Systematic Hybrid Analog/Digital Signal Coding

by

Richard J. Barron

B.S., University of Illinois at Urbana-Champaign (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

Author _____
Department of Electrical Engineering and Computer Science
June, 2000

Certified by _____
Alan V. Oppenheim
Ford Professor of Engineering
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Systematic Hybrid Analog/Digital Signal Coding

by

Richard J. Barron

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

Abstract

This thesis develops low-latency, low-complexity signal processing solutions for systematic source coding, or source coding with side information at the decoder. We consider an analog source signal transmitted through a hybrid channel that is the composition of two channels: a noisy analog channel through which the source is sent unprocessed and a secondary rate-constrained digital channel; the source is processed prior to transmission through the digital channel. The challenge is to design a digital encoder and decoder that provide a minimum-distortion reconstruction of the source at the decoder, which has observations of analog and digital channel outputs.

The methods described in this thesis have importance to a wide array of applications. For example, in the case of in-band on-channel (IBOC) digital audio broadcast (DAB), an existing noisy analog communications infrastructure may be augmented by a low-bandwidth digital side channel for improved fidelity, while compatibility with existing analog receivers is preserved. Another application is a source coding scheme which devotes a fraction of available bandwidth to the analog source and the rest of the bandwidth to a digital representation. This scheme is applicable in a wireless communications environment (or any environment with unknown SNR), where analog transmission has the advantage of a gentle roll-off of fidelity with SNR.

A very general paradigm for low-latency, low-complexity source coding is composed of three basic cascaded elements: 1) a space rotation, or transformation, 2) quantization, and 3) lossless bitstream coding. The paradigm has been applied with great success to conventional source coding, and it applies equally well to systematic source coding. Focusing on the case involving a Gaussian source, Gaussian channel and mean-squared distortion, we determine optimal or near-optimal components for each of the three elements, each of which has analogous components in conventional source coding. The space rotation can take many forms such as linear block transforms, lapped transforms, or subband decomposition, all for which we derive conditions of optimality. For a very general case we develop algorithms for the design of locally optimal quantizers. For the Gaussian case, we describe a low-complexity scalar quantizer, the nested lattice scalar quantizer, that has performance very near that of the optimal systematic scalar quantizer. Analogous to entropy coding for conventional source coding, Slepian-Wolf coding is shown to be an effective lossless bitstream coding stage for systematic source coding.

Thesis Supervisor: Alan V. Oppenheim

Title: Ford Professor of Engineering

Acknowledgments

I would like first to acknowledge Al Oppenheim, my advisor and mentor, for taking me on board while I was green, and providing the guidance and encouragement for me to follow the learning curve. You gave me faith all along that we would find that "nugget," and I think we found a good one.

To my parents, *a ghrá*. I honor you for living the American dream and for providing for me to do the same. It is only through your tireless and selfless attention to my development that I have achieved this long-sought goal.

To all of my friends in the DSPG, I could always rely on you to ease the anxiety and frustration that goes along with a journey such as ours. I would especially like to thank my close friends Stark Draper, Nick Laneman, Matt Secor, and bigal Seefeldt for all of our extracurriculars, e.g., runs to the Muddy, and for listening and contributing to the ideas in this thesis.

I will remember my time here with great fondness. From the first days that I arrived, I knew that the friends and memories made here would last a lifetime.

Contents

List of Figures

Chapter 1

Introduction and Motivation

1.1 Systematic signal coding

Many important signal estimation problems involve the use of distributed networks of sensors connected via digital communications links. Based on all sensor measurements, a global estimate of a desired source signal is determined. Due to low power considerations or other limitations, the rate of transmission from each of the sensors will be constrained to some finite value. Given rate constraints for each sensor, we face the challenge of reconstructing a source, at a minimum distortion, from rate-reduced representations of signals correlated to the source, *i.e.,* the digitally encoded sensor measurements. This distributed source coding problem, well-known as the CEO problem [10], is still quite open. Even for the two sensor problem, the research community has yet to determine a rate-distortion region, the region of rates (R_1, R_2) within which a system can achieve a desired distortion, and outside of which it is impossible to achieve the desired distortion [92]. In addition to the investigations into the asymptotic limits of performance, there has been growing interest in practical digital signal processing solutions for distributed source coding.

In this thesis we explore *systematic signal coding*, an important special case of the CEO problem using two sensors, a near-field sensor, and a far-field sensor. The near-field sensor has a strong measurement of the source, so strong that we assume the measurement is noise-free. The far-field sensor, on the other hand, measures the output of a distortion-inducing channel through which the source is transmitted. A receiver observes the output of the far-field sensor and the rate-reduced digital encoding of the near-field sensor. For the receiver to observe the far-field sensor measurement

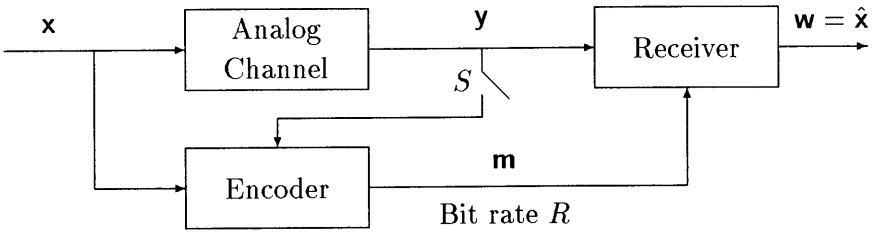


Figure 1-1: Source coding with side information. The signals $\mathbf{x}$, m, $\mathbf{y}$, and $\mathbf{w} = \hat{\mathbf{x}}$ are respectively the source, the encoded signal, the channel output, and the decoded source. The switch S may be opened or closed, depending on the application.

exactly, we must have that either the receiver is resident at the far-field sensor, or that the far-field sensor communicates to the receiver at such a high rate that it is transmitted distortion-free. For most applications the former scenario is most accurate, and, hence, for simplicity we will assume that this is the case. We conveniently view the receiver observations, therefore, as the outputs of parallel digital and analog channels. Combined, the digital and analog channels form what we call a *hybrid channel.* The block diagram for the systematic source coding scenario is given in Fig. 1-1. The N-dimensional source vector $\mathbf{x}$, is sent uncoded through a noisy analog channel, he output of which is $\mathbf{y}$. A feedback path is activated when the switch S is closed, thereby supplying the encoder with an observation of the side information. This thesis focuses on the case where the switch S is open, because it is more relevant to practical source coding applications. The source is coded by the encoder, creating an information signal $\mathbf{m}$, which is received undistorted by the decoder. From the signals $\mathbf{x}$ and $\mathbf{m}$, the decoder outputs a signal $\mathbf{w} = \hat{\mathbf{x}}$, which is intended to be a close approximation of the source $\mathbf{x}$. Throughout this thesis, we will use the symbols $\mathbf{x}$, $\mathbf{m}$, $\mathbf{y}$, and $\mathbf{w} = \hat{\mathbf{x}}$ to denote respectively the source, the encoded signal, the channel output, and the decoded source. As shown in Fig 1-1 the signals are vectors, but in some cases they may be more appropriately written as processes, $x[n]$, $y[n]$ and $w[n]$. For vector $\mathbf{x}$ and $\mathbf{y}$, the lengths will be assumed to be N and M, respectively, unless otherwise noted.

Throughout this thesis we will use the names "digital" and "analog" channel output, to mean, respectively, the output of the encoding of the near-field sensor measurement, and the measurement at the far-field sensor. In some cases, the naming convention may not be consistent with the

traditional definitions of "digital" and "analog", but, as it will be consistent for the majority of important cases, we use the nomenclature for the entire thesis.

The objective at the receiver is to use the measurements from both channels to reconstruct the source with minimum distortion. There are a variety of distortion criteria that we explore in this thesis, but we focus on mean-squared error. In some cases, the encoding of the near-field sensor may be arbitrarily fixed, and we do not have the liberty to design the encoder. For such scenarios, we view the receiver as simply an estimator that has observations of two channel outputs, the analog channel output, which usually has a stochastic relationship to the source, and the digital channel output, which is deterministically related to the source. In this context it is convenient to think of the receiver as performing analog estimation with digital, or deterministic, side information (from the digital channel). In other applications we have the liberty to choose the algorithm for encoding the near-sensor measurement, and the design of the encoder and the receiver are coupled. The receiver again has observations from the analog and digital channels, but in this case we view the receiver as a digital decoder with analog side information at the decoder. Thus, depending on the context, we will use the terms "digital side information" referring to the output of the digital channel, and "analog side information", referring to the output of the analog channel. Furthermore, we will use the terms "receiver," "estimator," and "decoder" interchangeably.

The term "systematic source coding" was coined by Shamai *et al* in [95] to describe source coding with analog side information at the receiver as an extension of a concept from error-correcting channel codes. A *systematic* error-correcting code is one whose codewords are the concatenation of the uncoded information source string and a string of parity-check bits. Similarly, in the systematic hybrid source coding scenario, we have an uncoded analog transmission (analogous to the uncoded source bits) and a source-coded digital transmission (analogous to the parity bits).

1.2 Applications

There are myriad applications for which systematic source coding design solutions are useful. The first is a near-field/far-field sensor scenario, exactly as described in the problem formulation at the outset of this chapter. A second powerful application is for the backward-compatible upgrade of an existing analog communications infrastructure. An existing full-band, but noisy analog commu-

nications infrastructure may be augmented by a low bandwidth digital side channel. Both analog and digital signals are used at properly enabled decoders to obtain improved fidelity, while legacy analog receiver are unaffected by the upgrade. Signal coding solutions for this source coding problem are clearly applicable to in-band on-channel (IBOC) digital audio broadcast (DAB) [14, 60], in which the digital audio information is multiplexed in the same bandwidth as the legacy AM or FM analog broadcast.

A physical realization of the hybrid channel may take several forms. In one scenario, the analog source is transmitted through a distortion-inducing channel such as a broadcast channel, and the digitally processed source is communicated through another medium such as fiber optic cable or a data network such as the internet. In another setting, a fixed channel bandwidth may be divided into an analog channel and digital channel. On the analog channel the source is amplitude or frequency modulated and transmitted. On the digital channel the digitally processed source is transmitted "error-free" through the use of forward error-correcting codes.

The latter example suggests that coding for a hybrid channel is a general framework for coding signals by using both digital and analog information on a given channel. This coding method is potentially applicable in a wireless communications environment (or any environment with unknown SNR), where reception of digital signals using forward error correction is typically in one of two states: either error-free or suffering unrecoverable errors due to significant fading between the transmitter and the receiver. There is a signal-to-noise ratio (SNR) threshold at which a source can no longer be communicated. In contrast, analog communication, *e.g.,* amplitude modulation, has a gentle "roll-off" of source fidelity with channel SNR. To achieve the high fidelity of digital source coding and slow roll-off of analog transmission, one may apply systematic hybrid analog/digital source coding. While SNR is at acceptable levels, a joint analog/digital decoding will be used. At a certain threshold SNR, when the digital signal is lost, the receiver will use only the analog signal. Consider a mobile phone unit on a battlefield. If the mobile unit is experiencing significant fading between it and a transmitter, a digital signal might be totally lost. An analog signal, although very noisy, may be comprehensible enough for a human listener to extract important information.

Another potential application is speech enhancement. Consider having available at the receiver an uncorrupted full-band recording of a particular speaker in which most of the common phonemes of the language are spoken. If the receiver observes a noise-corrupted utterance from the same

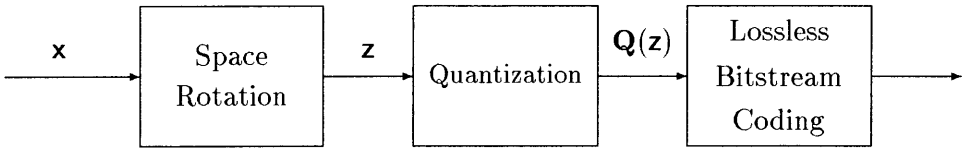


Figure 1-2: Source coding paradigm, for both standard source coding and systematic source coding. Space rotation, or transformation, usually takes the form of linear prediction, linear block transformation, or subband decomposition. Quantization is usually some simple form of digital encoding, often involving scalar operators. Lossless bitstream coding is entropy coding for standard source coding and is Slepian-Wolf Coding for systematic source coding.

speaker, and for a given segment of speech the receiver has knowledge of the uttered phoneme (using, for example, a hidden Markov model) the linear predictive coefficients obtained from the clean speech can potentially be used as side information to assist the enhancement. Although this side information may not be entirely accurate, a properly chosen estimator may yield perceptually pleasing results. This is an example where the form of the encoder is fixed, based on average statistics of the clean waveform, and it cannot be designed for minimum distortion. This type of application is not the focus of this thesis.

1.3 Thesis objective

Practical source coding systems are designed with constraints on two main quantities, latency and complexity. Latency must be kept especially low for real-time systems, such as two-way voice communications systems, and complexity impacts all practical source coding algorithms, as the algorithms are implemented on processors with limited computational capacities. The main purpose of this thesis is to provide a framework for systematic source coding using low-latency and low-complexity digital signal processing components. A very general paradigm for low-latency, low-complexity source coding, is shown in Fig. 1-2, which applies to both conventional source coding and systematic source coding. In this thesis we use the terms "conventional" and "standard" to refer to source coding methods used when there is no analog side information at the decoder, *i.e.*, when there exists only a digital channel. The paradigm diagrammed in Fig. 1-2, often referred

to as waveform coding [1], is composed of three basic components: 1) a space rotation, or transformation, 2) quantization, and 3) lossless bitstream coding. Clearly, Fig. 1-2 only describes the encoding stage; the decoder inverts the process with three steps: 1) bitstream decoding, 2) coefficient reconstruction ("inverse" quantization), and 3) inverse transformation. The space rotation and quantization stage can arguably be combined as a single vector quantizer. They are most often separated, however, as the quantization stage is usually composed of low-complexity quantizers, like scalar quantizers. The space rotation is selected to minimize the reconstruction distortion under the constraints on the quantizers. The space rotation can take many forms such as linear prediction, linear block transforms, or a subband decomposition [40]. Assuming a mean-squared error criterion, in standard coding, the optimal space rotation decorrelates the source samples, that is, it removes source redundancy. For systematic coding, we will establish that the optimal space rotation will decorrelate the error that is the result of minimum mean-squared error (MMSE) estimation of the source from the analog channel observation. The quantizers in most conventional source coding systems are often simple uniform quantizers. The analogous design element for systematic source coding is the nested lattice scalar quantizer (NLSQ) which is the composition of a uniform quantizer with a modulo element. The final block in Fig. 1-2 is the lossless bitstream coding element. For conventional coding, this stage is an entropy coder. It is well-known that following uniform quantization with entropy coding results in a very effective method of coding [32, 96]. Analogously, we show that an effective method for systematic coding uses an NLSQ followed by a Slepian-Wolf code; Slepian-Wolf codes are described in Chap. 3. Hence, the lossless bitstream stage for systematic coding is a Slepian-Wolf coding stage.

Although this thesis focuses primarily on digital signal processing solutions that lend themselves to practical implementations, we do have some theoretical results pertaining to the asymptotic limits of performance of systematic source coding systems, highlighted in Chap. 3. Wyner and Ziv give the first thorough theoretical treatment of source coding with side information at the decoder in [90], in which they derive the information theoretic rate-distortion limit. In Chap. 3 we review the results in [90] and related theoretical results, including some new discoveries. A final theoretical exposition in the thesis occurs in Chap. 6, where we establish the duality of source coding with side information at the decoder with information embedding. Some of the results of Chap. 6 are

[1] Another paradigm for source coding is so-called *parametric source coding*[40], which is not a focus of this thesis.

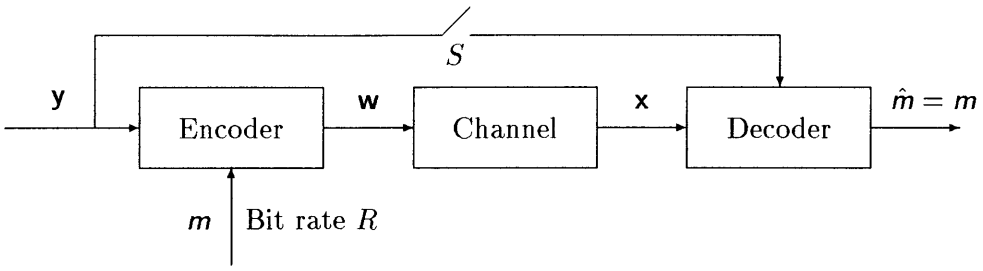


Figure 1-3: Information embedding. The signals $\mathbf{y}$, m, $\mathbf{w}$, and $\mathbf{x}$ are respectively the host, information signal, composite signal, and channel output. The switch S may be opened or closed, depending on the application

the product of a collaboration with B. Chen and G. W. Wornell, and appear in [2].

1.4 Duality with information embedding

Information embedding is the robust communication of information embedded into a host signal and has importance for applications such as digital watermarking for copyrighting material and information hiding, or steganography, for covert communications [72, 15]. We show the basic information embedding scenario in Fig. 1-3, in which we have the signals $\mathbf{y}$, m, $\mathbf{w}$, and $\mathbf{x}$, respectively the host, information signal, composite signal, and channel output. The encoder uses both the host and the information signal to create a *composite signal* $\mathbf{w}$, which is assumed to closely approximate the host $\mathbf{y}$. We assume the composite signal is passed through a channel, which is used to model an attempt by an adversary to remove the watermark. The output of the channel, $\mathbf{x}$, is decoded to yield an estimate $\hat{m} \approx m$ of the information signal. In some applications the host is known at the decoder, *i.e.,* the switch S is closed, but we focus on the case where the host is known only at the encoder.

The duality between systematic source coding and information embedding, is simply put. A good encoder (respectively, decoder) for systematic source coding is a good decoder (respectively, encoder) for information embedding. Comparing Figs. 1-1 and 1-3 we see that, as labeled, the encoder (respectively, decoder) for one case has the same input and output variables as the decoder (respectively, encoder) for the other case. One of the easiest ways to see the duality is through

example. One of the simplest forms of information embedding is so-called *low-bits modulation* [72]. At the encoder, the bits of the base-2 expansion of the information signal, $\boldsymbol{m}$, are extracted and used to replace the least significant bits of the digitized host. At the decoder, the channel output is digitized, *i.e.* quantized, and the least-significant bits of the quantizer output, are used to reconstruct the information signal. A dual system to low-bits modulation is *low-bits coding*, for systematic coding. At the encoder the source $\mathbf{x}$ is digitized, and the least-significant bits are extracted and sent down the digital channel. Let us assume that the channel is an additive white Gaussian noise channel (AWGN). As the channel noise likely corrupts only the LSBs, the decoder replaces the LSBs of the digital representation with those from the digital channel (signal $\boldsymbol{m}$) to form the final estimate of the source. Clearly the operation of the low-bits modulation encoder (respectively, decoder) is the same as the low-bits coding decoder (respectively, encoder). The precise implications and applications of the duality are developed in Chap. 6.

1.5 Notation

In terms of notation, we use a sans serif font $\boldsymbol{u}$ to denote a random variable, and a serif font u to denote a realization or a constant. Likewise, we use a bold sans serif font $\mathbf{u}$ to denote a random vector, a column vector, and a bold serif font $\mathbf{u}$ to denote a vector realization. The components of a vector $\mathbf{u}$ are denoted u_i, $i = 0, ..., N\text{-}1$. A matrix $\mathbf{U}$ is a denoted as a bold capital letter. The notation $\mathbf{U}^T$ implies the matrix transpose operation. The notation $\mathbf{U}^\dagger$ implies the Hermitian transpose operation. The functions $H(\cdot)$, and $I(\cdot;\cdot)$ are entropy and Shannon mutual information respectively. The function $h(\cdot)$ is either the differential entropy function or binary entropy function, depending on the argument.

1.6 Outline of thesis

The outline of the thesis is as follows. In Chap. 2 we derive the optimal receiver structure for the maximum *aposteriori* probability, minimum mean-squared error, and maximum likelihood estimation criteria, given observations from the analog and digital channels. We follow with some examples for particular sources and channels. Chap. 2 is less concerned with source coding paradigms than it is with exercising some of the basic principles of estimation with deterministic side information

at the receiver. In Chap. 3 we review the most important information theoretical results concerning rate-distortion theory with side information at the decoder, and develop some new theoretical results. Extending the basic Gaussian result from [89] we derive the rate-distortion function for jointly Gaussian source $x[n]$ and channel output $y[n]$. We also generalize the results of [95], and give rate-distortion achieving codes for the Gaussian case at all channel SNRs. In Chap. 4 we begin the development of the signal processing solution for systematic source coding, with results on systematic quantization. We develop design algorithms for the design of locally optimal scalar and vector quantizers. For the Gaussian case, we compare the performance of optimal quantizers with very low complexity quantizers, called nested lattice scalar quantizers. We establish the efficiency of a coding system comprised of an NLSQ and Slepian-Wolf coding. We also develop quantizers with partial feedback of the channel output. In Chap. 5 we develop the design of systematic transform and subband coders for the Gaussian case. Based on the results of Chap. 4 we derive optimal bit allocation strategies, and determine the optimal transforms and subband decompositions in terms of the statistics of the MMSE error process. In Chap. 6 we establish the theoretical duality of source coding with side information and information embedding, and use previous results for the Wyner-Ziv problem to create optimal solutions for information embedding. In Chap. 7 we assess the impact of our results and suggest directions for future work.

Chapter 2

Basic Principles and Design Elements

2.1 Introduction

In this chapter we introduce the basic signal processing design elements for the systematic encoder and decoder. The digital side information is a deterministic function of the source, which, depending on the application, may be designed to achieve a certain fidelity, or could simply be an arbitrary deterministic function. In any case, as long as the encoding function is known at the decoder, the observations of both the analog information and digital output of the encoder can be used together to obtain an optimal signal estimate based on some optimality criterion. In this chapter, we derive the receiver equations that yield the minimum mean-squared error (MMSE), maximum aposteriori probability (MAP), and maximum likelihood (ML) signal estimates.

We consider several important examples of optimal decoder operation. For a Gaussian source and channel and *linear side information*, we derive the MMSE, MAP, and ML decoders. For the case of minimum mean-squared error, we derive the optimum linear side information basis vectors at the encoder, based on the operation of the MMSE decoder. A final example involves a speech source signal with an AWGN analog channel and spectral envelope side information. We develop an algorithm which produces the ML signal estimate at the receiver.

As a prelude to Chapter 4 we describe the basic concept of digital encoding at a fixed rate given analog side information at the decoder. Inherent in digital coding of waveforms, certain relevant coefficients are quantized, or represented by a fixed number of codebook entries, and there exists a rule for signal reconstruction from the codebook values. The concepts of quantization encoding

and reconstruction are appropriately tailored to the case where we have analog side information at the decoder. The description we provide suggests the existence of a duality between systematic source coding and information embedding, which is explicitly detailed in Chap. 6.

2.2 Receiver Equations

Given observations from the digital channel and the uncoded analog channel, a receiver will produce a source estimate that is optimal according to some criterion. This section derives the commonly used MMSE, MAP, and ML estimators for the hybrid channel given an arbitrary encoding function. Letting $\mathbf{x}$ denote the zero-mean source vector of length N and $\mathbf{y}$ denote the noisy observation of $\mathbf{x}$, we derive the MMSE and MAP receivers from the *a posteriori* density function, $f_{\mathbf{x}|\mathbf{y},\mathbf{m}}(\mathbf{x}|\mathbf{y},\mathbf{m})$ written in shorthand as $f(\mathbf{x}|\mathbf{y},\mathbf{m})$. Let $\mathbf{m} = \mathbf{g}(\mathbf{x})$ be the digital channel information, which is a deterministic function of $\mathbf{x}$. Denoting the inverse image of $\mathbf{m}$ as the region in signal space $\mathbf{S} = \{\mathbf{x}|\mathbf{g}(\mathbf{x}) = \mathbf{m}\}$, the *a posteriori* density is reduced as follows:

$$f(\mathbf{x}|\mathbf{y},\mathbf{m}) = \frac{f(\mathbf{x}|\mathbf{y})f(\mathbf{m}|\mathbf{x},\mathbf{y})}{f(\mathbf{m}|\mathbf{y})} \tag{2.1}$$

$$= \frac{f(\mathbf{x}|\mathbf{y})f(\mathbf{m}|\mathbf{x})}{f(\mathbf{m}|\mathbf{y})} \tag{2.2}$$

$$= \begin{cases} \frac{f(\mathbf{x}|\mathbf{y})}{f(\mathbf{m}|\mathbf{y})} & \text{if } \mathbf{x} \in \mathbf{S} \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Equation (2.2) follows from the fact that $\mathbf{m}$ and $\mathbf{y}$ are conditionally independent given $\mathbf{x}$. Equation (2.3) follows from the fact that the side channel information $\mathbf{m}$ is deterministically related to the source $\mathbf{x}$. Conditioned on a particular value of $\mathbf{x} = \mathbf{x}$, $\mathbf{m} = \mathbf{g}(\mathbf{x})$ with probability 1. Note that the *a posteriori* density in (2.3) is simply the pdf of the conditional random vector $\mathbf{x}|\mathbf{y} \sim f(\mathbf{x}|\mathbf{y})$, further conditioned on the event that $\mathbf{x} \in \mathbf{S}$. To within the constant normalization factor $f(\mathbf{m}|\mathbf{y})$, the density $f(\mathbf{x}|\mathbf{y},\mathbf{m})$ is identical to $f(\mathbf{x}|\mathbf{y})$ inside the region of support described by $\mathbf{g}(\mathbf{x}) = \mathbf{m}$ and zero otherwise. Given this understanding, the nature of the two estimators is quite clear.

The MMSE estimator is given by $\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x}|\mathbf{y},\mathbf{m}]$. The form of the estimator can be simplified using equation (2.3):

$$\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x}|\mathbf{y},\mathbf{m}] \tag{2.4}$$

$$= \int_{-\infty}^{\infty} \mathbf{x} f(\mathbf{x}|\mathbf{y}, \mathbf{m}) d\mathbf{x} \tag{2.5}$$

$$= \frac{1}{f(\mathbf{m}|\mathbf{y})} \int_{\mathbf{S}} \mathbf{x} f(\mathbf{x}|\mathbf{y}) d\mathbf{x}. \tag{2.6}$$

The MMSE estimator is simply the centroid of the density $f(\mathbf{x}|\mathbf{y})$ in constraint region $\mathbf{S}$. Note that $\hat{\mathbf{x}}_{MMSE}$ is not in general an element of $\mathbf{S}$. For the case where $\mathbf{S}$ is convex, it is clear that $\hat{\mathbf{x}}_{MMSE} \in \mathbf{S}$. Generalizing (2.4) for all distortion measures $D(.,.)$, we have a generalized conditional centroid rule for minimum mean distortion (MMD) estimation:

$$\hat{\mathbf{x}}_{MMD} = \arg\min_{h(\cdot)} E[D(\mathbf{x}, h(\mathbf{y}))|\mathbf{y}, \mathbf{x} \in \mathbf{S}]. \tag{2.7}$$

We assume in (2.7) that the minimum exists, which is the case for most distortion measures and distributions seen in practice. In some cases, there will be many (perhaps infinitely many) estimates that satisfy the minimization in (2.7).

The MAP estimator, $\hat{\mathbf{x}}_{MAP}$, is the value of $\mathbf{x}$ that maximizes the *a posteriori* density $f(\mathbf{x}|\mathbf{y}, \mathbf{m})$, which is given by:

$$\hat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x} \in \mathbf{S}} f(\mathbf{x}|\mathbf{y}). \tag{2.8}$$

The estimator is the solution to a constrained optimization problem. An example of a tractable problem is if $f(\mathbf{x}|\mathbf{y})$ is a Gaussian distribution and $\mathbf{S}$ is convex. The MAP estimate is then a maximum of a concave function over a convex set, which can be solved by a number of numerical algorithms.

The ML signal estimate assumes no prior statistics for the source signal. The ML estimate of the source is the $\mathbf{x}$ that maximizes the following likelihood function:

$$f(\mathbf{y}, \mathbf{m}|\mathbf{x}) = f(\mathbf{y}|\mathbf{m}, \mathbf{x}) f(\mathbf{m}|\mathbf{x}) \tag{2.9}$$

$$= f(\mathbf{y}|\mathbf{x}) f(\mathbf{m}|\mathbf{x}) \tag{2.10}$$

$$= \begin{cases} f(\mathbf{y}|\mathbf{x}) & \text{if } \mathbf{x} \in \mathbf{S} \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

The ML estimate is thus the result of maximizing the likelihood $f(\mathbf{y}|\mathbf{x})$ over the constraint region $\mathbf{S}$.

For the case of additive white Gaussian noise, the surfaces of equal likelihood, as a function

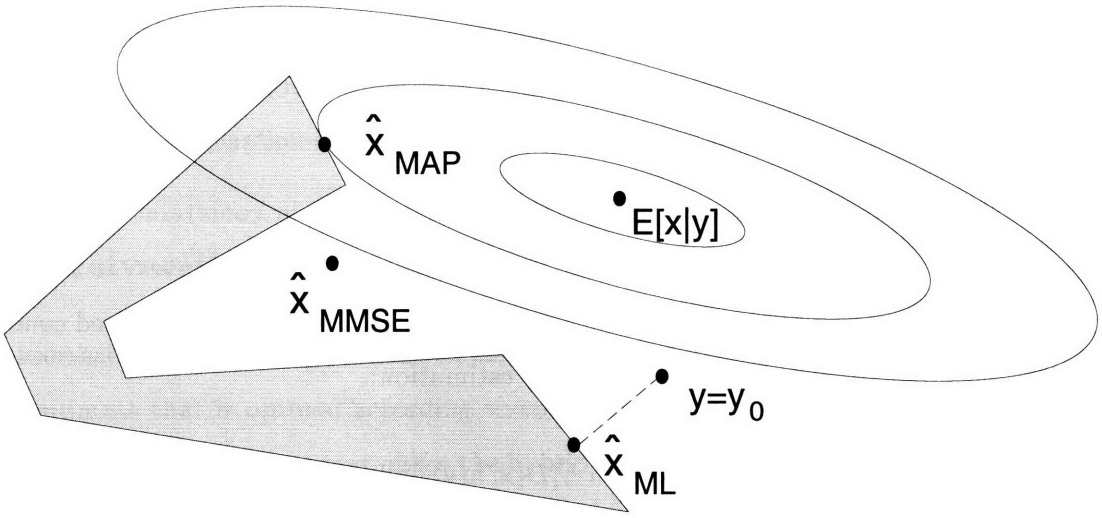


Figure 2-1: MMSE, MAP, and ML estimates. The shaded region is the constraint region **S**. The ellipses are the contours of equal probability for the density $f(\mathbf{x}|\mathbf{y}_0)$, the maximum of which is at $E[\mathbf{x}|\mathbf{y} = \mathbf{y}_0]$.

of $\mathbf{x}$, are equidistant from the point $\mathbf{x} = \mathbf{y}$, which implies that the ML estimate is the minimum distance projection of $\mathbf{y} = \mathbf{y}$ onto the constraint set **S**:

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \arg\min_{\mathbf{u}\in\mathbf{S}} \sum_{i=1}^{N} (u_i - y_i)^2. \tag{2.12}$$

In the frequency domain, the equation is written as:

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \arg\min_{x\in\mathbf{S}} \int_0^{2\pi} |X(e^{j\omega}) - Y(e^{j\omega})|^2 d\omega, \tag{2.13}$$

where $X(e^{j\omega})$ and $Y(e^{j\omega})$ denote the discrete Fourier transforms of the vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. In section 2.3.2 we examine the the case where $\mathbf{y}$ is speech in additive Gaussian noise and the digital channel carries spectral envelope parameters of $\mathbf{x}$. An algorithm is developed that yields the ML signal estimate.

Figure 2-1 shows an example of the estimators for a two-dimensional Gaussian random vector $\mathbf{x}$ and an additive white Gaussian channel with a noisy realization from the channel $\mathbf{y} = \mathbf{y}_0$. The shaded region is the constraint region **S**, representing all signals $\mathbf{x}$ meeting the constraints. The ellipses are the contours of equal probability for the density $f(\mathbf{x}|\mathbf{y})$, the maximum of which is at $E[\mathbf{x}|\mathbf{y} = \mathbf{y}_0]$. Note that $\hat{\mathbf{x}}_{\mathrm{ML}}$ and $\hat{\mathbf{x}}_{MAP}$ are in **S**, while $\hat{\mathbf{x}}_{MMSE}$ is not. Understanding the receiver

structures will assist us in the design of the systematic encoder.

2.3 Examples of optimal decoders

In this section we derive the optimal decoders for two simple, but useful forms of digital side information: linear side information and spectral envelope side information. Linear side information is simply defined as the linear projection of the source vector $\mathbf{x}$ onto one or more basis vectors $\mathbf{h}_i$, $i = 1, ..., L$ where L is the number of coefficients transmitted to the receiver. For the case of jointly Gaussian $\mathbf{x}$ and $\mathbf{y}$ with linear side information at the receiver, the MMSE estimate (which coincides with the MAP estimate) has a simple solution. We derive the optimum basis vectors as a function of the second order statistics for $\mathbf{x}$ and $\mathbf{y}$. This result is a gives an indication of the types of solutions that we can expect for optimal transform and subband coder designs in Chap. 5. Assuming an additive Gaussian channel we also derive the ML estimate with linear side information.

When the source that is being coded is speech, spectral envelope parameters (*e.g.*, linear predictive coefficients or zero phase coefficients) capture a great deal of information about the source, and are therefore appropriate for use as side information. In Section 2.3.2 we describe an iterative algorithm that attains the ML estimate of speech given an AWGN channel and spectral envelope side information.

Note that in both of these examples the side information sent across the supplemental digital channel is not truly digital, as the coefficients are assumed unquantized. These results are, however, instructive as they indicate the receiver behavior in the limit of low quantization distortion. Although the side information is not digital in an exact sense, we maintain the naming convention "digital channel" for the supplementary channel through which deterministic information about the source is sent.

2.3.1 Gaussian source/channel with linear side information

In this section, we explore a familiar problem of signal reconstruction with linear side information. Typically the problem is approached by constraining the estimator to have a particular form. In this section we impose no such restriction as we seek the unconstrained MMSE estimate. Depite the more general problem formulation, the results concerning optimal basis selection are consistent

with the less general approach. Linear side information is generated at the encoder by applying a matrix operation to the source vector $\mathbf{x}$. The length-N basis vectors, $\mathbf{h}_i$, $i = 1, ..., L$ form the columns of encoding matrix $\mathbf{H}_{N\times L}$. Without loss of generality, the $\mathbf{h}_i$ are assumed orthogonal:

$$\mathbf{h}_i^\dagger \mathbf{h}_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

The vector $\mathbf{m} = \mathbf{H}^\dagger \mathbf{x}$ is output from the encoder and sent over the digital channel. We examine the MMSE, MAP and ML estimates when the side information is linear.

Denoting the MMSE estimate by $\hat{\mathbf{x}}$, we let the error be given by $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$. For a general joint distribution $f_{\mathbf{xy}}(\mathbf{x}, \mathbf{y})$ on the source and channel output, the MMSE receiver estimate has no convenient closed form solution, besides the usual $E[\mathbf{x}|\mathbf{y}, \mathbf{m}]$. The estimate is the centroid of the pdf $f(\mathbf{x}|\mathbf{y})$ in the affine subspace $S = \{\mathbf{x}|\mathbf{H}^\dagger\mathbf{x} = \mathbf{m}\}$. Clearly the region S is convex, which implies that $\hat{\mathbf{x}}_{\text{MMS}} \in S$, but this fact does not simplify the problem. In fact, we must not be mislead by this observation to assume that the MMSE estimate is obtained by minimizing the MSE $E[(\mathbf{x} - \hat{\mathbf{x}})^\dagger(\mathbf{x} - \hat{\mathbf{x}})|\mathbf{y}]$ subject to the constraint $\mathbf{H}^\dagger\mathbf{x} = \mathbf{m}$. Indeed, this solution yields an estimate that has MSE greater than or equal to the MMSE error estimate using analog information alone.

