Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems

by

Brian Chen

B.S.E., University of Michigan (1994)
S.M., Massachusetts Institute of Technology (1996)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 2000

Signature of Author: _____

Department of Electrical Engineering and Computer Science
February 15, 2000

Certified by: _____

Gregory W. Wornell
Associate Professor of Electrical Engineering
Thesis Supervisor

Accepted by: _____

Arthur C. Smith
Chairman, Department Committee on Graduate Students

Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems

by

Brian Chen

Submitted to the Department of Electrical Engineering and Computer Science
on February 15, 2000, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Digital watermarking, information embedding, and data hiding systems embed information, sometimes called a digital watermark, inside a host signal, which is typically an image, audio signal, or video signal. The host signal is not degraded unacceptably in the process, and one can recover the watermark even if the composite host and watermark signal undergo a variety of corruptions and attacks as long as these corruptions do not unacceptably degrade the host signal.

These systems play an important role in meeting at least three major challenges that result from the widespread use of digital communication networks to disseminate multimedia content: (1) the relative ease with which one can generate perfect copies of digital signals creates a need for copyright protection mechanisms, (2) the relative ease with which one can alter digital signals creates a need for authentication and tamper-detection methods, and (3) the increase in sheer volume of transmitted data creates a demand for bandwidth-efficient methods to either backwards-compatibly increase capacities of existing legacy networks or deploy new networks backwards-compatibly with legacy networks.

We introduce a framework within which to design and analyze digital watermarking and information embedding systems. In this framework performance is characterized by achievable rate-distortion-robustness trade-offs, and this framework leads quite naturally to a new class of embedding methods called quantization index modulation (QIM). These QIM methods, especially when combined with postprocessing called distortion compensation, achieve provably better rate-distortion-robustness performance than previously proposed classes of methods such as spread spectrum methods and generalized low-bit modulation methods in a number of different scenarios, which include both intentional and unintentional attacks. Indeed, we show that distortion-compensated QIM methods can achieve capacity, the information-theoretically best possible rate-distortion-robustness performance, against both additive Gaussian noise attacks and arbitrary squared error distortion-constrained attacks. These results also have implications for the problem of communicating over broadcast channels. We also present practical implementations of QIM methods called dither modulation and demonstrate their performance both analytically and through empirical simulations.

Thesis Supervisor: Gregory W. Wornell
Title: Associate Professor of Electrical Engineering

Acknowledgments

First, of course, I would like to thank my advisor, Prof. Greg Wornell, for his guidance, helpful insights, and useful advice throughout the entire course of my graduate studies. He has truly been a model advisor.

I also gratefully acknowledge the Office of Naval Research, the Air Force Office of Scientific Research, the U.S. Army Research Laboratory, the MIT Lincoln Laboratory Advanced Concepts Committee, and the National Defense Science and Engineering Graduate Fellowship Program for their financial contributions.

I would like to thank the two readers on my committee, Prof. Robert Gallager and Prof. David Forney, for their efforts. Discussions with Prof. Amos Lapidoth were also of great help, especially when he referred me to the paper by Costa.

I am also grateful to my instructors at MIT. Much of the background for the work in this thesis was taught to me by the professors and TAs in my courses here. I thank all of them for their dedication to teaching.

My fellow graduate students of the Digital Signal Processing Group are also certainly deserving of recognition and acknowledgment for the many perspectives and insights they provided during informal conversations. It has been a privilege to interact with such bright, talented, and engaging people.

I would also like to acknowledge the professors at my undergraduate institution, the University of Michigan. They helped me get started in electrical engineering and taught me the fundamentals. Furthermore, my friends and fellow alumni in the University of Michigan Club of Greater Boston have provided welcome social relief from the day-to-day rigors of graduate student life. I would especially like to thank Robin, Julie, Nick, Brian P., and Nikki. I am very proud to be part of the Michigan family. Go Blue!

Finally, and most importantly, I would like to thank my family for all of their help and support. My sister Nancy has always set a good example for me, leading the way and showing me the ropes. My mom has given me all the loving care that only a mother can provide. Her Thanksgiving and Christmas dinners kept me going all year long. Finally, my dad has always supported me in all of my endeavors. He is my role model, confidant, and most trusted advisor.

To my parents,

my source of seed funding in human capital.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Digital watermarking and information embedding systems have a number of important multimedia applications [20, 41]. These systems embed one signal, sometimes called an "embedded signal" or "watermark", within another signal, called a "host signal". The embedding must be done such that the embedded signal causes no serious degradation to its host. 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.

1.1 Information Embedding Applications

One such application — copyright notification and enforcement — arises due to the relative ease with which one can create perfect copies of digital signals. Digital watermarking is one way to help prevent or reduce unintentional and intentional copyright infringement by either notifying a recipient of any copyright or licensing restrictions or inhibiting or deterring unauthorized copying. In some cases the systems need to be robust only against so-called unintentional attacks, common signal corruptions from sources such as lossy compression, format changes, and digital-to-analog-to-digital conversion. In other cases the systems must also resist deliberate attacks by "hackers". Typically, the digital watermark is embedded into multimedia content — an audio signal, a video signal, or an image, for example — and (1) identifies the content owner or producer, (2) identifies the recipient or purchaser, (3) enables a standards-compliant device to either play or duplicate the content, or (4) prevents

a standards-compliant device from playing or duplicating the content.

For example, a watermark embedded in a digital photograph could identify the photographer and perhaps include some contact information such as email address or phone number. Popular commercial image-editing software could include a watermark decoder and could notify the user that the photograph is copyrighted material and instruct the user to contact the photographer for permission to use or alter the photograph. Alternatively, a web crawler could look for the photographer's watermark in images on the Web and notify the photographer of sites that are displaying his or her photographs. Then, the photographer could contact the website owners to negotiate licensing arrangements.

Instead of identifying the content owner, the watermark could uniquely identify the purchaser, acting as a kind of "digital fingerprint" that is embedded in any copies that the purchaser creates. Thus, if the content owner obtains any versions of his or her content that were distributed or used in an unauthorized fashion, he or she can decode the watermark to identify the original purchaser, the source of the unauthorized copies.

The digital watermark could also either enable or disable copying by some duplication device that checks the embedded signal before proceeding with duplication. Such a system has been proposed for allowing a copy-once feature in digital video disc recorders [16]. Alternatively, a standards-compliant player could check the watermark before deciding whether or not to play the disc [28].

In addition to being easily duplicated, digital multimedia signals are also easily altered and manipulated, and authentication of, or detection of tampering with, multimedia signals is another application of digital watermarking methods [24]. So-called "fragile" watermarks change whenever the composite signal is altered significantly, thus providing a means for detecting tampering. Alternatively, one could embed a robust watermark, a digital signature, for example, within a military map. If the map is altered, the watermark may survive, but will not match the altered map. In contrast to traditional authentication methods, in both the fragile and robust cases, the watermark is embedded directly into the host signal. Thus, no side channel is required, and one can design the watermarking algorithm such that one can authenticate signals in spite of common format changes or lossy compression.

In addition to authentication, a number of national security applications are described in [1] and include covert communication, sometimes called "steganography" [33] or low probability of detection communication, and so-called traitor tracing, a version of the digital

fingerprinting application described above used for tracing the source of leaked information. In the case of covert communication, the host signal conceals the presence of the embedded signal, which itself may be an encrypted message. Thus, steganographic techniques hide the existence of the message, while cryptographic techniques hide the message's meaning.

Although not yet widely recognized as such, bandwidth-conserving hybrid transmission is yet another information embedding application, offering the opportunity to re-use and share existing spectrum to either backwards-compatibly increase the capacity of an existing communication network, *i.e.*, a "legacy" network, or allow a new network to be backwards-compatibly overlayed on top of the legacy network. In this case the host signal and embedded signal are two different signals that are multiplexed, *i.e.*, transmitted simultaneously over the same channel in the same bandwidth, the host signal being the signal corresponding to the legacy network. Unlike in conventional multiplexing scenarios, however, the backwards-compatibility requirement imposes a distortion constraint between the host and composite signals.

So-called hybrid in-band on-channel digital audio broadcasting (DAB) [5, 32] is an example of such a multimedia application where one may employ information embedding methods to backwards-compatibly upgrade the existing commercial broadcast radio system. In this application one would like to simultaneously transmit a digital signal with existing analog (AM and/or FM) commercial broadcast radio without interfering with conventional analog reception. Thus, the analog signal is the host signal, and the digital signal is the watermark. Since the embedding does not degrade the host signal too much, conventional analog receivers can demodulate the analog host signal. In addition, next-generation digital receivers can decode the digital signal embedded within the analog signal. This embedded digital signal may be all or part of a digital audio signal, an enhancement signal used to refine the analog signal, or supplemental information such as station identification. More generally, the host signal in these hybrid transmission systems could be some other type of analog signal such as video [43] or even a digital waveform. For example, a digital pager signal could be embedded within a digital cellular telephone signal.

Automated monitoring of airplay of advertisements on commercial radio broadcasts is one final example of a digital watermarking application. Advertisers can embed a digital watermark within their ads and count the number of times the watermark occurs during a given broadcast period, thus ensuring that their ads are played as often as promised.

In this case, however, the watermark is embedded within the baseband source signal (the advertisement), whereas in the bandwidth-conserving hybrid transmission applications discussed above, the digital signal may be embedded in either the baseband source signal or the passband modulated signal (a passband FM signal, for example).

1.2 Thesis Summary

A large number of information-embedding algorithms have been proposed [20, 33, 41] in this still emerging field. As will be developed in Sec. 2.4, one can classify these methods according to whether or not the host signal interferes with one's ability to recover the embedded watermark. A simple example of a host-interference rejecting method is the quantization-and-perturbation method of [43], which may be viewed as a type of generalized low-bit(s) modulation (LBM). These LBM methods range from simple replacement of the least significant bit(s) of the pixels of an image with a binary representation of the watermark to more sophisticated methods such as the one in [43] that involve transformation of the host signal before quantization and adjustment of the quantization step sizes. Host-interference non-rejecting methods include linear classes of methods such as spread-spectrum methods, which embed information by linearly combining the host signal with a small pseudo-noise signal that is modulated by the embedded signal. Although these methods have received considerable attention in the literature [4, 15, 22, 37, 44, 45], these methods are limited by host-signal interference when the host signal is not known at the decoder, as is typical in many of the applications mentioned above. Intuitively, the host signal in a spread spectrum system is an additive interference that is often much larger, due to distortion constraints, than the pseudo-noise signal carrying the embedded information.

In this thesis we examine information embedding problems from the highest, most fundamental level. Based on first principles we arrive at a general class of host-interference rejecting embedding methods called quantization index modulation (QIM) that perform provably better than the methods mentioned above in a wide variety of different contexts. We also examine the fundamental performance limits of information embedding methods.

We begin our discussion in Chap. 2 by developing formal mathematical models of the information embedding applications discussed above. In particular, one can view information embedding either as distortion-constrained multiplexing or as communication over

a super-channel with side information that is known at the encoder. Depending on the context, one view may be more convenient than the other, and we develop mathematically equivalent models from both of these perspectives. We also develop a framework in which the performance of an information embedding method may be characterized based on its achievable rate-distortion-robustness trade-offs and discuss how previously proposed digital watermarking algorithms fit into this framework.

The framework we develop in Chap. 2 leads quite naturally to the QIM class of embedding methods introduced in Chap. 3. In QIM information embedding, each quantizer within an ensemble of quantizers is associated with an index. The watermark modulates an index sequence, and the associated quantizer sequence is used to quantize the host signal, *i.e.*, the host signal is mapped to a sequence of reconstruction points. By ensuring that the sets of reconstruction points of the different quantizers in the ensemble are non-intersecting, one obtains a host-interference rejection property. Also, as discussed in more detail in Chap. 3, QIM methods are convenient from an engineering perspective because one can easily trade-off rate, distortion, and robustness by adjusting a few system parameters. Finally, we also describe so-called distortion compensation, which is a type of post-quantization processing that provides enhanced rate-distortion-robustness performance.

Not only is the QIM structure convenient from an engineering perspective, but such a structure also has a theoretical basis from an information-theoretic perspective, as we discuss in Chap. 4. In this chapter, we examine the fundamental rate-distortion-robustness performance limits, *i.e.*, capacity, of information embedding methods in general and show that one can achieve capacity against any fixed attack with a type of "hidden" QIM. We also discuss conditions under which distortion-compensated QIM can achieve capacity.

In general, though, one achieves capacity only asymptotically with long signal lengths, so we develop practical implementations of QIM called dither modulation in Chap. 5. The QIM quantizer ensembles in a dither modulation system are dithered quantizers, and modulating the quantization index is equivalent to modulating the dither signal. Such a structure allows for implementations with low computational complexity, especially if the quantizers are uniform, scalar quantizers. We also discuss spread-transform dither modulation in this chapter, a form of dither modulation that can easily be shown to outperform both so-called amplitude-modulation spread-spectrum methods and generalized LBM methods.

After having introduced a general framework in which to analyze digital watermarking

problems and having presented some novel embedding methods, in the remaining chapters we apply the framework to particular scenarios of interest, starting with a discussion of Gaussian scenarios in Chap. 6. Here, we derive information embedding capacities for Gaussian host signals and additive Gaussian noise channels, which may be good models for unintentional degradations to the composite signal. Our results apply to arbitrary host covariance matrices and arbitrary noise covariance matrices, and hence, apply to a large number of multimedia application scenarios, as discussed in Sec. 6.2. One of the more interesting results in this chapter is that one can embed at a rate of about 1/3 b/s per Hertz of host signal bandwidth per dB drop in received host signal-to-noise ratio (SNR) and that this capacity is independent of whether or not the host signal is available during watermark decoding. As we also discuss in Sec. 6.2, results in this chapter have important connections to the problem of communicating over broadcast channels, even in non-watermarking contexts. We conclude the chapter with a discussion of the gaps to capacity of QIM and spread spectrum methods and show that spread spectrum methods generally have a large gap, QIM methods have a small gap, and distortion-compensated QIM (DC-QIM) methods have no gap, *i.e.*, capacity-achieving DC-QIM methods exist in the Gaussian case.

We focus on intentional attacks in Chap. 7, considering both attacks on systems protected by a private key and worst-case attacks, where the attacker may know everything about the embedding and decoding processes, including any keys. Just as in the Gaussian case, in the case of squared error distortion-constrained attacks on private-key systems, one can achieve capacity with DC-QIM. In the no-key scenarios, QIM methods are provably better than spread-spectrum and generalized LBM methods.

To supplement our analytical results, we present simulation results in Chap. 8 for additive Gaussian noise channels and for JPEG compression channels. We also provide some sample host, composite, and channel output images in this chapter and demonstrate practically achievable gains from error correction coding and distortion compensation.

We conclude the thesis in Chap. 9, where we also discuss some possible directions for future work.

Chapter 2

Mathematical Modeling of Digital Watermarking

A natural starting point in the discussion of information embedding systems is to develop mathematical models that suitably describe information embedding applications such as those discussed in Chap. 1. Such models facilitate a precise consideration of the issues involved in the design and performance evaluation of information embedding systems. We present two mathematically equivalent models in this chapter from two different perspectives. We conclude the chapter with a discussion of classes of embedding methods. The reader is referred to App. A for notational conventions used throughout this thesis.

2.1 Distortion-constrained Multiplexing Model

Our first model is illustrated in Fig. 2-1. 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

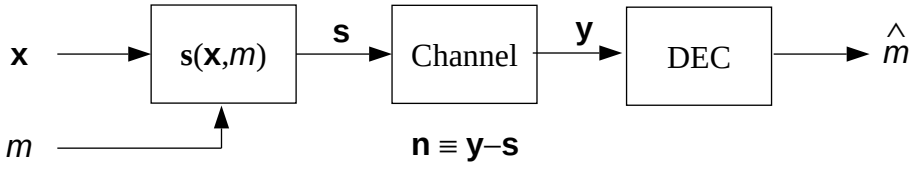

Figure 2-1: General information-embedding problem model. An integer message m is embedded in the host signal vector $\mathbf{x}$ using some embedding function $\mathbf{s}(\mathbf{x}, m)$. A perturbation vector $\mathbf{n}$ corrupts the composite signal $\mathbf{s}$. The decoder extracts an estimate $\hat{m}$ of m from the noisy channel output $\mathbf{y}$.

values, audio samples, or speech samples, for example. Alternatively, $\mathbf{x}$ could be a vector of coefficients from a linear transform of the host signal, such as the Discrete Cosine Transform (DCT) or the Discrete Fourier Transform (DFT), or from some non-linear transform. We emphasize the our framework is general enough to accommodate any representation of the host signal that involves real numbers.

We wish to embed at a rate of R_m bits per dimension (bits per host signal sample) so we can think of $\boldsymbol{m}$ as an integer, where

$$\boldsymbol{m} \in \left\{1, 2, \ldots, 2^{NR_m}\right\}. \tag{2.1}$$

The integer $\boldsymbol{m}$ can represent, for example, the watermark or digital fingerprint in a copyright application, an authentication signal, a covert message, or a digital signal communicated using an existing analog communication system. Again, our model applies to any type of digital embedded signal, including of course a digitized version of an analog signal. Although we focus in this thesis on the digital case since it is the one of interest in most applications, many of the embedding algorithms considered in later chapters can also be extended to include embedding of analog data as well, as discussed in Sec. 5.3.

These embedding algorithms embed $\boldsymbol{m}$ in the host signal $\mathbf{x}$ by mapping $\boldsymbol{m}$ and $\mathbf{x}$ to a composite signal vector $\mathbf{s} \in \Re^N$ using some embedding function $\mathbf{s}(\mathbf{x}, \boldsymbol{m})$. As explained in Chap. 1, the embedding must be done such that any degradations to the host signal are acceptable. This degradation is measured quantitatively by some distortion function between the host and composite signals. For example, two convenient distortion measures that are amenable to analysis are the well-known squared error distortion measure

$$D(\mathbf{s}, \mathbf{x}) = \frac{1}{N}\|\mathbf{s} - \mathbf{x}\|^2, \tag{2.2}$$

and the weighted squared error distortion measure

$$D(\mathbf{s}, \mathbf{x}) = \frac{1}{N}(\mathbf{s} - \mathbf{x})^T W (\mathbf{s} - \mathbf{x}), \tag{2.3}$$

where W is some weighting matrix. Also, the expectations $D_{\mathbf{s}}$ of these distortions taken over a probability distribution of the host signal and/or the embedded information are yet another set of distortion measures.

After the embedding, the composite signal $\mathbf{s}$ is typically subjected to a variety of signal corrupting manipulations such as lossy compression, addition of random noise, and resampling, as well as deliberate attempts to remove the embedded information. These manipulations occur inside some channel, which produces an output vector $\mathbf{y} \in \Re^N$. For convenience, we define a perturbation vector to be the difference $\mathbf{n} \stackrel{\Delta}{=} \mathbf{y} - \mathbf{s}$. Thus, this model is sufficiently general to include both random and deterministic perturbation vectors and both signal-independent and signal-dependent perturbation vectors. We restrict attention to cases where the degradations caused by the channel are not too large for at least two reasons. First, if we allow arbitrary degradations, for example, where the channel output is totally independent of the channel input, then clearly one cannot hope to reliably extract the embedded information from the channel output. Second, in most applications only this bounded degradation channel case is of interest. For example, it is of no value for an attacker to remove a copyright-protecting watermark from an image if the image itself is destroyed in the process. In Sec. 2.3, we give some examples of channel models and corresponding degradation measures that will be of interest in this thesis.

A decoder extracts, or forms an estimate $\hat{m}$ of, the embedded information m based on the channel output $\mathbf{y}$. We focus in this thesis on the "host-blind" case, where $\mathbf{x}$ is not available to the decoder. In some applications the decoder can also observe the original host signal $\mathbf{x}$. We comment on these less typical "known-host" cases throughout this thesis, where appropriate. The decoder ideally can reliably[1] extract the embedded information as long as the channel degradations are not too severe. Thus, the tolerable severity of the degradations is a measure of the robustness of an information embedding system.

One would like to design the embedding function $\mathbf{s}(\mathbf{x}, m)$ and corresponding decoder to achieve a high rate, low distortion, and high robustness. However, in general these three goals are conflicting. Thus, one evaluates the performance of the embedding system in terms of its achievable trade-offs among these three parameters. Such a characterization of the achievable rate-distortion-robustness trade-offs is equivalent to a notion of provable robustness at a given rate and distortion.

[1] "Reliably" can mean either that one can guarantee that $\hat{m} = m$ or that the probability of error is small, $\Pr[\hat{m} \neq m] \leq \epsilon$.

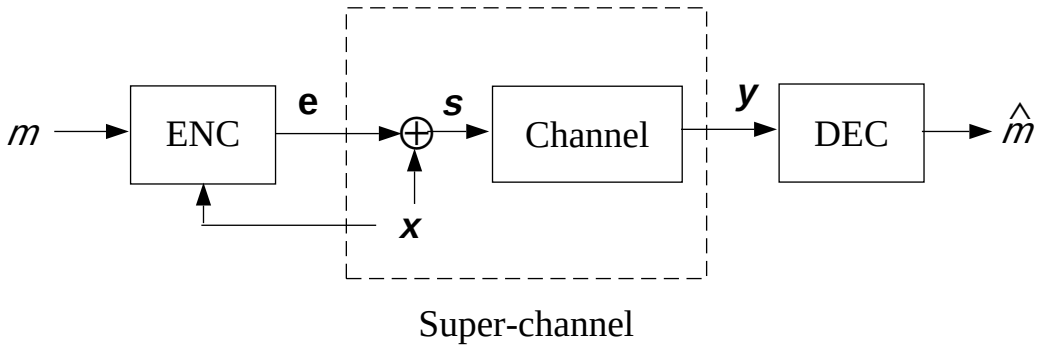

Figure 2-2: Equivalent super-channel model for information embedding. The composite signal is the sum of the host signal, which is the state of the super-channel, and a host-dependent distortion signal.

2.2 Equivalent Super-channel Model

An alternative representation of the model of Fig. 2-1 is shown in Fig. 2-2. The two models are equivalent since any embedding function $\mathbf{s}(\mathbf{x}, m)$ can be written 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 $\mathbf{e}(\mathbf{x}, m) \triangleq \mathbf{s}(\mathbf{x}, m) - \mathbf{x}$. Thus, one can view $\mathbf{e}$ as the input to a super-channel that consists of the cascade of an adder and the true channel. 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, squared error distortion (2.2) equals the power of $\mathbf{e}$,

$$\frac{1}{N}\|\mathbf{s} - \mathbf{x}\|^2 = \frac{1}{N}\|\mathbf{e}\|^2.$$

Therefore, one can view information embedding problems as power-limited communication over a super-channel with a state that is known at the encoder.[2] This view can be convenient for determining achievable rate-distortion-robustness trade-offs of various information embedding and decoding methods, as will become apparent in Chap. 4.

[2] Cox, *et al.*, have also recognized that one may view watermarking as communications with side information known at the encoder [17].

2.3 Channel Models

The channel model precisely describes the degradations that can occur to the composite signal. From this perspective, the channel is like a black box, to which one may or may not have access when formulating a model, and the objective of channel modeling is to describe the input/output relationship of this black box. From an alternative viewpoint, however, the channel model could simply describe the class of degradations against which one wishes the embedder and decoder to be robust, *i.e.*, the system is designed to work against all degradations described by this particular model. Although the difference between these two views may be subtle, it can be quite important when dealing with intentional attacks by a human attacker. From the first viewpoint, accurately describing all degradations that a human could possibly conceive using a tractable mathematical model could be quite difficult, if not impossible. However, from the second viewpoint, the channel model is more like a design specification: "Design an embedder and decoder that are robust against the following attacks."

Regardless of which viewpoint one adopts, in this thesis we describe the channel either probabilistically or deterministically. In the probabilistic case, we specify the channel input-output relationship in terms of the conditional probability law $p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s})$. Implicitly, this specification also describes the conditional probability law of the perturbation vectors against which the system must be robust since $p_{\mathbf{n}|\mathbf{s}}(\mathbf{n}|\mathbf{s}) = p_{\mathbf{y}|\mathbf{s}}(\mathbf{s}+\mathbf{n}|\mathbf{s})$. In the deterministic case, one can in general describe the channel input-output relationship in terms of the set of possible outputs $\mathcal{P}\{\mathbf{y}|\mathbf{s}\}$ for every given input, or equivalently, in terms of the set of desired tolerable perturbation vectors $\mathcal{P}\{\mathbf{n}|\mathbf{s}\}$ for every given input.

Some examples of families of such channel models are given below. These model families have parameters that naturally capture the severity of the associated set of perturbation vectors, and thus, these parameters also conveniently characterize the robustness of embedding methods as discussed in Sec. 2.1.

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,

$$\|\mathbf{y} - \mathbf{s}\|^2 = \|\mathbf{n}\|^2 \leq N\sigma_n^2. \tag{2.4}$$

This channel model, which describes a maximum distortion[3] 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. We also consider in this thesis the probabilistic counterpart to this channel, where $E[\|\mathbf{n}\|^2] \leq N\sigma_n^2$.

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 $D_{\mathbf{y}} = E[D(\mathbf{y}, \mathbf{x})]$, where this expectation is taken over the conditional probability density $p_{\mathbf{x}|\mathbf{s}}(\mathbf{x}|\mathbf{s})$.

3. **Additive noise channels:** In this case the perturbation vector $\mathbf{n}$ is modeled as random and statistically independent of $\mathbf{s}$. An additive white Gaussian noise (AWGN) channel is an example of such a channel, and the natural robustness measure in this case is the maximum noise variance σ_n^2 such that the probability of error is sufficiently low. These additive noise channel models may be appropriate for scenarios where one faces unintentional or incidental attacks, such as those that arise in hybrid transmission and some authentication applications. These models may also capture the effects of some lossy compression algorithms, as discussed in App. B.

[3]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 preprocessing 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.

2.4 Classes of Embedding Methods

An extremely large number of embedding methods have been proposed in the literature [20, 33, 41]. Rather than discussing the implementational details of this myriad of specific algorithms, in this section we focus our discussion on the common performance characteristics of broad classes of methods. One common way to classify watermarking algorithms is based on the types of host signals that the algorithms are designed to watermark [20, 41]. However, in this thesis we often examine watermarking at the highest, most fundamental level in which the host signal is viewed simply as a vector of numbers. At this level, the behavior (in terms of achievable rate-distortion-robustness trade-offs) of two audio watermarking algorithms may not necessarily be more alike than, say, the behavior of an audio watermarking algorithm and a similar video watermarking algorithm, although the measure of distortion and robustness may, of course, be different for video than for audio. Our classification system, therefore, is based on the types of behaviors that different watermarking systems exhibit as a result of the properties of their respective embedding functions. In particular, our taxonomy of embedding methods includes two classes: (1) host-interference non-rejecting methods and (2) host-interference rejecting methods.

2.4.1 Host-interference non-rejecting methods

A large number of embedding algorithms are designed based on the premise that the host signal is like a source of noise or interference. This view arises when one neglects the fact that the encoder in Fig. 2-2 has access to, and hence can exploit knowledge of, the host signal $\mathbf{x}$.

The simplest of this class have purely additive embedding functions of the form

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{w}(m), \tag{2.5}$$

where $\mathbf{w}(m)$ is typically a pseudo-noise sequence. Embedding methods in this class are often referred to as spread spectrum methods and some of the earliest examples are given by Tirkel, *et al.*[44, 45], Bender,[4] *et al.*[4], Cox, *et al.*[14, 15], and Smith and Comiskey [37].

[4]The "Patchwork" algorithm [4] of Bender, *et al.*, involves adding a small amount δ to some pseudo-randomly chosen host signal samples and subtracting a small amount δ from others. Thus, this method is equivalent to adding a pseudorandom sequence $\mathbf{w}(m)$ of $\pm\delta$ to the host signal, and hence, we consider the Patchwork algorithm to be a spread spectrum method.

From (2.5), we see that for this class of embedding methods, the host signal $\mathbf{x}$ acts as additive interference that inhibits the decoder's ability to estimate m. Consequently, even in the absence of any channel perturbations ($\mathbf{n} = \mathbf{0}$), one can usually embed only a small amount of information. Thus, these methods are useful primarily when either the host signal is available at the decoder or when the host signal interference is much smaller than the channel interference. Indeed, in [14] Cox, *et al.*, assume that $\mathbf{x}$ is available at the decoder.

The host-interference-limited performance of purely additive (2.5) embedding methods is embodied in Fig. 2-3 as the upper limit on rate of the dashed curve, which represents the achievable rate-robustness performance of host-interference non-rejecting methods, at a fixed level of embedding-induced distortion. Although the numerical values on the axes of Fig. 2-3 correspond to the case of white Gaussian host signals and additive white Gaussian noise channels, which is discussed in Chap. 6,[5] the upper rate threshold of the dashed curve is actually representative of the *qualitative* behavior of host-interference non-rejecting methods in general. Indeed, Su has derived a similar upper rate threshold for the case of so-called power-spectrum condition-compliant additive watermarks and Wiener attacks [39].

Many embedding methods exploit characteristics of human perceptual systems by adjusting the (squared error) distortion between $\mathbf{x}$ and $\mathbf{s}$ according to some perceptual model. When the model indicates that humans are less likely to perceive changes to $\mathbf{x}$, the host signal is altered a greater amount (in terms of squared error) than in cases when the perceptual model indicates that humans are more likely to perceive changes. One common method for incorporating these principles is to amplitude-weight the pseudo-noise vector $\mathbf{w}(m)$ in (2.5). The resulting embedding function is weighted-additive:

$$s_i(\mathbf{x}, m) = x_i + a_i(\mathbf{x}) w_i(m), \tag{2.6}$$

where the subscript i denotes the i-th element of the corresponding vector, *i.e.*, the i-th element of $\mathbf{w}(m)$ is weighted with an amplitude factor $a_i(\mathbf{x})$. An example of an embedding function within this class is proposed by Podilchuk and Zeng [34], where the amplitude

[5] To generate the curve, robustness is measured by the ratio in dB between noise variance and squared error embedding-induced distortion, the rate is the information-theoretic capacity (Eqs. (6.1) and (6.26) for host-interference rejecting and non-rejecting, respectively) in bits per host signal sample, and the ratio between the host signal variance and the squared error embedding-induced distortion is fixed at 20 dB.

