[From *Proc. of SPIE: Security and Watermarking of Multimedia Contents II*, vol. 3971.]

Preprocessed and postprocessed quantization index modulation methods for digital watermarking

Brian Chen and Gregory W. Wornell

Research Laboratory of Electronics and
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

Quantization index modulation (QIM) methods, a class of digital watermarking and information embedding methods, achieve very efficient trade-offs among the amount of embedded information (rate), the amount of embedding-induced distortion to the host signal, and the robustness to intentional and unintentional attacks. For example, we show that against independent additive Gaussian attacks, which are good models for at least some types of uninformed and unintentional attacks, QIM methods exist that achieve the best possible rate-distortion-robustness trade-offs (*i.e.*, capacity) asymptotically at high rates and achieve performance within a few dB of capacity at all finite rates. Furthermore, low-complexity realizations of QIM methods, such as so-called dither modulation, have also been shown to achieve favorable rate-distortion-robustness trade-offs.

We further develop preprocessing and postprocessing techniques that enable QIM to fully achieve capacity, not only against Gaussian attacks but also against other types of attacks as well. One practical postprocessing technique we develop we refer to as distortion compensation. Distortion compensation has the property that when suitably optimized it is sufficient for use in conjunction with QIM to achieve capacity against Gaussian attacks and against square-error distortion-constrained attacks. More generally, we present the results of a comparative information theoretic analysis of the fundamental performance limits of QIM, distortion-compensated QIM, and other watermarking methods and demonstrate practically achievable gains with experimental results.

Keywords: digital watermarking, dither modulation, quantization index modulation, distortion compensation, capacity, information embedding, data hiding, steganography

1. INTRODUCTION

Copyright notification and enforcement, authentication, and covert communication — these are just a few of the emerging multimedia security applications for digital watermarking and information embedding methods,[1,2] methods for embedding one signal, an "embedded signal" or "watermark", within another signal, a "host signal". The embedding must be done such that the embedded signal causes no serious degradation to its host, *i.e.*, the embedding-induced distortion must be small. At the same time, the embedding must be robust to common degradations to the composite host and watermark signal, which in some applications result from deliberate attacks. Ideally, whenever the host signal survives these degradations, the watermark also survives. Finally, for given distortion and robustness levels, one would like to embed as much data as possible in a given host signal, or equivalently, one would like to maximize the embedding rate.

Thus, one can evaluate a digital watermarking method or class of methods by its achievable rate-distortion-robustness trade-offs, and quantization index modulation (QIM) methods[3–5] are one class of methods that have been shown to achieve very favorable rate-distortion-robustness trade-offs. For example, due to their host-interference rejection properties, they perform provably better than additive spread spectrum methods[6–8] against no-key, square-error distortion-constrained attacks.[5] Furthermore, these methods also outperform other host-interference rejecting methods such as quantization-and-perturbation,[9] which may be viewed as a form of generalized low-bit modulation.[4]

The authors' email addresses and web pages are:
B. Chen: bchen@mit.edu, http://web.mit.edu/bchen/www/home.html
G. W. Wornell: gww@allegro.mit.edu, http://allegro.mit.edu/dspg/gww.html

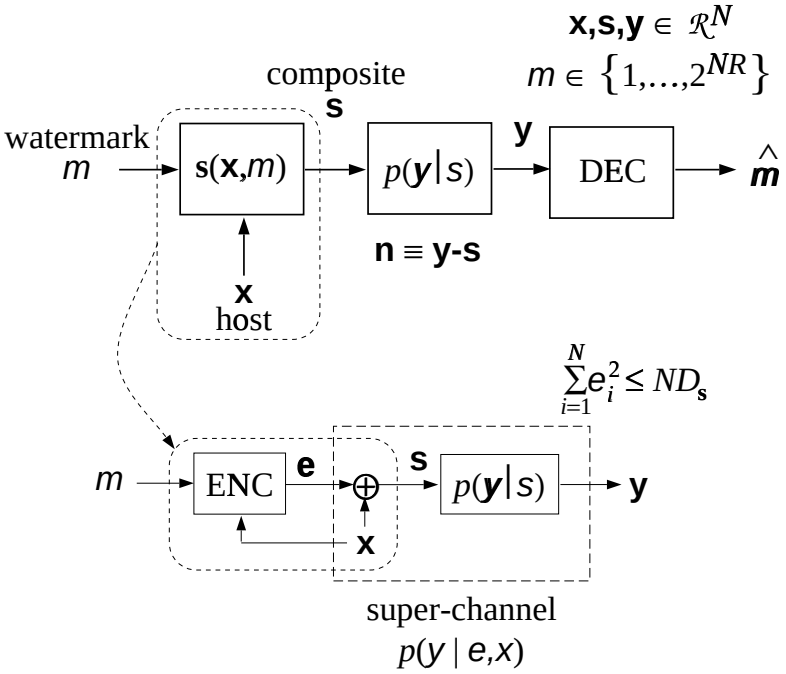

Figure 1. Information-embedding problem models. A message m is embedded in the host signal vector $\mathbf{x}$ using some embedding function $\mathbf{s}(\mathbf{x}, m)$. Equivalently, we may view the host signal as the state of a super-channel, with a host-dependent distortion signal $\mathbf{e}$ as the input. The decoder extracts an estimate $\hat{m}$ of m from the noisy channel output $\mathbf{y}$.

In fact, as we show in this paper, with the proper preprocessing and postprocessing, QIM methods exist that achieve the best possible rate-distortion-robustness performance, *i.e.*, capacity, against *any* fixed attack. We also show that distortion-compensated QIM[3,4] methods, a special subclass of postprocessed QIM methods, exist that achieve capacity against both additive Gaussian noise attacks and active, square-error distortion-constrained attacks on private-key systems. Finally, we present experimental results demonstrating the achievable performance of practical, low complexity, distortion-compensated QIM methods called distortion-compensated spread-transform dither modulation.

