[From *Proc. of SPIE: Security and Watermarking of Multimedia Contents*, vol. 3657.]

Dither modulation: a new approach to digital watermarking and information embedding

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

We consider the problem of embedding one signal (e.g., a digital watermark), within another "host" signal to form a third, "composite" signal. The embedding must be done in such a way that minimizes distortion between the host signal and composite signal, maximizes the information-embedding rate, and maximizes the robustness of the embedding. In general, these three goals are conflicting, and the embedding process must be designed to efficiently trade-off the three quantities.

We propose a new class of embedding methods, which we term quantization index modulation (QIM), and develop a convenient realization of a QIM system that we call dither modulation in which the embedded information modulates a dither signal and the host signal is quantized with an associated dithered quantizer. QIM and dither modulation systems have considerable performance advantages over previously proposed spread-spectrum and low-bit(s) modulation systems in terms of the achievable performance trade-offs among distortion, rate, and robustness of the embedding. We also demonstrate these performance advantages in the context of "no-key" digital watermarking applications, in which attackers can access watermarks in the clear.

We also examine the fundamental limits of digital watermarking from an information theoretic perspective and discuss the achievable limits of QIM and alternative systems.

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

1. INTRODUCTION

A variety of related applications have emerged recently[1] that require the design of systems for embedding one signal, sometimes called an "embedded signal" or "watermark", within another signal, called a "host signal". These applications include copyright notification and enforcement, authentication, and transmission of auxiliary information. Digital "fingerprinting" and enforcement of copy-once features in digital video disc recorders[2] are two commonly cited copyright enforcement applications, for example. In each of the proposed applications, the embedding must be done such that the embedded signal causes no serious degradation to its host. At the same time, the host always carries the embedded signal, which can only be removed by causing significant damage to the host.

Various information-embedding algorithms have been proposed[1] in this still emerging field. Some of the earliest proposed systems[3,4] employ a quantize-and-replace strategy: after first quantizing the host signal, these systems change the quantization value to embed information. A simple example of such a system is so-called low-bit(s) modulation (LBM), where the least significant bit(s) in the quantization of the host signal are replaced by a binary representation of the embedded signal. Recently, spread-spectrum based systems, which embed information by adding to the host signal a small pseudo-noise signal that is modulated by the embedded signal, have received considerable attention in the literature. (Several references are provided, for example, by Swanson, et al.[1]) However, as we demonstrate in this paper, spread-spectrum based systems offer relatively little robustness to noise when the host signal is not known at the decoder. Intuitively, when the host signal is not known at the decoder, as is typical in many applications of interest, it is a source of noise. With a spread-spectrum system, the host signal is an additive

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`

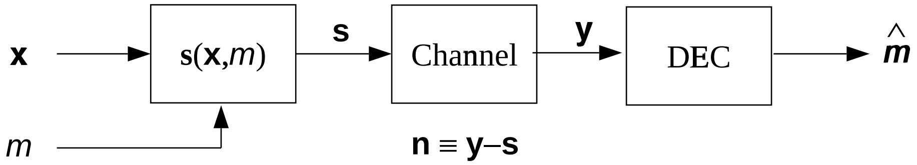

Figure 1. General information-embedding problem model. An integer message m is embedded in the host signal vector $\mathbf{x}$ using some embedding function $\mathbf{s}(\mathbf{x}, m)$. A perturbation vector $\mathbf{n}$ corrupts the composite signal $\mathbf{s}$. The decoder extracts an estimate $\hat{m}$ of m from the noisy channel output $\mathbf{y}$.

noise that is often much larger, due to distortion constraints, than the pseudo-noise signal carrying the embedded information.

While a number of embedding strategies have been proposed in the literature, much work remains to characterize the inherent trade-offs among the robustness of the embedding, the degradation to the host signal caused by the embedding, and the amount of data embedded. In this paper we introduce a framework for characterizing these trade-offs and develop a new family of information-embedding techniques based on ensembles of quantizers that perform these trade-offs efficiently. We refer to this family of techniques as "quantization index modulation" (QIM),[5] and we also explore a special member of this family, "dither modulation". As we will show, this new method of embedding information offers significant advantages over previously proposed spread-spectrum and LBM techniques.

In Sec. 2 we formalize the information-embedding problem using a general problem model applicable in many scenarios of interest. This characterization of the problem leads quite naturally to the class of information-embedding systems that we present in Sec. 3, namely quantization index modulation systems. Dither modulation is discussed in Sec. 4, including a demonstration of its performance advantages over spread-spectrum and LBM techniques. In Sec. 5 we show that QIM systems also have attractive performance advantages in the context of no-key digital watermarking since they are robust to in-the-clear attacks. We discuss the limits of digital watermarking over random channels from an information-theoretic perspective in Sec. 6. Finally, some concluding remarks are presented in Sec. 7.

2. PROBLEM MODEL

Although a variety of information-embedding applications exist, many of these can be described by Fig. 1. We have some host signal vector $\mathbf{x} \in \Re^N$ in which we wish to embed some information m. 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

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

An embedding function maps the host signal $\mathbf{x}$ and embedded information m to a composite signal $\mathbf{s} \in \Re^N$ subject to some distortion constraint. For example, one might choose the squared-error distortion constraint

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

The composite signal $\mathbf{s}$ is passed through a channel, where it 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. We let $\mathbf{y} \in \Re^N$ denote the output of the channel and define a noise or perturbation vector to be the difference $\mathbf{n} \triangleq \mathbf{y} - \mathbf{s}$. Thus, this model is sufficiently general to include both random and deterministic perturbation vectors and both signal-independent and signal-dependent perturbation vectors. Specific channels that will be of interest in this paper are:

1. **bounded perturbation channels:** A key requirement in the design of information-embedding systems is that the decoder must be capable of reliably extracting the embedded information *as long as the signal is not severely degraded.* Thus, it is reasonable to assume that the channel output $\mathbf{y}$ is a fair representation of

the original signal. One way to express this concept of "fair representation" is to bound the energy of the perturbation vector,

$$\|\mathbf{y} - \mathbf{s}\|^2 = \|\mathbf{n}\|^2 \leq N\sigma_n^2. \tag{3}$$

This channel model, which describes a maximum distortion* or minimum SNR constraint between the channel input and output, may be an appropriate model for either the effect of a lossy compression algorithm or attempts by an active attacker to remove the embedded signal, for example.