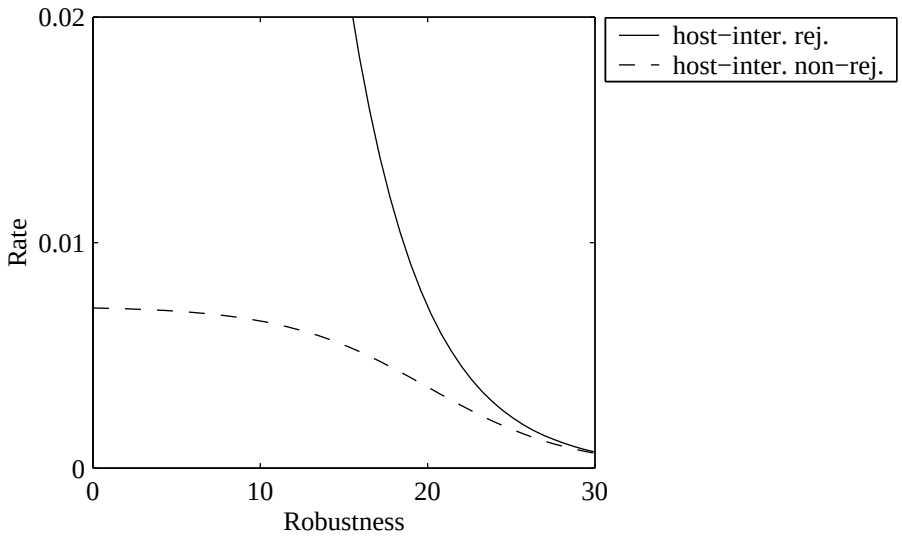

Figure 2-3: Qualitative behavior of host-interference rejecting (solid curve) and non-rejecting (dashed curve) embedding methods. The dashed curve's upper rate threshold at low levels of robustness (low levels of channel interference) indicates host-interference-limited performance.

factors $a_i(\mathbf{x})$ are set according to just noticeable difference (JND) levels computed from the host signal.

A special subclass of weighted-additive embedding functions, given in [14], arise by letting the amplitude factors be proportional to $\mathbf{x}$ so that

$$a_i(\mathbf{x}) = \lambda x_i,$$

where λ is a constant. Thus, these embedding functions have the property that large host signal samples are altered more than small host signal samples. This special subclass of embedding functions are purely additive in the log-domain since

$$s_i(\mathbf{x}, \boldsymbol{m}) = x_i + \lambda x_i w_i(\boldsymbol{m}) = x_i(1 + \lambda w_i(\boldsymbol{m}))$$

implies that

$$\log s_i(\mathbf{x}, \boldsymbol{m}) = \log x_i + \log(1 + \lambda w_i(\boldsymbol{m})).$$

Since the log function is invertible, if one has difficulty in recovering $\boldsymbol{m}$ from the composite signal in the log-domain due to host signal interference, then one must also encounter difficulty in recovering $\boldsymbol{m}$ from the composite signal in the non-log-domain. Thus,

host-proportional amplitude weighting also results in host signal interference, although the probability distributions of the interference $\log \mathsf{x}_i$ and of the watermark pseudo-noise $\log(1 + \lambda w_i(\mathsf{m}))$ are, of course, in general different than the probability distributions of x_i and $w_i(\mathsf{m})$. Although in the more general weighted-additive case (2.6), the encoder in Fig. 2-2 is not ignoring $\mathbf{x}$ since

$$e_i(\mathbf{x}, \mathsf{m}) = a_i(\mathbf{x}) w_i(\mathsf{m}),$$

in general unless the weighting functions $a_i(\mathbf{x})$ are explicitly designed to reject host interference in addition to exploiting perceptual models, host interference will still limit performance and thus this class of systems will still exhibit the qualitative behavior represented by the dashed curve in Fig. 2-3. We remark that in proposing the weighted-additive and log-additive embedding functions, Podilchuk and Zeng [34] and Cox, *et al.*[14], respectively, were actually considering the case where the host signal was available at the decoder, and hence, host interference was not an issue.

2.4.2 Host-interference rejecting methods

Having seen the inherent limitations of embedding methods that do not reject host interference by exploiting knowledge of the host signal at the encoder, we discuss in this section some examples of host-interference rejecting methods. In Chap. 3 we present a novel subclass of such host-interference rejecting methods called quantization index modulation (QIM). This QIM class of embedding methods exhibits the type of behavior illustrated by the solid curve in Fig. 2-3, while providing enough structure to allow the system designer to easily trade off rate, distortion, and robustness, *i.e.*, to move from one point on the solid curve of Fig. 2-3 to another.

Generalized low-bit modulation

Swanson, Zhu, and Tewfik [43] have proposed an example of a host-interference rejecting embedding method that one might call "generalized low-bit modulation (LBM)", although Swanson, *et al.*, do not use this term explicitly. The method consists of two steps: (1) linear projection onto a pseudorandom direction and (2) quantization-and-perturbation, as illustrated in Fig. 2-4. In the first step the host signal vector $\mathbf{x}$ is projected onto a pseudorandom

vector $\mathbf{v}$ to obtain

$$\tilde{x} = \mathbf{x}^T \mathbf{v}.$$

Then, information is embedded in $\tilde{x}$ by quantizing it with a uniform, scalar quantizer of step size Δ and perturbing the reconstruction point by an amount that is determined by m. (No information is embedded in components of $\mathbf{x}$ that are orthogonal to $\mathbf{v}$.) Thus, the projection $\tilde{s}$ of the composite signal onto $\mathbf{v}$ is

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

where $q(\cdot)$ is a uniform, scalar quantization function of step size Δ and $d(m)$ is a perturbation value, and the composite signal vector is

$$\mathbf{s} = \mathbf{x} + (\tilde{s} - \tilde{x})\mathbf{v}.$$

For example, suppose $\tilde{x}$ lies somewhere in the second quantization cell from the left in Fig. 2-4 and we wish to embed 1 bit. Then, $q(\tilde{x})$ is represented by the solid dot ($\bullet$) in that cell, $d(m) = \pm\Delta/4$, and $\tilde{s}$ will either be the $\times$-point (to embed a 0-bit) or the $\circ$-point (to embed a 1-bit) in the same cell. In [43] Swanson, *et al.*, note that one can embed more than 1 bit in the N-dimensional vector by choosing additional projection vectors $\mathbf{v}$. One could also, it seems, have only one projection vector $\mathbf{v}$, but more than two possible perturbation values $d(1), d(2), \ldots, d\left(2^{NR_m}\right)$.

We notice that all host signal values $\tilde{x}$ that map onto a given $\times$ point when a 0-bit is embedded also map onto the same $\circ$ point when a 1-bit is embedded. As a result of this condition, one can label the $\times$ and $\circ$ points with bit labels such that the embedding function is equivalent to low-bit modulation. Specifically, this quantization-and-perturbation process is equivalent to the following:

1. Quantize $\tilde{x}$ with a quantizer of step size $\Delta/2$ whose reconstruction points are the union of the set of $\times$ points and set of $\circ$ points. These reconstruction points have bit labels as shown in Fig. 2-4.

2. Modulate (replace) the least significant bit in the bit label with the watermark bit to arrive at a composite signal bit label. Set the composite signal projection value $\tilde{s}$ to the reconstruction point with this composite signal bit label.

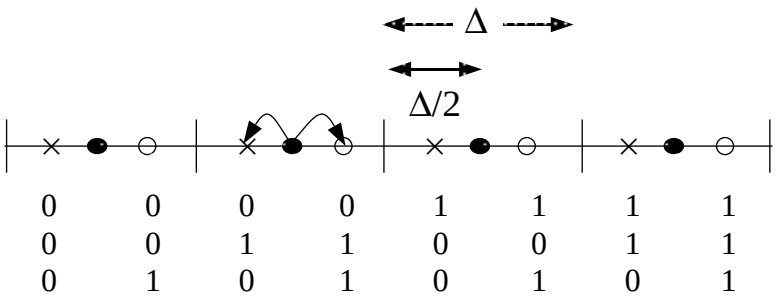

Figure 2-4: Equivalence of quantization-and-perturbation to low-bit modulation. Quantizing with step size Δ and perturbing the reconstruction point is equivalent to quantizing with step size $\Delta/2$ and modulating the least significant bit. In general, the defining property of low-bit modulation is that the embedding intervals for $\times$ points and $\circ$ points are the same.

Thus, the quantization-and-perturbation embedding method in [43] is low-bit modulation of the quantization of $\check{x}$.

An earlier paper [42] by Swanson, *et al.*, gives another example of generalized low-bit modulation, where a data bit is repeatedly embedded in the DCT coefficients of a block rather than in the projections onto pseudorandom directions. One can view the DCT basis vectors, then, as the projection vectors $\mathbf{v}$ in the discussion above. The actual embedding occurs through quantization and perturbation, which we now recognize as low-bit modulation.

Some people may prefer to use the term "low-bit modulation" only to refer to the modulation of the least significant bits of pixel values that are already quantized, for example, when the host signal is an 8-bit grayscale image. This corresponds to the special case when the vectors $\mathbf{v}$ are "standard basis" vectors, *i.e.*, $\mathbf{v}$ is a column of the identity matrix, and $\Delta = 2$. To emphasize that the quantization may occur in any domain, not just in the pixel domain, and that one may adjust the step size Δ to any desired value, we used the term "generalized LBM" above when first introducing the technique of Swanson, *et al.*. However, in this thesis the term LBM, even without the word "generalized" in front of it, refers to low-bit modulation in its most general sense.

In general, low-bit modulation can be defined by its *embedding intervals*, where the embedding interval $\mathcal{I}_m(\mathbf{s}_0)$ of a composite signal value $\mathbf{s}_0$ is the set of host signal values $\mathbf{x}$ that map onto $\mathbf{s}_0$ when embedding m, *i.e.*,

$$\mathcal{I}_m(\mathbf{s}_0) = \{\mathbf{x} | \mathbf{s}(\mathbf{x}, m) = \mathbf{s}_0\}.$$

Low-bit modulation embedding functions have the defining property that the set of embedding intervals corresponding to a given value of $\boldsymbol{m}$ are the same as the set of embedding intervals corresponding to all other values of $\boldsymbol{m}$, *i.e.*,

$$\{\mathcal{I}_i(\mathbf{s}_i)|\mathbf{s}_i \in \mathcal{S}_i\} = \{\mathcal{I}_j(\mathbf{s}_j)|\mathbf{s}_j \in \mathcal{S}_j\}, \qquad \forall i,j \in \left\{1,\ldots,2^{NR_m}\right\},$$

where $\mathcal{S}_i$ is the set of all possible composite signal values when $\boldsymbol{m} = i$. This point is discussed in more detail in Chap. 3 and illustrated in Fig. 3-3. For now, we return our attention to the special case of LBM with uniform, scalar quantization shown in Fig. 2-4.

Because the $\times$ and $\circ$ points in Fig. 2-4 are separated by some nonzero distance, we see that these LBM methods do, in fact, reject host-signal interference. The host signal $\check{\boldsymbol{x}}$ determines the particular $\times$ or $\circ$ point that is chosen as the composite signal value $\tilde{\boldsymbol{s}}$, but does not inhibit the decoder's ability to determine whether $\tilde{\boldsymbol{s}}$ is a $\times$ point or a $\circ$ point and, hence, to determine whether the embedded bit is a 0-bit or 1-bit.

However, LBM methods have the defining property that the embedding intervals for the $\times$ points and $\circ$ points are the same. This condition is an unnecessary constraint on the embedding function $\mathbf{s}(\mathbf{x}, \boldsymbol{m})$. As will become apparent throughout this thesis, by removing this constraint, one can find embedding functions that result in better rate-distortion-robustness performance than that obtainable by LBM.

Another host-interference rejecting method

Another host-interference rejecting method is disclosed in a recently issued patent [47]. Instead of embedding information in the quantization levels, information is embedded in the number of host signal "peaks" that lie within a given amplitude band. For example, to embed a 1-bit, one may force the composite signal to have exactly two peaks within the amplitude band. To embed a 0-bit, the number of peaks is set to less than two. Clearly, the host signal does not inhibit the decoder's ability to determine how many composite signal peaks lie within the amplitude band. The host signal does, however, affect the amount of embedding-induced distortion that must be incurred to obtain a composite signal with a given number of peaks in the amplitude band. For example, suppose the host signal has a large number of peaks in the amplitude band. If one tries to force the number of peaks in the band to be less than two in order to embed a 0-bit, then the distortion between

the resulting composite signal and host signal may be quite significant. Thus, even though this method rejects host-interference, it is not clear that it exhibits the desired behavior illustrated by the solid curve in Fig. 2-3. For example, to achieve a high rate when the channel noise is low, one needs to assign at least one number of signal peaks to represent $m = 1$, another number of signal peaks to represent $m = 2$, another number of signal peaks to represents $m = 3$, etc. Thus, one could potentially be required to alter the number of host signal peaks to be as low as 1 or as high as 2^{NR_m}. It is unclear whether or not one can alter the number of host signal peaks within the amplitude band by such a large amount without incurring too much distortion.

Quantization index modulation

As one can see, "bottom-up" approaches to digital watermarking abound in the literature in the sense that much of the literature is devoted to the presentation and evaluation of specific implementations of algorithms. One drawback of this approach is that by restricting one's attention to a particular algorithm, one imposes certain implicit structure and constraints on the embedding function and decoder, and often these constraints are not only unnecessary but also may lead to suboptimal performance. For example, if one restricts attention to an embedding function that happens to belong to the class of purely additive embedding functions (2.5), then host interference will inherently limit performance, as discussed above. Similarly, if one implements a method that can be classified as a low-bit modulation method, then one has implicitly imposed the constraint that the embedding intervals are invariant with respect to the watermark value.

In the next chapter, we take a "top-down" approach, where we examine watermarking from the highest, most fundamental level. Based on first principles we impose only enough structure as necessary to understand and control rate-distortion-robustness behavior. The result is a general class of host-interference rejecting embedding methods called quantization index modulation. As we show in Chap. 4, we incur no loss of optimality from an information-theoretic perspective by restricting attention to this class.

Chapter 3

Quantization Index Modulation

When one considers the design of an information-embedding system from a first principles point of view, so-called quantization index modulation (QIM) [6, 7] methods arise quite naturally, as we explain in this chapter. One can exploit the structure of these QIM methods to conveniently trade off rate, distortion, and robustness. Furthermore, as we shall see in later chapters, the QIM class is broad enough to include very good, and in some cases optimal, embedders and decoders, *i.e.*, there exist QIM methods that achieve the best possible rate-distortion-robustness trade-offs of any (QIM or non-QIM) method. We devote the rest of this chapter to describing the basic principles behind QIM.

3.1 Basic Principles

In Chap. 2, we considered 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 (2.1), the rate R_m determines the number of possible values for m, and hence, the number of functions in the ensemble. If the embedding-induced distortion is to be small, then each function in the ensemble must be close to an identity function in some sense so that

$$\mathbf{s}(\mathbf{x}; m) \approx \mathbf{x}, \qquad \forall m. \tag{3.1}$$

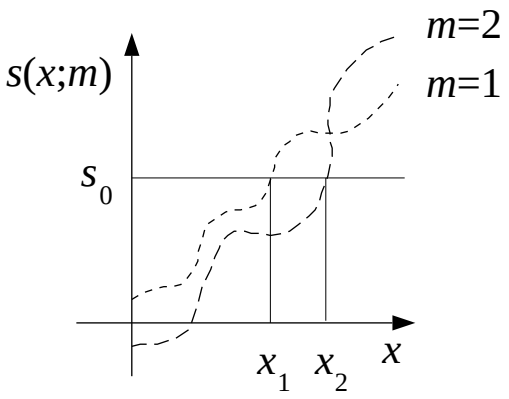

Figure 3-1: Embedding functions with intersecting ranges. The point s_0 belongs to the ranges of both continuous embedding functions. Thus, even with no perturbations ($y = s_0$) the decoder cannot distinguish between $m = 1$ (and $x = x_1$) and $m = 2$ (and $x = x_2$). Using discontinuous functions allows one to make the ranges non-intersecting.

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. For example, one might desire at the very least that the ranges 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, as illustrated in Fig. 3-1. This non-intersection property along with the approximate-identity property (3.1), which suggests that the ranges of each of the functions "cover" the space of possible (or at least highly probable) host signal values $\mathbf{x}$, 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. 3-2 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. 3-2 with $\times$'s and $\circ$'s. If $m = 1$, for example, the host signal is quantized with the $\times$-quantizer, *i.e.*, $\mathbf{s}$ is chosen to be the $\times$ closest to $\mathbf{x}$. If $m = 2$, $\mathbf{x}$ is quantized with the $\circ$-quantizer. Here, we see the non-intersecting nature of the ranges of the two quantizers as no $\times$ point is the same as any $\circ$ point. This non-intersection property leads to host-signal interference rejection. As $\mathbf{x}$ varies, the composite signal value $\mathbf{s}$ varies from one $\times$ point ($m = 1$) to another or

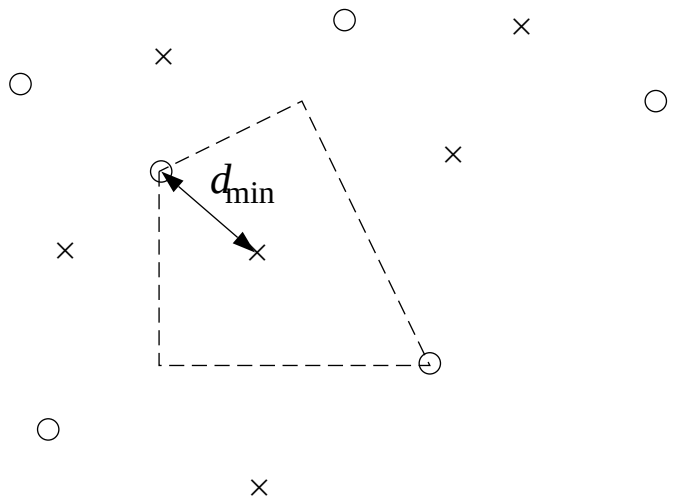

Figure 3-2: Quantization index modulation for information embedding. The points marked with $\times$'s and $\circ$'s belong to two different quantizers, each with its associated index. The minimum distance $d_{\min}$ measures the robustness to perturbations, and the sizes of the quantization cells, one of which is shown in the figure, determine the distortion. If $m = 1$, the host signal is quantized to the nearest $\times$. If $m = 2$, the host signal is quantized to the nearest $\circ$.

from one $\circ$ point ($m = 2$) to another, but it never varies between a $\times$ point and a $\circ$ point. Thus, even with an infinite energy host signal, one can determine m if channel perturbations are not too severe. We also see the discontinuous nature of the quantizers. The dashed polygon represents the quantization cell for the $\times$ 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 $\times$ in the cell interior to a $\times$ in the cell exterior. The $\times$ points and $\circ$ points are both quantizer reconstruction points and signal constellation points,[1] and we may view design of QIM systems as the simultaneous design of an ensemble of source codes (quantizers) and channel codes (signal constellations).

The structure of QIM systems is convenient from an engineering perspective since properties of the quantizer ensemble can be connected to the performance parameters of rate, distortion, and robustness. For example, as noted above the number of quantizers in the ensemble determines the information-embedding rate. The sizes and shapes of the quantization cells determine the embedding-induced distortion, all of which arises from quantization error. Finally, for many classes of channels, the minimum distance $d_{\min}$ between the sets of reconstruction points of different quantizers in the ensemble determines the robustness of

[1] One *set* of points, rather than one individual point, exists for each value of m.

the embedding. We define the minimum distance to be

$$d_{\min} \stackrel{\Delta}{=} \min_{(i,j):i\neq j} \min_{(\mathbf{x}_i,\mathbf{x}_j)} \|\mathbf{s}(\mathbf{x}_i;i) - \mathbf{s}(\mathbf{x}_j;j)\|. \tag{3.2}$$

Alternatively, if the host signal is known at the decoder, as is the case in some applications of interest, then the relevant minimum distance may be more appropriately defined as either

$$d_{\min}(\mathbf{x}) \stackrel{\Delta}{=} \min_{(i,j):i\neq j} \|\mathbf{s}(\mathbf{x};i) - \mathbf{s}(\mathbf{x};j)\|, \tag{3.3}$$

or

$$d_{\min} \stackrel{\Delta}{=} \min_{\mathbf{x}} \min_{(i,j):i\neq j} \|\mathbf{s}(\mathbf{x};i) - \mathbf{s}(\mathbf{x};j)\|. \tag{3.4}$$

The important distinction between the definition of (3.2) and the definitions of (3.3) and (3.4) is that in the case of (3.3) and (3.4) the decoder knows $\mathbf{x}$ and, thus, needs to decide only among the reconstruction points of the various quantizers in the ensemble corresponding to the particular value of $\mathbf{x}$. In the case of (3.2), however, the decoder needs to choose from all reconstruction points of the quantizers.

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 (2.4) implies that a minimum distance decoder is guaranteed to not make an error as long as

$$\frac{d_{\min}^2}{4N\sigma_n^2} > 1. \tag{3.5}$$

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 [26],

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

where $Q(\cdot)$ is the Gaussian Q-function,

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt. \tag{3.6}$$

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

$$\hat{m}(\mathbf{y}) = \arg\min_m \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{s}(\mathbf{x}; m)\|. \tag{3.7}$$

If, which is often the case, the quantizers $\mathbf{s}(\mathbf{x}; m)$ map $\mathbf{x}$ to the nearest reconstruction point, then (3.7) can be rewritten as

$$\hat{m}(\mathbf{y}) = \arg\min_m \|\mathbf{y} - \mathbf{s}(\mathbf{y}; m)\|. \tag{3.8}$$

Alternatively, if the host signal $\mathbf{x}$ is known at the decoder,

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

For general deterministic channels $\mathcal{P}\{\mathbf{y}|\mathbf{s}\}$ to guarantee error-free decoding, one needs to place the quantizer reconstruction points such that the sets of possible channel outputs for different values of $\boldsymbol{m}$ are non-intersecting, *i.e.*,

$$\left(\bigcup_{\mathbf{s}\in\mathcal{S}_i} \mathcal{P}\{\mathbf{y}|\mathbf{s}\}\right) \cap \left(\bigcup_{\mathbf{s}\in\mathcal{S}_j} \mathcal{P}\{\mathbf{y}|\mathbf{s}\}\right) = \emptyset, \qquad \forall i \neq j, \tag{3.9}$$

where, $\mathcal{S}_i$ again represents the set of all possible composite signal values when $\boldsymbol{m} = i$. In the case of the bounded perturbation channel, these sets of possible channel outputs are unions of spheres of radius $\sigma_n\sqrt{N}$ centered around each reconstruction point, and the non-intersection condition (3.9) reduces to the condition (3.5).

A natural alternative decoder to the minimum-distance decoder (3.7) in the general deterministic case is what one might call the possible-set decoder:

$$\hat{\boldsymbol{m}} = i, \quad \text{if } \mathbf{y} \in \textstyle\bigcup_{\mathbf{s}\in\mathcal{S}_i} \mathcal{P}\{\mathbf{y}|\mathbf{s}\},$$

assuming there is only one such i. Otherwise, an error is declared. Similarly, for general

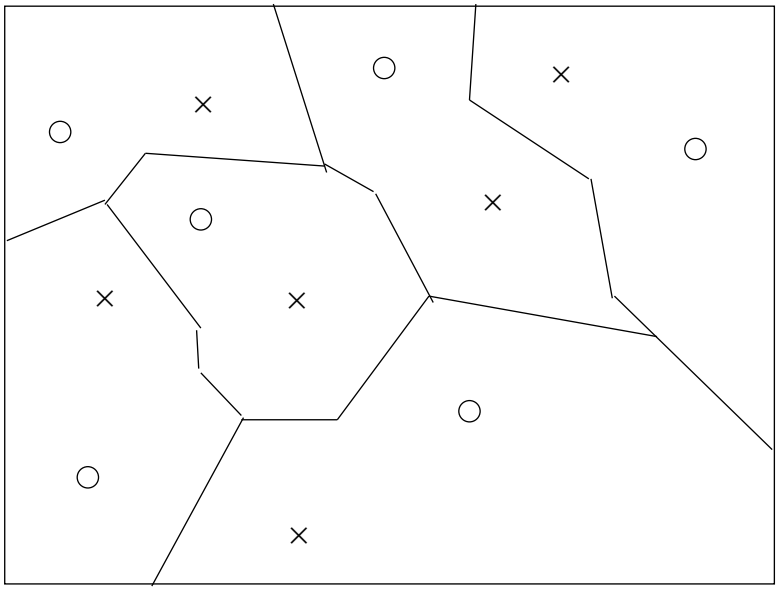

Figure 3-3: Embedding intervals of low-bit modulation. The $\times$ and $\circ$ points within an embedding interval (coarse cell), which is the union of two finer cells (not shown), differ in only the least significant bit. Thus, one can view the points as reconstruction points of two different coarse quantizers, each having the embedding intervals as quantization cells.

probabilistic channels $p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s})$, one might use the generalized maximum likelihood (ML) decoder

$$\hat{m} = \arg\max_i \max_{\mathbf{x}} p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s}(\mathbf{x}; i))$$

if the host signal $\mathbf{x}$ is deterministic, but unknown at the decoder. This decoder is, of course, the same as the minimum distance decoder (3.7) for the additive white Gaussian noise channel. If the host signal is random, then one might use the maximum likelihood decoder

$$\hat{m} = \arg\max_m \sum_{\mathbf{s}_i \in \mathcal{S}_m} \Pr[\mathbf{x} \in \mathcal{R}_i(m)] p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s}_i),$$

where $\mathcal{R}_i(m)$ is the i-th quantization cell of the m-th quantizer, which has $\mathbf{s}_i \in \mathcal{S}_m$ as a reconstruction point.

3.2 QIM vs. Generalized LBM

Although generalized LBM systems have nonzero minimum distance, there always exists a QIM system whose achievable performance is at least as good as, and usually better than, any given generalized LBM system. This concept is illustrated by Fig. 3-3. The $\times$ and $\circ$

points are reconstruction points of an LBM quantizer that is used to quantize the host signal. One bit is embedded by modulating the least significant bit (lsb). After modulation of the lsb, the corresponding reconstruction point is the composite signal. The $\times$ points represent reconstruction points that have a lsb of 0, and $\circ$ points represent points that have a lsb of 1. The unions of the two quantization cells corresponding to reconstruction points that differ only in their lsb are also shown in Fig. 3-3. We refer to these regions as coarse quantization cells. Due to modulation of the lsb, any host signal point within a given coarse quantization cell may be mapped to either the $\times$ point or the $\circ$ point within the coarse cell. Hence, these coarse quantization cells are the embedding intervals of the $\times$ and $\circ$ points contained within them. One may also view the $\times$ points and $\circ$ points as the reconstruction points of two different quantizers in an equivalent QIM system. These two quantizers have the same set of quantization cells, the coarse quantization cells. Clearly, then, this QIM system achieves the same performance as the LBM system. In general, though, the quantizers in a QIM system need not have the same quantization cells. Keeping the same reconstruction points as in Fig. 3-3, which preserves the minimum distance between quantizers, but exploiting the freedom to choose different quantization cells for the two quantizers usually results in lower embedding-induced distortion (except in rare, degenerate cases), and thus, the resulting QIM system achieves better rate-distortion-robustness performance than the original LBM system.

Another way of seeing the advantages of QIM over generalized LBM is shown in Figs. 3-4 and 3-5, which show one-dimensional embedding functions for embedding one bit in one host signal sample using LBM and QIM, respectively, with uniform scalar quantizers. Although the minimum distances (3.2) in the two cases are the same ($d_{\min} = 1/2$), the two functions $s(x;1)$ and $s(x;2)$ in the QIM case more closely approximate the identity function than the two functions in the LBM case, and thus the embedding-induced distortion in the QIM case is smaller than in the LBM case. This difference is quantified in Sec. 5.2.3.

3.3 Distortion Compensation

Distortion compensation is a type of post-quantization processing that can improve the achievable rate-distortion-robustness trade-offs of QIM methods. Indeed, with distortion compensation one can achieve the information-theoretically best possible rate-distortion-

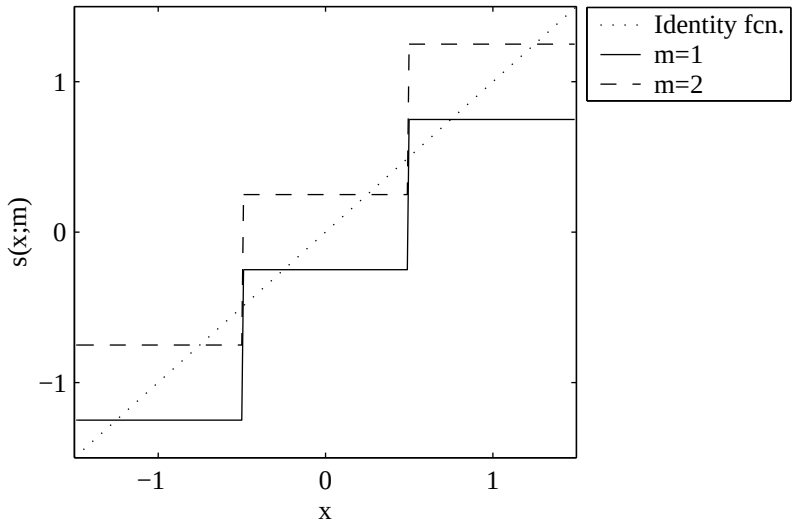

Figure 3-4: Embedding functions for LBM with uniform scalar quantization. Each of the two approximate-identity functions have a bias relative to the identity function, which increases the embedding-induced distortion.