2. PROBLEM MODEL

The two equivalent problem models of Fig. 1 capture the fundamental features of most digital watermarking applications from two different perspectives. The top model represents the view that digital watermarking is the simultaneous communication or multiplexing of two signals, a watermark m and a host signal $\mathbf{x}$, subject to a distortion constraint between the host signal and composite signal $\mathbf{s}$. The second view, represented by the bottom model of Fig. 1, is that digital watermarking is communication of a distortion signal $\mathbf{e}$ over a channel with a state variable known at the encoder, where the state variable or side information is the host signal $\mathbf{x}$.* We expound on these views below and show that the two models are mathematically equivalent. However, depending on the context, one model may be more convenient than the other, so it is helpful to keep both in mind.

We wish to embed some digital information or watermark m in some host signal vector $\mathbf{x} \in \Re^N$. This host signal could be a vector of pixel values or Discrete Cosine Transform (DCT) coefficients from an image, for example. Alternatively, the host signal could be a vector of samples or transform coefficients, such as Discrete Fourier Transform (DFT) or linear prediction coding coefficients, from an audio or speech signal. We wish to embed at a rate of R_m bits per dimension (bits per host signal sample) so we can think of m as an integer, where

$$m \in \left\{1, 2, \ldots, 2^{NR_m}\right\}. \tag{1}$$

*Cox, *et al.*, have also recognized that one can view watermarking in this way.[10]

An embedding function, denoted $\mathbf{s}(\mathbf{x}, m)$ in Fig. 1, maps the host signal $\mathbf{x}$ and embedded information m to a composite signal $\mathbf{s} \in \Re^N$. The embedding should not unacceptably degrade the host signal, so we have some distortion measure $D(\mathbf{s}, \mathbf{x})$ between the composite and host signals. For example, one might choose the square-error distortion measure

$$D(\mathbf{s}, \mathbf{x}) = \frac{1}{N}\|\mathbf{s} - \mathbf{x}\|^2. \tag{2}$$

In some cases we may measure the expected distortion $D_{\mathbf{s}} = E[D(\mathbf{s}, \mathbf{x})]$. Alternatively, we can write the embedding function $\mathbf{s}(\mathbf{x}, m)$ as the sum of the host signal $\mathbf{x}$ and a host-dependent distortion signal $\mathbf{e}(\mathbf{x}, m)$,

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{e}(\mathbf{x}, m),$$

simply by defining the distortion signal to be $\mathbf{e}(\mathbf{x}, m) \stackrel{\Delta}{=} \mathbf{s}(\mathbf{x}, m) - \mathbf{x}$. Thus, as shown in the bottom portion of Fig. 1, one can view $\mathbf{e}$ as the input to a super-channel that consists of the cascade of an adder and the true channel, which we describe below. As stated above, the host signal $\mathbf{x}$ is a state of this super-channel that is known at the encoder. The measure of distortion $D(\mathbf{s}, \mathbf{x})$ between the composite and host signals maps onto a host-dependent measure of the size $P(\mathbf{e}, \mathbf{x}) = D(\mathbf{x} + \mathbf{e}, \mathbf{x})$ of the distortion signal $\mathbf{e}$. For example, square-error distortion (2) equals the power of $\mathbf{e}$ since $\|\mathbf{s} - \mathbf{x}\|^2 = \|\mathbf{e}\|^2$.

The composite signal $\mathbf{s}$ is subjected to various common signal processing manipulations such as lossy compression, addition of random noise, and resampling, as well as deliberate attempts to remove the embedded information. These manipulations occur in some channel, which produces an output signal $\mathbf{y} \in \Re^N$ according to some conditional probability density function (pdf) $p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s})$. (In this paper we may use the alternative notation $p(\mathbf{y}|\mathbf{s})$ when there is no risk of confusion between random variables and sample values.) For convenience, we define a perturbation vector $\mathbf{n} \in \Re^N$ to be the difference $\mathbf{y} - \mathbf{s}$. Also, in this paper we consider memoryless channels, which have pdfs of the form

$$p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s}) = \prod_{i=1}^{N} p_{y|s}(y_i|s_i),$$

where y_i and s_i are the i-th components of $\mathbf{y}$ and $\mathbf{s}$, respectively. Generalization of the results in this paper to blockwise, memoryless channels, where the y_i and s_i are subvectors rather than scalars, are straightforward,[3,4] and in previous work we have considered deterministic channels.[5] Similarly, although in this paper we confine our attention to the case of independent and identically distributed (iid) host signals, where $p_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{N} p_x(x_i)$, generalization of our results to the case of blockwise iid host signals is also straightforward.[3,4]

The decoder forms an estimate $\hat{m}$ of the embedded information m based on the channel output $\mathbf{y}$. The robustness of the overall embedding-decoding method is characterized by the maximum noisiness of the channel, as measured by the variance σ_n^2 of the perturbations for example, for which one can decode the watermark with small probability of error. (More generally, as is done in other related work,[3–5] one can characterize the robustness in terms of the class of perturbation vectors over which the estimate $\hat{m}$ is reliable, either in a probabilistic sense or in a deterministic sense.) Specific channels of interest in this paper include (1) additive Gaussian noise channels and (2) arbitrary, square-error distortion-constrained attack channels, where the attacker can choose any channel law $p_{y|s}(y|s)$ subject to the constraint $E[(y - s)^2] \leq \sigma_n^2$.

One desires the embedding system to have high rate, low distortion, and high robustness, but in general these three goals tend to conflict. Thus, the performance of an information embedding system is characterized in terms of its achievable rate-distortion-robustness trade-offs.

3. DISTORTION-COMPENSATED QUANTIZATION INDEX MODULATION

In this section we review quantization index modulation (QIM)[3–5] and introduce a type of postprocessing called distortion compensation. We focus our discussion on a particular implementation of QIM called spread-transform dither modulation (STDM) with uniform, scalar quantization.