2. **bounded host-distortion channels:** Some attackers may work with distortion constraint between the host signal, rather than the channel input, and the channel output since this distortion is the most direct measure of degradation to the host signal. For example, if an attacker has partial knowledge of the host signal, which may be in the form of a probability distribution, so that he or she can calculate this distortion, then it may be appropriate to bound the expected distortion $E[D(\mathbf{y}, \mathbf{x})]$.

3. **probabilistic channels:** In some contexts it may be convenient to assume some probability distribution for $\mathbf{n}$. Two examples of probabilistic channels are discrete, memoryless channels and Gaussian channels.

The decoder forms an estimate $\hat{m}$ of the embedded information m based on the channel output $\mathbf{y}$. In the case of the bounded perturbation channel, we can characterize the robustness of the system by the maximum allowable σ_n^2 such that we can still guarantee that $\hat{m} = m$. Alternatively, in the case of probabilistic channels, we can characterize the reliability of the system by the probability of message error $\Pr[\hat{m} \neq m]$ or bit-error rate. The problem we face is to design an embedding function $\mathbf{s}(\mathbf{x}, m)$ that achieves the best possible trade-off among the three parameters rate, distortion, and robustness (or reliability).

3. QUANTIZATION INDEX MODULATION

Specifying the performance requirements of an information-embedding system in terms of rate, distortion, and robustness leads quite naturally to the notion of quantizer ensembles and a new method of information embedding, as we develop in this section. In the last section, we consider the embedding function $\mathbf{s}(\mathbf{x}, m)$ to be a function of two variables, the host signal and the embedded information. However, we can also view $\mathbf{s}(\mathbf{x}, m)$ to be a collection or ensemble of functions of $\mathbf{x}$, indexed by m. We denote the functions in this ensemble as $\mathbf{s}(\mathbf{x}; m)$ to emphasize this view. As one can see from (1), the rate R_m determines the number of possible values for m, and hence, the number of functions in the ensemble. The distortion constraint suggests that each function in the ensemble is close to an identity function so that

$$\mathbf{s}(\mathbf{x}; m) \approx \mathbf{x}, \qquad \forall m. \tag{4}$$

That the system needs to be robust to perturbations suggests that the points in the range of one function in the ensemble should be "far away" in some sense from the points in the range of any other function. At the very least, the ranges should be non-intersecting. Otherwise, even in the absence of any perturbations, there will be some values of $\mathbf{s}$ from which one will not be able to uniquely determine m. This property along with (4) suggests that the functions be discontinuous. Quantizers are just such a class of discontinuous, approximate-identity functions. Then, "quantization index modulation (QIM)" refers to embedding information by first modulating an index or sequence of indices with the embedded information and then quantizing the host signal with the associated quantizer or sequence of quantizers.

Fig. 2 illustrates this QIM information-embedding technique. In this example, one bit is to be embedded so that $m \in \{1, 2\}$. Thus, we require two quantizers, and their corresponding sets of reconstruction points in $\Re^N$ are represented in Fig. 2 with ×'s and ∘'s. If $m = 1$, for example, the host signal is quantized with the ×-quantizer, i.e., $\mathbf{s}$ is chosen to be the × closest to $\mathbf{x}$. If $m = 2$, $\mathbf{x}$ is quantized with the ∘-quantizer. Here, we see the non-intersecting nature of the ranges of the two quantizers as no × point is the same as any ∘ point. We also see the discontinuous nature of the quantizers. The dashed polygon represents the quantization cell for the × in its interior. As we move across the cell boundary from its interior to its exterior, the corresponding value of the quantization function jumps from the × in the cell interior to a × in the cell exterior.

*Some types of distortion, such as geometric distortions can be large in terms of squared error, yet still be small perceptually. However, in some cases these distortions can be mitigated either by pre-processing at the decoder or by embedding information in parameters of the host signal that are less affected (in terms of squared error) by these distortions. For example, a simple delay or shift may cause large squared error, but the magnitude of the DFT coefficients are relatively unaffected.

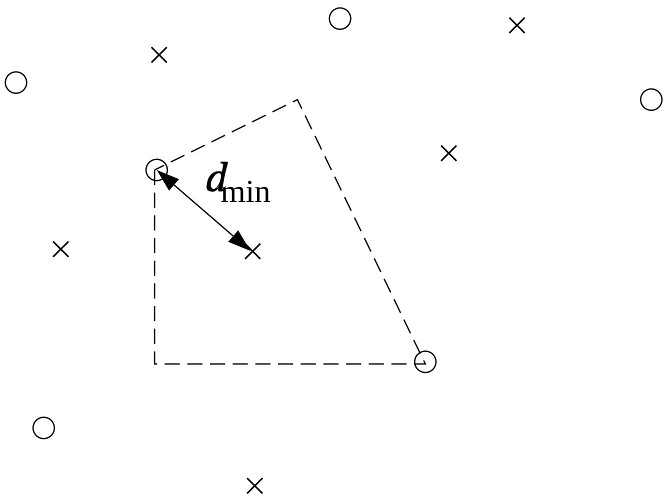

Figure 2. Quantization index modulation for information embedding. The points marked with ×'s and ∘'s belong to two different quantizers, each with its associated index. The minimum distance $d_{\min}$ measures the robustness to perturbations, and the sizes of the quantization cells, one of which is shown in the figure, determine the distortion. If $m = 1$, the host signal is quantized to the nearest ×. If $m = 2$, the host signal is quantized to the nearest ∘.

