at the root of the equation $R_Q - \zeta = \frac{1}{2}\log(1 + \frac{P_W(\zeta)}{D_A})$. Then, (25) gives the curved line segment $(B, C)$ of Figure 2. Note that, by virtue of the converse, (25) has to be a concave function of $R_Q$ (we will show in the sequel that this upper bound is also achievable). (iii) $R_Q \ge \frac{1}{2}\log(1 + \frac{P_I}{D} + \frac{P_I + D}{D_A})$. Then, the minimum in (25) is always less than (or equal to) the value of $\frac{1}{2}\log(1+\frac{P_W(\zeta)}{D_A})$ at $\zeta = \frac{1}{2}\log(1+\frac{P_I}{D})$, which is equal to $\frac{1}{2}\log(1+\frac{D}{D_A})$ (see Figure 4(c)). Hence, (25) is equal to $\frac{1}{2}\log(1+\frac{D}{D_A})$ in this case, and corresponds to the horizontal half-line which starts at point $C$ in Figure 2. This upper bound $R_W \leq \frac{1}{2}\log(1+\frac{D}{D_A})$ could also have been obtained using a plausible argument; that $R_W$ cannot be higher than the capacity of an AWGN channel with signal power $D$ and noise power $D_A$, obtained when $R_Q = \infty$.

**Note:** In the special case $D_A = 0$ (no attack) only the proof for the linear leftmost boundary $R_W \leq R_Q - 1/2\log(P_I/D)$ makes sense; the proofs of the other upper bounds give infinite results and should be ignored. The converse then coincides with the channel-coding part of the converse found in [16], or the converse of [14] for $R_F = 0$. ■

We are now going to prove that $\mathcal{R}_{D,D_A}$ is achievable.

**Proof:** (Achievability) Let $\epsilon > 0$. We are going to prove that $\mathcal{R}_{D,D_A}$ is achievable using a random coding argument. Let $W$ be the watermark index, uniformly distributed in $\{1,\ldots,2^{nR_W}\}$, and let $\hat{W}$ be the output of the decoder. We assume that the quantizer operates at a rate $R_Q \geq \frac{1}{2}\log\left(\frac{P_I}{D}\right)$.

First, we will show that the region that lies below the curved line segment $(B,C)$ of Figure 2 is achievable. Consequently, we assume that

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

Our approach uses an idea similar to the private version of the regular QIM [12]: generation of $2^{nR_W}$ quantizers, each one indexed by a different watermark.

