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

Damianos Karakos, *Member, IEEE,* and Adrian Papamarcou, *Member, IEEE*

***Abstract*—A system which embeds watermarks in n-dimensional Gaussian data and distributes them in compressed form is studied. The watermarked/compressed data have to satisfy a distortion constraint, and the watermark has to be recoverable in a private scenario (in which the original data are available at the watermark detector). The performance of the system in the presence of additive Gaussian attacks is considered, and the region of achievable quantization and watermarking rate pairs (R_Q, R_W) is established. Moreover, two surprising facts are demonstrated: 1) at low R_Q, the maximum achievable R_W is the same as when there are no attacks; and 2) at high (but finite) R_Q, the maximum achievable R_W is the same as when there is no compression ($R_Q = \infty$). Finally, the performance of related schemes is also discussed.**

***Index Terms*—Capacity, compression, copyright protection, Gaussian noise, Gaussian source, quantization, rate-distortion theory, rate region, watermarking.**

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]–[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.

In general, an information hider (or *watermarker*) embeds a message (known as *watermark*) into an original document (also referred to as the *covertext* [6]). The result is a watermarked document also known as *stegotext* [6]. The stegotext is subject to manipulation by a malicious *attacker*, who produces a *forgery*. The goal of the attacker is to make the watermark undetectable from the forgery. Careful design of the watermarking system can minimize the chance that such an attack will be successful.

Two key issues in the design of watermarking schemes are as follows.

- **Transparency:** The hidden message should not interfere perceptually with the covertext. The quality of the stegotext must thus be comparable to that of the covertext, a requirement which is often expressed in terms of a distortion constraint.
- **Robustness:** Although an attacker could possibly introduce distortion (e.g., through additive noise, quantization, digital-to-analog (D/A) conversion, etc.) into the stegotext and thus create a forgery, the hidden message should still be detectable. In the *private* detection scenario, the covertext 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 [6]–[15]. The model treated in this paper, which involves *joint* watermarking and compression, has received less attention in the literature. A brief summary of our model follows.

Due to bandwidth or storage constraints, a compressed digital version of the watermarked data is desirable. To that end, the watermarker encodes the covertext and the watermark index jointly as a representation vector in a source codebook, which becomes the (compressed) stegotext. The number of possible watermarks is 2^{nR_W}, while the size of the source codebook is 2^{nR_Q}; we call R_W the *watermarking rate* and R_Q the *quantization rate*. The index of the stegotext in the source codebook is then transmitted[1] to the customer, who has either access to the source codebook and can reconstruct the stegotext, or else obtains the reconstructed stegotext through 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 recoverable after an attack on the stegotext. Our analysis assumes that the attack is in the form of additive Gaussian noise, and that the watermark decoder has access to the original covertext (private detection). The main objective of this paper is the determination of the set of all allowable rate pairs (R_Q, R_W).

Our study of this problem is motivated by the following scenario (see Fig. 1). A news source owns a large number of high-resolution images and video sequences which it subsequently distributes to various news outlets such as newspapers, television broadcast stations, and other media organizations. The watermark identifying the source is embedded into each item, which is then converted into digital form for storage and transmission. Since multiple transmissions (to different outlets) of items are likely to take place, compression is of crucial importance. Upon request, an item is delivered electronically to a news

Manuscript received February 1, 2002; revised December 1, 2002. This work was supported in part by the National Science Foundation under Grant EIA-9700866. The material in this paper was presented in part at the 36th Annual Conference on Information Sciences and Systems, Princeton, NJ, March 2002.

The authors are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: karakos@eng.umd.edu; adrian@eng.umd.edu).

Communicated by A. Kavčić, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2003.814474

[1]We assume that the transmission is error free; a more general model could incorporate protection against packet losses.

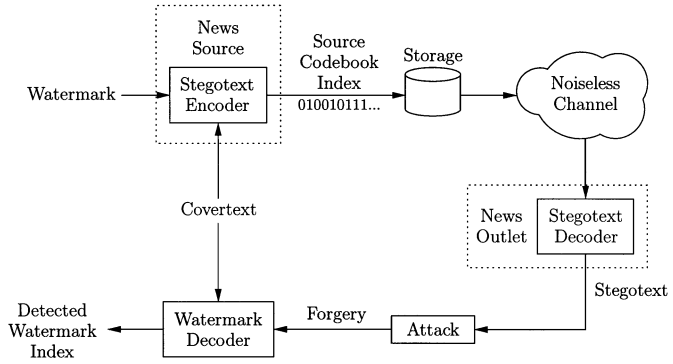

Fig. 1. An application of joint watermarking and compression.

outlet in the form of a source codebook index. The news outlet decodes the index using the source codebook, thereby obtaining the (compressed and watermarked) image or video sequence for eventual publication. The source wishes to ensure that its watermark is tamperproof, i.e., it is resilient to deliberate attacks on the published stegotext.

A key assumption of the above attack model is that the attacker has no access to the source codebook used in generating the stegotext from the covertext. If the attacker uses continuous distributions (e.g., the additive Gaussian noise in our case), then, with probability one, the resulting forgery will be different from every representation vector in the source codebook. Thus, the occurrence of an attack will be easily verifiable. A successful attack is one that renders the embedded watermark undetectable, thereby challenging the ownership of the document.

The main problem treated in this paper combines elements of source and channel coding. Indeed, the covertext and the watermark are jointly compressed using a source codebook; and the watermark, viewed as a message, must be recoverable after an attack, which can be easily modeled as a transmission channel (the additive Gaussian noise channel in our case). Since a single encoder–decoder pair is used for the entire system, it is important to differentiate between this model and the general class of problems treated in *joint source-channel coding* (JSCC) [16].

In JSCC, the key objective is to design an encoder which maps source sequences directly into channel input sequences, and a decoder which maps channel output sequences directly into source reconstructions that satisfy an average distortion constraint. Joint watermarking and compression, on the other hand, has two objectives: 1) reconstruction of the *source* sequence within specified distortion at the *encoder output*; and 2) recovery of an embedded *message* at the *decoder output*. Also note that, at least in the private scenario, the decoder has the covertext (source) available as side information. Information about the source is typically not available at the decoder in most JSCC problem formulations; recent work [17] on JSCC with side information (about the original source) at the decoder may have some implications on watermarking.

Previous work involving joint watermarking and compression [9], [14], focused on the case where the watermarked/compressed data were not subject to attacks (compression inherently introduces degradation, but cannot be construed as a malicious attack of the type studied in, e.g., [6], [7]). It was shown that, when the covertext is independent and identically distributed (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

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

$$R_W \leq R_Q - \frac{1}{2}\log\left(\frac{P_U}{D}\right)$$

where R_Q is the quantization rate, R_W is the watermarking rate, P_U is the covertext variance (per dimension or pixel), and D is the average quadratic distortion between the covertext and the compressed stegotext. 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 covertext, where

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

By standard rate-distortion theory for i.i.d. Gaussian sources, there are enough bits to represent the covertext with average distortion equal to D. Yet this scheme is entirely inadequate from a watermarking (or information hiding) perspective, since the reconstructed data do 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 covertext. The *regular* version of QIM, in which the stegotext 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. Analyses of other compression/watermarking techniques can be found in [18], [19].

In summary, this paper contains final versions of results in [9], [14], together with extensions to the important case where the compressed data 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 covertext and the compressed stegotext does not exceed a threshold D, and the watermark index is detectable with high probability in a *private* scenario, i.e., assuming that the covertext 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 II. The coding theorem that establishes $\mathcal{R}_{D,\,D_A}$ is proved in Section III. Section IV contains achievability results for other schemes that combine watermarking and compression. Extensions, conclusions and directions for further research are given in Section V.

