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

Damianos Karakos
Department of Electrical and Computer Engineering
University of Maryland
College Park, MD 20742
e-mail: karakos@eng.umd.edu

Adrian Papamarcou
Department of Electrical and Computer Engineering
University of Maryland
College Park, MD 20742
e-mail: adrian@eng.umd.edu

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.

I. 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 *covert* [6]).

The quality of the watermarked data must thus be comparable to that of the *covert* signal, a requirement which is often expressed in terms of a distortion constraint.

- **Robustness:** The message must be detectable in the watermarked image (the *covert* 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, 13, 6, 14, 15]. 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 either has access to the source codebook or obtains the reproduced image from a local, high-speed link. 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, 15], 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, 14]). 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

\[
R_Q \geq \frac{1}{2} \log \left( \frac{P_1}{D} \right) \\
R_W \leq R_Q - \frac{1}{2} \log \left( \frac{P_2}{D} \right)
\]

where \( R_Q \) is the quantization rate, \( R_W \) is the watermarking rate, \( P_1 \) 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

\[
R_Q - R_W > \frac{1}{2} \log \left( \frac{P_2}{D} \right)
\]

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 [13] 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 IV.

In summary, this paper contains final versions of results in [9, 15], 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 QM 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 II. Section III contains the proof of the coding theorem. Achievability results for other schemes that combine watermarking and compression are presented in Section IV. Finally, conclusions and directions for further research are given in Section V.

II. THE RATE REGION

The watermarking/authentication 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 $Y^n$ is the watermarked/quantized image which can be found in a 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 $Y^n$. The watermark decoder outputs $\hat{W}$, its estimate of $W$. The transparency and robustness requirements are expressed via the following constraints:

$$n^{-1}E||I^n - \hat{I}^n||^2 \leq D,$$  
$$\Pr\{\hat{W} \neq W\} \to 0, \text{ as } n \to \infty$$

By means of a coding theorem, we have established the region $\mathcal{R}_{D, D_A}$ of achievable rates $(R_Q, R_W)$:

$$\mathcal{R}_{D, D_A} = \left\{ (R_Q, R_W) : \right.$$ 

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

$$R_W \leq \left\{ \begin{array}{ll} \max & \left\{ \frac{1}{2} \log \gamma + P_W(\gamma) \right\} \\
\gamma \in \left[ \frac{P_I}{2D}, 1 \right] \\
\min & \left\{ R_Q - \frac{1}{2} \log \gamma \right\} \\
\end{array} \right.$$

where

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

The proof of the coding theorem (both the direct and converse parts) can be found in Section III. $\mathcal{R}_{D, D_A}$ is the shaded region in Figure 2. Its upper boundary is composed of:

- 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

$$R_W = \frac{1}{2} \log \left( \frac{P_I}{D} \right),$$

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

for $R_Q$ in the interval $\left[ \frac{1}{2} \log \left( \frac{P_I}{D} \right), \frac{1}{2} \log \left( 1 + \frac{P_I + D}{D_A} \right) \right]$, i.e., the projection of $BC$ on the $R_Q$-axis.

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

Two key conclusions can be drawn from Figure 2:

- For quantization rates $R_Q \in \left[ \frac{1}{2} \log \left( \frac{P_I}{D} \right), \frac{1}{2} \log \left( 1 + \frac{P_I + D}{D_A} \right) \right]$, 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 $R_Q \geq \frac{1}{2} \log \left( 1 + \frac{P_I + D}{D_A} \right)$, the maximum watermarking rate is constant and equal to $\frac{1}{2} \log \left( 1 + \frac{P_I + D}{D_A} \right)$. 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.

III. 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.

Converse Theorem

The converse theorem states that any rate pair $(R_Q, R_W)$ that satisfies constraints (1) and (2) must lie in $\mathcal{R}_{D, D_A}$ defined in the previous section.

Proof: Let $\epsilon > 0$. We assume that the watermark index $W$ is uniformly distributed in $\{1, \ldots, 2^{nR_W} \}$, that $\Pr\{W \neq \}$
\[ \frac{1}{n} I(I^n; \hat{Y}^n) \geq \frac{1}{2} \log \left( \frac{P_I}{P_{\hat{Y}}} \right) \]

This establishes the lower bound on \( R_Q \) in the definition of \( R_{D,A} \).