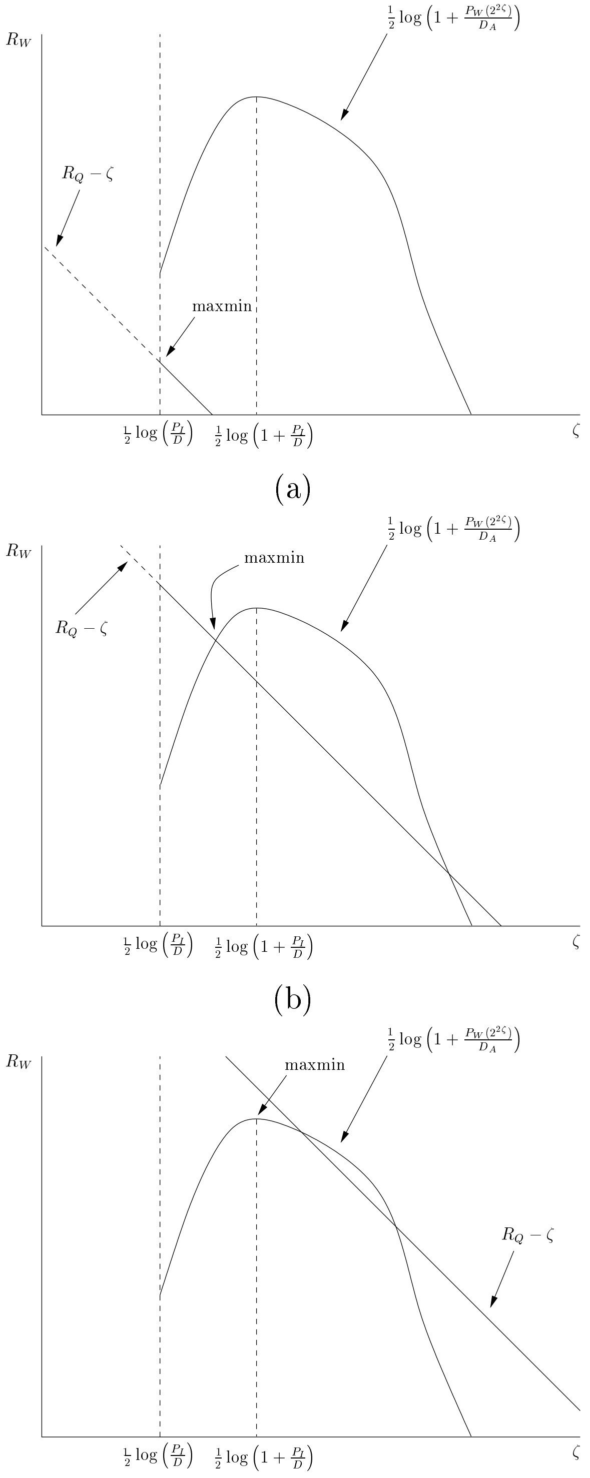
**Generation of codebook:** The encoder generates $2^{nR_W}$ sets, each consisting of $2^{nR_1}$ sequences $\tilde{Y}^n$ each, such that

$$R_Q = R_W + R_1 \tag{26}$$

We denote the sequences of set $w$ by $\{\tilde{Y}^n_w(1),\ldots,\tilde{Y}^n_w(2^{nR_1})\}$. Each one of these sequences is generated i.i.d. $\sim \mathcal{N}(0,P_{\tilde{Y}})$. Furthermore, we require that

$$R_1 \geq \frac{1}{2}\log\left(\frac{P_I}{D}\right) \tag{27}$$

**Embedding:** Given $I^n$ and $W$, the embedder looks into the set indexed by $W$ and tries to find a typical $\tilde{Y}^n_W(q)$ such that $|n^{-1}||I^n - \tilde{Y}^n_W(q)||^2 - D| < \epsilon$ (i.e., $I^n$ and



(a)

(b)

(c)

Figure 4: Plots of $R_Q-\zeta$ and $\frac{1}{2}\log(1+\frac{P_W(\zeta)}{D_A})$ and determination of the maximin point for various values of $R_Q$.

$\tilde{Y}_W^n(q)$ are distortion-typical). We call $T_{I,\hat{Y}}(\epsilon)$ the set of such typical sequences, that corresponds to a joint Gaussian distribution $p_{I,\hat{Y}}$ with covariance matrix

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

where $\gamma \triangleq 2^{2R_1}$ and

$$G\gamma = P_{\hat{Y}} \tag{28}$$

If there exists such a $\tilde{Y}_W^n(q)$, then the encoder outputs $\hat{Y}_W^n = \tilde{Y}_W^n(q)$ (if there are more than one sequences, the encoder will pick the one with the smallest index), otherwise the encoder outputs $\hat{Y}^n = 0$.

**Decoding:** The decoder receives $Z^n = \hat{Y}_W^n + V^n$. Given $I^n$, he tries to find a $\tilde{Y}_{\hat{w}}^n$ such that $(I^n, \tilde{Y}_{\hat{w}}^n, Z^n)$ is jointly typical with respect to some trivariate Gaussian distribution $p_{I,\hat{Y},Z}$ with covariance matrix

$$K_{I,\hat{Y},Z} = \begin{bmatrix} P_I & \sqrt{GP_I(\gamma-1)} & \sqrt{GP_I(\gamma-1)} \\ \sqrt{GP_I(\gamma-1)} & G\gamma & G\gamma \\ \sqrt{GP_I(\gamma-1)} & G\gamma & G\gamma + D_A \end{bmatrix}$$

Let $T_{I,\hat{Y},Z}^n(\epsilon)$ be the set of all typical sequences with respect to $p_{I,\hat{Y},Z}$. If there exists such a $\tilde{Y}_{\hat{w}}^n$, then the decoder outputs $\hat{W} = \hat{w}$, otherwise declares an error.