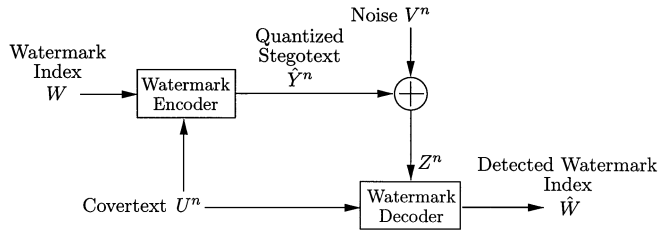

Fig. 2. The watermarking/authentication system with quantization.

II. The Rate Region

The watermarking/quantization system under consideration is shown in Fig. 2. In the embedding process, U^n is the i.i.d. n-dimensional Gaussian covertext of (per-symbol) variance P_U; W is the watermark index (independent of U^n) which is uniformly distributed over a set of size 2^{nR_W}; and $\hat{Y}^n$ is the (quantized) stegotext which can be found in a source codebook of size 2^{nR_Q}. (Since delivery of the watermarked data to the customer is noiseless, there is no need to explicitly show the source codebook index in Fig. 2.) 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:

$$n^{-1}E\|U^n - \hat{Y}^n\|^2 \le D \tag{1}$$

and

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

The converse and achievability results of Section III establish the region $\mathcal{R}_{D,\,D_A}$ of achievable rates $(R_Q,\, R_W)$, as expressed in (3), shown at the bottom of the page, where

$$P_W(\gamma) \triangleq \frac{\gamma(P_U+D) - 2P_U + 2\sqrt{P_U(\gamma D - P_U)(\gamma-1)}}{\gamma^2} \tag{4}$$

and $[\cdot]^+ \triangleq \max\{0, \cdot\}$. $\mathcal{R}_{D,\,D_A}$ is the shaded region in Fig. 3. Its upper boundary is composed of the following elements.

- The segment AB on the straight line
$$R_W = R_Q - \left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+.$$
- The curved segment BC defined by the equation
$$R_W = \max_{\gamma \in \left[\max\left\{1,\, \frac{P_U}{D}\right\},\, 2^{2R_Q}\right]} \cdot \min\left\{R_Q - \frac{1}{2}\log(\gamma),\, \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)\right\}$$
for R_Q in the interval
$$\left[\frac{1}{2}\log\left(\max\left\{1,\, \frac{P_U}{D}\right\} + \frac{|P_U - D|}{D_A}\right),\ \frac{1}{2}\log\left(1 + \frac{P_U+D}{D_A} + \frac{P_U}{D}\right)\right]$$
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 $\frac{1}{2}\log\left(1 + \frac{D}{D_A}\right)$.

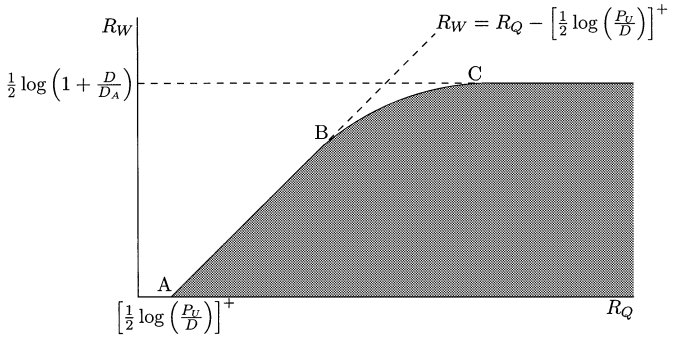

Fig. 3. The rate region $\mathcal{R}_{D,\,D_A}$ of achievable rate pairs $(R_Q,\, R_W)$.

Two key conclusions can be drawn from Fig. 3.

- For quantization rates
$$R_Q \in \left[\left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+,\ \frac{1}{2}\log\left(\max\left\{1,\, \frac{P_U}{D}\right\} + \frac{|P_U - D|}{D_A}\right)\right]$$
the watermarking rate R_W can be as high as $R_Q - \left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+$, 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
$$R_Q \ge \frac{1}{2}\log\left(1 + \frac{P_U+D}{D_A} + \frac{P_U}{D}\right)$$
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 covertext is solely due to watermarking, and where D represents the "signal" power in the additive white Gaussian noise (AWGN) 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

$$\mathcal{R}_{D,\,D_A} = \left\{(R_Q,\, R_W):\, R_Q \ge \left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+,\ R_W \le \max_{\gamma \in \left[\max\left\{1,\, \frac{P_U}{D}\right\},\, 2^{2R_Q}\right]} \min\left\{R_Q - \frac{1}{2}\log(\gamma),\, \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)\right\}\right\} \tag{3}$$

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. First, some definitions are in order.

Definition 1: A $(2^{nR_Q}, 2^{nR_W}, n)$ quantization/watermarking code consists of the following elements.

- A watermark set $\mathcal{M}_n = \{1, \ldots, 2^{nR_W}\}$.
- An encoding function f: $\mathcal{M}_n \times \mathcal{U}^n \rightarrow \hat{\mathcal{Y}}^n$ which maps a watermark index w and the covertext u^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: $\mathcal{Z}^n \times \mathcal{U}^n \rightarrow \mathcal{M}_n$ which maps the output of the channel z^n and the covertext u^n to an estimate $\hat{w}$ of w.

Here, $\mathcal{U} = \hat{\mathcal{Y}} = \mathcal{Z} = \mathbb{R}$. For random W and U^n, we have the random quantities $\hat{Y}^n = f(W, U^n)$ and $\hat{W} = g(Z^n, U^n)$. 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.

Definition 2: The probability of error in detecting watermark w is given by

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

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

$$\overline{\mathcal{D}} = E\left[n^{-1} \sum_{i=1}^{n} (U_i - f(W, U^n)_i)^2\right]$$

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

Definition 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 $\overline{\mathcal{D}} \leq D$. Moreover, a rate region $\mathcal{R}$ of pairs (R_Q, R_W) is achievable if every element of $\mathcal{R}$ is achievable.

The coding theorem is stated as follows.

Theorem 3.1: A quantization/watermarking code $(2^{nR_Q}, 2^{nR_W}, n)$ satisfies the transparency and robustness requirements (1) and (2), respectively, if and only if $(R_Q, R_W) \in \mathcal{R}_{D, D_A}$ (where $\mathcal{R}_{D, D_A}$ is defined in (3)).

The proof of Theorem 3.1 consists of two parts; the converse and the direct part. The converse part states that no rates outside $\mathcal{R}_{D, D_A}$ are achievable; the direct part states that $\mathcal{R}_{D, D_A}$ is indeed achievable. The two proofs can be found in Sections III-A and III-C, respectively.

A. Converse Theorem

The converse theorem is stated as follows.

Theorem 3.2: For any $(2^{nR_Q}, 2^{nR_W}, n)$ code that satisfies (1) and (2), the rate pair (R_Q, R_W) must lie in $\mathcal{R}_{D, D_A}$.

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

$$\frac{1}{n} \sum_{i=1}^{n} E(U_i - \hat{Y}_i)^2 = D. \tag{5}$$

By virtue of the monotonicity of the region $\mathcal{R}_{D, D_A}$ in D, the constraint can then be relaxed to an inequality, as in (1).

Since U^n is i.i.d. Gaussian, a standard converse rate-distortion theorem (e.g., [20]) yields

$$R_Q \geq \frac{1}{n} I(U^n; \hat{Y}^n) \geq \left[\frac{1}{2} \log\left(\frac{P_U}{D}\right)\right]^+. \tag{6}$$

This establishes the lower bound on R_Q in the definition of $\mathcal{R}_{D, D_A}$.