Uncorrelated x and y

In general, we are considering jointly Gaussian $\mathbf{x}$ and $\mathbf{y}$ distributed as $\mathcal{N}([\mu_\mathbf{x}\ \mu_\mathbf{y}], \mathbf{\Lambda_{xy}})$. To assist with the general solution, we first consider the case where the analog observation $\mathbf{y}$ is uncorrelated with $\mathbf{x}$, implying that our MMSE estimator, which is linear for the Gaussian case, can safely ignore $\mathbf{y}$. The only observation upon which we base our estimate is thus $\mathbf{H}^\dagger\mathbf{x} = \mathbf{m}$. Using the standard LLSE estimator equations, the optimal estimate is given by

$$\hat{\mathbf{x}} = \mu_\mathbf{x} + \mathbf{\Lambda_x H}(\mathbf{H}^\dagger\mathbf{\Lambda_x H})^{-1}(\mathbf{y} - \mu_\mathbf{y}), \tag{2.14}$$

where $\mathbf{\Lambda_x}$ is the autocovariance matrix for $\mathbf{x}$, and the error covariance matrix is given by

$$\mathbf{\Lambda_e} = \mathbf{\Lambda_x} - \mathbf{\Lambda_x H}(\mathbf{H}^\dagger\mathbf{\Lambda_x H})^{-1}\mathbf{H}^\dagger\mathbf{\Lambda_x}. \tag{2.15}$$

The trace of $\mathbf{\Lambda_e}$ is the MSE. Note that $\mathbf{\Lambda_e}$ is unaffected by the mean of $\mathbf{x}$, implying that a given affine subspace defined by the matrix $\mathbf{H}$ yields the same mean-squared error no matter what the mean.

For many applications, an encoder may be optimized for minimum error. For the case of linear side information the vectors $\mathbf{h}_i$, $i = 1, ..., L$, may be selected to minimize MSE. We present a solution that yields a local minimum in MSE with respect to the $\mathbf{h}_i$s. We first consider the case where $\mathbf{H}$ is made up of only one vector $\mathbf{h}$, *i.e.,* we transmit only one coefficient. We let the eigenvalue decomposition of $\mathbf{\Lambda_x}$ be given by

$$\mathbf{\Lambda_x} = \mathbf{UQU}^\dagger, \tag{2.16}$$

where $\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ ... \ \mathbf{u}_]$ is the unitary matrix whose columns are the unit eigenvectors of $\mathbf{\Lambda_x}$, and $\mathbf{Q} = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_n\}$ is a diagonal matrix with the eigenvalues of $\mathbf{\Lambda_x}$ as diagonal entries. Without loss of generality, we assume the eigenvalues are ordered

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n. \tag{2.17}$$

The error covariance matrix is thus

$$\mathbf{\Lambda_e} = \mathbf{U}(\mathbf{Q} - \frac{\mathbf{QU}^\dagger\mathbf{hh}^\dagger\mathbf{UQ}}{\mathbf{h}^\dagger\mathbf{UQU}^\dagger\mathbf{h}})\mathbf{U}^\dagger \tag{2.18}$$

Defining $\mathbf{z} = \mathbf{Q}^{1/2}\mathbf{U}^\dagger\mathbf{h}$ and using the fact that $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$, we see that the MSE is

$$\begin{aligned}
E[\mathbf{e}^\dagger\mathbf{e}] &= \text{Tr}(\mathbf{\Lambda_e}) && (2.19)\\
&= \text{Tr}(\mathbf{Q}) - \frac{\text{Tr}(\mathbf{Q}^{1/2}\mathbf{zz}^\dagger\mathbf{Q}^{1/2})}{\mathbf{z}^\dagger\mathbf{z}} && (2.20)\\
&= \text{Tr}(\mathbf{Q}) - \frac{\text{Tr}(\mathbf{z}^\dagger\mathbf{Qz})}{\mathbf{z}^\dagger\mathbf{z}} && (2.21)\\
&= \text{Tr}(\mathbf{Q}) - \frac{1}{\mathbf{z}^\dagger\mathbf{z}}\sum_{i=1}^{n}\lambda_i z_i^2 && (2.22)\\
&\geq \text{Tr}(\mathbf{Q}) - \lambda_1. && (2.23)
\end{aligned}$$

The inequality in (2.23) is met with equality for $\tilde{\mathbf{z}} = \frac{\mathbf{z}}{||\mathbf{z}||}$ equaling the elementary vector [1 0 ... 0], thus implying that minimum MSE is achieved for $\mathbf{h} = \mathbf{u}_1$, the eigenvector corresponding to the

largest eigenvalue of $\mathbf{\Lambda_e}$.

We consider the generalization of the result for one vector $\mathbf{h}$ to the case where $\mathbf{H}$ has L vectors which we select to minimize MSE. We assume the matrix $\mathbf{H}^*$ is comprised of column vectors $\mathbf{h}_i^* = \mathbf{u}_i$, $i = 1, ..., L$, the eigenvectors corresponding to the m largest eigenvalues of $\mathbf{\Lambda_e}$, and show that this matrix is a point of zero gradient of the MSE surface with respect to the matrix entries. Consider minimizing MSE with respect to the j^{th} vector $\mathbf{h}_j$, assuming the other vectors in $\mathbf{H} = \mathbf{H}^*$ are fixed. Given our assumptions, the receiver observes the coefficients $\mathbf{u}_i^\dagger \mathbf{x}$, $i = 1, ..., L$, $i \neq j$. The aposteriori density is thus $N - (L-1)$ dimensional, and its autocovariance matrix has eigenvalues $\{\lambda_j, \lambda_{N-L+1}, \lambda_{N-L+2}, ..., \lambda_N\}$, of which λ_j is the largest. By the above result for the one vector case, the MSE cannot be reduced by perturbing the vector $\mathbf{h}_j$ away from $\mathbf{u}_j$. This holds for all j implying that $\mathbf{h}_i = \mathbf{u}_i$, $i = 1, ..., m$, achieves at least a local optimum (zero-gradient point) of the MSE. We use this result to solve for the optimum matrix $\mathbf{H}$ for the case of general correlation between $\mathbf{x}$ and $\mathbf{y}$.

Correlated x and y

For $\mathbf{x}$ and $\mathbf{y}$ jointly Gaussian and correlated, we again use the standard LLSE equations to derive the optimal estimate given $\mathbf{y}$ and $\mathbf{m} = \mathbf{Hx}$ at the receiver. Denoting $\tilde{\mathbf{y}} = [\mathbf{x}^\dagger H \;\; y^\dagger]^\dagger$ as the observed vector at the receiver, the MMSE receiver estimate is

$$\hat{\mathbf{x}}(\mathbf{y}, \mathbf{m}) = \mu_{\mathbf{x}} + \mathbf{\Lambda}_{\mathbf{x}\tilde{\mathbf{y}}} \mathbf{\Lambda}_{\tilde{\mathbf{y}}}^{-1} (\tilde{\mathbf{y}} - \mu_{\tilde{\mathbf{y}}}). \tag{2.24}$$

Defining the error vector $\tilde{\mathbf{e}} = \mathbf{x} - \tilde{\mathbf{x}}$, we have the error covariance matrix

$$\mathbf{\Lambda}_{\tilde{\mathbf{e}}} = \mathbf{\Lambda}_{\mathbf{x}} - \mathbf{\Lambda}_{\mathbf{x}\tilde{\mathbf{y}}} \mathbf{\Lambda}_{\tilde{\mathbf{y}}}^{-1} \mathbf{\Lambda}_{\mathbf{x}\tilde{\mathbf{y}}}^\dagger. \tag{2.25}$$

The matrices $\mathbf{\Lambda}_{\mathbf{x}\tilde{\mathbf{y}}}$ and $\mathbf{\Lambda}_{\tilde{\mathbf{y}}}$ are respectively given by the following expressions:

$$\mathbf{\Lambda}_{\mathbf{x}\tilde{\mathbf{y}}} = [\mathbf{\Lambda}_{\mathbf{x}}\mathbf{H} \;\; \mathbf{\Lambda}_{\mathbf{xy}}], \tag{2.26}$$

$$\mathbf{\Lambda}_{\tilde{\mathbf{y}}} = \begin{bmatrix} \beta & \mathbf{\Psi} \\ \mathbf{\Psi}^\dagger & \mathbf{\Theta} \end{bmatrix}, \tag{2.27}$$

where $\beta = \mathbf{H}^\dagger \mathbf{\Lambda_x} \mathbf{H}$, $\mathbf{\Psi} = \mathbf{H}^\dagger \mathbf{\Lambda_{xy}}$, and $\mathbf{\Theta} = \mathbf{\Lambda_y}$. As the expression in (2.24) and (2.25) involve the term $\mathbf{\Lambda_{\tilde{y}}}^{-1}$, we determine the matrix inverse using a common matrix inversion formula. First we let $\bar{\mathbf{x}}(\mathbf{y})$ be the MMSE estimate of $\mathbf{x}$ given only the observation $\mathbf{y}$ at the decoder, and we define the error $\bar{\mathbf{e}} = \mathbf{x} - \bar{\mathbf{x}}(\mathbf{y})$. Given $\mathbf{\Lambda_{\tilde{y}}}$ as in (2.27), we determine matrix inverse to be [49]

$$\mathbf{\Lambda_{\tilde{y}}}^{-1} = \begin{bmatrix} \tilde{\beta} & \tilde{\mathbf{\Psi}} \\ \tilde{\mathbf{\Psi}}^\dagger & \tilde{\mathbf{\Theta}} \end{bmatrix}, \tag{2.28}$$

where the component matrices are

$$\tilde{\beta} = (\beta - \mathbf{\Psi}\mathbf{\Theta}^{-1}\mathbf{\Psi}^\dagger)^{-1}, \tag{2.29}$$

$$\tilde{\mathbf{\Psi}} = -(\beta - \mathbf{\Psi}\mathbf{\Theta}^{-1}\mathbf{\Psi}^\dagger)^{-1} \cdot (\mathbf{\Psi}\mathbf{\Theta}^{-1}), \tag{2.30}$$

$$\tilde{\mathbf{\Theta}} = \mathbf{\Theta}^{-1} + (\mathbf{\Psi}\mathbf{\Theta}^{-1})^\dagger \cdot (\beta - \mathbf{\Psi}\mathbf{\Theta}^{-1}\mathbf{\Psi}^\dagger)^{-1} \cdot (\mathbf{\Psi}\mathbf{\Theta}^{-1}), \tag{2.31}$$

which simplify to

$$\tilde{\beta} = (\mathbf{H}^\dagger \mathbf{\Lambda_{\bar{e}}} \mathbf{H}^\dagger)^{-1} \tag{2.32}$$

$$\tilde{\mathbf{\Psi}} = -(\mathbf{H}^\dagger \mathbf{\Lambda_{\bar{e}}} \mathbf{H})^{-1} (\mathbf{H}^\dagger \mathbf{\Lambda_{xy}} \mathbf{\Lambda_y}^{-1}) \tag{2.33}$$

$$\tilde{\mathbf{\Theta}} = \mathbf{\Lambda_y}^{-1} + \mathbf{\Lambda_y}^{-1} \mathbf{\Lambda_{xy}}^\dagger \mathbf{H} \cdot (\mathbf{H}^\dagger \mathbf{\Lambda_{\bar{e}}} \mathbf{H}^\dagger)^{-1} \cdot (\mathbf{H}^\dagger \mathbf{\Lambda_{xy}} \mathbf{\Lambda_y}^{-1}) \tag{2.34}$$

where

$$\mathbf{\Lambda_{\bar{e}}} = \mathbf{\Lambda_x} - \mathbf{\Lambda_{xy}} \mathbf{\Lambda_y}^{-1} \mathbf{\Lambda_{xy}}^\dagger \tag{2.35}$$

is the error covariance matrix for $\bar{\mathbf{x}}(\mathbf{y})$, the MMSE estimate given only $\mathbf{y}$ at the decoder.

We approach the design of the optimal linear encoder given observations $\mathbf{y}$ and $\mathbf{m} = \mathbf{Hx}$ at the decoder. We derive the basis vectors $\mathbf{h}_i$, $i = 1, ..., L$, such that MSE is minimized. The equation $\mathbf{H}^\dagger \mathbf{x} = \mathbf{m}$ defines $\mathbf{S}$, an $N - L$ dimensional affine subspace of $\mathcal{X} = \mathcal{R}^N$ that contains the realization of $\mathbf{x}$. Recall that the MMSE estimate of $\mathbf{x}$ given $\mathbf{H}^\dagger \mathbf{x}$ and $\mathbf{y}$ is the centroid of $f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ in the subspace S. The aposteriori density is normal $\mathcal{N}(\bar{\mathbf{x}}(\mathbf{y}), \mathbf{\Lambda_{\bar{e}}})$.

We use the results of Sec. 2.3.1, exploiting the fact that optimal basis selection is independent of mean, to derive the optimal basis for correlated $\mathbf{x}$ and $\mathbf{y}$. To do so we define the following problem. Let $\mathbf{r}$ be a random vector distributed as $\mathcal{N}(\mu_\mathbf{r}, \mathbf{\Lambda_{\bar{e}}})$, and let $\mathbf{m} = \mathbf{H}^\dagger \mathbf{r}$ be observed at the

decoder. The MMSE estimate $\hat{\mathbf{r}}(\mathbf{m})$ is the centroid of $f_{\mathbf{r}}(\mathbf{r})$ in the subspace defined by $\mathbf{H}^{\dagger}\mathbf{r} = \mathbf{m}$, and the MSE is independent of the mean $\mu_{\mathbf{r}}$. If we assume, then, that $\mathbf{mu_r} = \bar{\mathbf{x}}(\mathbf{y})$, we have the equivalence $f_{\mathbf{r}}(\mathbf{r}) = f_{\mathbf{x}|\mathbf{y}}(\mathbf{r}|\mathbf{y})$. As a consequence, the vectors $\mathbf{h}_i$, $i = 1, ..., L$, that minimize MSE for $\hat{\mathbf{r}}(\mathbf{m})$ also minimize the MSE for $\hat{\mathbf{x}}(\mathbf{y}, \mathbf{m})$. Thus the vectors $\mathbf{h}_i$, $i = 1, ..., L$ that minimize $\mathbf{y}$ and $\mathbf{m} = \mathbf{Hx}$ at the decoder are the eigenvectors of $\mathbf{\Lambda_{\tilde{e}}}$ (the error covariance matrix given only $\mathbf{y}$ at the decoder) corresponding to the L largest eigenvalues of $\mathbf{\Lambda_{\tilde{e}}}$, $\lambda_1 \geq ... \geq \lambda_L$. If we use these optimal vectors, then the component elements of $\mathbf{\Lambda_{\tilde{y}}}^{-1}$ given in (2.32)-(2.34) simplify by substituting $(\mathbf{H}^{\dagger}\mathbf{\Lambda_e}\mathbf{H})^{-1} = \mathrm{diag}(1/\lambda_1, ...1/\lambda_L)$.

Non-Gaussian x and y

The assumption that $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian is necessary for the eigenvectors of $\mathbf{\Lambda_{\tilde{e}}}$ to be optimal linear encoding vectors, a fact we confirm by counterexample. Consider the two-dimensional random vector $\mathbf{x} = [x_1 \;\; x_2]^{\dagger} \in \mathcal{R}^2$ with discrete pmf

$$p_{\mathbf{x}}(\mathbf{x}) = \begin{cases} \frac{1}{4}, & (x_1, x_2) \in \{(0,1), (0,-1), (1,0), (-1,0)\} \\ 0, & \text{otherwise} \end{cases} \tag{2.36}$$

We examine the case where the encoding matrix $\mathbf{H}$ is a one-dimensional vector $\mathbf{h}$, and the only observed information at the encoder is $\mathbf{h}^{\dagger}\mathbf{x}$. We easily confirm that the autocorrelation matrix for $\mathbf{x}$ is $\mathbf{\Lambda_x} = \frac{1}{2}\mathbf{I}$. Therefore, all unit vectors $\mathbf{u}$ are eigenvalues of $\mathbf{\Lambda_x}$, each with the same eigenvalue $\lambda = \frac{1}{2}$. Because both eigenvalues are equal, we assume by hypothesis, that $\mathbf{h}$ equaling any unit eigenvector (hence any unit vector) will be a locally optimum encoding vector with respect to MSE. We will show by contradiction that this can not be true.

Because $\mathbf{h}$ is two dimensional and is constrained to be a unit vector, it may be parameterized by a single scalar parameter α:

$$\mathbf{h}(\alpha) = \begin{bmatrix} \alpha \\ \sqrt{1-\alpha^2} \end{bmatrix}. \tag{2.37}$$

It follows that the MMSE, which we denote $E(\alpha)$, is also a function of α. Consider the two encoding unit vectors, $\mathbf{h}^1 = \mathbf{h}(0) = [0 \;\; 1]^{\dagger}$ and $\mathbf{h}^1 = \mathbf{h}(3/5) = 1/5[3 \;\; 4]^{\dagger}$. Given $\mathbf{m} = (\mathbf{h}^i)^{\dagger}\mathbf{x}$, $i = 0, 1$, at the decoder, the MMSE estimate is denoted $\hat{\mathbf{x}}_i$, and the error is given by $\mathbf{e}_i = \mathbf{x} - \hat{\mathbf{x}}_i$.

We determine the MSE for $\hat{\mathbf{x}}_0$ by noting that the observation vector $(\mathbf{h}^0)^{\dagger}\mathbf{x}$ takes on one of three

values, 0, 1, and -1, with probabilities $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{4}$, respectively and for each of these values, the error $\mathbf{e}_0$ is zero mean. Thus the MSE is given by

$$\begin{aligned} E(0) &= E[||\mathbf{e}_0||^2] && (2.38)\\ &= \mathrm{P}(\mathbf{m}=0)E[||\mathbf{e}_0||^2|\mathbf{m}=0]+ && (2.39)\\ &\quad \mathrm{P}(\mathbf{m}=1)E[||\mathbf{e}_0||^2|\mathbf{m}=1]+\mathrm{P}(\mathbf{m}=-1)E[||\mathbf{e}_0||^2|\mathbf{m}=-1] && (2.40)\\ &= \frac{1}{2}\cdot 1+\frac{1}{4}\cdot 0+\frac{1}{4}\cdot 0=\frac{1}{2} && (2.41)\end{aligned}$$

The MMSE $E(3/5)$ for $\hat{\mathbf{x}}_1$ clearly equals zero, as it takes on one of four values, $\pm 3/5$, $\pm 4/5$, each corresponding to a unique $\mathbf{x}$.

The mean value theorem implies that there exists an α_3, $0 < \alpha_3 < 3/5$, for which

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}E(\alpha)|_{\alpha=\alpha_3}=\frac{E(3/5)-E(0)}{3/5-0}=\frac{-1/2}{3/5}=-\frac{5}{6}\neq 0 \qquad (2.42)$$

The vector $\mathbf{h}(\alpha_3)$ does not attain a zero gradient on the MMSE surface. Thus, there exists an eigenvector that is not locally optimal, which is a contradiction to our hypothesis.

ML estimate for additive Gaussian channel

For the case of linear digital side information $\mathbf{m}=\mathbf{H}^\dagger\mathbf{x}$ and a channel output $\mathbf{y}=\mathbf{x}+\mathbf{v}$, where $\mathbf{v}$ is additive Gaussian noise $\sim\mathcal{N}(\mathbf{0},\mathbf{\Lambda})$, we derive a convenient closed-form solution for the ML source estimate. A generalization of (2.12), the ML estimate in this case is given by

$$\hat{\mathbf{x}}_{\mathrm{ML}}=\arg\min_{\mathbf{u}\in\mathrm{S}}(\mathbf{y}-\mathbf{u})^\dagger\mathbf{\Lambda}^{-1}(\mathbf{y}-\mathbf{u}), \qquad (2.43)$$

where the set $\mathbf{S}$ is the affine subspace defined by the equation $\mathbf{H}^\dagger\mathbf{u}=\mathbf{m}$. The expression on the right hand side of (2.43) is the weighted-norm distance between $\mathbf{y}$ and $\mathbf{u}$, where $\mathbf{\Lambda}^{-1}$ is the weighting matrix.

We define

$$\mathbf{z}=\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{y}-\mathbf{u}), \qquad (2.44)$$

so that,

$$\mathbf{z}^\dagger\mathbf{z} = (\mathbf{y}-\mathbf{u})^\dagger\mathbf{\Lambda}^{-1}(\mathbf{y}-\mathbf{u}), \tag{2.45}$$

and

$$\mathbf{H}^\dagger\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z} = \mathbf{H}^\dagger\mathbf{y} - \mathbf{H}^\dagger\mathbf{u}, \tag{2.46}$$

which, under the constraint $\mathbf{u} \in \mathbf{S}$, becomes

$$\mathbf{H}^\dagger\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z} = \mathbf{H}^\dagger\mathbf{y} - \mathbf{m} = \mathbf{a}, \tag{2.47}$$

where $\mathbf{a}$ is a known constant vector at the receiver. Noting from (2.44), that $\mathbf{u} = \mathbf{y} - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z}$, (2.43) can be rewritten as

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \mathbf{y} - \mathbf{\Lambda}^{\frac{1}{2}}\Big(\arg \min_{\mathbf{z}:\mathbf{H}^\dagger\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z}=\mathbf{a}} \mathbf{z}^\dagger\mathbf{z}\Big). \tag{2.48}$$

Or equivalently, we are solving the under-constrained set of equations

$$\mathbf{F}^\dagger\mathbf{z} = \mathbf{a}, \tag{2.49}$$

where $\mathbf{F}^\dagger = \mathbf{H}^\dagger\mathbf{\Lambda}^{\frac{1}{2}}$, for the minimum norm solution $\mathbf{z}$. This problem has a well-known solution which is

$$\hat{\mathbf{z}} = \mathbf{F}(\mathbf{F}^\dagger\mathbf{F})^{-1}\mathbf{a}. \tag{2.50}$$

Thus we have a closed form for the ML estimate:

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \mathbf{y} - \mathbf{\Lambda}\mathbf{H}(\mathbf{H}^\dagger\mathbf{\Lambda}\mathbf{H})^{-1}(\mathbf{H}^\dagger\mathbf{y} - \mathbf{m}). \tag{2.51}$$

2.3.2 Spectral envelope side information

For many useful applications the source vector $\mathbf{x}$ being coded is a short-time windowed section of speech. Linear predictive (LP) coefficients provide an efficient representation of windowed speech, and are therefore an appropriate choice of side information for transmission through a low bit-rate side channel. In this section, we will use the coefficients $\{\alpha_i,\ i = 0, 1, ..., M\}$ of the LP filter of order M as the digital side information $\mathbf{m}$ describing the inverse image region $\mathbf{S}$. The analog side information is $\mathbf{y} = \mathbf{x} + \mathbf{v}$, where $\mathbf{v}$ is AWGN of variance $\sigma_{\mathbf{v}}^2$. In this section, we develop an algorithm

that yields the ML estimate given the above analog and digital information at the receiver. Recall that for the AWGN channel, the ML estimate is the signal point in $\mathbf{S}$ of minimum two-norm distance to $\mathbf{y}$.

Because it is derived from values of the autocorrelation function of the clean speech, the side information represents a constraint only on the Fourier transform magnitude of the estimate. We first show that if we impose no constraints on phase, the minimum-distance element to $\mathbf{y}$ will have the same phase as $Y(e^{j\omega})$. In equation 2.13 note that the integral is minimized if the distance between $X(e^{j\omega})$ and $Y(e^{j\omega})$ is minimized for all ω. Consider a particular frequency $\omega = \omega_0$, and let $X(e^{j\omega_0}) = |X|e^{j\theta}$ and $Y(e^{j\omega_0}) = |Y|e^{j\phi}$. The squared distance between $X(e^{j\omega_0})$ and $Y(e^{j\omega_0})$ is

$$\begin{aligned} J &= |X(e^{j\omega_0}) - Y(e^{j\omega_0})|^2 && (2.52)\\ &= (|X|\cos\theta - |Y|\cos\phi + j(|X|\sin\theta - |Y|\sin\phi)) \times && (2.53)\\ &\quad (|X|\cos\theta - |Y|\cos\phi - j(|X|\sin\theta - |Y|\sin\phi)) && (2.54)\\ &= |X|^2 + |Y|^2 - 2|X||Y|(\cos\theta\cos\phi + \sin\theta\sin\phi) && (2.55)\\ &= |X|^2 + |Y|^2 - 2|X||Y|(\cos(\theta - \phi)), && (2.56) \end{aligned}$$

which is minimized when $\theta = \phi$. Thus the minimum distance estimate will have the same phase as the noisy realization.

Knowing that the estimate shares the same phase as the noisy realization allows us to specify a new constraint set $\mathbf{S}' = \{x \in \mathbf{S} \mid \angle X(e^{j\omega}) = \angle Y(e^{j\omega})\}$ and to simplify the expression in equation (2.13):

$$\hat{x}_{\mathrm{ML}} = \arg\min_{\mathbf{x}\in\mathbf{S}'} \int_0^{2\pi} ||X(e^{j\omega})| - |Y(e^{j\omega})||^2 d\omega. \tag{2.57}$$

There is no clear solution to the constrained minimization in equation (2.57). A simple solution results, however, if we consider a slightly modified distance measure:

$$\hat{x} = \arg\min_{\mathbf{x}\in\mathbf{S}'} \int_0^{2\pi} ||X(e^{j\omega})|^2 - |Y(e^{j\omega})|^2|^2 d\omega. \tag{2.58}$$

The time domain expression for equation (2.58) is

$$\hat{x} = \arg\min_{\mathbf{x}\in\mathbf{S}'} \sum_i (R_{\mathbf{x}}[i] - R_{\mathbf{y}}[i])^2, \tag{2.59}$$

where $R_{\mathbf{x}}$ and $R_{\mathbf{y}}$ are the autocorrelation functions of $\mathbf{x}$ and $\mathbf{y}$ respectively. For each $\mathbf{x} \in \mathbf{S}'$ there is a one-to-one correspondence between the LP coefficients $\{\alpha_i,\ i = 0, ..., M\}$ and $\{R_{\mathbf{x}}[i],\ i = 0, ..., M\}$ [59]. For this reason, we can map the constraint set $\mathbf{S}'$ on $\mathbf{x}$ to a constraint set $\mathbf{S}_R$ on $R_{\mathbf{x}}$: $\mathbf{S}_R = \{R_{\mathbf{x}} \mid R_{\mathbf{x}}[i] = R_{\mathbf{x}}[i], i = 0, ..., M\}$. The estimate in equation (2.59) is therefore the one whose autocorrelation function is the minimum-distance projection of $R_{\mathbf{y}}$ onto $\mathbf{S}_R$, and whose phase is the same as $\mathbf{y}$.

Projection onto convex sets

In this section we describe the minimum-distance projection of $R_{\mathbf{y}}$ onto $\mathbf{S}_R$ using projection onto convex sets. The constraint set $\mathbf{S}_R$ can be described as the intersection of two convex sets. We define the sets on the Hilbert space l^2 of finite-norm discrete sequences (which is automatically satisfied by the vector length N being finite):

$$C_1 = \{u \in l^2 \mid u[i] = R_{\mathbf{x}}[i],\ i = -M, ..., M\} \tag{2.60}$$

$$C_2 = \{u \in l^2 \mid U(e^{j\omega}) \text{ real, positive } \forall\omega\} \tag{2.61}$$

The set C_2 ensures that the signals are legitimate autocorrelation functions. We denote the projection operators P_1 and P_2 which perform the minimum-distance projections onto the sets C_1 and C_2, respectively. P_1 is best described as operations in time, while P_2 is best described as an operation in frequency:

$$P_1 u[n] = \begin{cases} R_{\mathbf{x}}[n] & n = -M, ..., M \\ u[n] & \text{otherwise} \end{cases} \tag{2.62}$$

$$P_2 U(e^{j\omega}) = \begin{cases} U(e^{j\omega}) & \text{if } U(e^{j\omega}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.63}$$

Letting $\mathbf{u}_0 = R_{\mathbf{y}}$, the conventional projection onto convex sets (POCS) algorithm is given by the iteration $\mathbf{u}_{i+1} = P\mathbf{u}_i$, where $P = P_1 P_2$. The sequence $\{\mathbf{u}_i\}_{i=0}^{\infty}$ converges to some point in $\mathbf{S}_R$ [18]. In order to converge on the minimum-distance projection of $R_{\mathbf{y}}$ onto S_R, the algorithm must

be modified [11]. Let $\mathbf{u}_0 = R_\mathbf{y}$, and define the sequence $\{\mathbf{u}_i\}_{i=0}^{\infty}$ by

$$\begin{aligned}
\mathbf{u}_1 &= P_1\mathbf{u}_0, & \mathbf{v}_1 &= \mathbf{u}_1 - \mathbf{u}_0 \\
\mathbf{u}_2 &= P_2\mathbf{u}_1, & \mathbf{v}_2 &= \mathbf{u}_2 - \mathbf{u}_1 \\
\mathbf{u}_3 &= P_1(\mathbf{u}_2 - \mathbf{v}_1), & \mathbf{v}_3 &= \mathbf{v}_1 + \mathbf{u}_3 - \mathbf{u}_2 \\
\mathbf{u}_4 &= P_2(\mathbf{u}_3 - \mathbf{v}_2), & \mathbf{v}_4 &= \mathbf{v}_2 + \mathbf{u}_4 - \mathbf{u}_3 \\
\mathbf{u}_5 &= P_1(\mathbf{u}_4 - \mathbf{v}_3), & \mathbf{v}_5 &= \mathbf{v}_3 + \mathbf{u}_5 - \mathbf{u}_4 \\
\mathbf{u}_6 &= P_2(\mathbf{u}_5 - \mathbf{v}_4), & \mathbf{v}_6 &= \mathbf{v}_4 + \mathbf{u}_6 - \mathbf{u}_5 \\
&\vdots & &\vdots
\end{aligned} \tag{2.64}$$

The sequence $\{\mathbf{u}_i\}_{i=0}^{\infty}$ converges to the minimum-distance projection of $R_\mathbf{y}$ onto $\mathbf{S}_R$. The Fourier transform of the vector $\mathbf{u} = \lim_{i\to\infty} \mathbf{u}_i$ is the squared magnitude response of the desired estimate. The phase of the estimate is assigned the phase of $\mathbf{y}$, which is justified above. This estimate is not the ML estimate, however, because of the approximation made in equation (2.58).

Zero-phase impulse response coefficients

Equation (2.57) indicates an exact ML solution subject to a modification of the side information. The time domain expression for equation (2.57) is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbf{S}'} \sum_i (x_{zp}[i] - y_{zp}[i])^2, \tag{2.65}$$

where $\mathbf{x}_{zp}$ and $\mathbf{y}_{zp}$ are the zero-phase impulse responses of $\mathbf{x}$ and $\mathbf{y}$ respectively. Let the side information be $\{\mathbf{x}_{zp}[i],\ i = 0, ..., M\}$, the first $M+1$ zero-phase impulse response coefficients of $\mathbf{x}$. The side information describes a constraint set on $\mathbf{x}_{zp}$: $\mathbf{S}_{zp} = \{\mathbf{x}_{zp} \mid \mathbf{x}_{zp}[i] = \mathbf{x}_{zp}[i],\ i = 0, ..., M\}$. The set $\mathbf{S}_{zp}$ is the intersection of the following two convex sets on l^2:

$$C_1 = \{\mathbf{u} \in l^2 \mid u[i] = \mathbf{x}_{zp}[i],\ i = -M, ..., M\} \tag{2.66}$$

$$C_2 = \{\mathbf{u} \in l^2 \mid U(e^{j\omega}) \text{ real, positive } \forall\omega\}. \tag{2.67}$$

The set C_2 ensures that the sequences in $\mathbf{S}_{zp}$ are legitimate zero-phase impulse responses.

The ML estimate is calculated by means of a POCS algorithm. The projection operators P_1

and P_2, which perform the minimum-distance projections onto the sets C_1 and C_2 respectively, are given by

$$P_1 u[n] = \begin{cases} \mathbf{x}_{zp}[n] & n = -M, ..., M \\ u[n] & \text{otherwise} \end{cases} \tag{2.68}$$

$$P_2 U(e^{j\omega}) = \begin{cases} U(e^{j\omega}) & \text{if } U(e^{j\omega}) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2.69}$$

To obtain the zero-phase response of the ML estimate, the minimum-distance projection algorithm proceeds as in equation (2.64). The phase of the ML estimate is assigned the phase of $\mathbf{y}$, which is justified above.

Experiments and Results

We have implemented the algorithms presented in this section on speech and have conducted informal listening tests. Enhancement experiments have been performed using spectral envelope side information that is derived from the clean speech. There are also preliminary results for single-sensor enhancement with no side information, using LP coefficients estimated from the noisy speech. In all experiments the speech is sampled at 10kHz, and the processing is done on 20ms frames with 50% overlap.

Using LP coefficients determined from the clean speech as side information, we used the algorithm in section 2.3.2 to enhance noisy speech at several SNRs. The number of coefficients used was varied from 4 to 20. Relative to perceived output quality, there are significantly diminished returns for additional coefficients beyond 12. This is not surprising, because the spectral shaping of most speech is captured by a 12-pole model. The algorithm significantly improves intelligibility, even for very low SNRs (-10dB), and the enhanced speech is more perceptually pleasing than the noisy speech. The algorithm functions at such low SNRs because of the use of information from the clean speech. One negative aspect of the algorithm is that for low SNRs, there is a slight harshness to the enhanced speech. At high SNRs (>40dB), the enhanced speech is perceptually the same as the noisy speech. For the purposes of comparison, we consider speech enhanced by a Wiener filter, where the power spectrum of the source is approximated by a 12-pole model calculated from the clean speech. The Wiener filter approach has fewer artifacts and is more perceptually pleasing.

The same experiments were performed using zero-phase coefficients determined from the clean speech as side information. For these experiments we used the algorithm in section 2.3.2. Diminished returns in perceived quality occurs beyond 14 coefficients. The performance characteristics of the algorithm are similar to the LP algorithm: improved intelligibility at very low SNRs and no effect at high SNRs. The enhanced speech has the same perceptual properties as the speech from the LP algorithm, including a slight harshness at low SNRs.

The final experiment involves single-sensor noise reduction without side information. The LP coefficients are estimated from the noisy speech using ML estimation [45, 46]. Now, considering the speech as the unknown parameter and the LP coefficients as side information, we find an approximate ML estimate of the speech using the algorithm in section 2.3.2. Preliminary results suggest that this is a promising approach.

2.4 Overview of Digital Encoding

We have seen examples of systems in which the decoder receives deterministic side information $\mathbf{m}$ that is a vector of real numbers of infinite precision. In a realistic application, the side information is the output of a digital channel of finite rate, which requires that the side information be from a set of a finite number of levels, $I = \{1, ..., K\}$, where $K = 2^{nR}$ with R being the rate in bits/sample. To digitally encode a source we map $\mathbf{x} \in \mathcal{R}^n$ to one of K levels, or indices, $i \in I$ in a codebook and transmit the index. A decoder reconstructs the source from the index. The general term for digital encoding/decoding is *quantization.* If no analog information were present at the decoder, we have the conventional source coding scenario, and the decoder takes the level $i \in I$ from the digital channel and reconstructs the source as $\hat{\mathbf{x}}_i$, the i^{th} entry in the decoder codebook. Clearly, for $\mathbf{x} \in \mathcal{R}^n$ digitization induces distortion.

If analog information is present at the decoder, the codebook is no longer just an ensemble of reconstruction points, but rather is an ensemble of reconstruction functions $\hat{\mathbf{x}}_i(\mathbf{y})$. We refer to the case when analog information is at the decoder as *systematic quantization.* Of course, even though $\hat{\mathbf{x}}_i(\mathbf{y})$ may have infinite precision, distortion is still present at the decoder by the fact that $\mathbf{y}$ is a noisy version of $\mathbf{x}$ and $\mathbf{m}$ is of finite precision. In Chap. 4 we develop algorithms which yield systematic encoder/decoder pairs that are optimal with respect to a fidelity criterion. In this

section we introduce the concept of systematic digital encoding/decoding and show by intuitive argument, the duality of systematic coding with information embedding. A theoretical exposition of this duality is given in detail in Chap. 6.

Despite the intrinsic difference in operation of the decoder with and without analog information, the basic operation of the *encoder* does not change. The source is mapped to one of M levels. Depending the statistics of $\mathbf{x}$ and $\mathbf{y}$, the actual encoder map may vary. An encoder map may be characterized by the partitioning of $\mathcal{R}^n$ into a set of M non-overlapping regions $A = \{A_1, ...A_M\}$, each of which may or may not be a connected set. For each i, we let A_i be a union of disjoint sets, $A_i = \bigcup_j A_{ij}$, where the A_{ij} is connected for all j, and for all $k \neq j$, $A_{ik} \bigcup A_{ij}$ is not connected. In other words, the A_{ij}s are the largest connected pieces in A_i.

2.4.1 Systematic quantization

Fig. 2-2 shows an example of a partition for $N = 2$ and $K = 2$, where the lines represent the boundaries between A_1 and A_2. The region A_1, represented by the bit 0, is the union of the areas containing an $\circ$, and the region A_2, represented by the bit 1, is the union of the areas containing an $\times$. The $\times$s and $\circ$s are placed at the center of mass of their respective areas. The encoder simply transmits an $i = 0$ (respectively, an $i = 1$) if $\mathbf{x}$ lies in a region with an $\circ$ (respectively, an $\times$). In Fig. 2-2(a) the encoder transmits a 0. The basic concept behind systematic quantization is that the analog information gives a coarse description of the source, and the digital information refines the description. Thus in Fig. 2-2 we desire that the $\circ$s (respectively, $\times$s) be separated far enough apart, so that given the analog information and the index $i = 0$ (respectively, i=1), it is clear which region A_{ik} the source $\mathbf{x}$ comes from.