**Probability of error:** Without loss of generality, we assume that $W = 1$. We have the following error events:

- $E_1$: $\hat{Y}^n = 0$, i.e., there exists no $q \in \{1, \ldots, \gamma\}$ such that $(I^n, \tilde{Y}_1^n(q)) \in T_{I,\hat{Y}}$ (encoding error).
- $E_2$: There exists a $\tilde{Y}_1^n(q) = \hat{Y}_1^n$ such that $(I^n, \hat{Y}_1^n) \in T_{I,\hat{Y}}$ but $(I^n, \hat{Y}_1^n, Z^n) \notin T_{I,\hat{Y},Z}$.
- $E_3$: $(I^n, \hat{Y}_1^n, Z^n) \in T_{I,\hat{Y},Z}$ but there also exists a $k > 1$ such that $(I^n, \hat{Y}_k^n, Z^n) \in T_{I,\hat{Y},Z}$ (decoding error).

So, the probability of error is

$$P_e = \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$$

From the rate-distortion theorem [15] we know that if $R_1 > I(I, \hat{Y})$ then the probability of encoding error $\Pr(E_1)$ goes to zero as $n \to \infty$. From the construction of $\hat{Y}^n$ we know that $I(I, \hat{Y})$ is within $o(1)$ of

$$\frac{1}{2}\log\left(\frac{1}{\sin^2(\phi)}\right) = \frac{1}{2}\log\left(\frac{P_{\hat{Y}}}{G}\right) \tag{29}$$

(see Figure 3). Thus, from (28) and (29) we have that $R_1 = I(I, \hat{Y}) + o(1)$, hence $\Pr(E_1) \to 0$ as required. Observe that (similarly to the converse part of the theorem) the range of $R_1$ is

$$\frac{1}{2}\log\left(\frac{P_I}{D}\right) < R_1 < R_Q \tag{30}$$

and hence the range of $\gamma$ is similar to (22).

In order to prove that $\Pr(E_2) \to 0$, we need to prove the following lemma.

**Lemma 1** *With probability approaching unity, the triplet* $(I^n, \hat{Y}_1^n, Z^n)$ *belongs to* $T_{I,\hat{Y},Z}$.

**Proof:** We showed above that $\Pr\{(I^n, \hat{Y}_1^n) \in T_{I,\hat{Y}}\} \to 1$. Since $Z^n = \hat{Y}_1^n + V^n$ and $V^n$ is independent of $I^n, \hat{Y}_1^n$, it is straightforward to show that the empirical correlations obtained from $(I^n, \hat{Y}_1^n, Z^n)$ are within $\epsilon$ (or a factor thereof) of the corresponding entries of $K_{I,\hat{Y},Z}$ with probability that goes to 1. Since $p_{I,\hat{Y},Z}$ is Gaussian with covariance matrix $K_{I,\hat{Y},Z}$, typicality is thus immediatelly established with probability approaching unity. ■

Hence, $\Pr(E_2) \to 0$ as $n \to \infty$.