At this point, we define the inner product between two random vectors S^n, T^n as

$$\langle S^n, T^n \rangle \triangleq \frac{1}{n} \sum_{i=1}^{n} E[S_i T_i]. \tag{7}$$

We further define the following quantities:

$$P_{\hat{Y}} \triangleq \langle \hat{Y}^n, \hat{Y}^n \rangle \tag{8}$$

$$\mu_0 \triangleq \frac{\langle U^n, \hat{Y}^n \rangle}{P_{\hat{Y}}} \tag{9}$$

$$\lambda_0 \triangleq \frac{\langle U^n, \hat{Y}^n \rangle}{P_U} \tag{10}$$

$$\begin{aligned} P_{U|\hat{Y}} &\triangleq \frac{1}{n} \sum_{i=1}^{n} E\left[(U_i - \mu_0 \hat{Y}_i)^2\right] \\ &= P_U - \frac{(\langle U^n, \hat{Y}^n \rangle)^2}{P_{\hat{Y}}} \end{aligned} \tag{11}$$

$$\begin{aligned} P_{\hat{Y}|U} &\triangleq \frac{1}{n} \sum_{i=1}^{n} E\left[(\hat{Y}_i - \lambda_0 U_i)^2\right] \\ &= P_{\hat{Y}} - \frac{(\langle U^n, \hat{Y}^n \rangle)^2}{P_U} \end{aligned} \tag{12}$$

$$\gamma \triangleq \frac{P_U}{P_{U|\hat{Y}}} = \frac{P_{\hat{Y}}}{P_{\hat{Y}|U}} \tag{13}$$

where the second equalities in (11)–(13) are straightforward consequences of previous definitions.

The upper bound on R_W will follow from four lemmas. The first lemma provides a lower bound on $I(U^n, \hat{Y}^n)$ in terms of γ defined above.

Lemma 3.3: Let γ be as defined in (13). Then for U^n i.i.d. Gaussian, we have

$$\frac{1}{n} I(U^n; \hat{Y}^n) \geq \frac{1}{2} \log(\gamma)$$

with equality iff $(U_1, \hat{Y}_1), \ldots, (U_n, \hat{Y}_n)$ are i.i.d. Gaussian.

Proof: We have the following inequalities:

$$\begin{aligned} I(U^n; \hat{Y}^n) &= h(U^n) - h(U^n|\hat{Y}^n) \\ &= h(U^n) - h(U^n - \mu_0 \hat{Y}^n|\hat{Y}^n) \\ &\geq h(U^n) - h(U^n - \mu_0 \hat{Y}^n) \end{aligned} \tag{14}$$

where μ_0 was defined in (9), and (14) is true because conditioning reduces entropy. Next, we have

$$\begin{aligned} & n^{-1} h(U^n - \mu_0 \hat{Y}^n) \\ & \leq \frac{1}{n} \sum_{i=1}^{n} h(U_i - \mu_0 \hat{Y}_i) && (15) \\ & \leq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log(2\pi e) E\left[(U_i - \mu_0 \hat{Y}_i)^2\right] && (16) \\ & \leq \frac{1}{2} \log(2\pi e) \left(\frac{1}{n} \sum_{i=1}^{n} E\left[(U_i - \mu_0 \hat{Y}_i)^2\right] \right) && (17) \\ & = \frac{1}{2} \log(2\pi e) P_{U|\hat{Y}} && (18) \end{aligned}$$

where (16) is true because the differential entropy of a continuous random variable is upper-bounded by the differential entropy of a Gaussian variable with the same variance [20]; (17) is a consequence of the concavity of the $\log(\cdot)$ function; and (18) is due to (11). Hence, from (14), (18), and (13), we have

$$\frac{1}{n} I(U^n; \hat{Y}^n) \geq \frac{1}{2} \log\left(\frac{P_U}{P_{U|\hat{Y}}}\right) = \frac{1}{2} \log(\gamma). \tag{19}$$

Note that all relationships up to (17) hold with any scalar replacing μ_0. The value defined in (9) satisfies the orthogonality condition $\langle \hat{Y}^n, U^n - \mu_0 \hat{Y}^n \rangle = 0$ and, thus, yields the minimum mean-squared error (MMSE) (equal to $P_{U|\hat{Y}}$) in estimating U^n by a scalar multiple of $\hat{Y}^n$. Thus, among all choices of scalars, μ_0 as defined in (9) yields the tightest possible bound in (19).

We also have the following conditions for equality: in (14), iff the random vector $U^n - \mu_0 \hat{Y}^n$ is independent of $\hat{Y}^n$; in (15), iff the variables $U_i - \mu_0 \hat{Y}_i$ are independent; in (16), iff $U_i - \mu_0 \hat{Y}_i$ is Gaussian for every i; and in (17), iff $E[(U_i - \mu_0 \hat{Y}_i)^2]$ takes the same value for every i. It is straightforward to show that these four conditions are jointly equivalent to $(U_1, Y_1), \ldots, (U_n, Y_n)$ being i.i.d. Gaussian, and, thus, the proof is complete. □

The second lemma establishes the range of possible values of γ.

Lemma 3.4: Under the distortion constraint (5), the range of γ is the interval $\left[\max\{1, \frac{P_U}{D}\}, 2^{2R_Q}\right]$.

Proof: From (6) and Lemma 3.3 we have

$$R_Q \geq \frac{1}{n} I(U^n; \hat{Y}^n) \geq \frac{1}{2} \log(\gamma).$$

Hence, $\gamma \leq 2^{2R_Q}$, thus establishing the upper bound on γ.

For establishing the lower bound on γ, it suffices to consider (5) and (13). Specifically, from (13) we can easily see that

$$\gamma = \left(1 - \frac{(\langle U^n, \hat{Y}^n \rangle)^2}{P_U P_{\hat{Y}}}\right)^{-1} \tag{20}$$

and from (5) we have

$$\langle U^n, \hat{Y}^n \rangle = \frac{1}{2}(P_U + P_{\hat{Y}} - D). \tag{21}$$

Substituting (21) into (20) we obtain

$$\gamma = \left(1 - \frac{(P_U + P_{\hat{Y}} - D)^2}{4 P_U P_{\hat{Y}}}\right)^{-1}. \tag{22}$$

It can be easily shown that (22) is minimized when $P_{\hat{Y}} = |P_U - D|$, and the minimum value is $\max\left\{1, \frac{P_U}{D}\right\}$, as required. □

The third lemma establishes a relationship between $P_{\hat{Y}|U}$ and $P_W(\gamma)$; note that $P_{\hat{Y}|U}$ is defined in (12) while $P_W(\gamma)$ appears in (4).

Lemma 3.5: For any allowable value of γ

$$P_{\hat{Y}|U} \leq P_W(\gamma) \tag{23}$$

where $P_W(\gamma)$ and $P_{\hat{Y}|U}$ were defined in (4) and (12), respectively.

Proof: From (13), we have that $P_{\hat{Y}} = \gamma P_{\hat{Y}|U}$. Substituting into (22) we obtain

$$\gamma = \left(1 - \frac{\left(P_U + \gamma P_{\hat{Y}|U} - D\right)^2}{4 P_U \gamma P_{\hat{Y}|U}}\right)^{-1}. \tag{24}$$

Solving for $P_{\hat{Y}|U}$ yields

$$P_{\hat{Y}|U} = \frac{\gamma(P_U + D) - 2P_U \pm 2\sqrt{P_U(\gamma D - P_U)(\gamma - 1)}}{\gamma^2}. \tag{25}$$

The larger of these two values is equal to $P_W(\gamma)$, thereby completing the proof. □

The fourth lemma involves two chains of inequalities corresponding to the two upper bounds on R_W.