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

\[
\langle U^n, V^n \rangle \triangleq \frac{1}{n} \sum_{i=1}^{n} E[U_i V_i].
\]

The geometry of this space is shown in Figure 3, where the circle \( C \) has radius \( \sqrt{D} \) corresponding to the distortion constraint (4). The lengths of \( I^n \) and \( \hat{Y}^n \) are given by \( \sqrt{P_I} \) and \( \sqrt{P_{\hat{Y}}} \), respectively, where \( P_I \triangleq n^{-1} \sum_{i=1}^{n} E[I_i^2] \); while the angle between the two vectors is denoted by \( \phi \). The maximum \( \phi_{\text{max}} \) of \( \phi \) is obtained when \( \hat{Y}^n \) is tangent to \( C \), in which case

\[ \sin^2(\phi_{\text{max}}) = \frac{D}{P_I} \]

Note that for any value of \( \phi \) less than \( \phi_{\text{max}} \), there are two possible positions of \( \hat{Y}^n \) on the circle \( C \) (the position farther from the origin is shown in Figure 3).

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 \( I^n \):

\[ \lambda_0 \triangleq \arg \min_{\lambda \epsilon \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} E[(\hat{Y}_i - \lambda I_i)^2] \]

Let \( P_{\gamma|\mu} \) denote the resulting MMSE error, and note that

\[ \sin^2(\phi) = \frac{P_{\gamma|\mu}}{P_{\gamma}} \]

The geometry of Figure 3, it easily follows that for \( \gamma \triangleq \sin^{-2}(\phi) \),

\[ P_{\gamma|\mu}(\gamma - 1) = \frac{(P_I + P_{\hat{Y}} - D)^2}{4P_I} \]

\[ P_\gamma = \gamma P_{\gamma|\mu}(\gamma - 1) \]

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

\[ P_{\gamma|\mu} = \frac{\gamma(P_I + D) - 2P_I \pm 2\sqrt{P_I(\gamma D - P_I)(\gamma - 1)}}{\gamma^2} \]

Consistent with our earlier observation, there are two possible values of \( P_{\gamma|\mu} \) for every \( \phi < \phi_{\text{max}} \), or equivalently, for every \( \gamma > P_I/D \). The larger value is precisely \( R_W(\gamma) \), as defined in (3).

The mutual information between \( I^n \) and \( \hat{Y}^n \) is also related to the geometry of Figure 3. Specifically, as we prove in [16],

\[ \frac{1}{n} I(I^n; \hat{Y}^n) \geq \frac{1}{2} \log \left( \frac{P_I}{P_{\hat{Y}}} \right) = \frac{1}{2} \log(\gamma) \]

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

\[ R_W = n^{-1} H(W|I^n, V^n) \]

(10)

\[ = n^{-1} I(W; \hat{Y}^n) + n^{-1} H(W|I^n) \]

(11)

\[ \leq n^{-1} I(W; \hat{Y}^n) + n^{-1} H(W|I^n, Z^n) + \epsilon \]

(12)

\[ = n^{-1} H(\hat{Y}^n|I^n) + \epsilon \]

(13)

\[ \leq n^{-1} H(\hat{Y}^n) + \epsilon \]

(14)

where (10) holds because \( I^n \) and \( V^n \) are independent of \( W \); (11) follows from H(W|I^n, Z^n) \( \leq \) H(W|I^n, Z^n); (12) is a consequence of Fano’s inequality; (13) holds because \( H(\hat{Y}^n|W, I^n) = 0 \) (since \( \hat{Y}^n \) is a function of \( W \) and \( I^n \)) and (14) follows from \( R_{\gamma} \geq n^{-1} H(\hat{Y}^n) \).

The second chain of inequalities is as follows (where \( \lambda_0 \) was defined in (6)):

\[ R_W = n^{-1} H(W|I^n) \]

(15)

\[ = n^{-1} I(W; Z^n) + n^{-1} H(W|I^n, Z^n) \]

(16)

\[ \leq n^{-1} I(W; Z^n) + \epsilon \]

(17)

\[ = n^{-1} h(Z^n) - n^{-1} h(Z^n|I^n, W) + \epsilon \]

(18)

\[ \leq n^{-1} h(Z^n) - \lambda_0 R + V^n + \epsilon \]

(19)

where (15) holds because \( I^n \) is independent of \( W \); (16) follows from Fano’s inequality; (17) holds because \( \lambda_0 I^n \) is a function of \( I^n \) and \( W \); (18) follows from the independence of \( V^n \) and \( (I^n, W) \); and (19) holds by the usual Gaussian bound on differential entropy.