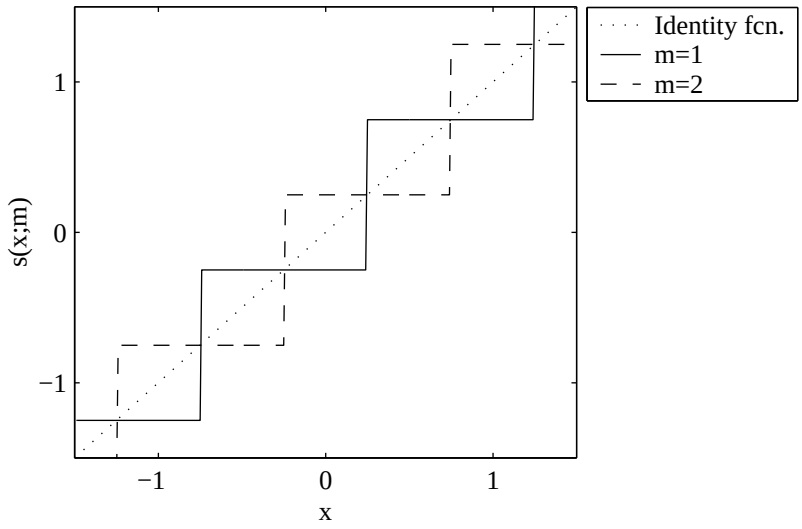

Figure 3-5: Embedding functions for QIM with uniform scalar quantization. The two approximate-identity functions do not share the same embedding intervals and thus more closely approximate the identity function than do the LBM approximate-identity functions.

robustness performance in many important cases, as discussed in Chaps. 4, 6, and 7. We explain the basic principles behind distortion compensation in this section.

As explained above, increasing the minimum distance between quantizers leads to greater robustness to channel perturbations. For a fixed rate and a given quantizer ensemble, scaling[2] all quantizers by $\alpha \leq 1$ increases $d^2_{\min}$ by a factor of $1/\alpha^2$. However, the embedding-induced distortion also increases by a factor of $1/\alpha^2$. Adding back a fraction $1-\alpha$ of the quantization error to the quantization value removes, or compensates for, this additional distortion. The resulting embedding function is

$$\mathbf{s}(\mathbf{x}, \boldsymbol{m}) = \mathbf{q}(\mathbf{x}; \boldsymbol{m}, \Delta/\alpha) + (1-\alpha)[\mathbf{x} - \mathbf{q}(\mathbf{x}; \boldsymbol{m}, \Delta/\alpha)], \tag{3.10}$$

where $\mathbf{q}(\mathbf{x}; \boldsymbol{m}, \Delta/\alpha)$ is the $\boldsymbol{m}$-th quantizer of an ensemble whose reconstruction points have been scaled by α so that two reconstruction points separated by a distance Δ before scaling are separated by a distance Δ/α after scaling. The first term in (3.10) represents normal QIM embedding. We refer to the second term as the distortion-compensation term.

Typically, the probability density functions of the quantization error for all quantizers in the QIM ensemble are similar. Therefore, the distortion compensation term in (3.10) is statistically independent or nearly statistically independent of $\boldsymbol{m}$ and can be treated as noise or interference during decoding. Thus, decreasing α leads to greater minimum distance, but for a fixed embedding-induced distortion, the distortion-compensation interference at the decoder increases. One optimality criterion for choosing α is to maximize a "signal-to-noise ratio (SNR)" at the decision device,

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

where this SNR is defined as the ratio between the squared minimum distance between quantizers and the total interference energy from both distortion-compensation interference and channel interference. Here, d_1 is the minimum distance when $\alpha = 1$ and is a characteristic of the particular quantizer ensemble. One can easily verify that the optimal scaling

[2] If a reconstruction point is at $\mathbf{q}$, it is "scaled" by α by moving it to $\mathbf{q}/\alpha$.

parameter α that maximizes this SNR is

$$\alpha_{\mathrm{SNR}} = \frac{\mathrm{DNR}}{\mathrm{DNR}+1}, \tag{3.11}$$

where DNR is the (embedding-induced) distortion-to-noise ratio $D_{\mathbf{s}}/\sigma_n^2$. Such a choice of α also maximizes the information-theoretically achievable rate when the channel is an additive Gaussian noise channel and the host signal $\mathbf{x}$ is Gaussian, as discussed in Chap. 6. Finally, as discussed in Chap. 7, one can asymptotically achieve capacity with this choice of α in the high-fidelity limit of small embedding-induced distortion and small perturbation energy.

Chapter 4

Information Theoretic Perspectives

In this chapter 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 "hidden QIM" 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, we derive sufficient conditions under which only distortion compensation postprocessing is required. As we discuss in Chaps. 6 and 7, these conditions are satisfied in the following three important cases: (1) an additive Gaussian noise channel and a Gaussian host signal, (2) squared error distortion-constrained attacks and a Gaussian host signal, and (3) squared error distortion-constrained attacks, a non-Gaussian host signal, and asymptotically small embedding-induced distortion and attacker's distortion.

4.1 Communication over Channels with Side Information

The super-channel model of Sec. 2.2 and Fig. 2-2 facilitates our analysis, *i.e.*, we view information embedding as the transmission of a host-dependent distortion signal $\mathbf{e}$ over a super-channel with side information or state $\mathbf{x}$ that is known at the encoder. In this chapter

we also assume a squared error distortion constraint

$$\frac{1}{N}\sum_{i=1}^{N} \mathsf{e}^2 \leq D_{\mathbf{s}}.$$

and a memoryless channel with known probability density function (pdf)

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

where y_i and s_i are the i-th components of $\mathbf{y}$ and $\mathbf{s}$, respectively.[1] Then, the super-channel is also memoryless and has probability law

$$p_{\mathbf{y}|\mathbf{e},\mathbf{x}}(\mathbf{y}|\mathbf{e},\mathbf{x}) = p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{x}+\mathbf{e}) = \prod_{i=1}^{N} p_{y|\mathsf{s}}(y_i|x_i+e_i) = \prod_{i=1}^{N} p_{y|\mathsf{e},x}(y_i|e_i,x_i).$$

The capacity [13] of this super-channel is the reliable information-embedding rate R_m that is asymptotically achievable with long signal lengths N.

In non-watermarking contexts Gel'fand and Pinsker [19] and Heegard and El Gamal [21] have determined the capacity of such a channel in the case of a random state vector $\mathbf{x}$ with independent and identically distributed (iid) components when the encoder sees the entire state vector before choosing the channel input $\mathbf{e}$. In this case the capacity is

$$C = \max_{p_{u,e|x}(u,e|x)} I(u;y) - I(u;x), \tag{4.1}$$

where $I(\cdot;\cdot)$ denotes mutual information and u is an auxiliary random variable. In the case of watermarking, the maximization (4.1) is subject to a distortion constraint $E[\mathsf{e}^2] \leq D_{\mathbf{s}}$. A formal proof of the extension of (4.1) to include the distortion constraint has been given by Barron [2, 3]. Others [9, 30] are working on extending or have extended these results to the case where the channel law $p_{y|\mathsf{s}}(y|s)$ is not fixed but rather is chosen by an attacker subject to a distortion constraint. A related information-theoretic formulation can be found in [31].

As we shall see in the next section, one way to interpret (4.1) is that $I(u;y)$ is the total number of bits per host signal sample that can be transmitted through the channel

[1] Extension of results in this chapter to the case where the channel is only blockwise memoryless is straightforward by letting y_i and s_i be the i-th blocks, rather than i-th scalar components, of $\mathbf{y}$ and $\mathbf{s}$. In this case, information rates are measured in bits per block, rather than bits per sample.

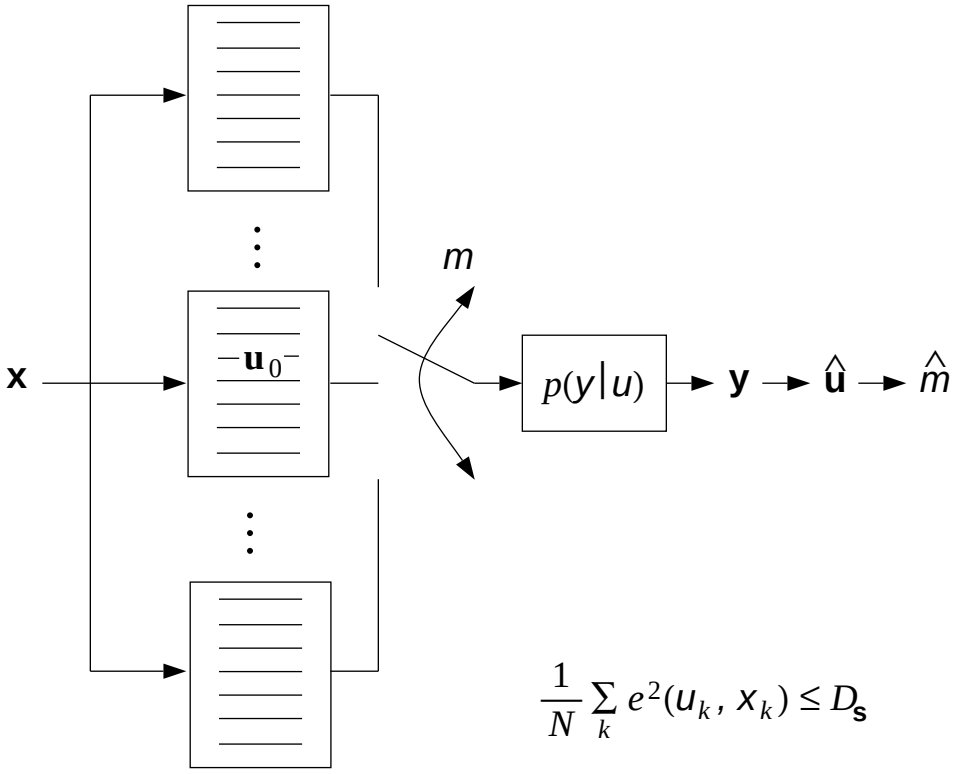

Figure 4-1: Capacity-achieving "hidden QIM". One embeds by choosing a codeword $\mathbf{u}_0$ that is jointly distortion-typical with $\mathbf{x}$ from the m-th quantizer's codebook. The distortion function is $e^2(u, x)$. The decoder finds a codeword that is jointly typical with $\mathbf{y}$. If this codeword is in the i-th subset, then $\hat{m} = i$.

and $I(u; x)$ is the number of bits per sample that are allocated to the host signal x. The difference between the two is the number of bits per host signal sample that can be allocated to the embedded information m.

4.1.1 Optimality of "hidden" QIM

As we show in this section, one can achieve the capacity (4.1) 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 the proof of the achievability of capacity by Gel'fand and Pinsker [19], with added interpretation in terms of quantization (source coding). Fig. 4-1 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 maximizing conditional distribution $p_{u,e|x}(u,e|x)$ from (4.1). 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 $\mathbf{u}$-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[2]

$$\mathbf{e}(\mathbf{u}_0, \mathbf{x}) = [e(u_{0,1}, x_1) \ \cdots \ e(u_{0,N}, x_N)]^T.$$

By distortion-typical, we mean that $\mathbf{u}_0$ and $\mathbf{x}$ are jointly typical and $\|\mathbf{e}(\mathbf{u}_0, \mathbf{x})\|^2 \leq N(D_{\mathbf{s}} + \epsilon)$, *i.e.*, the function $e^2(u,x)$ is the distortion function in the $\mathbf{u}$-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 principle is one of the main ideas behind the rate-distortion theorem [13, Ch. 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 principles are two of the main ideas behind the classical channel coding theorem [13, Ch. 8].) Thus, the probability of error $\Pr[\hat{m} \neq m]$ is small, and we can indeed achieve the capacity (4.1) with QIM in the $\mathbf{u}$-domain.

The remaining challenge, therefore, is to determine the right preprocessing and post-processing 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.

[2] From convexity properties of mutual information, one can deduce that the maximizing distribution in (4.1) always has the property that e is a deterministic function of (u,x) [19].

4.1.2 Optimality of distortion-compensated QIM

We show in this section that distortion-compensated QIM (DC-QIM) can achieve capacity whenever the maximizing distribution $p_{u,e|x}(u,e|x)$ in (4.1) is of a form such that

$$u = e + \alpha x. \tag{4.2}$$

As mentioned in the introduction to this chapter, this condition is satisfied in at least three important cases, which are discussed in Chaps. 6 and 7.

To see that DC-QIM can achieve capacity when the maximizing pdf in (4.1) satisfies (4.2), we show that one can construct an ensemble of random DC-QIM codebooks that satisfy (4.2). First, 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.*,

$$\mathbf{q}(\mathbf{x}; m, \Delta/\alpha) = \frac{1}{\alpha}\mathbf{q}(\alpha\mathbf{x}; m, \Delta). \tag{4.3}$$

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 the DC-QIM embedding function (3.10) and substituting (4.3) into the result, we obtain

$$\begin{aligned} \mathbf{s}(\mathbf{x}, m) &= \mathbf{q}(\mathbf{x}; m, \Delta/\alpha) + (1-\alpha)[\mathbf{x} - \mathbf{q}(\mathbf{x}; m, \Delta/\alpha)] \\ &= \alpha\mathbf{q}(\mathbf{x}; m, \Delta/\alpha) + (1-\alpha)\mathbf{x} \\ &= \mathbf{q}(\alpha\mathbf{x}; m, \Delta) + (1-\alpha)\mathbf{x}. \end{aligned} \tag{4.4}$$

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 (4.2). (Equivalently, we choose the codewords of $\mathbf{q}(\cdot; m, \Delta/\alpha)$ in (3.10) 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. 4.1.1, we can

achieve a rate $I(u;y) - I(u;x)$. From (4.4), 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 (4.1) satisfies (4.2), our DC-QIM codebooks can also have this distribution and, hence, achieve capacity (4.1).

4.2 Noise-free Case

In the noise-free case ($\mathbf{y} = \mathbf{s}$), which arises, for example, when a discrete-valued composite signal is transmitted over a digital channel with no errors, QIM is optimal even without distortion compensation, *i.e.*, one can achieve capacity with $u = x + e = s$. To see this, we first note that the rate $R_m = H(y|x)$ is achievable with $u = s$ since

$$\begin{aligned} R_m &= I(u;y) - I(u;x) \\ &= I(y;y) - I(y;x) \\ &= H(y) - [H(y) - H(y|x)] \\ &= H(y|x), \end{aligned} \tag{4.5}$$

where we have used $u = s = y$ in the second line. Now, we shall show that the capacity (4.1) cannot exceed $H(y|x)$:

$$\begin{aligned} I(u;y) - I(u;x) &= H(u) - H(u|y) - H(u) + H(u|x) \\ &= H(u|x) - H(u|y) \\ &\leq H(u|x) - H(u|y,x) \\ &= I(u;y|x) \\ &= H(y|x) - H(y|u,x) \\ &\leq H(y|x). \end{aligned} \tag{4.6}$$

The third line follows since conditioning decreases entropy. The final line arises since entropy is nonnegative. Thus, we see that QIM is optimal in the noise-free case and achieves the

capacity

$$C_{\text{noise-free}} = \max_{p_{\mathsf{e}|\mathsf{x}}(e|x)} H(\mathsf{y}|\mathsf{x}), \tag{4.7}$$

where we have replaced a maximization over $p_{\mathsf{y}|\mathsf{x}}(y|x)$ ($H(\mathsf{y}|\mathsf{x})$ depends only on $p_{\mathsf{y}|\mathsf{x}}(y|x)p_{\mathsf{x}}(x)$ and $p_{\mathsf{x}}(x)$ is given.) with an equivalent maximization over $p_{\mathsf{e}|\mathsf{x}}(e|x) = p_{\mathsf{y}|\mathsf{x}}(x+e|x)$ since $\mathsf{y} = \mathsf{s} = \mathsf{x} + \mathsf{e}$.

4.3 Known-host Case

In some information embedding applications the host signal may be available at the decoder, and one may view the known-host information embedding problem as one of communication with side information known both at the encoder and decoder, a scenario for which Heegard and El Gamal have determined the capacity to be [21]

$$C_{\text{known}} = \max_{p_{\mathsf{e}|\mathsf{x}}(e|x)} I(\mathsf{e};\mathsf{y}|\mathsf{x}). \tag{4.8}$$

Once again, one can achieve this capacity with QIM in the $\mathbf{u}$-domain, except that the total number of codewords $\mathbf{u}$ is $2^{N[I(u;y,x)-\epsilon]}$ instead of $2^{N[I(u;y)-\epsilon]}$ and decoding involves finding a $\mathbf{u}$ that is jointly typical with the pair $(\mathbf{x},\mathbf{y})$ rather than with only $\mathbf{y}$. (There are still $2^{N[I(u;x)+\epsilon]}$ codewords per quantizer.) Thus, the achievable rate is

$$\begin{aligned} I(\mathsf{u};\mathsf{y},\mathsf{x}) - I(\mathsf{u};\mathsf{x}) - 2\epsilon &= I(\mathsf{u};\mathsf{x}) + I(\mathsf{u};\mathsf{y}|\mathsf{x}) - I(\mathsf{u};\mathsf{x}) - 2\epsilon \\ &= H(\mathsf{y}|\mathsf{x}) - H(\mathsf{y}|\mathsf{u},\mathsf{x}) - 2\epsilon \\ &= H(\mathsf{y}|\mathsf{x}) - H(\mathsf{y}|\mathsf{u},\mathsf{x},\mathsf{e}) - 2\epsilon \\ &= H(\mathsf{y}|\mathsf{x}) - H(\mathsf{y}|\mathsf{x},\mathsf{e}) - 2\epsilon \\ &= I(\mathsf{e};\mathsf{y}|\mathsf{x}) - 2\epsilon, \end{aligned}$$

where the first line follows from the chain rule for mutual information [13, Sec. 2.5], the third line since e is a deterministic function of x and u, and the fourth line from the fact that $\mathsf{u} \to (\mathsf{x},\mathsf{e}) \to \mathsf{y}$ forms a Markov chain.

Thus, we see that with the appropriate choice of domain, or equivalently with the appropriate preprocessing and postprocessing, QIM is optimal in the sense that capacity-

achieving QIM systems exist. In the next chapter we discuss practical implementations of QIM with reasonable delay and complexity.

4.4 Conditions for Equivalence of Host-blind and Known-host Capacities

Before we discuss practical implementations, however, we derive in this section a necessary and sufficient condition for the equivalence of the host-blind and known-host capacities.[3] When this condition is satisfied, information embedding methods exist that completely reject host interference since no loss in rate-distortion-robustness performance results from not having the host signal at the decoder.

To derive this equivalence condition, we write the following equalities and inequalities, where all mutual informations and entropy expressions are with respect to the *host-blind* capacity-achieving distribution, the maximizing distribution in (4.1):

$$\begin{aligned} C_{\text{known}} &\geq I(e;y|x) && (4.9)\\ &\geq I(u;y|x) && (4.10)\\ &= H(u|x) - H(u|y,x) \\ &\geq H(u|x) - H(u|y) && (4.11)\\ &= H(u) - H(u|y) - [H(u) - H(u|x)] \\ &= I(u;y) - I(u;x) \\ &= C_{\text{host-blind}}. \end{aligned}$$

The first line follows from (4.8) and the second from the Data Processing Inequality [13, Sec. 2.8] since $u \to e \to y$ is a Markov chain given x. The fourth line arises since conditioning never increases entropy. The final line arises since all mutual informations are with respect to the maximizing distribution in (4.1). Thus, we obtain the obvious result that the host-blind capacity cannot be greater than the known-host capacity.

We arrive at conditions for equivalence of the two capacities by finding necessary and sufficient conditions for the inequalities (4.9), (4.10), and (4.11) to hold with equality. The

[3] The results we report in this section are from joint work with Richard Barron of MIT [3].

Data Processing Inequality (4.10) holds with equality if and only if

$$I(\mathsf{y};\mathsf{e}|\mathsf{u},\mathsf{x}) = 0. \tag{4.12}$$

This condition is always satisfied since the maximizing distribution in (4.1) has the property that e is a deterministic function of (u,x) [19]. The expression (4.9) holds with equality if and only if the conditional pdf

$$p_{\mathsf{e}|\mathsf{x}}(e|x) = \sum_{u} p_{\mathsf{u},\mathsf{e}|\mathsf{x}}(u,e|x)$$

corresponding to the maximizing distribution in (4.1) is also the maximizing distribution in (4.8). The final inequality (4.11) holds with equality if and only if $H(\mathsf{u}|\mathsf{y},\mathsf{x}) = H(\mathsf{u}|\mathsf{y})$, or equivalently if and only if x and u are conditionally independent given y,

$$I(\mathsf{x};\mathsf{u}|\mathsf{y}) = 0. \tag{4.13}$$

An intuitive interpretation for this condition is that if this condition holds, observing $\mathbf{x}$ does not give any more information than that obtained from observing $\mathbf{y}$ alone about which codeword $\mathbf{u}$ was chosen by the encoder. Since the decoder estimates m by determining the codebook to which $\mathbf{u}$ belongs, if (4.13) holds, then the decoder's job is not made any easier if it is allowed to observe the host signal x.

Chapter 5

Dither Modulation

Viewing an embedding function as an ensemble of approximate identity functions and restricting these approximate identity functions to be quantizers leads to a convenient structure in which one can achieve rate-distortion-robustness trade-offs by adjusting the number of quantizers, the quantization cells, and the minimum distance, as discussed in Chap. 3. This structure reduces the information-embedding system design problem to one of constructing an ensemble of quantizer reconstruction points that also form a good signal constellation. Furthermore, as discussed in Chap. 4 imposing such structure need not lead to any loss of optimality provided one chooses the proper domain for quantization and, as we discuss in Chaps. 6 and 7, when used in conjunction with distortion compensation (Sec. 3.3), such structure need not result in any loss of optimality in certain important cases even when quantizing in the composite-signal domain.

Imposing additional structure on the quantizer ensembles themselves leads to additional insights into the design, performance evaluation, and implementation of QIM embedding methods, particularly when one is concerned with low-complexity implementations. A convenient structure to consider is that of so-called dithered quantizers [23, 49], 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 "dither modulation". We discuss several low-complexity realizations of such dither modulation methods in the rest of this chapter.

5.1 Coded Binary Dither Modulation with Uniform Scalar Quantization

Coded binary dither modulation with uniform, scalar quantization is one such realization. (By scalar quantization, we mean that the high dimensional base quantizer $\mathbf{q}(\cdot)$ is the Cartesian product of scalar quantizers.) 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 $k_u/k_c = 1$.) We divide the host signal $\mathbf{x}$ into N/L non-overlapping blocks of length L and embed the i-th coded bit z_i in the i-th block, as described below.

- Two length-L dither sequences $d[k, 0]$ and $d[k, 1]$ and one length-L sequence of uniform, scalar quantizers with step sizes $\Delta_1, \ldots, \Delta_L$ are constructed with the constraint
$$d[k, 1] = \begin{cases} d[k, 0] + \Delta_k/2, & d[k, 0] < 0 \\ d[k, 0] - \Delta_k/2, & d[k, 0] \geq 0 \end{cases}, \quad k = 1, \ldots, L,$$
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 $\pm\Delta_k/4$ and its negative satisfy this constraint. One could alternatively

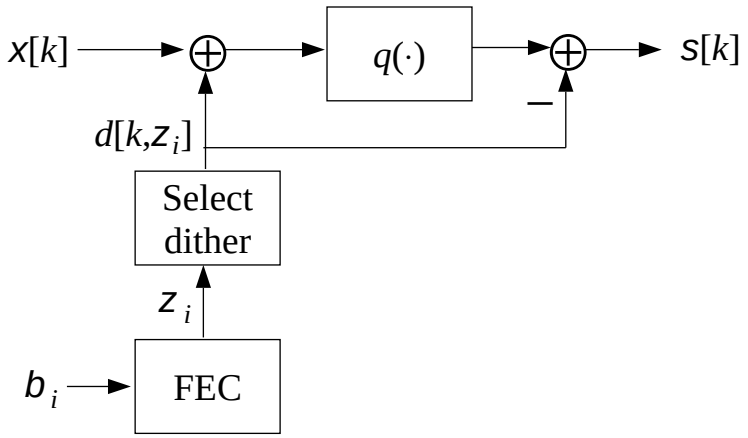

Figure 5-1: Embedder for coded binary dither modulation with uniform, scalar quantization. The only required computation beyond that of the forward error correction (FEC) code is one addition, one scalar quantization, and one subtraction per host signal sample.

choose $d[k,0]$ pseudorandomly with a uniform distribution over $[-\Delta_k/2, \Delta_k/2]$.[1] Also, the two dither sequences need not be the same for each length-L block.

- The i-th block of $\mathbf{x}$ is quantized with the dithered quantizer using the dither sequence $d[k, z_i]$.

5.1.1 Computational complexity

A block diagram of one implementation of the above embedding process is shown in Fig. 5-1, where we use the sequence notation $x[k]$ to denote the k-th element of the host signal vector $\mathbf{x}$. The actual embedding of the coded bits z_i requires only two adders and a uniform, scalar quantizer.

An implementation of the corresponding minimum distance decoder (3.8) is shown in Fig. 5-2. One can easily find the nearest reconstruction sequence of each quantizer (the 0-quantizer and the 1-quantizer) to the received sequence $y[k]$ using a few adders and scalar quantizers. For hard-decision forward error correction (FEC) decoding, one can make decisions on each coded bit z_i using the rule:

$$\hat{z}_i = \arg\min_{l\in\{0,1\}} \sum_{k=(i-1)L+1}^{iL} (y[k] - s_{\mathbf{y}}[k;l])^2 , \qquad i = 1, \ldots, N/L.$$

[1] 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 [23].

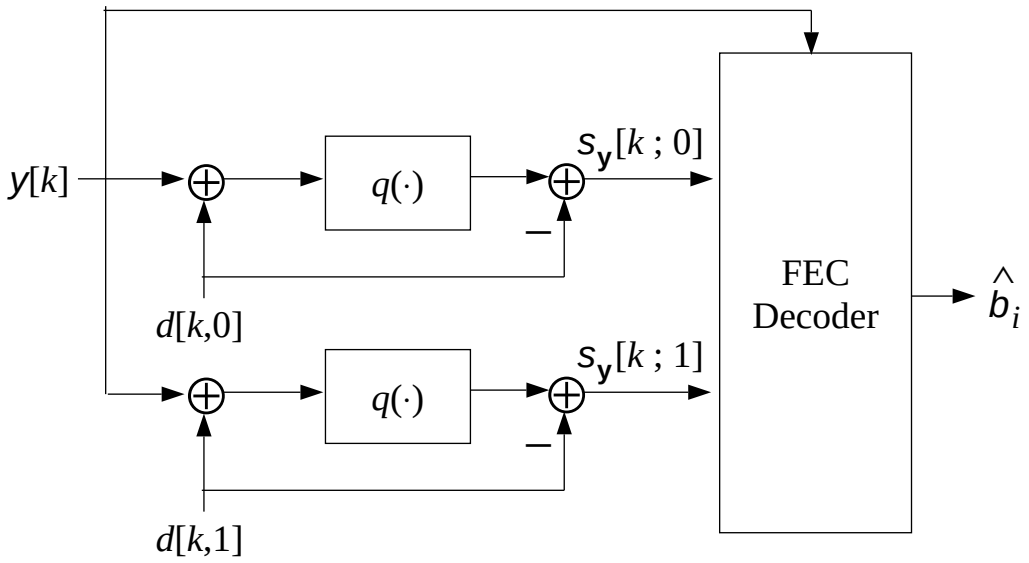

Figure 5-2: Decoder for coded binary dither modulation with uniform, scalar quantization. The distances between the received sequence $y[k]$ and the nearest quantizer reconstruction sequences $s_y[k;0]$ and $s_y[k;1]$ from each quantizer are used for either soft-decision or hard-decision forward error correction (FEC) decoding.

Then, the FEC decoder can generate the decoded information bit sequence $\left\{\hat{b}_1, \ldots, \hat{b}_{NR_m}\right\}$ from the estimates of the coded bits $\left\{\hat{z}_1, \ldots, \hat{z}_{N/L}\right\}$. Alternatively, one can use the metrics

$$\text{metric}(i,l) = \sum_{k=(i-1)L+1}^{iL} \left(y[k] - s_y[k;l]\right)^2, \qquad i = 1, \ldots, N/L, \quad l = 0, 1,$$

for soft-decision decoding. For example, one can use these metrics as branch metrics for a minimum squared Euclidean distance Viterbi decoder [26].

5.1.2 Minimum distance

Any two distinct coded bit sequences differ in at least d_H places, where d_H is the minimum Hamming distance of the error correction code. For each of these d_H blocks, the reconstruction points of the corresponding quantizers are shifted relative to each other by $\pm\Delta_k/2$ in the k-th dimension. Thus, the square of the minimum distance (3.2) over all N dimensions

is

$$\begin{aligned} d^2_{\min} &= d_H \sum_{k=1}^{L} \left(\frac{\Delta_k}{2}\right)^2 \\ &= \left(d_H \frac{k_u}{k_c}\right) \frac{1}{4LR_m} \sum_k \Delta_k^2 \\ &= \gamma_c \frac{1}{4LR_m} \sum_k \Delta_k^2, \end{aligned} \tag{5.1}$$

where γ_c is often referred to as the gain of the error correction code,

$$\gamma_c \stackrel{\Delta}{=} d_H(k_u/k_c). \tag{5.2}$$

If the quantization cells are sufficiently small such that the host signal can be modeled as uniformly distributed within each cell, the expected squared error distortion of a uniform, scalar quantizer with step size Δ_k is

$$\frac{1}{\Delta_k} \int_{-\Delta_k/2}^{\Delta_k/2} x^2 \, dx = \frac{\Delta_k^2}{12}.$$

Thus, the overall average expected distortion (2.2) is

$$D_{\mathbf{s}} = \frac{1}{12L} \sum_k \Delta_k^2. \tag{5.3}$$

Combining (5.1) and (5.3) yields the distortion-normalized squared minimum distance,

$$d^2_{\text{norm}} \equiv \frac{d^2_{\min}}{D_{\mathbf{s}}} = \frac{3\gamma_c}{R_m}, \tag{5.4}$$

which can be used to characterize the achievable performance of particular QIM realizations, as is done in later chapters.

5.2 Spread-transform Dither Modulation

Spread-transform dither modulation (STDM) is a special case of coded binary dither modulation. Some advantages of STDM over other forms of dither modulation, over a class of spread-spectrum methods we call amplitude-modulation spread spectrum (AM-SS), and over the generalized LBM method in [43] are discussed below.