Lemma 3.6: For all $\epsilon > 0$ such that $\Pr\{W \neq \hat{W}\} < \epsilon$, we have

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

and

$$R_W \leq \frac{1}{2} \log\left(1 + \frac{P_{\hat{Y}|U}}{D_A}\right) + \epsilon.$$

Proof: The first chain of inequalities is as follows:

$$\begin{aligned} R_W &= n^{-1} H(W|U^n, V^n) && (26) \\ &= n^{-1} I(W; \hat{Y}^n|U^n, V^n) + n^{-1} H(W|U^n, \hat{Y}^n, V^n) \\ &\leq n^{-1} I(W; \hat{Y}^n|U^n, V^n) + n^{-1} H(W|U^n, Z^n) && (27) \\ &\leq n^{-1} I(W; \hat{Y}^n|U^n, V^n) + \epsilon && (28) \\ &= n^{-1} H(\hat{Y}^n|U^n, V^n) - n^{-1} H(\hat{Y}^n|W, U^n, V^n) + \epsilon \\ &= n^{-1} H(\hat{Y}^n|U^n) + \epsilon && (29) \\ &= n^{-1} H(\hat{Y}^n) - n^{-1}(H(\hat{Y}^n) - H(\hat{Y}^n|U^n)) + \epsilon \\ &\leq R_Q - n^{-1} I(\hat{Y}^n; U^n) + \epsilon && (30) \end{aligned}$$

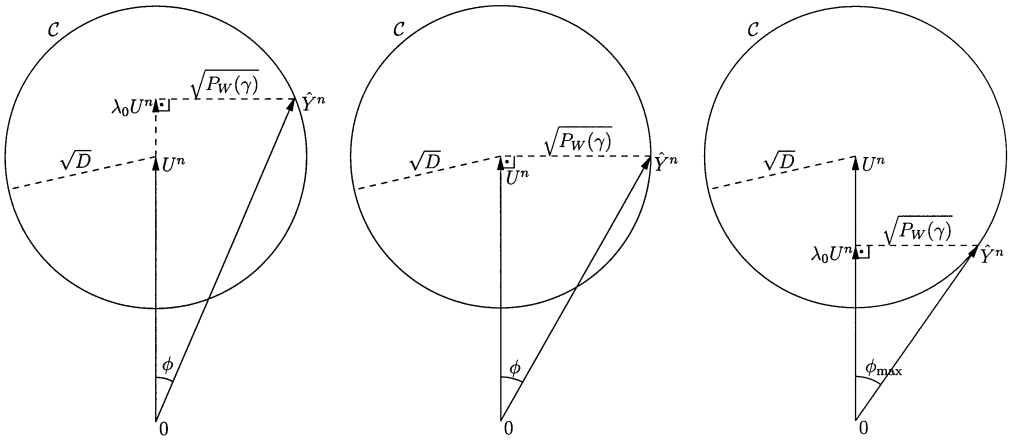

Fig. 4. The second-moment space $\mathcal{L}_2$ spanned by vectors U^n and $\hat{Y}^n$, shown for three different values of ϕ. The circle $\mathcal{C}$ is the locus of all $\hat{Y}^n$ such that $n^{-1}E\|U^n - \hat{Y}^n\|^2 = D$. As ϕ increases from 0, $P_W(\gamma)$ increases monotonically (case (a)) until it reaches its maximum value D (case (b)), then decreases monotonically until $\phi = \phi_{\max}$ (case (c)).

where (26) holds because U^n and V^n are independent of W; (27) follows from

$$H(W|U^n, \hat{Y}^n, V^n) = H(W|U^n, Z^n, \hat{Y}^n, V^n) \le H(W|U^n, Z^n);$$

(28) is a consequence of Fano's inequality; (29) holds because

$$H(\hat{Y}^n|W, U^n, V^n) = 0$$

(since $\hat{Y}^n$ is a function of W and U^n) and V^n is independent of U^n, $\hat{Y}^n$, and (30) follows from $R_Q \ge n^{-1}H(\hat{Y}^n)$.

The second chain of inequalities is as follows (where λ_0 was defined in (10)):

$$\begin{aligned}
R_W &= n^{-1}H(W|U^n) && (31)\\
&= n^{-1}I(W; Z^n|U^n) + n^{-1}H(W|U^n, Z^n)\\
&\le n^{-1}I(W; Z^n|U^n) + \epsilon && (32)\\
&= n^{-1}h(Z^n|U^n) - n^{-1}h(Z^n|U^n, W) + \epsilon\\
&= n^{-1}h(Z^n|U^n) - n^{-1}h(Z^n - \hat{Y}^n|U^n, W) + \epsilon && (33)\\
&= n^{-1}h(\hat{Y}^n - \lambda_0 U^n + V^n|U^n) - n^{-1}h(V^n) + \epsilon && (34)\\
&\le n^{-1}h(\hat{Y}^n - \lambda_0 U^n + V^n) - \frac{1}{2}\log(2\pi e)D_A + \epsilon\\
&\le \frac{1}{2}\log(2\pi e)\left(P_{\hat{Y}|U} + D_A\right) - \frac{1}{2}\log(2\pi e)D_A + \epsilon && (35)\\
&= \frac{1}{2}\log\left(1 + \frac{P_{\hat{Y}|U}}{D_A}\right) + \epsilon && (36)
\end{aligned}$$

where (31) holds because U^n is independent of W; (32) follows from Fano's inequality; (33) holds because $\hat{Y}^n$ is a function of U^n and W; (34) follows from the independence of V^n and (U^n, W); and (35) is easily established using a chain of inequalities similar to (15)–(18).

From (30) and (36), the lemma follows. □

The proof of the converse result can now be completed as follows:

$$\begin{aligned}
R_W &\le \min\left\{R_Q - \frac{1}{n}I(U^n; \hat{Y}^n),\ \frac{1}{2}\log\left(1 + \frac{P_{\hat{Y}|U}}{D_A}\right)\right\} + \epsilon && (37)\\
&\le \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)\right\} + \epsilon && (38)\\
&\le \max_{\gamma \in \left[\max\left\{1, \frac{P_U}{D}\right\}, 2^{2R_Q}\right]} \cdot \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)\right\} + \epsilon && (39)
\end{aligned}$$

where (37) is due to Lemma 3.6; (38) follows from Lemmas 3.3 and 3.5; and the maximum in (39) is taken over the range of values of γ established by Lemma 3.4. By taking $\epsilon \to 0$, the converse is proved. □

B. Explanatory Remarks

1) Geometrical Considerations: A geometrical interpretation of the arguments presented in the proof of the converse theorem can be obtained by considering the L_2-space spanned by the vectors U^n and $\hat{Y}^n$, with inner product defined in (7).

The geometry of this space is depicted in Fig. 4. The vector $\hat{Y}^n$ lies on a circle $\mathcal{C}$ centered at U^n and having radius $\sqrt{D}$ corresponding to the distortion constraint (5). The lengths of U^n and $\hat{Y}^n$ are given by $\sqrt{P_U}$ and $\sqrt{P_{\hat{Y}}}$, respectively, where $P_{\hat{Y}}$ was defined in (8). The angle between the two vectors is denoted by ϕ and is assumed to take values in $[0, \pi]$, as usual.

Note that Fig. 4 is drawn for the case $D \le P_U$, which is a safe assumption in most practical applications. The maximum value of ϕ is then $\phi_{\max} < \pi/2$, and is obtained when $\hat{Y}^n$ is tangent to $\mathcal{C}$. The maximum value of $\sin^2(\phi)$ is also

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

In the case $D > P_U$, the circle $\mathcal{C}$ encloses the origin and, thus, the range of values of ϕ is the entire interval $[0, \pi]$. Then the maximum value of $\sin^2\phi$ is unity. Combining the two cases together, we have that the maximum value of $\sin^2(\phi)$ is

$$\min\left\{1, \frac{D}{P_U}\right\}$$

and thus the minimum value of $\gamma = \sin^{-2}(\phi)$ is

$$\max\left\{1, \frac{P_U}{D}\right\}$$

as in Lemma 3.4.

As we argued in the proof of Lemma 3.3, the values λ_0 and μ_0, as defined in (10) and (9), respectively, have the following interpretation: $\lambda_0 U^n$ is the projection of $\hat{Y}^n$ on U^n, or equivalently, the MMSE estimator of $\hat{Y}^n$ among all scalar multiples on U^n; and similarly, $\mu_0\hat{Y}^n$ is the scalar MMSE estimator of U^n given $\hat{Y}^n$. $P_{\hat{Y}|U}$ and $P_{U|\hat{Y}}$, as defined in (12) and (11), respectively, are the resulting MMSE errors. As was seen in the proof of Lemma 3.5, $P_W(\gamma)$ is the larger of two possible values of $P_{\hat{Y}|U}$ corresponding to a particular value of γ (other than 1 or P_U/D). Fig. 4 clearly illustrates that, in the case $D < P_U$, there are two possible positions of $\hat{Y}^n$ on $\mathcal{C}$ for each value of ϕ smaller than $\phi_{\max}$ (correspondingly, for each γ greater than P_U/D). The position farthest from the origin is marked in Fig.4 (a) and (b), and the length of the error vector $\hat{Y}^n - \lambda_0 U^n$ is shown as $\sqrt{P_W(\gamma)}$.