Fig. 2-2(b) and (c) illustrate the operation of a decoder for the given encoder map. We assume that at the decoder we wish to minimize MSE. Given only $\mathbf{y}$ the MMSE estimate is $\tilde{\mathbf{x}}(\mathbf{y}) = E[\mathbf{x}|\mathbf{y}]$. Recall from Sec. 2.2 that given i and $\mathbf{y}$, MSE is minimized if we take the centroid of $f(\mathbf{x}|\mathbf{y})$ in the region A_i. For most sources and channels of interest (including the Gaussian case), $f(\mathbf{x}|\mathbf{y})$ has most of it probability mass weighted around its mean $\tilde{\mathbf{x}}(\mathbf{y})$. We assume for most cases that for a fixed i the A_{ij}s are separated far enough such that most of the probability mass of $f(\mathbf{x}|\mathbf{y})$ is contained in A_{ik} for some k. Also we assume that A_{ik} is small enough that $f(\mathbf{x}|\mathbf{y})$ is relatively constant across A_{ik}. Note that these assumptions require that M be large; for clarity of illustration in Fig. 2-2 we

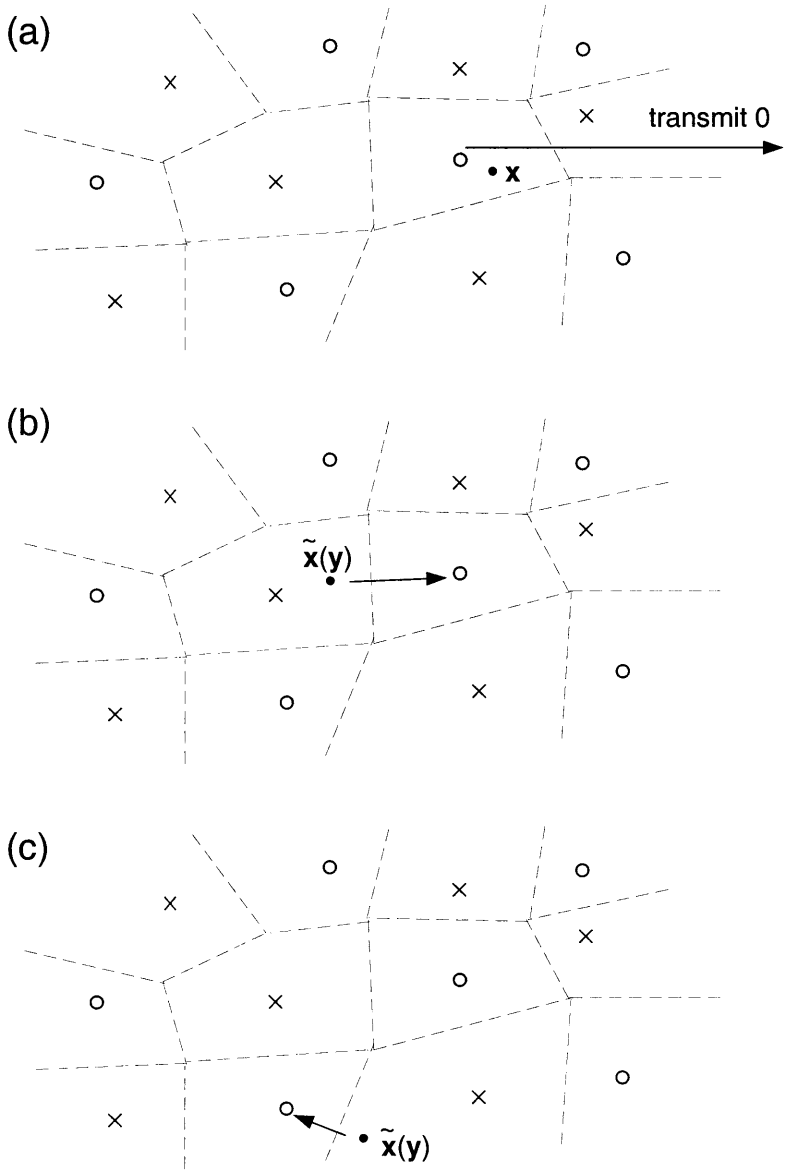


Figure 2-2: Example of systematic quantization. The region A_1, represented by the bit 0, is the union of the areas containing an $\circ$, and the region A_2, represented by the bit 1, is the union of the areas containing an $\times$. (a) The encoder. (b) The decoder with no channel error. (c) The decoder with a channel error.

have not used a large M. Given our assumptions, it is clear that a near-MMSE decoder simply maps $\tilde{\mathbf{x}}(\mathbf{y})$ to the center of mass of A_{ik}. We have illustrated this in Figs. 2-2(b) and (c) for two cases, in (b) where we have correctly identified the A_{ik} from which $\mathbf{x}$ has come, and in (c) where we have chosen the wrong A_{ik}. We call the case depicted by (c) as a *channel error* to reflect the similarity to digital communications when a receiver decodes the wrong point in a signal constellation.

The generalization of this quantization concept to larger dimensions and higher rates is clear. Low complexity implementations may be accomplished by adding structure to the quantizers. For example, consider a lattice of dimension n, which is the union of M sublattices. Define A_i as the union of the Voronoi regions around the i^{th} lattice. Given the appropriate lattice selection, we show in Chap. 3, as an extension of the results of [95], that such a structure can attain the theoretical rate-distortion bound for a Gaussian source and channel.

2.4.2 Duality to information embedding

We introduced the information embedding problem in Chap. 1. Recall that information is modulated by embedding it into a host signal, denoted $\mathbf{y}$. A recently developed method for information embedding, called quantization index modulation (QIM), is developed in [15], and has been shown to achieve the capacity of the information embedding channel for the Gaussian case [15, 2]. We show that QIM, as a result of joint work with B. Chen and G. W. Wornell, is the dual to the systematic quantization method shown in this section, in the sense that the systematic encoder serves as the QIM decoder and the systematic decoder serves as the QIM encoder. The QIM method embeds information into $\mathbf{y}$ by quantizing $\mathbf{y}$ with one of M quantizers; the information modulates which quantizer is used. Thus the QIM codebook partitions signal space into a set of non-overlapping regions $A = \{A_1, ...A_K\}$, with each $A_i = \bigcup_j A_{ij}$ defined as above. Within each A_{ij} is a reconstruction point. We modulate the information signal $i \in I$ by mapping $\mathbf{y}$ to the reconstruction point in the nearest A_{ij}; we choose the nearest point because we want to induce as little distortion as necessary.

In Fig. 2-3, we show the QIM encoder and decoder for $N = 2$ and $K = 2$. Again the region A_1, represented by the bit 0, is the union of the areas containing an $\circ$, and the region A_2, represented by the bit 1, is the union of the areas containing an $\times$. In Fig. 2-3(a) the embedded bit is a 0, and we thus map $\mathbf{y}$ to the nearest $\circ$. This operation is the same as the decoder operation in Sec. 2.4.1,

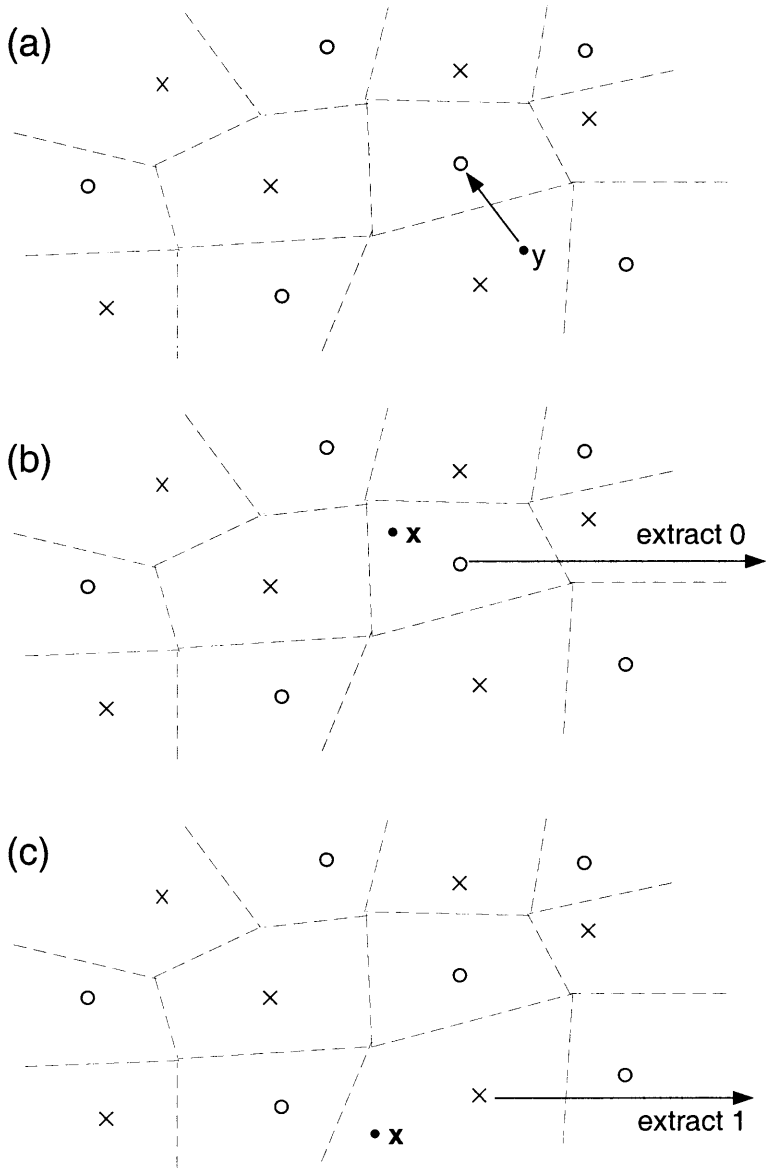


Figure 2-3: Example of QIM. The region A_1, represented by the bit 0, is the union of the areas containing an $\circ$, and the region A_2, represented by the bit 1, is the union of the areas containing an $\times$. (a) The encoder embedding a 0. (b) The decoder with no channel error. (c) The decoder with a channel error.

except in this case, the function $\tilde{\mathbf{x}}(\mathbf{y}) = \mathbf{y}$. In Figs. 2-3(b) and (c) we show QIM decoding, in which an information bit is extracted, for two cases. In (b) the decoder has correctly identified the signal point to which $\mathbf{x}$ is mapped, and hence has no error in extracting the information bit. In (c) the receiver has decoded the wrong signal point (a channel error has occurred), which induces an error in identifying the information bit. It is clear that the QIM decoder behaves as the systematic encoder in Sec. 2.4.1, as the bit extracted is a function of the index of the region A_i in which $\mathbf{x}$ resides.

2.5 Conclusion

This chapter established some of the basic principles for encoder and decoder design for source coding with side information at the decoder. We derived the optimal receivers with respect to MMSE, MAP, and ML criteria. We examined receiver design for the case of linear side information and jointly Gaussian $\mathbf{x}$ and $\mathbf{y}$. This example also allowed us to take a glimpse at optimal encoder design, as we derived the MMSE encoding matrix. A second example looked at a more practical scenario with a speech source and spectral envelope side information. Finally as a prelude to Chap. 4, we discussed the basics of the K-level quantizer design , the building block for a digital encoder, and suggested a duality between this structure and a structure used for information embedding, the QIM method. The duality between source coding with side information at the decoder and information embedding is discussed in detail at a more theoretical level in Chap. 6.

Chapter 3

Literature Review: Rate-Distortion Theory with Side Information

3.1 Introduction

In their celebrated paper [90], Wyner and Ziv established the theoretical limits of performance of source coding with analog side information at the decoder. Referring to Fig. 1-1, this is the case in which the switch S is open. Wyner and Ziv pose the question, now known as the *Wyner-Ziv problem*, given that the analog side information is known only at the decoder, what is the minimum rate R required for $\mathbf{m}$ such that the source can be reconstructed with average distortion less than some d, i.e., what is the rate-distortion function with side information at the decoder? Unless otherwise noted, in this chapter we consider the special case in which the source vector $\mathbf{x}$ is drawn iid from $p_x(x)$ and the channel is memoryless with transition density $p_{y|x}(y|x)$. Under these assumptions, Wyner and Ziv derive an elegant single-letter expression for the rate-distortion function. Prior to the publication of [90], there had only been results concerning the much simpler case where side information is known at the encoder and decoder [7, 33] (switch S is closed in Fig. 1-1). Wyner and Ziv detail the relationship between their problem and this simpler one.

The results in [90] are impractically asymptotic in nature, requiring that block lengths approach infinity. In contrast, this thesis focuses on the development of implementable, low-latency, low-complexity signal processing algorithms. The main purpose of this chapter is to establish the prior art for source coding with side information at the decoder, the cast majority of which is theoretical

in nature. Although not a focus of this thesis, in this chapter we derive some extensions and generalizations of the basic theory pertaining to the Wyner-Ziv problem. Also in Chap. 6 we refer to this chapter to explore the information theoretic duality between information embedding and rate-distortion theory with side information at the decoder. Those wishing to focus on practical signal processing implementations may choose to skip this chapter. However, we advise the reader to understand the asymptotic results as they are a mark against which to compare the signal processing algorithms developed in this thesis. Furthermore, these theoretical results and the development of these results suggest practical approaches to the systematic source coding problem.

In addition to providing an expression for the general rate-distortion function, Wyner and Ziv derive the expression for the quadratic Gaussian case [89] and the doubly binary symmetric channel with Hamming distortion [90]. In Sec. 3.3.2, we use the basic Gaussian result to derive an inverse waterpouring result for a vector of independent Gaussians. Any papers subsequent to [90] have been extensions of the basic results of Wyner and Ziv. Shamai *et al* [66] construct codes that achieve the rate distortion bound for the doubly binary symmetric case, which we review in Sec. 3.4.2. In the same paper, Shamai *et al* derive the necessary and sufficient conditions for systematic coding to be the most efficient coding method. Zamir and Shamai [95] construct nested lattice codes that achieve capacity for the quadratic Gaussian case in the limit of high SNR. We develop a generalization of this construction which achieves capacity for all SNRs in Sec. 3.3.4.

3.2 Rate-distortion functions

In this section we present the expression for the Wyner-Ziv rate-distortion function, and compare it to the expression for the case in which the analog side information $\mathbf{y}$ is known at the encoder and decoder.

3.2.1 Wyner-Ziv rate-distortion function

In their original paper, Wyner and Ziv defined the rate-distortion function with side information at the decoder, denoted $R_{x|y}^{\mathrm{WZ}}(d)$, as the minimum data rate at which $\mathbf{m}$ can be transmitted such that when N is large the average distortion $E[\frac{1}{N}\sum_{k=1}^{N} D(\mathbf{x}_k, \mathbf{w}_k)]$ is arbitrarily close to d. Their

main theorem is the following:

$$R_{x|y}^{\mathrm{WZ}}(d) = \inf I(x;u) - I(y;u) \tag{3.1}$$

where the infimum is taken over all random variables u satisfying

$$y \to x \to u \text{ is a Markov chain,} \tag{3.2}$$

and functions $f : \mathcal{U} \times \mathcal{X} \to \mathcal{W}$ satisfying

$$E[D(x,w)] \leq d \text{ where } w = f(u,y). \tag{3.3}$$

Condition (3.2), *i.e.*, y and u are conditionally independent given x, implies

$$I(y;u|x) = 0, \tag{3.4}$$

which simplifies (3.1) to be

$$R_{x|y}^{\mathrm{WZ}}(d) = \inf I(x;u|y). \tag{3.5}$$

3.2.2 Slepian-Wolf codes ($d = 0$)

An interesting special case of Wyner-Ziv encoding is when $d = 0$, *i.e.*, $\mathbf{x}$ is perfectly reconstructed at the decoder. The required rate for this operating point can be determined by the well-known Slepian-Wolf result [69], which we summarize here. Consider two discrete-alphabet iid sequences $\mathbf{u}$ and $\mathbf{v}$ jointly distributed as $p_{\mathbf{uv}}(\mathbf{u},\mathbf{v}) = \prod_{i=1}^{n} p_{uv}(u_i, v_i)$. A Slepian-Wolf code is a method for the lossless encoding of $\mathbf{u}$ and $\mathbf{v}$, individually at two separate encoders, to be decoded by a single decoder. Let R_u and R_v be the rates of the encoders of $\mathbf{u}$ and $\mathbf{v}$ respectively. If both $\mathbf{u}$ and $\mathbf{v}$ were encoded at a single encoder the total rate of transmission for lossless recovery of $\mathbf{u}$ and $\mathbf{v}$ is clearly $H(u,v)$, the joint entropy of the sources. The startling result by Slepian and Wolf is that for $R_u > H(U|V)$ and $R_v > H(V|U)$, any two rates, R_u and R_v, will allow perfect reconstruction at the decoder for some code as long as $R_u + R_v \geq H(u,v)$. Therefore there exists no rate loss for encoding the two sources separately versus encoding them together. The achievable rate region for a Slepian-Wolf code is shown in Fig. 3-1.

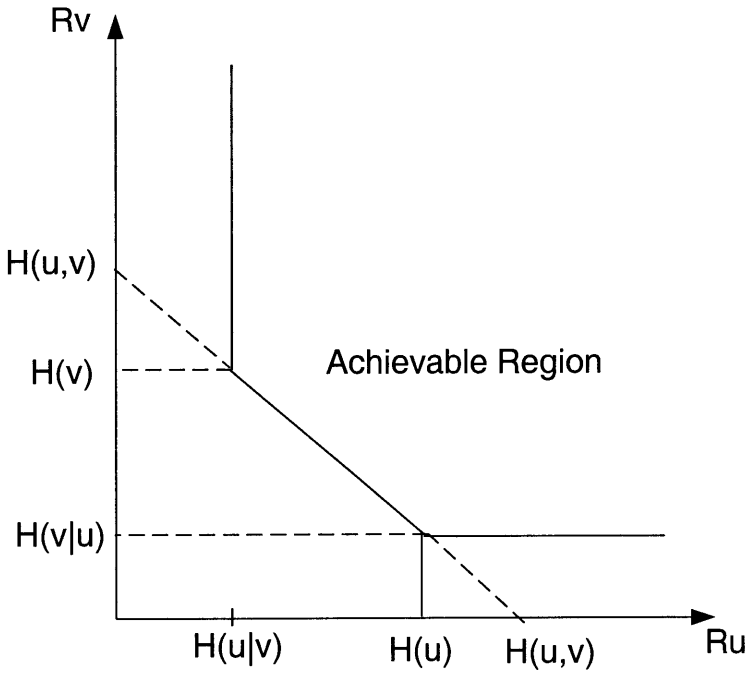


Figure 3-1: The Slepian-Wolf Achievable Rate Region

In the Wyner-Ziv setting, we have **u** equaling **x**, and **v** equaling the finely quantized output of the analog channel; the channel output **y** must be quantized (introducing a degree of suboptimality) because Slepian-Wolf coding only works with discrete-alphabet sources. By the problem construction, **v** is sent (via the analog channel observation) at rate $R_v = H(v)$. Thus the minimum rate required for perfect reconstruction of **u** is $H(u|v)$ (because $H(u|v) + H(v) = H(u, v)$. By application of Fano's inequality, it is shown in [90] that

$$R_{x|y}^{\mathrm{WZ}}(0) = H(x|y), \tag{3.6}$$

which is consistent with the Slepian-Wolf result.

The proof of the Slepian and Wolf result relies on the asymptotic behavior of random codes, which are not constructive, and hence indicate no meaningful implementation. There has been some recent work on practical Slepian-Wolf codes as found in [76, 64], but for arbitrary joint statistics on u and v no general method has been found that approaches the theoretical limit of performance. An insightful method is given by Wyner [87] for the case where u is Bernoulli($\frac{1}{2}$), and $v = u \oplus z$ is observed undistorted at the decoder, where z is Bernoulli(p), and $\oplus$ denotes modulo-2 addition. A

Bernoulli(p) source is sequence of independent binary random variables where p is the probability of a 1. We interpret $\boldsymbol{v}$ as the output of a binary symmetric channel (BSC) with crossover probability p. The quantity $H(\boldsymbol{u}|\boldsymbol{v})$ equals $h(p)$. The Slepian-Wolf code for this case employs a parity-check code, defined by an $N \times L$ parity check matrix $\mathbf{H}$, where $L = nh(p)$. The encoder transmits the syndrome $\mathbf{H}^{\dagger}\mathbf{u}$, thereby requiring the rate $h(p)$. Utilizing the asymptotic behavior of linear codes Wyner shows that $\boldsymbol{u}$ is constructed exactly. We see in Sec.3.4.2 that the nested linear code for Wyner-Ziv coding in the doubly binary symmetric case is a generalization of this Slepian-Wolf code for all reconstruction distortions $D \geq 0$.

3.2.3 Conditional rate-distortion function

For certain source coding applications the side information $\mathbf{y}$ may be known at the decoder *and* the encoder; this scenario is shown in to Fig. 1 with the switch S closed. Derived in [33, 7], the rate-distortion function for this scenario is given by the conditional rate-distortion function:

$$R_{\boldsymbol{x}|\boldsymbol{y}}(d) = \inf_{p(w|x) \in \mathcal{P}^{\mathrm{RD}}_{\boldsymbol{w}|\boldsymbol{y}}} I(\boldsymbol{x}; \boldsymbol{w}|\boldsymbol{y}). \tag{3.7}$$

where

$$\mathcal{P}^{\mathrm{RD}}_{\boldsymbol{w}|\boldsymbol{y}} = \{p_{\boldsymbol{w}|\boldsymbol{x},\boldsymbol{y}}(w|x,y) : E[D(\boldsymbol{x},\boldsymbol{w})] \leq d\}. \tag{3.8}$$

By problem construction it is clear that

$$R_{\boldsymbol{x}|\boldsymbol{y}}(d) \leq R^{\mathrm{WZ}}_{\boldsymbol{x}|\boldsymbol{y}}(d) \tag{3.9}$$

The necessary and sufficient conditions under which (3.9) holds with equality are derived in [90]; we outline the derivation here for comparison to information embedding. We let $\boldsymbol{u}$ satisfy (3.2) and let $\boldsymbol{w} = f(\boldsymbol{u}, \boldsymbol{y})$, where f satisfies (3.3). By application of the data processing inequality, it is easily shown that

$$I(\boldsymbol{x}; \boldsymbol{w}|\boldsymbol{y}) \leq I(\boldsymbol{x}; \boldsymbol{u}|\boldsymbol{y}), \tag{3.10}$$

Equality holds in (3.10) if and only if

$$I(\boldsymbol{x}; \boldsymbol{u}|\boldsymbol{w}, \boldsymbol{y}) = 0. \tag{3.11}$$

Thus by (3.5), (3.7), and (3.10), we have that (3.9) holds with equality if and only if the minimizing distribution for y, x, u, w in (3.1) also minimizes the argument on the right hand side of (3.7) and, in addition, the condition in (3.11) holds for the minimizing distribution.

3.3 Quadratic Gaussian case

Using the general expressions for rate-distortion function, in this section we specify the results to the case of Gaussian source, memoryless Gaussian channel, and quadratic distortion metric.

3.3.1 Wyner-Ziv rate-distortion function

For jointly Gaussian $\mathbf{x}$ and $\mathbf{y}$ whose elements are all drawn independently from the Gaussian density $f_{xy}(x, y) \sim \mathcal{N}(0, \Lambda_{xy})$, the Wyner-Ziv rate distortion function is [89]

$$R_{y|x}^{\mathrm{WZ}}(d) = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma_{y|x}^2}{d}\right) & 0 \le d < {\sigma_{x|y}}^2 \\ 0, & d \ge c{\sigma_{x|y}}^2, \end{cases} \tag{3.12}$$

where $\sigma_{y|x}^2$ is the linear least squared error variance.

To obtain the equivalence in (3.12), we first establish notation. Without loss of generality we write

$$y = \beta(x + n), \tag{3.13}$$

where $\beta > 0$, and $n \sim \mathcal{N}(0, \sigma_n^2)$ is independent of y. Defining

$$c = \sigma_y^2/(\sigma_x^2 + \sigma_n^2), \tag{3.14}$$

the minimum mean-squared estimate for x from y is

$$\hat{x}_{\mathrm{MMSE}} = E[x|y = y] = \rho x, \tag{3.15}$$

where

$$\rho = c/\beta \tag{3.16}$$

is the correlation coefficient. The corresponding error variance is

$$\sigma^2_{x|y} = c\sigma^2_n. \tag{3.17}$$

Wyner derived (3.12) by first showing that the conditional rate-distortion function (with y at the encoder) is

$$R_{x|y}(d) = \begin{cases} \frac{1}{2}\log c\sigma^2_n/d, & 0 \le d < c\sigma^2_n \\ 0, & d \ge c\sigma^2_n, \end{cases} \tag{3.18}$$

He then proceeds to show that with no side information at the encoder, there is a test channel that achieves the same rate-distortion function as (3.18). Because $R_{x|y}(d) \le R^{\mathrm{WZ}}_{x|y}(d)$, (3.18) is also the Wyner-Ziv rate distortion function.

In Wyner's formulation, the test channel encoder simply assigns the auxiliary random variable u to be a linear combination of the source and an independent zero-mean Gaussian variable: $u = \alpha x + e$, where $\alpha = (\sigma^2_n c - d)/\sigma^2_n c$, and $e \sim \mathcal{N}(0, \alpha d)$. The test channel decoder function is also a linear function: $w = f(u, y) = u + (c/\beta)(1-\alpha)y$. We easily confirm that the distortion is d, and $I(x; u|y)$ equals $R_{x|y}(d)$ in (3.18). Thus $R_{x|y}(d) = R^{\mathrm{IE}}_{x|y}(d)$, implying (3.12). For the special case where $\beta = 1$ (additive noise channel) and SNR$\rightarrow \infty$, the decoder function is

$$w = u + (1-\alpha)y. \tag{3.19}$$

We will see in Chap. 6 that this special case test channel decoding function is the same as the information embedding encoding function for the Gaussian case.

3.3.2 Coding a sequence of independent Gaussians

In this section we treat the case in which $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, distributed as $f_{\mathbf{xy}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} f_{x_i y_i}(x_i, y_i)$. The pairs (x_i, y_i) and (x_j, y_j) are independent, but not necessarily identically distributed. We determine the rate-distortion function for this case (assuming a quadratic distortion metric), which lends itself nicely to the solution for the rate-distortion function for a stationary Gaussian source process $x(t)$ with the jointly stationary Gaussian process $y(t)$ known at the decoder. Our solution closely parallels [23, pp. 348-349] which determines the rate-distortion function of a sequence of correlated Gaussian source variables (with no analog side information at the decoder).

We use the notation $\mathbf{u}_i^j$ to denote the vector $[u_i\ u_{i+1}...u_j]^T$, a subvector of the vector $\mathbf{u}$; the absence of a subscript implies $i = 1$. We lower bound $I(\mathbf{x};\mathbf{u}|\mathbf{y})$, and show the achievability of the bound:

$$\begin{aligned}
I(\mathbf{x};\mathbf{u}|\mathbf{y}) &= h(\mathbf{x}|\mathbf{y}) - h(\mathbf{x}|\mathbf{u},\mathbf{y}) && (3.20)\\
&= \sum_{i=1}^{N} h(x_i|y_i) - \sum_{i=1}^{N} h(x_i|x^{i-1},y,u) && (3.21)\\
&\geq \sum_{i=1}^{N} h(x_i|y_i) - \sum_{i=1}^{N} h(x_i|u_i,y_i) && (3.22)\\
&= \sum_{i=1}^{N} I(x_i;u_i|y_i) && (3.23)\\
&\geq \sum_{i=1}^{N} R(d_i) && (3.24)\\
&= \sum_{i=1}^{N} \max\left(\frac{1}{2}\log\frac{\sigma^2_{x_i|y_i}}{d_i}, 0\right), && (3.25)
\end{aligned}$$

where we have defined $d_i = E[(x_i - f_i(u_i,y_i))^2]$ for some decoding function $f_i()$. Equality is met in (3.22) if we have $f_{\mathbf{x}|\mathbf{uy}}(\mathbf{x}|\mathbf{u},\mathbf{y}) = \prod_{i=1}^{n} f_{x|uy}(x_i|u_i,y_i)$. It is straightforward to confirm that this condition is met if we let

$$f_{\mathbf{u}|\mathbf{x}}(\mathbf{u}|\mathbf{x}) = \prod_{i=1}^{n} f_{u|x}(u_i|x_i). \tag{3.26}$$

As $\mathbf{u}$ is simply an auxiliary random variable determined at the encoder, $f_{\mathbf{u}|\mathbf{x}}(\mathbf{u}|\mathbf{x})$ can be set as desired. Equality is met in (3.24) with optimal selection of $f_{u|x}(u_i|x_i)$ and decoder functions $f_i()$, as per (3.1).

The rate-distortion function is thus given by

$$R^{\mathrm{WZ}}_{x|y} = \min_{\{d_i\}:\sum d_i = d} \sum_{i=1}^{N} \max\left(\frac{1}{2}\log\frac{\sigma^2_{x_i|y_i}}{d_i}, 0\right), \tag{3.27}$$

which can be solved using a simple functional with Lagrange multipliers identically to the case with no analog side information [23, pg. 348] to yield

$$R^{\mathrm{WZ}}_{x|y}(d) = \sum_{i=1}^{n} \frac{1}{2}\log\frac{\sigma^2_{x_i|y_i}}{d_i}, \tag{3.28}$$

where

$$d_i = \begin{cases} \lambda, & \text{if } \lambda < \sigma^2_{x_i|y_i}, \\ \sigma_i^2 & \text{if } \lambda \geq \sigma^2_{x_i|y_i}, \end{cases} \tag{3.29}$$

and λ is chosen so that $\sum_{i=1}^{n} d_i = d$.

From (3.28) we can interpret a rate-distortion achieving code for the case of independent Gaussian pairs, as one which performs inverse-waterpouring on the variances $\sigma^2_{x_i|y_i}$. This result leads us to believe that an optimum decoder first performs MMSE estimation of x_i from y_i and then inverse-waterpours bits on the remaining MMSE error, giving only bits to samples with error variance greater than some λ. The most bits go to the sample with the largest mean-squared estimation error.

If we consider the stationary Gaussian processes $x(t)$ and $y(t)$ in the limit of long observation times, the Fourier basis is the limiting Karhunen-Loeve basis for both processes [86]. In a similar manner, projection onto the Fourier basis removes crosscorrelation between coefficients corresponding to different basis elements. Thus the $x(t)$ and $y(t)$ projected onto Fourier basis function elements satisfy our assumptions. We define the MMSE estimation error $e(t) = x(t) - \hat{x}_{\text{MMSE}}(t)$. The rate-distortion function is thus achieved by inverse-waterpouring on frequency components of the function

$$E(e^{j\omega}) = S_{xx}(e^{j\omega}) - \frac{S_{xy}(e^{j\omega})}{S_{yy}(e^{j\omega})} \tag{3.30}$$

$$= E[(X(e^{j\omega}) - X(e^{j\omega}))^2], \tag{3.31}$$

which is the Wiener filter error as a function of frequency. Of course, for a real-world implementation we consider $x(t)$ and $y(t)$ observed only on a finite interval $[T_0,\ T_0+T]$, where T is large compared to the coherence time of the processes, and inverse-waterpour on a only finite number of frequencies.

3.3.3 Geometric interpretation

The rate-distortion function for the Gaussian case has a geometric interpretation as a form of sphere covering that indicates the nature of codes that achieve the rate-distortion limit for this case. Given a side information vector $\mathbf{y}$ of length N at the decoder, a minimum mean-squared error estimate of the source is $\hat{\mathbf{x}} = \rho\mathbf{y}$, where ρ is defined in (3.16). The remaining mean-squared error about the

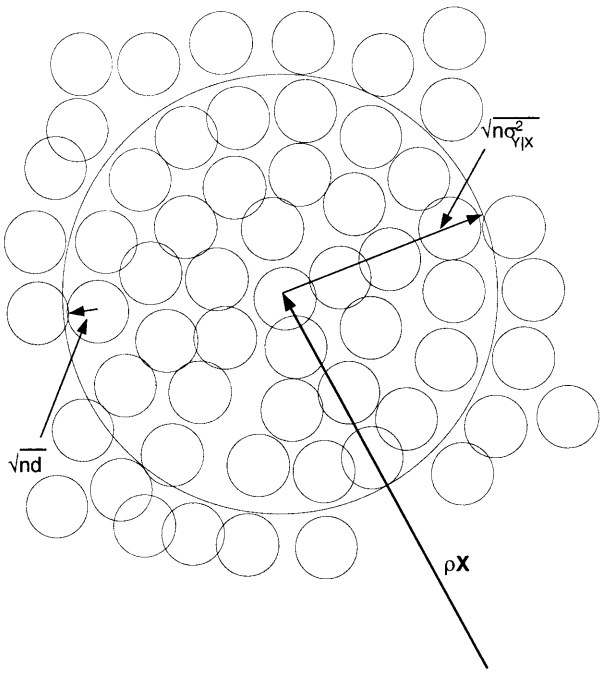


Figure 3-2: Sphere-covering for Wyner-Ziv rate-distortion coding in the quadratic Gaussian case. The vector $\rho\mathbf{y}$ is estimate of the source from the channel output.

estimate is zero-mean Gaussian with variance $\sigma^2_{x|y}$, implying that the source must lie in a sphere $S_{x|y}$ of radius $\sqrt{N\sigma^2_{x|y}}$ about $\rho\mathbf{y}$. A Wyner-Ziv codebook for a distortion d will contain $2^{NR(d)}$ code vectors in $\mathcal{R}^N$ such that most source sequences of length N lying in $S_{x|y}$ are within a distance $\sqrt{Nd}$ of a codeword. Rate-distortion coding for the Gaussian case, therefore, amounts to covering the sphere $S_{x|y}$ with smaller spheres of radius $\sqrt{Nd}$, which we illustrate in Fig. 3-2. Clearly the number of codewords is lower bounded by the ratio of the volumes of the large to the small spheres:

$$2^{NR(d)} \geq \left(\frac{\sigma^2_{x|y}}{d}\right)^{N/2}, \tag{3.32}$$

and this lower bound is met by a code that achieves the rate-distortion bound given by (3.12).

3.3.4 Lattice code for Wyner-Ziv encoding

In [95] Zamir and Shamai construct codes using a pair of nested lattices that achieve the Wyner-Ziv rate-distortion bound for the quadratic Gaussian case, where $\mathbf{y} = \beta\mathbf{x} + \mathbf{n}$ for $n \sim \mathcal{N}(\mathbf{0}, \sigma^2_n\mathbf{I})$ uncorrelated with $\mathbf{x}$, in the limit of high SNR. Using the same basic construction, we offer a

straightforward generalization which yields a code that achieves the bound for all SNRs. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be a pair of unbounded lattices in $\mathcal{R}^n$ such that $\mathcal{L}_2 \subset \mathcal{L}_1$, *i.e.,* $\mathcal{L}_2$ is nested in lattice $\mathcal{L}_1$. An N-dimensional lattice is defined by $\mathcal{L} = \{\mathbf{l}_i\}$ where

$$\mathbf{l}_i \in \mathcal{R}^n, \;\; \mathbf{l}_0 = \mathbf{0}, \;\; \mathbf{l}_i + \mathbf{l}_j \in \mathcal{L} \quad \forall i, j. \tag{3.33}$$

The lattice decoding functions for each lattice are nearest neighbor quantizing functions $Q_i : \mathcal{R}^n \rightarrow \mathcal{L}_i$, defined by:

$$Q_i(x) \triangleq \arg\min_{\mathbf{l} \in \mathcal{L}} \|\mathbf{x} - \mathbf{l}\|, \;\; i = 1, 2. \tag{3.34}$$

The basic Voronoi region for each lattice $\mathcal{L}_i$, which specifies the shape of the nearest-neighbor decoding region, is $\mathcal{V}_i = \{x : Q(x) = 0\}$. A Voronoi region has several significant quantities: volume V_i, second moment σ_i^2, and normalized second moment G_i given respectively by

$$V_i = \int_{\mathcal{V}_i} \mathrm{d}x \tag{3.35}$$

$$\sigma_i^2 = \frac{1}{NV_i} \int_{\mathcal{V}_i} \|x\|^2 \mathrm{d}x \tag{3.36}$$

$$G_i = \frac{\sigma_i^2}{V_i^{2/N}}. \tag{3.37}$$

We use the same terminology used in [95] regarding lattice quantization and nested lattices. For some $l \in \mathcal{L}_1$, the quantity $l - Q_2(l)$ is termed the "coset shift of l with respect to $\mathcal{L}_2$." The function $k(l) : \mathcal{L}_1 \rightarrow \{1, ..., 2^{NR}\}$ indexes the coset shifts, where R is the rate of the code and $2^{NR} = V_1/V_2$. The inverse function of $k(\cdot)$ is $g(k(l)) = l - Q_2(l)$. A random dither vector $\mathbf{z}_1$, known at the encoder and decoder, is defined to be independent of $(\mathbf{x}, \mathbf{n})$ and uniformly distributed over $\mathcal{V}_1$. Its variance is therefore equal to σ_1^2 and by the results in [94], $\mathbf{z}_1$ is a "Gaussian-like" vector for large N.