A few parameters of the quantizer ensemble conveniently characterize the performance of a QIM system. As noted above the number of quantizers in the ensemble determines the information-embedding rate. The size and shape of the quantization cells determine the distortion due to the embedding. Finally, the minimum distance $d_{\min}$ between the sets of reconstruction points of different quantizers in the ensemble determines the robustness of the embedding. We define the minimum distance to be

$$d_{\min} \triangleq \min_{(i,j):i\neq j} \min_{(\mathbf{x}_i,\mathbf{x}_j)} ||\mathbf{s}(\mathbf{x}_i;i) - \mathbf{s}(\mathbf{x}_j;j)||. \tag{5}$$

Intuitively, the minimum distance measures the size of perturbation vectors that can be tolerated by the system. For example, in the case of the bounded perturbation channel, the energy bound of Eq. (3) implies that a minimum distance decoder is guaranteed to not make an error as long as

$$\frac{d_{\min}^2}{4N\sigma_n^2} > 1. \tag{6}$$

In the case of an additive white Gaussian noise channel with a noise variance of σ_n^2, at high signal-to-noise ratio the minimum distance also characterizes the error probability of the minimum distance decoder,[6]

$$\Pr[\hat{m} \neq m] \sim Q\left(\sqrt{\frac{d_{\min}^2}{4\sigma_n^2}}\right).$$

The minimum distance decoder to which we refer simply chooses the reconstruction point closest to the received vector, i.e.,

$$\hat{m}(\mathbf{y}) = \arg\min_m \min_{\mathbf{x}} ||\mathbf{y} - \mathbf{s}(\mathbf{x};m)||. \tag{7}$$

If, which is often the case, the quantizers $\mathbf{s}(\mathbf{x};m)$ map $\mathbf{x}$ to the nearest reconstruction point, then (7) can be rewritten as

$$\hat{m}(\mathbf{y}) = \arg\min_m ||\mathbf{y} - \mathbf{s}(\mathbf{y};m)||. \tag{8}$$

From the preceding discussion, we see that the nonzero minimum distance of QIM systems offers quantifiable robustness to perturbations, even when the host signal is not known at the decoder. In contrast, spread-spectrum based systems that have been proposed recently offer relatively little robustness to perturbations if the host signal is not known at the decoder. These systems embed information by adding a pseudo-noise vector $\mathbf{w}(m)$ to the host signal,

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{w}(m). \tag{9}$$

From the definition of minimum distance (5),

$$\begin{aligned} d_{\min} &= \min_{(i,j):i\neq j} \min_{(\mathbf{x}_i,\mathbf{x}_j)} \|\mathbf{x}_i + \mathbf{w}(i) - \mathbf{x}_j - \mathbf{w}(j)\| \\ &= \min_{(i,j):i\neq j} \|\mathbf{x}_i + \mathbf{w}(i) - (\mathbf{x}_i + \mathbf{w}(i) - \mathbf{w}(j)) - \mathbf{w}(j)\| \\ &= 0. \end{aligned}$$

Thus, although these systems may be effective when the host signal is known at the decoder, when the host signal is not known, they offer no guaranteed robustness to perturbations. As alluded to in Sec. 1, in a spread-spectrum system (9), $\mathbf{x}$ is an additive noise that is often much larger than $\mathbf{w}$ due to the distortion constraint. The quantization that occurs with quantization index modulation, however, removes much of the noisiness introduced by $\mathbf{x}$.[†] We shall see this robustness advantage in Sec. 4 in the context of dither modulation, a special case of quantization index modulation.

4. DITHER MODULATION: A SPECIAL CASE

For ease of implementation and analysis, one may want to impose some structure on the quantizer ensembles discussed in the last section. A convenient ensemble to consider are so-called dithered quantizers,[7,8] which have the property that the quantization cells and reconstruction points of any given quantizer in the ensemble are shifted versions of the quantization cells and reconstruction points of any other quantizer in the ensemble. In non-watermarking contexts, the shifts typically correspond to pseudorandom vectors called dither vectors. For information-embedding purposes, the dither vector can be modulated with the embedded signal, i.e., each possible embedded signal maps uniquely onto a different dither vector $\mathbf{d}(m)$. The host signal is quantized with the resulting dithered quantizer to form the composite signal. Specifically, we start with some base quantizer $\mathbf{q}(\cdot)$, and the embedding function is

$$\mathbf{s}(\mathbf{x}; m) = \mathbf{q}(\mathbf{x} + \mathbf{d}(m)) - \mathbf{d}(m).$$

We call this type of information embedding, which is a special case of quantization index modulation, "dither modulation".

As a simple example, we consider the case of coded binary dither modulation and uniform, scalar quantization with step size Δ. We assume that $1/N \leq R_m \leq 1$. The dither vectors in a coded binary dither modulation system are constructed in the following way:

- The NR_m information bits $\{b_1, b_2, \ldots, b_{NR_m}\}$ representing the embedded message m are error correction coded using a rate-k_u/k_c code to obtain a coded bit sequence $\{z_1, z_2, \ldots, z_{N/L}\}$, where

  $$L = \frac{1}{R_m}(k_u/k_c).$$

  (In the uncoded case, $z_i = b_i$ and $k_u/k_c = 1$.)