2) The Upper Boundary of the Rate Region: In this part, we examine the behavior of the upper bound on R_W as a function of R_Q

$$r_W(R_Q) \triangleq \max_{\gamma \in \left[\max\left\{1, \frac{P_U}{D}\right\}, 2^{2R_Q}\right]} \cdot \min\left\{R_Q - \frac{1}{2}\log(\gamma), \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)\right\}. \quad (40)$$

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

The second argument of $\min\{\cdot, \cdot\}$ in (40) 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 $\hat{Y}^n - \lambda_0 U^n$ when U^n and $\hat{Y}^n$ are as shown in Fig. 4, with $\sin^{-2}(\phi) = \gamma$. Clearly, $\sqrt{P_W(\gamma)}$ increases monotonically as ϕ increases from $\phi = 0$ to

$$\phi = \arctan(\sqrt{D/P_U}) = \arcsin(\sqrt{D/(P_U + D)})$$

then decreases monotonically as ϕ increases to

$$\phi_{\max} = \arcsin(\min\{1, \sqrt{D/P_U}\}).$$

Equivalently (but in the reverse direction), as γ increases from $\gamma = \max\left\{1, \frac{P_U}{D}\right\}$ to $\gamma = 1 + \frac{P_U}{D}$ and then on to infinity, $P_W(\gamma)$ increases from $|P_U - D| \cdot \min\left\{1, \frac{D}{P_U}\right\}$ to D (its maximum value), and then decreases to 0. The function $\frac{1}{2}\log(1 + \frac{P_W(\gamma)}{D_A})$ has similar behavior, and is plotted in Fig. 5 against $\frac{1}{2}\log(\gamma)$. The initial (leftmost) and maximum R_W-ordinates on the curve are $\frac{1}{2}\log(1 + \frac{|P_U - D|}{D_A}\min\{1, \frac{D}{P_U}\})$ and $\frac{1}{2}\log(1 + \frac{D}{D_A})$, respectively.

The first argument of $\min\{\cdot, \cdot\}$ in (40) involves R_Q and decreases monotonically from $R_Q - \left[\frac{1}{2}\log(\frac{P_U}{D})\right]^+$ to zero as γ ranges over the interval $[\max\{1, \frac{P_U}{D}\}, 2^{2R_Q}]$ of maximization in (40). Plotted against $\frac{1}{2}\log(\gamma)$, it yields a line segment of slope -1 (in Fig. 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 Fig. 5. There are three regimes of interest:

a) In the first regime, the straight-line segment lies entirely below the curve (Fig. 5(a)). The maximin in (40) is then given by the maximum ordinate on the line segment, i.e.,

$$r_W(R_Q) = R_Q - \left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+.$$

This occurs for

$$R_Q \in \left[\left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+, \frac{1}{2}\log\left(\max\left\{1, \frac{P_U}{D}\right\} + \frac{|P_U - D|}{D_A}\right)\right].$$

b) In the second regime, the straight-line segment intersects the rising portion of the curve (Fig. 5(b)). The maximin in (40) 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

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

c) The third regime corresponds to all other values of R_Q, namely

$$R_Q > \frac{1}{2}\log\left(1 + \frac{P_U}{D} + \frac{(P_U + D)}{D_A}\right).$$

In this case, the straight-line segment either intersects the curve on its falling portion only (as in Fig. 5(c)), or does not intersect it at all. The maximin value in (40) is then given by the maximum ordinate on the curve, i.e.,

$$r_W(R_Q) = \frac{1}{2}\log\left(1 + \frac{D}{D_A}\right).$$

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 D and noise power D_A (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 $\mathcal{R}_{D, D_A}$ described in Section II.

Note: In the special case $D_A = 0$ (no attack), the curve in Fig. 5 is displaced to $+\infty$ and only the first regime remains, i.e., the bound on R_W is simply $R_W \le R_Q - \left[\frac{1}{2}\log(\frac{P_U}{D})\right]^+$. The converse theorem then reduces to the channel coding part of the converse theorem in [21], and also the converse theorem of [14] for $R_F = 0$.

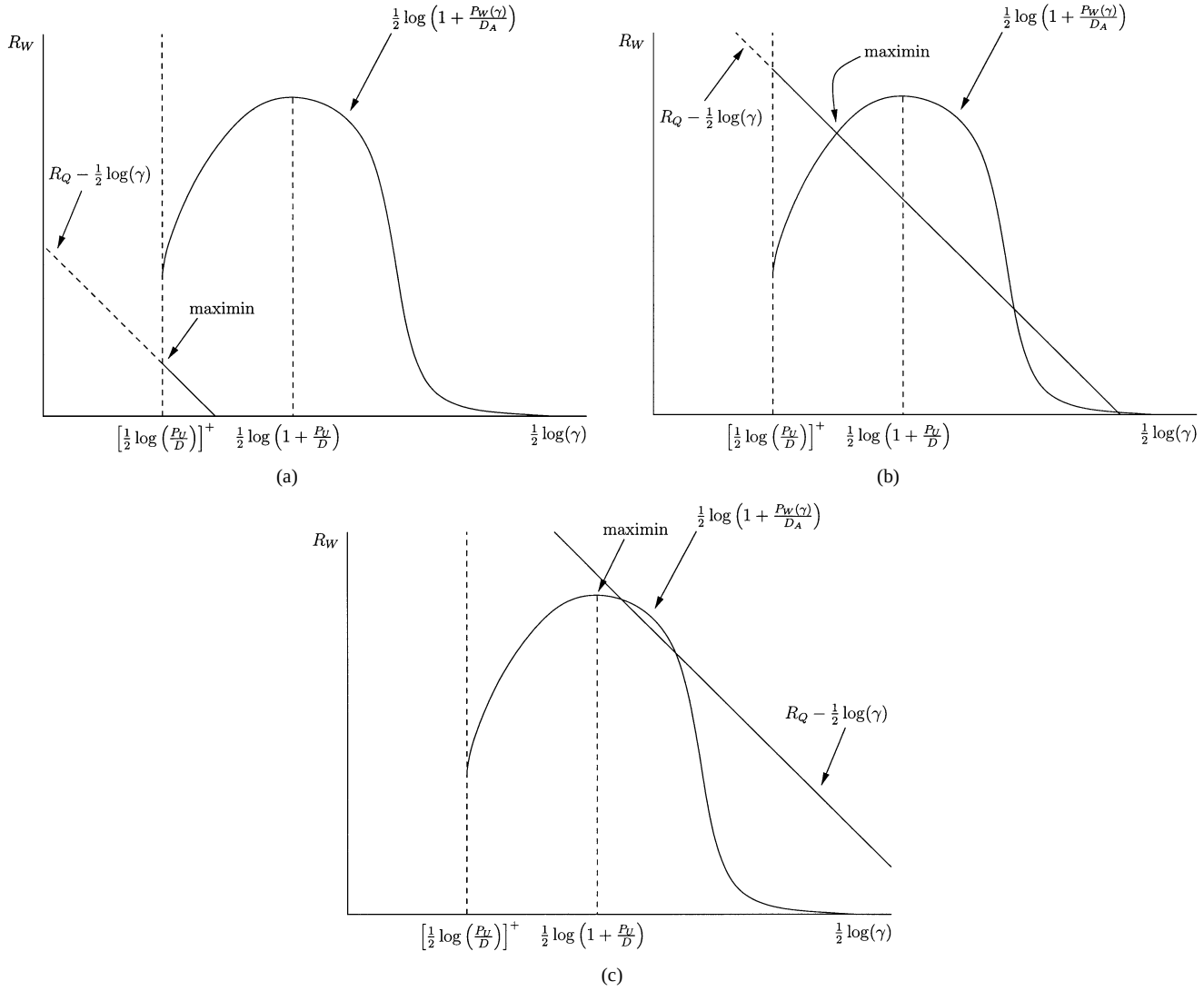

Fig. 5. Plots of $R_Q - \frac{1}{2}\log(\gamma)$ and $\frac{1}{2}\log(1+\frac{P_W(\gamma)}{D_A})$ and determination of the maximin point for various values of R_Q.