"Spread-transform" refers to first transforming the host signal vector $\mathbf{x}$ by projecting it onto a collection of orthogonal (usually pseudorandom) projection vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{N/L_{\text{STDM}}}$ to obtain a set of transformed host signal

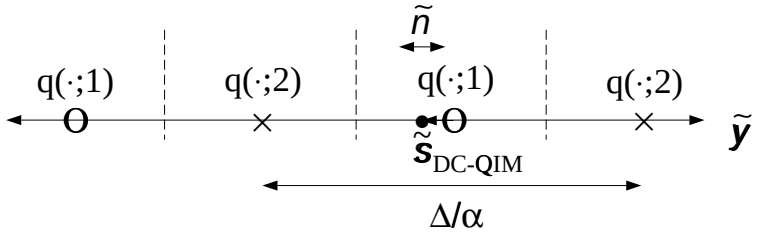

Figure 2. Decoder decision regions for distortion-compensated spread-transform dither modulation. Host signal projections are quantized using one of two quantizers, and some of the quantization error is added back (compensated).

components $\tilde{x}_1, \ldots, \tilde{x}_{N/L_{\text{STDM}}}$. When embedding information, we modify only these N/L_{STDM} coefficients to obtain transformed composite signal components $\tilde{s}_1, \ldots, \tilde{s}_{N/L_{\text{STDM}}}$ and set

$$\mathbf{s} = \mathbf{x}^{\perp} + \sum_{i=1}^{N/L} \tilde{s}_i \mathbf{v}_i,$$

where $\mathbf{x}^{\perp}$ is the component of $\mathbf{x}$ orthogonal to the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{N/L_{\text{STDM}}}$. We call L_{STDM} the spreading length.

QIM embedding methods embed information in the spread-transformed host signal components by quantizing them with a quantizer chosen from an ensemble of quantizers. The watermark m determines the choice of quantizer. For example, if one wishes embed one bit ($m = 1$ or $m = 2$) in one host signal component $\tilde{x}$, then $\tilde{s}_{\text{QIM}} = q(\tilde{x}; m)$, where $q(\cdot; 1)$ and $q(\cdot; 2)$ are two different quantizers. In Fig. 2 these quantizers are uniform, scalar quantizers with step size Δ/α. In this case both the reconstruction points, which are shown as $\circ$ points and $\times$ points in Fig. 2, and the quantization cells, which are not shown in Fig. 2, of the two quantizers are shifted versions of each other so the quantizers are dithered quantizers, and we refer to this type of QIM as dither modulation.[3–5] In general QIM quantizer ensembles need not be dithered quantizers, scalar quantizers, nor uniform quantizers, and the host signal need not be spread-transformed first.[3–5]

One way to decode QIM embedded data is to use minimum distance decoding, which corresponds to finding the quantizer reconstruction point that is closest in distance to the channel output and setting $\hat{m} = i$ if this reconstruction point belongs to the i-th quantizer in the ensemble. In the case discussed above, where one bit is embedded in one host signal sample, the decision region boundaries of the minimum distance decoder are shown in Fig. 2 as dashed lines. If $\tilde{y}$ falls within a decision region containing a $\circ$ point, then $\hat{m} = 1$. Otherwise, $\hat{m} = 2$. Therefore, the distance between the reconstruction points and their respective decision region boundaries determines the amount of interference that can be tolerated before decoding errors occur. Significantly, and in contrast to additive spread-spectrum methods, the host signal $\tilde{x}$ does not interfere with decoding since $\tilde{x}$ may affect which $\circ$ point ($m = 1$), or which $\times$ point ($m = 2$) is selected during encoding, but $\tilde{x}$ does not cause $\tilde{y}$ to move from a $\circ$ point towards a $\times$ point or vice-versa. Thus, we say that QIM methods are host-interference rejecting methods. (This host-interference rejection property leads to a so-called "signal-to-noise ratio (SNR) advantage", as we discuss in other work.[3,4,11])

One can increase this distance by decreasing α. However, since Δ/α is the quantizer step size, decreasing α also increases the quantization error, which is the QIM embedding-induced distortion. Therefore, to keep the embedding-induced distortion fixed as one decreases α, one must compensate for this additional quantization error. One way to do so, which we call distortion-compensation, is to add part of the quantization error to the reconstruction point to form the composite signal. Specifically, if the embedding function is

$$\tilde{s}_{\text{DC}-\text{QIM}} = q(\tilde{x}; m, \Delta/\alpha) + (1 - \alpha)[\tilde{x} - q(\tilde{x}; m, \Delta/\alpha)], \tag{3}$$

where $q(\cdot; m, \Delta/\alpha)$ denotes the m-th quantizer with step size Δ/α, then the square-error embedding-induced distortion (2) is independent of α for α between 0 and 1.[3,4] If the quantizers are dithered quantizers, as they are in Fig. 2, then we call this type of information embedding distortion-compensated dither modulation. The generalization of these methods, where the quantizers need not be dithered quantizers, is called distortion-compensated QIM.[3,4] The resulting composite signal is shown in Fig. 2, where the deflection from the quantizer reconstruction point to the

composite signal point $\tilde{\mathsf{s}}$ is due to the second term in (3). Thus, this deflection is a source of interference, which we refer to as distortion-compensation interference, during decoding, along with the channel perturbation interference $\tilde{n}$. Since the embedding-induced distortion and embedding rate in (3) are independent of α, we choose α to maximize the robustness. For example, one optimization criterion is to choose α to maximize a SNR at the decision device,

$$\mathrm{SNR}(\alpha) = \frac{d_1^2/\alpha^2}{(1-\alpha)^2 \frac{D_{\mathsf{s}}}{\alpha^2} + \sigma_n^2} = \frac{d_1^2}{(1-\alpha)^2 D_{\mathsf{s}} + \alpha^2 \sigma_n^2},$$

where, in the case of Fig. 2, this SNR is defined as the ratio between the squared length of the decoder decision regions and the total interference energy from both distortion-compensation interference and channel interference. Here, $d_1 = \Delta/2$ is the decoder decision region length when $\alpha = 1$ (no distortion compensation). It is straightforward to verify that the optimal scaling parameter α that maximizes this SNR is

$$\alpha_{\mathrm{SNR}} = \frac{\mathrm{DNR}}{\mathrm{DNR}+1}, \tag{4}$$

where DNR is the (embedding-induced) distortion-to-noise ratio $D_{\mathsf{s}}/\sigma_n^2$. As discussed in Sec. 4, such a choice of α also maximizes the information-embedding capacity in the case when the host signal $\mathbf{x}$ is Gaussian and the channel is an additive Gaussian noise channel and in high-fidelity cases even if the host signal is non-Gaussian and the channel represents arbitrary attacks.