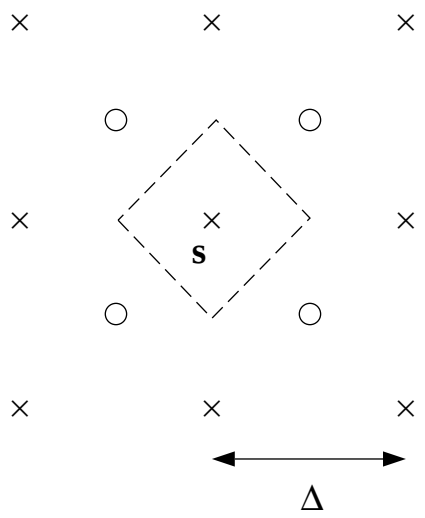

Figure 5-3: Dither modulation with uniform quantization step sizes.

5.2.1 Basic description and principles

The distortion-normalized squared minimum distance (5.4) of binary dither modulation with uniform scalar quantization does not depend on the sequence Δ_k, *i.e.*, on the distribution of the distortion across samples within the length-L block. Thus, one is free to choose any distribution without sacrificing d^2_{norm}, which completely characterizes the performance of dither modulation (and QIM in general) against bounded perturbation attacks and bounded host-distortion attacks, as we show in Chap. 7.

It may be advantageous in other contexts, though, to concentrate the distortion in a small number of samples, for example, in the first sample of every length-L block. Fig. 5-3 shows the reconstruction points of two quantizers for embedding one bit in a block of two samples, where the quantization step sizes are the same for both samples. Fig. 5-4 shows the case where a unitary transform has first been applied before embedding one bit. The first transform coefficient is the component of the host signal in the direction of $\mathbf{v}$, and the second transform coefficient is the component orthogonal to $\mathbf{v}$. The step size for quantizing the first transform coefficient is larger than in Fig. 5-3, but the step size for quantizing the second transform coefficient is zero. In this case to embed a 0-bit, the host signal is quantized to the nearest point on a line labeled with a $\times$. To embed a 1-bit, the host signal is quantized to the nearest point on a line labeled with a $\circ$. The minimum distance in both cases is $\Delta/\sqrt{2}$, and the average squared error distortion is $\Delta^2/12$ per sample. Thus, the robustness against bounded perturbations is the same in both cases. However, the *number*

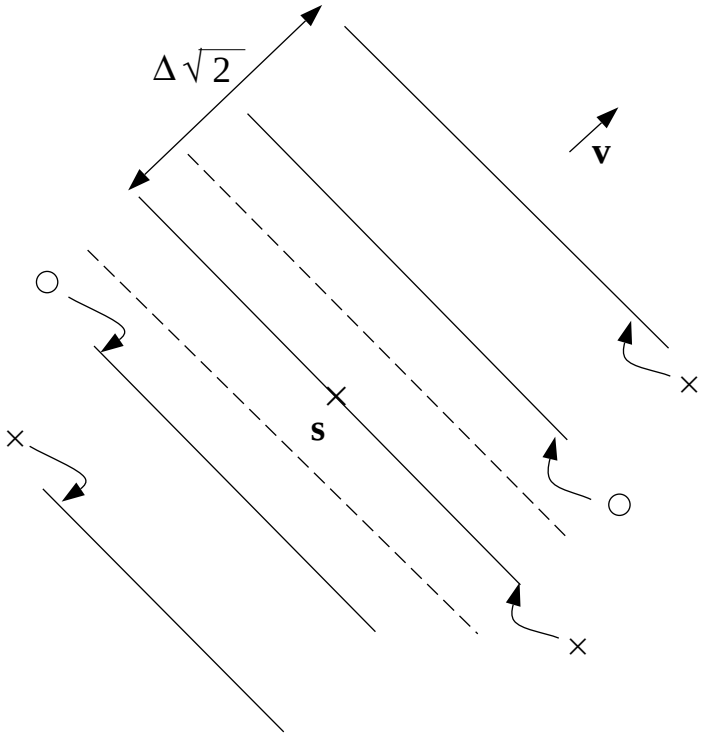

Figure 5-4: Transform dither modulation with quantization of only a single transform component. The quantization step size for the component of the host signal orthogonal to **v** is zero.

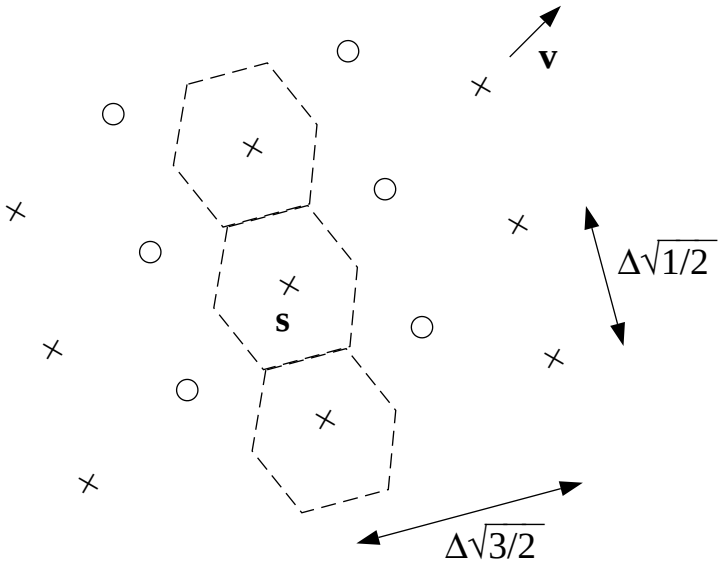

Figure 5-5: Transform dither modulation with non-uniform quantization step sizes.

of perturbation vectors of length $d_{\min}/2$ that cause decoding errors is higher for the case of Fig. 5-3 than for the case of Fig. 5-4. (For intermediate cases such as the one shown in Fig. 5-5, where quantization step sizes in different dimensions are different but non-zero, the number of perturbation vectors of length $d_{\min}/2$ that cause decoding errors is the same as in Fig. 5-3, but these vectors are not orthogonal.) Thus, for probabilistic channels such as additive noise channels, the *probability* of error may be different in the different cases. For example, suppose a 0-bit is embedded and the composite signal is the $\times$ point labeled with $\mathbf{s}$ in Figs. 5-3 and 5-4. If the channel output lies in the decision region defined by the dashed box in Fig. 5-3 and defined by the two dashed lines in Fig. 5-4, then the decoder will correctly determine that a 0-bit was embedded. If the perturbation vector places the channel output outside the decision region, however, the decoder will make an error with very high probability. (There is some possibility that the channel output is outside the decision region but is still closer to a $\times$ point other than $\mathbf{s}$ than to the closest $\circ$. These events, however, are very unlikely for many perturbation probability distributions that are of practical interest.) Since the decision region of Fig. 5-4 contains the decision region of Fig. 5-3, we conclude that the probability of a correct decision in the case of non-uniform quantization step sizes is higher.

The unitary transform in the case of Fig. 5-4 not only facilitates a comparison of Figs. 5-3 and 5-4, but also may be necessary to spread any embedding-induced distortion over frequency and space, in the case of an image, and over frequency and time, in the case of an audio signal, to meet a peak distortion constraint, for example. Although, the distortion is concentrated in only one transform coefficient, if the energy of $\mathbf{v}$ is spread over space/time and frequency — for example, $\mathbf{v}$ is chosen pseudorandomly — then the distortion will also be spread. Thus, we call this type of dither modulation, which is illustrated in Fig. 5-6, "spread-transform dither modulation (STDM)".

Later in this thesis, we show that dither modulation methods have considerable performance advantages over previously proposed spread spectrum methods in a variety of contexts. However, much effort has already been invested in optimizing spread spectrum systems, for example, by exploiting perceptual properties of the human visual and auditory systems or designing receiver front-ends to mitigate effects of geometric distortion. An advantage of spread-transform dither modulation over other forms of dither modulation is that one can easily convert existing amplitude-modulation spread spectrum (AM-SS) systems,

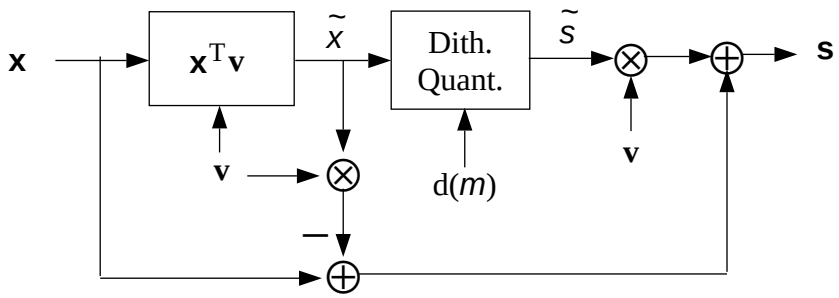

Figure 5-6: Spread-transform dither modulation. Information is embedded in the projection of a block $\mathbf{x}$ of the host signal onto $\mathbf{v}$, which is typically a pseudorandom vector. Components of $\mathbf{x}$ orthogonal to $\mathbf{v}$ are added back to the signal after dithered quantization to form the corresponding block of the composite signal $\mathbf{s}$.

a class of previously proposed spread spectrum methods that have embedding functions of the form

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

into spread-transform dither modulation systems since the embedding function can be re-written in the form

$$\mathbf{s}(\mathbf{x}, m) = (\tilde{x} + a(m))\mathbf{v} + (\mathbf{x} - \tilde{x}\mathbf{v}),$$

where $\tilde{x} = \mathbf{x}^T\mathbf{v}$. We see that AM-SS is equivalent to adding a value $a(m)$ to the projection $\tilde{x}$ of the host signal onto the spreading vector $\mathbf{v}$. Thus, if one has spent considerable effort in designing a good spread spectrum system, for example, by designing a $\mathbf{v}$ that has good perceptual distortion properties, but would like to gain the advantages of dither modulation, one can do so simply by replacing the addition step of AM-SS,

$$\tilde{s} = \tilde{x} + a(m), \tag{5.5}$$

by the quantization step of STDM,

$$\tilde{s} = q(\tilde{x} + d(m)) - d(m). \tag{5.6}$$

5.2.2 SNR advantage of STDM over AM spread spectrum

The close coupling of STDM and AM spread spectrum allows a direct comparison between the performance of the two methods that suggests that STDM has important performance

advantages over AM spread spectrum in a broad range of contexts, as we show in this section. This performance advantage results from the host signal interference rejection properties of QIM methods in general.

We consider embedding one bit in a length-L block $\mathbf{x}$ using STDM and AM spread spectrum methods with the same spreading vector $\mathbf{v}$, which is of unit length. Because the embedding occurs entirely in the projections of $\mathbf{x}$ onto $\mathbf{v}$, the problem is reduced to a one-dimensional problem with the embedding functions (5.5) and (5.6). For AM-SS (5.5), $a(m) = \pm\sqrt{LD_{\mathbf{s}}}$ so that

$$|a(1) - a(2)|^2 = 4LD_{\mathbf{s}}. \tag{5.7}$$

For STDM (5.6),

$$\min_{(\tilde{x}_1,\tilde{x}_2)} |\tilde{s}(\tilde{x}_1, 1) - \tilde{s}(\tilde{x}_2, 2)|^2 = \Delta^2/4 = 3LD_{\mathbf{s}}, \tag{5.8}$$

where $\Delta = \sqrt{12LD_{\mathbf{s}}}$ so that the expected distortion in both cases is the same, and where we have used the fact that $d(1)$ and $d(2)$ are chosen such that $|d(1) - d(2)| = \Delta/2$. Because all of the embedding-induced distortion occurs only in the direction of $\mathbf{v}$, the distortion in both cases also has the same time or spatial distribution and frequency distribution. Thus, one would expect that any perceptual effects due to time/space masking or frequency masking are the same in both cases. Therefore, squared error distortion may be a more meaningful measure of distortion when comparing STDM with AM-SS than one might expect in other more general contexts where squared error distortion may fail to capture certain perceptual effects.

The decoder in both cases makes a decision based on $\tilde{y}$, the projection of the channel output $\mathbf{y}$ onto $\mathbf{v}$. In the case of AM-SS,

$$\tilde{y} = a(m) + \tilde{x} + \tilde{n},$$

while in the case of STDM,

$$\tilde{y} = \tilde{s}(\tilde{x}, m) + \tilde{n},$$

where $\tilde{n}$ is the projection of the perturbation vector $\mathbf{n}$ onto $\mathbf{v}$. We let $P(\cdot)$ be some measure of energy. For example, $P(x) = x^2$ in the case of a deterministic variable x, or $P(\mathbf{x})$ equals

the variance of the random variable x. The energy of the interference or "noise" is $P(\tilde{x}+\tilde{n})$ for AM-SS, but only $P(\tilde{n})$ for STDM, *i.e.*, the host signal interference for STDM is zero. Thus, the signal-to-noise ratio at the decision device is

$$\mathrm{SNR}_{\mathrm{AM-SS}} = \frac{4LD_{\mathrm{s}}}{P(\tilde{x}+\tilde{n})}$$

for AM-SS and

$$\mathrm{SNR}_{\mathrm{STDM}} = \frac{3LD_{\mathrm{s}}}{P(\tilde{n})}$$

for STDM, where the "signal" energies $P(a(1)-a(2))$ and $P\left(\min_{(\tilde{x}_1,\tilde{x}_2)}|\tilde{s}(\tilde{x}_1,1)-\tilde{s}(\tilde{x}_2,2)|\right)$ are given by (5.7) and (5.8). Thus, the advantage of STDM over AM-SS is

$$\frac{\mathrm{SNR}_{\mathrm{STDM}}}{\mathrm{SNR}_{\mathrm{AM-SS}}} = \frac{3}{4}\frac{P(\tilde{x}+\tilde{n})}{P(\tilde{n})}, \tag{5.9}$$

which is typically very large since the channel perturbations $\tilde{n}$ are usually much smaller than the host signal $\tilde{x}$ if the channel output $\tilde{y}$ is to be of reasonable quality. For example, if the host signal-to-channel noise ratio is 30 dB and $\tilde{x}$ and $\tilde{n}$ are uncorrelated, then the SNR advantage (5.9) of STDM over AM spread spectrum is 28.8 dB.

Furthermore, although the SNR gain in (5.9) is less than 0 dB ($3/4 = -1.25$ dB) when the host signal interference is zero ($\tilde{x}=0$), for example, such as would be the case if the host signal $\mathbf{x}$ had very little energy in the direction of $\mathbf{v}$, STDM may not be worse than AM-SS even in this case since (5.9) applies only when $\tilde{x}$ is approximately uniformly distributed across the STDM quantization cell so that $D_{\mathrm{s}} = \Delta^2/(12L)$. If $\tilde{x}=0$, however, and one chooses the dither signals to be $d(m) = \pm\Delta/4$, then the distortion is only $D_{\mathrm{s}} = \Delta^2/(16L)$ so that STDM is just as good as AM-SS in this case.

5.2.3 SNR advantage of STDM over generalized LBM

Spread-transform dither modulation methods also have an SNR advantage over generalized low-bit(s) modulation methods such as the quantization-and-perturbation [43] embedding method. As we show in App. C, the distortion-normalized squared minimum distance (5.4) of LBM is $7/4 \approx 2.43$ dB worse than that of dither modulation in the case of uniform, scalar quantization. Thus, for a fixed rate and embedding-induced distortion, the squared-minimum distance, and hence the SNR at the decision device, for LBM will be 2.43 dB

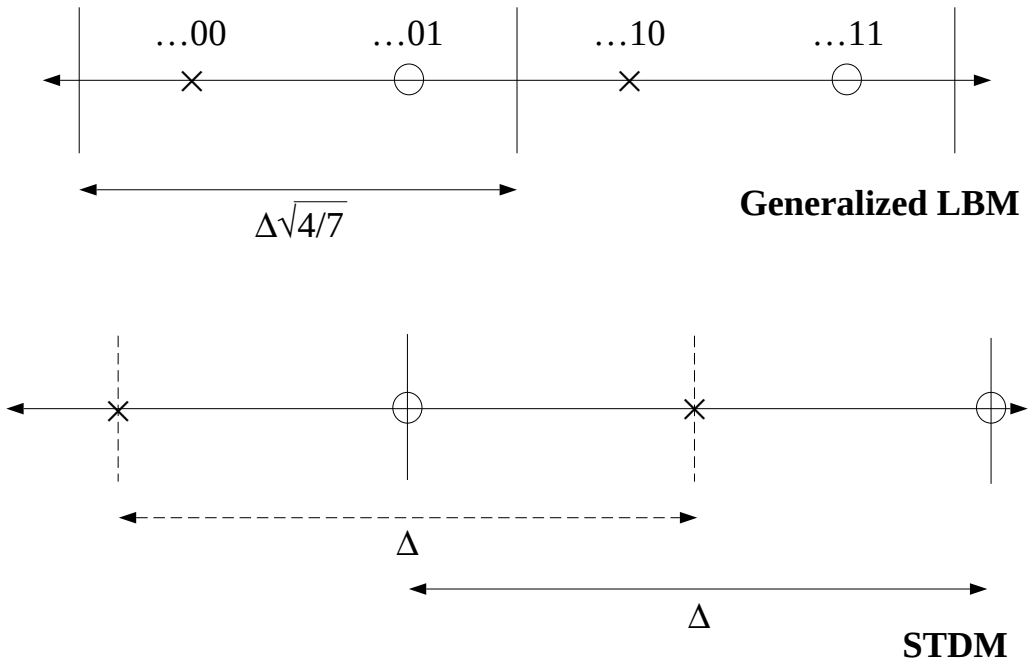

Figure 5-7: Spread-transform dither modulation vs. generalized low-bit modulation. The embedding interval boundaries of generalized LBM, which are shown with solid lines, are the same for both $\times$ points and $\circ$ points. In contrast, in the case of STDM, the $\times$-point embedding intervals, shown by solid lines, differ from the $\circ$-point embedding intervals, shown by dashed lines. An SNR advantage of $7/4 = 2.43$ dB for STDM results.

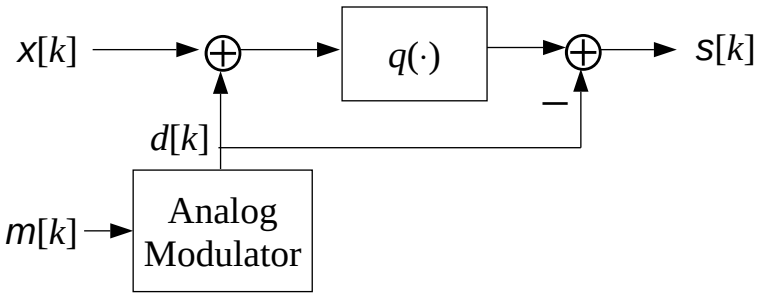

Figure 5-8: Analog dither modulation with uniform, scalar quantization. An analog modulation technique such as amplitude modulation generates a dither sequence. Dithered quantization follows.

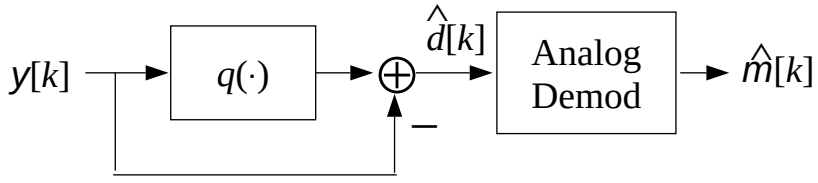

Figure 5-9: Analog dither demodulation with uniform, scalar quantization. The first stage, estimation of the dither sequence, is followed by second stage analog demodulation.

worse than that of STDM.

This SNR advantage is illustrated in Fig. 5-7, where the quantizer reconstruction points and embedding intervals for both generalized LBM and STDM are shown. The embedding-induced squared error distortion is the same for both cases, but the squared minimum distance for generalized LBM is a factor of 4/7 smaller than that of STDM.

5.3 Embedding Analog Data

In some potential applications, one may desire to use some of the embedding methods discussed in this thesis to embed analog data as well as digital data. In this section we briefly discuss some aspects of analog embedding using dither modulation.

If $m[k]$ is a sequence of real numbers rather than a sequence of bits, one can still use it to modulate the dither vector, as illustrated in Fig. 5-8. For example, one could modulate

the amplitude of a signature sequence $v[k]$,[2]

$$d[k] = m[k]v[k].$$

Using vector notation, the analog dither modulation embedding function is

$$\mathbf{s}(\mathbf{x}, \mathbf{m}) = \mathbf{q}(\mathbf{x} + \mathbf{d}(\mathbf{m})) - \mathbf{d}(\mathbf{m}). \tag{5.10}$$

One method for decoding the embedded message $m[k]$ is the two-stage demodulation method illustrated in Fig. 5-9. First, one constructs an estimate $\hat{\mathbf{d}}$ of the dither vector. Then, one demodulates the embedded message $\hat{\mathbf{m}}$ from this estimated dither vector. Typically, one modulates the dither vector $\mathbf{d}$ such that it will not carry a reconstruction point out of its quantization cell, *i.e.*,

$$\mathbf{q}(\mathbf{q}_0 - \mathbf{d}) = \mathbf{q}_0$$

for every reconstruction point $\mathbf{q}_0$. For example, in the case of uniform scalar quantization, this condition is satisfied if $|d[k]| < \Delta_k/2$, where Δ_k is the quantization step size of the k-th scalar quantizer. In these typical cases

$$\begin{aligned} \hat{\mathbf{d}}(\mathbf{s}) &= \mathbf{q}(\mathbf{s}) - \mathbf{s} \\ &= \mathbf{q}(\mathbf{q}(\mathbf{x} + \mathbf{d}(\mathbf{m})) - \mathbf{d}(\mathbf{m})) - (\mathbf{q}(\mathbf{x} + \mathbf{d}(\mathbf{m})) - \mathbf{d}(\mathbf{m})) \\ &= \mathbf{q}(\mathbf{x} + \mathbf{d}(\mathbf{m})) - \mathbf{q}(\mathbf{x} + \mathbf{d}(\mathbf{m})) + \mathbf{d}(\mathbf{m}) \\ &= \mathbf{d}(\mathbf{m}). \end{aligned} \tag{5.11}$$

Thus, the dither estimation stage (5.11) of Fig. 5-9 is the inverse of the dithered quantization stage (5.10) of Fig. 5-8, and if the analog demodulation stage is the inverse of the analog modulation stage, this decoder perfectly reconstructs $\mathbf{m}$ in the noiseless case. In the noisy case if the perturbation vector $\mathbf{n}$ is small enough such that

$$\mathbf{q}(\mathbf{s} + \mathbf{n}) = \mathbf{q}(\mathbf{s}),$$

[2] We include the possibility that the embedded information is a single number that has been expanded into a sequence $m[k]$ through repetition.

then

$$\begin{aligned}\hat{\mathbf{d}}(\mathbf{y}) &= \mathbf{q}(\mathbf{s}+\mathbf{n}) - (\mathbf{s}+\mathbf{n}) \\ &= (\mathbf{q}(\mathbf{s}) - \mathbf{s}) - \mathbf{n} \\ &= \mathbf{d}(\mathbf{m}) - \mathbf{n},\end{aligned}$$

where we have used (5.11) in the last line. Thus, in this small perturbation case the dithered quantization and dither estimation stages are transparent to the analog modulator and analog demodulator (to within a sign change). The effective channel connecting the analog modulator to the analog demodulator produces the same perturbation vector, to within a sign change, as the perturbation vector of the actual channel.

Chapter 6

Gaussian Channels

As discussed in Chaps. 1 and 2, a number of information embedding applications arise in which robustness against only unintentional attacks is required. In many of these cases, an additive Gaussian noise model for the channel may be appropriate, especially if we allow an arbitrary covariance matrix or, equivalently, an arbitrary noise power spectrum. Furthermore, although many of the host signals that arise in multimedia applications — speech, audio, images, and video signals, for example — may not be precisely Gaussian, a Gaussian model for these signals can still capture the correlation among signal samples, provided that we allow an arbitrary covariance matrix. Thus, even if the host signal is not actually Gaussian, if we have only a second-order characterization of the host signal, a Gaussian model allows us to incorporate all of this information. Also, given the host-signal interference rejection properties of good information embedding systems, the non-Gaussianity of the host signal may not play a significant role in the ultimate performance of such systems.

Thus, in this chapter we examine the ultimate performance limits of various information embedding methods when both the host signal is Gaussian and the channel is an additive Gaussian noise channel. Specifically, we consider the case where the host signal vector $\mathbf{x}$ and the noise vector $\mathbf{n}$ are statistically independent and can be decomposed into

$$\mathbf{x} = [x_1 \cdots x_{N/L}]^T \qquad \text{and} \qquad \mathbf{n} = [n_1 \cdots n_{N/L}]^T,$$

where the x_i are independent and identically distributed (iid), L-dimensional, zero-mean, Gaussian vectors with covariance matrix $K_x = Q_x \Lambda_x Q_x^T$ and the n_i are iid, L-dimensional,

zero-mean, Gaussian vectors with covariance matrix $K_n = Q_n \Lambda_n Q_n^T$. The columns of the matrices Q_x and Q_n are the eigenvectors of their respective covariance matrices and Λ_x and Λ_n are diagonal matrices of the respective eigenvalues. This model is appropriate when the power spectra of the host signal and channel noise are sufficiently smooth that one can decompose the channel into L parallel, narrowband subchannels, over each of which the host signal and channel noise power spectra are approximately flat. Many bandwidth-conserving hybrid transmission applications are examples of such a scenario, and this model may also apply to optimal, *i.e.*, rate-distortion achieving [13], lossy compression of a Gaussian source, as discussed in App. B.

When the channel noise is not white, issues arise as to how to measure distortion and how to define distortion-to-noise ratio (DNR). One may want to make the embedding-induced distortion "look like" the channel noise so that as long as the channel noise does not cause too much perceptible degradation to the host signal, then neither does the embedding-induced distortion. One can impose this condition by choosing distortion measures that favor relatively less embedding-induced distortion in components where the channel noise is relatively small and allow relatively more distortion in components where the channel noise is relatively large. Then, the embedding-induced distortion will look like a scaled version of the channel noise, with the DNR as the scaling factor. If the DNR is chosen small enough, then the embedding-induced distortion will be "hidden in the noise".

In this chapter we consider two ways to measure distortion and DNR and show that in each case when we impose this constraint that the embedding-induced distortion signal look like a scaled version of the channel noise, the information-embedding capacity is independent of the host and noise statistics and depends only on the DNR. After presenting these capacity results in Sec. 6.1, we discuss in Sec. 6.2 their implications for bandwidth-conserving hybrid transmission applications where a digital data signal is embedded within a multimedia host signal. We also discuss some connections between information embedding and broadcast communication problems in this section. We conclude the chapter in Sec. 6.3 by comparing different types of information embedding methods in terms of their gaps to capacity.

6.1 Capacities

As explained in Chap. 4, viewing information embedding as communication with side information allows one to apply earlier results of Gel'fand and Pinsker [19] to conclude that the information embedding capacity is given by (4.1). In this section we specialize this result to the Gaussian case described above. Our main results are:

1. The capacity is

$$C_{\text{Gauss}} = \frac{1}{2}\log_2(1 + \text{DNR}), \tag{6.1}$$

   when one uses one of two squared error based distortion measures and constrains the embedding-induced distortion to look like a scaled version of the channel noise. This capacity is the same as in the case when the host signal is known at the decoder.

2. Preprocessing the host signal with a linear transform that whitens the channel noise and decorrelates the host signal samples, embedding with distortion-compensated QIM, and postprocessing the result with the inverse linear transform is an optimal (capacity-achieving) embedding strategy in the Gaussian case.

We arrive at these results by considering first the case of a white host signal and white noise. After determining the capacity in that simplest case, we show that one can transform the case of a colored host signal and white noise into the white host, white noise case. Finally, we show that one can transform the most general case of a colored host signal and colored noise into the colored host, white noise case.

6.1.1 White host and white noise

We consider first the case of a white host signal ($K_x = \sigma_x^2 I$), white noise ($K_n = \sigma_n^2 I$), and the distortion constraint

$$\frac{L}{N}\sum_{i=1}^{N/L} \boldsymbol{e}_i^T \boldsymbol{e}_i \le L D_{\mathbf{s}}, \tag{6.2}$$

This case is equivalent to the $L = 1$ case, and the equivalent distortion constraint is

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{e}_i^2 \le D_{\mathbf{s}},$$

with the corresponding constraint on $p_{u,e|x}(u,e|x)$ in (4.1) being $E[\mathsf{e}^2] \leq D_{\mathbf{s}}$. We see that squared error distortion-constrained, Gaussian information embedding is equivalent to power-constrained communication over a Gaussian channel with Gaussian side information known at the encoder, a case for which Costa [11] has determined the capacity to be

$$C_{\text{AWGN}} = \frac{1}{2}\log_2\left(1+\frac{D_{\mathbf{s}}}{\sigma_n^2}\right) = \frac{1}{2}\log_2(1+\text{DNR}), \tag{6.3}$$

as asserted in (6.1). Remarkably, as we discuss in Sec. 6.1.4, the capacity is the same as in the case when the host signal $\mathbf{x}$ is known at the decoder, implying that an infinite energy host signal causes no decrease in capacity in this Gaussian case, *i.e.*, good information embedding systems can *completely* reject host-signal interference in the Gaussian case.