C. Direct Theorem

The direct theorem is stated as follows.

Theorem 3.7: For any rate pair $(R_Q, R_W) \in \mathcal{R}_{D,D_A}$, there exists a $(2^{nR_Q}, 2^{nR_W}, n)$ code such that (1) and (2) are satisfied.

Proof: As required for $\mathcal{R}_{D,D_A}$, we limit the quantization rate to

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

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 [12], in that 2^{nR_W} quantizers, each one indexed by a different watermark, are employed.

Codebook Generation: Let

$$\gamma \in \left[\max\left\{1, \frac{P_U}{D}\right\}, 2^{2R_Q}\right].$$

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.,

$$R_Q = R_W + R_1. \tag{41}$$

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

Watermark Embedding: Given U^n and a deterministic w, the embedder identifies within the wth codebook the first codeword $\tilde{Y}^n(w, q)$ such that the pair $(U^n, \tilde{Y}^n(w, q))$ lies in the set $T_{U,\hat{Y}}(\epsilon)$ of typical pairs with respect to a bivariate Gaussian distribution $p_{U,\hat{Y}}$ having mean zero and covariance

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

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 wth codebook is jointly typical with U^n, then the embedder outputs $\hat{Y}^n(w) = 0$. In this manner, 2^{nR_W} watermarked versions of the covertext U^n

are obtained: $\hat{Y}^n(1), \ldots, \hat{Y}^n(2^{nR_W})$. Clearly, for random W, the embedder output is $\hat{Y}^n(W)$.

Note that the second moments in $K_{U,\hat{Y}}$ are consistent with the geometry of Fig. 4, with $\gamma = \sin^{-2}(\phi)$. Specifically, if the pair $(U^n, \hat{Y}^n)$ lies in $T_{U,\hat{Y}}(\epsilon)$, then the empirical second moments

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

are within ϵ (or a factor thereof) of the average values shown implicitly in Fig. 4. This also means that the distortion constraint (1) is essentially met (since ϵ-differences can be safely ignored).

Decoding: Again, the decoder has access to the covertext U^n. 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 U^n a single $\hat{Y}^n(\hat{w})$ such that the triplet $(U^n, \hat{Y}^n(\hat{w}), Z^n)$ lies in $T^n_{U,\hat{Y},Z}(\epsilon)$, the set of typical triplets with respect to the trivariate Gaussian distribution $p_{U,\hat{Y},Z}$ having zero mean and covariance matrix $K_{U,\hat{Y},Z}$ shown at the bottom of the page. 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 $p_{U,\hat{Y},Z}(u, \hat{y}, z) = p_{U,\hat{Y}}(u, \hat{y})p_V(z - \hat{y})$, where p_V is the marginal of the attack noise V^n. The quantities of interest here are the determinants $|K_{U,\hat{Y}}| = P_U P_W(\gamma)$, $|K_{U,Z}| = P_U(P_W(\gamma) + D_A)$, and $|K_{U,\hat{Y},Z}| = P_U D_A P_W(\gamma)$, as well as the mutual information values

$$I(U;\hat{Y}) = \frac{1}{2}\log\left(\frac{P_U\,\gamma P_W(\gamma)}{|K_{U,\hat{Y}}|}\right) = \frac{1}{2}\log(\gamma) \quad (42)$$

and

$$I(\hat{Y};Z|U) = \frac{1}{2}\log\left(\frac{|K_{U,Z}|\,|K_{U,\hat{Y}}|}{P_U\,|K_{U,\hat{Y},Z}|}\right) = \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right). \quad (43)$$

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 $(U^n, \hat{Y}^n(1, q)) \in T_{U,\hat{Y}}$.
- E_2: There exists a $\tilde{Y}^n(1, q) = \hat{Y}^n(1)$ such that $(U^n, \hat{Y}^n(1)) \in T_{U,\hat{Y}}$, but $(U^n, \hat{Y}^n(1), Z^n) \notin T_{U,\hat{Y},Z}$.
- E_3: $(U^n, \hat{Y}^n(1), Z^n) \in T_{U,\hat{Y},Z}$ but there also exists a $k > 1$ such that $(U^n, \hat{Y}^n(k), Z^n) \in T_{U,\hat{Y},Z}$.

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 [20], we know that if $R_1 > I(U;\hat{Y})$ (the mutual information of the bivariate $p_{U,\hat{Y}}$ defined above), then $\Pr(E_1) \to 0$ as $n \to \infty$. Since $R_1 = R_Q - R_W$ from (41), and $I(U;\hat{Y}) = \frac{1}{2}\log(\gamma)$ from (42), it follows that $\Pr(E_1) \to 0$ provided that

$$R_W < R_Q - \frac{1}{2}\log(\gamma). \quad (44)$$

Behavior of $\Pr(E_2)$: To show that $\Pr(E_2) \to 0$, it suffices to show that the triplet $(U^n, \hat{Y}^n(1), Z^n)$ lies in $T_{U,\hat{Y},Z}$ with probability approaching unity asymptotically. In the previous paragraph, we showed that

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

Since $Z^n = \hat{Y}^n(1) + V^n$ and V^n is independent of $(U^n, \hat{Y}^n(1))$, it follows easily that the empirical correlations obtained from $(U^n, \hat{Y}^n(1), Z^n)$ are within ϵ (or a factor thereof) of the corresponding entries of $K_{U,\hat{Y},Z}$ with probability approaching unity asymptotically. Typicality with respect to $p_{U,\hat{Y},Z}$ thus holds (also with probability approaching unity).

Behavior of $\Pr(E_3)$:

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

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

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

and by construction, $Z^n = \hat{Y}^n(1) + V^n$ is independent of $\hat{Y}^n(2)$ given U^n, a standard argument (cf. the proof of [20, Theorem 8.6.1]) yields

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

where the conditional mutual information is computed with respect to the trivariate $p_{U,\hat{Y},Z}$ defined earlier. From (43), we have that

$$I(Z;\hat{Y}|U) = \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right)$$

and, therefore, $\Pr(E_2) \to 0$ provided

$$R_W < \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A}\right). \quad (45)$$

$$K_{U,\hat{Y},Z} = \begin{bmatrix} P_U & \sqrt{(\gamma-1)P_U P_W(\gamma)} & \sqrt{(\gamma-1)P_U P_W(\gamma)} \\ \sqrt{(\gamma-1)P_U P_W(\gamma)} & \gamma P_W(\gamma) & \gamma P_W(\gamma) \\ \sqrt{(\gamma-1)P_U P_W(\gamma)} & \gamma P_W(\gamma) & \gamma P_W(\gamma) + D_A \end{bmatrix}.$$

From (44) and (45) it follows that R_W is achievable provided

$$R_W < \min\left\{R_Q - \frac{1}{2}\log(\gamma),\ \frac{1}{2}\log\left(1+\frac{P_W(\gamma)}{D_A}\right)\right\}. \tag{46}$$

Choosing $\gamma \in \left[\max\left\{1, \frac{P_U}{D}\right\}, 2^{2R_Q}\right]$ so as to maximize the right-hand side of (46), we can achieve the whole region $\mathcal{R}_{D,D_A}$.

We have thus proved that if $(R_Q, R_W) \in \mathcal{R}_{D,\,D_A}$ 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 $\mathcal{R}_{D,\,D_A}$ 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. □

IV. 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, for the following systems.

- Regular QIM [12], where no knowledge of the covertext is available at the decoder (public scenario).
- Additive watermarking, where the embedder computes the weighted sum of the covertext and a watermark-dependent signal and then compresses the resulting vector using a universal (watermark nonspecific) quantizer. A private detection scenario is assumed in this case.