For any value of \( \gamma \) (corresponding to the geometry of \( I^n \) and \( \hat{Y}^n \)), we obtain from (14) and (9)

\[ R_W \leq R_Q - \frac{1}{2} \log(\gamma) + \epsilon \]
and from (19) and (8) 
\[ R_W \leq \frac{1}{2} \log \left( 1 + \frac{P_w(\gamma)}{D_A} \right) + \epsilon \]

An upper bound on the range of \( \gamma \) can be deduced from (5), (9), resulting in \[ \frac{P_I}{D} \leq \gamma \leq 2^{2R_Q} \]

Thus 
\[ R_W \leq \max_{\gamma \in \left( \frac{P_I}{D}, 2^{2R_Q} \right)} \left\{ \min \left\{ R_Q - \frac{1}{2} \log(\gamma), \frac{1}{2} \log \left( 1 + \frac{P_w(\gamma)}{D_A} \right) \right\} + \epsilon \right\} \]  \( (20) \)

and taking \( \epsilon \to 0 \), we obtain the upper bound on \( R_W \) in the definition of \( R_{D_{D_A}} \).

**Note:** In the special case \( D_A = 0 \) (no attack), the bound on \( R_W \) is simply \( R_W \leq R_Q - \frac{1}{2} \log(P_I/D) \). The converse theorem then reduces to the channel coding part of the converse theorem in [18], and also the converse theorem of [15] for \( R_F = 0 \).

**Direct Theorem**

We now show that \( R_{D_{D_A}} \) is achievable.

**Proof:** As required for \( R_{D_{D_A}} \), we limit the quantization rate to \( R_Q \geq \frac{1}{2} \log \left( \frac{P_I}{D} \right) \). We use a random coding argument, where the watermark index \( W \) is assumed uniformly distributed in \( \{1, \ldots, 2^{2R_W}\} \). The technique is similar to the private version of regular QIM [13], in that \( 2^{2R_W} \) quantizers, each one indexed by a different watermark, are employed.

**Codebook Generation:** Let \( \gamma \in \left( \frac{P_I}{D}, 2^{2R_Q} \right) \). A set of \( 2^{2R_Q} \) i.i.d. \( \sim N(0, \gamma P_w(\gamma)) \) Gaussian sequences \( \tilde{Y}^n \), is generated and partitioned into \( 2^{2R_W} \) subsets of \( 2^{R_W} \) sequences each, i.e., 
\[ R_Q = R_W + R_1 \]  \( (21) \)

The \( w \)th subset, consisting of sequences \( \tilde{Y}^n(w, 1), \ldots, \tilde{Y}^n(w, 2^{R_W}) \), 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 \( T_{I,\tilde{Y}}(\epsilon) \) of typical pairs with respect to a bivariate Gaussian distribution \( P_{I,\tilde{Y}} \) having mean zero and covariance 
\[ K_{I,\tilde{Y}} = \begin{bmatrix} \frac{P_I}{\sqrt{(\gamma - 1) P_I P_w(\gamma)}} & \frac{\sqrt{(\gamma - 1) P_I P_w(\gamma)}}{\gamma P_w(\gamma)} \\ \frac{\sqrt{(\gamma - 1) P_I P_w(\gamma)}}{\gamma P_w(\gamma)} & \frac{\sqrt{(\gamma - 1) P_I P_w(\gamma)}}{\gamma P_w(\gamma)} \end{bmatrix} \]

The output of the encoder (decoder) is denoted by \( \tilde{Y}^n(w) = \tilde{Y}^n(w, q) \). If none of the codewords in the \( w \)th codebook is jointly typical with \( I^n \), then the embedder outputs \( \tilde{Y}^n(w) = 0 \).

In this manner, \( 2^{2R_W} \) watermarked versions of the image \( I^n \) are obtained: \( \tilde{Y}^n(1), \ldots, \tilde{Y}^n(2^{2R_W}) \). Clearly, for random \( W \), the embedder output is \( \tilde{Y}^n(W) \).

Note that the second moments in \( K_{I,\tilde{Y}} \) are consistent with the geometry of Figure 3, with \( \gamma = \sin^{-2}(\phi) \). In particular, if the pair \( (I^n, \tilde{Y}^n) \) lies in \( T_{I,\tilde{Y}}(\epsilon) \), then the empirical second moments: 
\[ \frac{1}{n} \sum_{i=1}^{n} I_i^2, \quad \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i^2 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} I_i \tilde{Y}_i \]

are within \( \epsilon \) (or a factor thereof) of the average values shown implicitly in Figure 3. This also means that the distortion constraint (1) is essentially met (since \( \epsilon \)-differences can be safely ignored).