Before proceeding to the colored host signal case, we briefly discuss the proof [11] of (6.3). As discussed in Chap. 4, one wishes to find the pdf that maximizes (4.1). One distribution to try is the one implied by [11]

$$\mathsf{u} = \mathsf{e} + \alpha\mathsf{x}, \tag{6.4}$$

where $\mathsf{e} \sim \mathcal{N}(0, D_{\mathbf{s}})$ and e and x are independent.[1] For a fixed value of α, an achievable rate $I(\mathsf{u};\mathsf{y}) - I(\mathsf{u};\mathsf{x})$ is [11]

$$R(\alpha) = \frac{1}{2}\log_2\left(\frac{D_{\mathbf{s}}(D_{\mathbf{s}}+\sigma_x^2+\sigma_n^2)}{D_{\mathbf{s}}\sigma_x^2(1-\alpha)^2+\sigma_n^2(D_{\mathbf{s}}+\alpha^2\sigma_x^2)}\right),$$

which can also be written in terms of the DNR and the host-signal-to-noise ratio ($\text{SNR}_x = \sigma_x^2/\sigma_n^2$),

$$R(\alpha) = \frac{1}{2}\log_2\left(\frac{\text{DNR}(1+\text{DNR}+\text{SNR}_x)}{\text{DNR}\,\text{SNR}_x(1-\alpha)^2+(\text{DNR}+\alpha^2\text{SNR}_x)}\right). \tag{6.5}$$

This rate is maximized by setting

$$\alpha_{\text{cap}} = \frac{\text{DNR}}{\text{DNR}+1} \tag{6.6}$$

to obtain (6.3). Clearly, since (6.3) is the maximum achievable rate when x is known at the

[1] We emphasize that while the sequences $\mathbf{e}$ and $\mathbf{x}$ may be of independent type, the distortion signal $\mathbf{e}$ is still chosen as a function of the host signal $\mathbf{x}$, as described in Chap. 4.

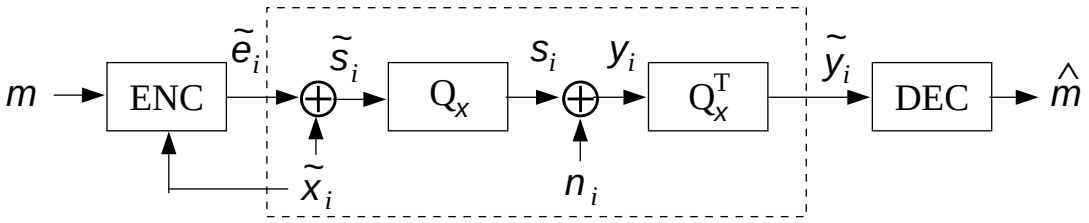

Figure 6-1: Embedding in transform domain for colored host signal and white noise. The dashed box is the equivalent transform-domain channel.

decoder (See Sec. 6.1.4.), one cannot exceed this rate when $\boldsymbol{x}$ is not known at the decoder, and this achievable rate is the capacity.

6.1.2 Colored host and white noise

We now consider the case of an arbitrary host signal covariance matrix $K_x = Q_x \Lambda_x Q_x^T$ and white noise ($K_n = \sigma_n^2 I$). The distortion constraint is still (6.2) with the corresponding constraint on $p_{u,e|x}(u,e|x)$ in (4.1) being $E[\boldsymbol{e}^T\boldsymbol{e}] \leq LD_{\mathbf{s}}$. Thus, $LD_{\mathbf{s}}$ is the maximum average energy of the L-dimensional vectors $\boldsymbol{e}_i$, so $D_{\mathbf{s}}$ is still the maximum average energy per dimension.

One way to determine the capacity in this case is to consider embedding in a linear transform domain, where the covariance matrix of the host signal is diagonal. Because the transform is linear, the transformed host signal vector remains Gaussian. One such orthogonal transform is the well-known Karhunen-Loeve transform [46], and the resulting transformed host signal vector is

$$\tilde{\boldsymbol{x}} = Q_x^T \boldsymbol{x},$$

with covariance matrix $K_{\tilde{x}} = \Lambda_x$. The distortion constraint (6.2) in the transform domain on the vectors $\tilde{\boldsymbol{e}} = Q_x^T \boldsymbol{e}$ is

$$\frac{L}{N} \sum_{i=1}^{N/L} \tilde{\boldsymbol{e}}_i^T \tilde{\boldsymbol{e}}_i \leq LD_{\mathbf{s}},$$

since

$$\tilde{\boldsymbol{e}}_i^T \tilde{\boldsymbol{e}}_i = \boldsymbol{e}_i^T Q_x Q_x^T \boldsymbol{e}_i = \boldsymbol{e}_i^T \boldsymbol{e}_i.$$

An overall block diagram of the transformed problem is shown in Fig. 6-1. The transform-domain channel output $\tilde{\boldsymbol{y}}$ is

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{e}} + \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{n}},$$

where the transform-domain noise $\tilde{\boldsymbol{n}}$ has the same covariance matrix as $\boldsymbol{n}$,

$$K_{\tilde{\boldsymbol{n}}} = Q_{\boldsymbol{x}}^T \left(\sigma_n^2 I\right) Q_{\boldsymbol{x}} = \sigma_n^2 I = K_{\boldsymbol{n}}.$$

Since both $K_{\tilde{\boldsymbol{x}}}$ and $K_{\tilde{\boldsymbol{n}}}$ are diagonal, in the transform domain we have L parallel, *independent* subchannels, each of which is an AWGN channel with noise variance σ_n^2 and each of which has a white, Gaussian host signal. Thus, as we show formally in App. D, the overall capacity is simply the sum of the capacities of the individual subchannels (6.3),

$$C_L = \sum_{i=1}^{L} \frac{1}{2} \log_2(1 + \text{DNR}) = \frac{L}{2} \log_2(1 + \text{DNR}). \tag{6.7}$$

This capacity is in bits per L-dimensional host signal vector, so the capacity in bits per dimension is

$$C = \frac{1}{2} \log_2(1 + \text{DNR}), \tag{6.8}$$

the same as the capacity when the host signal is white (6.3). Thus, not only is the capacity independent of the host signal power for white Gaussian host signals as discussed above in Sec. 6.1.1, but in the more general case where the host signal has any arbitrary covariance matrix, the capacity is independent of *all* host signal statistics. (The statistics of a Gaussian random vector are completely characterized by its mean and covariance.)

6.1.3 Colored host and colored noise

We now extend our results to the case of arbitrary host signal and noise covariance matrices $K_{\boldsymbol{x}} = Q_{\boldsymbol{x}} \Lambda_{\boldsymbol{x}} Q_{\boldsymbol{x}}^T$ and $K_{\boldsymbol{n}} = Q_{\boldsymbol{n}} \Lambda_{\boldsymbol{n}} Q_{\boldsymbol{n}}^T$, respectively. We assume that the eigenvalues of $K_{\boldsymbol{n}}$ are non-zero, *i.e.*, $K_{\boldsymbol{n}}$ is invertible.

As discussed in the introduction to this chapter, when the channel noise is not white, one may want to constrain the embedding-induced distortion signal to "look like" a scaled version of the channel noise. As mentioned during that discussion, we consider two such ways to impose this constraint through our definition of distortion and DNR. The first distortion measure is a weighted average squared error measure, and in the second case, we use multiple distortion constraints, one on each of the components. Below, we show that both cases can be transformed into a colored host, white noise case, and thus, the capacity

is (6.1).

Weighted squared error distortion

We consider the distortion measure and constraint

$$\frac{L}{N}\sum_{i=1}^{N/L} \mathbf{e}_i^T K_n^{-1} \mathbf{e}_i \leq L\,\mathrm{DNR}, \tag{6.9}$$

so that the corresponding constraint on $p_{u,\mathbf{e}|x}(u,e|x)$ in (4.1) is $E[\mathbf{e}^T K_n^{-1}\mathbf{e}] \leq L\,\mathrm{DNR}$. The weighting matrix K_n^{-1} more heavily penalizes distortion in the directions of eigenvectors corresponding to small eigenvalues (noise variances). Thus, the embedding-induced distortion will tend to be large only in those components where the channel noise is also large, and the distortion will tend to be small in the components where the channel noise is also small. The equivalence between this case and the colored host, white noise case discussed in the last section will be made apparent through an invertible, linear transform.

The transform required in this case not only diagonalizes the noise covariance matrix, but also makes the transformed noise samples equivariant. Specifically, the transform matrix is $\Lambda_n^{-1/2}Q_n^T$, and the transformed host signal vector

$$\tilde{\mathbf{x}} = \Lambda_n^{-1/2}Q_n^T\mathbf{x}$$

has covariance matrix

$$K_{\tilde{x}} = \Lambda_n^{-1/2}Q_n^T K_x Q_n \Lambda_n^{-1/2}.$$

A block diagram for the overall problem is similar to the one in Fig. 6-1, with the transform matrix Q_x^T replaced by $\Lambda_n^{-1/2}Q_n^T$ and the inverse transform matrix Q_x replaced by $Q_n\Lambda_n^{1/2}$. Because the transform is invertible, there is no loss of optimality from embedding in this transform domain. The transform-domain channel output $\tilde{\mathbf{y}}$ is

$$\tilde{\mathbf{y}} = \tilde{\mathbf{e}} + \tilde{\mathbf{x}} + \tilde{\mathbf{n}},$$

where the transform-domain noise $\tilde{\mathbf{n}}$ has covariance matrix

$$K_{\tilde{n}} = \Lambda_n^{-1/2}Q_n^T\left(Q_n\Lambda_n Q_n^T\right)Q_n\Lambda_n^{-1/2} = I. \tag{6.10}$$

Thus, the components of $\tilde{\boldsymbol{n}}$ are uncorrelated (and independent since $\tilde{\boldsymbol{n}}$ is Gaussian) and have unit variance.

The distortion constraint (6.9) in the transform domain is

$$\frac{L}{N}\sum_{i=1}^{N/L}\tilde{\boldsymbol{e}}_i^T\tilde{\boldsymbol{e}}_i \leq L\,\mathrm{DNR}$$

since

$$\begin{aligned}\boldsymbol{e}_i^T K_{\boldsymbol{n}}^{-1}\boldsymbol{e}_i &= \boldsymbol{e}_i^T\left(Q_{\boldsymbol{n}}\Lambda_{\boldsymbol{n}}^{-1}Q_{\boldsymbol{n}}^T\right)\boldsymbol{e}_i \\ &= \left(\boldsymbol{e}_i^T Q_{\boldsymbol{n}}\Lambda_{\boldsymbol{n}}^{-1/2}\right)\left(\Lambda_{\boldsymbol{n}}^{-1/2}Q_{\boldsymbol{n}}^T\boldsymbol{e}_i\right) \\ &= \tilde{\boldsymbol{e}}_i^T\tilde{\boldsymbol{e}}_i.\end{aligned}$$

Thus, the transform-domain distortion constraint in this case is the same as the non-transform domain distortion constraint (6.2) of the last section. In both cases the host signal is colored and Gaussian, and the channel noise is white and Gaussian. Thus, the capacity in both cases is the same (6.1),

$$C = \frac{1}{2}\log_2(1+\mathrm{DNR}), \tag{6.11}$$

and was determined in the last section.

Multiple, simultaneous squared error distortion

An alternative, and more restrictive, distortion constraint to (6.9) arises by strictly requiring that the embedding-induced distortion in components corresponding to small noise eigenvalues be small rather than simply weighting these distortions more heavily. Specifically, we consider the set of constraints

$$\frac{L}{N}\sum_{i=1}^{N/L}\left(\boldsymbol{q}_j^T\boldsymbol{e}_i\right)^2 \leq \mathrm{DNR}\,\lambda_j, \qquad j=1,\ldots,L, \tag{6.12}$$

where $\boldsymbol{q}_j$ and λ_j are the j-th eigenvector and eigenvalue, respectively, of K_n. Any distortion signal that satisfies (6.12) also satisfies (6.9) since

$$\begin{aligned}
\frac{L}{N}\sum_{i=1}^{N/L} \boldsymbol{e}_i^T K_n^{-1} \boldsymbol{e}_i &= \frac{L}{N}\sum_{i=1}^{N/L} \left(Q_n^T \boldsymbol{e}_i\right)^T \Lambda_n^{-1} \left(Q_n^T \boldsymbol{e}_i\right) \\
&= \frac{L}{N}\sum_{i=1}^{N/L}\sum_{j=1}^{L} \left(\boldsymbol{q}_j^T \boldsymbol{e}_i\right)^2 \frac{1}{\lambda_j} \\
&= \sum_{j=1}^{L}\left[\frac{L}{N\lambda_j}\sum_{i=1}^{N/L}\left(\boldsymbol{q}_j^T \boldsymbol{e}_i\right)^2\right] \\
&\leq L\,\mathrm{DNR},
\end{aligned}$$

where the first line follows from the factorization $K_n^{-1} = Q_n \Lambda_n^{-1} Q_n^T$ and where the final line follows from (6.12). Thus, the constraint (6.12) is indeed more restrictive than (6.9).

To determine the information-embedding capacity in this case, we again consider the noise-whitening linear transform $\Lambda_n^{-1/2} Q_n^T$. The j-th component of the transform-domain distortion vector $\tilde{\boldsymbol{e}}_i = \Lambda_n^{-1/2} Q_n^T \boldsymbol{e}_i$ is

$$\left[\tilde{\boldsymbol{e}}_i\right]_j = \frac{1}{\sqrt{\lambda_j}} \boldsymbol{q}_j^T \boldsymbol{e}_i.$$

Thus, the transform-domain distortion constraint equivalent to (6.12) is

$$\frac{L}{N}\sum_{i=1}^{N/L} \left[\tilde{\boldsymbol{e}}_i\right]_j^2 \leq \mathrm{DNR}, \qquad j = 1, \ldots, L. \tag{6.13}$$

By (6.10), the transform-domain noise covariance matrix is the identity matrix. Thus, if we treat each of the L subchannels independently, each with its own distortion constraint (6.13) and a noise variance of unity, then on the j-th subchannel we can achieve a rate

$$C_j = \frac{1}{2}\log_2(1 + \mathrm{DNR}),$$

so the total rate across all L channels in bits per dimension is

$$C = \frac{1}{L}\sum_{j=1}^{L} C_j = \frac{1}{2}\log_2(1 + \mathrm{DNR}). \tag{6.14}$$

Since this rate equals the capacity (6.11) corresponding to a less restrictive distortion con-

straint (6.9), we cannot hope to achieve a rate higher than this one. Thus, treating the L subchannels independently does not result in any loss of optimality, and the achievable rate (6.14) is indeed the capacity.

Thus, for Gaussian host signals and additive Gaussian noise channels, with the constraint that the embedding-induced distortion signal "look like" the channel noise, the information-embedding capacity is independent of the host and noise covariance matrices (Since the signals are Gaussian, the capacity is actually independent of all host signal and noise statistics.) and is given by (6.1).

6.1.4 Non-interfering host signal

There are some scenarios in which host-signal interference is either small or non-existent. For example, the watermark may be embedded only in host signal components that have a small amount of energy, especially if robustness to intentional attacks or lossy compression is not required. Alternatively, the host signal may be large, but available to the decoder. We treat both of these cases below.

In the limit of small host signals ($\mathsf{x} \to 0$), Fig. 2-2 reduces to the classical communication problem considered in many textbooks [13] since $\mathsf{s} \to \mathsf{e}$. In this limit, of course, the capacity is the mutual information between $\mathsf{e} = \mathsf{s}$ and y maximized over all $p_{\mathsf{e}}(\cdot)$ such that $E[\mathsf{e}^2] \leq D_{\mathsf{s}}$. In the additive white Gaussian noise channel case, the capacity is well known to be [13]

$$C_{\mathsf{x}\to 0} = \frac{1}{2}\log_2(1 + \mathrm{DNR}),$$

which, again, equals the capacity (6.1) when the host signal is not small. By examining (6.5), (6.18), and (6.26) in the limit of small host signals ($\mathrm{SNR}_x \to 0$), we see that distortion-compensated QIM with any α, regular QIM, and additive spread spectrum, respectively, are all optimal in this case.

As discussed in Chap. 4, when the host signal is not necessarily small but is known at the decoder, then the capacity is given by (4.8). Again, the maximization is subject to a distortion constraint, which in the case of white noise is $E[\mathsf{e}^2] \leq D_{\mathsf{s}}$. Because subtracting a known constant from y does not change mutual information, we can equivalently write

$$C = \max_{p_{\mathsf{e}|\mathsf{x}}(e|x)} I(\mathsf{e}; \mathsf{y} - \mathsf{x}|\mathsf{x}).$$

We note that $\mathbf{y} - \mathbf{x} = \mathbf{e} + \mathbf{n}$, so in the case of an AWGN channel the capacity is again

$$C_{\text{Gauss,known}} = \frac{1}{2}\log_2(1 + \text{DNR}),$$

where the maximizing distribution $p_{\mathbf{e}|x}(e|x)$ is a zero mean Gaussian distribution with variance $D_{\mathbf{s}}$. Again, both QIM and spread spectrum are optimal in this case. Quantizers of optimal QIM systems have reconstruction sequences $\mathbf{s}_i$ chosen iid from a zero mean Gaussian distribution with variance $\sigma_x^2 + D_{\mathbf{s}}$, and optimal spread spectrum systems add zero mean iid Gaussian sequences with variance $D_{\mathbf{s}}$ to the host signal.

6.2 Capacities for Embedding Data within Multimedia Host Signals

The capacity expressions in Sec. 6.1 apply to arbitrary host and noise covariance matrices and, thus, these achievable rate-distortion-robustness expressions are quite relevant to many of the multimedia applications mentioned in Chap. 1, especially those where one faces incidental channel degradations (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, thermal noise, adjacent channel and co-channel interference (which may be encountered in digital audio broadcasting (DAB) applications, for example), and residual noise after appropriate equalization of intersymbol interference channels or slowly varying fading channels. 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 FM DAB applications, conventional receivers demodulate the host analog FM signal from the channel output, not from the composite signal, which is available only at the transmitter. Similarly, 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 use the capacity expressions of this chapter to conveniently determine the achievable

embedded rate per unit of host signal bandwidth and per unit of received host signal degradation. In particular, we show in this section that this capacity is about 1/3 bit per second (b/s) for every Hertz (Hz) of host signal bandwidth and every dB drop in received host signal-to-noise ratio (SNR).

We examine two cases, one where the host signal is an analog signal and one where the host signal is a digital signal. We also point out some connections between our results and the problem of communication over the Gaussian broadcast channel [13, Ch. 14].

6.2.1 Analog host signals

In each of the cases considered in Sec. 6.1, the measure of distortion, and hence the DNR, is defined to make the embedding-induced distortion signal "look like" the channel noise, again the idea being that if channel noise distortion to the host signal is perceptually acceptable, then an embedding-induced distortion signal of the same power spectrum will also be perceptually acceptable. As discussed in those sections, 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

$$10 \log_{10}(1 + \text{DNR}) \text{ dB}. \tag{6.15}$$

In the white noise case ($K_{\boldsymbol{n}} = \sigma_n^2 I$), for example, the embedding-induced distortion looks like white noise with variance $D_{\mathbf{s}}$. With no embedding, one would have had a received host signal-to-noise ratio of $\text{SNR}_x = \sigma_x^2/\sigma_n^2$. Due to the additional interference from the embedding-induced distortion, however, the received host SNR drops to

$$\frac{\sigma_x^2}{D_{\mathbf{s}} + \sigma_n^2} = \frac{\text{SNR}_x}{1 + \text{DNR}},$$

a drop of $1 + \text{DNR}$.

Since the capacity in bits per dimension (bits per host signal sample) is given by (6.1), and there are two independent host signal samples per second for every Hertz of host signal

Host Signal	Bandwidth	Capacity
NTSC video	6 MHz	2.0 Mb/s/dB
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 6.1: 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.

bandwidth [26], the capacity in bits per second per Hertz is

$$C = \log_2(1 + \mathrm{DNR}) \text{ b/s/Hz}. \tag{6.16}$$

Taking the ratio between (6.16) and (6.15), we see that the "value" in embedded rate of each dB drop in received host signal quality is

$$C = \frac{\log_2(1 + \mathrm{DNR})}{10 \log_{10}(1 + \mathrm{DNR})} = \frac{1}{10} \log_2 10 \approx 0.3322 \text{ b/s/Hz/dB} \tag{6.17}$$

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 6.1.

6.2.2 Coded digital host signals

When the host signal is a coded digital signal, one could, of course, apply the above analysis to determine the achievable embedding rate for a given SNR degradation to this coded digital host signal. An alternative measure of the received host signal quality is the capacity of the corresponding host digital channel. For example, in the case of white noise and a white host signal,[2] if there were no embedding, the capacity corresponding to a host digital signal power of σ_x^2 and a noise variance of σ_n^2 is

$$R_0 = \frac{1}{2} \log_2(1 + \mathrm{SNR}_x).$$

[2] As is well known [13], white Gaussian coded signals are capacity-achieving for transmission over additive white Gaussian noise channels, so a white, Gaussian model for the coded host digital signal is actually a pretty good model in this case.

Embedding an additional digital signal within the host digital signal drops the host digital capacity to

$$R_1 = \frac{1}{2}\log_2\left(1 + \frac{\mathrm{SNR}_x}{1+\mathrm{DNR}}\right)$$

due to the drop in received host signal-to-noise ratio of $1 + \mathrm{DNR}$. Unlike in the case of an analog host signal, if one must actually lower the rate of the coded host digital signal as a result of the embedding, then one may have to redesign both the digital encoder that generates this coded digital host signal and the corresponding decoder. However, there may still be an advantage to this configuration of embedding one digital signal within another over simply designing a single new digital system that encodes both the host message and embedded message into a single digital signal. For example, the decoder for the host digital signal is different from the decoder for the embedded digital signal so information in the embedded channel is kept secret from those with decoders for the host signal. The embedded digital channel rate is given by (6.1),

$$R_2 = \frac{1}{2}\log_2(1 + \mathrm{DNR})$$

so that the combined rate of the two channels is

$$R_1 + R_2 = \frac{1}{2}\log_2(1 + \mathrm{DNR} + \mathrm{SNR}_x) \geq R_0.$$

Because the combined rate is greater than the original rate R_0 of the no-embedding case, the rate R_2 of the embedded signal is actually higher than the loss in rate of the host signal, *i.e.*, one bit in the host signal buys more than one bit in the embedded signal. Of course, this apparent increase in total capacity comes at the cost of increased total power, which is $D_{\mathbf{s}} + \sigma_x^2$. Still, the combined rate $R_1 + R_2$ is as large as the achievable rate using a *single* digital signal with this same total power, indicating that creating two signals that can be decoded separately results in no loss.

6.2.3 Connections to broadcast communication

We conclude this section by commenting on the connection between information embedding and broadcast communication [13, Ch. 14], where a single transmitter sends data to multiple receivers (decoders). For example, the downstream (base-to-mobile) channel in a cellular

telephone system is a broadcast channel. Our analyses in Secs. 6.2.1 and 6.2.2 apply to two special cases of the broadcast communication problem. In each case a single transmitter sends data to a host decoder and an embedded information decoder.

In Sec. 6.2.1 the host decoder is the identity function ($\hat{x} = y$), the host signal is an analog message, and the distortion measure is (possibly weighted) squared error distortion. The constraint on the host decoder to be the identity function arises, for example, from a backwards-compatibility requirement.

In contrast, we drop the backwards-compatibility requirement in Sec. 6.2.2, allowing a different host decoder (with rate R_1) in the broadcast case than the decoder (with rate R_0) in the single user case. Also, both the host signal and embedded information are digital signals. Indeed, the rate pair (R_1, R_2) is the achievable rate pair for the Gaussian broadcast channel [13, Ch. 14] when broadcasting independent information to two different receivers with the same noise variance. However, if one were to use superposition coding [12], the usual method for achieving capacity on broadcast channels, the embedded information decoder would need to know the host signal codebook so that the decoder could decode the host signal and subtract it from the channel output before decoding the embedded signal. This type of decoding is sometimes called "onion peeling" or successive cancellation. The above discussion in Sec. 6.2.2 shows that one can actually achieve the same rates without requiring that the embedded information decoder have access to the host signal codebook.

6.3 Gaps to Capacity

The capacity (6.1) gives the ultimate performance limit that is achievable by any information embedding system. When designing these systems, we often impose certain structure on the embedding function $\mathbf{s}(\mathbf{x}, m)$ so that we can understand how the system will behave. For example, as discussed in Chap. 3, the structure of QIM embedding functions allow us to conveniently trade off rate, distortion, and robustness by adjusting the number of quantizers, the quantization cell sizes and shapes, and the minimum distance between quantizers, respectively. Similarly, one achieves rate-distortion-robustness trade-offs in an amplitude-modulation spread spectrum system by adjusting the number of different amplitudes, the magnitudes of the amplitudes, and the differences between amplitudes. Imposing such structure allows one to search for the best embedding functions within a restricted class.

Although finding the best embedding function within a restricted class may be easier than finding the best embedding function over a large class, one incurs some risk, of course, that the restricted class may not contain very good embedding functions.

In this section, therefore, we discuss the "goodness" of the best possible embedding functions that lie within certain embedding function classes. In particular, we examine the performance gaps to capacity of the best possible embedding functions within the distortion-compensated QIM, regular QIM (QIM without distortion compensation), and additive spread spectrum classes. We also consider the gap to capacity of uncoded STDM and uncoded generalized LBM with uniform scalar quantization. We restrict our attention to the white host, white noise case since, as discussed in Sec. 6.1, one can transform the more general colored host, colored noise case into the white host, white noise case.

6.3.1 Optimality of distortion-compensated QIM

In Sec. 4.1.2 we showed that the condition (4.2) on the maximizing distribution in (4.1) is a sufficient condition for the existence of a capacity-achieving DC-QIM codebook. As discussed in Sec. 6.1.1, the maximizing distribution in the white Gaussian host, white Gaussian noise case satisfies (6.4), which is indeed the same condition as (4.2). Therefore, there is no gap between DC-QIM and capacity in this case. Furthermore, the capacity-achieving distortion compensation parameter α is given by (6.6), which is the same as the SNR-maximizing α given by (3.11).

6.3.2 Regular QIM gap to capacity

If one sets $\alpha = 1$, one obtains a regular QIM embedding function with no distortion compensation. Then, if one chooses reconstruction points from the pdf implied by (6.4),[3] one can achieve a rate (6.5):

$$R_{\text{QIM}} \geq \frac{1}{2}\log_2\left(\text{DNR}\frac{1+\text{DNR}+\text{SNR}_x}{\text{DNR}+\text{SNR}_x}\right). \tag{6.18}$$

However, the converse is not true, *i.e.*, one cannot show that a QIM system cannot achieve a rate greater than (6.18), and thus (6.18) is only a lower bound on the capacity of QIM.

[3] The pdf of the reconstruction points $u = s$ in this case is $\mathcal{N}(0, D_{\mathbf{s}} + \sigma_x^2)$, which is not the same as the well-known rate-distortion optimal pdf [13] for quantizing Gaussian random variables, which is $\mathcal{N}(0, \sigma_x^2 - D_{\mathbf{s}})$.

One can quantify the gap between regular QIM and the Gaussian capacity (6.1) in terms of the additional DNR required by a regular QIM system to achieve the same rate as a capacity-achieving system. We show below that regular QIM asymptotically achieves capacity at high embedding rates and that at finite rates the gap is never more than $e \approx 4.3$ dB.

A QIM system can achieve the rate (6.18), but this lower bound on capacity of QIM is not tight. In fact, the expression (6.18) actually approaches $-\infty$ in the limit of low DNR. However, we can determine a tighter lower bound on the capacity of spread-transform QIM, a subclass of QIM methods. Since these spread-transform QIM methods are special cases of QIM methods, this tighter lower bound is also a lower bound on the capacity of QIM.

Spread-transform QIM is a generalization of spread-transform dither modulation (See Sec. 5.2.) in which the host signal vector $\mathbf{x} = [x_1 \cdots x_N]^T$ is projected onto $N/L_{\rm ST}$ orthonormal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{N/L_{\rm ST}} \in \Re^N$ to obtain transformed host signal samples $\tilde{x}_1, \ldots, \tilde{x}_{N/L_{\rm ST}}$, which are quantized using QIM. Because projection onto the vectors $\mathbf{v}_i$ represents a change of orthonormal basis, the transformed host signal samples and the transformed noise samples $\tilde{n}_1, \ldots, \tilde{n}_{N/L_{\rm ST}}$, which are the projections of the original noise vector $\mathbf{n} = [n_1 \cdots n_N]^T$ onto the orthonormal vectors $\mathbf{v}_i$, are still independent, zero-mean, Gaussian random variables with the same variance as the original host signal and noise samples, respectively. However, if the distortion per original host signal sample is $D_{\rm s}$, then the distortion per transformed host signal sample is $L_{\rm ST} D_{\rm s}$. Thus, we obtain a "spreading gain" of $L_{\rm ST}$ in terms of DNR, but the number of bits embedded per original host signal sample is only $1/L_{\rm ST}$ times the number of bits embedded per transformed host signal sample. Thus, one can determine an achievable rate $R_{\rm STQIM}$ of spread-transform QIM by appropriately modifying (6.18) to obtain

$$\begin{aligned} R_{\rm STQIM} &\geq \frac{1}{2L_{\rm ST}} \log_2 \left(L_{\rm ST} \cdot {\rm DNR} \frac{1 + L_{\rm ST} \cdot {\rm DNR} + {\rm SNR}_x}{L_{\rm ST} \cdot {\rm DNR} + {\rm SNR}_x} \right) \\ &\geq \frac{1}{2L_{\rm ST}} \log_2 (L_{\rm ST} \cdot {\rm DNR}). \end{aligned} \tag{6.19}$$

To upper bound the gap between QIM and capacity we first recognize from (6.19) that the minimum DNR required for QIM to achieve a rate R asymptotically with large N is

$${\rm DNR}_{\rm QIM} \leq \frac{2^{2L_{\rm ST}R}}{L_{\rm ST}}, \tag{6.20}$$

which is minimized at $L_{\rm ST} = 1/(2R\ln 2)$. One may wonder if one can actually obtain this spreading gain, however, since the description of the spread-transform operation above requires that $N/L_{\rm ST}$ be a positive integer less than or equal to N. If $N/L_{\rm ST}$ must be rounded to the nearest integer, the spreading gain has lower and upper bounds

$$\frac{N}{\left(\frac{N}{L_{\rm ST}}+0.5\right)} \le \frac{N}{\text{round}\left(\frac{N}{L_{\rm ST}}\right)} \le \frac{N}{\left(\frac{N}{L_{\rm ST}}-0.5\right)}$$

and, thus, still approaches $L_{\rm ST}$ in the limit of large N. Therefore, the rounding operation has an asymptotically negligible effect on the spreading gain. However, $L_{\rm ST} \ge 1$ even in the limit of large N to have $N/L_{\rm ST} \le N$. Thus, if one sets

$$L_{\rm ST} = \max\left\{\frac{1}{2R\ln 2}, 1\right\}, \tag{6.21}$$

then (6.20) remains a valid upper bound on the required DNR for a QIM method to achieve a rate R. From (6.1) we see that the minimum DNR required for a capacity-achieving method to achieve a rate R is

$$\text{DNR}_{\rm opt} = 2^{2R} - 1.$$

Combining this expression with (6.20), we see that the gap between QIM and the Gaussian capacity is at most

$$\frac{\text{DNR}_{\rm QIM}}{\text{DNR}_{\rm opt}} \le \frac{2^{2L_{\rm ST}R}}{L_{\rm ST}\left(2^{2R}-1\right)}. \tag{6.22}$$

This expression is plotted in Fig. 6-2, where $L_{\rm ST}$ is given by (6.21).