Although our focus is on achievability results, the rate region $\mathcal{R}_{D,\,D_A}$ derived in Section III can be taken as an outer bound on the achievable rate region of both schemes considered in this section.

A. Regular QIM, 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 stegotext (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 covertext space with $2^{n(R_Q-R_W)}$ representation vectors (codewords). The watermark W is embedded into a covertext U^n by quantizing U^n using the Wth quantizer, yielding a representation vector $\hat{Y}^n$. Detection of the watermark W in forgery 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 III mapped Z^n to one of 2^{nR_W} representation vectors, each taken from a *different* codebook.)

As discussed in [12], achievable rates for regular QIM (also called "hidden" QIM) under constraints (1) and (2) can be found using a single-letter formula developed by Gel'fand and Pinsker [22]

$$R = I(T;\,Z) - I(T;\,U).$$

Here, R is an achievable channel code rate for a memoryless channel with input variable A (not shown in the above formula), output variable Z, and i.i.d. channel side information U^n available to the transmitter (only). T is an auxiliary i.i.d. random variable which can be chosen (together with A) so as to maximize R subject to the Markov constraint $T \to A,\ U \to Z$. The proof of the direct (achievability of R) result in [22] employs approximately $2^{nI(T;Z)}$ auxiliary sequences T^n generated randomly in an i.i.d. fashion. In the context of QIM, the memoryless channel is none other than the attack channel; the auxiliary sequences are the source codewords themselves; and the side information U^n is the covertext. This leads to the following relationships:

$$R_Q = I(\hat{Y};\,Z) = I(\hat{Y};\,\hat{Y}+V) \tag{47}$$

$$R_W = [I(\hat{Y};\,Z) - I(\hat{Y};\,U)]^+. \tag{48}$$

The trivariate distribution $p_{U,\hat{Y},Z}(u,\,\hat{y},\,z)$ can be taken as the Gaussian in the proof of the direct theorem in Section III. Thus, $p_{U,\hat{Y},Z}(u,\,\hat{y},\,z) = p_{U,\hat{Y}}(u,\,\hat{y})p_V(z-\hat{y})$, where U and $V = Z - \hat{Y}$ are independent with mean zero and variances P_U and D_A, respectively; and $\hat{Y}$ also has mean zero and satisfies $E(\hat{Y}-U)^2 = D$. It should be noted again that the second moments of $p_{U,\hat{Y},Z}(u,\,\hat{y},\,z)$ are consistent with the geometry of Fig. 4.

We briefly investigate the behavior of (48) as R_Q (given by (47)) varies. Letting $P_{\hat{Y}} = \gamma P_W(\gamma) = E(\hat{Y}^2)$, we have from (47)

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

and thus,

$$P_{\hat{Y}} = D_A(2^{2R_Q}-1). \tag{49}$$

Also, (48) gives

$$R_W = \left[R_Q - \frac{1}{2}\log(\gamma)\right]^+. \tag{50}$$

Setting $P_{\hat{Y}|U} = P_W(\gamma) = P_{\hat{Y}}/\gamma$ in (25) and expressing γ in terms of P_U, $P_{\hat{Y}}$, and D, we obtain (with the aid of (49)) (51), shown at the bottom of the page. The achievable region implied

$$R_W = \left[R_Q - \frac{1}{2}\log\left(\frac{P_U D_A(2^{2R_Q}-1)}{P_U D_A(2^{2R_Q}-1) - \frac{1}{4}(P_U + D_A(2^{2R_Q}-1) - D)^2}\right)\right]^+. \tag{51}$$

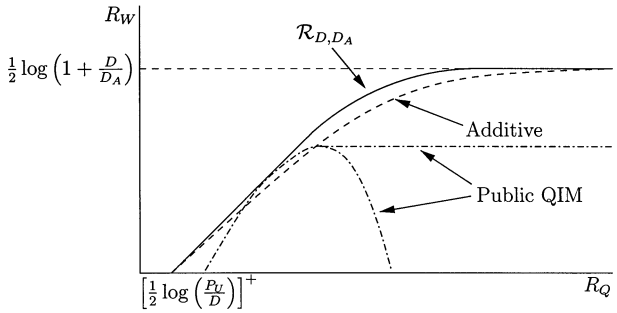

Fig. 6. Inner bounds on the achievable rate regions for public QIM and private additive schemes. $\mathcal{R}_{D,\,D_A}$ is an outer bound on the achievable rate regions of both schemes.

by (51) is shown in Fig. 6. The range of values of R_Q for which R_W in (51) is nonzero is a subinterval of

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

whose exact endpoints are given by the roots of a cubic. Expression (51) is shown in Fig. 6 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 Fig. 6, 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). Whether there exists an encoding scheme such that the maximum achievable watermarking rate is equal to $\frac{1}{2}\log(1+\frac{D}{D_A})$ (for a finite value of R_Q) is still an open question.

B. Additive Watermarking, Private Scenario

Additive watermarking schemes (see, e.g., [9], [14]) 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 covertext 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 the data.

In general, additive watermarking reduces to the computation of

$$Y^n=\alpha U^n+\beta x^n(W) \tag{52}$$

where W is the index of the watermark and $x^n(\cdot)$ is an n-dimensional signal that does not depend on the covertext U^n. α, β are nonzero 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 ((1) in this case). The decoder attempts to detect W given $\hat{Y}^n$ and U^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 since the distortion constraint is between $\hat{Y}^n$ and U^n (not between $\hat{Y}^n$ and the argument of the quantizer f), compressing Y^n is equivalent to compressing $\alpha^{-1}Y^n$, i.e., $\hat{Y}^n=f(Y^n)=g(\alpha^{-1}Y^n)$. This effectively eliminates the parameter α in (52). Also, the parameter β can be absorbed in the power of the watermark. Thus, we can use the simpler form[2]

$$Y^n=U^n+x^n(W).$$

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

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

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 U^n and $\hat{Y}^n$ is, with probability approaching unity, no larger than $D+\epsilon$, which, in turn, implies that the distortion constraint (1) is essentially satisfied. By the usual rate-distortion argument, taking

$$R_Q=I(Y;\hat{Y})+\epsilon \tag{53}$$

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, $I(Y;\hat{Y})\geq\left[\frac{1}{2}\log(\frac{P_U}{D})\right]^+$ with equality iff $P_X=0$ and $P_U=P_{\hat{Y}}+D$.)

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

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

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

$$R_W=I(X;U,Z)-\epsilon \tag{54}$$

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

Solving (53) for P_X, then substituting into (54) and letting $\epsilon\to 0$, we obtain the achievable watermarking rate of expression (55), which is nonnegative for $R_Q\geq\left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+$. Equation (55) is maximized when $P_{\hat{Y}}$ is equal to (56), yielding

[2]In the case of no compression, scaling of the covertext plays an important role in determining capacity; see [6], [7]. Here, however, compression may include whatever scaling of Y^n is required in order to satisfy the distortion constraint; and the power of the watermark X^n is optimized to maximize the watermarking rate. Hence, additional scaling of U^n by α does not provide an additional degree of freedom.

final expression (57) (see (55)–(57) at the bottom of the page). The corresponding curve is also shown in Fig. 6 (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 $Y^n = U^n + X^n$ 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 covertexts and distributes them in compressed form, such that an average distortion constraint is met. We assumed that the compressed stegotexts are further corrupted by Gaussian attacks. By means of a coding theorem, we established the region $\mathcal{R}_{D,\,D_A}$ of achievable quantization and watermarking rates such that the error probability in decoding (in a private scenario) the embedded message from a forgery approaches zero asymptotically in n. Moreover, we presented achievability results for the public version of the regular QIM scheme, as well as for additive watermarking/quantization schemes.