In order to achieve the Wyner-Ziv rate-distortion bound, the lattices $\mathcal{L}_1$ and $\mathcal{L}_2$ must satisfy the following conditions for all $\epsilon > 0$:

a) $\mathcal{L}_2 \subset \mathcal{L}_1$

b) $\sigma_1^2 = \frac{d\sigma_{x|y}^2}{\sigma_{x|y}^2 - d}$

c) $\sigma_2^2 \leq \sigma_1^2 + \sigma_{x|y}^2 + \epsilon$

d) $\mathrm{P}\{Q_2(\mathbf{g}+\mathbf{v}) \neq \mathbf{g}\} < \epsilon \quad \forall \mathbf{g} \in \mathcal{L}_2$

where $\mathbf{v}$ is a Gaussian of variance $\sigma^2_{x|y} + \sigma_1^2$, respectively

e) $\log(2\pi e G_i) < \epsilon \quad i = 1, 2.$

Note that properties d) and e) rely on both lattices $\mathcal{L}_1$ and $\mathcal{L}_2$ having good sphere-packing properties [21]. By random coding arguments, such lattices likely exist, but the specific lattices satisfying these properties have yet been determined. For a good lattice $\mathcal{L}_2$, property d) requires that the variance of $\mathbf{v}$ be slightly less than σ_2^2. Property e) is satisfied for any good lattice, because in the limit of large N the normalized second moment of a sequence of good lattices approaches $\frac{1}{2\pi e}$, as shown by Poltyrev as referenced in [94].

Properties b), c), and e) prescribe the rate of the code to be within at most $1/n$ bit from

$$R = \frac{1}{n}\log\left(\frac{V_1}{V_2}\right) = \frac{1}{2}\log\left(\frac{\sigma_2^2 G_1}{\sigma_1^2 G_2}\right) \tag{3.38}$$

$$\leq \frac{1}{2}\log\left(\frac{\sigma^2_{x|y}}{d}\right) + O(\epsilon), \tag{3.39}$$

which equals the rate distortion function for all SNRs. We have altered Properties b) and c) slightly from [95] in order to achieve the rate-distortion function for all SNRs. We will see how the decoder is modified to accommodate this change in the properties.

The encoder simply calculates $\mathbf{x}_q = Q_1(\mathbf{x} + \mathbf{z}_1)$ and $\mathbf{l} = \mathbf{x}_q - Q_2(\mathbf{x}_q)$, and transmits the index $k(\mathbf{l})$. The decoder observes $\mathbf{y} = \beta(\mathbf{x} + \mathbf{n})$, $\mathbf{z}_1$, and $k(\mathbf{l})$, and calculates the coset shift $\mathbf{s} = g(k(\mathbf{l}))$. The reconstructed source at the decoder has the following form:

$$\mathbf{w} = \hat{\mathbf{x}} = \frac{a}{\beta}\mathbf{y} + b\{Q_2(\rho\mathbf{y} + \mathbf{z}_1 - \mathbf{s}) - \mathbf{z}_1 + \mathbf{s}\} \tag{3.40}$$

where $\mathbf{e_q}$ is independent of $\mathbf{x}$ and $\mathbf{n}$ and is distributed as $\mathbf{z}_1$, and the correlation coefficient ρ is given by (3.16). Note that the only change in the decoder from [95] is the inclusion of the estimator gain ρ prior to quantization by $Q_2(\cdot)$. Intuitively, prior to quantization at the decoder, we are simply improving our estimate of $\mathbf{x}$, by performing MMSE estimation. The intuition also follows from our derivation of the MMSE systematic decoder in Chap. 2. We argued that for large N the centroid decision rule is closely approximated by mapping $\rho\mathbf{y}$ to the nearest lattice point. Using the equality

$\mathbf{s} = \mathbf{x}_q - Q_1(\mathbf{x} + \mathbf{z}_1)$, the argument of $Q_2(\cdot)$ in (3.40), denoted by $\mathbf{t}$ is

$$\begin{aligned} \mathbf{t} &= c(\mathbf{x} + \mathbf{n}) + \mathbf{z}_1 - (\mathbf{x}_q - Q_2(\mathbf{x}_q)) && (3.41) \\ &= \mathbf{g} + \mathbf{e}_q + (c-1)\mathbf{x} + \rho\mathbf{n}, && (3.42) \end{aligned}$$

where c is given by (3.14); $\mathbf{e}_q = \mathbf{x} + \mathbf{z}_1 - Q_1(\mathbf{x} + \mathbf{z}_1)$ is independent of $\mathbf{y}$ and distributed as $\mathbf{z}_1$ (by the properties of subtractive dithered quantization [94]); and $\mathbf{g} \in \mathcal{L}_2$ is defined by $\mathbf{g} = Q_2(\mathbf{x}_q)$. The last three terms on the right hand side of (3.42) are "noise terms", and the variance their sum is $\sigma_1^2 + \sigma_{x|y}^2$. By Property d), $Q_2(\mathbf{t}) = \mathbf{g}$ with probability greater than $1 - \epsilon$, which implies that with arbitrarily high probability we have

$$\hat{\mathbf{x}} = a(\mathbf{x} + \mathbf{n}) + b(\mathbf{x} - \mathbf{e}_q). \tag{3.43}$$

Recalling that $\frac{1}{N}E[||\mathbf{e}_q||^2] = \frac{1}{N}E[||\mathbf{z}_1||^2] = \sigma_1^2$, we select a and b to minimize mean-squared error. Thus, Property b) yields

$$\frac{1}{N}E[||\hat{\mathbf{x}} - \mathbf{x}||^2] = d + O(\epsilon), \tag{3.44}$$

which implies that we have met the rate-distortion bound.

3.4 Binary symmetric channel and source with Hamming distortion metric

In this section we consider the scenario where the signals being communicated and the channels over which they are being communicated are binary symmetric. The source $\mathbf{x}$ is Bernoulli($\frac{1}{2}$), and the channel is a binary symmetric channel with crossover probability p. Clearly, the output of the channel is also Bernoulli($\frac{1}{2}$), which gives rise to the term *doubly binary symmetric case* [90]. The distortion metric $D(\cdot,\cdot)$ is bit error rate, or Hamming distortion metric.

3.4.1 Rate-distortion functions

The Wyner-Ziv rate-distortion function for this scenario is determined in [90] to be the lower convex envelope of the function $h(p * d) - h(d)$ and the point $(R, d) = (0, p)$, where $h(x) = -x\log(x) - (1 -$

$x)\log(1-x)$ and $p * d$ denotes binary convolution: $p * d = p(1-d) + d(1-p)$. Written in closed form, we have

$$R_{x|y}^{\mathrm{WZ}}(d) = \begin{cases} g(d), & 0 \le d \le d_c \\ g(d_c)\left(1 - \frac{(d-d_c)}{(p-d_c)}\right), & d_c < d \le p \end{cases} \tag{3.45}$$

$$g(d) = \begin{cases} h(p * d) - h(d), & 0 \le d < p \\ 0, & d = p, \end{cases} \tag{3.46}$$

where d_c is the solution to the equation

$$\frac{g(d_c)}{d_c - p} = g'(d_c). \tag{3.47}$$

As a comparison we show the conditional rate-distortion function ($\mathbf{y}$ known at the encoder and decoder) for the binary symmetric case [7]:

$$R_{x|y}(d) = \begin{cases} h(p) - h(d), & 0 \le d \le p \\ 0, & d \ge p \end{cases} \tag{3.48}$$

Fig. 3-3 shows an example of $R_{x|y}^{\mathrm{WZ}}(d)$ and $R_{x|y}(d)$ for channel transition probability $p = 0.25$.

3.4.2 Nested linear codes that achieve Wyner-Ziv rate-distortion function

In [66], Shamai *et al* construct codes that achieve the Wyner-Ziv rate-distortion limit for the binary symmetric case. We present the construction here in order to explicitly show the duality with capacity-achieving information embedding codes in Chap. 6. The codes operate in the bit error-rate regime $0 \le d \le d_c$; time sharing with no coding can be used to achieve all other operating points on the rate-distortion curve.

The codes are constructed from two linear codes $\mathcal{C}_1$ and $\mathcal{C}_2 \subset \mathcal{C}_1$, both of which have codewords of length N. As every codeword of $\mathbf{c} \in \mathcal{C}_2$ is a codeword of $\mathcal{C}_1$, $\mathcal{C}_2$ is referred to as "nested" in $\mathcal{C}_1$. It is assumed that both codes are good source/channel codes. Associated with $\mathcal{C}_1$ and $\mathcal{C}_2$ is a parity check matrix, respectively $\mathbf{H}_1$ of dimension $m_1 \times n$ and $\mathbf{H}_2$ of dimension $m_2 \times n$. The values m_1 and m_2 satisfy $\frac{m_1}{n} = h(d)$ and $\frac{m_2}{n} = h(p * d)$. A parity check matrix H defines a code $\mathcal{C}$ by the equation

$$\mathbf{Hc} = \mathbf{0} \quad \forall \mathbf{c} \in \mathcal{C}, \tag{3.49}$$

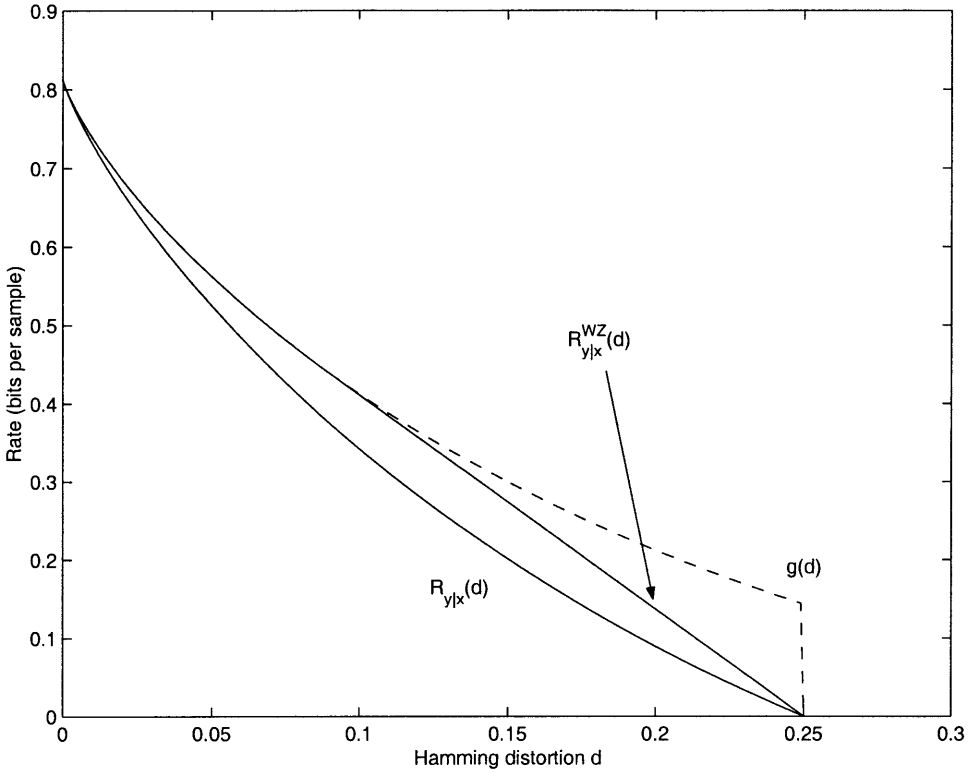


Figure 3-3: The binary symmetric case where the channel transition probability is $p = 0.25$. The dashed line is the function $g(d) = h(p * d) - h(d)$. Assuming x at the decoder, the solid lines are $R_{x|y}(d)$ and $R^{\mathrm{WZ}}_{x|y}(d)$, the rate distortion functions respectively with and without x at the encoder.

where the superscript T denotes the transpose operation. As $\mathcal{C}_2$ is a subcode of $\mathcal{C}_1$ we can write $\mathbf{H}_2$ as

$$\mathbf{H}_2 = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_a \end{bmatrix} \tag{3.50}$$

Every codeword $\mathbf{c} \in \mathcal{C}_1$ satisfies

$$\mathbf{H}_1\mathbf{c} = \mathbf{0}. \tag{3.51}$$

If in addition $\mathbf{H}_a\mathbf{c} = \mathbf{0}$, then $\mathbf{c}$ is also a codeword of $\mathcal{C}_2$. Assume $\mathbf{c} \in \mathcal{C}_2$. Let the channel output be given by $\mathbf{y} = \mathbf{c} \oplus \mathbf{u}$, where $\mathbf{u}$ is the binary error sequence induced by a channel and $\oplus$ denotes modulo-2 addition. The decoder for $\mathcal{C}_1$ and $\mathcal{C}_2$, first calculates a syndrome $\mathbf{H}_1\mathbf{y} = \mathbf{H}_1\mathbf{u}$ and $\mathbf{H}_2\mathbf{y} = \mathbf{H}_2\mathbf{u}$ respectively. To every syndrome value there corresponds an estimate of the error sequence $\mathbf{u}$, given by the decoding functions $f_1(\mathbf{H}_1\mathbf{u})$ for $\mathcal{C}_1$ and $f_2(\mathbf{H}_2\mathbf{u})$ for $\mathcal{C}_2$. If the estimate $\hat{\mathbf{u}}$ equals $\mathbf{u}$, then the codeword can be recovered by $\mathbf{c} = \mathbf{y} \oplus \hat{\mathbf{u}}$. Using optimal linear codes and decoding functions, we

can be assured for large N that

$$f_1(\mathbf{H}_1\mathbf{u}_1) = \mathbf{u}_1 \tag{3.52}$$

$$f_2(\mathbf{H}_2\mathbf{u}_2) = \mathbf{u}_2 \tag{3.53}$$

for most realizations $\mathbf{u}_1$ of a Bernoulli(d) process and most realizations $\mathbf{u}_2$ of a Bernoulli($p * d$) process.

The Wyner-Ziv encoder. We encode the host $\mathbf{x}$ in two steps:

1) Select the codeword $\mathbf{x}_q \in \mathcal{C}_1$ that is closest in Hamming distance to $\mathbf{x}$. Let $\mathbf{e}_q = \mathbf{x} \oplus \mathbf{x}_q$.

2) Transmit the length $m_2 - m_1$ vector $\mathbf{H}_a\mathbf{x}_q$.

The rate of the encoder is thus

$$\frac{m_2 - m_1}{n} = h(p * d) - h(d) = R_{x|y}^{\mathrm{WZ}}(d). \tag{3.54}$$

The Wyner-Ziv Decoder. The decoder observes $\mathbf{H}_a\mathbf{x}_q$ and $\mathbf{y} = \mathbf{x} \oplus \mathbf{n}$, where $\mathbf{n}$ is Bernoulli(p). After constructing the vector,

$$\mathbf{H}_2\mathbf{x}_q = \begin{bmatrix} \mathbf{0} \\ \mathbf{H}_a\mathbf{x}_q \end{bmatrix}, \tag{3.55}$$

the decoder then calculates

$$\mathbf{w} = \mathbf{y} \oplus f_2(\mathbf{H}_2\mathbf{x}_q \oplus \mathbf{H}_2\mathbf{y}) = \mathbf{y} \oplus f_2(\mathbf{H}_2(\mathbf{e}_q \oplus \mathbf{n})). \tag{3.56}$$

For large N, $\mathbf{e}_q$ resembles a Bernoulli(d) process, implying that $\mathbf{e}_q \oplus \mathbf{n}$ resembles a Bernoulli($p * d$) process. By (3.53) Shamai *et al* argue that with high probability

$$\mathbf{w} = \mathbf{y} \oplus \mathbf{e}_q \oplus \mathbf{n} = \mathbf{x} \oplus \mathbf{e}_q = \mathbf{x}_q. \tag{3.57}$$

The reconstruction error $\mathbf{x} \oplus \mathbf{x}_q = \mathbf{e}_q$ has Hamming distortion d as desired.

Chapter 4

Systematic Quantization

Many signal processing source coding applications have constraints on two quantities: latency, too much of which is intolerable for applications like the coding of speech for two-way communications, and complexity, which is limited by processing capabilities. The previous chapter describes the theoretical lower bound on the rate of a source coding system required to achieve a prescribed fidelity. For most sources and analog channels of interest, these limits of performance are achievable only if we process a block of source samples and let the block length get arbitrarily large. Such a coding method can only exist in a theoretical realm, of course, as latency and complexity are potentially unbounded. In this chapter we consider the basic building block of latency- and complexity-constrained systematic source coding systems, the N-dimensional systematic quantizer. We develop the systematic scalar quantizer (SSQ) design algorithm ($N = 1$), and the systematic vector quantizer (SVQ) design algorithm for higher dimensions which employs Monte-Carlo methods. The SVQ and the SVQ design algorithm are highly complex. Due to this fact and the fact that the relatively low-complexity SSQ is a provably effective coding method, we focus most of our attention in this chapter on the SSQ. The SSQ and SVQ algorithms yield quantizers that achieve a local optimum (local minimum or saddle point) in the expected error surface. We focus on the case of minimizing mean-squared error. Reminiscent of the Lloyd-Max algorithm [47, 48] and the generalized Lloyd-Max algorithm [31], respectively, the SSQ and SVQ algorithms are iterative, involving two step main steps. Similar iterative design techniques are used for the scalar quantizer design in [27] and [77] for different problem constructions. The rate-distortion performance of the algorithms is analyzed for a Gaussian source and channel and mean-squared distortion constraint.

Allowing for the partial feedback of the channel output to the encoder, we develop a further generalization of the algorithm for the design of quantizers at the encoder and decoder that yield a local optimum in expected error. Using analytical techniques, in Sec. 4.2.5 we determine the gain from using feedback with scalar quantizers over not using feedback for the Gaussian case.

In general the output of the SSQ and SVQ design algorithms have no simple structure and could be difficult to implement. Observing the resultant structures from the quantizer design algorithms for the Gaussian case, however, we develop very low-complexity quantizer designs, called nested lattice scalar quantizers (NLSQs), using two nested lattices that closely approximate optimal quantizers. In fact these lattice designs are the low-dimensional equivalents of those used in Chap. 3 that are used to attain the rate-distortion function for the Gaussian case.

Ideally, we would like to be able to use a simple component, like an SSQ, and achieve a rate close to the rate-distortion bound. Processing only a small number of source samples at a time, however, is a suboptimal approach in terms of rate-distortion performance. In Sec. 4.3 we show that we can overcome this pitfall by post-processing the SSQ output. For $\mathbf{x}$ and $\mathbf{y}$ drawn iid from $f_{\mathbf{xy}}(x, y)$, we show that if, at the encoder, the output of an uniform scalar quantizer is coded by a Slepian-Wolf code (see Sec. 3.2.2), we achieve performance within .255 bits/sample of the rate-distortion bound. Thus, we show that scalar quantizers can potentially be an efficient low-complexity method for source coding.

4.1 Systematic Scalar Quantization

The challenge in scalar quantization, with or without side information, is to code a scalar random variable $\mathbf{x}$ with a constrained representation of K discrete levels. For a digital encoding, we usually have $K = 2^R$ where R is the number of bits per sample of the representation. In order to establish a point of reference, we first describe the operation of a standard scalar quantizer (SQ). A K-level SQ is function of a scalar random variable $\mathbf{x}$ that is represented by the composition of an encoder map $f : \mathcal{X} \to \{0, ..., K-1\}$, whose output is a codeword index, and a decoder map $g : \{0, ..., K-1\} \to \hat{\mathcal{X}}$, whose output is the lossy reconstruction $\hat{\mathbf{x}}$ of the $\mathbf{x}$. The signal $\hat{\mathbf{x}}$ is a rate-reduced representation (encoding) of $\mathbf{x}$. The encoder map partitions $\mathcal{X}$ into the set $\mathcal{A} = \{A_1, ..., A_K\}$, where we define $A_i = \{\mathbf{x} : f(\mathbf{x}) = i\}$. Thus the SQ is defined completely by the partition $\mathcal{A}$ and the codebook. For

the given metric $D(.,.)$, an optimal scalar quantizer design is one which minimizes $E[D(x,\hat{x})]$.

An SSQ is likewise represented by an encoder map, whose output is a codeword index, and decoder map, whose output is the lossy reconstruction $\hat{x}$ of x. The encoder map $f : \mathcal{X} \rightarrow \{0, ..., K-1\}$, in fact, has an identical structure to that of an SQ, in that it is defined by the partition $\mathcal{A} = \{A_1, ..., A_K\}$, where we define the decision regions $A_i = \{x : f(x) = i\}$. Although the encoder maps for the SQ and SSQ are conceptually similar, we will see that the exact partition for an optimal SSQ is significantly different than that for the corresponding optimal SQ without side information at the decoder. At the decoder we have the channel output $\mathbf{y}$ (which we allow to be an M-dimensional vector quantity) to aid in the estimation of x. We allow for $\mathbf{y}$ to be an M-dimensional vector quantity, but for the complexity of the SSQ design to be manageable, we must have M small ($\approx$ 1 to 2). In the case of large M the SVQ design algorithm of Sec. 4.5 should be used. The domain of the decoder map for the SSQ differs from that for the SQ. For the SSQ, the decoder map $g : \{0, ..., K-1\} \times \mathcal{Y}^n \rightarrow \hat{\mathcal{X}}$, uses the codeword index, which we denote by the random variable $k = f(x)$, and $\mathbf{y}$ to determine the estimate $\hat{x}$. Thus, in contrast to the SQ codebook, which corresponds indices k to constant reconstruction points $\hat{x}_k$, the SSQ codebook corresponds indices k to reconstruction functions, or estimators, $\hat{x}_k(\mathbf{y})$. In order to emphasize the interpretation of the codebook as an indexed set of estimators we use $\hat{x}_k(\mathbf{y})$ to denote $g(k, \mathbf{y})$. We define an optimal SSQ as one which minimizes $E[D(x, \hat{x}_k(\mathbf{y}))]$, where the expectation is taken over all x and $\mathbf{y}$.

4.1.1 Ad hoc approaches

Before approaching the formal SSQ design algorithms, we briefly describe a few obvious *ad hoc* approaches to scalar quantization with side information at the decoder to gain intuitive insight, and to later illustrate the advantages of our methods. One approach is to simply use the same encoder map as a standard SQ, say a uniform SQ or perhaps a Lloyd-Max SQ. At the decoder, we consider one of two methods of reconstruction. In the first case, we use the decoder map corresponding to the selected SQ. This method obviously gives us the same rate-distortion performance as if the analog information were not at the decoder, as it completely ignores $\mathbf{y}$ at the encoder and decoder. Considering a different approach for decoder operation, we can employ an optimal decoder rule, *e.g.* the conditional centroid rule for MMSE reconstruction, for the chosen encoder design. A simple example, illustrated in Fig. 4-1 for 2-bit quantization shows that this method still falls short.

Consider a squared distortion metric, a Gaussian source x and channel observation $y = x + n$, where n is a Gaussian random variable, independent of x, with variance σ_n^2. Because our encoder uses a standard encoder map function, the decision regions A_k, $k = 0, ..., K-1$, are simply intervals. We show the standard partition in Fig. 4-1(a). Assuming a very high signal-to-noise ratio (SNR), we have that $\sigma_{x|y}^2$ is small compared to the length of the smallest of these intervals. Thus, as shown in Fig. 4-1(b) for most observations y, $f_{x|y}(x|y)$ will be contained in a single interval A_k for some k. Assuming that the level k is transmitted down the digital channel, the centroid of $f_{x|y}(x|y)$ in A_k is very close to $E[x|y]$, which is the MMSE estimate given only the observation y. Thus we have gained nothing from the use of the digital information. An alternate partition $\mathcal{A}$ is shown in Fig. 4-1(c) in which a decision region A_i is not an interval, but rather the union of uniformly spaced intervals. Note that there exists granularity with which to improve the analog estimate with the digital information. We will discuss this type of quantizer, the NLSQ, in Sec. 4.2.

A second ad hoc approach to systematic scalar coding, which we refer to as *low-bits coding*, uses the bits transmitted down the digital channel to represent the least significant bits (LSBs) of the source. Assuming the scalar source sample has a dynamic range on the interval $[-P/2, P/2)$, it can be digitized using $C > R$ bits, with a granularity of $P2^{-C}$. The encoder takes the R LSBs and transmits them, thus achieving a rate of R bits per sample. The observation at the decoder y is usually a rough approximation of the source, say if the channel is additive noise. Based on statistics of the channel and source, the decoder can generate an estimate $\tilde{x}$ of the source. The decoder performs analog-to-digital (A/D) conversion on $\tilde{x}$ using the same C-bit digitization as used on the source. As the channel noise likely corrupts only the LSBs, the decoder replaces the LSBs of the digital representation to form the final estimate of the source. We analyze this method of coding in Sec. 4.6, and show that it is suboptimal compared to the low-complexity SSQs designed by our methods.

The low-bits coding method of systematic source-coding is the dual to the method of information embedding known as low-bits modulation [72, 6, 73]. For low-bits modulation, at the encoder, a host is quantized and information is embedded by replacing the least significant bit(s) of the digitized host with bits from an information sequence. The decoder reads the information bits from the LSBs of the composite signal, and, if necessary, applies error correction decoding to obtain the desired information. It is clear that the low-bits modulation encoder (respectively, decoder) operates as the

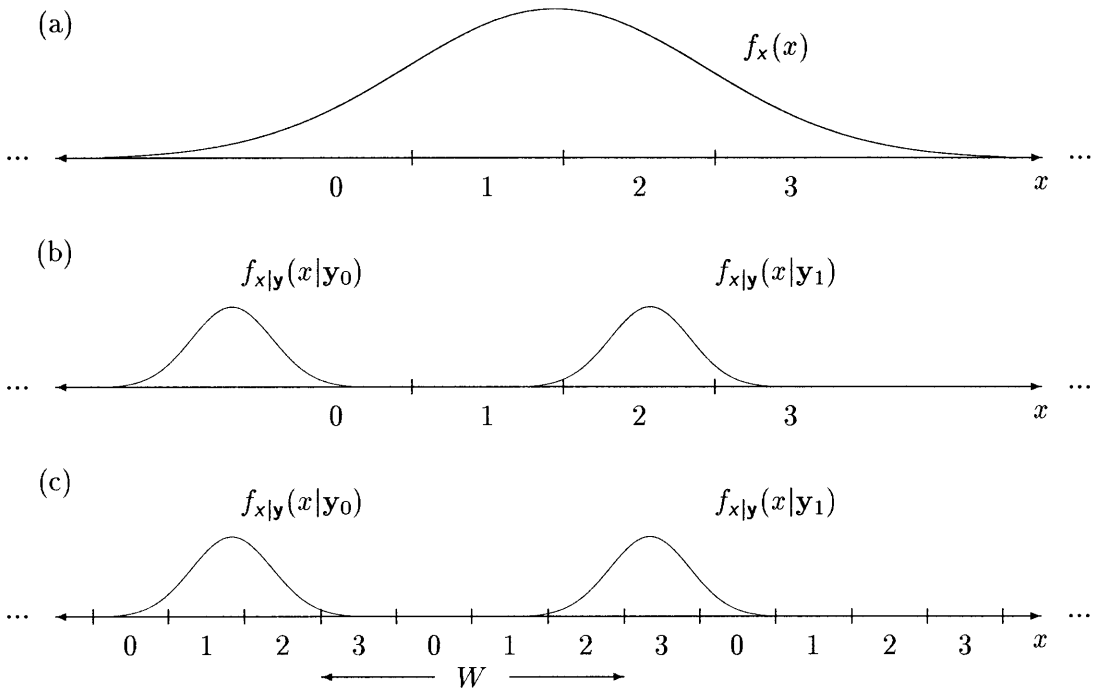


Figure 4-1: Quantizer decision regions (4 level quantizer). The domain of x (the real line) is divided into cells. Each cell labeled with an i, $i = 0, ..., 3$, belongs to an encoder decision region A_i. (a) Source density $f_x(x)$, with the optimal partition for standard quantization (no side information). (b) Aposteriori densities for two channel realization $\mathbf{y} = \mathbf{y}_0$ and $\mathbf{y} = \mathbf{y}_1$. Standard partition has insufficient granularity to improve the analog estimate. (c) Alternate partition offers needed granularity to improve decoder estimate.

low-bits coding decoder (respectively, encoder). Just as we prove in Sec. 4.6 that low-bits coding is a suboptimal approach to systematic coding of scalars, it is shown in [15] that low-bits modulation is a suboptimal approach to information embedding; the optimal approach is QIM.

4.1.2 Necessary conditions for optimality

In this section, we derive two conditions that any optimal SSQ must satisfy, one for the encoder partition $\mathcal{A}$ and one for the decoder map g, both of which are not satisfied by the above ad hoc techniques. Following the derivations of the necessary conditions, we develop an SSQ design algorithm in Sec. 4.1.5 whose output does satisfy the conditions. Recall that an optimal SSQ is one which minimizes the objective function

$$J = E[D(x, \hat{x}_{f(x)}(\mathbf{y}))]., \tag{4.1}$$

where the expectation is taken over all x and $\mathbf{y}$.

4.1.3 Optimal encoder

For a fixed decoder function g, we derive the *necessary condition at the encoder*, which specifies the partition $\mathcal{A}$. Defining

$$J_x = E_{\mathbf{y}}[D(x, \hat{x}_k(\mathbf{y}))|x], \tag{4.2}$$

we use iterated expectations and rewrite (4.1) as

$$J = E[J_x]. \tag{4.3}$$

Clearly, the objective function J is minimized only if J_x is minimized for all x. Assume the decoder is fixed, defining the set of reconstruction functions $\hat{x}_k(\mathbf{y})$, which are indexed by $k \in \{0, ..., K-1\}$; each k corresponds to a partition set A_k. For a given $x = x$, the optimal encoder will minimize J_x by selecting the best function $\hat{x}_k(\mathbf{y})$ to be used at the decoder, hence implying the decision region A_k to which x belongs. Thus for a fixed decoder g, a necessary condition at the encoder for an

optimal systematic quantizer is

$$f(x) = \arg\min_k E_{\mathbf{y}}[D(x, \hat{x}_k(\mathbf{y}))|x = x], \tag{4.4}$$

which simply states that encoder selects the index k (which corresponds to the reconstruction function $\hat{x}_k(\mathbf{y})$) that, averaged over all $\mathbf{y}$, is the minimum distance to the input x. In terms of the optimal encoder decision regions, (4.4) is rewritten as

$$A_k = \{x : E_{\mathbf{y}}[D(x, \hat{x}_k(\mathbf{y}))|x = x] \leq E_{\mathbf{y}}[D(x, \hat{x}_j(\mathbf{y}))|x = x] \ \forall j \neq k\}. \tag{4.5}$$

Unfortunately, (4.5) cannot be simplified, even if D is a squared distortion metric. We see this by expanding (4.4) for a squared distortion metric:

$$\begin{aligned} A_k &= \{x : E[(x - \hat{x}_k(\mathbf{y}))^2|x = x] \leq E[(x - \hat{x}_j(\mathbf{y}))^2|x = x] \ \forall j \neq k\} & (4.6) \\ &= \{x : 2x(\int_{\mathbf{y}} (\hat{x}_j(\mathbf{y}) - \hat{x}_k(\mathbf{y}))f_{\mathbf{y}|x}(\mathbf{y}|x)d\mathbf{y} \leq \int_{\mathbf{y}} (\hat{x}_j^2(\mathbf{y}) - \hat{x}_k^2(\mathbf{y}))f_{\mathbf{y}|x}(\mathbf{y}|x)d\mathbf{y}. \ \forall j \neq k\} & (4.7) \end{aligned}$$

The main difficulty preventing the simplification of (4.7) is that the integrals depend on x, specifically in the term $f_{\mathbf{y}|x}(\mathbf{y}|x)$. Similar expressions to (4.7) are given in [77], for the problem of multiple description scalar quantizer design, and [27], for the problem of quantizer design for combined source-channel coding, without this dependence on x. Without the dependence of the expectation on x, it is clear that the partition $\mathcal{A}$ is simply a collection of intervals. For the case of systematic quantization, however, $\mathcal{A}$ is not so easily described, because as we shall see by example, in some cases a decision region A_i is the (potentially infinite) union of disjoint intervals. The intractable nature of (4.7) adds considerable complexity to the SSQ design algorithm, but this is to be expected because of the complex nature of the decision regions.

4.1.4 Optimal decoder

The necessary condition at the decoder is also derived by rewriting (4.1) through the use of iterated expectation, but in the reverse order from (4.3). Defining

$$J_{\mathbf{y}} = E_x[D(x, \hat{x}_{f(x)}(\mathbf{y}))|\mathbf{y}],$$

$$= \sum_{i=1}^{K} \mathrm{P}(x \in A_i) E[D(x, \hat{x}_i(\mathbf{y}))|\mathbf{y},\ x \in A_i] \tag{4.8}$$

$$= \sum_{i=1}^{K} \int_{x \in A_i} D(x, \hat{x}_i(\mathbf{y})) f_{x|\mathbf{y}}(x|\mathbf{y}) dx$$

we have

$$J = E_{\mathbf{y}}[J_{\mathbf{y}}]. \tag{4.9}$$

Clearly, J is minimized only if $J_{\mathbf{y}}$ is minimized for each each $\mathbf{y}$, and $J_{\mathbf{y}}$ is minimized only if $E[D(x, \hat{x}_k(\mathbf{y}))|\mathbf{y}, x \in A_k]$ is minimized for all $\mathbf{y}$ and all $k \in \{0, ..., K-1\}$. Thus, assuming a fixed encoder partition $\mathcal{A}$, for a given $\mathbf{y}$ and k, a necessary condition at the decoder is that $\hat{x}_k(\mathbf{y})$ be chosen to minimize $E[D(x, \hat{x}_k(\mathbf{y}))|\mathbf{y}, x \in A_k]$. What we have at the decoder, then, is the straightforward generalized conditional centroid rule (2.7) for minimum mean distortion estimation:

$$\hat{x}_k(\mathbf{y}) = g(k, \mathbf{y}) = \arg\min_{\hat{x}} E[D(x, \hat{x})|\mathbf{y}, x \in A_k]. \tag{4.10}$$

For a squared error criterion we have the estimate

$$\hat{x}_k(\mathbf{y}) = g(k, \mathbf{y}) \tag{4.11}$$

$$= E[x|\mathbf{y}, x \in A_k] \tag{4.12}$$

$$= \frac{1}{\mathrm{P}(x \in A_k|\mathbf{y})} \int_{x \in A_k} x f_{x|\mathbf{y}}(x|\mathbf{y}) dx, \tag{4.13}$$

which is simply the centroid of $f_{x|\mathbf{y}}(x|\mathbf{y})$ in A_k.

4.1.5 Design Algorithm

In this section we state the SSQ design algorithm, for a squared distortion metric, and show its convergence to a local optimum on the objective function surface. The SSQ design algorithm is iterative, with two main steps to the iteration, each based on the necessary conditions at the encoder and decoder, respectively.

As stated earlier, the regions A_k, $k \in \{0, ..., K-1\}$, that determine the encoder function for SSQs are not in general intervals, but rather are the union of several disjoint intervals: $A_i = \bigcup_j A_{ij}$, where A_{ij} is an interval and $\bar{A}_{ij} \bigcap \bar{A}_{ik} = \emptyset$ for all $j \neq k$ ($\bar{A}$ indicates the closure of A). We refer to

any interval A_{ij} as a *cell* of the SSQ. For the SSQ design algorithm, we start by uniformly dividing the region of support for $\mathbf{x}$ into several equally sized small intervals, and approximate a given A_i by the union of several of these small intervals. More formally, we assume that a large interval $\mathcal{C} = [-C/2 \;\; C/2]$ of the real line contains virtually all of the region of support for $f_{\mathbf{x}}(x)$ and that this interval is divided into T equally spaced intervals of length C/T, forming a partition $\mathcal{T}$. The intervals are denoted $[t_i,\ t_{i+1}]$, $i = 0, ..., T-1$, where $t_0 = -C/2$, $t_T = C/2$, and $t_{i+1} = t_i + C/T$. The point in center of the interval $[t_i,\ t_{i+1}]$ is denoted x_i. We will see that in order for the algorithm to converge, we must allow the sampling of $\mathcal{C}$ to vary dynamically. If we start with a fine enough partition $\mathcal{T}$, however, the algorithm will converge adequately with a fixed partition.