- Two dither subvectors of length-L are constructed with the constraint

  $$d_i(2) = \begin{cases} d_i(1) + \Delta/2, & d_i(1) < 0 \\ d_i(1) - \Delta/2, & d_i(1) \geq 0 \end{cases}, \quad i = 1, \ldots, L,$$

  where $d_i(1)$ and $d_i(2)$ are the i-th components of the two dither subvectors. This constraint ensures that the two corresponding L-dimensional dithered quantizers are the maximum possible distance from each other. For example, a pseudorandom sequence of $\pm\Delta/4$ and its negative satisfy this constraint. One could alternatively choose $d_i(1)$ pseudorandomly with a uniform distribution over $[-\Delta/2, \Delta/2]$.[‡]

[†] Consider, for example, that a quantized random variable has finite entropy while a continuous random variable has infinite entropy.

[‡] A uniform distribution for the dither sequence implies that the quantization error is statistically independent of the host signal and leads to fewer "false contours", both of which are generally desirable properties from a perceptual viewpoint.[7]

- One dither subvector is associated with a 0, and the other is associated with a 1.
- The sequence of N/L dither subvectors associated with the coded bit sequence $z_1, z_2, \ldots, z_{N/L}$ are concatenated to form $\mathbf{d}(m) \in \Re^N$.

If the error correction code is a binary block code with a minimum Hamming distance of d_H, then the reconstruction points of any given quantizer in the resulting ensemble are shifted by $\pm\Delta/2$ in each dimension relative to the points of any other quantizer over at least Ld_H dimensions. Thus, the minimum distance squared (5) is

$$d_{\min}^2 = Ld_H \left(\frac{\Delta}{2}\right)^2 = \left(d_H \frac{k_u}{k_c}\right) \frac{1}{R_m} \left(\frac{\Delta}{2}\right)^2 = \gamma_c \frac{1}{R_m} \left(\frac{\Delta}{2}\right)^2, \tag{10}$$

where $\gamma_c = d_H(k_u/k_c)$.

If the quantization cells are sufficiently small such that $\mathbf{x}$ can be modeled as uniformly distributed within each cell, the expected squared-error distortion per sample (2) of a uniform, scalar quantizer is

$$E[D(\mathbf{s},\mathbf{x})] = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} x^2\, dx = \frac{\Delta^2}{12}. \tag{11}$$

Thus, with bounded perturbation energy and a minimum distance decoder (8), the guaranteed error-free decoding condition (6) can be used to compactly express the trade-off among distortion, robustness and rate:

$$\begin{aligned} &\frac{d_{\min}^2}{4N\sigma_n^2} > 1 \\ \Longrightarrow\ &\gamma_c \frac{\Delta^2}{16NR_m\sigma_n^2} > 1 \\ \Longrightarrow\ &\gamma_c \frac{3}{4} \frac{1}{NR_m} \frac{\Delta^2/12}{\sigma_n^2} > 1 \\ \Longrightarrow\ &\boxed{\gamma_c \frac{3}{4} \frac{1}{NR_m} \frac{E[D(\mathbf{s},\mathbf{x})]}{\sigma_n^2} > 1.} \end{aligned} \tag{12}$$

The second line follows by substituting the expression in (10) for $d_{\min}^2$. The third line is a re-grouping of factors from the second line. The fourth line follows from (11). Thus, for example, at a fixed rate R_m to tolerate more perturbation energy σ_n^2 requires that we accept more expected distortion $E[D]$. Eq. (12) conveniently relates design specifications to design parameters for dither modulation systems. For example, if the design specifications require an embedding rate of at least R_m and robustness to noise of at least σ_n^2 in energy per sample, then (12) gives the minimum embedding-induced distortion that must be introduced into the host signal, or equivalently via (11) the minimum quantization step size Δ, to achieve these specifications. Finally, we see that γ_c is the improvement or gain in the trade-off among distortion, robustness, and rate due to the error correction code. For example, an uncoded system has $\gamma_c = 1 = 0$ dB.

As mentioned in Sec. 3, spread-spectrum systems have $d_{\min} = 0$, so no condition analogous to (12) exists under which error-free decoding is guaranteed. The nonzero minimum distance of QIM systems leads to a performance advantage over spread spectrum in the case of random additive white Gaussian noise channels as well.[9] Although LBM systems have nonzero minimum distance, the achievable performance trade-offs are not as good as those of dither modulation (12). We show in App. A that the LBM system corresponding to the coded binary dither modulation system of this section is worse by 2.43 dB (Eq. (18)).

5. ROBUSTNESS TO IN-THE-CLEAR ATTACKS

As mentioned in Sec. 2, some attackers may exploit partial knowledge of the host signal. In these cases a bounded host-distortion channel model, rather than a bounded perturbation channel model, may be appropriate.

In addition, these attackers may also exploit knowledge about the embedding and decoding processes. To limit the attackers' knowledge, some digital watermarking systems use keys, parameters that allow appropriate parties to

Table 1. Attacker's distortion penalties. The distortion penalty is the additional distortion that an attacker must incur to successfully remove a watermark. A distortion penalty less than 1 (0 dB) indicates that the attacker can actually improve the signal quality and remove the watermark simultaneously. In the quantization index modulation case, reconstruction points are assumed to lie at centroids of quantization cells.