4. INFORMATION-THEORETIC PERSPECTIVES

In this section we consider from an information theoretic perspective the best possible rate-distortion-robustness performance that one could hope to achieve with any information embedding system. Our analysis leads to insights about some properties and characteristics of good information embedding methods, *i.e.*, methods that achieve performance close to the information-theoretic limits. In particular, a canonical structure emerges for information embedding that consists of (1) preprocessing of the host signal, (2) QIM embedding, and (3) postprocessing of the quantized host signal to form the composite signal. One incurs no loss of optimality by restricting one's attention to this simple structure. Also, only distortion compensation postprocessing is required in the following three cases: (1) an additive Gaussian noise channel and a Gaussian host signal, (2) square-error distortion-constrained attacks and a Gaussian host signal, and (3) square-error distortion-constrained attacks, a non-Gaussian host signal, and asymptotically small embedding-induced distortion D_{s} and attacker's distortion σ_n^2 (*i.e.*, high fidelity case). We emphasize that in this section we consider QIM and distortion-compensated QIM in their most general senses, where the quantizers are not necessarily dithered quantizers, uniform quantizers, or scalar quantizers. In fact, capacity-achieving performance is generally achievable only asymptotically with long signal lengths N.

The bottom model of Fig. 1, the one which models information embedding problems as communication with state information known at the encoder, is the most convenient one for this information-theoretic analysis. In non-watermarking contexts Gel'fand and Pinsker[12] have determined the capacity of such a channel in the case when the encoder sees the entire iid state vector $\mathbf{x}$ before choosing the channel input $\mathbf{e}$. In particular, the capacity is

$$C = \max_{p_{u,\mathsf{e}|\mathsf{x}}(u,e|x)} I(u;y) - I(u;x), \tag{5}$$

where $I(\cdot;\cdot)$ denotes mutual information and u is an auxiliary random variable. In the case of watermarking, the maximization (5) is subject to a distortion constraint $E[\mathsf{e}^2] \leq D_{\mathsf{s}}$.

4.1. Optimality of preprocessed and postprocessed QIM

In this section we show that one can achieve the capacity (5) by a type of "hidden" QIM, *i.e.*, QIM that occurs in a domain represented by the auxiliary random variable u. One moves into and out of this domain with pre- and post-quantization processing.

Our discussion here is basically a summary of Gel'fand and Pinsker's capacity-achievability proof,[12] with added interpretation in terms of quantization (source coding). Fig. 3 shows an ensemble of 2^{NR_m} quantizers, where $R_m = I(u;y) - I(u;x) - 2\epsilon$, where each source codeword (quantizer reconstruction vector) $\mathbf{u}$ is randomly drawn from the iid distribution $p_u(u)$, which is the marginal distribution corresponding to the host signal distribution $p_x(x)$ and the

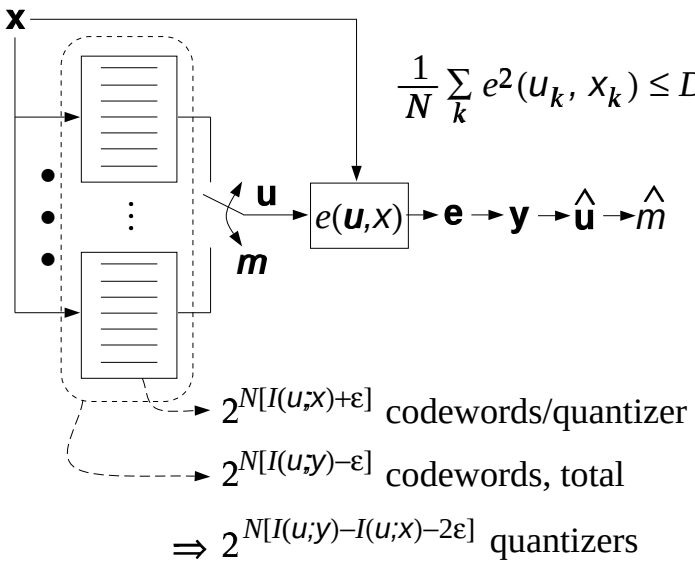

Figure 3. "Hidden" Quantization Index Modulation. QIM in the "**u**-domain" can achieve capacity. Preprocessing and postprocessing is used to move into and out of this domain.

maximizing conditional distribution $p_{u,e|x}(u,e|x)$ from (5). (Although the source codebooks are therefore random, both the encoder and decoder, of course, know the codebooks.) Each codebook contains $2^{N[I(u;x)+\epsilon]}$ codewords so there are $2^{N[I(u;y)-\epsilon]}$ codewords total.

QIM embedding in this **u**-domain corresponds to finding a vector $\mathbf{u}_0$ in the m-th quantizer's codebook that is jointly distortion-typical with $\mathbf{x}$ and generating $\mathbf{e}(\mathbf{u}_0, \mathbf{x}) = [e(u_{0,1}, x_1) \cdots e(u_{0,N}, x_N)]^T$. (From convexity properties of mutual information, one can deduce that the maximizing distribution in (5) always has the property that e is a deterministic function of (u, x).[12]) By distortion-typical, we mean that $\mathbf{u}_0$ and $\mathbf{x}$ are jointly typical and $\|\mathbf{e}(\mathbf{u}_0, \mathbf{x})\|^2 \le N(D_s + \epsilon)$, *i.e.*, the function $e^2(u, x)$ is the distortion function in the **u**-domain. Since the m-th quantizer's codebook contains more than $2^{NI(u;x)}$ codewords, the probability that there is no $\mathbf{u}_0$ that is jointly distortion-typical with $\mathbf{x}$ is small. (This is one of the main ideas behind the rate-distortion theorem.[13]) Thus, the selection of a codeword from the m-th quantizer is the quantization part of QIM, and the generation of $\mathbf{e}$, and therefore $\mathbf{s} = \mathbf{x} + \mathbf{e}$, from the codeword $\mathbf{u}_0$ and $\mathbf{x}$ is the post-quantization processing.