The probability of error $\Pr(E_3)$, is upper-bounded as follows:

$$\begin{aligned}
&\Pr\{\exists w \neq 1 : (I^n, \hat{Y}_w^n, Z^n) \in T_{I,\hat{Y},Z}\} \\
&\leq \sum_{w=2}^{2^{nR_W}} \Pr\{(I^n, \hat{Y}_w^n, Z^n) \in T_{I,\hat{Y},Z}\} \\
&\leq 2^{nR_W} \Pr\{(I^n, \hat{Y}_2^n, Z^n) \in T_{I,\hat{Y},Z}\}]
\end{aligned} \tag{31}$$

where the last inequality is due to the symmetry of the construction of the $\hat{Y}^n$ sequences. It can be easily shown that the quantity $\Pr\{(I^n, \hat{Y}_2^n, Z^n) \in T_{I,\hat{Y},Z}\}$ is upper-bounded by $2^{-n(I(Z;\hat{Y}_2|I)-\epsilon)}$, since

- $(I^n, \hat{Y}_2^n) \in T_{I,\hat{Y}}$; and

- by construction, given $I^n$, $Z^n = \hat{Y}_1^n + V^n$ is independent of $\hat{Y}_2^n$.

It can be easily shown that $I(Z;\hat{Y}_2|I) = \frac{1}{2}\log\left(1+\frac{G}{D_A}\right)$. So, in order for (31) to vanish asymptotically, it suffices

$$R_W \leq \frac{1}{2}\log\left(1+\frac{G}{D_A}\right) - \epsilon \tag{32}$$

The quantity $G$ has the same operational meaning as in the converse; based on Figure 3 it is defined as $G = \min_{\lambda\in\mathbf{R}} \frac{1}{n}\sum_{i=1}^n E(\hat{Y}_i - \lambda I_i)^2$. Together with (28) and assuming that (4) is satisfied with equality, we get the quadratic equation

$$G + \frac{1}{4P_I}(P_I - D + \gamma G)^2 - \gamma G = 0$$

which has the root (the maximum of the two)

$$G = \frac{\gamma(P_I + D) - 2P_I + 2\sqrt{P_I(\gamma D - P_I)(\gamma - 1)}}{\gamma^2} = P_W(\gamma) \tag{33}$$

We then substitute (33) into (32) to get:

$$R_W \leq \tfrac{1}{2}\log\left(1+\tfrac{P_W(\gamma)}{D_A}\right) - o(1) \tag{34}$$

where $\gamma$ satisfies equation (26). It can be shown relatively easily that when

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

then there is always a $\gamma$ (that equals the solution to a 3rd degree polynomial equation) that satisfies both (26) and (34) (case (b) in Figure 4). In this case, (34) coincides with (24) (the "curved" segment $(B,C)$ of Figure 2 is achieved). In order to achieve the rest of $\mathcal{R}_{D,D_A}$, we observe the following:

- When $R_Q = R_Q^* \triangleq \frac{1}{2}\log\left(\frac{P_I}{D}+\frac{P_I-D}{D_A}\right)$ then (34) shows that $R_W = R_W^* \triangleq \frac{1}{2}\log\left(1+\frac{D}{D_A}+\frac{D^2}{P_I D_A}\right)$ is achievable. This particular $(R_Q^*, R_W^*)$ pair (point $B$ of figure 2) lies on the $R_W = R_Q - \frac{1}{2}\log\left(\frac{P_I}{D}\right)$ line.

- When $R_Q = R_Q^{**} \triangleq \frac{1}{2}\log\left(1+\frac{P_I+D}{D_A}+\frac{P_I}{D}\right)$ then $R_W = R_W^{**} \triangleq \frac{1}{2}\log\left(1+\frac{D}{D_A}\right)$ is achievable. The $(R_Q^{**}, R_W^{**})$ pair (point $C$ of figure 2) lies on the $R_W = \frac{1}{2}\log\left(1+\frac{D}{D_A}\right)$ line.

- The whole line $R_Q \geq \frac{1}{2}\log\left(\frac{P_I}{D}\right)$, $R_W = 0$ is trivially achievable (by just compressing $I^n$ up to average distortion $D$). Call this line $\mathcal{L}_0$.

It is straightforward to see now that by timesharing an $(R_Q^*, R_W^*)$-rate code with a $(R_Q^{**}, R_W^{**})$-rate code and a code with rates that lie on the $\mathcal{L}_0$ line, we can achieve the whole rate region $\mathcal{R}_{D,D_A}$ (it can be easily verified that the codes obtained from the timesharing satisfy the distortion constraint (1)).