Embedding System	Distortion Penalty ($D_{\mathbf{y}}/D_{\mathbf{s}}$)
Quant. Index Mod.	$1+\frac{1}{4}\frac{d_{\min}^2/N}{D_{\mathbf{s}}} > 0$ dB
Binary Dith. Mod.	$1+\gamma_c\frac{3/4}{NR_m} > 0$ dB
Spread Spectrum	$-\infty$ dB
LBM	≤ 0 dB

embed and/or decode the embedded signal. The locations of the modulated bits in a LBM system and the pseudo-noise vectors in a spread-spectrum system are examples of keys. If only certain parties privately share the keys to both embed and decode information, and no one else can do either of these two functions, then the watermarking system is a private-key system. Alternatively, if some parties possess keys that allow them to either embed or decode, but not both, then the system is a public-key system since these keys can be made available to the public for use in one of these two functions without allowing the public to perform the other function. However, in some scenarios it may be desirable to allow everyone to embed and decode watermarks without the use of keys. For example, in a copyright ownership notification system, everyone could embed the ASCII representation of a copyright notice such as, "Property of ..." in their copyrightable works. Such a system is analogous to the system currently used to place copyright notices in (hardcopies of) books, a system in which there is no need for a central authority to store, register, or maintain separate keys — there are none — or watermarks — all watermarks are English messages — for each user. The widespread use of such a "no-key" system in which the watermark is "in the clear" requires only standardization of the decoder so that everyone will agree on the decoded watermark, and hence, the owner of the copyright.

In this section, we examine the robustness of QIM, spread spectrum, and LBM systems to in-the-clear attacks from adversaries that have a distortion constraint, partial knowledge of the host signal, and full knowledge of the embedding and decoding processes including any keys. We show that of the three systems considered, only QIM systems are robust enough such that the attacker must degrade the host signal quality to remove the watermark.

The measure of robustness is $D_{\mathbf{y}}$, the minimum expected squared-error per letter distortion between $\mathbf{y}$ and $\mathbf{x}$ that an attacker would need to impose in order to cause a decoding error. We use $D_{\mathbf{s}}$ to denote the expected distortion between $\mathbf{s}$ and $\mathbf{x}$. The ratio between $D_{\mathbf{y}}$ and $D_{\mathbf{s}}$ is the distortion penalty that the attacker must pay to remove the watermark, and hence, is a figure of merit measuring the trade-off between robustness and embedding-induced distortion at a given rate. Distortion penalties for QIM, spread-spectrum, and LBM systems are derived below and are shown in Table 1.

5.1. Quantization Index Modulation

We first consider the robustness of quantization index modulation. We assume that all reconstruction points $\mathbf{s}$ lie at the centroids of their respective quantization cells.

We use $\mathcal{R}$ to denote the quantization cell containing $\mathbf{x}$ and $p_{\mathbf{x}}(\mathbf{x})$ to denote the conditional probability density function of $\mathbf{x}$ given that $\mathbf{x} \in \mathcal{R}$. Again, for sufficiently small quantization cells, this probability density function can often be approximated as uniform over $\mathcal{R}$, for example. Since $\mathbf{s}$ is the centroid of $\mathcal{R}$,

$$\int_{\mathcal{R}} (\mathbf{s}-\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} = \mathbf{0}. \tag{13}$$

Also, the average squared-error per letter distortion due to the embedding given $\mathbf{x} \in \mathcal{R}$ is

$$D_{\mathbf{s}} = \frac{1}{N}\int_{\mathcal{R}} ||\mathbf{s}-\mathbf{x}||^2 p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x}. \tag{14}$$

The most general attack can always be represented as $\mathbf{y} = \mathbf{s} + \mathbf{n}$, where $\mathbf{n}$ may be a function of $\mathbf{s}$. The resulting distortion is

$$\begin{aligned} D_{\mathbf{y}} &= \frac{1}{N}\int_{\mathcal{R}} \|\mathbf{y}-\mathbf{x}\|^2 p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} = \frac{1}{N}\int_{\mathcal{R}} \|(\mathbf{s}-\mathbf{x})+\mathbf{n}\|^2 p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \\ &= \frac{1}{N}\int_{\mathcal{R}} \|\mathbf{s}-\mathbf{x}\|^2 p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} + \frac{1}{N}\|\mathbf{n}\|^2 \int_{\mathcal{R}} p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} + \frac{2}{N}\mathbf{n}^{\mathrm{T}}\int_{\mathcal{R}} (\mathbf{s}-\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} \\ &= D_{\mathbf{s}} + \frac{\|\mathbf{n}\|^2}{N}, \end{aligned}$$

where we have used (14), the fact that $p_{\mathbf{x}}(\mathbf{x})$ is a probability density function and thus integrates to one, and (13) to obtain the last line. For a successful attack, $\|\mathbf{n}\| \geq d_{\min}/2$ so our figure of merit for a quantization index modulation system is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} \geq 1 + \frac{1}{4}\frac{d_{\min}^2/N}{D_{\mathbf{s}}}. \qquad (15)$$

Thus, for any QIM system of nonzero $d_{\min}$, the attacker's distortion penalty is always greater than 1 (0 dB), indicating that to remove the watermark, the attacker must degrade the host signal quality beyond the initial distortion caused by the embedding of the watermark.