The decoder finds a $\mathbf{u}$ that is jointly typical with the channel output $\mathbf{y}$ and declares $\hat{m} = i$ if this $\mathbf{u}$ is in the i-th quantizer's codebook. Because the total number of codewords $\mathbf{u}$ is less than $2^{NI(u;y)}$, the probability that a $\mathbf{u}$ other than $\mathbf{u}_0$ is jointly typical with $\mathbf{y}$ is small. Also, the probability that $\mathbf{y}$ is jointly typical with $\mathbf{u}_0$ is close to 1. (These are two of the main ideas behind the classical channel coding theorem.[13]) Thus, the probability of error $\Pr[\hat{m} \neq m]$ is small, and we can indeed achieve the capacity (5) with QIM in the **u**-domain.

The remaining challenge, therefore, is to determine the right preprocessing and postprocessing given a particular channel (attack) $p_{y|s}(y|s)$. As mentioned above, for a number of important cases, it turns out that the only processing required is post-quantization distortion-compensation. We discuss these cases in the next section.

4.2. Optimality of distortion-compensated QIM

In this section we show that distortion-compensated QIM (DC-QIM) can achieve capacity whenever the maximizing distribution $p_{u,e|x}(u,e|x)$ in (5) is of a form such that

$$u = e + \alpha x. \tag{6}$$

This condition is satisfied in at least three important cases: (1) the case of a Gaussian host signal and an additive Gaussian noise channel[3,4]; (2) the case of a Gaussian host signal and arbitrary† square-error distortion-constrained

†In each of the arbitrary attack cases considered in this section, we assume that the attacker knows the codebook distribution $p_u(u)$, but not the codebook. Since both the encoder and decoder do know the codebook, these cases are private-key scenarios.

attacks[14]; and (3) the case of arbitrary square-error distortion-constrained attacks, a zero-mean, finite variance host signal with bounded and continuous pdf, and asymptotically small embedding-induced distortion D_s and attacker's distortion σ_n^2.[14] The values of α, which are the information-theoretically optimal distortion compensation parameters in (3), in these three cases are, respectively,

$$\begin{aligned} \alpha_1 &= \frac{\mathrm{DNR}}{\mathrm{DNR}+1}, \\ \alpha_2 &= \frac{\mathrm{DNR}}{\mathrm{DNR}+\beta}, \quad \beta = \frac{\mathrm{SNR}_x + \mathrm{DNR}}{\mathrm{SNR}_x + \mathrm{DNR} - 1}, \\ \alpha_3 &= \frac{\mathrm{DNR}}{\mathrm{DNR}+1} \end{aligned}$$

where $\mathrm{SNR}_x = \sigma_x^2/\sigma_n^2$ is the ratio between the host signal variance and the attacker's distortion and, thus, is a kind of host "signal-to-noise ratio". We see that the SNR-maximizing α in (4) is also capacity-achieving in the first and third cases.

To see that DC-QIM can achieve capacity when the maximizing pdf in (5) satisfies (6), we show that one can construct an ensemble of random DC-QIM codebooks that satisfy (6). We begin by writing the generalized version of (3):

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{q}(\mathbf{x}; m, \Delta/\alpha) + (1-\alpha)[\mathbf{x} - \mathbf{q}(\mathbf{x}; m, \Delta/\alpha)], \tag{7}$$

where $\mathbf{q}(\cdot; m, \Delta/\alpha)$ is the m-th quantizer, which is possibly a vector, non-uniform, or non-dithered quantizer, in the QIM ensemble. The parameter Δ/α is some measure of scale, for example, the distance between the first and second reconstruction points, and thus represents a generalization of its meaning (quantizer step size) in Sec. 3 to reflect the possibly vector, non-uniform nature of the quantizers. Next, we observe that quantizing $\mathbf{x}$ is equivalent to quantizing $\alpha\mathbf{x}$ with a scaled version of the quantizer and scaling back, *i.e.*,

$$\mathbf{q}(\mathbf{x}; m, \Delta/\alpha) = \frac{1}{\alpha}\mathbf{q}(\alpha\mathbf{x}; m, \Delta). \tag{8}$$

This identity simply represents a change of units to "units of $1/\alpha$" before quantization followed by a change back to "normal" units after quantization. For example, if $\alpha = 1/1000$, instead of quantizing $\mathbf{x}$ volts we quantize $\alpha\mathbf{x}$ kilovolts (using the same quantizer, but relabeling the reconstruction points in kilovolts) and convert kilovolts back to volts by multiplying by $1/\alpha$. Then, rearranging terms in (7) and substituting (8) into the result, we obtain

$$\begin{aligned} \mathbf{s}(\mathbf{x}, m) &= \mathbf{q}(\mathbf{x}; m, \Delta/\alpha) + (1-\alpha)[\mathbf{x} - \mathbf{q}(\mathbf{x}; m, \Delta/\alpha)] \\ &= \alpha\mathbf{q}(\mathbf{x}; m, \Delta/\alpha) + (1-\alpha)\mathbf{x} \\ &= \mathbf{q}(\alpha\mathbf{x}; m, \Delta) + (1-\alpha)\mathbf{x}. \end{aligned} \tag{9}$$