We now examine the asymptotic limits of (6.22) at low and high rates. Eq. (6.21) implies $L_{\rm ST} = 1/(2R\ln 2)$ in the limit of small R, so in this limit (6.22) approaches

$$\begin{aligned}\frac{\text{DNR}_{\rm QIM}}{\text{DNR}_{\rm opt}} &\le \frac{2^{2L_{\rm ST}R}}{L_{\rm ST}\left(2^{2R}-1\right)} \\ &= \frac{2^{1/\ln 2}(2R\ln 2)}{2^{2R}-1} \\ &= e\,\frac{2R\ln 2}{2^{2R}-1} \to e, \qquad \text{as } R\to 0.\end{aligned}$$

The third line follows from the identity $x^{1/\ln x} = e$ for any x, which one can derive by noting that $\ln x^{1/\ln x} = (1/\ln x)\ln x = 1$. Thus, the gap is at most a factor of e (approximately

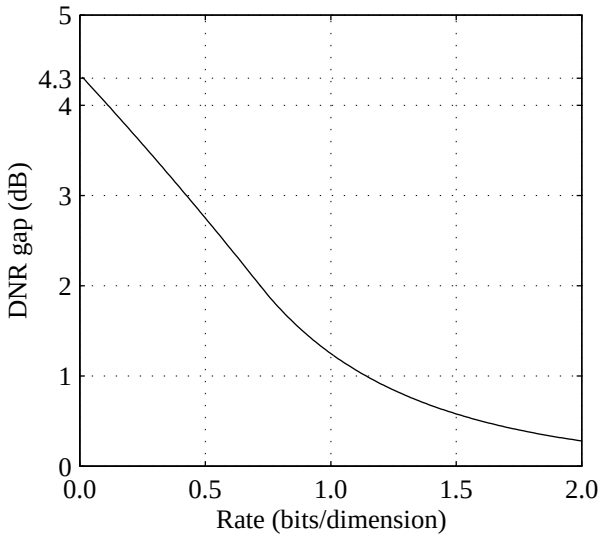

Figure 6-2: DNR gap between spread-transform QIM and Gaussian capacity. The spreading length is restricted to be greater than or equal to 1. The maximum gap is a factor of e, which is approximately 4.3 dB.

4.3 dB) in the limit of low rates. In the limit of large R, (6.21) implies $L_{\rm ST} = 1$ so (6.22) approaches

$$\frac{\mathrm{DNR}_{\rm QIM}}{\mathrm{DNR}_{\rm opt}} = \frac{2^{2R}}{2^{2R}-1} \to 1, \qquad \text{as } R \to \infty.$$

Thus, QIM asymptotically achieves capacity at high embedding rates.

As we described in Sec. 6.2, in many applications one may be concerned about the degradation to the received host signal, which is $(1 + \mathrm{DNR})$ rather than DNR. The gap in DNR (6.22) is larger than the gap in $(1 + \mathrm{DNR})$, which has a corresponding upper bound

$$\frac{1 + \mathrm{DNR}_{\rm QIM}}{1 + \mathrm{DNR}_{\rm opt}} \leq \frac{1 + \frac{2^{2RL_{\rm ST}}}{L_{\rm ST}}}{2^{2R}}.$$

This gap is plotted in Fig. 6-3 as a function of $2R$, the rate in b/s/Hz. Again, $L_{\rm ST}$ is given by (6.21) since minimizing $\mathrm{DNR}_{\rm QIM}$ also minimizes $1 + \mathrm{DNR}_{\rm QIM}$. Thus, for example, a digital rate of 1 b/s/Hz using QIM requires at most 1.6 dB more drop in analog channel quality than the approximately 3-dB drop required for a capacity achieving method (Sec. 6.2).

6.3.3 Uncoded STDM gap to capacity

The performance of the best QIM methods can approach the Gaussian capacity at high rates and is within 4.3 dB of capacity at low rates, indicating that the QIM class is large

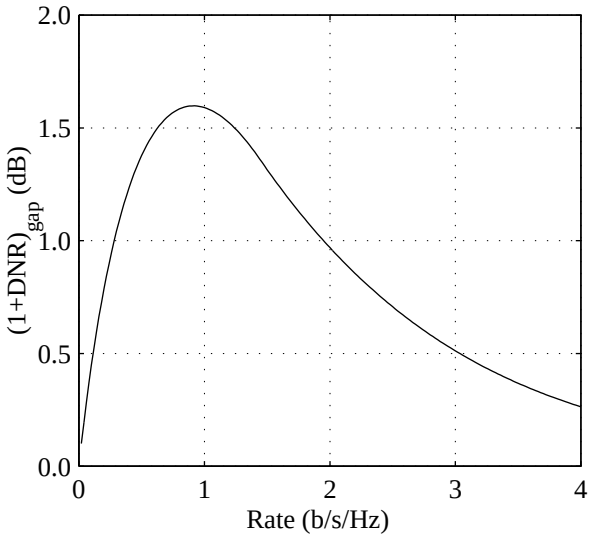

Figure 6-3: Received host SNR gap (1+DNR) between spread-transform QIM and capacity. The spreading length is restricted to be greater than or equal to 1. One bit/dimension equals 2 b/s/Hz.

enough to include very good embedding functions and decoders. In this section we consider the achievable performance of uncoded spread-transform dither modulation (STDM) with uniform scalar quantization since STDM is an important, low-complexity realization of QIM.

The gap between uncoded STDM and the Gaussian capacity (6.1) can easily be quantified for low rates ($R_m \leq 1$), which are typical in many applications, at a given probability of error. From Fig. 5-4, we see that an upper bound on the bit-error probability of uncoded STDM is

$$P_b \leq 2\,Q\left(\sqrt{\frac{d_{\min}^2}{4\sigma_n^2}}\right),$$

where, as in Chap. 3, $Q(\cdot)$ is the Gaussian Q-function (3.6). This bound is reasonably tight for low error probabilities, and from (5.4) we can write this probability of error in terms of the rate-normalized distortion-to-noise ratio $\mathrm{DNR}_{\mathrm{norm}} = \mathrm{DNR}/R_m$,

$$P_b \approx 2\,Q\left(\sqrt{\frac{3\cdot\mathrm{DNR}}{4R_m}}\right) = 2\,Q\left(\sqrt{\frac{3}{4}\mathrm{DNR}_{\mathrm{norm}}}\right). \tag{6.23}$$

From (6.1), a capacity-achieving method can achieve arbitrarily low probability of error as long as $R_m \leq C_{\mathrm{Gauss}}$ or

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

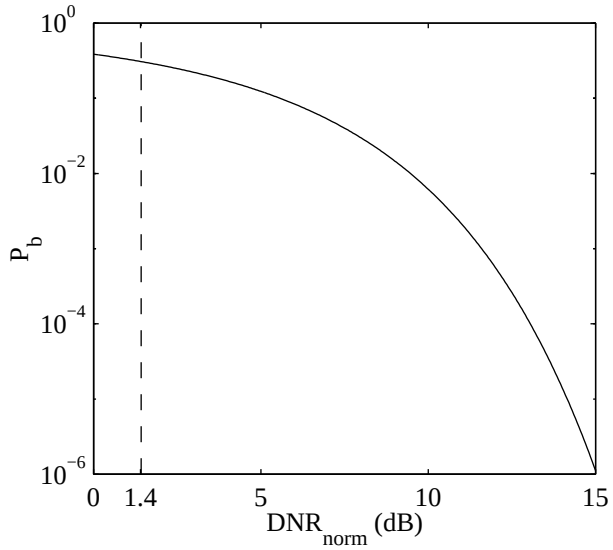

Figure 6-4: Uncoded spread-transform dither modulation (STDM) gap to Gaussian capacity. The solid curve shows the bit-error probability for uncoded STDM as a function of rate-normalized distortion-to-noise ratio ($\mathrm{DNR}_{\mathrm{norm}}$). The dashed curve is the minimum required $\mathrm{DNR}_{\mathrm{norm}}$ for reliable information-embedding for any embedding method.

For small R_m, $2^{2R_m} - 1 \approx 2R_m \ln 2$ so the minimum required $\mathrm{DNR}_{\mathrm{norm}}$ for arbitrarily low probability of error is

$$\mathrm{DNR}_{\mathrm{norm}} \geq 2 \ln 2 \approx 1.4 \text{ dB}. \tag{6.24}$$

The probability of error P_b of STDM is plotted as a function of $\mathrm{DNR}_{\mathrm{norm}}$ in Fig. 6-4. The required $\mathrm{DNR}_{\mathrm{norm}}$ for a given P_b can be compared to (6.24) to determine the gap to capacity. For example, at an error probability of 10^{-6}, uncoded STDM is about 13.6 dB from capacity. One can reduce this gap by at least 9.3 dB through channel coding, vector quantization, and non-dithered quantization. The remaining gap (at most 4.3 dB) is the gap between QIM and capacity and can be closed with distortion compensation. In Chap. 8 we illustrate that one can fairly easily close the gap between uncoded STDM (with uniform scalar quantizers) and capacity by about 6 dB using practical channel codes and distortion compensation.

6.3.4 Uncoded LBM gap to capacity

Again, from App. C the distortion-normalized minimum distance for LBM with uniform scalar quantization is a factor of $7/4 \approx 2.43$ dB worse than that of STDM (5.4). Thus, the

LBM counterpart to (6.23) is that the bit-error probability of uncoded LBM is

$$P_b \approx 2\,Q\left(\sqrt{\frac{3}{7}\mathrm{DNR}_{\mathrm{norm}}}\right). \tag{6.25}$$

Then, the gap to capacity of uncoded LBM at an error probability of 10^{-6} is about 16 dB, 2.4 dB more than the 13.6-dB gap of uncoded STDM.

6.3.5 Spread spectrum gap to capacity

Additive methods such as spread spectrum linearly combine a watermark signal with the host signal, $\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{w}(\boldsymbol{m})$, so that the distortion signal in Fig. 2-2

$$\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{m}) = \boldsymbol{w}(\boldsymbol{m})$$

is not a function of the host signal. Thus,

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} = \boldsymbol{e} + \boldsymbol{x} + \boldsymbol{n}.$$

The distortion constraint still constrains the size of $\boldsymbol{e}$ to $E[\boldsymbol{e}^2] = D_{\mathbf{s}}$ so that in the Gaussian case considered here, the achievable rate of a spread spectrum method is the well-known [13] Gaussian channel capacity, treating both $\boldsymbol{x}$ and $\boldsymbol{n}$ as interference sources,

$$R_{\mathrm{SS}} = \frac{1}{2}\log_2\left(1 + \frac{D_{\mathbf{s}}}{\sigma_x^2 + \sigma_n^2}\right) = \frac{1}{2}\log_2\left(1 + \frac{\mathrm{DNR}}{\mathrm{SNR}_x + 1}\right), \tag{6.26}$$

where, again, SNR_x is the ratio between the host signal variance and the channel noise variance. (This rate is also the capacity when $\boldsymbol{n}$ is non-Gaussian, but still independent of $\boldsymbol{s}$, and a correlation detector is used for decoding [25].) By comparing (6.26) to (6.1) we see that the gap to capacity of spread-spectrum is

$$\frac{\mathrm{DNR}_{\mathrm{ss}}}{\mathrm{DNR}_{\mathrm{opt}}} = \mathrm{SNR}_x + 1.$$

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 regular QIM.

In the high signal-to-distortion (SDR) limit where $\sigma_x^2/D_{\mathbf{s}} \gg 1$, which is of interest

for many high-fidelity applications, the achievable rate of spread spectrum (6.26) clearly approaches zero. This result is one more example of the inability of spread spectrum methods to reject host signal interference, in contrast to dither modulation, QIM, and other optimal or near optimal embedding methods.

6.3.6 Known-host case

As discussed in Sec. 6.1.4, both capacity-achieving QIM and capacity-achieving spread spectrum methods exist when the host signal is known at the decoder. Although coded binary dither modulation with uniform, scalar quantization is not optimal in this case, for AWGN channels one can achieve performance within $\pi e/6 \approx 1.53$ dB of capacity as we show below. We consider the case of dither signals with a uniform distribution over the interval $[-\Delta/2, \Delta/2]$. In this case,

$$s = q(x + d) - d = x + e,$$

where the quantization error e is uniformly distributed over the interval $[-\Delta/2, \Delta/2]$ and statistically independent of x (even though e is a function of x and d) [23]. Thus, the achievable rate $I(e; e + n)$ is slightly lower than the case where e is Gaussian. The entropy power inequality can be used to show that the decrease in achievable rate is bounded by [40]

$$C_{\text{Gauss,known}} - R_{\text{dith}} \leq \frac{1}{2} \log_2 \frac{1 + \text{DNR}}{1 + (6/\pi e)\text{DNR}}. \tag{6.27}$$

This gap approaches the upper limit of $\frac{1}{2} \log_2 \frac{\pi e}{6} \approx 0.2546$ bits/dimension as the distortion-to-noise ratio gets large. For any finite DNR, the gap is smaller. By subtracting the upper bound on the gap (6.27) from the capacity (6.1), one obtains a lower bound on the achievable rate of this type of dither modulation:

$$R_{\text{dith}} \geq \frac{1}{2} \log_2 \left(1 + \frac{6}{\pi e} \text{DNR} \right). \tag{6.28}$$

Thus, dither modulation with uniform scalar quantization in this case is at most $\pi e/6 \approx 1.53$ dB from capacity.

Chapter 7

Intentional Attacks

Intentional, distortion-constrained attacks may be encountered in copyright, authentication, and covert communication applications. In a digital video disc (DVD) copyright application, for example, the attacker may try to remove the watermark from illegally copied video so that a standards compliant DVD player will not recognize the video as watermarked and will thus play the disc. In authentication applications, if an attacker can successfully remove authentication watermarks so that even authentic documents are rejected as unauthentic, then the authentication system will be rendered useless. In the context of covert communications, even if an adversary is unable to detect the presence of a hidden message, an attacker can disrupt its communication by degrading the composite signal carrying the message.

In each of these examples, the attacker faces a distortion constraint on his or her signal manipulations. In the DVD copyright application, the distortion constraint arises because the attacker desires a copy of the video that is of acceptable quality. In the case of authentication, degrading the signal too much to remove the authentication watermark results in a signal that may indeed be no longer authentic due to its unacceptably low quality. In the covert communication application, the attacker may be prohibited from degrading the host signal so severely that it will no longer be useful for its intended purpose. For example, the host signal may communicate useful information over a network to a group of only partially trusting allies of which the attacker is a member. That attacker suspects that two other members of this group wish to covertly communicate additional information to each other by embedding the information in the host signal. The attacker wishes to disrupt such

potential covert communication, but cannot destroy the host signal in the process.

An attacker's ability to prevent reliable watermark decoding depends on the amount of knowledge that the attacker has about the embedding and decoding processes. To limit such knowledge, some digital watermarking systems use keys, parameters that allow appropriate parties to 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" or "universally-accessible" system requires only standardization of the decoder so that everyone will agree on the decoded watermark, and hence, the owner of the copyright.

Although the attacker does not know the key in a private-key scenario, he or she may know the basic algorithm used to embed the watermark. In [30], Moulin and O'Sullivan model such a scenario by assuming that the attacker knows the codebook distribution, but not the actual codebook. As we discuss below, in this private-key scenario the results of Moulin and O'Sullivan imply that distortion-compensated QIM methods are optimal (capacity-achieving) against squared error distortion-constrained attackers. In the absence of keys, however, the attacker does know the codebook, and the bounded perturbation channel and the bounded host-distortion channel models of Chap. 2 are better models for attacks in these no-key scenarios. As we show in this chapter, QIM methods in general, and dither modulation in particular, achieve provably better rate-distortion-robustness trade-offs than both spread spectrum and generalized LBM techniques against these classes of

attacks on no-key systems.

7.1 Attacks on Private-key Systems

Moulin and O'Sullivan have derived both the capacity-achieving distribution and an explicit expression for the capacity (4.1) in the case where the host is white and Gaussian and the attacker faces an expected perturbation energy constraint $E[\|\mathbf{n}\|^2] \leq \sigma_n^2$. In this case the capacity is [30]

$$C_{\text{Gauss,private}} = \frac{1}{2}\log_2\left(1 + \frac{\text{DNR}_{\text{attack}}}{\beta}\right), \qquad \beta = \frac{\text{SNR}_{x,\text{attack}} + \text{DNR}_{\text{attack}}}{\text{SNR}_{x,\text{attack}} + \text{DNR} - 1},$$

where $\text{DNR}_{\text{attack}} = D_{\mathbf{s}}/\sigma_n^2$ is the distortion-to-perturbation ratio and $\text{SNR}_{x,\text{attack}} = \sigma_x^2/\sigma_n^2$ is the host signal-to-perturbation ratio. The maximizing distribution is [30]

$$\boldsymbol{u} = \boldsymbol{e} + \alpha_{\text{Gauss,private}}\boldsymbol{x},$$

where $\boldsymbol{e} \sim \mathcal{N}(0, D_{\mathbf{s}})$ is statistically independent of $\boldsymbol{x}$ and

$$\alpha_{\text{Gauss,private}} = \frac{\text{DNR}_{\text{attack}}}{\text{DNR}_{\text{attack}} + \beta}. \tag{7.1}$$

Since this distribution satisfies the condition (4.2), distortion-compensated QIM can achieve capacity against these attacks. Eq. (7.1) gives the optimal distortion-compensation parameter.

Moulin and O'Sullivan have also considered the case of host signals that are not necessarily Gaussian but that have zero mean, finite variance, and bounded and continuous pdfs. In the limit of small $D_{\mathbf{s}}$ and σ_n^2, a limit of interest in high-fidelity applications, the capacity approaches

$$C_{\text{high-fidelity}} \to \frac{1}{2}\log_2\left(1 + \text{DNR}_{\text{attack}}\right),$$

and the capacity-achieving distribution approaches

$$\boldsymbol{u} = \boldsymbol{e} + \alpha_{\text{high-fidelity}}\boldsymbol{x},$$

where, again, $\boldsymbol{e} \sim \mathcal{N}(0, D_{\mathbf{s}})$ is statistically independent of $\boldsymbol{x}$ [30]. Again, since this distribu-

tion satisfies the condition (4.2), distortion-compensated QIM can achieve capacity in this high-fidelity limit. The capacity-achieving distortion-compensation parameter is [30]

$$\alpha_{\text{high-fidelity}} = \frac{\text{DNR}_{\text{attack}}}{\text{DNR}_{\text{attack}} + 1}.$$

7.2 Attacks on No-key Systems

In this section, we examine worst case in-the-clear attacks, attacks that arise when the attacker has full knowledge of the embedding and decoding processes including any keys. We consider two models for such attackers from Sec. 2.3: (1) the bounded perturbation channel model in which the squared error distortion between the channel input and channel output is bounded and (2) the bounded host-distortion channel model in which the squared error distortion between the host signal and channel output is bounded.

7.2.1 Bounded perturbation channel

In this section we characterize the achievable performance of binary dither modulation with uniform scalar quantization, spread spectrum, and low-bit(s) modulation when one wants guaranteed error-free decoding against all bounded perturbation attacks.

Binary dither modulation with uniform scalar quantization

One can combine the guaranteed error-free decoding condition (3.5) for a minimum distance decoder (3.8) with the distortion-normalized minimum distance (5.4) of binary dither modulation with uniform scalar quantization to compactly express its achievable performance as

$$\frac{(d^2_{\min}/D_{\mathbf{s}})D_{\mathbf{s}}}{4N\sigma_n^2} = \gamma_c \frac{3/4}{NR_m}\frac{D_{\mathbf{s}}}{\sigma_n^2} > 1, \tag{7.2}$$

or, equivalently, its achievable rate[1] as

$$R_m < \frac{3\gamma_c}{4N}\frac{D_{\mathbf{s}}}{\sigma_n^2}. \tag{7.3}$$

[1]One can view these achievable rates (7.3) as the deterministic counterpart to the more conventional notions of achievable rates and capacities of random channels discussed in Chaps. 4 and 6.

Thus, for example, at a fixed rate R_m to tolerate more perturbation energy σ_n^2 requires that we accept more expected distortion $D_{\mathbf{s}}$. Eq. (7.2) conveniently relates design specifications to design parameters for dither modulation methods. For example, if the design specifications require an embedding rate of at least R_m and robustness to perturbations of at least σ_n^2 in energy per sample, then (7.2) gives the minimum embedding-induced distortion that must be introduced into the host signal, or equivalently via (5.3) the minimum average squared quantization step size $\frac{1}{L}\sum_k \Delta_k^2$ to achieve these specifications. Finally, we see that γ_c is the improvement or gain in the achievable rate-distortion-robustness trade-offs due to the error correction code.

Spread spectrum

The nonzero minimum distance of QIM methods offers quantifiable robustness to perturbations, even when the host signal is not known at the decoder. In contrast, spread-spectrum methods offer relatively little robustness if the host signal is not known at the decoder. As discussed in Sec. 2.4, these methods have linear embedding functions of the form

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{w}(m), \tag{7.4}$$

where $\mathbf{w}(m)$ is a pseudo-noise vector. From the definition of minimum distance (3.2),

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

This zero minimum distance property of spread spectrum methods is illustrated in Fig. 7-1.

Thus, although these methods 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, *i.e.*, no achievable rate expression analogous to (7.3) exists for additive spread spectrum. As is evident from (7.4), in a spread-spectrum system, $\mathbf{x}$ is an additive interference, which is often much larger than $\mathbf{w}$ due to the distortion constraint. In contrast, the quantization that occurs with quantization index modulation, provides immunity against

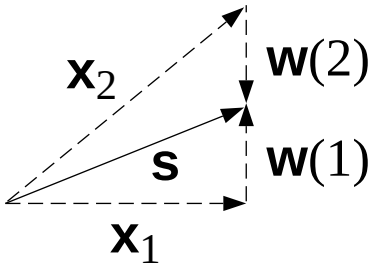

Figure 7-1: Zero minimum distance of spread spectrum embedding methods. The composite signal vector $\mathbf{s}$ lies in both signal sets, and thus, even with no perturbations ($\mathbf{y} = \mathbf{s}$) one cannot distinguish between $(\mathbf{x}, m) = (\mathbf{x}_1, 1)$ and $(\mathbf{x}, m) = (\mathbf{x}_2, 2)$.

this host signal interference, as discussed in Chap. 3.[2]

Low-bit modulation

As shown in App. C, the distortion-normalized minimum distance of LBM is about 2.43 dB worse (Eq. (C.3)) than that of dither modulation. Therefore, its achievable rate-distortion-robustness performance is also about 2.43 dB worse than (7.2).

7.2.2 Bounded host-distortion channel

As mentioned in Sec. 2.3, the bounded host-distortion channel model arises when the attacker's distortion is measured between the host signal and channel output. Unlike in the case of the bounded perturbation channel, considering performance against the worst possible channel output $\mathbf{y}$ satisfying the attacker's host-distortion constraint does not provide much insight. The channel output $\mathbf{y} = \mathbf{x}$ results in $D_{\mathbf{y}} = 0$, and this channel output contains no information about the embedded information m. Thus, this channel output is the worst case output, but it is not clear that an attacker can produce this output without knowledge of $\mathbf{x}$. The attacker can, however, exploit partial knowledge of the host signal, where such partial knowledge may be described, for example, by the conditional probability density $p_{\mathbf{x}|\mathbf{s}}(\mathbf{x}|\mathbf{s})$ of the host signal given observation of the channel input (composite signal).

Thus, in this section we measure robustness to attacks by the minimum expected distortion $D_{\mathbf{y}}$ for a successful attack, where the expectation is taken with respect to $p_{\mathbf{x}|\mathbf{s}}(\mathbf{x}|\mathbf{s})$. The ratio between $D_{\mathbf{y}}$ and the expected embedding-induced distortion $D_{\mathbf{s}}$ is the distortion

[2] Another way to understand this host-signal interference rejection is to consider, for example, that a quantized random variable has finite entropy while a continuous random variable has infinite entropy.

Table 7.1: 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 0 dB indicates that the attacker can actually improve the signal quality and remove the watermark simultaneously.

Embedding Method	Distortion Penalty $(D_{\mathbf{y}}/D_{\mathbf{s}})$
Regular QIM	$1 + \frac{d^2_{\text{norm}}}{4N} > 0$ dB
Binary Dith. Mod. w/uni. scalar quant.	2.43 dB $\geq 1 + \gamma_c \frac{3/4}{NR_m} > 0$ dB
DC-QIM	$-\infty$ dB
Spread Spectrum	$-\infty$ dB
LBM	≤ 0 dB
Binary LBM w/uni. scalar quant.	-2.43 dB

penalty that the attacker must pay to remove the watermark and, hence, is a figure of merit measuring the robustness-distortion trade-off at a given rate. Distortion penalties for regular QIM, binary dither modulation with uniform scalar quantization, distortion-compensated QIM, spread-spectrum, LBM, and binary LBM with uniform scalar quantization are derived below and are shown in Table 7.1. We see that of the methods considered, only QIM methods (including binary dither modulation with uniform scalar quantization) are robust enough such that the attacker must degrade the host signal quality to remove the watermark.

Regular QIM

We first consider the robustness of regular quantization index modulation. For any distortion measure, as long as each reconstruction point $\mathbf{s}$ lies at the minimum distortion point of its respective quantization cell, the QIM distortion penalty is greater than or equal to 1 since any output $\mathbf{y}$ that an attacker generates must necessarily lie away from this minimum distortion point. Equality occurs only if each quantization cell has at least two minimum distortion points, one of which lies in the incorrect decoder decision region. For expected squared-error distortion, the minimum distortion point of each quantization cell is its centroid, and one can express this distortion penalty in terms of the distortion-normalized

minimum distance and the signal length N, as we show below.

We use $\mathcal{R}$ to denote the quantization cell containing $\mathbf{x}$ and $p_{\mathbf{x}}(\mathbf{x}|\mathcal{R})$ 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 $\mathcal{R}$,

$$\int_{\mathcal{R}} (\mathbf{s} - \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}|\mathcal{R})\, d\mathbf{x} = \mathbf{0}. \tag{7.5}$$

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

$$D_{\mathbf{s}|\mathcal{R}} = \frac{1}{N} \int_{\mathcal{R}} \|\mathbf{s} - \mathbf{x}\|^2 p_{\mathbf{x}}(\mathbf{x}|\mathcal{R})\, d\mathbf{x}. \tag{7.6}$$

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

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

where we have used (7.6), the fact that $p_{\mathbf{x}}(\mathbf{x}|\mathcal{R})$ is a probability density function and, thus, integrates to one, and (7.5) to obtain the last line. For a successful attack, $\|\mathbf{n}\| \geq d_{\min}/2$ so

$$D_{\mathbf{y}|\mathcal{R}} \geq D_{\mathbf{s}|\mathcal{R}} + \frac{d_{\min}^2}{4N}.$$

Averaging both sides of this expression over all quantization cells $\mathcal{R}$ yields

$$D_{\mathbf{y}} \geq D_{\mathbf{s}} + \frac{d_{\min}^2}{4N}$$

so that our figure of merit for quantization index modulation methods is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} \geq 1 + \frac{d_{\min}^2/D_{\mathbf{s}}}{4N} = 1 + \frac{d_{\mathrm{norm}}^2}{4N}. \tag{7.7}$$

Thus, for any QIM method of nonzero distortion-normalized minimum distance d_{norm}, 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.

Binary dither modulation with uniform, scalar quantization

In the special case of coded binary dither modulation with uniform, scalar quantization, Eq. (5.4) gives d^2_{norm}. Due to the uniformity of the quantizers, the bound (7.7) is met with equality so that the attacker's distortion penalty (7.7) that must be paid to defeat the watermark in this case is

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} = 1 + \gamma_c \frac{3/4}{NR_m}. \tag{7.8}$$

Because the Hamming distance d_H of a block code cannot exceed the number of coded bits $NR_m(k_c/k_u)$,

$$\frac{\gamma_c}{NR_m} = \frac{d_H}{NR_m(k_c/k_u)} \leq 1,$$

where the first equality follows from the definition (5.2) of γ_c. Thus, an upper bound for the distortion penalty (7.8) in this case is

$$1 + \gamma_c \frac{3/4}{NR_m} \leq \frac{7}{4} \approx 2.43 \text{ dB}.$$

Although this penalty may seem modest, it is larger than that obtainable by either spread spectrum or low-bit(s) modulation, as we show below. The difficulty in obtaining large distortion penalties arises from the fact that an in-the-clear attacker can concentrate all of his or her distortion in the minimum distance direction in N-dimensional space.

As a final note, (7.8) implies that binary dither modulation with uniform, scalar quantization can defeat any attacker as long as

$$\left(1 + \gamma_c \frac{3/4}{NR_m}\right) \frac{D_{\mathbf{s}}}{D_{\mathbf{y}}} > 1,$$

an expression that is analogous to (7.2), which applied for the bounded perturbation channel, rather than the bounded host-distortion channel. In each case some multiple of the

ratio between the embedding-induced distortion and the attacker's distortion, a "distortion-to-noise ratio", must be greater than 1.