We proved that if $(R_Q, R_W) \in \mathcal{R}_{D,D_A}$ then the average probability of error, over the ensemble of the random codes, vanishes asymptotically with $n$. Hence, we argue that there exist deterministic codes that achieve $\mathcal{R}_{D,D_A}$ with arbitrarily small probability of error (averaged over all the messages). Finally, we conclude the proof by making the maximal probability of error arbitrarily small, through an appropriate expurgation of the codebook. ■

# 4 Performance of Other Schemes

In this section we will 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, for the following systems:

- Regular Quantization Index Modulation (QIM) [12], where no knowledge of the original image is available at the decoder (public version).
- Additive watermarking, where the embedder computes the weighted sum of the original image and a watermark-dependent signal and then compresses the result 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 $\mathcal{R}_{D,D_A}$ 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 [12] (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. The watermark $W$ is embedded into an original image $I^n$ by quantizing $I^n$

using the $W^{\text{th}}$ 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 $W$.

As discussed in [12], achievable pairs $(R_Q, R_W)$ for regular QIM (also called "hidden" QIM) under constraints (1) and (2) can be found using a well-known formula due to Gel'fand and Pinsker [17]:

$$\begin{aligned} R_Q &= I(\hat{Y}; Z) = I(\hat{Y}; \hat{Y} + V) & (35) \\ R_W &= I(\hat{Y}; Z) - I(\hat{Y}; I) & (36) \end{aligned}$$

where $I$ and $V$ are independent Gaussian variables distributed as before, and $\hat{Y}$ is such that $E(\hat{Y} - I)^2 \leq D$ (also note that $Z = \hat{Y} + V$).

We have investigated the behavior of (36) as $R_Q$ varies, expressing $R_W$ in terms of $R_Q$ and the system parameters $P_I$, $D$ and $D_A$. In the random coding argument for (35) and (36), all codewords are i.i.d. Gaussian with variance $P_{\hat{Y}}$. Hence (35) becomes

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

and therefore

$$P_{\hat{Y}} = D_A(2^{2R_Q} - 1) \qquad (37)$$

Also, (36) gives

$$R_W = R_Q - \frac{1}{2} \log \left( \frac{P_I P_{\hat{Y}}}{P_I P_{\hat{Y}} - (E(I\hat{Y}))^2} \right) \qquad (38)$$

From (1) we get $P_I + P_{\hat{Y}} - 2E(I\hat{Y}) \leq D$, so, since $D \leq P_I$, (38) is maximized when $E(I\hat{Y}) = (P_I + P_{\hat{Y}} - D)/2$ for all $P_{\hat{Y}}$. Hence, (38) becomes

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

and by substituting (37) we obtain

$$R_W = R_Q - \frac{1}{2} \log \left( \frac{P_I D_A (2^{2R_Q} - 1)}{P_I D_A (2^{2R_Q} - 1) - \frac{1}{4}(P_I + D_A(2^{2R_Q} - 1) - D)^2} \right) \qquad (39)$$