The selection of a good value of T is important to approximate the optimal decision regions adequately. Given an observation $\mathbf{y}$, for an optimal encoder, the region of support for $f_{\mathbf{x}|\mathbf{y}}(x|\mathbf{y})$ will likely intersect most of the A_is, because an optimal encoder will minimize the distortion due to quantizer "granularity". We wish to have our intervals $[t_i,\ t_{i+1}]$ to be several times smaller that the smallest A_{ij}, in order to approximate the A_{ij} adequately. Thus a good value of T is one for which KC/T is, on average, several times smaller that the standard deviation of the posteriori densities $f_{\mathbf{x}|\mathbf{y}}(x|\mathbf{y})$. Note that we are assuming that for a fixed decoder g, there exists at least one interval $A_{lm} \in A_l \in \mathcal{A}$ such that $|A_{lm}| \leq |A_{ij}| \forall i \neq l, j \neq m$. This is a safe assumption for most sources and channels of interest, a fact supported by the following argument. Assume there is some density $f_{\mathbf{xy}}(x, y)$ such that the optimum encoder does not satisfy the assumption that there is some minimum length $|A_{lm}|$. In this case the variance of $f_{\mathbf{x}|\mathbf{y}}(x|y)$ must approach zero for some region of $\mathcal{Y}^N$ while still maintaining considerable probability mass in that region – such an example is clearly pathological. The assumption of a minimum length interval assures the convergence of the algorithm.

Given a suitably small stopping threshold δ, the SSQ design algorithm is described by the following steps:

1) Set the iteration counter $l = 0$. Set the initial number of intervals $T^{(l)}$ to some suitably large T. Select an initial partition $\mathcal{A}^{(l)}$. Set the objective function $J^{(l)} = \infty$.

2) $l \leftarrow (l+1)$. $T^{(l)} \leftarrow T^{(l-1)}$. Assuming $\mathcal{A}^{(l-1)}$ governs the operation of the encoder, determine the optimal decoder function $g^{(l)}$ from (4.13).

3) Assuming $g^{(l)}$ is the fixed decoder function, for each $x \in \{x_0, ..., x_{T^{(l)}-1}\}$ determine the optimal encoder function $f(x)$ from (4.4). Determine the elements of $\mathcal{A}^{(l)}$ by the following: $A_i^{(l)} = \bigcup_{\{j:f(x_j)=i\}}[t_j,\ t_{j+1}]$, $i = 0, ..., K-1$.

4) Compute the objective function $J^{(l)}$ from (4.9). If $J^{(l)} \leq J^{(l-1)}$, go to step 5). Otherwise set the counter $m = 0$; set $L^{(m)} = J^{(l)}$; and execute the following

   a) $T^{(l)} \leftarrow 2T^{(l)}$. Compute $A_i^{(l)}$, $i = 0, ..., K-1$ as in step 3).

   b) $m \leftarrow (m+1)$. Compute the objective function $L^{(m)}$ from (4.9). If $L^{(m)} \leq J^{(l-1)}$ then $J^{(l)} \leftarrow L^{(m)}$; go to step 5). Otherwise go to step 4a).

5) If $(J^{(l-1)} - J^{(l)})/J^{(l-1)} < \delta$ then stop. Otherwise go to step 2).

The algorithm clearly converges, as the objective function decreases at every step of the algorithm and is bounded by zero. Step 4) is in place to ensure that the objective function will decrease at each iteration despite our approximation that the A_{ij}s are the union of uniformly spaced intervals in $\mathcal{C}$. Step 4) is usually not required by other quantizer design algorithms [47, 27, 77], as, unlike the systematic design problem, these problems have simple descriptions of the regions A_i, and thus their algorithms for determining the encoder do not require sampling of the region of support for x. Empirical evaluation has shown that Step 4a), b), and c) are usually unnecessary if the initial sampling of $\mathcal{C}$ is sufficiently fine ($T^{(0)} = T$ is sufficiently large).

The fixed point of the algorithm will be a point at which the gradient of the objective function ∇J is zero with respect to the encoder and decoder parameters. This fact is true because at the fixed point the necessary conditions at the encoder and decoder are clearly satisfied. Note that the fixed point is not necessarily a global minimum as the zero gradient condition implies only a local minimum or saddle point. In general, the direct calculation of $J^{(l)}$ from (4.9) in step 4) is difficult, because it entails solving an integral involving $\hat{x}_i(\mathbf{y})$, which is often a complicated expression. We suggest approximating the integral in (4.9) with a Riemann sum.

4.1.6 Partial feedback of analog side information

In some scenarios, there may exist feedback from the source decoder to the source encoder of a form of the side information $\mathbf{y}$. The feedback can potentially improve the rate-distortion performance for

the forward path. As discussed in Sec. 3.3, in the rate-distortion limit, even full feedback does not improve the rate-distortion performance in the Gaussian case, but for finite-dimensional quantizers, the feedback may still have a positive effect. In Sec. 4.2.5, for SSQs we determine analytically the coding gain using feedback over no feedback for the Gaussian case. In many cases the feedback channel is a finite-rate digital channel, which requires that $\mathbf{y}$ be quantized prior to transmission. In this section, for simplicity we assume that $\mathbf{y}$ is a scalar y, and the feedback is a scalar quantized version of y. In the feedback scenario, there exist encoders at both the transmitter and receiver ends, and a single source decoder at the receiver end. We distinguish the two encoders by calling them the *source encoder* and *feedback encoder*, respectively. Note that for this feedback scenario, we do not have an issue with instantaneous feedback, as the source encoder may buffer the source (after it has been transmitted over the analog channel) before transmitting its encoded version over the analog channel. In this section we develop algorithms, which are natural extensions of the SSQ design algorithms, that yield locally optimal scalar encoders (source and feedback) and the corresponding optimal decoder, where an optimal design is again defined as one which minimizes J as given by (4.1). Prior to the work in this thesis there have only been asymptotic results regarding systematic coding with partial feedback [9, 8].

The scalar coding system with feedback is represented by several functions. The decoder map $g : \{0, ..., K-1\} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$ operates as in the standard SSQ, with the source codeword index k and side information y as input, and the source estimate $\hat{x}_k(y)$ as output. The feedback encoder map $h : \mathcal{Y} \rightarrow \{1, ..., P\}$, has y as input and a codeword index as output to be used by the source encoder. The feedback encoder map is completely defined by the partition of $\mathcal{Y}$ into the set $\mathcal{B} = \{B_1, ..., B_P\}$, where we define $B_i = \{y : h(y) = i\}$, $i = 1, ...P-1$. The source encoder map $f : \{1, ..., P\} \times \mathcal{R} \rightarrow \{0, ..., K-1\}$, has the source as input, as with standard SSQ, and it has additional input from the feedback path; the output of the encoder map is the codeword index used by the decoder. As in the case of standard SSQ, the source encoder map is completely defined by a partition $\mathcal{A}$ of $\mathcal{X}$, but with feedback, the partition is a function of the feedback codeword index p: $\mathcal{A}^p = \{A_1^p, ..., A_K^p\}$, where we define $A_i^p = \{x : f(p, x) = i\}$, $i = 0, ..., K-1$.

Similar to standard SSQ design, design of SSQ with feedback (SSQF) is accomplished by deriving necessary conditions at the two encoders and the decoder. First we note that the input to the decoder, k and y, is unchanged from the standard SSQ problem. Since the decoder knows y, it can

determine $p = h(y, k)$, and the decision region A_k^p in which x lies. Thus, the necessary condition at the decoder is unchanged from standard SSQ. Letting $A_k = A_k^p$, the necessary condition is given by (4.10) for the general case and (4.13) for the case of squared distortion. To derive the necessary condition at the source encoder we define

$$J_{x,p} = E_y[D(x, \hat{x}_{f(x,p)}(y))|y \in B_p, x], \tag{4.14}$$

$$= \frac{1}{\mathrm{P}(y \in B_p)} \int_{y \in B_p} D(x, \hat{x}_{f(x,p)}(y)) f_{y|x}(y|x) dy, \tag{4.15}$$

and rewrite J as

$$J = \sum_{p=1}^{P} \mathrm{P}(y \in B_p) E[D(x, \hat{x}_{f(x,p)}(y))|y \in B_p] \tag{4.16}$$

$$= \sum_{p=1}^{P} \mathrm{P}(y \in B_p) E_x[E_y[D(x, \hat{x}_{f(x,p)}(y))|y \in B_p, x]] \tag{4.17}$$

$$= \sum_{p=1}^{P} \mathrm{P}(y \in B_p) E[J_{x,p}]. \tag{4.18}$$

Clearly, J is minimized only if $J_{x,p}$ is minimized for all x and p. Assume the decoder g is fixed, defining the set of reconstruction functions $\hat{x}_k(y)$, which are indexed by $k \in \{0, ..., K-1\}$. For a given $x = x$ and $p = p$, the optimal encoder will minimize $J_{x,p}$ by selecting the k corresponding to the optimal set A_k^p to which to assign x. Thus for a fixed decoder g, a necessary condition at the encoder for an optimal systematic quantizer with feedback is

$$f(p, x) = \arg\min_k E_y[D(x, \hat{x}_k(y))|y \in B_p, x = x]. \tag{4.19}$$

The only difference between (4.19) and (4.4) is that in case of feedback, the source density is modified to reflect the feedback observation. In terms of the optimal encoder decision regions, (4.19) is rewritten as

$$A_k^p = \{x : E_y[D(x, \hat{x}_k(y))|y \in B_p, x = x] \leq E_y[D(x, \hat{x}_j(y))|y \in B_p, x = x] \ \forall j \neq k\}. \tag{4.20}$$

Defining

$$J_y = E_x[D(x, \hat{x}_{f(x,h(y))}(y))|y], \tag{4.21}$$

we obtain the necessary condition at the feedback encoder by rewriting J as

$$J = E_y[J_y]. \tag{4.22}$$

Clearly, J is minimized only if J_y is minimized for all y. We assume a fixed decoder g which defines the set of functions $\hat{x}_k(y)$, and a fixed source encoder f which defines the partition $\mathcal{A}^p$ (for each p), containing the decision regions A_k^p. Given these assumptions, an optimal feedback encoder selects $h(y)$ to satisfy:

$$h(y) = \arg \min_{p \in \{1,...,P\}} E_x[D(x, \hat{x}_{f(x,p)}(y))|y], \tag{4.23}$$

which defines the partition $\mathcal{B}$ in the obvious manner.

The SSQF design algorithm is iterative with three main steps, in contrast to the two main steps for the SSQ design algorithm. We assume the region of support for x is sampled as in Sec. 4.1.5, with T_x total samples. The sampling intervals are denoted $[t_i^x t_{i+1}^x]$, $i = 0, ..., T-1$, and the centers of the intervals are denoted x_i. In order to approximate the optimal feedback decision regions B_p we assume similar sampling of the region of support for y, with T_y total samples. The sampling intervals are denoted $[t_i^y t_{i+1}^y]$, $i = 0, ..., T-1$, and the centers of the intervals are denoted y_i.

Because there are three necessary conditions for the optimal feedback coding system, there are many ways to implement a convergent algorithm; we present one here. For simplicity, we assume at each iteration l that $T_x^{(l)} = T_y^{(l)} = T^{(l)}$. Given a suitably small stopping threshold δ, the SSQF design algorithm is described by the following steps:

1) Set the iteration counter $l = 0$. Set the initial number of intervals $T^{(l)}$ to some suitably large T. Select the initial partitions $\mathcal{A}^{p,(l)}$, $p = 1, ..., P$ and $\mathcal{B}^{(l)}$. Set the objective function $J^{(l)} = \infty$.

2) $l \leftarrow (l+1)$. $T^{(l)} \leftarrow T^{(l-1)}$. Assuming $\mathcal{A}^{p,(l-1)}$ and $\mathcal{B}^{(l-1)}$ govern the operation of the source encoder and feedback encoder respectively, determine the optimal decoder function $g^{(l)}$ from (4.13).

3) Assuming $g^{(l)}$ is the fixed decoder function and $\mathcal{A}^{p,(l-1)}$ governs the operation of the source

encoder, for each $y \in \{y_0, ..., y_{T^{(l)}-1}\}$ determine the optimal feedback encoder function $h(y,k)$ for all y and k from (4.23). Determine the elements of $\mathcal{B}^{(l)}$ by the following: $B_i^{k,(l)} = \bigcup_{\{j:h(y_j,k)=i\}}[t_j^y,\ t_{j+1}^y]$, $i = 1, ..., P$.

4) Compute the objective function $J^{(l)}$ from (4.22). If $J^{(l)} \leq J^{(l-1)}$, go to step 5). Otherwise set the counter $p = 0$; set $L^{(p)} = J^{(l)}$; and execute the following

   a) $T^{(l)} \leftarrow 2T^{(l)}$. Compute $B_i^{(l)}$, $i = 1, ..., P,$, as in step 3).

   b) $p \leftarrow (p+1)$. Compute the objective function $L^{(p)}$ from (4.22). If $L^{(p)} \leq J^{(l-1)}$ then go to step 5). Otherwise go to step 4a).

5) Assuming $g^{(l)}$ is the fixed decoder function and $\mathcal{B}^{(l-1)}$ governs the operation of the feedback encoder, for each $x \in \{x_0, ..., x_{T^{(l)}-1}\}$ determine the optimal feedback encoder function $f(p,x)$ for all p and x from (4.19). Determine the elements of $\mathcal{A}^{p,(l)}$ by the following: $A_i^{p,(l)} = \bigcup_{\{j:f(p,x_j)=i\}}[t_j^x,\ t_{j+1}^x]$, $i = 0, ..., K-1$.

6) Compute the objective function $J^{(l)}$ from (4.22). If $J^{(l)} \leq J^{(l-1)}$, go to step 7). Otherwise set the counter $q = 0$; set $L^{(q)} = J^{(l)}$; and execute the following

   a) $T^{(l)} \leftarrow 2T^{(l)}$. Compute $A_k^{p,(l)}$, $k = 0, ..., K-1,$, $p = 1, ..., P$, as in step 5).

   b) $q \leftarrow (q+1)$. Compute the objective function $L^{(q)}$ from (4.22). If $L^{(q)} \leq J^{(l-1)}$ then $J^{(l)} \leftarrow L^{(q)}$; go to step 7). Otherwise go to step 6a).

7) If $(J^{(l-1)} - J^{(l)})/J^{(l-1)} < \delta$ then stop. Otherwise go to step 2).

As with the standard SSQ algorithm, the SSQF algorithm converges to a fixed point that is either a local minimum or a saddle point on the surface of J with respect to the system parameters. Note that including feedback adds another "nested" layer of complexity to operation of the source encoder and to the SSQ design algorithm. The source encoder is comprised of P encoder maps (as opposed to just one for SSQ), indexed by the feedback codebook index p. Similarly, we can think of the SSQF design algorithm as the SSQ design algorithm carried out for each $p \in \{1, ..., P\}$, with the additional update step of a feedback encoder as each iteration. Even though we are using simple scalar design elements the design complexity is quite formidable.

4.2 SSQ: the Gaussian case

In this section we investigate the application of SSQ to the important case of zero-mean jointly Gaussian x and $\mathbf{y}$, and a square distortion metric. We are assuming that there is no feedback unless otherwise noted. For many sources, including images and audio, the Gaussian source model accurately represents the given source and has been applied successfully to many applications, including source coding. The Gaussian channel model is widely encountered in practical applications. It very accurately represents the AM channel and closely approximates the FM channel (in the baseband domain) in the high SNR case using phase-locked loop receivers [85]. We begin with a qualitative discussion of the behavior of a good SSQ encoder and decoder for the Gaussian case, and determine that the nested lattice scalar quantizer (NLSQ) satisfies those properties. Given the good properties of the NLSQ, we use a suitably selected NLSQ as an initial condition for the SSQ design algorithm. We run the SSQ design algorithm with $\mathbf{y} = x + n$, $n \sim \mathcal{N}(0, \sigma_n^2)$, for a variety of SNRs, and observe that the algorithm, not surprisingly, converges to quantizer that is approximately an NLSQ. In Sec. 4.2.5 we perform analysis to determine the optimal parameterization of the NLSQ for a given $\sigma^2_{x|\mathbf{y}}$. This analysis yields a convenient expression for the operational rate-distortion function.

4.2.1 The NLSQ encoder map

A well known property of a zero-mean jointly Gaussian pair $(x, \mathbf{y})$ is that the conditional density $f_{x|\mathbf{y}}(x, \mathbf{y})$ is Gaussian with a constant variance

$$\sigma^2_{x|\mathbf{y}} = \sigma_x^2 - \mathbf{\Lambda}_{x\mathbf{y}} \mathbf{\Lambda}_{\mathbf{y}}^{-1} \mathbf{\Lambda}_{x\mathbf{y}}^T \tag{4.24}$$

independent of y and a conditional mean

$$E[x|\mathbf{y}] = \mathbf{\Lambda}_{x\mathbf{y}} \mathbf{\Lambda}_{\mathbf{y}}^{-1} \mathbf{y}, \tag{4.25}$$

which, of course, is the MMSE estimate of x from $\mathbf{y}$. As detailed at the outset of this chapter, the SSQ encoder map is defined by the partition $\mathcal{A}$, comprised of decision regions A_i, $i = 0, ..., K-1$ each of which is the union of disjoint intervals, or cells. A desirable property for any SSQ to have is

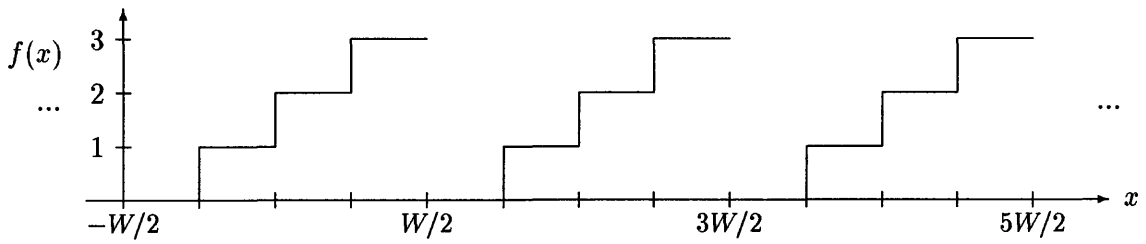


Figure 4-2: NLSQ encoder map, a 2 bit example.

the ability at the decoder to distinguish which cell $\boldsymbol{x}$ came from, given $\boldsymbol{k}$ and $\mathbf{y}$. Given that $f_{\boldsymbol{x}|\mathbf{y}}(x|\mathbf{y})$ has infinite extent over x, exact determination of the cell is not possible, but if an encoder partition such as that illustrated in Fig. 4-1(c) is used, it is clear the cell can be determined with high probability. The partition shown in Fig. 4-1(c), called the *nested lattice scalar quantizer* (NLSQ) encoder map for reasons described below, can be represented by a simple encoder map that is the composition of a modulo operation with a standard uniform quantizer encoder map:

$$f(x) = U_{\Delta,W}(x \bmod W) \tag{4.26}$$

$$U_{\Delta,W}(\nu) = k, \quad k\Delta \leq \nu + \frac{W}{2} < (k+1)\Delta \tag{4.27}$$

where $U(\nu)$ is a standard uniform quantizer encoder map with a domain $[-W/2, W/2]$ and output level $k \in \{0, ..., K-1\}$, and $\Delta = W/K$. From Fig. 4-1(c) we see that W is on the order of $\sigma_{\boldsymbol{x}|\mathbf{y}}$. Assuming that the density $f_{\boldsymbol{x}}(x)$ is supported on the (very large) interval $[-C/2,\ C/2]$, an equivalent definition of $f(x)$ to (4.26) reverses the order of the modulo and uniform quantizer functions:

$$f(x) = U_{\Delta,C}(x) \bmod K, \tag{4.28}$$

where in (4.28) the modulo function operates on whole numbers, as opposed to reals in (4.26). The function $f(x)$ is illustrated in Fig. 4-2 for $K = 4$, and it shows a characteristic repeated staircase structure, where the staircase width is W.

As shown in Fig. 4-1(c), because the variance of the aposteriori density $f_{\boldsymbol{x}|\mathbf{y}}(x, \mathbf{y})$ is independent of $\mathbf{y}$, the density can be supported almost entirely by K distinct cells, each corresponding to a different decision region. Thus, the decoder can determine which cell the source $\boldsymbol{x}$ came from with high probability. This appealing feature suggests that using an encoder map defined by (4.26)

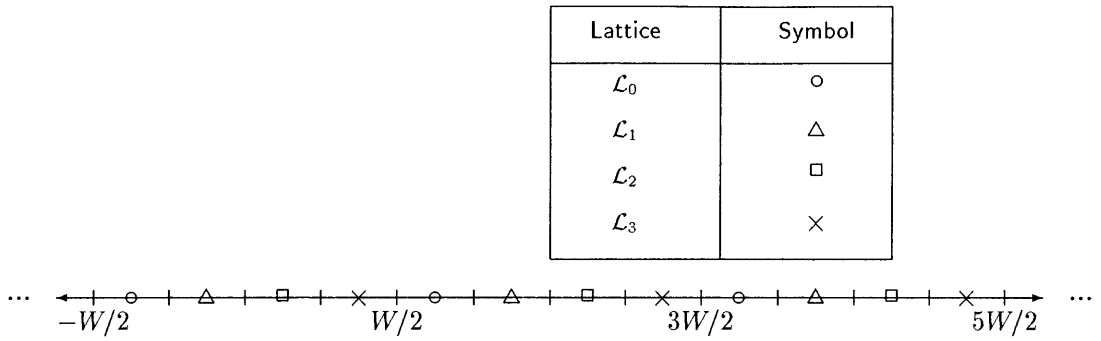


Figure 4-3: Lattice interpretation of NLSQ encoder map.

would be a good initial condition for the SSQ design algorithm.

For implementational and descriptive purposes, it is convenient to view (4.26) and (4.28) as operations on nested lattices, as illustrated in Fig.4-3, hence the meaning behind the designation NLSQ. Let $\mathcal{C}$ be a uniform scalar lattice with infinite extent on the real line and lattice points spaced W/K units apart. In Fig.4-3, $\mathcal{C}$ is the collection of all the lattice points in the figure. Let $\mathcal{L}_0$ be a uniform lattice nested in $\mathcal{C}$ with lattice points spaced W units apart, as shown in Fig.4-3, *i.e.* $\mathcal{L}_0$ is a lattice that is made up of every K^{th} lattice point of $\mathcal{C}$. For a given lattice $\mathcal{K}$ we define

$$Q_{\mathcal{K}}(x) \triangleq \arg\min_{l \in \mathcal{K}} \|x - l\|. \tag{4.29}$$

We recognize that the NLSQ is identical to the coding system using nested lattices described in Sec. 3.3.4 specialized to dimension $N = 1$. The NLSQ encoder simply calculates $l = Q_{\mathcal{C}}(x)$ and transmits the coset shift $l - Q_{\mathcal{L}_0}(l)$ with respect to the lattice $\mathcal{L}_0$. For $k = 0, ..., K\text{-}1$ we define the lattices $\mathcal{L}_k = \{l : l - Q_{\mathcal{L}_0}(l) = k\}$, comprised of all the points corresponding to the coset shift k. Using this definition, we see that the encoder simply sends the index k of the lattice $\mathcal{L}_k$ that is closest to x.

4.2.2 SSQ design algorithm

In this section we use the SSQ design algorithm to determine the optimal SSQ for the case where $\mathbf{y}$ is scalar and is given by $y = x + n$, where $n \sim \mathcal{N}(0, \sigma_n^2)$ is uncorrelated with x. Fixing $\sigma_x^2 = 20$ we run the algorithm for a variety of SNRs, from 0dB to 30dB, and several numbers of levels K, from 2 to 64. Empirical results suggest that even for the Gaussian case, the mean-squared error surface is multi-modal with respect to the quantizer parameters. Thus the selection of initial condition $\mathcal{A}^{(0)}$

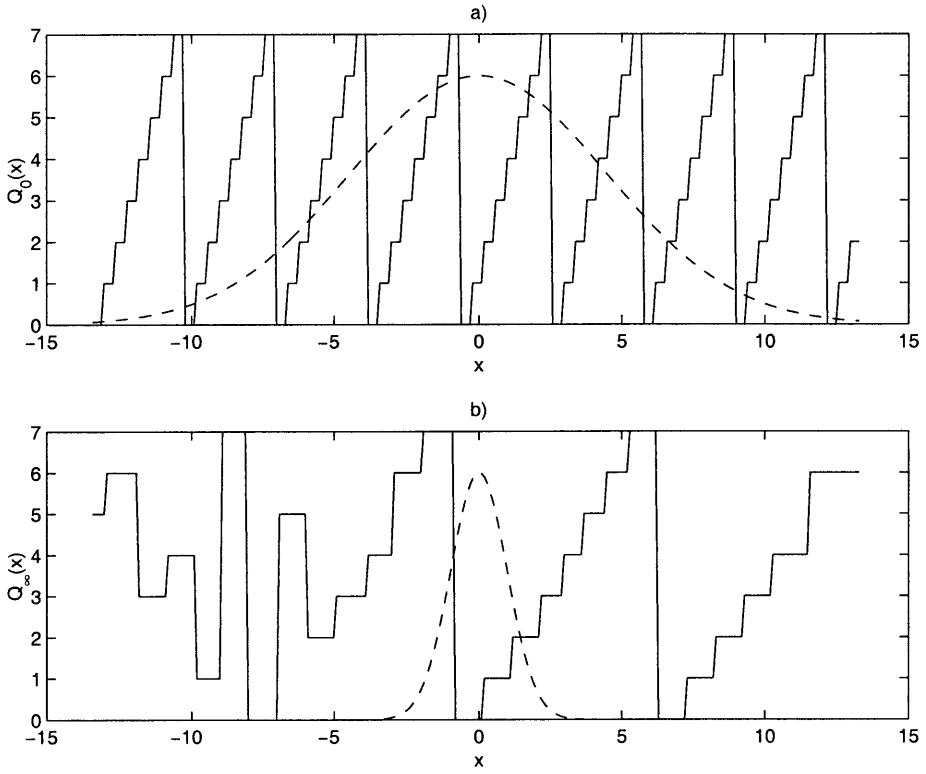


Figure 4-4: Quantizer design, $\sigma_x^2 = 20$, 13 dB SNR. a) initial condition $f^{(0)}(x)$. The dashed line is a scaled version of $f_x(x)$. b) MMSE quantizer $f^{(\infty)}$. The dashed line is a scaled version of $f_{x|y}(x|y=0)$

may impact the performance of the SSQ design. This is in contrast to standard scalar quantization in which case the Lloyd-Max quantizer is the globally optimal quantizer for any density that satisfies log-convexity [28, 12, 68], a property satisfied by the Gaussian case. Using the intuition gained from Sec. 4.2.1 we initialize the algorithm with an NLSQ encoder map $f^{(0)}$ with $W = 3\sigma_{x|y}$. For all cases tested, convergence occurs in three to five iterations, with more iterations required for the higher SNRs and larger K. In Fig. 4-4 we show the results of the design procedure for $\sigma_x^2 = 20$, SNR=13dB, and $K = 8$. Fig. 4-4(a) shows a scaled version of $f_x(x)$ superimposed on the initial encoder map $f^{(0)}(x)$. Fig. 4-4(b) shows the fixed point of the algorithm $f^{(\infty)}(x)$ with a scaled version of $f_{x|x}(x|y=0)$ superimposed. Interestingly the fixed point partition of the algorithm has approximately a repeated staircase structure. In contrast to the initial partition, the fixed point has $W \approx 7.5\sigma_{x|y}$, and the MSE is significantly improved. We also performed the experiment at 13 dB SNR using a more random initial condition, by assigning an index selected uniformly from

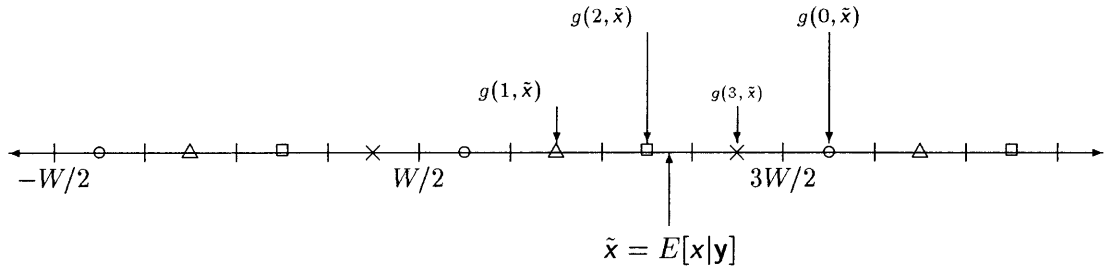


Figure 4-5: The NLSQ decoder, a 2 bit example

$\{0, ...K-1\}$ to each of the intervals $[t_i, t_{i+1}]$, $i = 0, ...T$-1, in $\mathcal{T}$. The results were very similar to those in Fig. 4-4(b).

For reasonably high SNR (>3 dB) and moderate rates, we observe that the output of the SSQ design algorithm has a similar structure to the one observed in Fig. 4-4(b). Thus it is safe to assume that a good approximation for the optimal SSQ encoder map for the Gaussian case is given by (4.26), the NLSQ encoder map, for SNRs greater than 3dB, where the staircase width depends on $\sigma_{x|\mathbf{y}}$ and R. Indeed, we have confirmed empirically through many trials that an NLSQ encoder map with a suitable W is very close to a fixed point of the SSQ design algorithm. Finding the optimal W via analysis is the focus of Sec. 4.2.5.

4.2.3 NLSQ decoder

The NLSQ encoder is a very low-complexity encoding mechanism, given by (4.26) or (4.28). For the design of efficient signal processing systems, we would like the operation of the NLSQ decoder map to be equally simple. Recall that given the index $\boldsymbol{k}$, the MMSE decoder function is the centroid of $f_{x|\mathbf{y}}(x|\mathbf{y})$ in $A_{\boldsymbol{k}}$. We observe from the output of the SSQ design algorithm that W is many times $\sigma_{x|\mathbf{y}}$, implying that the vast majority of the $f_{x|\mathbf{y},x\in A_{\boldsymbol{k}}}(x|\mathbf{y})$ is contained in only one cell of $A_{\boldsymbol{k}}$, say $A_{\boldsymbol{k}j}$, the nearest cell in $A_{\boldsymbol{k}}$ to $E[x|\mathbf{y}] = \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{y}$. Assuming a high rate R, $f_{x|\mathbf{y}}(x|\mathbf{y})$ is relatively constant over $A_{\boldsymbol{k}j}$, implying the MMSE estimate is approximately the center of $A_{\boldsymbol{k}j}$. We use this approximation to define a low-complexity NLSQ decoder map. In terms of the lattices defined in Sec. 4.2.1, the decoder simply maps $\tilde{x} = E[x|\mathbf{y}]$ to the nearest point in the lattice $\mathcal{L}_{\boldsymbol{k}}$, as illustrated in Fig. 4-5 for $R = 2$. The NLSQ decoder map can be viewed simply as a standard quantizer at the decoder, operating on $E[x|\mathbf{y}]$, where the quantizer is selected by the index $\boldsymbol{k}$. The reconstruction points are the lattice points of $\mathcal{L}_{\boldsymbol{k}}$ and the decision region boundaries are half-way points between the lattice points. Note that this is exactly the same operation as the *encoder* for

the QIM information embedding method [15]. For QIM, the encoder quantizes the host $\mathbf{y}$ with a quantizer that is selected (or modulated) by the information index. The intuition gained in this section is the same intuition that leads to the discovery in Sec. 3.3.4 that the rate-distortion function for the Gaussian case can be attained for all SNRs with nested lattices of high dimension, only if at the decoder, $\mathbf{y}$ is attenuated by the MMSE estimation gain ρ prior to quantization.

4.2.4 Properties of NLSQ quantization noise

With the NLSQ properly defined, we comment on the statistical properties the quantization noise induced by the NLSQ. We assume that x has the vast majority of its probability mass supported by a large interval $[-C/2, C/2]$. With the staircase width W adequately large, the NLSQ decoder will clearly identify the correct cell from which x originates with high probability. Thus the effective operation of the NLSQ is approximately that of a standard uniform quantizer with resolution $\Delta = W/K$, operating at rate of $\log_2(KC/W)$, over the interval [-C/2, C/2]. We can therefore apply results from standard SQ [31], which conclude that scalar quantization noise can be modeled accurately as being uncorrelated with $\mathbf{x}$, having a uniform marginal density over $[-\Delta/2, \Delta/2]$ and having a flat power spectral density.

4.2.5 Optimal NLSQ staircase width W

In this section we determine the optimal W for an NLSQ assuming that $K = 2^R$ is large, where R is the allocated number of bits per sample of the quantizer. First, we consider qualitatively the impact on the optimum W of increasing the number of bits R. There are two types of distortion that impact the MSE, which we name *intracell distortion* and *intercell distortion.* Intracell distortion is the granular distortion induced by the quantization step size Δ and it occurs when the decoder correctly identifies which cell x comes from. Intercell distortion is distortion induced when the decoder incorrectly identifies the cell from which x comes, and is therefore off by some multiple of W. Consider increasing the number bits R while keeping W fixed. The intracell distortion decreases with each bit, while the intercell stays relatively constant, eventually dominating the MSE and making each new bit ineffective. Thus, with increasing R, the staircase width, W must increase for optimum overall MSE. Interestingly, this insight about the increase of W with rate indicates that in the limit as $R \to \infty$, for a fixed SNR, the optimal SSQ is, in fact, not an

NLSQ. As R increases, in order to keep intercell error from dominating the total distortion, the width between cells of each decision region A_k, $k = 0, ..., K$-1 must increase, eventually to the point where the source density $f_x(x)$ is virtually all supported by K distinct cells. In this case, the results of standard SQ indicate that the optimal encoder map will be that of a Lloyd-Max quantizer. In the Gaussian case, this quantizer is non-uniform, and hence cannot be represented by an NLSQ. For moderate bit rates, however, we will see that the NLSQ is in fact near-optimal, and it is for this case that it is important to calculate the optimal NLSQ staircase width W.

The optimal W, as a function of R, is derived by determining a expression for MSE, in terms of only W and R and minimizing with respect to W. We will make a few approximations in determining the MSE, but they prove to have little impact in the determination of the optimal W. Because K is large, we assume that the cell size $\Delta = W2^{-R}$ is small relative to $\sigma^2_{x|\mathbf{y}}$. Our objective function is MSE, which is given by (4.1), $J = E_{\mathbf{y}}[E_x[J_{\mathbf{y}}]]$, where $J_{\mathbf{y}}$ is given by (4.8).