Distortion-compensated QIM

An in-the-clear attacker of a DC-QIM system knows the quantizers and can determine the watermark m after observing the composite signal $\mathbf{s}$. If the quantization cells are contiguous so that the distortion-compensation term in (3.10) does not move $\mathbf{s}$ out of the cell containing $\mathbf{x}$, then an attacker can recover the original host signal with the following attack:

$$\begin{aligned} \mathbf{y} &= \frac{\mathbf{s} - \alpha \mathbf{q}(\mathbf{s}; m, \Delta/\alpha)}{1-\alpha} \\ &= \frac{\mathbf{s} - \alpha \mathbf{q}(\mathbf{x}; m, \Delta/\alpha)}{1-\alpha} \\ &= \mathbf{x}. \end{aligned}$$

The final line follows simply by inverting (3.10). Thus, the attacker's distortion penalty $D_{\mathbf{y}}/D_{\mathbf{s}}$ is $-\infty$ dB. We see that although DC-QIM is optimal against additive Gaussian noise attacks and against squared error distortion-constrained attacks in private-key scenarios, it is in some sense "maximally suboptimal" against in-the-clear (no-key) attacks. Regular QIM, on the other hand, is almost as good as DC-QIM against additive Gaussian noise attacks (Chap. 6) and also resistant to in-the-clear attacks as discussed above. Thus, regular QIM methods may offer an attractive compromise when one requires resistance to both intentional attacks and unintentional attacks and one cannot employ a private key.

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 methods are not very attractive for universally-accessible digital watermarking applications. In contrast, the quantization that occurs in quantization index modulation methods effectively hides the exact value of the host signal even when the embedded information m is known, thus allowing universal access with a positive (in dB) attacker's distortion penalty.

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,

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

Without loss of generality, we assume that the reconstruction points of $\mathbf{q}(\cdot)$ are at the centroids of the quantization cells. One attack that completely removes information about m is to output these reconstruction points,

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

Since $\mathbf{y}$ is at a minimum distortion point of the quantization cell,

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

with equality only if both $\mathbf{s}$ and $\mathbf{y}$ are minimum distortion points. Thus, 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 7.1.), error correction coding does not improve low-bit(s) modulation in this context.

Binary low-bit modulation with uniform, scalar quantization

When the least significant bit of a uniform, scalar quantizer is modulated, the results in App. C imply that

$$D_{\mathbf{s}} = \frac{7}{48L} \sum_k \Delta_k^2,$$

while

$$D_{\mathbf{y}} = \frac{1}{12L} \sum_k \Delta_k^2.$$

Thus,

$$\frac{D_{\mathbf{y}}}{D_{\mathbf{s}}} = \frac{4}{7} \approx -2.43 \text{ dB}.$$

Chapter 8

Simulation Results

In Chap. 4 we argued that QIM, with the right preprocessing and postprocessing to move into and out of the correct domain, is a capacity-achieving information embedding structure. As discussed in Chap. 6, the "right" postprocessing in the Gaussian case is distortion compensation and no preprocessing is required. Furthermore, even with no distortion compensation, QIM methods can achieve performance within a few dB of capacity.

In general, though, one can achieve capacity only asymptotically with long signal lengths N, and hence, with large complexity and delay. In Chap. 5, therefore, we introduced low-complexity realizations of QIM methods that, of course, can be combined with distortion compensation, and in this chapter we present several simulation results to demonstrate the practically achievable performance of such realizations.

8.1 Uncoded Methods

In this section we present results for *uncoded* dither modulation with uniform, scalar quantization. These methods have extremely low computational complexity as discussed in Chap. 5. In the section following this one, we demonstrate the additional gains that one can achieve with practical error correction codes and distortion compensation.

8.1.1 Gaussian channels

As discussed in Chap. 6, the bit-error probability of uncoded spread-transform dither modulation (STDM) with uniform, scalar quantization is (6.23)

$$P_b \approx 2\,Q\left(\sqrt{\frac{3}{4}\mathrm{DNR}_{\mathrm{norm}}}\right)$$

for additive white Gaussian noise (AWGN) channels, where again

$$\mathrm{DNR}_{\mathrm{norm}} \triangleq \frac{\mathrm{DNR}}{R_m}. \tag{8.1}$$

For example, one can achieve a bit-error probability of about 10^{-6} at a $\mathrm{DNR}_{\mathrm{norm}}$ 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

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

where $\mathrm{DNR}_{\mathrm{norm}}$ is the minimum $\mathrm{DNR}_{\mathrm{norm}}$ necessary in (6.23) for a given P_b and where DNR is determined by channel conditions and the embedding-induced distortion.

This case is illustrated in Fig. 8-1, where despite the fact that the channel has degraded the composite image by over 12 dB, all 512 embedded bits are recovered without any errors from the 512-by-512 image. The actual bit-error probability is about 10^{-6}.

8.1.2 JPEG channels

The robustness of digital watermarking algorithms to common lossy compression algorithms such as JPEG is of considerable interest. A natural measure of robustness is the worst tolerable JPEG quality factor[1] for a given bit-error rate at a given distortion level and embedding rate. We experimentally determined achievable rate-distortion-robustness operating points for particular uncoded implementations of both STDM and "unspread dither modulation (UDM)", where we use UDM to refer to the case where there is no projection onto a spreading vector $\mathbf{v}$ and all host signal components are quantized with the same step size ($\Delta_1 = \Delta_2 = \cdots = \Delta_L$ in Sec. 5.1).

[1]The JPEG quality factor is a number between 0 and 100, 0 representing the most compression and lowest quality, and 100 representing the least compression and highest quality.

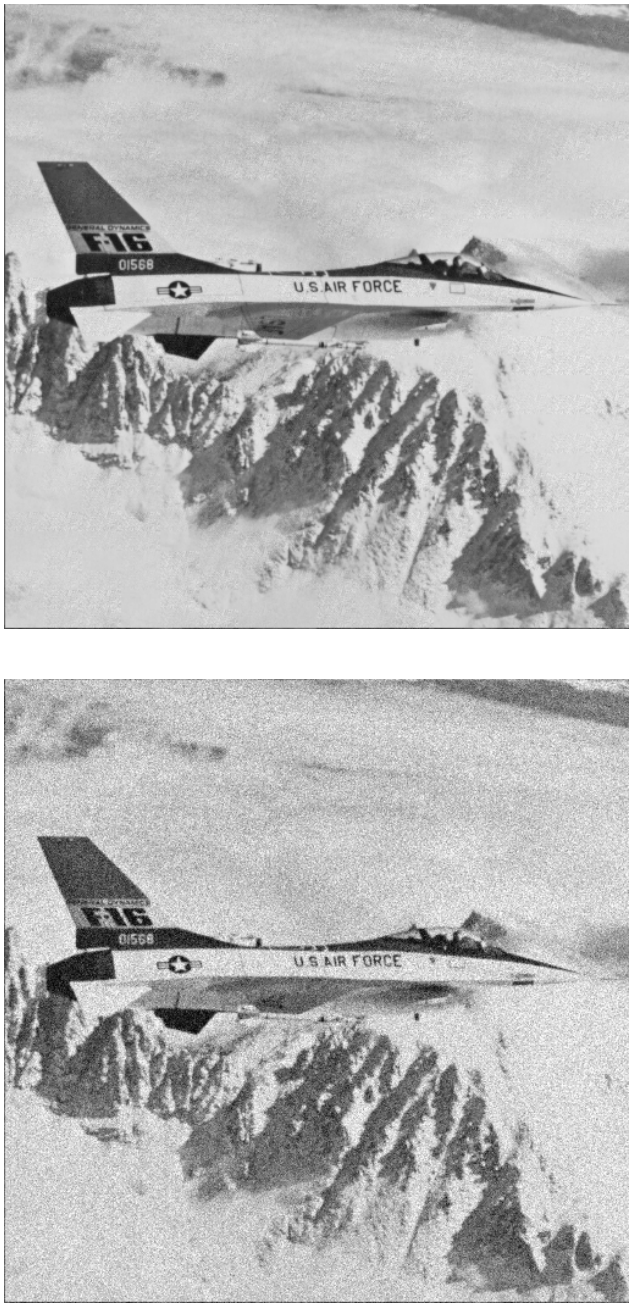

Figure 8-1: Composite (top) and AWGN channel output (bottom) images. The composite and channel output images have peak signal-to-distortion ratios of 34.9 dB and 22.6 dB, respectively. DNR $= -12.1$ dB, yet all bits were extracted without error. $R_m = 1/512$ and $\text{DNR}_{\text{norm}} = 15.0$ dB so the actual bit-error probability is 10^{-6}.

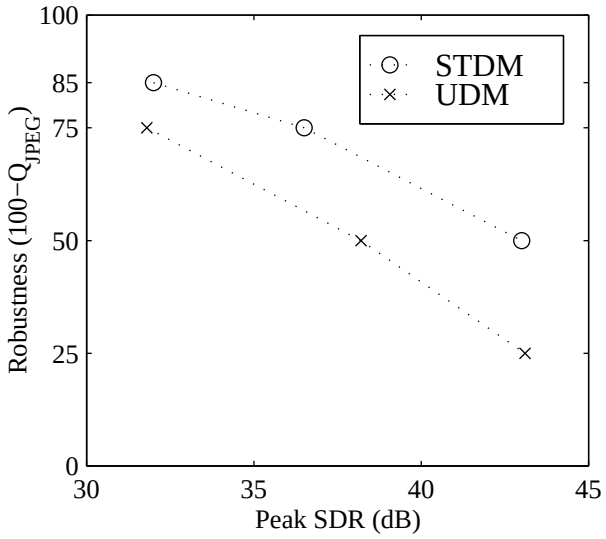

Figure 8-2: Achievable robustness-distortion trade-offs of uncoded dither modulation on the JPEG channel. $R_m = 1/320$. The bit-error rate is less than 5×10^{-6}.

These achievable distortion-robustness trade-offs at an embedding rate of $R_m = 1/320$ bits per grayscale pixel are shown in Fig. 8-2 at various JPEG quality factors ($\mathrm{Q_{JPEG}}$). The peak signal-to-distortion ratio (SDR) is defined as the ratio between the square of the maximum possible pixel value and the average embedding-induced distortion per pixel. Sample host and composite signals, both 512-by-512 images, are shown in Fig. 8-3. The actual embedding is performed in the DCT domain using 8-by-8 blocks ($f_1, f_2 \in \{0, 1/16, \ldots, 7/16\}$) and low frequencies $\left(\sqrt{f_1^2 + f_2^2} \leq 1/4\right)$, with 1 bit embedded across 5 DCT blocks. STDM is better than unspread dither modulation by about 5 dB at $(100 - \mathrm{Q_{JPEG}})$ of 50 and 75. As discussed in Sec. 5.2, one explanation for this performance advantage is the lower number of "nearest neighbors", the number of directions in which large perturbation vectors can cause decoding errors, of STDM relative to UDM.

Although no bit errors occurred during the simulations used to generate Fig. 8-2, we estimate the bit-error rate to be at most 5×10^{-6}. At an embedding rate of 1/320, one can only embed 819 bits in the host signal image, which is not enough to measure bit-error rates this low. However, one can estimate an upper bound on the bit-error rate by measuring the bit-error rate ϵ at an embedding rate five times higher ($R_m = 1/64$) and calculating the coded bit-error probability of a rate-1/5 repetition code when the uncoded error probability is ϵ assuming independent errors, which can approximately be obtained by embedding the

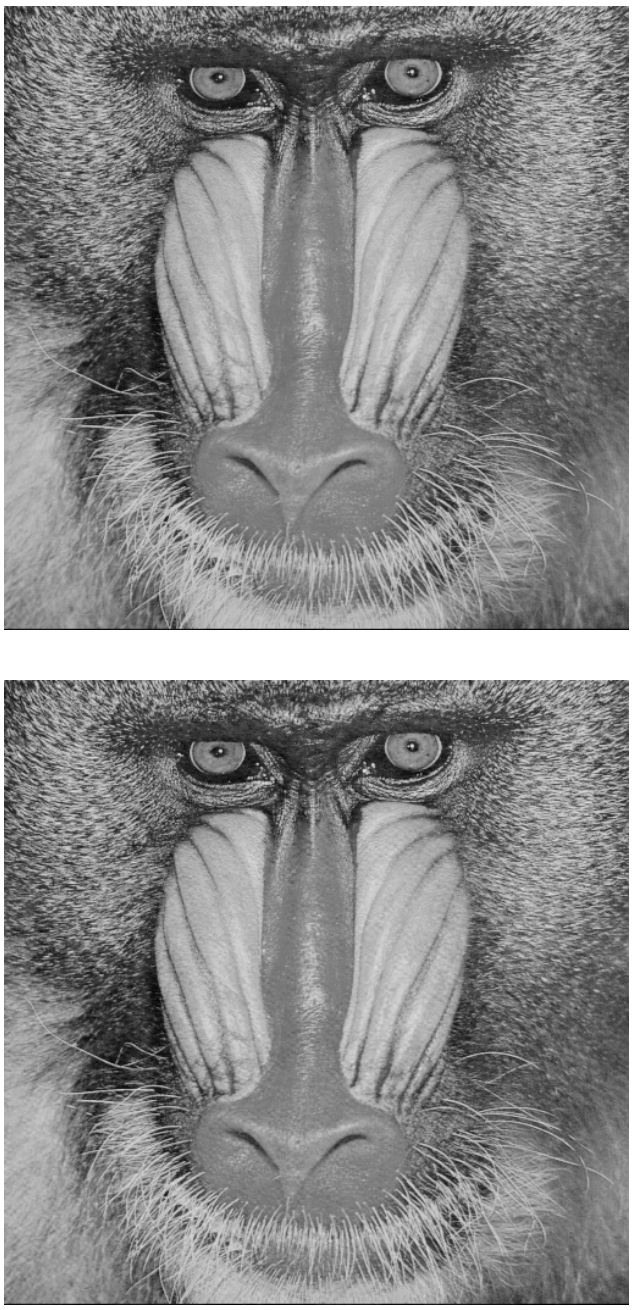

Figure 8-3: Host (top) and composite (bottom) images. After 25%-quality JPEG compression of the composite image, all bits were extracted without error. $R_m = 1/320$. Peak SDR of composite image is 36.5 dB.

Rate (R_{conv})	Generators (octal)	d_{free}
1/2	561, 753	12
1/4	463, 535, 733, 745	24

Table 8.1: Convolutional code parameters. Each code has a memory of 8 (constraint length of 9).

repeated bits in spatially separated places in the image. This coded bit-error probability is

$$P_{\text{rep}} = \sum_{k=3}^{5} \binom{5}{k} \epsilon^k (1-\epsilon)^{5-k} \tag{8.2}$$

If $\epsilon \leq 32/4096$, then (8.2) implies $P_{\text{rep}} \leq 4.7 \times 10^{-6}$. Thus, to obtain Fig. 8-2, we first embedded at a rate of 1/64 adjusting the SDR until $\epsilon \leq 32/4096$. Then, we embedded at a rate of 1/320 using a rate-1/5 repetition code to verify that no bit errors occurred.

8.2 Gains from Error Correction Coding and Distortion Compensation

As mentioned in Chap. 6, the gap between uncoded STDM (with uniform, scalar quantization and without distortion compensation) and the Gaussian capacity is about 13.6 dB in terms of DNR_{norm} at a bit-error rate of 10^{-6}. In this section we investigate how much of this gap can be closed with practical error correction codes and distortion compensation. From the definition of DNR_{norm} (8.1), we see that a gain factor g in the minimum DNR_{norm} required for a given bit-error rate translates into

1. a factor of g increase in rate for fixed levels of embedding-induced distortion and channel noise (robustness),
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 DNR_{norm} required for a given bit-error rate is the fundamental parameter of interest, and in the Gaussian case the DNR_{norm} also completely determines the bit-error probability (6.23) for uncoded STDM for $R_m \leq 1$.

In our experiment we embedded 10^7 bits in a pseudorandom white Gaussian host using memory-8, rate-1/2 and rate-1/4, convolutional codes with maximal free distance. Table 8.1

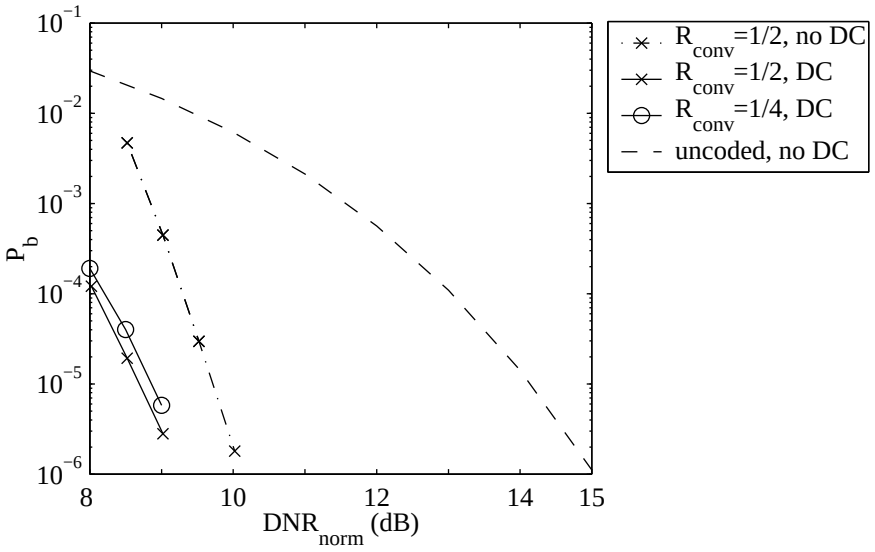

Figure 8-4: Error-correction coding and distortion-compensation gains. With common, memory-8 convolutional codes one can obtain gains of about 5 dB over uncoded STDM. Distortion compensation (DC) yields about 1 dB additional gain. Generalized LBM curves for non-distortion-compensated cases lie 2.43 dB to the right of the corresponding STDM curves.

contains the generators and free distances of these codes [27, Tbl. 11.1]. One coded bit was embedded in each spread-transformed host signal component using uniform, scalar quantizers as described in Chap. 5. 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 [26] of the convolutionally encoded data, as illustrated in Fig. 5-2. Experimentally measured bit-error rate (BER) curves are plotted in Fig. 8-4. 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 analysis in App. C, we know that the corresponding BER curves for generalized LBM with uniform, scalar quantization lie 2.43 dB to the right of their STDM counterparts in cases where no distortion compensation is performed. In distortion compensated cases, however, the distortion-compensation interference in the LBM case has a different probability distribution than in the STDM case because the quantizer reconstruction points are not in the centroids of their respective embedding intervals. Thus, it is unclear where the BER curves for distortion-compensated LBM lie.

Another set of experiments was performed to illustrate to advantages of distortion-compensated STDM over regular STDM against JPEG compression attacks. A rate-1/5

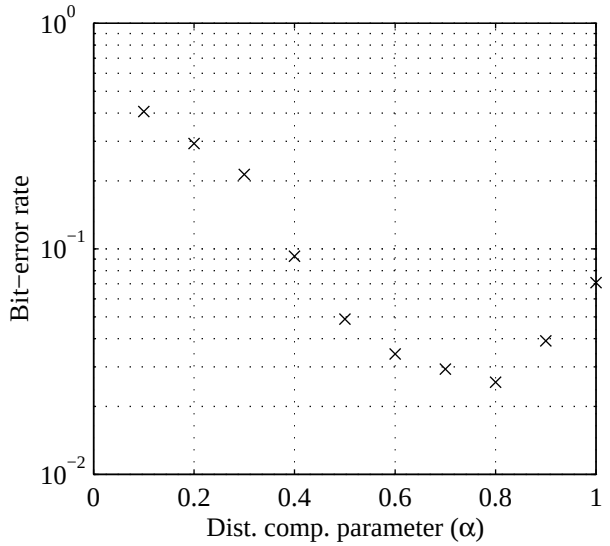

Figure 8-5: Bit-error rate for various distortion compensation parameters for JPEG compression channel of 25%-quality. $R_m = 1/320$. The peak SDR, between 43.3-43.4 dB, is chosen high enough to obtain a measurable bit-error rate.

repetition code was used to embed 1 bit in the low frequencies of five 8-by-8 DCT blocks for an overall embedding rate of 1/320. Using Fig. 8-2, we chose a low enough embedding-induced distortion level (SDR = 43 dB) such that we would be able to observe errors in the 819 decoded bits after 25-percent quality JPEG compression. Then, we measured the decoded bit-error rate with different distortion compensation parameters α in (3.10). The results are shown in Fig. 8-5.

We see that distortion compensation is helpful, provided that one chooses α to obtain an efficient trade-off between minimum distance and distortion-compensation interference, both of which are increased by decreasing α, as discussed in Sec. 3.3. The measured distortion-to-noise ratios in the projections of the received signals onto the STDM pseudo-random vectors were between 3.2 dB and 3.6 dB. For DNRs in this range, the α given by (3.11), which maximizes "SNR at the decision device" and is optimal for AWGN channels, is between 0.67 and 0.69. Although the measured bit-rate error in Fig. 8-5 is lower for $\alpha = 0.8$ than for $\alpha = 0.7$ (21/819 vs. 24/819), these measurements are within statistical uncertainty.

Chapter 9

Concluding Remarks

We conclude this thesis by briefly summarizing many of the main results and commenting on several promising directions for future research.

9.1 Concluding Summary

Digital watermarking, information embedding, and data hiding systems play a key role in addressing at least three major challenges that have arisen from the widespread distribution of multimedia content over digital communication networks. In particular, these systems are enabling technologies for (1) enforcing and protecting copyrights, (2) authenticating and detecting tampering of multimedia signals, and (3) backwards-compatibly upgrading existing legacy communication networks. We have introduced a framework in which to design and analyze these digital watermarking, information embedding, and data hiding systems, characterizing the goodness of such systems by their rate-distortion-robustness performance. A class of embedding methods that we call quantization index modulation (QIM) emerges quite naturally from this framework, and such QIM methods possess a number of attractive features from both a practical engineering perspective and an information-theoretic perspective.

From an engineering perspective, the structure of QIM systems reduces the problem of engineering information embedding systems to one of simultaneously designing good source codes and good channel codes, and the system designer can exploit such structure to conveniently trade-off rate, distortion, and robustness by tuning system parameters such as quantization step size. At the same time, the structure is sufficiently general that one can

achieve capacity, the information-theoretically best possible rate-distortion-robustness performance, against any type of fixed attack with QIM in the proper domain. For example, for (even possibly colored) Gaussian host signals distortion-compensated QIM (DC-QIM), which we have also introduced in this thesis, can achieve capacity against additive Gaussian noise attacks, which may be good models for unintentional attacks. Even for arbitrary (non-fixed) attacks, capacity-achieving DC-QIM systems exist for squared error distortion-constrained attacks, if the host signal is Gaussian and the embedder and decoder share a private key. For non-Gaussian host signals DC-QIM systems can achieve capacity asymptotically in the limit of small embedding-induced and attacker's distortions, which is the limit of interest in high-fidelity applications.

We have also presented practical, low complexity implementations of QIM called dither modulation with uniform scalar quantization. These methods achieve quantifiably better rate-distortion-robustness performance than previously proposed classes of methods such as amplitude-modulation spread spectrum and quantization-and-perturbation [43], which one may view as a generalized form of low-bit(s) modulation. One can quantify this performance gain in terms of SNR at the decoder decision device and in terms of achievable rate-distortion-robustness performance against worst-case squared error distortion-constrained intentional attacks, where the attacker knows everything about the encoding and decoding processes, including any keys. Furthermore, one can conveniently convert previously developed amplitude-modulation spread spectrum systems to spread-transform dither modulation systems by replacing addition with quantization.

Information-embedding capacities in the case of Gaussian host signals and additive Gaussian noise with arbitrary statistics have also been presented in this thesis. The capacities in these cases are the same for both the host-blind and known-host cases and are independent of host signal statistics, indicating that an infinite-energy host signal interferer causes absolutely no performance degradation. When applied to multimedia applications such as bandwidth-conserving hybrid transmission, these results imply a capacity of about 1/3 b/s for every Hertz of host signal bandwidth and dB drop in received host signal quality. When quantization occurs in the composite signal domain, QIM methods exist that achieve performance within 1.6 dB of these capacities, and one can eliminate even this small gap with distortion compensation.

We have also presented a number of empirical simulation results that complement our

mathematical analyses. These simulation results demonstrate practically achievable performance of dither modulation against additive Gaussian noise and JPEG compression attacks and demonstrate practically achievable gains from error correction coding and distortion compensation.

9.2 Future Work and Extensions

In this section we comment on several directions for extending the results of this thesis and discuss possible approaches to tackling some remaining open problems.

9.2.1 General attack models

In many applications, particularly in those involving intentional attacks, one may encounter attacks that cause a large amount of squared error distortion but that do not cause an unacceptably large degradation in *perceptible* signal quality. Such attacks may not be well captured by the models discussed in this paper. Although Sec. 3.1 contains some general comments about dealing with general deterministic and probabilistic channels, application of these deterministic and probabilistic models and their corresponding decoding structures to the myriad number of exotic attacks such as scaling, rotation, cropping, and column replacement remains as an interesting direction for future work.

9.2.2 Incorporation of other aspects of digital communication theory

As discussed in Chap. 3, one may view the quantizer reconstruction points in a QIM ensemble as signal constellation points for communication of the embedded information $\boldsymbol{m}$. Thus, one can incorporate a number of well-known digital communication techniques into QIM information embedding techniques for scenarios that have not explicitly been considered in this thesis. In many cases, one may also be able to further analyze QIM methods by applying results from classical (non-watermarking) digital communications theory. We give several examples below.

Amplitude-scaling invariant QIM

In some applications such as the digital audio broadcasting (DAB) application mentioned in Chap. 1, the channel may scale the amplitude of the channel input. In these cases one

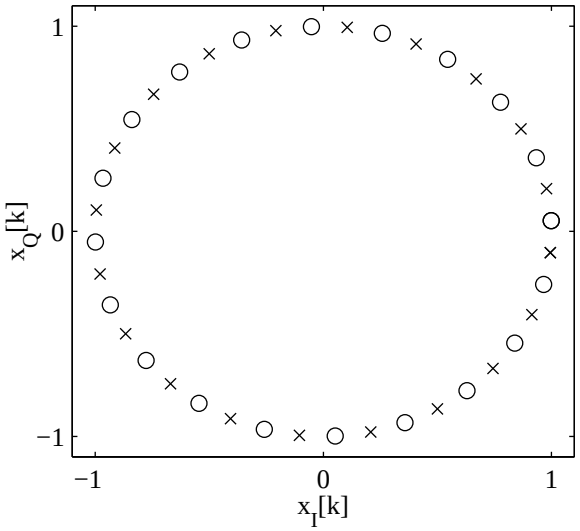

Figure 9-1: Signal constellation and quantizer reconstruction points for phase quantization and dither modulation with analog FM host signal. $x_I[k]$ and $x_Q[k]$ are the in-phase and quadrature signal components, respectively. The quantizer step size Δ is $\pi/10$. The $\times$-quantizer dither value is $\Delta/3$. The $\circ$-quantizer dither value is $-\Delta/6$.

may want to embed information only in the phase[1] of the host signal so that the decoder will not need to estimate changes in amplitude.

In the case of FM DAB, the analog FM host signal has constant amplitude, and thus, an example of a resulting signal constellation and/or ensemble of quantizer reconstruction points is shown in Fig. 9-1. As one can see, the resulting constellation is a phase-shift keying (PSK) [26] constellation. Thus, a natural starting point for future work in analyzing the performance of this type of "phase quantization index modulation" in detail might be the wide body of PSK analyses that has already been developed for classical communication.

Multirate QIM

Since digital watermarking involves communicating a digital message, digital watermarking methods suffer from the well-known threshold effect or "cliff" effect that is typical of digital communication in general. If the interference at the decoder is smaller than some threshold, then the decoder can successfully decode the message. However, if the interference is larger than the threshold, then the decoder fails. This inherent property of digital communication

[1] If the host signal is real valued, one can group host signal samples into pairs, and treat each pair as the real and imaginary part of a complex number to obtain phase samples.

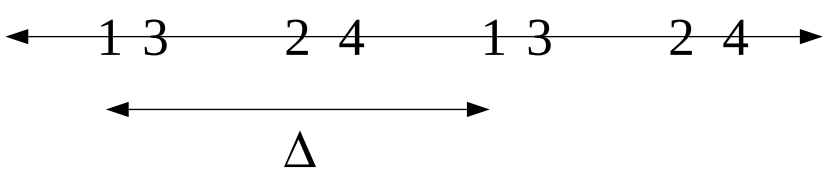

Figure 9-2: Broadcast or multirate digital watermarking with spread-transform dither modulation. In high-noise scenarios the decoder determines if m is even or odd to extract one bit. In low-noise scenarios the decoder determines the precise value of m to extract two bits and, hence, double the rate.

systems creates some challenges if either the amount of channel interference is not known when designing the system or if there are multiple decoders, each facing a different amount of interference.

One solution, of course, is to choose the information embedding rate low enough so that the worst case interference is below the failure threshold. However, if the interference turns out to be smaller than the worst case amount, then one might desire that the decoder have the capability to extract more than this minimum rate of embedded information. To accommodate such "graceful degradation" (or "graceful improvement", depending on one's perspective) in rate, one can replace individual quantizer reconstruction points with "clouds" of points, as described in [12, 35] for broadcast communication in non-watermarking contexts.

An example of such a "broadcast" or multirate STDM quantizer ensemble for digital watermarking is shown in Fig. 9-2. The reconstruction points of four quantizers are labeled 1, 2, 3, and 4, respectively. The minimum distance between an even and an odd point is larger than the minimum distance between any two points and is set large enough such that the decoder can determine if an even or an odd quantizer was used, and hence extract one bit, even under worst-case channel noise conditions. However, if the channel noise is smaller, then the decoder can determine the precise quantizer used, and hence, extract two bits. Again, we leave detailed analysis and performance optimization of such a multirate QIM method for future work.

Closing the gap to the Gaussian capacity

Simulation results (Fig. 8-4) in Chap. 8 demonstrated about a 6-dB gain from distortion compensation and reasonable-complexity error correction coding compared to uncoded STDM with uniform scalar quantization. As shown in Fig. 6-4 of Chap. 6, the total gap

between uncoded STDM and capacity is about 13.6 dB in the Gaussian case. Thus, about a 7.6-dB gap to capacity remains.

As discussed in Chap. 6, distortion-compensated QIM systems exist that achieve the Gaussian capacity. Thus, the remaining 7.6-dB gap must be due to the remaining structure of the systems that we implemented in our simulations, structure that inherently places limiting constraints on the systems. These constraints are: (1) using convolutional codes of memory 8 and (2) using dithered, uniform, scalar quantizers.