We construct our random DC-QIM codebooks by choosing the codewords of $\mathbf{q}(\cdot; m, \Delta)$ from the iid distribution $p_u(u)$, the one corresponding to (6). (Equivalently, we choose the codewords of $\mathbf{q}(\cdot; m, \Delta/\alpha)$ in (7) from the distribution of u/α, *i.e.*, the iid distribution $\alpha p_u(\alpha u)$.) Our quantizers $\mathbf{q}(\cdot; m, \Delta)$ choose a codeword $\mathbf{u}_0$ that is jointly distortion-typical with $\alpha\mathbf{x}$. The decoder looks for a codeword in all of the codebooks that is jointly typical with the channel output. Then, following the achievability argument of Sec. 4.1, we can achieve a rate $I(u;y) - I(u;x)$. From (9), we see that

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + [\mathbf{q}(\alpha\mathbf{x}; m, \Delta) - \alpha\mathbf{x}] = \mathbf{x} + (\mathbf{u}_0 - \alpha\mathbf{x}).$$

Since $\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{e}$, we see that $\mathbf{e} = \mathbf{u}_0 - \alpha\mathbf{x}$. Thus, if the maximizing distribution in (5) satisfies (6), our DC-QIM codebooks can also have this distribution and, hence, achieve capacity (5).

5. GAUSSIAN CASE

We now focus our attention on the case of a Gaussian host signal and an additive Gaussian noise channel. When both the host signal and channel noise are white, the capacity (5) is[15]

$$C_{\mathrm{Gauss}} = \frac{1}{2}\log_2(1 + \mathrm{DNR}), \tag{10}$$

No-key scenarios have been considered using alternative analysis methods.[3,5]

Host Signal	Bandwidth	Capacity
Analog FM	200 kHz	66.4 kb/s/dB
Analog AM	30 kHz	10.0 kb/s/dB
Audio	20 kHz	6.6 kb/s/dB
Telephone voice	3 kHz	1.0 kb/s/dB

Table 1. Information-embedding capacities for transmission over additive Gaussian noise channels for various types of host signals. Capacities are in terms of achievable embedded rate per dB drop in received host signal quality.

In the case of blockwise-iid host signals and channel noise, where we allow some correlation between host signal samples within blocks and between noise samples within blocks, the expression (10) also gives the capacity,[4] except in this case the embedding-induced distortion measure is a weighted square-error distortion measure, where the weights are chosen to force the embedding distortion signal $\mathbf{e}$ to be "hidden" by the channel noise $\mathbf{n}$. In particular, very little distortion is allowed in components where the channel noise is small and relatively more distortion is allowed in components where the channel noise is large.

5.1. Capacities of multimedia host signals

Because the capacity expression (10) applies to arbitrary host and noise covariance matrices, it is quite relevant to many multimedia information embedding applications, especially those where one faces unintentional attacks. For example, these capacities do not depend on the power spectrum of the host signal and thus these results apply to audio, video, image, speech, analog FM, analog AM, and coded digital signals, to the extent that these signals can be modeled as Gaussian. Also, the additive Gaussian noise with arbitrary covariance model may be applicable to lossy compression, printing and scanning noise, and transmission noise. Furthermore, when considering the amount of embedding-induced distortion, in many applications one is most concerned with the quality of the *received* host signal, *i.e.*, the channel output, rather than the quality of the composite signal. For example, in many authentication applications, the document carrying the authentication signal may be transmitted across some channel to the intended user. In these cases one can conveniently express the achievable embedded rate per unit of host signal bandwidth and per unit of received host signal degradation, as we show in this section.

One can view the DNR as the amount by which one would have to amplify the noise to create a noise signal with the same statistics as the embedding-induced distortion signal. Thus, if one views the received channel output as a noise-corrupted version of the host signal, then the effect of the embedding is to create an additional noise source DNR times as strong as the channel noise, and therefore, the received signal quality drops by a factor of $(1 + \text{DNR})$ or

$$10 \log_{10}(1 + \text{DNR}) \text{ dB}. \tag{11}$$

Since the capacity in bits per host signal sample is given by (10), and there are two independent host signal samples per second for every Hertz of host signal bandwidth, the capacity in bits per second per Hertz is

$$C = \log_2(1 + \text{DNR}) \text{ b/s/Hz}. \tag{12}$$

Taking the ratio between (12) and (11), we see that the "value" in embedded rate of each dB drop in received host signal quality is

$$C = \frac{\log_2(1 + \text{DNR})}{10 \log_{10}(1 + \text{DNR})} = \frac{1}{10} \log_2 10 \approx 0.3322 \text{ b/s/Hz/dB} \tag{13}$$

Thus, the available embedded digital rate in bits per second depends only on the bandwidth of the host signal and the tolerable degradation in received host signal quality. Information-embedding capacities for several types of host signals are shown in Table 1.

5.2. Gaps to capacity

Rather remarkably, the capacity (10) is independent of host signal statistics (For example, σ_x^2 does not appear in (10).), implying that an infinite energy host signal causes no decrease in capacity in the Gaussian case and that one can do just as well when the host signal is not known at the decoder as when the host signal is known at the decoder. This principle suggests that optimal and near-optimal digital watermarking methods have the type of host-signal interference rejection capability mentioned earlier in Sec. 3. Because QIM methods (even without distortion

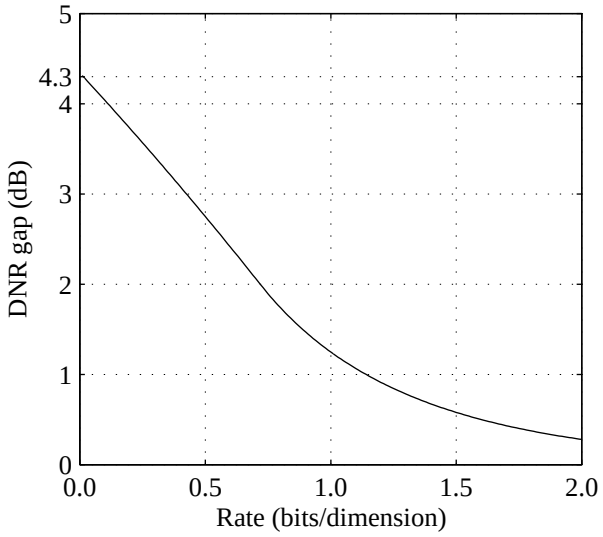

Figure 4. Gap between spread-transform QIM and Gaussian capacity. The spreading length is restricted to be greater than or equal to 1. The maximum gap is a factor of e, or about 4.3 dB.