In the special case of coded binary dither modulation with uniform, scalar quantization considered in Sec. 4, Eq. (10) gives $d_{\min}^2$, and Eq. (11) gives the distortion $D_{\mathbf{s}}$. Thus, the attacker's distortion penalty (15) that must be paid to defeat the watermark in this case is

$$\boxed{\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} \geq 1 + \gamma_c \frac{3/4}{N R_m}.}$$

5.2. Spread-spectrum Modulation

The embedding function of a spread-spectrum system is

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

so the resulting distortion is

$$D_{\mathbf{s}} = \|\mathbf{w}\|^2/N > 0.$$

An attacker with full knowledge of the embedding and decoding processes can decode the message m, and hence, reproduce the corresponding pseudo-noise vector $\mathbf{w}$. Therefore, the attacker can completely remove the watermark by subtracting $\mathbf{w}$ from $\mathbf{s}$ to obtain the original host signal,

$$\mathbf{y} = \mathbf{s} - \mathbf{w}(m) = \mathbf{x}.$$

Hence, the resulting distortion penalty is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} = \frac{0}{D_{\mathbf{s}}} = -\infty \text{ dB}.$$

Because the spread-spectrum embedding function combines the host signal $\mathbf{x}$ and watermark $\mathbf{w}(m)$ in a simple linear way, anyone that can extract the watermark, can easily remove it. Thus, these systems are not very robust to in-the-clear attacks. In contrast, the quantization that occurs in quantization index modulation systems effectively hides the exact value of the host signal even when the embedded information m is known, thus allowing no-key digital watermarking with a positive (in dB) attacker's distortion penalty.

5.3. Low-bit(s) Modulation

The embedding function of a LBM system can be written as

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

where $\mathbf{q}(\cdot)$ represents the coarse quantizer that determines the most significant bits and $\mathbf{d}$ represents the effect of the (modulated) least significant bits. Because the embedding never alters the most significant bits of the host signal,

$$\mathbf{q}(\mathbf{s}) = \mathbf{q}(\mathbf{x}).$$

One possible attack is to simply remodulate the least significant bits with some other message m',

$$\mathbf{y} = \mathbf{q}(\mathbf{s}) + \mathbf{d}(m') = \mathbf{q}(\mathbf{x}) + \mathbf{d}(m').$$

Since both $\mathbf{s}$ and $\mathbf{y}$ are both low-bit(s) modulated versions of $\mathbf{x}$, the distortions must be equal, particularly if the distortions are averaged over all possible choices of m and m'. Thus, the attacker's distortion penalty in this case is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} = 1 = 0 \text{ dB},$$

i.e., an attacker can remove the watermark without causing additional distortion to the host signal. This result applies regardless of whether error correction coding is used. Thus, in contrast to dither modulation (See Table 1.), error correction coding does not improve low-bit(s) modulation in this context. As a final note, although the distortion penalty for this particular attack is 0 dB, this attack is not necessarily the best that an attacker could choose. Thus, the argument above shows only that 0 dB is an upper bound on the distortion penalty, a fact that is reflected in Table 1.

6. FUNDAMENTAL PERFORMANCE LIMITS FOR RANDOM CHANNELS

In previous sections we consider primarily deterministic channels using worst-case analyses, determining the performance limits of digital watermarking that is robust to all attacks belonging to a given class such as bounded perturbation or bounded host-distortion attacks. These analyses rely on very few assumptions about the channel. In some scenarios, however, one may wish to incorporate additional knowledge about the relationship between the channel input and output. A common approach to modeling this relationship is to assume a conditional probability law of the channel output given the input. In this section, we consider the fundamental limits of information embedding over these random channels.

We restrict our analysis to the case of a blockwise independently and identically distributed (iid) host, a blockwise-memoryless channel, a block embedding function, and a random embedded message. Specifically, the assumptions are:

1. **blockwise-iid host signal:** The host signal is $\mathbf{x} = [x_1 \cdots x_{N/L}]^T$, where the x_i are a collection of iid, L-dimensional subvectors with a probability density function $p_x(x_i)$.
2. **blockwise-memoryless channel:** We assume a probabilistic channel model of the form
$$p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s}) = \prod_{j=1}^{N/L} p_{y|s}(y_j|s_j),$$
where $\mathbf{y} = [y_1 \cdots y_{N/L}]^T$ and $\mathbf{s} = [s_1 \cdots s_{N/L}]^T$.
3. **block embedding function:** The overall N-dimensional embedding function $\mathbf{s}(\mathbf{x}, m)$ can be decomposed into a sequence of N/L L-dimensional embedding functions so that s_i is a function of only x_i and z_i, the i-th subvector of an encoding of m. This model is sufficiently general to include both QIM systems with finite-dimensional quantizers and spread-spectrum systems, as explained below.
4. **random message:** In this section the information-embedding rate R_m is the amount of information, as measured by its entropy, that is embedded per host signal sample. Thus, we consider m to be a random integer chosen uniformly from the set $\{1, \ldots, 2^{NR_m}\}$.

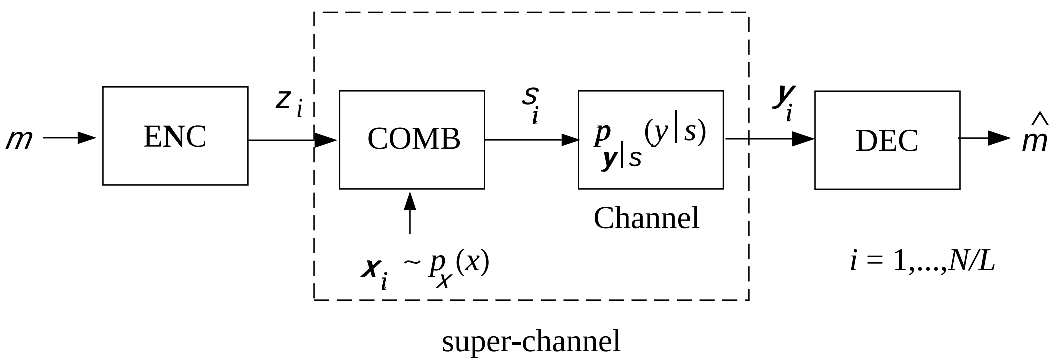

Figure 3. Equivalent super-channel for information embedding. The embedding function can be decomposed into two stages, encoding and combining. The cascade of the combiner with the true channel forms a super-channel.