**Decoding:** Again, the decoder has access to the original image \( I^n \). Upon receiving \( Z^n = Y^n(W) + V^n \), the decoder seeks among all watermarked versions \( Y^n(1), \ldots, Y^n(2^{2R_W}) \) of \( I^n \) a single \( \tilde{Y}^n(w) \) such that the triplet \( (I^n, \tilde{Y}^n(w), Z^n) \) lies in \( T_{I,\tilde{Y},Z}(\epsilon) \), the set of typical triplets with respect to the trivariate Gaussian distribution \( P_{I,\tilde{Y},Z} \) having zero mean and covariance matrix \( K_{I,\tilde{Y},Z} \) with entries again consistent with the geometry of Figure 3. If a unique such sequence \( \tilde{Y}^n(w) \) exists, then the decoder outputs \( W = \hat{w} \); otherwise, the decoder declares an error.

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

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

- **Error 1:** \( \hat{Y}^n(1) = 0 \), i.e., there exists no \( q \in \{1, \ldots, 2^{2R_W}\} \) such that \( (I^n, \hat{Y}^n(1), q) \in T_{I,\hat{Y}} \).

- **Error 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 \( (I^n, \hat{Y}^n(1), Z^n) \notin T_{I,\hat{Y},Z} \).

- **Error 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 
\[ \Pr\{W \neq 1\} = \Pr(E_1) + \Pr(E_2) + \Pr(E_3) \]

**Behavior of Pr(E1):** From standard rate-distortion theorems [17], we know that if \( R_1 > I(I; \hat{Y}) \) (the mutual information of the bivariate \( P_{I,\tilde{Y}} \) defined above), then \( \Pr(E_1) \to 0 \) as \( n \to \infty \).

Since \( R_1 = R_Q - R_W \) from (21), and \( I(I; \hat{Y}) = \frac{1}{2} \log(\gamma) \), it follows that \( \Pr(E_1) \to 0 \) provided that 
\[ R_W < R_Q - \frac{1}{2} \log(\gamma) \]  \( (22) \)

**Behavior of Pr(E2):** 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 \( Z^n = \tilde{Y}^n(1) + V^n \) and \( V^n \) is independent of \( (I^n, \hat{Y}^n(1)) \), it follows easily that the empirical correlations obtained from \( (I^n, \hat{Y}^n(1), Z^n) \) are within \( \epsilon \) (or a factor thereof) of the corresponding entries of \( K_{I,\tilde{Y},Z} \) with probability approaching unity asymptotically. Typicality with respect to \( P_{I,\tilde{Y},Z} \) thus holds (also with probability approaching unity).

**Behavior of Pr(E3):**
\[ \Pr(E_3) = \Pr\{\exists w \neq 1 : (I^n, \hat{Y}^n(w), Z^n) \in T_{I,\hat{Y},Z}\} \leq \sum_{w=2}^{2^{2R_W}} \Pr\{(I^n, \hat{Y}^n(w), Z^n) \in T_{I,\hat{Y},Z}\} \]
\[ = (2^{2R_W} - 1) \Pr\{(I^n, \hat{Y}^n(2), Z^n) \in T_{I,\hat{Y},Z}\} \]

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, \( Z^n = \hat{Y}^n(1) + V^n \) is independent of \( \hat{Y}^n(2) \) given \( \Gamma^n \), a standard argument (cf. the proof of Theorem 8.6.1 in [17]) yields

\[
\Pr\{ (I^n, \hat{Y}^n(2), Z^n) \in T_I, \hat{Y}, x \} \leq 2^{-n(I(\hat{Y}; Z) - (\gamma D)/2)}
\]

where the conditional mutual information is computed with respect to the trivariate \( p_{I, \hat{Y}, Z} \) defined earlier. In our case, \( I(Z; \hat{Y}|I) = \frac{1}{2} \log \left( 1 + \frac{P_W(\gamma)}{D_A} \right) \) and therefore \( \Pr(E_2) \to 0 \) provided

\[
R_W < \frac{1}{2} \log \left( 1 + \frac{P_W(\gamma)}{D_A} \right)
\]  

(24)

From (22) and (24) it follows that \( R_W \) is achievable provided

\[
R_W < \min \left\{ R_Q - \frac{1}{2} \log (\gamma), \frac{1}{2} \log \left( 1 + \frac{P_W(\gamma)}{D_A} \right) \right\}
\]  

(25)

Choosing \( \gamma \in \left[ \frac{P_Q}{D_A}, 2^{R_Q} \right] \) so as to maximize the right-hand side of (25), we can achieve the whole region \( R_{D,D_A} \).