In Fig. 4-6 we show a representative aposteriori density $f_{x|\mathbf{y}}(x|\mathbf{0})$ whose domain is partitioned as an NLSQ encoder map for a given W and Δ. Each cell labeled with an i, $i = 0, ..., K$-1, belongs to an encoder decision region A_i. An important observation is that because all densities $f_{x|\mathbf{y}}$ are Gaussian with the same variance (differing only in mean) and the granularity of the partition is fine, the region of support for $f_{x|\mathbf{y}}(x|\mathbf{y})$ will be partitioned in approximately the same way as $f_{x|\mathbf{y}}(x|\mathbf{0})$ in Fig. 4-6 for all $\mathbf{y}$. The spacing between cells will be identical for all $\mathbf{y}$; it is only the labelling of the cells that will be permuted. Thus the MSE conditioned on $\mathbf{y}$ is constant for all $\mathbf{y}$, *i.e.,* $J_{\mathbf{y}}$ is constant for all $\mathbf{y}$, simplifying the objective function to be $J = J_{\mathbf{y}=\mathbf{0}}$. For the rest of this analysis, we assume a fixed $\mathbf{y} = \mathbf{0}$. In Fig. 4-6 there are shown for each k, $k = 0, ..., K$-1, corresponding to a decision region A_k, three relevant cells (each labeled with by k), one in the center, the left and the right. Given $\mathbf{y}$, the probability of x begin outside these cells for any k is negligible. We denote the center cell by A_k^{C} and the union of the left and right cells by A_k^{LR} and rewrite $J_{\mathbf{y}=\mathbf{0}}$ as

$$\begin{aligned} J_{\mathbf{y}=\mathbf{0}} &= \sum_{k=0}^{K-1} \mathrm{P}(x \in A_k^{\mathrm{C}}|\mathbf{y}=\mathbf{0}) E[(x-\hat{x}_k)^2|\mathbf{y}=\mathbf{0}, x \in A_k^{\mathrm{C}}] \\ &\quad + \sum_{k=0}^{K-1} \mathrm{P}(x \in A_k^{\mathrm{LR}}|\mathbf{y}=\mathbf{0}) E[(x-\hat{x}_k)^2|\mathbf{y}=\mathbf{0}, x \in A_k^{\mathrm{LR}}] \qquad (4.30) \\ &= \sum_{k=0}^{K-1} \Big(\int_{-W/2+\Delta k}^{-W/2+\Delta(k+1)} f_{x|\mathbf{y}}(x|\mathbf{0}) dx \Big) E[(x-\hat{x}_k)^2|\mathbf{y}=\mathbf{0}, x \in A_k^{\mathrm{C}}] \end{aligned}$$

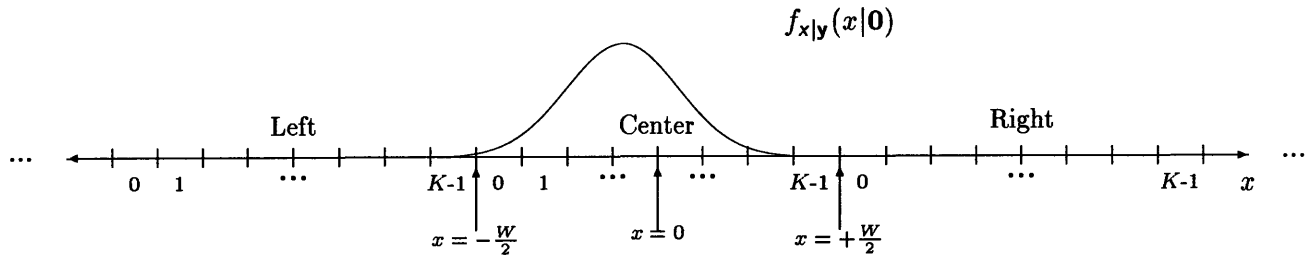


Figure 4-6: The aposteriori density $f_{x|y=0}$ whose domain is partitioned by an NLSQ encoder map partition. The MSE averaged over all $\mathbf{y}$ equals the MSE for the case $\mathbf{y} = \mathbf{0}$. The regions of the domain that contribute significantly to the MSE are labeled Left, Center, and Right, corresponding to separate staircases of width W.

$$+2\sum_{k=0}^{K-1}\Big(\int_{+W/2+\Delta k}^{+W/2+\Delta(k+1)} f_{x|y}(x|\mathbf{0})dx\Big)E[(x-\hat{x}_k)^2|\mathbf{y}=\mathbf{0}, x\in A_k^{\mathrm{LR}}]. \tag{4.31}$$

Recall that for a given $k = k$ and $\mathbf{y}$ the NLSQ decoder selects as its reconstruction point the center of the nearest cell in A_k to $E[x|\mathbf{y}]$, which equals zero for $\mathbf{y} = 0$. Given our high rate assumption, we hence have the following very accurate approximations:

$$E[(x-\hat{x}_k)^2|\mathbf{y}=\mathbf{0}, x\in A_k^{\mathrm{C}}] \approx \Delta^2/12,\ k=0,...,K-1 \tag{4.32}$$

and

$$E[(x-\hat{x}_k)^2|\mathbf{y}=\mathbf{0}, x\in A_k^{\mathrm{LR}}] \approx W^2+\Delta^2/12.\ k=0,...,K-1 \tag{4.33}$$

The independence of (4.32) and (4.33) on the digital index k allow us to simplify $J_{\mathbf{y}=\mathbf{0}}$ even further:

$$\begin{aligned} J_{\mathbf{y}=\mathbf{0}} &\approx \frac{\Delta^2}{12}\int_{-W/2}^{W/2} f_{x|y}(x|\mathbf{0})\,dx + 2(W^2+\frac{\Delta^2}{12})\int_{W/2}^{3W/2} f_{x|y}(x|\mathbf{0})\,dx & (4.34)\\ &= \frac{\Delta^2}{12}\int_{-3W/2}^{3W/2} f_{x|y}(x|\mathbf{0})\,\mathrm{d}x + 2W^2\int_{W/2}^{3W/2} f_{x|y}(x|\mathbf{0})\,dx & (4.35)\\ &\approx \frac{\Delta^2}{12}1 + 2W^2Q(W/(2\sigma_{x|y})), & (4.36)\end{aligned}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-x^2/2}\,dx. \tag{4.37}$$

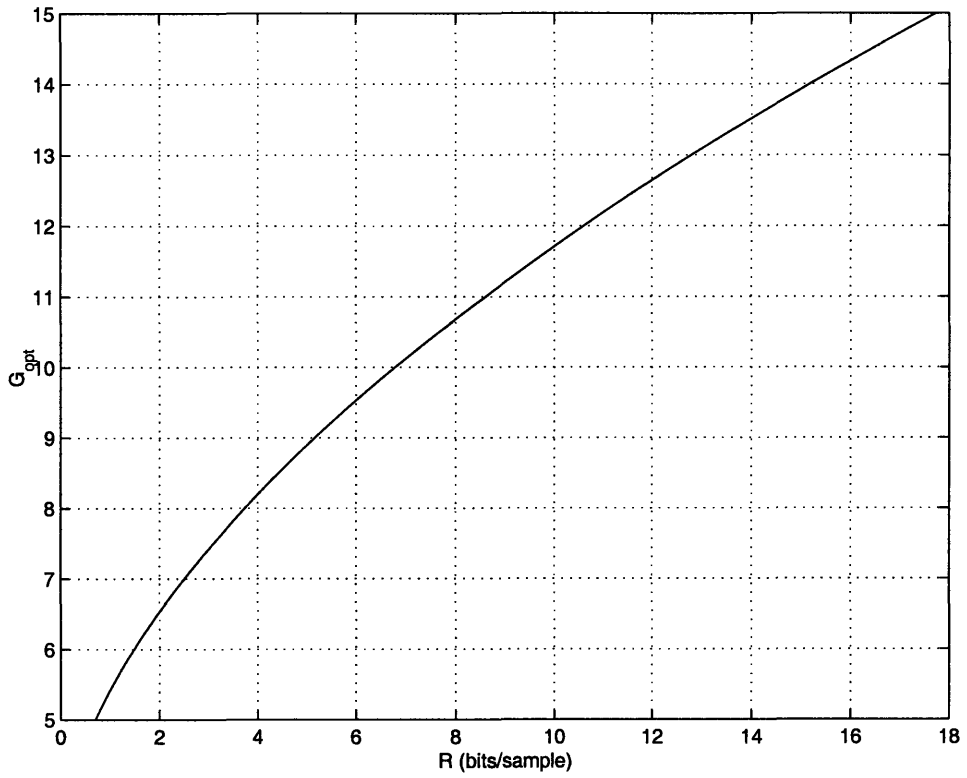


Figure 4-7: Optimal gain G_{opt} as a function of bit rate R. The NLSQ staircase width is $W = G_{\text{opt}}\sigma_{x|y}$.

To arrive at (4.36) we have assumed that W is large enough so that the integrals in (4.35) approach their asymptotic values. Letting $W = G\sigma_{x|y}$ for some constant factor G and $\Delta = W2^{-R}$ and assuming our approximate equalities are met with equality, we have

$$J_{y=0} = \sigma^2_{x|y}\Big(\frac{G^2}{12}2^{-2R} + 2G^2Q(G/2)\Big), \tag{4.38}$$

whose derivative with respect to G is

$$\frac{dJ}{dG} = \sigma^2_{x|y}\Big(\frac{G}{6}2^{-2R} - \frac{1}{\sqrt{2\pi}}2G^2e^{-G^2/8} + 4GQ(G/2)\Big) \tag{4.39}$$

Setting (4.39) equal to zero, we cannot readily solve for G in terms of R, but we can solve for R in terms of G:

$$R = -\frac{1}{2}\log_2\Big(\frac{6}{\sqrt{2\pi}}Ge^{-G^2/8} - 24Q(G/2)\Big). \tag{4.40}$$

We argued at the outset of Sec. 4.2.5 that G should be a monotonically increasing function of R, which implies that (4.40) is invertible and yields the desired $G_{\text{opt}}(R)$. Thus, we compute R from (4.40) for a range of G, and plot $G_{\text{opt}}(R)$ in Fig. 4-7, which is indeed a monotonically increasing function of R. Note that G_{opt} is always greater than 5, which implies that an interval of length $W_{\text{opt}}(R) = G_{\text{opt}}(R)\sigma_{x|\mathbf{y}}$ will support the vast majority of $f_{x|\mathbf{y}}(x|\mathbf{y})$. Thus, $Q(W/(2\sigma_{x|\mathbf{y}})) \approx 0$, which together with (4.38) yields the operational distortion-rate function for the NLSQ:

$$d \triangleq J_{\mathbf{y}=\mathbf{0}} = \frac{G^2_{\text{opt}}(R)}{12}\sigma^2_{x|\mathbf{y}}2^{-2R}. \tag{4.41}$$

We use (4.41) to compare the performance of the NLSQ to the optimal SSQ designed by the MMSE design algorithm.

4.2.6 Operational rate-distortion functions

In this section we evaluate the operational rate-distortion performance of the SSQ and the NLSQ for a unit variance source x, and a channel output $y = x + n$, where n is additive Gaussian noise independent of x. For a given SNR, we consider a range of rates between 1 and 6 bits, and for each rate, design the SSQ, appropriately parameterize the NLSQ, and evaluate the distortion for both SQs. The SSQ is designed with a initial partition that is a suboptimal NLSQ encoder partition with $W = 3\sigma x|\mathbf{y}$. The rate-distortion function for the NLSQ is simply calculated from (4.41).

We show the operational rate-distortion curves in Fig.s 4-8(a)-(c) as solid lines for 0, 10, and 20 dB respectively, with the corresponding Wyner-Ziv rate distortion curves, given by (3.12), as dashed lines. In Fig. 4-8(d) we compare the rate-distortion performances across the three analog SNRs. In Fig. 4-9, we compare the operational rate-distortion function of the SQ that uses y, a measurement at 8 dB SNR, optimally at the decoder versus the Lloyd-Max quantizer that ignores the existence y. As expected, we have significant gains from using $\mathbf{y}$, with higher gains for high SNRs. Also plotted in Fig. 4-9 is the standard Gaussian rate-distortion function, whose performance, in the case where $\mathbf{y}$ is ignored, can only be approached by complex N-dimensional quantizers in the limit of large N. We see that using only an 8dB analog signal at decoder and a simple NLSQ, we beat the bound. Of course, the comparison is not a fair one because the bound assumes no analog information at the decoder.

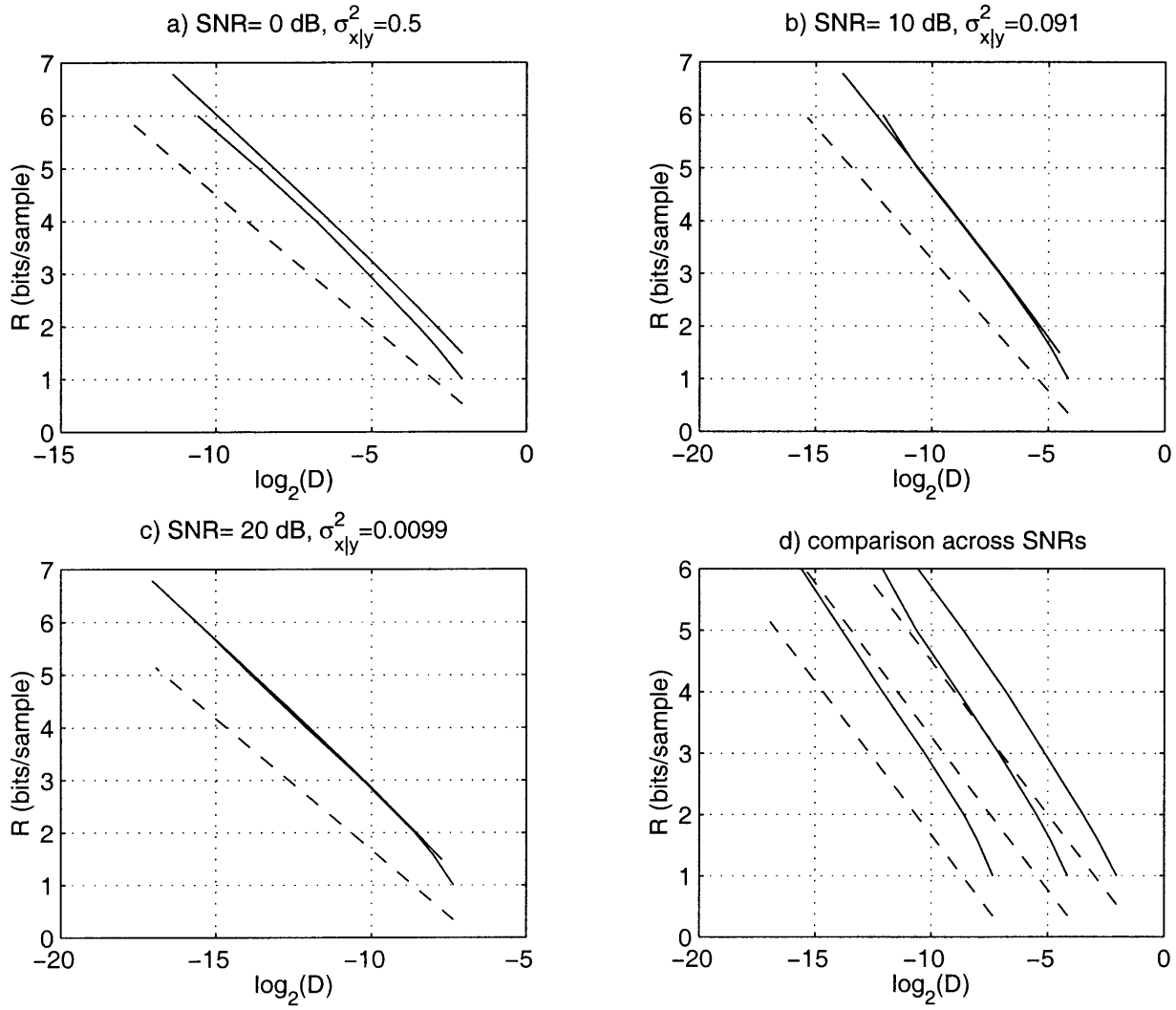


Figure 4-8: The operational rate-distortion functions of the optimal SSQ and the optimal NLSQ, where rate is plotted as a function of $\log_2(d)$. The solid lines are the operational curves, and the dashed lines are the theoretical lower bounds. The results shown are for $\sigma_x^2 = 1$ and analog SNR of (a) 0dB (b) 10 dB and (c) 20 dB. (d) shows a comparison across all SNRs of the optimal SSQs, where the curve from top to bottom correspond to 0, 10, and 20dB respectively.

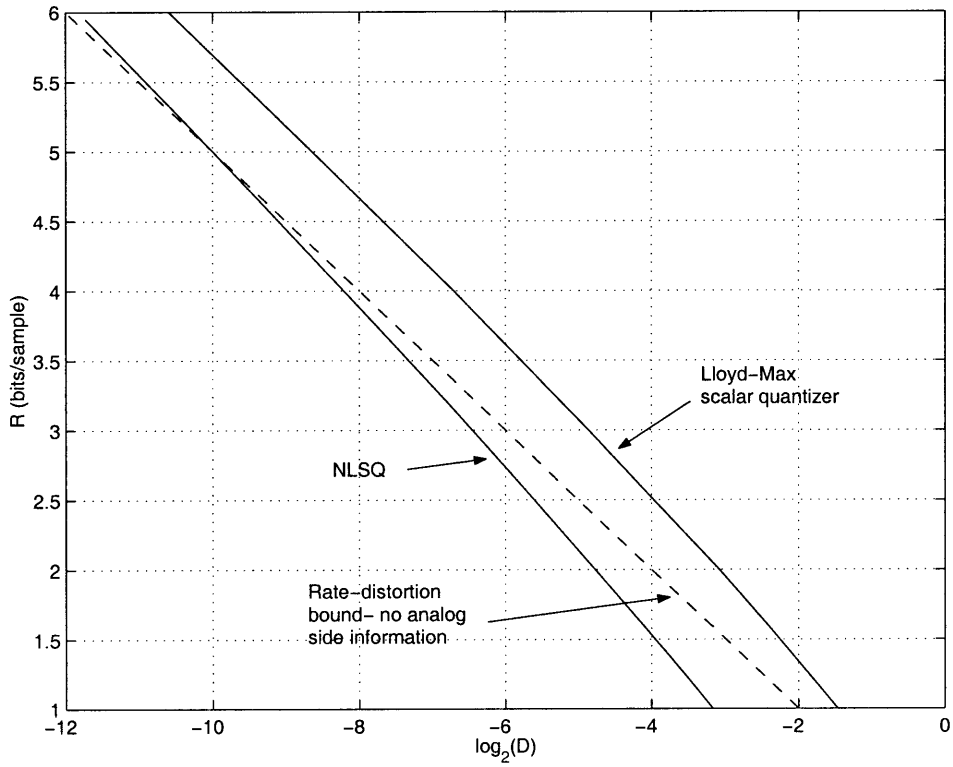


Figure 4-9: Comparison of operational rate-distortion performance of optimal SSQ versus quantizers that ignore y. The SNR is 8 dB. Shown as solid lines are the operational rate distortion functions of the optimal SQ with and without the analog channel output. The dashed line is the standard Gaussian rate-distortion function, which ignores $\mathbf{y}$.

There are several items worthy of mention about Fig. 4-8. First, note that the NLSQ performance matches that of the optimal SSQ almost exactly for SNRs of 10 and 20dB, and at 0dB the optimal SSQ outperforms the NLSQ by about a quarter bit per sample. Fig. 4-8 illustrates that at moderate to high SNRs the NLSQ is a very good approximation for the optimal SSQ while at low SNRs it is not a good approximation. A simple thought experiment reveals why the NLSQ encoder map is not optimal for low SNR. Assuming the limiting case where SNR approaches $-\infty$, the decoder gleans no information from the analog channel output and has only the digital encoding to use for reconstruction. Thus the problem is reduced to standard SQ, in which case the Lloyd-Max quantizer is the MMSE quantizer. In the Gaussian case, this quantizer is non-uniform, and hence cannot be represented by an NLSQ.

A second point to note is that for all SNRs the operational rate-distortion function for the optimal SSQ is about 1 to 1.5 bits per sample from the rate-distortion limit for most rates, which is quite large for some source coding applications. This gap lessens for very low bit rates near 1 bit per sample - the curves tend to tail off near 1 bit per sample.

Another interesting characteristic of Fig. 4-8 is that the rates of the NLSQ and optimal SSQ are essentially linear versus $\log_2(d)$, and the slope of the line in both cases is very near $-1/2$. This fact is especially important in that we can accurately approximate (4.41), the distortion-rate function of the NLSQ, by

$$d = \theta \sigma^2_{x|y} \, 2^{-2R}, \tag{4.42}$$

where

$$\theta = G^2_{\text{opt}}(R_{\text{ave}})/12, \tag{4.43}$$

for some R_{ave} that is a rate that takes on a value in the middle of the desired range of performance. For most systematic source coding applications, an SSQ will require at most 7 bits, which from Fig. 4-7 implies that $G_{\text{opt}}(R_{\text{ave}}) \approx 8$ implying that θ is about 16/3. Close inspection of Fig. 4-8 indicates that the slope of R versus $\log(d)$ for the NLSQ is slightly less than $-1/2$, which is accounted for by the variation of W_{opt} in (4.41) with R. Therefore, a more accurate approximation of $d(R)$ would replace the term 2^{-2R} in (4.42) with $2^{-(2-\epsilon)R}$ for some small ϵ. For our purposes, however, (4.42) is an accurate approximation. We shall see in Chap. 5 that (4.42) has important implication on optimal transform and subband coding and the accompanying optimal bit allocation.

Equation (4.42) also facilitates a comparison of SSQ with and without feedback. For the case with feedback, we can come up with an expression similar to (4.42) for the distortion-rate function in the limit of high feedback rate R_{F}, which implies full feedback. Recall that the optimal decoder function with feedback is given by (4.13). Using (4.19) in the case of full feedback, we see that the optimal encoder map with feedback is given by,

$$f(p, x) = \arg\min_k D(x, \hat{x}_k(y)) \tag{4.44}$$

Defining $z(\mathbf{y})$ as a random variable whose probability distribution is given by $f_z(z) = f_{x|\mathbf{y}}(z|\mathbf{y})$, we see that (4.13) and (4.19), are the necessary conditions that must be satisfied for a scalar quantizer acting on $z(\mathbf{y})$. Because z is Gaussian, these conditions are also sufficient for optimality. Thus the optimal quantizer with feedback is a Max quantizer for a Gaussian of mean $E[x|\mathbf{y}]$ and variance $\sigma^2_{x|\mathbf{y}}$. It is well known that the distortion-rate function for this Max quantizer, in the limit of high forward rate R, is given by [31]:

$$d_{\text{feedback}} = \theta_{\mathrm{F}} \sigma^2_{x|\mathbf{y}}\, 2^{-2R}, \tag{4.45}$$

where for the Gaussian case $\theta_{\mathrm{F}} = \sqrt{3}\pi/2$. Defining the coding gain γ as the ratio of the distortion without feedback, $d_{\text{no feedback}}$ to the distortion with feedback, we have

$$\begin{aligned}
\gamma &\triangleq \frac{d_{\text{no feedback}}}{d_{\text{feedback}}} && (4.46)\\
&= \frac{\theta \sigma_{x|\mathbf{y}} 2^{-2R},}{\theta_{\mathrm{F}} \sigma^2_{x|\mathbf{y}} 2^{-2R}} && (4.47)\\
&= \frac{16/3}{\sqrt{3}\pi/2} = 1.96 && (4.48)\\
&= 2.92\ \text{dB}, && (4.49)
\end{aligned}$$

nearly a 3dB gain. Equation (4.49) is equivalently expressed as a rate gain of $\frac{1}{2}\log_2(1.96) = .485$ bit per sample. This is in sharp contrast to the case of optimal N-dimensional vector quantization, in the limit of large N, where the use of feedback yields no coding gain.

4.3 Scalar Quantization and Slepian-Wolf Coding

As reflected by the Gaussian example in the previous section, there is a significant gap between an optimal SSQ operational rate-distortion function and the Wyner-Ziv rate-distortion limit. In order to approach the rate-distortion bound, we can employ vector techniques such as the ones using nested lattices described in [95] and Sec. 3.3.4. These methods, however, are impractical for large vector lengths, which are required to approach the rate-distortion bound. An alternative approach, offering potentially lower complexity, uses a uniform scalar quantizer followed by post-processing to improve the rate-distortion performance. For conventional source coding, it is well-known that for a mean-squared error distortion criterion, using a uniform scalar quantizer followed by entropy coding, a rate can be achieved that is within .255 bits per sample of the rate-distortion bound. An ad hoc extension of this result to systematic coding would employ an NLSQ followed by an entropy coder, but a simple example shows that this method offers little reduction in rate. Consider a Gaussian source $\mathbf{x}$ and analog channel output $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is additive white Gaussian noise (AWGN). Assuming a very high SNR, the MMSE error variance $\sigma^2_{x|y}$ is very small relative to σ^2_x, implying that the region of support for a single staircase of the NLSQ is small relative to σ_x. Thus within a given staircase width, the density $f_x(x)$ is relatively uniform, implying that the output of the NLSQ will be a discrete uniform random variable. As a sequence of uniform random variables is already coded to its entropy rate, entropy coding offers no reduction in rate for this example.

In this section we propose an analogous scheme to entropy-constrained scalar quantization, for systematic source coding where we follow a uniform SQ with Slepian-Wolf coding. Slepian-Wolf coding is described in Sec. 3.2.2. We prove that for a mean-squared distortion metric, for $(\mathbf{x}, \mathbf{y})$ drawn iid from $p_{xy}(x, y)$, using ideal Slepian-Wolf codes, this method achieves a rate that is within .255 bits per sample of the rate distortion bound. The assumption that $(\mathbf{x}, \mathbf{y})$ is iid is made only to facilitate a formal proof and may at first glance seem stringent. Through use of the rate-distortion bound for a sequence of independent Gaussians described in Sec. 3.3.2, we show that our performance bound implies similar low-complexity strategies that achieve the rate-distortion bound for jointly stationary Gaussian sources $x[n]$ and $y[n]$ with general cross correlation.

4.3.1 Encoding and decoding

The encoder applies a uniform scalar quantizer encoder map, $U(x)$, to each of the elements of the source $\mathbf{x}$, producing the output vector $\mathbf{q}$. The function $U(x)$ is an ordinary K-level uniform quantizer encoder map, $U : \mathcal{X} \to \{1, ..., K\}$, having granularity Δ across the interval $[-K\Delta/2 \;\; K\Delta/2]$. We have a one-to-one mapping between $\mathbf{q}$ and the vector $\mathbf{x_q}$, where we define each element $x_{q,i}$ as the center of the cell represented by q_i. The vector $\mathbf{q}$, or equivalently $\mathbf{x}_q$, is Slepian-Wolf encoded, assuming that the analog channel output $\mathbf{y}$ (unknown at the encoder) is at the decoder. Using the analog channel output and the coded bitstream, the decoder reconstructs $\mathbf{x_q}$ via Slepian-Wolf decoding. We form the source estimate $\hat{\mathbf{x}}$ by taking the centroid of the apostiori density in the cell represented by $\mathbf{q}$, or given a high-rate assumption, we use $\hat{\mathbf{x}} = \mathbf{x_q}$.

4.3.2 Slepian-Wolf coding

In this section we summarize some key points from Sec. 3.2.2 using notation that is relevant to this chapter. A Slepian-Wolf code losslessly encodes two iid sequences $\mathbf{u}$ and $\mathbf{v}$ (from discrete alphabets), jointly distributed as $p_{\mathbf{uv}}(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{n} p_{u_i v_i}(u_i, v_i)$, individually at two separate encoders, to be decoded by a single decoder. The main result of Slepian and Wolf is that for $R_u > H(U|V)$ and $R_v > H(V|U)$, any two rates, R_u and R_v, will allow perfect reconstruction at the decoder for some code as long as $R_u + R_v \geq H(u, v)$. Slepian-Wolf coding can be applied to signals from finite alphabets only, so we must assume that any continuously distributed signals, which occur in many relevant source coding applications, have been quantized, thereby introducing a degree of suboptimality. Thus, in the Wyner-Ziv setting considered in this section, we have $\mathbf{u} = \mathbf{x}_q$ and $\mathbf{v} = \mathbf{y}_q$, where $\mathbf{y}_q$ is a finely quantized version of the $\mathbf{y}$. Because $\mathbf{y}$ is observed directly at the decoder, $\mathbf{y}_q$ is effectively communicated losslessly to the decoder at its entropy rate $H(y_q)$. Note that the finer the quantization of $\mathbf{y}$, the higher this entropy rate (potentially approaching ∞), but this fact does not impact our coding method. The Slepian-Wolf result states that $\mathbf{x}_q$ can be coded losslessly at a rate $H(x_q|y_q)$, which in the limit of fine quantization of $\mathbf{y}$ equals $H(x_q|y)$.

The proof of the Slepian and Wolf result relies on the asymptotic behavior of random codes, which are not constructive, and hence indicate no meaningful implementation. There has been some recent work on practical Slepian-Wolf codes as found in [76, 64], but for arbitrary joint statistics on u and v no general method has been found that approaches the theoretical limit of performance. In

anticipation of the availability of good low-complexity Slepian-Wolf codes in the future, we present our result here.

4.3.3 Bound on performance

Consider a source $\mathbf{x}$ and channel output $\mathbf{y}$, distributed as $f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{n} f_{xy}(x_i, y_i)$. For the conventional case of the lossy coding of a single source, the well known .255 bit/sample upper bound on the deviation from the rate-distortion bound of entropy-constrained uniform quantization makes use of the Shannon lower bound (SLB) [34]. Similarly, we make use of the a conditional SLB [35, 92]:

$$R_{x|y}^{\mathrm{WZ}}(d) \geq R_{x|y}(d) \geq h(x|y) - \frac{1}{2}\log 2\pi e d \tag{4.50}$$

where $R_{x|y}(d)$ is the conditional rate-distortion function, described in Sec. 3.2.3 as the rate-distortion function given observations of $\mathbf{y}$ at the encoder and decoder.

For an observation $y = y$ of a scalar component of the channel output, we denote the corresponding scalar component by x, its quantized value by x_q, and its aposteriori density by $f_{x|y}(x|y)$. We assume that $f(x)$ has a high rate, *i.e.,* Δ is small with respect to the $\sigma_{x|y}$ for all y. This assumption may not hold in general but is true for many important estimation problems, including Gaussian estimation. For any receiver that can decode $\mathbf{x_q}$, the reconstruction distortion d is independent of the observation y and is given by:

$$d = E[(x - x_q)^2] \approx \frac{\Delta^2}{12}. \tag{4.51}$$

Given a vector observation $\mathbf{y}$, the n^{th}-order entropy of the quantized signal $\mathbf{x_q}$ is [34]

$$H(\mathbf{x_q}|\mathbf{y}) \approx h(\mathbf{x}|\mathbf{y}) - E[\log V(S_j)], \tag{4.52}$$

where $V(S_j)$ is the volume of the j^{th} Voronoi region defined by the quantizer. We will assume that (4.52) is nearly an equality, and from this point will claim equality. Because the Voronoi regions are n-dimensional cubes of side Δ, we have

$$H(\mathbf{x_q}|\mathbf{y} = \mathbf{y}) = h(\mathbf{x}|\mathbf{y} = \mathbf{y}) - \log \Delta^n, \tag{4.53}$$

which simplifies to

$$\sum_{i=1}^{n} H(\mathsf{x}_{q,i}|\mathsf{y}_i = y_i) = \sum_{i=1}^{n} h(\mathsf{x}_{q,i}|\mathsf{y}_i = y_i) - n \log \Delta. \tag{4.54}$$

Multiplying both sides of (4.54) by $1/n$ and averaging each term over all y, we have

$$\frac{1}{n}\sum_{i=1}^{n} H(\mathsf{x}_{q,i}|\mathsf{y}) = \frac{1}{n}\sum_{i=1}^{n} h(\mathsf{x}_{q,i}|\mathsf{y}) - \log \Delta. \tag{4.55}$$

Using (4.51) and taking (4.55) to the limit as $n \to \infty$ we have

$$H(\mathsf{x}_q|\mathsf{y}) = -\frac{1}{2}\log d - \frac{1}{2}\log 12 + h(\mathsf{x}|\mathsf{y}). \tag{4.56}$$

We combine (4.56) with (4.50) to get

$$\begin{aligned} H(\mathsf{x}_q|\mathsf{y}) &\leq R^{\mathrm{WZ}}_{\mathsf{x}|\mathsf{y}}(d) - \frac{1}{2}\log(12d) + \frac{1}{2}\log(2\pi e d) = \frac{1}{2}\log\left(\frac{\pi e}{6}\right) && (4.57) \\ &= R^{\mathrm{WZ}}_{\mathsf{x}|\mathsf{y}}(d) + .255. && (4.58) \end{aligned}$$

Thus we have proven that the cascade of a uniform quantizer and a Slepian-Wolf coder can yield performance within .255 bits/sample of the Wyner-Ziv rate-distortion bound for (x, y) drawn iid from $f_{xy}(x, y)$.

Note that the NLSQs of Sec. 4.2 for the Gaussian case, as expressed in (4.28), are effectively a uniform scalar quantizer followed by a suboptimal Slepian-Wolf-like code, the modulo element. The modulo element accounts for the presence of y at the receiver. The "code" is suboptimal because the apostiori density is not uniform over the support of a single staircase. If it were uniform for all y, then the uncoded output of the modulo quantizer would already be coded at a rate $H(\mathsf{x}_q|\mathsf{y})$, but this is not the case for jointly Gaussian x and y. To approach the rate-distortion bound more closely for the Gaussian case, we need a more sophisticated code, which will work on large blocks of quantizer output samples.

In [61], Pradhan *et al,* describe a systematic encoding system for the Gaussian case that effectively performs scalar quantization followed by Slepian-Wolf coding, although this fact is not acknowledged explicitly in the paper. Pradhan *et al,* use interleaved lattices at the encoder similar to nested lattices. The method differs from ours in that the component lattices are heuristically

based on Lloyd-Max reconstruction points. Pradhan *et al,* further code the SSQ outputs by employing a variation on the Slepian-Wolf code of Wyner for the binary symmetric case (see Sec. 3.2.2). At the decoder the quantized source is reconstructed using sequential decoding based on the Viterbi algorithm [29].

4.4 Jointly Gaussian x and y

For most relevant sources, $(\mathbf{x}, \mathbf{y})$ is not an iid sequence. We anticipate that future work will extend the proof of this result to general $(\mathbf{x}, \mathbf{y})$, likely exploiting the generalization of the Slepian-Wolf result to ergodic sources [24]. For now, we extend the iid result to coding a stationary jointly Gaussian processes $\mathbf{x}$ and $\mathbf{y}$ with general cross-correlation. Recall from Sec. 3.3.2 that for $(\mathbf{x}, \mathbf{y})$ a sequence of jointly Gaussian independent random variables, the rate-distortion function is given by (3.28). Intuitively, we achieve (3.28) by separately coding the x_i sequences of a given error variance, giving none of the rate to a sequence of error variance less than λ. For every other sequence, we apply a code that achieves the Wyner-Ziv rate-distortion limit at the distortion level λ, a type of coding which is termed inverse-waterpouring.

As noted in Sec. 3.3.2, if we consider the stationary Gaussian processes $x[n]$ and $y[n]$ in the limit of long observation times, the Fourier basis is the limiting Karhunen-Loeve basis for both processes. Furthermore there exists no cross-correlation between coefficients projected onto different basis functions. Thus $x[n]$ and $y[n]$ projected onto Fourier basis function elements satisfy our assumptions of independence. An approximation to Fourier decomposition is a time-frequency decomposition implemented by a cosine-modulated, critically sampled orthonormal filter bank [80]. We formally define a filter bank in Chapter 5.

We propose a coding system that uses a filter bank and uniform quantization and Slepian-Wolf coding applied to each of the subbands to achieve within .255 bits/sample of the Wyner-Ziv rate-distortion performance limit for a stationary Gaussian source and jointly Gaussian observation at the decoder. We do not discuss all of the details of subband coding in this section, as we defer to Chap. 5 for the full development. Applying the analysis filter bank separately to $x[n]$ and $y[n]$, we have the subband signals $X_i[m]$ and $Y_i[m]$, $i = 1, ..., M$, where M is the number of filters. We

assume the subband signals satisfy

$$E[X_i[m]X_j[n]] = \sigma_X^2\delta[i-j]\delta[m-n] \tag{4.59}$$

$$E[Y_i[m]Y_j[n]] = \sigma_Y^2\delta[i-j]\delta[m-n] \tag{4.60}$$

$$E[X_i[m]Y_j[n]] = \sigma_{XY}\delta[i-j]\delta[m-n], \tag{4.61}$$

where $\delta[i]$ is the Kroneker delta function. We note three important facts. First, by the fact that the filter bank is orthonormal, the MSE in the subband domain equals that in the time domain. Secondly, because the filter bank is critically sampled, we have the same number of subband domain samples as time domain samples. And finally, we note that the rate distortion function given in (3.28) is in units of bits per vector. Because we want the rate in bits per time sample, we must normalize the rate-distortion function by the "vector" length , or number of subband signals M. Using these observations we see that the Wyner-Ziv rate-distortion function for correlated Gaussian processes in terms of the subband signal variances is

$$\begin{aligned} R_{x|y}^{\mathrm{WZ}}(d) &= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\log\frac{\sigma_{X_i|Y_i}^2}{d_i}, \\ &= \frac{1}{M}\sum_{i=1}^{M}R_{X_i|Y_i}^{\mathrm{WZ}}(d_i) \end{aligned} \tag{4.62}$$

where we use $R_{X_i|Y_i}^{\mathrm{WZ}}(d)$ to denote the rate-distortion function for the subband signal $X_i[m]$;

$$d_i = \begin{cases} \lambda, & \text{if } \lambda < \sigma_{X_i|Y_i}^2, \\ \sigma_{X_i|Y_i}^2 & \text{if } \lambda \geq \sigma_{X_i|Y_i}^2, \end{cases} \tag{4.63}$$

and λ is chosen so that $\sum_{i=1}^{n} d_i = d$. We give zero rate to a subband with MMSE variance less than λ. To each of the other subband signals, we apply a uniform quantizer of step-size $\Delta = \lambda$, followed by a Slepian-Wolf code. For a given subband i requiring non-zero rate, the uniform quantizer achieves the desired distortion λ, and assuming $Y_i[m]$ at the decoder, the Slepian-Wolf code achieves a rate $R_i < R_{X_i|Y_i}^{\mathrm{WZ}}(d_i) + .255$ bits per sample. We have that the total average rate

for the system is

$$\begin{aligned} R(d) &= \frac{1}{M}\sum_{i=1}^{M} R_i && (4.64)\\ &< \frac{1}{M}\sum_{i=1}^{M}(R^{\mathrm{WZ}}_{X_i|Y_i}(d_i) + .255), && (4.65)\\ &= R^{\mathrm{WZ}}_{x|y}(d) + .255. && (4.66) \end{aligned}$$

thereby proving that for a simple system using scalar quantizers we achieve within .255 bits per sample of the Wyner-Ziv rate distortion function for stationary jointly Gaussian processes $x[n]$ and $y[n]$.