The embedding system can be decomposed into two stages as shown in Fig. 3. The first stage is an encoder, which takes the embedded information m as input and produces the N/L L-dimensional code subvectors z_i as output. The second stage combines each z_i with the corresponding x_i to form s_i. A QIM system combines by quantizing x_i with the quantizer associated with z_i. For example, z_i could be the dither subvector of a dither modulation system. A spread-spectrum system combines by adding to x_i the pseudo-noise subvector associated with z_i. We see that in both cases the cascade of the combiner with the true channel forms a blockwise-memoryless super-channel.

For a given combiner, reliable information-embedding (where $\Pr[\hat{m} \neq m]$ is arbitrarily small) is possible if and only if

$$R_m \leq \frac{1}{L} I(z; y),$$

where $I(\cdot;\cdot)$ denotes mutual information.[10] Thus, the optimal encoder and combiner maximize $I(z;y)$ subject to a distortion constraint $E[D(s,x)] \leq D_{\max}$. Also, in general, the optimal encoder involves long codes that achieve optimal performance asymptotically with large N/L.

Useful insight can be obtained by considering the case of a scalar combiner ($L = 1$) and an additive noise channel ($\mathbf{n} = \mathbf{y} - \mathbf{s}$ is independent of $\mathbf{s}$). Interestingly, in the small distortion limit (where the host signal variance σ_x^2 is much larger than the embedding-induced distortion $D_{\mathbf{s}}$), the achievable information-embedding rate with dither modulation can be nonzero even if the achievable rate with an additive spread-spectrum system is zero, as we show below. This result indicates that dither modulation is much more attractive than spread spectrum for applications where the host signal, rather than channel noise, is the dominant noise source at the decoder, a result that is closely related to the minimum distance properties of the associated systems developed in Sec. 3 and 4.

With an additive system such as spread spectrum, $y_i = s_i + n_i = z_i + x_i + n_i$. Thus, the achievable rate $I(z;y)$ equals the capacity of an additive noise channel with a power constraint $E[z^2] = D_{\mathbf{s}}$ on the input and an effective noise $x + n$. In the small distortion limit, the effective noise variance $\sigma_x^2 + \sigma_n^2$ is infinite, even if σ_n^2 is finite (not much larger than $D_{\mathbf{s}}$). Thus, if the effective noise has a Gaussian distribution, for example, the capacity of this effective additive noise channel is zero. Indeed, even if the effective noise is not Gaussian, but one uses, as is commonly done, a Gaussian codebook for z and a minimum distance decoder (or equivalently, a correlation decoder), then the capacity is still zero.[11]

In contrast, the achievable rate with dither modulation can be nonzero, which can be shown by first defining a random variable $v_i \triangleq q(y_i) - y_i$ and noting that

$$\begin{aligned} v_i &= q(q(x_i + z_i) - z_i + n_i) - [q(x_i + z_i) - z_i + n_i] \\ &= q(n_i - z_i) + q(x_i + z_i) - q(x_i + z_i) + z_i - n_i \\ &= z_i + q(n_i - z_i) - n_i, \end{aligned}$$

where z_i is the i-th element of the dither vector. The first two terms of the second line arise from the identity $q(q(a) + b) = q(a) + q(b)$ for a uniform scalar quantization function $q(\cdot)$. The third line shows that v_i is independent of x_i, and hence, $I(z;v)$ can be greater than zero even when σ_x^2 is infinite. (For example, when $\sigma_n^2 = 0$ and z_i has a uniform distribution over the quantization cell, $v_i = z_i$ and $I(z;v)$ is infinite.) The data processing inequality[10] states that $I(z;y) \geq I(z;v)$. Thus, if $I(z;v)$ can be nonzero in the small distortion limit, then so can the achievable embedding rate $I(z;y)$.

More general results on achievable information-embedding rates can be found in an upcoming paper.[12]

7. CONCLUDING REMARKS

Quantization index modulation (QIM) systems in general, and dither modulation in particular, offer significant performance advantages over previously proposed spread-spectrum and low-bit(s) modulation systems in terms of the achievable trade-offs among information-embedding rate, distortion, and robustness. For a given rate and embedding-induced distortion, the nonzero minimum distance of QIM systems makes the embedding considerably more robust than that of spread-spectrum systems, which have zero minimum distance, in scenarios where the host signal is not available at the decoder. We have demonstrated this performance advantage in the fairly general cases of bounded perturbation channels and bounded host-distortion channels, both of which may be useful for modeling effects of lossy compression or an adversary's attempts to remove the embedded signal with an SNR constraint, for example.

No-key digital watermarking systems may be useful for applications such as copyright identification, provided that these systems are robust to in-the-clear attacks. In these scenarios, an attacker of a QIM system incurs a distortion penalty despite the fact that the attacker can exploit full knowledge of the embedding and decoding processes. In contrast, an attacker incurs no distortion penalty when attacking a low-bit(s) modulation system and can actually completely remove the watermark with no distortion when attacking a spread-spectrum system due to its simple, linear nature.

Finally, information-theoretic analysis also reveals performance advantages of QIM systems in the commonly studied case of additive random noise channels.

APPENDIX A. LOW-BIT(S) MODULATION DISTORTION

In this appendix we calculate the expected squared-error distortion per sample of a low-bit(s) modulation system. We assume that the host signal and embedded signal are statistically independent.