where $R_Q$ is assumed to lie in a subinterval of

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

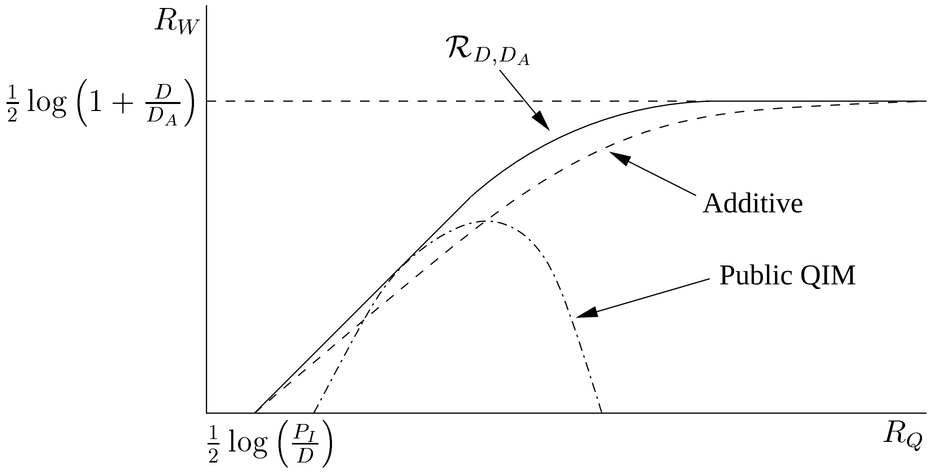


Figure 5: Upper boundaries of the achievable rate regions for the public QIM and the private additive schemes. $\mathcal{R}_{D,D_A}$ gives an outer bound to the achievable rate regions of these schemes.

which ensures that the argument of $\log(\cdot)$ in (39) is no less than unity and that the resulting value of $R_W$ is nonnegative (the exact expression for the range of $R_Q$ can be obtained by solving a 3rd degree polynomial). Figure 5 shows the achievable region described by (39).

As can be seen from Figure 5, the watermarking rate $R_W$ obtained using i.i.d. Gaussian codebooks is positive only for a finite range of values of $R_Q$. 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). Moreover, when QIM is applied to fingerprinting, there is an additional drawback: each user, who receives a different fingerprinted version of an image, will need to be provided a different source codebook in order to do the decoding. This entails higher cost and complexity than using a universal quantizer (or quantization algorithm) which is easily accessible by all users.

The two above problems can be circumvented with the use of additive schemes (e.g. see [9, 14]). The analysis of such an additive scheme follows.

*B. Additive Watermarking, Private Scenario*

In general, additive watermarking entails the computation of

$$Y^n = \alpha I^n + \beta X^n(W)$$

where $\alpha, \beta$ are non-zero scalars, $W$ is the index of the watermark and $X^n(W)$ is a $n$-dimensional signal that does not depend on the original image $I^n$. Since we require the output of the encoder to be a compressed image $\hat{Y}^n$, we use a universal quantizer $f$ (that does not depend on the watermark embedded) to produce

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

such that the distortion constraint (1) is satisfied. The decoder, given $Z^n, I^n$ tries to detect $W$ with vanishing probability of error.

We will use a random coding argument. The watermarker generates a channel codebook which consists of $2^{nR_W}$ signals $X^n(1), \ldots, X^n(2^{nR_W})$, each one i.i.d. Gaussian distributed with variance $P_X$. Since the distortion constraint is between $I^n$ and $\hat{Y}^n$ (and not between $Y^n$ and $\hat{Y}^n$) we consider a quantizer that scales $Y^n$ by $1/\alpha$ before quantizing. Furthermore, we assume that $P_X$ is a free parameter in our model, hence, we can equivalently set

$$\hat{Y}^n = f(I^n + X^n(W))$$

The source codebook consists of $2^{nR_Q}$ sequences $\tilde{Y}^n(1), \ldots, \tilde{Y}^n(2^{nR_Q})$, whose components are i.i.d. $\mathcal{N}(0, P_{\hat{Y}})$. For the embedding, $\hat{Y} = \tilde{Y}^n(q)$, where $q$ is the smallest index such that the pair $(Y^n, \tilde{Y}^n(q))$ is typical with respect to a joint Gaussian distribution $p_{Y,\hat{Y}}$. If no such $q$ can be found, then the encoder declares an error. The distribution $p_{Y,\hat{Y}}$ has zero mean and covariance matrix