4.5 Vector Quantization

4.5.1 General structure

In the interest of coding efficiency, some scenarios may warrant grouping several source samples together and performing vector quantization (VQ). In this section we propose an algorithm called the SVQ design algorithm for vector quantization with side information at the decoder. For simplicity we assume no feedback, but incorporating feedback is a straightforward application of the results from Sec. 4.1.6. Scalar quantization on each component of a vector is form of VQ that partitions $\mathcal{X}^N$ with rectangular elements, which is often inefficient. VQ offers the ability to partition the region of support for a length-N vector x more optimally than that which can be achieved with scalar quantization. The problem formulation for SVQ is as follows. The vectors $\mathbf{x}$ and $\mathbf{y}$, of lengths N and M respectively, are drawn from the distribution $f_{\mathbf{xy}}(\mathbf{x},\mathbf{y})$. The encoder acts on $\mathbf{x}$ and produces a codeword index, and the decoder uses the codeword index and $\mathbf{y}$ to form an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$. Apart from acting on vectors and not scalars, the operation of the SVQ is the same as the SSQ described in Sec. 4.1, with an encoder map $\mathbf{f} : \mathcal{X}^N \to \{0, ..., K-1\}$, and a decoder map $\mathbf{g} : \{0, ..., K-1\} \times \mathcal{Y}^M \to \hat{\mathcal{X}}^N$. The encoder map describes a partition of N-space, $\mathcal{A} = \{A_1, ..., A_K\}$ where we define $A_i = \{\mathbf{x} : f(\mathbf{x}) = i\}$.

Recalling the necessary conditions for the encoder in the SSQ case (see (4.5) and (4.10)), there is clearly no dependence of these conditions on the fact that x is scalar. Thus, some form of the

SSQ design algorithm could conceivably be used for SVQ design. The problem lies in the evaluation of the integrals in (4.5) and (4.10), which grow unwieldy for large dimension. Furthermore, often the exact source and/or channel statistics may be difficult to characterize, or are only partially described. One of the common standard VQ design techniques, called the generalized Lloyd algorithm [34] or the k-means algorithm [50] avoids the evaluation of integrals of large dimension by using statistically generated test sequences, or Monte Carlo trials, based on the source statistics. Using this approach the integrals are approximated stochastically. In addition, the source and channel statistics are effectively learned by the algorithm from the stochastic realizations, avoiding the need for a closed-form statistical characterization. Similar to the generalized Lloyd algorithm, the SVQ design algorithm that we propose here uses sequences of vector pairs $(\mathbf{x}, \mathbf{y})$ assumed drawn from $f_{\mathbf{xy}}(\mathbf{x}|\mathbf{y})$.

In the scalar case, we observed that the partition decision regions A_k, $k = 0, ..., K-1$ are complicated and non-regular. Furthermore, the reconstruction functions $\hat{x}_k(y)$ have no convenient analytical expressions. The same is true of course for SVQ, except that these complex structures are even further complicated by the increase in the number of dimensions. Note that for some particular source/channel characteristics, simple SVQs with a high degree of structure can be used effectively. For example, in the Gaussian case discussed in Sec. 3.3.4, simple nested lattice codes, in the limit of large dimension, can be used to achieve the rate-distortion limit. In general, however, the only tractable method for SVQ quantizer implementation is to have the encoder map and decoder map operate by table lookup. Clearly the memory requirements of a system of table lookup become unmanageable for large N and M, and we do not propose that this form of SVQ is usable for all applications. The algorithm for SVQ design we propose here serves three main purposes. First, for moderate N and M, a table lookup system may be implementable and effective. Second, even for large N and M, the SVQ design algorithm yields near-optimum quantizers, whose error performance is a gauge against which to compare other lower complexity systems. Finally, the qualitative structure of the SVQ output from the design algorithm may guide the development of effective, low-complexity SVQ implementations.

4.5.2 Design algorithm

We generate a sequence of stochastic realizations of the source $\mathbf{x}_i$, $i = 1, ..., P$. For each $\mathbf{x}_i$, we generate L channel realizations, forming the sequence $\mathbf{y}_{ij}$, $j = 1, ..., L$. We create a partition of $\mathcal{Y}$, defined by $\mathcal{B} = \{B_i\}$, $i = 1, ..., G$, where each B_i is a connected region. Usually B_i is an M-dimensional rectangular region. We consider the objective function

$$J_{\text{SVQ}} = \sum_{i=1}^{P} \sum_{j=1}^{L} \|\mathbf{x}_i - \hat{\mathbf{x}}_{\mathbf{f}(\mathbf{x}_i)}(\mathbf{y}_{ij})\|^2, \tag{4.67}$$

where we impose the constraint on the decoder function:

$$\hat{\mathbf{x}}_k(\mathbf{y}) = \mathbf{x}_{k,i} \forall \mathbf{y} \in B_i \text{ for a fixed } i \in \{1, ..., G\}, \tag{4.68}$$

We wish to optimize J_{SVQ} with respect to $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(k, \mathbf{y})$. By the law of large numbers, J_{SVQ} will converge to J in (4.1) under the constraint (4.68). Thus, an SVQ that is near-optimal with respect to (4.68), is likely to be near optimal with respect to (4.1) under the decoder function constraint. If the $\mathcal{B}$ is fine enough a partition, (4.68) will not create much disparity between J_{SVQ} unconstrained versus constrained.

By rewriting (4.68) as

$$J_{\text{SVQ}} = \sum_{k=0}^{K-1} \sum_{i:\mathbf{x}_i \in A_k} \sum_{j=1}^{L} \|\mathbf{x}_i - \hat{\mathbf{x}}_k(\mathbf{y}_{ij})\|^2, \tag{4.69}$$

we have the necessary condition at the encoder:

$$\mathbf{f}(\mathbf{x}_i) = \arg\min_k \sum_{j=1}^{L} \|\mathbf{x} - \hat{\mathbf{x}}_k(\mathbf{y}_{ij})\|^2. \tag{4.70}$$

By rewriting (4.68) as

$$J_{\text{SVQ}} = \sum_{j=1}^{L} \sum_{k=0}^{K-1} \sum_{i:\mathbf{x}_i \in A_k} \|\mathbf{x}_i - \hat{\mathbf{x}}_k(\mathbf{y}_{ij})\|^2 \tag{4.71}$$

$$= \sum_{m=1}^{M} \sum_{k=0}^{K-1} \sum_{i:\mathbf{x}_i \in A_k} \sum_{j:\mathbf{y}_{ij} \in B_m} \|\mathbf{x}_i - \hat{\mathbf{x}}_{k,m}\|^2 \tag{4.72}$$

we have the necessary condition at the decoder:

$$\hat{\mathbf{x}}_{k,m} = \text{cent}_1\{(\mathbf{x}_i, \mathbf{y}_{i,j}) : i,j \text{ s.t. } \mathbf{x}_i \in A_k, \mathbf{y}_{ij} \in B_m\}, \tag{4.73}$$

where $\text{cent}_1\{(\mathbf{x}, \mathbf{y})\}$ is the centroid with respect to the $\mathbf{x}$ component of the pair.

The SVQ design algorithm is a simple iterated application of (4.70) and (4.73) given by the following steps:

1) Set the iteration counter $l = 0$. For all $i = 1, ..., P$, set $\mathbf{f}^{(l)}(\mathbf{x}_i)$ to some suitable initial value. Set the objective function $J_{\text{SVQ}}^{(l)} = \infty$.

2) $l \leftarrow (l+1)$. Assuming $\mathbf{f}^{(l-1)}(\mathbf{x}_i)$ governs the operation of the encoder, determine the optimal decoder function $\hat{\mathbf{x}}_{k,m}^{(l)}$ from (4.73) for all $k = 0, ..., K$-1 and $m = 1, ..., M$.

3) Assuming $\hat{\mathbf{x}}_{k,m}^{(l)}$ is the fixed decoder function, for all $i = 1, ..., P$, determine $\mathbf{f}^{(l)}(\mathbf{x}_i)$ from (4.70).

4) Compute $J_{\text{SVQ}}^{(l)}$ from (4.69). If $(J^{(l-1)} - J^{(l)})/J^{(l-1)} < \delta$ for a suitable threshold δ then stop. Otherwise go to step 2).

Clearly, the algorithm converges, as J_{SVQ} decreases at every iteration and is bounded by zero. Furthermore, the fixed point of the algorithm is a local minimum or saddle point of the error surface J_{SVQ} with respect to the encoder and decoder parameters, as implied by the satisfaction of both the necessary conditions at the encoder and decoder.

Once the stochastic optimization procedure has concluded, the partition $\mathcal{A}$ with respect to the continuous space $\mathcal{X}^N$ can be determined by one of many heuristic methods. A reasonable method is the following. Partition each component of $\mathcal{X}^N$ with a scalar partition $\mathcal{C}$ comprised of successive intervals of the same length. We have thereby defined a partition of $\mathcal{X}^N$: $\mathcal{C}^N = \{C_j\}$, $j = 1, ..., H$, which is the N-dimensional cross product of $\mathcal{C}$; each C_j is an N-dimensional cube. We assign the codebook index k to a $C \in \mathcal{C}^N$ as follows:

$$k(C) = \arg\max_l |\{x_i \in C : f(x_i) = l,\ i = 1, ..., P\}|. \tag{4.74}$$

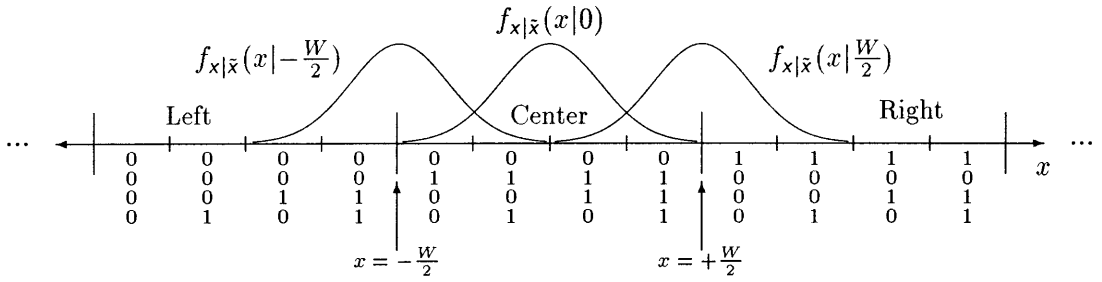


Figure 4-10: A partition used for low-bits coding. The bits are labeled most significant to least significant bit, top to bottom.

The decision regions A_i, $i = 0, ..., K\text{-}1$ are defined by

$$A_i = \bigcup_{j=1}^{H} \{C_j : k(C_j) = i\}. \tag{4.75}$$

Using this method the encoder can be implemented by scalar quantization of the elements of $\mathbf{x}$, followed by a table lookup. The decoder can similarly be designed by table lookup.

4.6 Low-Bits Coding

We conclude this chapter by evaluating, for the Gaussian case, the ad hoc scheme for systematic coding, called low-bits coding (LBC), described at the outset of this chapter in Sec. 4.1.1. For simplicity we will assume that $\mathbf{y} = y$ is scalar, but the results of this section are easily generalized to $\mathbf{y}$ a vector. We assume that R is large, and that the SNR is large enough that $f_x(x)$ can be assumed constant over several standard deviations $\sigma_{x|y}$. We first show that LBC encoding is equivalent to NLSQ encoding, which is near-optimal for high SNR, and that LBC decoding is a suboptimal form of decoding that impairs the performance relative to an NLSQ. The main result of this section is the derivation of a formula for the coding gain of NLSQ over LBC, which is exponential in the rate R of the digital system.

The C-bit digitization of the source over the interval $[-P/2, P/2]$ creates a partition of the interval into uniform cells of width $P2^{-C}$, each cell corresponding to a C-bit index. We have shown the partition in Fig 4-10 where $C = 4$ and $R = 2$. For the purposes of discussion, in Fig. 4-10 we show that the region of support for x is $[-3P/8, 5P/8]$, with the mean of x equaling $P/8$, and we

have defined $W = P/4$. Clearly this is an equivalent setting to zero mean x and region of support $[-P/2, P/2]$. For $R = 2$, the encoder sends the 2 LSBs of the index corresponding to the cell in which x resides. Note that for a given 2 bit LSB transmission, say 00, represented by level k, the encoder specifies a partition A_k that is the union of intervals of width W/K, where $K = 4$, spaced uniformly by W. Indeed, the LBC encoder is an NLSQ, where in general $K = 2^R$, and the staircase width is $W = P2^{C-R}$. We will be assuming in this section that $W = W_{\text{opt}}(R)$, to have a basis of comparison to the NLSQ.

We will assume an LBC decoder that performs as follows. The decoder calculates the MMSE estimate $\tilde{x} = E[x|y]$ from y, digitizes the estimate with a C-bit A/D converter over the desired dynamic range, and replaces the R LSBs for each sample with those received from the digital channel to form the reconstruction $\hat{x}$. Because $\tilde{x} = \rho y$ is an invertible function of y we will perform all of our calculations with respect to our new channel observation $\tilde{x}$, and consider the aposteriori density $f_{x|\tilde{x}}(x|\tilde{x})$. Note that $\sigma_{\tilde{x}}^2 = \sigma_x^2$. Over an length-$W$ interval of values for $\tilde{x}$, the LBC decoder will assign the estimate $\hat{x}$ the same $C - R$ most significant bits. In Fig. 4-10 these intervals are delimited by the large vertical hash marks. The boundaries of the intervals occur at $-3W/2$, $-W/2$, $W/2$, $3W/2$, and $5W/2$. In general, we denote the intervals corresponding to a fixed set of $C - R$ most significant bits by $S_i = [-P/2 + iP2^{-(C-R)}, -P/2 + (i+1)P2^{-(C-R)}]$, $i = 0, ..., 2^{C-R}$-1. For any $\tilde{x}$ residing in one of these intervals, say S_i, the decoder will assign $\hat{x}$ to a some reconstruction point in the same interval. The actual point it assigns is the center of the cell corresponding to k the LSB index received on the digital channel. This method of decoding is suboptimal because when $\tilde{x}$ is at the edge of S_i, a closer reconstruction point from the lattice $\mathcal{L}_k$ will exist outside of S_i. The NLSQ, on the other hand, will choose the closest lattice point in $\mathcal{L}_k$, for the reconstruction.

From our analysis in Sec. 4.2.5 we know that given the high rate assumption, the MSE of the NLSQ is simply given by $J_{\text{NLSQ}} = \Delta^2/12$, where $\Delta = W/K$. We calculate the MSE of LBC by the following:

$$J_{\text{LBC}} = E_{\tilde{x}}[E_x[(x - \hat{x}_{f(x)}(\tilde{x}))^2|\tilde{x}]] \tag{4.76}$$

$$= \sum_{i=0}^{2^{C-R}-1} \int_{\tilde{x}\in S_i} E_x[(x - \hat{x}_{f(x)}(\tilde{x}))^2|\tilde{x}]] f_{\tilde{x}}(\tilde{x}) d\tilde{x} \tag{4.77}$$

$$= \sum_{i=0}^{2^{C-R}-1} \int_{\tilde{x}\in S_i} J_{\tilde{x}} f_{\tilde{x}}(\tilde{x}) d\tilde{x}, \tag{4.78}$$

where

$$J_{\tilde{x}} = E_x[(x - \hat{x}_{f(x)}(\tilde{x}))]^2|\tilde{x}]]. \tag{4.79}$$

As in Sec. 4.2.5 we focus our attention on three regions, labeled "Left", "Center", and "Right", in Fig. 4-10, corresponding to the intervals $[-3W/2, -W/2]$, $[-W/2, W/2]$, and $[W/2, 3W/2]$, respectively. We consider the quantity, $J_{\tilde{x},\mathrm{C}} = J_{\tilde{x}}$ for all $\tilde{x} \in [-W/2, W/2]$. We define $A_{k,\mathrm{L}}$, $A_{k,\mathrm{C}}$, $A_{k,\mathrm{R}}$ as the k^{th} cell of the left, center, and right regions, respectively. As shown in Fig. 4-10, we safely assume that for all $\tilde{x} \in [-W/2, W/2]$ that $f_{x|\tilde{x}}(x|\tilde{x})$ has the vast amount of it probability in $[-3W/2, 3W/2]$, and thus we expand $J_{\tilde{x},\mathrm{C}}$ as follows:

$$\begin{aligned} J_{\tilde{x},\mathrm{C}} &= E_x[(x, \hat{x}_{f(x)}(\tilde{x}))|\tilde{x}]] \qquad (4.80) \\ &= \sum_k \int_{A_{k,\mathrm{L}}} (x - \hat{x}_k(\tilde{x}))^2 f_{x|\tilde{x}}(x|\tilde{x}) dx \\ &\quad + \sum_k \int_{A_{k,\mathrm{C}}} (x - \hat{x}_k(\tilde{x}))^2 f_{x|\tilde{x}}(x|\tilde{x}) dx \\ &\quad + \sum_k \int_{A_{k,\mathrm{C}}} (x - \hat{x}_k(\tilde{x}))^2 f_{x|\tilde{x}}(x|\tilde{x}) dx. \qquad (4.81) \end{aligned}$$

We denote the centers of the cells $A_{k,\mathrm{R}}$, $A_{k,\mathrm{C}}$, and $A_{k,\mathrm{R}}$ by $\hat{x}_{k,\mathrm{L}}$, $\hat{x}_{k,\mathrm{C}}$, and $\hat{x}_{k,\mathrm{R}}$, respectively. By the high rate assumptions, we assume that $f_{x|\tilde{x}}(x|\tilde{x})$ is relatively constant across a given cell. Thus we have

$$\begin{aligned} J_{\tilde{x},\mathrm{C}} &= \sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{L}}|\tilde{x}) \int_{A_{k,\mathrm{L}}} (x - \hat{x}_{k,\mathrm{C}})^2 dx \\ &\quad + \sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{C}}|\tilde{x}) \int_{A_{k,\mathrm{C}}} (x - \hat{x}_{k,\mathrm{C}}))^2 dx \\ &\quad + \sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{R}}|\tilde{x}) \int_{A_{k,\mathrm{R}}} (x - \hat{x}_{k,\mathrm{R}}))^2 dx. \qquad (4.82) \\ &= \sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{L}}|\tilde{x}) \Big(\frac{\Delta^3}{12} + \Delta W^2\Big) \\ &\quad + \sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{C}}|\tilde{x}) \frac{\Delta^3}{12} \end{aligned}$$

$$+\sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{R}}|\tilde{x})\Big(\frac{\Delta^3}{12}+\Delta W^2\Big) \tag{4.83}$$

$$\begin{aligned} = \ & \Big(\frac{\Delta^2}{12}+W^2\Big)\sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{L}}|\tilde{x})\Delta \\ & +\frac{\Delta^2}{12}\sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{C}}|\tilde{x})\Delta \\ & +\Big(\frac{\Delta^3}{12}+\Delta W^2\Big)\sum_k f_{x|\tilde{x}}(\hat{x}_{k,\mathrm{R}}|\tilde{x})\Delta. \end{aligned} \tag{4.84}$$

Given that $\Delta = W/K$ is small, we can approximate the summations in (4.84) with integrals, yielding

$$\begin{aligned} J_{\tilde{x},\mathrm{C}} \ = \ & \Big(\frac{\Delta^2}{12}+W^2\Big)\int_{-3W/2}^{-W/2} f_{x|\tilde{x}}(x|\tilde{x})dx \\ & +\frac{\Delta^2}{12}\int_{-W/2}^{W/2} f_{x|\tilde{x}}(x|\tilde{x})dx \\ & +\Big(\frac{\Delta^3}{12}+\Delta W^2\Big)\int_{W/2}^{3W/2} f_{x|\tilde{x}}(x|\tilde{x})dx \end{aligned} \tag{4.85}$$

$$\begin{aligned} = \ & W^2\Big\{Q\Big(\frac{-3W/2-\tilde{x}}{\sigma_{x|y}}\Big)-Q\Big(\frac{-W/2-\tilde{x}}{\sigma_{x|y}}\Big)\Big\} \\ & +\frac{\Delta^2}{12}\Big\{Q\Big(\frac{-3W/2-\tilde{x}}{\sigma_{x|y}}\Big)-Q\Big(\frac{3W/2-\tilde{x}}{\sigma_{x|y}}\Big)\Big\} \\ & W^2\Big\{Q\Big(\frac{W/2-\tilde{x}}{\sigma_{x|y}}\Big)-Q\Big(\frac{3W/2-\tilde{x}}{\sigma_{x|y}}\Big)\Big\} \end{aligned} \tag{4.86}$$

In this treatment we are assuming a high SNR, high enough so that over any interval of length $W = G\sigma_{x|y}$, the density $f_{\tilde{x}}(\tilde{x})$ can be assumed constant. Given this assumption and denoting the center of each interval S_i by x_{S_i}, we write (4.78) as

$$J_{\mathrm{LBC}} = \sum_{i=0}^{2^{C-R}-1} f_{\tilde{x}}(x_{S_i})\int_{\tilde{x}\in S_i} J_{\tilde{x}}d\tilde{x}. \tag{4.87}$$

We note from Fig. 4-10 and (4.86) that the terms

$$\int_{\tilde{x}\in S_i} J_{\tilde{x}}d\tilde{x} \tag{4.88}$$

can be computed identically to

$$\int_{-W/2}^{W/2} J_{\tilde{x},\mathrm{C}} d\tilde{x} \tag{4.89}$$

simply by shifting the mean of x by some multiple of W. Hence (4.88) equals (4.89) for all S_i, implying from (4.87) that

$$J_{\mathrm{LBC}} = \Big(\int_{-W/2}^{W/2} J_{\tilde{x},\mathrm{C}} d\tilde{x} \Big) \sum_{i=0}^{2^{C-R}-1} f_{\tilde{x}}(x_{S_i}). \tag{4.90}$$

Using L to denote the term in (4.89), we have

$$\begin{aligned} L &= W^2 \int_{-W/2}^{W/2} \Big\{ Q\Big(\frac{-3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) - Q\Big(\frac{-W/2 - \tilde{x}}{\sigma_{x|y}}\Big) \Big\} d\tilde{x} \\ &\quad + \frac{\Delta^2}{12} \int_{-W/2}^{W/2} \Big\{ Q\Big(\frac{-3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) - Q\Big(\frac{3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) \Big\} \\ &\quad + W^2 \int_{-W/2}^{W/2} \Big\{ Q\Big(\frac{W/2 - \tilde{x}}{\sigma_{x|y}}\Big) - Q\Big(\frac{3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) \Big\} d\tilde{x} \end{aligned} \tag{4.91}$$

By symmetry, the first and third integrals in (4.91) are equal, which yields

$$\begin{aligned} L &= 2W^2 \int_{-W/2}^{W/2} \Big\{ Q\Big(\frac{-3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) - Q\Big(\frac{-W/2 - \tilde{x}}{\sigma_{x|y}}\Big) \Big\} d\tilde{x} \\ &\quad + \frac{\Delta^2}{12} \int_{-W/2}^{W/2} \Big\{ Q\Big(\frac{-3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) - Q\Big(\frac{3W/2 - \tilde{x}}{\sigma_{x|y}}\Big) \Big\} \end{aligned} \tag{4.92}$$

$$\begin{aligned} &= 2W^2 \int_{-W/2}^{W/2} \Big(\frac{1}{\sqrt{2\pi}} \int_{W/2}^{3W/2} e^{-\frac{(x-\tilde{x})^2}{2\sigma_{x|y}^2}} dx \Big) d\tilde{x} \\ &\quad + \frac{\Delta^2}{12} \int_{-W/2}^{W/2} \Big(\frac{1}{\sqrt{2\pi}} \int_{-3W/2}^{3W/2} e^{-\frac{(x-\tilde{x})^2}{2\sigma_{x|y}^2}} dx \Big) d\tilde{x} \end{aligned} \tag{4.93}$$

Given that W is several multiples of a $\sigma_{x|y}$, we simplify (4.93) by the very accurate approximation

$$L \approx 2W^2 \int_{-\infty}^{W/2} \Big(\frac{1}{\sqrt{2\pi}} \int_{W/2}^{\infty} e^{-\frac{(x-\tilde{x})^2}{2\sigma_{x|y}^2}} dx \Big) d\tilde{x} + \frac{\Delta^2}{12} \int_{-W/2}^{W/2} d\tilde{x} \tag{4.94}$$

$$= 2W^2 \int_{-\infty}^{W/2} \Big(\frac{1}{\sqrt{2\pi}} \int_{W/2}^{\infty} e^{-\frac{(x-\tilde{x})^2}{2\sigma_{x|y}^2}} dx \Big) d\tilde{x} + \frac{\Delta^2}{12} W \tag{4.95}$$

which we will assume to be equality. In the integral expression in (4.95), we let $W = G\sigma_{x|y}$ and

perform the change of variables

$$v = \frac{\tilde{x}}{\sigma_{x|y}} - \frac{G}{2} \tag{4.96}$$

and

$$u = \frac{x}{\sigma_{x|y}} - \frac{G}{2} - v, \tag{4.97}$$

to yield

$$\begin{aligned} L &= \frac{\Delta^2}{12}W + \frac{2W^2{\sigma_{x|y}}^2}{\sqrt{2\pi}} \int_{-\infty}^{0}\int_{-v}^{\infty} e^{\frac{-u^2}{2}} du\, dv & (4.98)\\ &= \frac{\Delta^2}{12}W + \frac{2W^2{\sigma_{x|y}}^2}{\sqrt{2\pi}}, & (4.99) \end{aligned}$$

where (4.99) follows from a change to polar coordinates. Finally combining (4.99) and (4.90), we have an expression for the distortion of the low-bits coding method:

$$\begin{aligned} J_{\text{LBC}} &= \left(\frac{\Delta^2}{12}W + \frac{2W^2{\sigma_{x|y}}^2}{\sqrt{2\pi}}\right) \sum_{i=0}^{2^{C-R}-1} f_{\tilde{x}}(x_{S_i}) & (4.100)\\ &= \frac{\Delta^2}{12} + \frac{2W{\sigma_{x|y}}^2}{\sqrt{2\pi}}, & (4.101) \end{aligned}$$

where in (4.101) we used

$$1 = \int_{-\infty}^{\infty} f_{\tilde{x}}\, dx \approx \sum_{i=0}^{2^{C-R}-1} f_{\tilde{x}}(x_{S_i})W. \tag{4.102}$$

Thus, we have shown that the distortion for the LBC system is a constant offset from the optimal NLSQ distortion, assuming W is relatively constant with respect to R. The coding gain γ of NLSQ over LBC is given by

$$\begin{aligned} \gamma &\triangleq \frac{J_{\text{LBC}}}{J_{\text{NLSQ}}} & (4.103)\\ &= \frac{\Delta^2}{12}\left(1 + \frac{24}{\sqrt{2\pi}}\sigma_{x|y}^2 2^{2R}\right)/\frac{\Delta^2}{12} & (4.104)\\ &= 1 + \frac{24}{\sqrt{2\pi}}\sigma_{x|y}^2 2^{2R} & (4.105) \end{aligned}$$

which is exponential in the rate R.

Chapter 5

MMSE Transform and Subband Coding

5.1 Introduction

In general, vector quantization is the only method that will approach the theoretical limit of rate-distortion performance for a systematic source coding system. Of course, performing systematic scalar quantization on each component of a length-N source $\mathbf{x}$ is a form of systematic VQ that imposes the constraint that the encoder decision regions are the union of several N-dimensional rectangular regions. SSQ has a low cost in terms of complexity, especially when implemented as an NLSQ, but as evidenced by the results of Chap. 4, in most cases SSQ is a suboptimal method of encoding in terms of rate-distortion performance. In the case of a highly correlated source $\mathbf{x}$, the performance of the SSQ is even more significantly impaired. In contrast, optimal SVQ will, in theory, approach the rate-distortion limit, but it can have overwhelming complexity, even for moderate dimension. In this chapter, we explore a middle ground between a pure SSQ system and optimal SVQ. We propose methods called *systematic block transform coding* (SBTC), for coding vectors , and *systematic subband coding* (SSC), for coding signals $x[n]$, that first transform the signal, *i.e.* perform a rotation of signal space, followed by component-wise SSQ of transformed signal. The rotation of signal space is equivalently viewed as a rotation of the encoder decision regions, into an orientation not allowed by straightforward SSQ. This method is especially useful for correlated sources and/or correlated channel output, in which case the SSQs can be "aligned"

with the correlation structure of the signals. Although not a general model for SVQ – the encoder decision regions are still unions of hyper-rectangles – significant coding gains relative to a mean-squared error criterion can be had by using SBTC over simple time-domain SSQ. Thus by using an easily implemented linear transformation and simple SSQ quantization structures, we create an effective SVQ method for the efficient, low-complexity encoding of a wide variety of sources.

Recalling our source coding paradigm stated in Chap. 1, the space-rotation, or redundancy-removal block in the system block diagram of Fig. 1-2 is the transform in this treatment. In Sec. 4.3 we discuss the advantages of Slepian-Wolf coding the quantizer outputs. after the transform. The analysis in this chapter concerning the optimization of the transform and SSQ bit allocation assumes no Slepian-Wolf coding after quantization, but this assumption does not preclude the use such codes in the system. For the remainder of the chapter, we will assume that no Slepian-Wolf coding is performed on the quantizer outputs. A full system implemented in practice may include all three components, a transform, a set of quantizers (usually SSQs) and a Slepian-Wolf code.

A compelling reason to adopt a block transform or subband coding approach to systematic source coding is that transform and subband coding have been applied with great success to many conventional source coding applications including audio coding [39], image coding [42, 81, 67], video coding [44], seismic processing, etc.. In the conventional scenario, the source is transformed, and the transform coefficients are quantized by conventional scalar quantizers. There are well-known results from conventional source coding regarding optimal bit allocation to the scalar quantizers, and selection of the optimal transform that have analogous incarnations for SBTC. In addition to providing significant coding gain, standard transform coding usually employs transforms that perform a time-frequency decomposition that affords the use of human perceptual models for applications such as audio or image coding. We will see that these same perceptual models can be used directly in many important SBTC and SSC systems.

Systematic transform coding can be applied to a wide variety of sources and channels for many important distortion metrics, but the main analytical results of this chapter rely on the assumption that $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian and that we use a squared distortion metric: $D(\mathbf{u}, \hat{\mathbf{u}}) = \|\mathbf{u} - \hat{\mathbf{u}}\|^2$. We therefore assume the Gaussian case throughout this chapter, and point out where certain results may be generalized to other cases. The chapter is arranged as follows. In the first portion of this chapter we consider the most elementary form of space rotation, multiplication of $\mathbf{x}$ by a

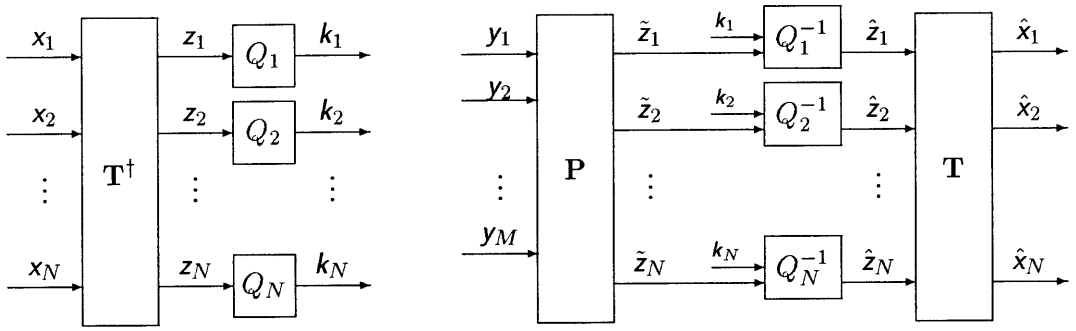
unitary transform $\mathbf{T}^\dagger$. The transform coefficients are coded by an optimal NLSQ encoder map so that these same coefficients are reconstructed at the decoder using $\mathbf{y}$ and the digital indices. The source is reconstructed by a simple inverse transformation. We derive the optimal transform and bit allocation strategies. We show the encouraging result that for a particular class of useful channel models, the optimal transform for SBTC is the same as the optimal transform for standard transform coding. For an even more general channel model, we show that the optimal transforms are the same in the case of a first-order autoregressive (AR-1) source process, in the limit of high correlation. Thus, for a broad class of sources and channels we can use the same fast transforms that have been developed for standard transform coding for SBTC. We extend the results for systematic block transform coding to systematic lapped transform coding.

For most coding applications the source is a discrete sequence $x[n]$, and not simply a finite length vector. Coding with block transforms, of course, can be applied to $x[n]$ simply by parsing the sequence into consecutive blocks, and applying $\mathbf{T}^\dagger$ to the blocks. A more general method for transforming $x[n]$ involves using a filter bank, which performs a subband decomposition of the source[80, 79]. Indeed block transforms are a special case of a filter bank. Using a filter bank and coding the subband signals with NLSQs, we have the basic structure for the systematic subband coding. We show that the optimal bit-allocation strategy is inverse-waterpouring on the analog estimation error variances of the subband signals. We show in the theoretical case of infinite length filters, two conditions that are both necessary and sufficient for a filter bank to be optimal for SSC. For the Gaussian channel – convolutional distortion and additive Gaussian noise – we show the equivalence of the optimal transforms for SSC and standard source coding. We also discuss the form of the optimal filter bank for finite length filters, which has the property of optimal error energy compaction.

5.2 Block transform coding

5.2.1 Basic structure

The structure of a systematic block transform coder is shown in Fig. 5-1. Fundamentally, what is depicted is the encoding of transform coefficients z_i at the encoder with an NLSQ encoder map,



(a) Encoder (b) Decoder

Figure 5-1: Systematic block transform block diagram. (a) Encoder. Matrix $\mathbf{T}^\dagger$ is unitary transform. Q_i are NLSQ encoder maps. (b) Decoder. Matrix $\mathbf{P}^\dagger = \mathbf{\Lambda_{zy}\Lambda_y}^{-1} = \mathbf{T}^\dagger\mathbf{\Lambda_{xy}\Lambda_y}^{-1}$ is analog estimation matrix. Q_i^{-1} are NLSQ decoder maps. $\mathbf{T}$ is the inverse transform.

and the MMSE reconstruction of those transform coefficients at the decoder, using the NLSQ encoder outputs and the analog observation $\mathbf{y}$, from which the vector estimate $\hat{\mathbf{x}}$ is formed by inverse transform. The encoder, Fig. 5-1(a), is comprised of a unitary transform $\mathbf{T}^\dagger$ acting on the source $\mathbf{x}$ by the equation, $\mathbf{z} = \mathbf{T}^\dagger\mathbf{x}$, and a set of NLSQ encoder functions Q_i. By the linearity of $\mathbf{T}^\dagger$, we have that $\mathbf{z}$ and $\mathbf{y}$ are also jointly Gaussian. Note that since there are multiple quantizers, we must index the quantizer output levels by $i = 1, ..., N$, forming the vector $\mathbf{k}$. A unitary transform is one for which $\mathbf{T}^\dagger\mathbf{T} = \mathbf{I}$. We have chosen to use $\mathbf{z} = \mathbf{T}^\dagger\mathbf{x}$ as opposed to $\mathbf{z} = \mathbf{T}\mathbf{x}$ to ensure that the i^{th} transform coefficient $z_i = \mathbf{t}_i^\dagger\mathbf{x}$, is the projection of $\mathbf{x}$ onto $\mathbf{t}_i$, the i^{th} column vector of $\mathbf{T}$. We use the notation $Q_i(.)$ for the quantizer encoding function, denoted by $f(.)$ in Chap. 4, to be consistent with the notation from many publications from standard transform coding.