The embedding function of such a system can be written as

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

where $\mathbf{q}(\cdot)$ is a coarse quantizer that determines the most significant bits in the quantization of $\mathbf{x}$, and $\mathbf{d}$ is determined by the modulated least significant bits. We define $\mathbf{q}(\cdot)$ such that its reconstruction points lie at the centroids of its quantization cells so that

$$E[\mathbf{q}(\mathbf{x}) - \mathbf{x}] = \mathbf{0}. \tag{16}$$

Then, the expected distortion is

$$\begin{aligned}\frac{1}{N}E\left[\|\mathbf{s}-\mathbf{x}\|^2\right] &= \frac{1}{N}E\left[\|\mathbf{q}(\mathbf{x})-\mathbf{x}+\mathbf{d}(m)\|^2\right] \\ &= \frac{1}{N}E\left[\|\mathbf{q}(\mathbf{x})-\mathbf{x}\|^2 + 2(\mathbf{q}(\mathbf{x})-\mathbf{x})^T\mathbf{d}(m) + \|\mathbf{d}(m)\|^2\right] \\ &= \frac{1}{N}E\left[\|\mathbf{q}(\mathbf{x})-\mathbf{x}\|^2\right] + \frac{1}{N}E\left[\|\mathbf{d}(m)\|^2\right],\end{aligned} \tag{17}$$

where we have used (16) and the independence of $\mathbf{x}$ and m to obtain the final line. Thus, the overall distortion is the distortion of the coarse quantizer plus the expected magnitude-squared per sample of the least significant bits adjustment vector $\mathbf{d}(m)$.

We consider the special case of a uniform, scalar quantizer with modulation of the least significant bit and compare its distortion to the dither modulation example of Sec. 4. The low-bit modulation system is illustrated in Fig. 4. The coarse quantizer $\mathbf{q}(\cdot)$ has a step size of Δ, and every component of $\mathbf{d}$ equals $\pm\Delta/4$. Consequently, the system has the same minimum distance as the dither modulation system of Sec. 4 and, hence, the same noise tolerance for a given rate. If we make the same assumption as in Sec. 4 that $\mathbf{x}$ can be modeled as uniformly distributed within each

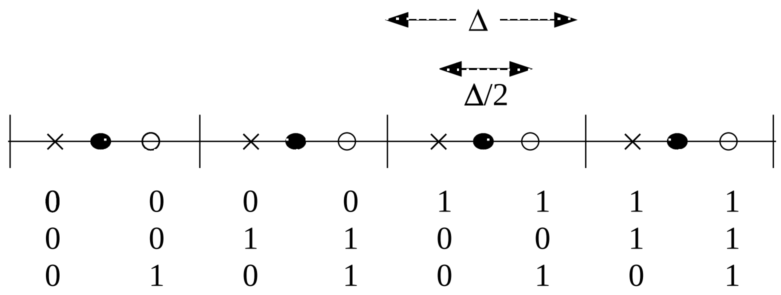

Figure 4. Low-bit modulation with a uniform, scalar quantizer. The quantizer has a step size of $\Delta/2$, and the least significant bit (lsb) is modulated. All reconstruction points marked with a $\times$ have a lsb of 0. Points marked with a $\circ$ have a lsb of 1. This process is equivalent to first quantizing using a quantizer with a step size of Δ, whose reconstruction points are marked with a $\bullet$, and adding $\pm\Delta/4$.

cell of $\mathbf{q}(\cdot)$, then the first term in (17) is $\Delta^2/12$, the same as the expected distortion (11) of the dither modulation system. The second term is $\Delta^2/16$ since every component of $\mathbf{d}$ is $\pm\Delta/4$. Thus, the overall expected distortion is

$$\left(\frac{1}{12}+\frac{1}{16}\right)\Delta^2 = \frac{7}{48}\Delta^2.$$

Therefore, the low-bit modulation system is worse than the dither modulation system by

$$\frac{7/48}{1/12} = \frac{7}{4} = 2.43\text{ dB}. \tag{18}$$

ACKNOWLEDGMENTS

This work has been supported in part by the Office of Naval Research under Grant No. N00014-96-1-0930, by the Air Force Office of Scientific Research under Grant No. F49620-96-1-0072, 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," *Proceedings of the IEEE* **86**, pp. 1064–1087, June 1998.
2. I. J. Cox and J.-P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE Journal on Selected Areas in Communications* **16**, pp. 587–593, May 1998.
3. J. M. Barton, "Method and apparatus for embedding authentication information within digital data." United States Patent #5,646,997. Issued July 8, 1997.
4. K. Tanaka, Y. Nakamura, and K. Matsui, "Embedding secret information into a dithered multi-level image," in *Proceedings of the 1990 IEEE Military Communications Conference*, pp. 216–220, 1990.
5. B. Chen and G. W. Wornell, "System, method, and product for information embedding using an ensemble of non-intersecting embedding generators." U.S. patent pending. Licensing info: MIT Technology Licensing Office.
6. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publishers, 2nd ed., 1994.
7. N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, 1984.
8. R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Transactions on Information Theory* **42**, pp. 1152–1159, July 1996.
9. B. Chen and G. W. Wornell, "Digital watermarking and information embedding using dither modulation," in *Proceedings of IEEE Workshop on Multimedia Signal Processing (MMSP-98)*, (Redondo Beach, CA), Dec. 1998.
10. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.
11. A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Transactions on Information Theory* **42**, pp. 1520–1529, Sept. 1996.
12. B. Chen and G. W. Wornell, "Dither modulation and quantization index modulation: New methods for digital watermarking and information embedding." Preprint.