We have thus proved that if \((R_Q, R_W) \in R_{D,D_A}\) 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 \( R_{D,D_A} \) 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.

\[ \square \]

IV. PERFORMANCE OF OTHER SCHEMES

In this section we present achievable 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) [13], 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 achieveability results, the rate region \( R_{D,D_A} \) 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 [13] (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 \( W^{th} \) 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 \( W \).

\[ Y^n = \alpha I^n + \beta X^n(\hat{W}) \]

As discussed in [13], 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 Gelfand and Pinsker [19]:

\[
R_Q = I(\hat{Y}; Z) = I(\hat{Y}; \hat{Y} + V)
\]

(26)

\[
R_W = \frac{1}{2} \log \left( 1 - \frac{P_Q}{D_A} \right)
\]

(27)

where \( [1]^+ = \max\{1, 0\} \). 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 (27) 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 (26) and (27), all codewords are i.i.d. Gaussian with variance \( P_I \). Optimization with respect to \( P_I \) yields

\[
R_W = \left[ R_Q - \frac{1}{2} \log \left( \frac{P_{I,D_A}(2^{R_Q}-1)}{P_{I,D_A}(2^{R_Q}-1)-\frac{P_{I,D_A}(2^{R_Q}-1)-D}{D_A}} \right) \right]^+
\]

(28)

Here, \( R_Q \) is assumed to lie in a subinterval of

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

which ensures that the argument of \( \log(\cdot) \) in (28) 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). Expression (28) is shown in Figure 4 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 4, 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 not arise when additive watermarking schemes (see, e.g., [9, 15]) are used. The analysis of such an additive scheme follows.

B. Additive Watermarking, Private Scenario

In general, additive watermarking entails the computation of

\[ Y^n = \alpha I^n + \beta X^n(\hat{W}) \]
\[ R_W = \frac{1}{2} \log \left( \frac{2^{2R_Q} \left( (4P_1(D+D_A)) - (D+P_1 + 2^{2R_Q} D_A - \sqrt{(2^{2R_Q} D_A + D)^2 + P_1(1 + 2D_A(2^{2R_Q} - 2) - 2D)} \right)^2}{4P_1 \sqrt{(2^{2R_Q} D_A + D)^2 + P_1(1 + 2D_A(2^{2R_Q} - 2) - 2D)}} \right) \] (29)

where \( W \) is the index of the watermark and \( X^n(W) \) is a \( n \)-dimensional signal that does not depend on the original image \( I^n \). \( \alpha, \beta \) 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:

\[ Y^n = f(Y^n) \]

subject to an appropriate distortion constraint (1) in this case). The decoder attempts to detect \( W \) given \( Y^n \) and \( I^n \) with vanishing probability of error.

The details of the achievability argument can be found in [16]. The watermarker generates a 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 source codebook \( \{Y^n(1), \ldots, Y^n(2^{nR_q})\} \), also i.i.d. Gaussian with variance \( P_Y \), where both \( P_X \) and \( P_Y \) are free parameters in the model.

Source encoding is done by means of a joint typicality criterion using the distortion constraint (1). The watermark detector also uses a joint typicality criterion. Optimizing with respect to \( P_X \) and \( P_Y \) yields expression (29) for the achievable watermarking rate, where \( R_Q \geq \frac{1}{2} \log \left( \frac{D}{2} \right) \). The upper boundary of the region of achievable rate pairs \( (R_Q, R_W) \) is shown in Figure 4. As expected, when \( R_Q \to \infty \), \( Y^n \) is negligibly different from \( Y^n = I^n + X^n \) and thus \( R_W \) approaches the capacity of an AWGN channel.

V. CONCLUDING REMARKS

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.

We are currently investigating a number of possible extensions to the problem considered in this paper. For example, one extension is the case of Gaussian attacks combined with scaling; i.e. \( Z^n = \beta X^n + V^n \), where \( \beta \) is a positive scalar. Also, we consider more general frameworks (arbitrary distortion metrics and image distributions, non-Gaussian attacks, etc) and will report our results in a forthcoming publication.

ACKNOWLEDGMENTS

The authors would like to thank Aaron Cohen and Neri Merhav for their comments on an earlier version of this paper.

REFERENCES