Therefore, possible directions for future work include using more powerful error correction codes such as low-density parity check codes [18, 29, 38] and using high dimensional vector quantizers [23], perhaps based on lattices [10]. Potentially fruitful directions for future work in this area include not only applying these coding and quantization methods to obtain enhanced performance but also drawing from related analyses to gain insights into understanding the achievable performance. For example, one can separate the achievable gains of lattice codes into a coding gain and a shaping gain [26]. Perhaps, one can develop similar analyses to determine separately achievable gains from channel coding and vector quantization. Similarly, determining the achievable performance of LBM with vector quantization also remains as an open area of future research.

9.2.3 System level treatments

Also, while the focus of this thesis has been on information embedding methods, "system level" issues such as how to best employ information embedding methods for, say, authentication remain largely unresolved. For example, given an information embedding method that is robust to all anticipated attacks and can embed imperceptibly at a given rate, what authentication signal should be embedded? More generally, a treatment of both watermarking-based and non-watermarking-based authentication methods, focusing on authentication performance rather than on information-embedding (rate-distortion-robustness) performance, would complement very well the analyses in this thesis. Similar statements apply for copyright protection mechanisms and for paradigms for backwards-compatible upgrading of communication networks.

9.2.4 Duality with Wyner-Ziv source coding

One final area for future exploration is the information-theoretic duality [3, 8] between information embedding and so-called Wyner-Ziv source coding [48], which is lossy source coding with side information that is known at the decoder but not at the encoder. For example, the information-embedding capacity expression (4.1) has a rate-distortion counterpart in the Wyner-Ziv problem [48, Eq. (15a)]. Similarly, the lossless version of Wyner-Ziv source coding, called Slepian-Wolf source coding [36], is the dual of noise-free information embedding, and the noise-free information-embedding capacity (4.7) is the dual result to the Slepian-Wolf source coding theorem [36, Eq. (5)]. Also, the conditions (4.12) and (4.13) under which having the host signal only at the encoder, and not the decoder, leads to no loss in achievable performance (relative to the known-host case) have Wyner-Ziv counterparts under which having side information only at the decoder, and not the encoder, leads to no loss in performance (relative to having the side information at both the encoder and decoder).

As a result of this duality, one can leverage work on the design and analysis of information embedding systems to better design and analyze Wyner-Ziv source coding systems, and vice versa. Thus, future research in this area may indeed prove to be quite fruitful. More detailed comments can be found in [3].

Appendix A

Notational Conventions

This appendix explains some notational conventions used in this thesis. Exceptions to these conventions are noted in the thesis, where appropriate. In general, scalar values are written in regular font, while vector values are written in **boldface** font. Random variables (scalar or vector) are written in sans serif font, while deterministic variables and sample values are written in regular Roman font. Matrices are denoted by capital letters, although not all variables denoted by capital letters are matrices. Here are some examples:

- x: Scalar, random variable
- x: Scalar, deterministic variable or sample value
- x: Vector, random variable
- $\mathbf{x}$: Vector, deterministic variable or sample value
- A: Deterministic matrix. Exceptions: $D(\mathbf{x}, \mathbf{s})$, N, and L.

The notation $p_{\mathsf{x}}(x)$ represents the probability density function (pdf) of random variable x evaluated at sample value x, although occasionally we use the shortened notation $p(\mathsf{x})$ when there is no risk of confusion between random variables and sample values.

Appendix B

Ideal Lossy Compression

The additive Gaussian noise channel model *may* be a good model for lossy compression, if the compression algorithm was designed ideally for Gaussian source signals, as we show in this appendix. Specifically, our goal is to develop a channel model for the case when the composite signal s is lossily compressed to produce the channel output. We emphasize in advance that our results here do not imply that every good compression algorithm can be modeled as adding Gaussian noise to the source. Instead, we simply derive an equivalent additive Gaussian noise model corresponding to the rate-distortion achieving distribution — the probability distribution from which one can choose random source codebooks, at least one of which achieves the best possible compression rate for a given distortion [13] — for Gaussian sources and squared error distortion.

We consider the case of a zero-mean, white, Gaussian source signal, which is the composite signal in this case, and a lossy compression algorithm that has maximum allowable squared error distortion of σ_z^2. The rate-distortion achieving distribution for this source coding problem is represented by the test channel [13, Chap. 13]

$$s = y' + z, \tag{B.1}$$

where $s \sim \mathcal{N}(0, \sigma_s^2)$ is a Gaussian source signal sample, $y' \sim \mathcal{N}(0, \sigma_s^2 - \sigma_z^2)$ is the corresponding sample from a source codeword, and $z \sim \mathcal{N}(0, \sigma_z^2)$ is the source coding error (quantization error) and is statistically independent of y'. The notation $x \sim \mathcal{N}(0, \sigma_x^2)$ means that x is a Gaussian random variable with mean 0 and variance σ_x^2. We assume, as is usu-

ally the case, that the distortion σ_z^2 is smaller than the source signal variance σ_s^2.[1] Thus, $\mathbf{s}$ and $\mathbf{y}'$ are jointly Gaussian [46] and the conditional probability density function (pdf) of $\mathbf{y}'$ given $\mathbf{s}$ is also a Gaussian density with[2]

$$\begin{aligned} E[\mathbf{y}'|\mathbf{s}] &= \frac{\sigma_s^2 - \sigma_z^2}{\sigma_s^2}\mathbf{s} \\ \text{var}[\mathbf{y}'|\mathbf{s}] &= \left(\sigma_s^2 - \sigma_z^2\right)\frac{\sigma_z^2}{\sigma_s^2}. \end{aligned} \tag{B.2}$$

Thus, an equivalent model for the test channel (B.1) is

$$\mathbf{y}' = a\mathbf{s} + \mathbf{z}', \qquad \mathbf{z}' \sim \mathcal{N}(0, \text{var}[\mathbf{y}'|\mathbf{s}]),$$

where

$$a = \frac{\sigma_s^2 - \sigma_z^2}{\sigma_s^2} \tag{B.3}$$

and $\mathbf{z}'$ is statistically independent of $a\mathbf{s}$. Hence, an equivalent model for this type of lossy compression is (1) multiplication by a scalar factor a followed by (2) addition of Gaussian noise $\mathbf{z}'$ that is independent of the scaled signal. Furthermore, from (B.3) we see that the scalar multiplication factor a is known if the amount of allowable compression-induced distortion σ_z^2 is known. (We assume that σ_s^2 is known since this composite signal variance depends only on the embedding function and not on the compression algorithm.) If we know a at the (watermarking) decoder, we can preprocess the source codeword $\mathbf{y}'$ by multiplying it by $1/a$ to get

$$\mathbf{y} = \frac{1}{a}\mathbf{y}' = \mathbf{s} + \mathbf{n},$$

where $\mathbf{n} = \mathbf{z}'/a$ is zero-mean, additive, Gaussian noise, independent of the composite signal $\mathbf{s}$. In this sense, we can model lossy compression (followed by front-end multiplication by a scalar at the decoder) as an additive Gaussian noise channel.

[1] If the allowable distortion is greater than σ_s^2, then one could simply use $\mathbf{y}' = \mathbf{0}$ as the source codeword. Clearly, no information embedding system can be robust to such compression.

[2] One simple way to derive the mean and variance of this pdf is to find, respectively, the linear least squares estimate of $\mathbf{y}'$ given $\mathbf{s}$ and the error variance of this estimate [46].

Finally, from (B.2) and (B.3), the noise variance is

$$\sigma_n^2 = \frac{\mathrm{var}[y'|s]}{a^2} = \frac{\sigma_z^2}{1 - \sigma_z^2/\sigma_s^2}.$$

This noise variance approaches the compression-induced distortion σ_z^2 in the limit of large composite signal to compression-distortion ratio σ_s^2/σ_z^2.

Appendix C

LBM Distortion-normalized Minimum Distance

In this appendix we calculate the distortion-normalized minimum distance of binary low-bit(s) modulation (LBM) with uniform, scalar quantization. We assume that the host signal and embedded signal are statistically independent.

The embedding function of any LBM method 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. The quantization cells of $\mathbf{q}(\cdot)$ are the coarse quantization cells illustrated in Fig. 3-3, and we define $\mathbf{q}(\cdot)$ such that its reconstruction points lie at the centroids of these quantization cells. Thus,

$$E[\mathbf{q}(\mathbf{x}) - \mathbf{x}] = \mathbf{0}. \tag{C.1}$$

Then, the expected distortion is

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

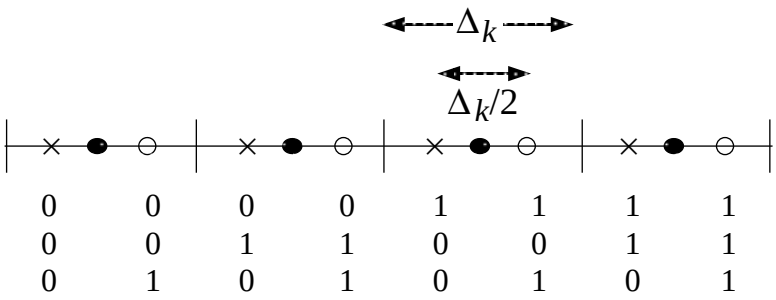

Figure C-1: Low-bit modulation with a uniform, scalar quantizer. The quantizer has a step size of $\Delta_k/2$, and the least significant bit (lsb) is modulated. All reconstruction points marked with a $\times$ have a lsb of 0. Points marked with a $\circ$ have a lsb of 1. This process is equivalent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

where we have used (C.1) and the independence of $\mathbf{x}$ and $\boldsymbol{m}$ 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 $\mathbf{d}(\boldsymbol{m})$.

We consider the uniform, scalar quantization case that is most directly comparable to the binary dither modulation case of Chap. 5, where we have a sequence of coded bits, each of which is repeated L times and embedded in a length-L block with a sequence of uniform, scalar quantizers. Instead of modulating the dither, however, we modulate the least significant bit of each quantizer. The k-th uniform, scalar quantizer is illustrated in Fig. C-1. The coarse quantizer $q_k(\cdot)$ has a step size of Δ_k, and the k-th least significant bit adjustment element d_k equals $\pm\Delta_k/4$. Consequently, the system has the same minimum distance (5.1) as the dither modulation systems of Chap. 5,

$$d^2_{\min} = \gamma_c \frac{1}{4LR_{\boldsymbol{m}}} \sum_k \Delta_k^2.$$

If we make the same assumption as in Chap. 5 that $\mathbf{x}$ can be modeled as uniformly distributed within each cell of $\mathbf{q}(\cdot)$, then the first term in (C.2) is

$$\frac{1}{L}\sum_k E\left[\|q(\mathbf{x}_k) - \mathbf{x}_k\|^2\right] = \frac{1}{12L}\sum_k \Delta_k^2,$$

the same as the expected distortion (5.3) of the dither modulation system. The second term is

$$\frac{1}{L}\sum_k d_k^2 = \frac{1}{16L}\sum_k \Delta_k^2.$$

Thus, the overall expected distortion is

$$D_{\mathbf{s}} = \left(\frac{1}{12L} + \frac{1}{16L}\right) \sum_k \Delta_k^2 = \frac{7}{48L} \sum_k \Delta_k^2,$$

and the distortion-normalized squared minimum distance is

$$d^2_{\text{norm}} = \frac{12\gamma_c}{7R_m}.$$

By comparing with (5.4), we see that LBM is worse than the corresponding dither modulation system by

$$\frac{3}{12/7} = \frac{7}{4} \approx 2.43\,\text{dB}. \tag{C.3}$$

Appendix D

Gaussian Capacity Proof: Colored Host, White Noise

In this appendix we formally complete the derivation of capacity (6.8) that is sketched in Sec. 6.1.2 for the case of a colored Gaussian host signal and white Gaussian noise. As described in that section, our goal is to find a probability density function (pdf) $p_{\tilde{u},\tilde{e}|\tilde{x}}(\tilde{u},\tilde{e}|\tilde{x})$ that maximizes the transform-domain version of (4.1),

$$C = \max_{p_{\tilde{u},\tilde{e}|\tilde{x}}(\tilde{u},\tilde{e}|\tilde{x})} I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{y}}) - I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{x}}), \tag{D.1}$$

subject to the constraint

$$E[\tilde{\boldsymbol{e}}^T\tilde{\boldsymbol{e}}] \leq LD_{\mathbf{s}}. \tag{D.2}$$

Our strategy is to hypothesize a pdf $p_{\tilde{u},\tilde{e}|\tilde{e}}(\tilde{u},\tilde{e}|\tilde{x})$ and show that with this choice of pdf $I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{y}}) - I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{x}})$ in (D.1) equals the expression in (6.7). Since this expression is also the capacity in the case when the host signal is known at the decoder (Sec. 6.1.4.), we cannot hope to achieve a higher rate, and hence, this pdf must indeed maximize (D.1).

We consider the pdf corresponding to the case where

$$\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{e}} + \alpha\tilde{\boldsymbol{x}}, \qquad \tilde{\boldsymbol{e}} \sim \mathcal{N}(0, D_{\mathbf{s}}I), \tag{D.3}$$

$\tilde{\boldsymbol{e}}$ and $\tilde{\boldsymbol{x}}$ are statistically independent, and α is given by (3.11). The notation $\boldsymbol{v} \sim \mathcal{N}(\mu, K)$

means that $\mathbf{v}$ is a Gaussian random vector with mean μ and covariance matrix K. Clearly, this choice of pdf satisfies the distortion constraint (D.2). Also, as explained in Sec. 6.1.2, $\tilde{\mathbf{x}} \sim \mathcal{N}(0, \Lambda_x)$ so $\tilde{\mathbf{u}} \sim \mathcal{N}(0, D_{\mathbf{s}} I + \alpha^2 \Lambda_x)$. The differential entropy $h(\mathbf{v})$ of an L-dimensional Gaussian random vector $\mathbf{v} \sim \mathcal{N}(\mu, K)$ is [13]

$$\frac{1}{2} \log_2 (2\pi e)^L |K|,$$

which for diagonal covariance matrices $K = \mathrm{diag}(k_1, \ldots, k_L)$ reduces to

$$\sum_{i=1}^{L} \frac{1}{2} \log_2 (2\pi e k_i). \tag{D.4}$$

Therefore,

$$\begin{aligned} I(\tilde{\mathbf{u}}; \tilde{\mathbf{x}}) &\stackrel{\Delta}{=} h(\tilde{\mathbf{u}}) - h(\tilde{\mathbf{u}}|\tilde{\mathbf{x}}) \\ &= h(\tilde{\mathbf{u}}) - h(\tilde{\mathbf{e}}) \\ &= \sum_{i=1}^{L} \frac{1}{2} \log_2 \left[2\pi e \left(D_{\mathbf{s}} + \alpha^2 \lambda_{x,i} \right) \right] - \sum_{i=1}^{L} \frac{1}{2} \log_2 (2\pi e D_{\mathbf{s}}) \\ &= \sum_{i=1}^{L} \frac{1}{2} \log_2 \frac{D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}}{D_{\mathbf{s}}}, \end{aligned} \tag{D.5}$$

where $\lambda_{x,i}$ denotes the i-th diagonal entry of Λ_x. The second line follows from (D.3) and the statistical independence of $\tilde{\mathbf{e}}$ and $\tilde{\mathbf{x}}$, and the third line follows since $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{e}}$ have diagonal covariance matrices and, hence, have entropies of the form (D.4). Thus, all that remains is to compute $I(\tilde{\mathbf{u}}; \tilde{\mathbf{y}})$ in (D.1).

The transform-domain channel output $\tilde{\mathbf{y}} = \tilde{\mathbf{e}} + \tilde{\mathbf{x}} + \tilde{\mathbf{n}}$ has a diagonal covariance matrix $K_{\tilde{\mathbf{y}}} = D_{\mathbf{s}} I + \Lambda_x + \sigma_n^2 I$ and via (D.3) can be written in the form

$$\tilde{\mathbf{y}} = \tilde{\mathbf{u}} + (1 - \alpha)\tilde{\mathbf{x}} + \tilde{\mathbf{n}}. \tag{D.6}$$

Thus, the differential entropy of $\tilde{\mathbf{y}}$ is given by (D.4),

$$h(\tilde{\mathbf{y}}) = \sum_{i=1}^{L} \frac{1}{2} \log_2 \left[2\pi e \left(D_{\mathbf{s}} + \lambda_{x,i} + \sigma_n^2 \right) \right]. \tag{D.7}$$

One can similarly determine $h(\tilde{\mathbf{y}}|\tilde{\mathbf{u}})$ after determining $K_{\tilde{\mathbf{y}}|\tilde{\mathbf{u}}}$. Since $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{u}}$ are jointly Gaus-

sian vectors, the conditional density of $\tilde{\boldsymbol{y}}$ given $\tilde{\boldsymbol{u}}$ is Gaussian with conditional covariance matrix [46, (Eq. 1.150)]

$$K_{\tilde{y}|\tilde{u}} = K_{\tilde{y}} - K_{\tilde{y}\tilde{u}} K_{\tilde{u}}^{-1} K_{\tilde{y}\tilde{u}}^{T}. \tag{D.8}$$

From (D.6), the statistical independence of $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{n}}$, and the statistical independence of $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{n}}$, one can infer that

$$K_{\tilde{y}} = K_{\tilde{u}} + (1-\alpha)^2 K_{\tilde{x}} + K_{\tilde{n}} + (1-\alpha)\left(K_{\tilde{x}\tilde{u}} + K_{\tilde{x}\tilde{u}}^{T}\right)$$

and

$$\begin{aligned} K_{\tilde{y}\tilde{u}} K_{\tilde{u}}^{-1} K_{\tilde{y}\tilde{u}}^{T} &= \left[K_{\tilde{u}} + (1-\alpha)K_{\tilde{x}\tilde{u}}\right] K_{\tilde{u}}^{-1} \left[K_{\tilde{u}} + (1-\alpha)K_{\tilde{x}\tilde{u}}^{T}\right] \\ &= K_{\tilde{u}} + (1-\alpha)\left(K_{\tilde{x}\tilde{u}} + K_{\tilde{x}\tilde{u}}^{T}\right) + (1-\alpha)^2 K_{\tilde{x}\tilde{u}} K_{\tilde{u}}^{-1} K_{\tilde{x}\tilde{u}}^{T}. \end{aligned}$$

Inserting these expressions into (D.8), we obtain

$$K_{\tilde{y}|\tilde{u}} = K_{\tilde{n}} + (1-\alpha)^2 \left[K_{\tilde{x}} - K_{\tilde{x}\tilde{u}} K_{\tilde{u}}^{-1} K_{\tilde{x}\tilde{u}}^{T}\right],$$

which is a diagonal matrix since $K_{\tilde{n}}$, $K_{\tilde{x}}$, $K_{\tilde{x}\tilde{u}}$, and $K_{\tilde{u}}$ are all diagonal. The i-th diagonal entry is

$$\begin{aligned} \left[K_{\tilde{y}|\tilde{u}}\right]_{ii} &= \sigma_n^2 + (1-\alpha)^2 \left[\lambda_{x,i} - \frac{\alpha^2 \lambda_{x,i}^2}{D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}}\right] \\ &= \frac{\sigma_n^2 \left(D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}\right) + (1-\alpha)^2 \lambda_{x,i} D_{\mathbf{s}}}{D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}}. \end{aligned}$$

Thus, the conditional entropy of this conditionally Gaussian random vector is (D.4)

$$h(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{u}}) = \sum_{i=1}^{L} \frac{1}{2} \log_2 \left[2\pi e \frac{\sigma_n^2 \left(D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}\right) + (1-\alpha)^2 \lambda_{x,i} D_{\mathbf{s}}}{D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}}\right], \tag{D.9}$$

and taking the difference between (D.7) and (D.9), one obtains

$$I(\tilde{\boldsymbol{u}}; \tilde{\boldsymbol{y}}) = \sum_{i=1}^{L} \frac{1}{2} \log_2 \left[\frac{\left(D_{\mathbf{s}} + \lambda_{x,i} + \sigma_n^2\right)\left(D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}\right)}{\sigma_n^2 \left(D_{\mathbf{s}} + \alpha^2 \lambda_{x,i}\right) + (1-\alpha)^2 \lambda_{x,i} D_{\mathbf{s}}}\right]. \tag{D.10}$$

Taking the difference between (D.10) and (D.5) yields

$$\begin{aligned} I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{y}}) - I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{x}}) &= \sum_{i=1}^{L} \frac{1}{2}\log_2\left[\frac{D_{\mathbf{s}}\left(D_{\mathbf{s}}+\lambda_{x,i}+\sigma_n^2\right)}{\sigma_n^2\left(D_{\mathbf{s}}+\alpha^2\lambda_{x,i}\right)+(1-\alpha)^2\lambda_{x,i}D_{\mathbf{s}}}\right] \\ &= \sum_{i=1}^{L} \frac{1}{2}\log_2\left[\mathrm{DNR}\frac{1+\mathrm{DNR}+\mathrm{SNR}_{x,i}}{\mathrm{DNR}+\alpha^2\mathrm{SNR}_{x,i}+(1-\alpha)^2\mathrm{DNR}\,\mathrm{SNR}_{x,i}}\right], \end{aligned}$$

where $\mathrm{SNR}_{x,i} = \lambda_{x,i}/\sigma_n^2$ is the host signal-to-noise ratio in the i-th channel. Finally, substituting (3.11) into this expression yields

$$\begin{aligned} & I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{y}}) - I(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{x}}) \\ &\quad = \sum_{i=1}^{L} \frac{1}{2}\log_2\left[(1+\mathrm{DNR})^2\mathrm{DNR}\frac{1+\mathrm{DNR}+\mathrm{SNR}_{x,i}}{\mathrm{DNR}(1+\mathrm{DNR})^2+\mathrm{DNR}^2\,\mathrm{SNR}_{x,i}+\mathrm{DNR}\,\mathrm{SNR}_{x,i}}\right] \\ &\quad = \sum_{i=1}^{L} \frac{1}{2}\log_2\left[(1+\mathrm{DNR})^2\frac{1+\mathrm{DNR}+\mathrm{SNR}_{x,i}}{(1+\mathrm{DNR})^2+\mathrm{SNR}_{x,i}\,(1+\mathrm{DNR})}\right] \\ &\quad = \sum_{i=1}^{L} \frac{1}{2}\log_2(1+\mathrm{DNR}), \end{aligned}$$

which equals the desired expression (6.7). □

Bibliography

[1] Gonzalo Arce, Charles G. Boncelet, Jr., Richard F. Graveman, and Lisa M. Marvel. Applications of information hiding. In *Proc. of Third Annual Federated Laboratory Symposium on Advanced Telecommunications & Information Distribution Research Program*, pages 423–427, College Park, MD, February 1999.

[2] Richard J. Barron. Private communication, 1999.

[3] Richard J. Barron, Brian Chen, and Gregory W. Wornell. The duality between information embedding and source coding with side information and its implications and applications. Submitted to *IEEE Transactions on Information Theory*, 2000.

[4] Walter Bender, Daniel Gruhl, Norishige Morimoto, and Anthony Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3-4):313–336, 1996.

[5] Brian Chen and Carl-Erik W. Sundberg. Broadcasting data in the FM band by means of adaptive contiguous band insertion and precancelling techniques. In *Proc. of IEEE International Conf. on Communications*, volume 2, pages 823–827, Vancouver, Canada, June 1999.

[6] Brian Chen and Gregory W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. Submitted to *IEEE Transactions on Information Theory*, 1999.

[7] Brian Chen and Gregory W. Wornell. Quantization index modulation methods for digital watermarking and information embedding. Accepted, pending revision, for publication in *Journ. of VLSI Signal Proc. Sys. for Signal, Image, and Video Tech., Special Issue on Multimedia Signal Proc.*, 2000.

[8] Jim Chou, S. Sandeep Pradhan, and Kannan Ramchandran. On the duality between distributed source coding and data hiding. In *Proceedings of the Thirty-third Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, October 1999.

[9] Aaron Cohen and Amos Lapidoth. Private communication, 1999.

[10] John H. Conway and Neil J. A. Sloane. *Sphere Packings, Lattices, and Groups.* Springer-Verlag, 1988.

[11] Max H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, IT-29(3):439–441, May 1983.

[12] Thomas M. Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18(1):2–14, January 1972.

[13] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[14] Ingemar J. Cox, Joe Killian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997.

[15] Ingemar. J. Cox, Joe Killian, Tom Leighton, and Talal Shamoon. A secure, robust watermark for multimedia. In *Information Hiding. First International Workshop Proceedings*, pages 185–206, June 1996.

[16] Ingemar J. Cox and Jean-Paul M. G. Linnartz. Some general methods for tampering with watermarks. *IEEE Journal on Selected Areas in Communications*, 16(4):587–593, May 1998.

[17] Ingemar J. Cox, Matthew L. Miller, and Andrew L. McKellips. Watermarking as communications with side information. *Proceedings of the IEEE*, 87(7):1127–1141, July 1999.

[18] Robert G. Gallager. *Low-Density Parity-Check Codes*. Monograph. M.I.T. Press, Cambridge, MA, 1963.

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

[20] Frank Hartung and Martin Kutter. Multimedia watermarking techniques. *Proceedings of the IEEE*, 87(7):1079–1107, July 1999.

[21] Chris Heegard and Abbas A. El Gamal. On the capacity of computer memory with defects. *IEEE Transactions on Information Theory*, IT-29(5):731–739, September 1983.

[22] Juan R. Hernandez, Fernando Perez-Gonzalez, Jose Manuel Rodriguez, and Gustavo Nieto. Performance analysis of a 2-D-multipulse amplitude modulation scheme for data hiding and watermarking of still images. *IEEE Journ. on Selected Areas in Communications*, 16(4):510–524, May 1998.

[23] N. S. Jayant and Peter Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice-Hall, 1984.

[24] Deepa Kundur and Dimitrios Hatzinakos. Digital watermarking for telltale tamper proofing and authentication. *Proceedings of the IEEE*, 87(7):1167–1180, July 1999.

[25] Amos Lapidoth. Nearest neighbor decoding for additive non-Gaussian noise channels. *IEEE Transactions on Information Theory*, 42(5):1520–1529, September 1996.

[26] Edward A. Lee and David G. Messerschmitt. *Digital Communication*. Kluwer Academic Publishers, second edition, 1994.

[27] Shu Lin and Daniel J. Costello, Jr. *Error Control Coding: Fundamentals and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[28] Jean-Paul Linnartz, Ton Kalker, and Jaap Haitsma. Detecting electronic watermarks in digital video. In *Proc. of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2071–2074, Phoenix, AZ, March 1999.

[29] David J. C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Transactions on Information Theory*, 45(2):399–431, March 1999.

[30] Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. Preprint. October 1999.

[31] Joseph A. O'Sullivan, Pierre Moulin, and J. Mark Ettinger. Information theoretic analysis of steganography. In *Proc. of the 1998 IEEE International Symposium on Information Theory*, page 297, Cambridge, MA, August 1998.

[32] Haralabos C. Papadopoulos and Carl-Erik W. Sundberg. Simultaneous broadcasting of analog FM and digital audio signals by means of adaptive precanceling techniques. *IEEE Transactions on Communications*, 46(9):1233–1242, September 1998.

[33] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding — a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999.

[34] Christine I. Podilchuk and Wenjun Zeng. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 16(4):525–539, May 1998.

[35] Kannan Ramchandran, Antonio Ortega, K. Metin Uz, and Martin Vetterli. Multiresolution broadcast for digital HDTV using joint source/channel coding. *IEEE Journal on Selected Areas in Communications*, 11(1):6–23, January 1993.

[36] David Slepian and Jack K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, IT-19(4):471–480, July 1973.

[37] Joshua R. Smith and Barrett O. Comiskey. Modulation and information hiding in images. In *Information Hiding. First International Workshop Proceedings*, pages 207–226, June 1996.

[38] Daniel A. Spielman. Finding good LDPC codes. In *Proceedings of the 36th Annual Allerton Conference on Communication, Control, and Computing*, September 1998.

[39] Jonathan K. Su. Power-spectrum condition-compliant watermarking. DFG V^3D^2 Watermarking Workshop, October 1999. Abstract and transparencies from this talk were obtained from `http://www.lnt.de/~watermarking`.

[40] Feng-Wen Sun and H. C. A. van Tilborg. Approaching capacity by equiprobable signaling on the Gaussian channel. *IEEE Transactions on Information Theory*, 39(5):1714–1716, September 1993.

[41] Mitchell D. Swanson, Mei Kobayashi, and Ahmed H. Tewfik. Multimedia data-embedding and watermarking technologies. *Proceedings of the IEEE*, 86(6):1064–1087, June 1998.

[42] Mitchell D. Swanson, Bin Zhu, and Ahmed H. Tewfik. Robust data hiding for images. In *Proceedings of the 1996 IEEE Digital Signal Processing Workshop*, pages 37–40, Loen, Norway, September 1996.

[43] Mitchell D. Swanson, Bin Zhu, and Ahmed H. Tewfik. Data hiding for video-in-video. In *Proceedings of the 1997 IEEE International Conference on Image Processing*, volume 2, pages 676–679, Piscataway, NJ, 1997.

[44] Andrew Z. Tirkel, G. A. Rankin, Ron van Schyndel, W. J. Ho, N. R. A. Mee, and Charles F. Osborne. Electronic water mark. In *Proceedings of Digital Image Computing, Technology and Applications*, pages 666–672, Sydney, Australia, December 1993.

[45] Ron van Schyndel, Andrew Z. Tirkel, and Charles F. Osborne. A digital watermark. In *Proceedings of the First IEEE International Conference on Image Processing*, volume 2, pages 86–90, Austin, TX, November 1994.

[46] Alan S. Willsky, Gregory W. Wornell, and Jeffrey H. Shapiro. Stochastic processes, detection and estimation. MIT 6.432 Supplementary Course Notes, Cambridge, MA, 1996.

[47] Jack Wolosewicz and Kanaan Jemili. Apparatus and method for encoding and decoding information in analog signals. United States Patent #5,828,325, October 1998.

[48] Aaron D. Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, IT-22(1):1–10, January 1976.

[49] Ram Zamir and Meir Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, July 1996.