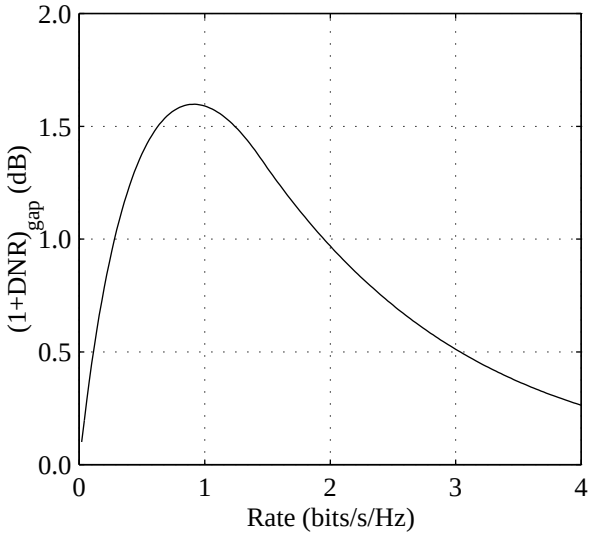

Figure 5. Gap in (1 + DNR) between spread-transform QIM and Gaussian capacity. The spreading length is restricted to be greater than or equal to 1. One bit/dimension equals 2 b/s/Hz.

compensation) possess such a capability, one can show[3] that there exist near capacity-achieving QIM methods, as illustrated in Fig. 4, which shows an upper bound on the "gap" between QIM and capacity.[3] This gap, which arises when $\alpha = 1$ instead of its optimal value (4), is the additional amount of DNR that the best possible QIM system needs to achieve the same rate as a capacity-achieving system. This gap is at most a factor of $e \approx 4.3$ dB at any finite rate and asymptotically approaches 0 dB at high rates. The gap in terms of $(1 + \text{DNR})$, which as we discussed in Sec. 5.1 is more relevant when one cares about received host signal quality, is shown in Fig. 5. Of course, even at all finite rates both of these gaps can be eliminated with distortion compensation, as discussed in Sec. 4.

In contrast, spread spectrum methods do not reject host signal interference. Thus, the achievable rate of a spread spectrum method is the Gaussian channel capacity, treating both x and n as interference sources. As is well-known,[3,16] when both x and n are white, this capacity is

$$C_{\text{ss}} = \frac{1}{2}\log_2\left(1 + \frac{D_{\text{s}}}{\sigma_x^2 + \sigma_n^2}\right) = \frac{1}{2}\log_2\left(1 + \frac{\text{DNR}}{\text{SNR}_x + 1}\right), \tag{14}$$

where $\text{SNR}_x = \sigma_x^2/\sigma_n^2$ is the host signal-to-noise ratio. (This rate is also the capacity when n is non-Gaussian, but still independent of s, and a correlation detector is used for decoding.[17]) By comparing (14) to (10) we see that the gap (in terms of DNR) to capacity of spread-spectrum is $\text{SNR}_x + 1$. Typically, SNR_x is very large since the channel noise is not supposed to degrade signal quality too much. Thus, in these cases the gap to capacity of spread-spectrum is much larger than the gap to capacity of QIM.

6. SIMULATION RESULTS

Having established the existence of capacity-achieving and near capacity-achieving embedding and decoding methods within the distortion-compensated QIM and regular QIM classes, respectively, we now present simulation results demonstrating practically achievable performance for low-complexity dither modulation implementations employing uniform, scalar quantization and practical error correction codes.

It can be shown fairly easily[3] that for additive white Gaussian noise (AWGN) channels and $R_m < 1$, the bit-error probability P_b of *uncoded* spread-transform dither modulation (STDM) with uniform, scalar quantization is upper bounded by

$$P_b \leq 2Q\left(\sqrt{\frac{3}{4}\text{DNR}_{\text{norm}}}\right), \tag{15}$$

where DNR_{norm} is the rate-normalized distortion-to-noise ratio

$$\text{DNR}_{\text{norm}} \triangleq \frac{\text{DNR}}{R_m}. \tag{16}$$

For example, one can achieve a bit-error probability of about 10^{-6} at a DNR_{norm} of 15 dB. Thus, no matter how noisy the AWGN channel, one can reliably embed using uncoded STDM by choosing sufficiently low rates. In particular, one needs to choose a rate satisfying

$$R_m \leq \frac{\text{DNR}}{\text{DNR}_{\text{norm}}},$$

where DNR_{norm} is the minimum DNR_{norm} necessary in (15) for a given P_b and DNR is determined by channel conditions and the embedding-induced distortion.

One can improve performance significantly using error correction coding and distortion compensation. In fact, from the capacity expression (10) for the case of white, Gaussian noise, we see that reliable information embedding is possible if