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

where $d = n^{-1}E||I^n - \hat{Y}^n||^2$. By setting

$$R_Q = I(Y;\hat{Y}) + \epsilon = \frac{1}{2}\log\left(\frac{(P_I + P_X)P_{\hat{Y}}}{|K_{Y,\hat{Y}}|}\right) + \epsilon \tag{40}$$

it can be shown that the distortion constraint (1) is satisfied. For the detection of the watermark, given $Z^n = \hat{Y}^n + V^n$ and $I^n$, the detector tries to find a $w$ such that $(I^n, X^n(w), Z^n)$ are typical with respect to a joint i.i.d. distribution $p_{I,X,Z}$. This distribution is Gaussian with covariance matrix

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

It can be proved that if $R_W = I(X; I, Z) - \epsilon$ then the probability of decoding error goes to zero as $n \to \infty$. Solving (40) for $P_X$ and substituting into $I(X; I, Z)$, we get an achievable rate:

$$R_W = \frac{1}{2}\log\left(\frac{2^{2R_Q}(2d(P_I + P_{\hat{Y}}) - d^2 - (P_I - P_{\hat{Y}})^2 + 4D_A P_I}{4P_I(2^{2R_Q}D_A + P_{\hat{Y}})}\right) - o(1) \tag{41}$$

Then, (41) is maximized for $d = D$ and $P_{\hat{Y}} = -2^{2R_Q}D_A + \sqrt{(2^{2R_Q}D_A + D)^2 + P_I(P_I + 2D_A(2^{2R_Q} - 2) - 2D)}$, and becomes:

$$R_W = \tfrac{1}{2}\log\left(\frac{2^{2R_Q}\left((4P_I(D+D_A) - \left(D+P_I+2^{2R_Q}D_A - \sqrt{(2^{2R_Q}D_A+D)^2+P_I(P_I+2D_A(2^{2R_Q}-2)-2D)}\right)^2\right)}{4P_I\sqrt{(2^{2R_Q}D_A+D)^2+P_I(P_I+2D_A(2^{2R_Q}-2)-2D)}}\right) \tag{42}$$

where $R_Q \geq \frac{1}{2}\log\left(\frac{P_I}{D}\right)$. The region of achievable rate pairs $(R_Q, R_W)$ can be seen in Figure 5. As expected, when $R_Q \to \infty$, $\hat{Y}^n$ becomes negligibly different from $Y^n = I^n + X^n$ and therefore $R_W$ approaches the capacity of an AWGN channel.

# 5 Conclusions

In this paper, we considered a system that watermarks $n$-dimensional i.i.d. Gaussian images and distributes them in compressed form, such that an average distortion constraint is met. We assumed that the watermarked images are further corrupted by Gaussian attacks. By means of a coding theorem, we established the region of achievable watermarking and quantization rates such that the error probability in decoding the embedded message in a watermarked/quantized image approaches zero asymptotically in $n$. We also presented achievability results for the public version of the regular Quantization Index Modulation scheme, as well as for additive watermarking/quantization schemes.

There are a number of possible extensions to the problem considered in this paper. For example, it would be interesting to establish the rate region in the case where both a watermark (identifying the agent) and a fingerprint (identifying the user) are embedded sequentially into an image by independent encoders. We suspect that, in the presence of attacks, the resultant rate region could be different than the rate region obtained from joint embedding, as is the case in a multiple-access channel. Moreover, we are investigating more general attack scenarios (e.g., combining our model with the one in [13]).

# References

[1] 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.

[2] F. Petitcolas, R. Anderson, and M. Kuhn. Information hiding - a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999.

[3] 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.

[4] S. Katzenbeisser and F. Petitcolas. *Information Hiding Techniques for Steganography and Digital Watermarking.* Artech House, 2000.

[5] I.Cox, J. Bloom, and M. Miller. *Digital Watermarking.* Morgan Kaufmann Publishers, 2001.

[6] 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.

[7] P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding, preprint, available at http://www.ifp.uiuc.edu/~moulin/paper.html. January 2001.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] A. Cohen and A. Lapidoth. The Gaussian watermarking game - parts I, II. *To appear in the IEEE Transactions on Information Theory*, January 2001.

[14] 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.

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

[16] 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.

[17] S. Gel'fand and M. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.