The expression of $\mathcal{R}_{D,\,D_A}$ reveals two surprising facts: 1) at low quantization rates, Gaussian attacks do not degrade the system performance; and 2) there exists a quantization rate threshold above which quantization does not hinder the watermark detection, i.e., the watermarking rate can be as high as in the case of no compression. Intuitively, these facts can be explained as follows: 1) At low R_Q, the quantization cells are large enough, thus, inducing large detection regions; Gaussian noise does not, therefore, cause a significant amount of error (in terms of the error exponent). In other words, a detection error is caused mainly due to the degradation introduced by the compression. 2) There exists a finite R_Q such that the representation sequences and the corresponding decoding regions (given the covertext U^n) have the geometrical properties of an optimal code, which achieves the capacity of an AWGN channel.

A. Extensions to Above Model

There are many possible extensions to the problem of joint watermarking and compression treated in this paper. In [23], the following problems are investigated and the corresponding rate regions are characterized.

1) Game Played Between Watermarker and Attacker: The attacker, who (presumably) knows the statistics of the watermarking strategy, is free to choose any memoryless attack that satisfies the average distortion constraint

$$\frac{1}{n} E\|Z^n - \hat{Y}^n\|^2 \le D_A$$

while the watermark decoder (but not the encoder) knows the statistics of the attack. As is proved in [23], the rate region becomes

$$\mathcal{R}^{game}_{D,\,D_A} = \left\{ (R_Q,\, R_W)\colon R_Q \ge \left[\frac{1}{2}\log\left(\frac{P_U}{D}\right)\right]^+ \right. \\ \left. R_W \le \max_{\gamma \in \Gamma(R_Q,\,D,\,D_A)} \cdot \min\left\{ R_Q - \frac{1}{2}\log(\gamma),\, \frac{1}{2}\log\left(1 + \frac{P_W(\gamma)}{D_A} - \frac{1}{\gamma}\right)\right\}\right\}$$

where

$$\Gamma(R_Q,\, D,\, D_A) \triangleq \left[\max\{1,\, P_U/D\},\, 2^{2R_Q}\right] \cap \{\gamma\colon \gamma P_W(\gamma) > D_A\}.$$

It is further proved that the optimal memoryless attack (from the attacker's point of view) corresponds to optimum compression of a Gaussian source with distortion D_A. Similar observations were made in [6], [7] for the case of no compression.

2) General Gaussian Distributions: Here, it is assumed that: 1) the attack noise is additive and Gaussian but not necessarily stationary (or ergodic), and 2) the covertext U^n is Gaussian but not stationary, either. Although the achievable rate region may not have a limit as $n \to \infty$, the probability of error can be made to approach zero for very large n. The proof of the coding theorem relies on the fact that general Gaussian processes satisfy a version of the asymptotic equipartition property; more details on this property can be found in [23]–[26].

B. Future Work

For future research directions, we briefly mention a number of problems. One interesting problem is the determination of the rate region in a public scenario. Such a region should be a subset of $\mathcal{R}_{D,\,D_A}$, but whether it is a proper subset is still an open question.

Another interesting problem would be to consider more general attacks, possibly with different distortion constraints (e.g., of the almost-sure type [6] or the large-deviations type [10]).

$$R_W = \frac{1}{2}\log\left(\frac{2^{2R_Q}(2D(P_U + P_{\hat{Y}}) - D^2 - (P_U - P_{\hat{Y}})^2 + 4D_A P_U)}{4P_U(2^{2R_Q}D_A + P_{\hat{Y}})}\right) \tag{55}$$

$$P_{\hat{Y}} = -2^{2R_Q}D_A + \sqrt{(2^{2R_Q}D_A + D)^2 + P_U(P_U + 2D_A(2^{2R_Q} - 2) - 2D)} \tag{56}$$

$$R_W = \frac{1}{2}\log\left(\frac{2^{2R_Q}\left(4P_U(D+D_A) - \left(D + P_U + 2^{2R_Q}D_A - \sqrt{(2^{2R_Q}D_A + D)^2 + P_U(P_U + 2D_A(2^{2R_Q} - 2) - 2D)}\right)^2\right)}{4P_U\sqrt{(2^{2R_Q}D_A + D)^2 + P_U(P_U + 2D_A(2^{2R_Q} - 2) - 2D)}}\right). \tag{57}$$

Also, as we pointed out in Section I, our attack model assumes that the attacker does not have access to the source codebook, and, thus, the occurrence of an attack is almost certainly detectable. It would be interesting to consider other models where the attacker has access to either the entire source codebook or a subset of it.

ACKNOWLEDGMENT

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

REFERENCES

[1] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, vol. 86, pp. 1064–1087, June 1998.

[2] F. Petitcolas, R. Anderson, and M. Kuhn, "Information hiding—A survey," *Proc. IEEE*, vol. 87, pp. 1062–1078, July 1999.

[3] M. Barni, F. Bartolini, I. Cox, J. Hernandez, and F. Perez-Gonzalez, "Digital watermarking for copyright protection: A communications perspective," *IEEE Commun. Mag.*, vol. 39, pp. 90–133, Aug. 2001.

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

[5] I. Cox, J. Bloom, and M. Miller, *Digital Watermarking*. San Mateo, CA: Morgan Kaufmann, 2001.

[6] A. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1639 –1667, June 2002.

[7] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, pp. 563–593, Mar. 2003.

[8] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, pp. 420–430, Mar. 2000.

[9] D. Karakos and A. Papamarcou, "A relationship between quantization and distribution rates of digitally watermarked data," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, June 2000, p. 47.

[10] A. Somekh-Baruch and N. Merhav, "On the error exponent and capacity games of private watermarking systems," *IEEE Trans. Inform. Theory*, vol. 49, pp. 537–562, Mar. 2003.

[11] Y. Steinberg and N. Merhav, "Identification in the presence of side information with application to watermarking," *IEEE Trans. Inform. Theory*, vol. 47, pp. 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 Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.

[13] A. Cohen and A. Lapidoth, "The capacity of the vector Gaussian watermarking game," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 5.

[14] D. Karakos and A. Papamarcou, "Fingerprinting, watermarking and quantization of Gaussian data," in *Proc. 39th Allerton Conf. Communication, Control and Computing (Invited Talk)*, Monticello, IL, Oct. 2001.

[15] A. Somekh-Baruch and N. Merhav, "On the capacity game of public watermarking systems," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, p. 223. Also, submitted to the *IEEE Trans. Inform. Theory*, [Online]. Available: http://tiger.technion.ac.il/users/merhav.

[16] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 827–838, Nov. 1987.

[17] N. Merhav and S. Shamai (Shitz). (2002, May) On joint source-channel coding for the Wyner-Ziv source and the Gel'fand–Pinsker channel. *IEEE Trans. Inform. Theory* [Online]. Available: http://tiger.technion.ac.il/users/merhav; submitted for publication

[18] M. Ramkumar and A. Akansu, "Theoretical capacity measures for data hiding in compressed images," *Proc. SPIE*, vol. 3528, pp. 482–492, Nov. 1998.

[19] D. Kundur, "Implications for high capacity data hiding in the presence of lossy compression," in *Proc. IEEE Int. Conf. Information Technology: Coding and Computing*, Las Vegas, NV, Mar. 2000, pp. 16–21.

[20] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[21] D. Karakos and A. Papamarcou, "A relationship between quantization and distribution rates of digitally watermarked data," Univ. Maryland, Inst. Syst. Res. Tech. Rep TR 2000-51, [Online]. Available: http://www.isr.umd.edu/TechReports., Dec. 2000.

[22] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[23] D. Karakos, "Digital watermarking, fingerprinting and compression: An information-theoretic perspective," Ph.D. dissertation, Univ. of Maryland, College Park, MD, 2002.

[24] W. Yu, A. Sutivong, D. Julian, T. Cover, and M. Chiang, "Writing on colored paper," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 302.

[25] ——, (2002) Writing on colored paper. [Online]. Available: http://www.comm.utoronto.ca/~weiyu/

[26] T. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inform. Theory*, vol. 35, pp. 37–43, Jan. 1989.