$$R_m \leq C = \frac{1}{2}\log_2(1 + \text{DNR})$$

or, equivalently,

$$\frac{\text{DNR}}{2^{2R_m} - 1} \geq 1.$$

For small R_m, $2^{2R_m} - 1 \approx 2R_m \ln 2$, so this condition becomes

$$\text{DNR}_{\text{norm}} \geq 2\ln 2 \approx 1.4 \text{ dB}.$$

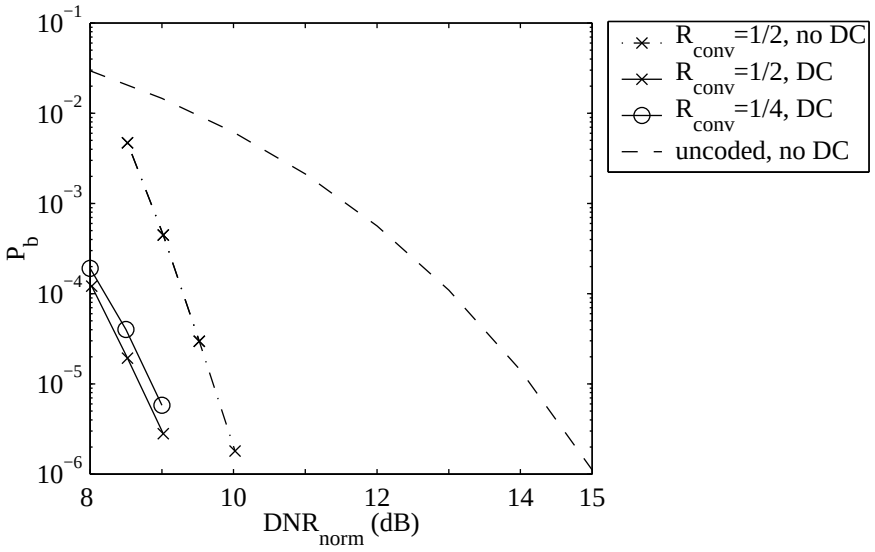

Figure 6. Error-correction coding and distortion-compensation (DC) gains. With common, memory-8 convolutional codes one can obtain gains of about 5 dB over uncoded STDM. Distortion compensation yields about 1 dB additional gain.

Since, as stated above, uncoded STDM with uniform, scalar quantization requires a $\mathrm{DNR}_{\mathrm{norm}}$ of 15 dB for a bit-error probability of 10^{-6}, there is a gap to capacity of about 13.6 dB.

We now report the results of one experiment designed to investigate how much of this gap can be closed with practical error correction codes and distortion compensation. In our experiment we embedded 10^7 bits in a pseudo-random white Gaussian host using memory-8, rate-1/2 and rate-1/4, convolutional codes with maximal free distance. The generators of these two codes in octal notation are $(561, 753)$ and $(463, 535, 733, 745)$, respectively.[18] One coded bit was embedded in each spread-transformed host signal component using uniform, scalar quantizers as described in Sec. 3 and Fig. 2. We used the squared Euclidean distances between the channel output samples $\tilde{y}$ and the nearest reconstruction point from each of the two quantizers to calculate branch metrics for Viterbi decoding[19] of the convolutionally encoded data. Experimentally measured bit-error rate (BER) curves are plotted in Fig. 6. We observe an error correction coding gain of about 5 dB at a BER of 10^{-6}. Distortion compensation provides an additional 1-dB gain.

From the definition of $\mathrm{DNR}_{\mathrm{norm}}$ (16), we see a gain factor of g in $\mathrm{DNR}_{\mathrm{norm}}$ translates directly into

1. a factor of g increase in rate for fixed levels of embedding-induced distortion and channel noise (robustness), or
2. a factor of g reduction in distortion for a fixed rate and robustness, or
3. a factor of g increase in robustness for a fixed rate and distortion.

Thus, the minimum $\mathrm{DNR}_{\mathrm{norm}}$ required for a given bit-error rate is, indeed, the fundamental parameter of interest and, as one can see from (15), in the Gaussian case the $\mathrm{DNR}_{\mathrm{norm}}$ also completely determines the bit-error probability for uncoded STDM for $R_m \leq 1$.

Other simulation results, including sample images, for both Gaussian channels and JPEG compression channels are reported elsewhere.[4,11]

ACKNOWLEDGMENTS

This work has been supported in part by the Air Force Office of Scientific Research under Grant No. F49620-96-1-0072, by the MIT Lincoln Laboratory Advanced Concepts Committee, and by a National Defense Science and Engineering Graduate Fellowship.

REFERENCES

1. M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. of the IEEE* **86**, pp. 1064–1087, June 1998.
2. F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. of the IEEE* **87**, pp. 1079–1107, July 1999.
3. B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," submitted to *IEEE Trans. on Information Theory*, 1999.
4. B. Chen and G. W. Wornell, "Quantization index modulation methods for digital watermarking and information embedding," to appear in *Journ. of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 2000.
5. B. Chen and G. W. Wornell, "Dither modulation: A new approach to digital watermarking and information embedding," in *Proc. of SPIE: Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 342–353, (San Jose, CA), Jan. 1999.
6. A. Z. Tirkel, G. A. Rankin, R. van Schyndel, W. J. Ho, N. R. A. Mee, and C. F. Osborne, "Electronic water mark," in *Proc. of Digital Image Computing, Technology and Applications*, pp. 666–672, (Sydney, Australia), Dec. 1993.
7. R. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proc. of the IEEE International Conference on Image Processing*, vol. 2, pp. 86–90, (Austin, TX), Nov. 1994.
8. I. J. Cox, J. Killian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. on Image Processing* **6**, pp. 1673–1687, Dec. 1997.
9. M. D. Swanson, B. Zhu, and A. H. Tewfik, "Data hiding for video-in-video," in *Proc. of the 1997 IEEE International Conference on Image Processing*, vol. 2, pp. 676–679, (Piscataway, NJ), 1997.
10. I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. of the IEEE* **87**, pp. 1127–1141, July 1999.
11. B. Chen and G. W. Wornell, "Provably robust digital watermarking," in *Proc. of SPIE: Multimedia Systems and Applications II*, vol. 3845, pp. 43–54, (Boston, MA), Sept. 1999.
12. S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory* **9**(1), pp. 19–31, 1980.
13. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.
14. P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *Preprint*, 1999.
15. M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory* **29**, pp. 439–441, May 1983.
16. J. R. Smith and B. O. Comiskey, "Modulation and information hiding in images," in *Information Hiding. First International Workshop Proceedings*, pp. 207–226, June 1996.
17. A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. on Information Theory* **42**, pp. 1520–1529, Sept. 1996.
18. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
19. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publishers, second ed., 1994.