The decoder is shown in Fig 5-1(b) as the composition of an analog estimation block $\mathbf{P}$, a set of NLSQ decoder functions, and an inverse transform $\mathbf{T}$. We use the notation Q_i^{-1} for the quantizer decoder map – obviously Q_i^{-1} is not a true inverse of Q_i – whose definition is slightly modified from the definition of $\mathbf{g}$ in Chap. 4. Given $\mathbf{y}$ and $\mathbf{k}$, the decoder function $\mathbf{g}$ takes the MMSE estimate $\tilde{\mathbf{z}}(\mathbf{y})$ and for all i maps z_i to the nearest lattice point in $\mathcal{L}_{k_i}$. In Fig. 5-1 $\mathbf{g}$ is shown as the composition of $\mathbf{P}$ and the block of Q^{-1}s. Clearly, from this definition

$$\mathbf{P} = \mathbf{\Lambda_{zy}\Lambda_y}^{-1} = \mathbf{T}^\dagger\mathbf{\Lambda_{xy}\Lambda_y}^{-1}, \tag{5.1}$$

the linear operator for the MMSE estimation of $\mathbf{z}$. The analog estimate is given by $\tilde{\mathbf{z}} = \mathbf{P}\mathbf{y} = \mathbf{T}^\dagger\tilde{\mathbf{x}}$,

where $\tilde{\mathbf{x}}$ is the MMSE estimate of the source given $\mathbf{y}$. We define $Q^{-1} : \mathcal{K} \times \tilde{\mathcal{Z}} \rightarrow \hat{\mathcal{Z}}$ as the function which maps $\tilde{z}$ to the nearest lattice point in $\mathcal{L}_k$. The transform $\mathbf{T}$ takes the reconstructed coefficients $\hat{\mathbf{z}}$ into the original source domain, by using the transform that is the inverse to that used at the encoder. The inverse is given by $\mathbf{T}$ because the transform is unitary. Consistent with conventional block transform coding principles, we constrain the transform $\mathbf{T}^\dagger$ to be unitary in order to exploit the norm preservation property: for any random vector $\mathbf{u}$, and $\mathbf{v} = \mathbf{T}^\dagger\mathbf{u}$, we have

$$E[\mathbf{v}^\dagger\mathbf{v}] = E[\mathbf{u}^\dagger\mathbf{T}\mathbf{T}^\dagger\mathbf{u}] = E[\mathbf{u}^\dagger\mathbf{u}]. \tag{5.2}$$

Thus if we define the error vectors

$$\mathbf{e_x} = \mathbf{x} - \tilde{\mathbf{x}} \tag{5.3}$$

and

$$\mathbf{e_z} = \mathbf{z} - \tilde{\mathbf{z}} = \mathbf{T}^\dagger(\mathbf{e_x}), \tag{5.4}$$

we see that the norm preserving property implies that the MSE in the transform domain, $E[\|\mathbf{e_z}\|^2]$, is the same as the MSE in the time domain, $E[\|\mathbf{e_x}\|^2]$. Another attribute of unitary matrices that will be of value to us is that the determinant of a unitary $\mathbf{T}$ is 1, and hence multiplication of another matrix by $\mathbf{T}$, has no effect on the determinant.

5.2.2 Quantizer characteristic

The results of this chapter are strongly dependent on the distortion-rate characteristic of the NLSQ under the high-resolution assumption, given by (4.42) for the Gaussian case. The i^{th} transform coefficient z_i is quantized by a given SSQ, with a given rate R_i. We define the variance

$$\sigma_i^2 = \sigma^2_{z_i|\mathbf{y}}. \tag{5.5}$$

Assuming that the NLSQ is optimized for the density $f_{z_i\mathbf{y}}(z_i, \mathbf{y})$, we have from (4.42) that

$$d_i = \theta\sigma_i^2\, 2^{-2R_i}, \tag{5.6}$$

where $\theta = 16/3$. This value of θ assumes that the NLSQ has a staircase width $W = 8\sigma_{x|y}$, which is near optimal for bit rates under 7 bits per sample. As a basis for comparison, it is well-known that for standard block transform coding, quantization distortion for the i^{th} subband is given by [40]

$$d_i = \phi \sigma_{z_i}^2 \, 2^{-2R_i}, \tag{5.7}$$

where ϕ is a constant that varies according to the type of quantizer being employed and the source density. If a uniform quantizer is used, which is often the case, ϕ is, independent of distribution, $H^2/12$ where $H\sigma_{z_i}$ is the width of the region of support of the quantizer. If the Max quantizer is used, the value of ϕ depends on the coefficient density $f_{z_i}(z_i)$, which can be calculated by a high rate integral approximation [31]. For the Gaussian case $\phi = \sqrt{3}\pi/2$. The equality in (5.7) is much more general than (5.6) in that it applies to all distributions, not just Gaussian.

Non-Gaussian case

Although not as general as (5.7), equation (5.6) is applicable to more than just the Gaussian case. For example consider just a single coefficient $\boldsymbol{z}$ and an additive channel, such that $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{n}$, with iid noise $\boldsymbol{n}$ independent of $\boldsymbol{z}$, and an *arbitrary* known distribution. We can design a simple, suboptimal scheme that has the same characteristic as (5.6). The encoder is an NLSQ encoder map, with W chosen to be several times σ_n, enough to support a vast majority of the density of $\boldsymbol{n}$. The decoder is a simplified NLSQ decoder map, which maps the unmodified $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{n}$ to the nearest lattice point $\mathcal{L}_k$. For a large enough $W = L\sigma_n$, the decoder will select the correct cell from which $\boldsymbol{z}$ originated with high probability. It is straightforward to verify that in this case for a given rate R,

$$d = \kappa \sigma_{\boldsymbol{n}}^2 \, 2^{-2R}. \tag{5.8}$$

where $\kappa = L^2/12$. We will see that optimal bit allocation to transform coefficients can be stated easily for all cases in which the distortion-rate function can be expressed in a similar manner.

5.3 Optimal bit allocation

5.3.1 Problem Description

Given a set of coefficients, z_i, $i = 1, ..., N$, our challenge is to code the z_i, each with an SSQ, with a fixed bit allocation, for minimum MSE, assuming $\mathbf{y}$ is at the decoder. The formal description of the problem is, under the constraint

$$\sum_{i=1}^{N} R_i = R, \tag{5.9}$$

minimize the objective function

$$J = \sum_{i=1}^{N} d_i = \sum_{i=1}^{N} \theta \sigma_i^2 \, 2^{-2R_i}, \tag{5.10}$$

with respect to the R_is. By the norm preserving property of $\mathbf{T}$, J is the same as the MSE of the reconstructed source. In all respects except for the terms σ_i^2 in place of $\sigma_{z_i}^2$, this is the same problem formulation as the standard optimal bit allocation problem first addressed in [37, 38]; these terms σ_i and σ_{z_i} are merely constant parameters in the optimization. The classical solution of the standard bit allocation problem [38] involves Lagrangian optimization, but a simpler one offered in [31] is more straightforward, based on the elementary arithmetic/geometric mean inequality. These techniques suffer from two primary weaknesses. First the solution for the R_is will be real and the number of bits per quantizer is an integer value. Secondly, a solution that minimizes the objective function may yield a negative value for R_i for some i.[1] In spite of these weaknesses, the optimal bit allocation result has been applied to conventional source coding with great success.

5.3.2 Optimal strategy

Given that we have the same classical formulation, but for a change of parameters, we state the optimal bit allocation result for systematic scalar quantization by using the standard bit allocation result [31] and replacing σ_{z_i} with σ_i. Defining

$$\bar{R} = R/N, \tag{5.11}$$

[1] See [65] for an optimization technique that averts this issue.

the average number of bits per coefficient and

$$\beta^2 = \Big(\prod_{i=1}^{N} \sigma_i^2\Big)^{\frac{1}{N}}, \tag{5.12}$$

the geometric mean of the analog estimation error variances, the optimal assignment of bits to the i^{th} SSQ is

$$R_i = \bar{R} + \frac{1}{2}\log_2 \frac{\sigma_i^2}{\beta^2}. \tag{5.13}$$

The resultant minimum MSE attained is

$$J = N\theta\beta^2 2^{-2\bar{R}}. \tag{5.14}$$

It is easily verified that the distortion the same across all quantizers, *i.e.*,

$$d_i = \theta\beta^2 2^{-2\bar{R}}, \tag{5.15}$$

implying that an optimal bit allocation strategy is to inverse waterpour on the analog estimation variances of the transform coefficients. In other words, for coefficients where the analog estimation from $\mathbf{y}$ will yield a poor estimate of the transform coefficient, the optimal algorithm will give more bits than to a coefficient that is well estimated. This is in contrast to the classical waterpouring result which calls for inverse waterpouring on the variances of the transform coefficients themselves. Note for some source/channel characteristics, the optimal bit allocation strategy for systematic and classical coding will differ dramatically.

5.3.3 Non-Gaussian case

It is straightforward infer, by observing the distortion-rate characteristic of equation (5.8), the optimal bit allocation strategy for using the suboptimal NLSQ in the non-Gaussian case as discussed in Sec. 5.2.2. The optimal strategy is to inverse waterpour on the noise variances σ_{n_i}. For many channels, such as the additive white noise channel, the σ_{n_i} will be the same, which implies that the optimal encoder gives the same number of bits to each transform coefficient. In this case, it is easily verified that the total MSE is not improved at all by taking a transform before applying the SSQ,

making this method ineffective. However, in the case where the noise has different variances across coefficients, a coding gain can be attained. As indicated by the next section, transform coding gain can also be obtained in the non-Gaussian case and the additive white noise case if the objective function is a weighted sum of coefficient distortion.

5.3.4 Weighted distortion measure

We consider the weighted objective function J_{w} under the usual constraint (5.9):

$$J_{\mathrm{w}} = \sum_{i=1}^{N} w_i d_i = \sum_{i=1}^{N} \theta w_i \sigma_i^2 \, 2^{-2R_i}. \tag{5.16}$$

In this case, again using the standard result, the optimization yields

$$R_i = \bar{R} + \frac{1}{2}\log_2 \frac{\sigma_i^2}{\beta^2} + \frac{1}{2}\log_2 \frac{w_i}{\omega^2}, \tag{5.17}$$

where ω^2 is the geometric mean of the weight values w_i. The overall MSE is given by

$$J_{\mathrm{w}} = N\theta\omega^2\beta^2 2^{-2\bar{R}}. \tag{5.18}$$

Thus the optimal bit allocation strategy is to inverse waterpour on the weighted analog estimation variances $w_i\sigma_i^2$. This result will prove useful when coding with weights assigned according to human perceptual sensitivities.

5.4 Optimal Transform

5.4.1 Derivation

Given the optimal bit allocation strategy, and resulting overall MSE, we derive the optimal $\mathbf{T}$ to minimize J in (5.14). It is well-known from standard coding that the optimal transform for MMSE coding is the Karhunen-Loeve (KL) transform of the source, or the matrix $\mathbf{S}^\dagger$ which diagonalizes the covariance matrix of the source,

$$\mathbf{\Lambda_x} = E[\mathbf{x}\mathbf{x}^\dagger]. \tag{5.19}$$

We will see that the optimal transform is that which diagonalizes the error covariance matrix

$$\Lambda_{\mathbf{e_x}} = \mathbf{\Lambda_x} - \mathbf{\Lambda_{xy}}\mathbf{\Lambda_y}^{-1}\mathbf{\Lambda_{xy}}, \tag{5.20}$$

where $\mathbf{e_x}$ is defined in (5.3). The proof of the optimal $\mathbf{T}$ for SBTC follows closely that for standard coding.

Using (5.4), we see that

$$\begin{aligned} \mathbf{\Lambda_{e_z}} &= E[\mathbf{T}^\dagger \mathbf{e_x} \mathbf{e_x}^\dagger \mathbf{T}] && (5.21) \\ &= \mathbf{T}^\dagger \mathbf{\Lambda_{e_x}} \mathbf{T} && (5.22) \end{aligned}$$

which is the error covariance when the analog observation is used to estimate the transformed vector $\mathbf{z}$. Thus if we estimate a transformed vector $\mathbf{T}^\dagger\mathbf{x}$ from $\mathbf{y}$ for minimum MSE, the error covariance matrix can indeed be made diagonal by selecting $\mathbf{T}^\dagger = \mathbf{U}^\dagger$ for

$$\mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 ... \mathbf{u}_N], \tag{5.23}$$

where $\mathbf{u}_i$, $i = 1, ..., N$ are eigenvectors of $\mathbf{\Lambda_{e_x}}$. We assume that the $\mathbf{u}_i$ are such that their corresponding eigenvalues λ_i are ordered:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq 0. \tag{5.24}$$

Because $\mathbf{T}$ is unitary, it is clear from (5.22) that

$$\det\mathbf{\Lambda_{e_x}} = \det\mathbf{\Lambda_{e_z}}. \tag{5.25}$$

The proof of the optimal $\mathbf{T}^\dagger$ relies on two well-known facts. First, the determinant of a matrix equals the product of its eigenvalues:

$$\det\mathbf{\Lambda_{e_x}} = \prod_{i=1}^{N} \lambda_i. \tag{5.26}$$

The second fact is that for any autocorrelation matrix $\mathbf{\Lambda_{e_x}}$ of a real valued random vector $\mathbf{e_x}$ with

zero mean and variances σ_i^2, we have,

$$\det\mathbf{\Lambda}_{\mathbf{e}_z} \leq \prod_{i=1}^{N} \sigma_i. \tag{5.27}$$

Thus, from (5.14), (5.25), (5.26), and (5.27) we have the following chain of inequalities:

$$\begin{aligned} J &= N\theta\Big(\prod_{i=1}^{N} \sigma_i^2\Big)^{\frac{1}{N}} 2^{-2\bar{R}} && (5.28)\\ &\geq N\theta\Big(\det\mathbf{\Lambda}_{\mathbf{e}_x}\Big)^{\frac{1}{N}} 2^{-2\bar{R}} && (5.29)\\ &= N\theta\Big(\prod_{i=1}^{N} \lambda_i^2\Big)^{\frac{1}{N}} 2^{-2\bar{R}}, && (5.30) \end{aligned}$$

which are met with equality when $\mathbf{\Lambda}_{\mathbf{e}_z}$ is diagonal, done by setting $\mathbf{T} = \mathbf{U}$ in (5.23).

5.4.2 Coding Gain

It is useful to quantify the gain achieved by scalar coding transformed coefficients over simply scalar coding the time-domain samples of $\mathbf{x}$. The coding gain, Γ, is defined as the ratio of $J_{\text{no transform}}$, the MSE of the system that simply uses an optimal NLSQ and optimal bit allocation on the samples of $\mathbf{x}$, to $J_{\text{opt SBTC}}$, the MSE of the SBTC system that uses the optimal transform and optimal bit allocation. It is clear that

$$J_{\text{no transform}} = N\theta\Big(\prod_{i=1}^{N} \sigma_{x_i|\mathbf{y}}^2\Big)^{\frac{1}{N}} 2^{-2\bar{R}}, \tag{5.31}$$

which yields the coding gain

$$\begin{aligned} \Gamma &= \frac{J_{\text{no transform}}}{J_{\text{opt SBTC}}} && (5.32)\\ &= \frac{\Big(\prod_{i=1}^{N} \sigma_{x_i|\mathbf{y}}^2\Big)^{\frac{1}{N}}}{\Big(\det\mathbf{\Lambda}_{\mathbf{e}_x}\Big)^{\frac{1}{N}}}, && (5.33) \end{aligned}$$

a quantity always greater than 1 by (5.27). As will be discussed in Sec. 5.6, when the source $\mathbf{x}$ is a block from a stationary process $x[n]$ and the analog channel output observation is a process $y[n]$ (jointly stationary with $x[n]$), with noncausal observations ($n = -\infty, ..., \infty$) or causal observations

$(n = -\infty, ..., n_0 + N)$, the analog estimation error $\mathbf{e}_x$ is stationary, implying that $\sigma^2_{x_i|\mathbf{y}} = \sigma^2_{x|\mathbf{y}}$, a constant for all i. In this case

$$\mathrm{Tr}\Big(\mathbf{\Lambda}_{\mathbf{e}_x}\Big) = N\sigma^2_{x|\mathbf{y}} = \sum_{i=1}^{N} \lambda_i^2, \tag{5.34}$$

which allows us to simplify (5.33) to be

$$\Gamma = \frac{\Big(\prod_{i=1}^{N} \sigma^2_{x|y}\Big)^{\frac{1}{N}}}{\Big(\prod_{i=1}^{N} \lambda_i\Big)^{\frac{1}{N}}} \tag{5.35}$$

$$= \frac{\sigma^2_{x|y}}{\Big(\prod_{i=1}^{N} \lambda_i\Big)^{\frac{1}{N}}} \tag{5.36}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N} \lambda_i}{\Big(\prod_{i=1}^{N} \lambda_i\Big)^{\frac{1}{N}}}. \tag{5.37}$$

Hence, the coding gain is the ratio of the arithmetic mean to the geometric mean of the eigenvalues of $\mathbf{\Lambda}_{\mathbf{e}_x}$. The coding gain in standard transform coding is, analogously, the ratio of the arithmetic to the geometric mean of $\mathbf{\Lambda}_{\mathbf{x}}$.

A different notion of coding gain is obtained when we compare optimal systematic transform coding to optimal standard transform coding, or in other words compare the optimal systems that use side information at the decoder versus those that ignore it. The optimal MSE of a standard coding system is [34]

$$J_{\mathrm{std}} = N\theta_{\mathrm{std}}\Big(\prod_{i=1}^{N} \alpha_i\Big)^{1/N} 2^{-2\bar{R}}, \tag{5.38}$$

where the α_i, $i = 1, ..., N$ are the eigenvalues of $\mathbf{\Lambda}_x$, and $\theta_{\mathrm{std}} = \sqrt{3}\pi/2$ if we use Max quantizers on the transform coefficients; if we use uniform quantizers in the standard codec, $\theta_{\mathrm{std}} \approx \theta$. The coding gain from using side information, Γ_{SI}, is the ratio of optimal error attained not using side information to the optimal error attained using side information:

$$\Gamma_{\mathrm{SI}} = \frac{J_{\mathrm{std}}}{J_{\mathrm{opt\ SBTC}}} \tag{5.39}$$

$$= \frac{\theta_{\mathrm{std}}}{\theta}\Big(\prod_{i=1}^{N} \frac{\alpha_i}{\lambda_i}\Big)^{1/N} \tag{5.40}$$

The coding gain in (5.40) should be greater than one for moderate to high SNRs, the regime for which NLSQs are optimal. In Sec. 5.5 we consider a special case in which the coding gain simplifies to a form that gives us a simple rule of thumb, in terms of SNR, that can be used to assess the gains to be had by using $\mathbf{y}$ at the decoder.

5.5 Special case: "graphic equalizer" channel

Over the many decades of interest in conventional digital source coding, there have been derived many transforms that closely approximate the KL transform for many sources of interest, from cosine-modulated transforms like the discrete cosine transform (DCT) [62], to multi-resolution wavelet transforms [84, 51]. Much effort has been made to create algorithms that execute these near optimal transforms with low computational complexity. Thus, it is in our interest to leverage these past efforts to derive fast, near optimal transforms for SBTC. In this section we show that for a broad class of channels, the optimal transform for standard coding is the optimal transform for SBTC, affording us the privilege to use well-known fast transform algorithms for coding.

5.5.1 Model for convolution

We denote the matrix that diagonalizes $\mathbf{\Lambda_x}$, the KL transform of the source, by

$$\mathbf{V} = [\mathbf{v}_1\mathbf{v}_2...\mathbf{v}_N], \tag{5.41}$$

where $\mathbf{v}_i$ is the eigenvector of $\mathbf{\Lambda_x}$ corresponding to the eigenvalue α_i. We assume the ordering

$$\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_N \geq 0 \tag{5.42}$$

Consider the following channel description. Defining $\mathbf{\Sigma}_h = \text{diag}\{h_1, ..., h_N\}$ and $\mathbf{\Sigma}_g = \text{diag}\{g_1, ..., g_N\}$, for some $\{h_1, ..., h_N\}$ and $\{g_1, ..., g_N\}$, we have

$$\mathbf{y} = \mathbf{H}^\dagger\mathbf{x} + \mathbf{G}^\dagger\mathbf{n}, \tag{5.43}$$

where

$$\mathbf{H}^\dagger = \mathbf{V}\mathbf{\Sigma}_h\mathbf{V}^\dagger, \tag{5.44}$$

and

$$\mathbf{G}^\dagger = \mathbf{V}\boldsymbol{\Sigma}_g\mathbf{V}^\dagger, \tag{5.45}$$

where $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$ is uncorrelated with $\mathbf{x}$. Letting $\nu = \mathbf{G}^\dagger\mathbf{n}$, we have

$$\boldsymbol{\Lambda}_\nu = \mathbf{G}^\dagger\mathbf{G} = \mathbf{V}\boldsymbol{\Sigma}_g^2\mathbf{V}^\dagger. \tag{5.46}$$

The matrices $\mathbf{H}^\dagger$,$\mathbf{G}^\dagger$, and $\boldsymbol{\Lambda}_\nu$ have the same spectral decomposition as $\boldsymbol{\Lambda}_\mathbf{x}$. Note that (5.43) models AWGN channels exactly.

For many classes of sources, a near-optimal transform decomposes the source into component frequency bins, effectively filtering the source by a set of narrowband bandpass filters. The aggregated passbands of all the filters cover the bandwidth of the source. The passband of one filter is essentially in the stopband of all of the other filters. An important one of these transforms is the DCT [62], whose basis functions are given by

$$t_{nk} = c(k)\sqrt{\frac{2}{N}}\cos\left[\left(n+\frac{1}{2}\right)\frac{(k)\pi}{N}\right], \quad n = 0, ..., N-1, \; k = 0, ..., N-1 \tag{5.47}$$

where t_{nk} is the n^{th} element of the vector $\mathbf{t}_k$, a column vector of $\mathbf{T}$, and

$$c(k) = \begin{cases} 1/\sqrt{2} & \text{if } k = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{5.48}$$

The DCT is one of a class of useful transforms with cosine-modulated basis functions [2], which takes a basic lowpass filter structure (a rectangular window in the case of the DCT) and modulates it by cosine functions, harmonically separated by π/N radians. The effect is to decompose the source by bandpass filters, of approximate passband bandwidth π/N, spaced by π/N radians from one another. Of course, these filters are not perfect bandpass filters, as the window can have significant sidelobe energy in the frequency-domain. The DCT is the KL transform for a first-order autoregressive (AR-1) process in the limit of high intersample correlation [16], which makes it useful for sources with slowly varying backgrounds, like natural images.

We will assume that the DCT or other "frequency-decomposing" transform is a KL basis for

[2] In general we can consider sine-modulated transforms, but we restrict our attention to cosine-modulated transforms

the source. Interpreting the transform as a bank of bandpass filters, we can then view (5.43) as approximating a channel with convolutional distortion and additive colored noise. Consider the action of $\mathbf{H}^\dagger = \mathbf{V}\mathbf{\Sigma}_h\mathbf{V}^\dagger$. The matrix $\mathbf{V}^\dagger$ projects $\mathbf{x}$ onto the i^{th} eigenvector $\mathbf{t}_i$, for all i. The matrix $\mathbf{\Sigma}_h$ then scales the i^{th} projection by h_i, for all i. The values h_i multiply a narrowband signal in the frequency domain, thereby approximating convolution in the frequency domain. The final matrix $\mathbf{V}$, weights each eigenvector by the scaled projections, and sums them to bring the signal back into the time domain. The matrix $\mathbf{H}^\dagger$ essentially performs a "graphic equalizer" approximation to convolution. Similarly, the action of $\mathbf{G}^\dagger$ on the white noise $\mathbf{n}$ is approximately that of convolution, thereby coloring the noise in a virtually arbitrary manner. Our approximations are further justified by the fact that the DCT has been shown to approximate cyclic convolution [19] for real, even filters.

The restriction to real, even filters in [19] highlights one caveat to our approximation. The DCT and other cosine-modulated transforms are real transforms. Thus, a particular basis function implement a narrowband, bandpass filter that is *two-sided* and *real.* Scaling the projection onto this basis function by a (real) gain, allows us only to modify the magnitude of the signal and not the phase. Thus we can only approximate real-even filters with our model, if a cosine-modulated basis is the KL transform for the source. In contrast, the discrete Fourier transform (DFT) has basis functions that implement one-sided, complex narrowband bandpass filters, and thus a complex gain can be applied to modify both magnitude and phase. The phase problem associated with cosine-modulated transforms is often not an issue in coding sources, like audio, that are perceptually resistant to phase distortion.

5.5.2 Equivalence of optimal transforms

We define $\mathbf{\Sigma}_\alpha = \text{diag}\{\alpha_1, ..., \alpha_N\}$. For the channel model given in (5.43), we write out the error covariance matrix

$$\begin{aligned}
\mathbf{\Lambda}_{\mathbf{e}_\mathbf{x}} &= \mathbf{\Lambda}_\mathbf{x} - \mathbf{\Lambda}_{\mathbf{xy}}\mathbf{\Lambda}_\mathbf{y}^{-1}\mathbf{\Lambda}_{\mathbf{xy}}^\dagger && (5.49)\\
&= \mathbf{\Lambda}_\mathbf{x} - \mathbf{\Lambda}_\mathbf{x}\mathbf{H}(\mathbf{H}^\dagger\mathbf{\Lambda}_\mathbf{x}\mathbf{H} + \mathbf{\Lambda}_\nu)^{-1}\mathbf{H}^\dagger\mathbf{\Lambda}_\mathbf{x} && (5.50)\\
&= \mathbf{V}\mathbf{\Sigma}_\alpha\mathbf{V}^\dagger - \mathbf{V}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h\mathbf{V}^\dagger(\mathbf{V}(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h^2 + \mathbf{\Sigma}_g^2)\mathbf{V}^\dagger)^{-1}\mathbf{V}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h\mathbf{V}^\dagger && (5.51)\\
&= \mathbf{V}(\mathbf{\Sigma}_\alpha - \mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h^2 + \mathbf{\Sigma}_g^2)^{-1}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h)\mathbf{V}^\dagger && (5.52)
\end{aligned}$$

$$= \mathbf{V}\text{diag}\{\lambda_1, ..., \lambda_N\}\mathbf{V}^\dagger, \tag{5.53}$$

where

$$\lambda_i = \alpha_i - \frac{\alpha_i^2 h_i^2}{\alpha_i h_i^2 + g_i^2} \tag{5.54}$$

$$= \frac{\alpha_i g_i^2}{\alpha_i h_i^2 + g_i^2} \quad i = 1, ..., N \tag{5.55}$$

are the eigenvalues of $\mathbf{\Lambda_{e_x}}$. Thus the matrix $\mathbf{V}^\dagger$ that diagonalizes $\mathbf{\Lambda_x}$ also diagonalizes $\mathbf{\Lambda_{e_x}}$. The optimal transform for standard transform coding is optimal for systematic transform coding for the general class of channel models given in (5.43).

5.5.3 Low-complexity decoder structure

There are some additional consequences of the special structure in (5.43), which suggest a form for the decoder that has remarkably low complexity. In addition to diagonalizing $\mathbf{\Lambda_{e_x}}$, $\mathbf{V}^\dagger$ also diagonalizes

$$\mathbf{\Lambda_{xy}} = \mathbf{V}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h\mathbf{V}^\dagger, \tag{5.56}$$

and

$$\mathbf{\Lambda_y} = \mathbf{V}(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h^2 + \mathbf{\Sigma}_g^2)\mathbf{V}^\dagger. \tag{5.57}$$

Using $\mathbf{z} = \mathbf{V}^\dagger\mathbf{x}$, the MMSE estimate of $\mathbf{z}$ from $\mathbf{y}$ is given by

$$\tilde{\mathbf{z}} = \mathbf{\Lambda_{zy}}\mathbf{\Lambda_y}^{-1}\mathbf{y} \tag{5.58}$$

$$= \mathbf{V}^\dagger\mathbf{V}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h\mathbf{V}^\dagger(\mathbf{V}(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_h^2 + \mathbf{\Sigma}_g^2)\mathbf{V}^\dagger)^{-1}\mathbf{y} \tag{5.59}$$

$$= \text{diag}\{\zeta_1, ..., \zeta_N\}\mathbf{V}^\dagger\mathbf{y}, \tag{5.60}$$

where

$$\zeta_i = \frac{\alpha_i h_i}{\alpha_i h_i^2 + g_i^2}. \tag{5.61}$$

The decoder analog estimation stage as described by (5.60) is accomplished as follows. The same transform $\mathbf{V}^\dagger$ that is used at the encoder on $\mathbf{x}$ is applied to the analog estimate to form

$$\mathbf{s} = \mathbf{V}^\dagger \mathbf{y}. \tag{5.62}$$

The estimate $\tilde{z}_i$ of the i^{th} transform coefficient simply a scalar operation on s_i:

$$\tilde{z}_i = \zeta_i s_i. \tag{5.63}$$

The decoder is of low complexity because the transform $\mathbf{V}^\dagger$ that acts on $\mathbf{y}$ can be implemented with efficient algorithms, such as that which calculates the DCT.

5.5.4 Coding Gain

Using (5.55) the coding gain Γ_{SI}, which compares coding performance with and without side information is

$$\Gamma_{\text{SI}} = \frac{\theta_{\text{std}}}{\theta}\Big(\prod_{i=1}^{N}(1+\frac{\alpha_i h_i^2}{g^2})\Big)^{1/N}, \tag{5.64}$$

$$= \frac{\theta_{\text{std}}}{\theta}\Big(\prod_{i=1}^{N}(1+h_i^2\text{SNR}_i)\Big)^{1/N}, \tag{5.65}$$

where SNR_i is the signal to noise ratio in the i^{th} frequency bin. We assume that for standard coding, uniform quantizers are used, and $\theta_{\text{std}} = \theta$. Assuming $h_i = 1$ (no filtering) and $g_i = \sigma_n^2$ (white noise), in (5.65) we see that the coding gain is proportional to the geometric mean of the terms $1 + \text{SNR}_i$. In the limit high SNR the 1 in each of these terms becomes negligible, and we have a further simplification:

$$\Gamma_{\text{SI}} = \Big(\prod_{i=1}^{N}\text{SNR}_i\Big)^{1/N} \tag{5.66}$$

$$= \frac{\sigma_x^2}{\sigma_n^2}\frac{\Big(\prod_{i=1}^{N}\alpha_i\Big)^{1/N}}{\sigma_x^2} \tag{5.67}$$

$$= \frac{\text{SNR}}{\Gamma_{\text{std}}}, \tag{5.68}$$

where Γ_{std} is the well-known standard coding gain, realized from standard transform coding over straight PCM, *i.e.,* the ratio of the arithmetic mean to the geometric mean of the eigenvalues of $\mathbf{\Lambda_x}$; SNR is the average signal to noise ratio. In the limit of high SNR, then, the improvement of systematic coding over standard coding is linear in the average SNR, with the slope of the increase equaling the reciprocal of the standard coding gain.

5.5.5 Overhead information for locally stationary sources

For many coding applications, the source is not accurately modeled as simply a stationary process with known second-order statistics. A better model for most sources is one in which the process is locally stationary, varying slowly with time in a fashion that is not known apriori at the decoder. For such a model applied to SBTC in its most general case, overhead bits must be transmitted to the decoder in order to communicate the autocovariance matrix of the source. Making the same assumptions about the local source and channel statistics as in Sec. 5.5, however, we see from (5.60) and (5.61) that the only required overhead are the α_i, $i = 1, ..., N$, the eigenvalues of $\mathbf{\Lambda_x}$, or the variances of the transform coefficients z_i. Thus the overhead is reduced from the description $N(N-1)/2$ coefficients to just N coefficients. Furthermore, in many cases the α_i can be described efficiently in a parametric manner, through the use of such quantities as linear predictive coefficients [40].

5.5.6 Cosine transforms and general convolutional distortion

In the case where $\mathbf{V}^\dagger$, the near-optimal transform for standard coding, is a cosine-modulated transform, as discussed in Sec. 5.5, (5.43) is only accurate for modeling real, even convolutional distortion. For many real systems the convolutional distortion will not be real and even. In these cases, although we do not prove this fact, $\mathbf{V}^\dagger$ may be a good transform to use for systematic encoding for reasons of computational efficiency and the removal of source redundancy. In this section we make three main assumptions, that 1) $\mathbf{T} = \mathbf{V}$, 2) the impulse response of the channel is not real and even, and 3) as before there is also additive colored noise distortion. In addition, we assume that the magnitude response of the channel can be considered flat across the positive-frequency passband and flat across the negative-frequency passband of any filter defined by a column of $\mathbf{T}$. Thus, we assume the same channel model as in (5.43), with $\mathbf{H}$ less constrained than

in (5.44). Because we are only interested in second order statistics, we can maintain an accurate model of colored noise by still assuming that the matrix $\mathbf{G}$ is constrained as in (5.45). The MMSE estimate of $\mathbf{z} = \mathbf{T}^\dagger \mathbf{x}$ from $\mathbf{y}$ is analytical, easily obtained from the standard formulas. However, for the sake of computational efficiency, at the decoder we would like to avoid an operation involving multiplication by an arbitrary matrix (N^2 multiplies). We desire a low-complexity decoder such as the one in (5.60). One could use a structure like (5.60) in this case, involving the operation $\mathbf{s} = \mathbf{T}^\dagger \mathbf{y}$ and optimal scalar estimation on each s_i, but such a method will not in general achieve the minimum MSE. Simply by applying gain to all of the s_i, we effectively apply a real, even filter to the channel output, leaving the phase unaffected. Clearly exploiting the magnitude and phase of the channel frequency response will improve the estimate – just consider the intuitive example of the non-causal Wiener filter operating in the Fourier domain:

$$W(e^{j\omega}) = \frac{S_{xy}(e^{j\omega})}{S_{yy}(e^{j\omega})}. \tag{5.69}$$

In general, $W(e^{j\omega})$ does not have zero phase, which implies that the estimator is extracting different information from the negative versus positive frequencies. We propose a method that will exploit both the magnitude and phase information of the channel output by using two vectors at the decoder, $\mathbf{s} = \mathbf{T}^\dagger \mathbf{y}$ and another vector $\bar{\mathbf{s}} = \bar{\mathbf{T}}^\dagger \mathbf{y}$, also calculated using a fast transform. Each scalar z_i is estimated by a 2-by-2 linear operation on s_i and a $\bar{s}_i$. We will see that the extra transform comes at little additional computational cost.

At the decoder we form a signal $\bar{\mathbf{s}} = \bar{\mathbf{T}}^\dagger \mathbf{y}$ in addition to $\mathbf{s}$, where the transform $\bar{\mathbf{T}}^\dagger$ is such that for every cosine basis vector $\mathbf{t}_i$ in $\mathbf{T}$ there is a corresponding sine basis vector $\bar{\mathbf{t}}_i$, a column vector in $\bar{\mathbf{T}}$. Thus if $\mathbf{T}^\dagger$ is the DCT then $\bar{\mathbf{T}}^\dagger$ is the discrete sine transform (DST), whose basis functions are given by

$$\bar{t}_{nk} = \sqrt{\frac{2}{N}} \cos\left[\left(n + \frac{1}{2}\right)\frac{(k)\pi}{N}\right], \quad n = 0, ..., N-1, \; k = 1, ..., N-1. \tag{5.70}$$

There is no DST basis function for $k = 0$, because the $k = 0$ DCT basis function is real and even. Plugging $k = 0$ into (5.70) gives zero. Sinusoidal (cosine or sine) transforms can usually be computed by way of the DFT [3] and a simple linear operation. Thus, at the decoder, the values

[3] Alternative methods can use the discrete Hartley transform [71, 53] or the DCT-IV transform, an efficient method

of the DFT that are generated to calculate the cosine transform vector $\mathbf{s}$ can be used for the easy calculation of the sine transform vector $\bar{\mathbf{s}}$; relative to the order $N \log N$ calculations required for the DFT, the amount of computation required for the sine transform is negligible. We will see, however, that we may incur extra costs as far as overhead bits required to communicate statistical information.

Projecting onto both sines and cosines, we have maintained the phase information of $\mathbf{y}$. The two values s_i and $\bar{s}_i$ represent, respectively, the real and imaginary component of the projection of $\mathbf{y}$ onto the *positive-frequency* component of the i^{th} basis function. We assume that frequency response of the channel is constant over the positive-frequency passband of each of the basis functions, equaling the complex value $h = h_R + jh_I$; our notation omits the implicit dependence on basis vector index i. Correspondingly, the frequency response in the negative frequency passband equals the constant $h^* = h_R - ih_I$ (where * is the complex conjugate operator). Given that $\mathbf{t}_i$ and $\bar{\mathbf{t}}_i$ are narrowband, with their passbands approximately disjoint from the passbands of other basis functions, most of the information about z_i will be contained in just the two coefficients s_i and $\bar{s}_i$. We therefore perform analog estimation with s_i and $\bar{s}_i$.

We take a moment to note that our assumption that $\mathbf{t}_i$ and $\bar{\mathbf{t}}_i$ have little energy in the passbands of any other basis function is inaccurate for block transforms, because of the presence of significant sidelobes. We will see, however, in Sec. 5.8 and Sec. 5.9 that when lapped transforms or perfect-reconstruction filterbanks are used as the method of signal transformation, sidelobes can be adequately attenuated for this assumption to be accurate. All of the analysis is this section can be applied directly to systems using lapped transform and filterbanks.

We define $\bar{z}_i = \bar{\mathbf{t}}_i^{\dagger}\mathbf{x}$, to be the imaginary component (or sine component) of $\mathbf{x}$ in the *one-sided, positive-frequency* passband of the i^{th} basis function. We assume that the effect of convolution via the channel on the i^{th} transform coefficient is a gain of h in positive frequency and a gain of h^* in negative frequency. By the definition of complex multiplication we have the following channel output equations:

$$\begin{bmatrix} s_i \\ \bar{s}_i \end{bmatrix} = \begin{bmatrix} h_R & -h_I \\ h_I & h_R \end{bmatrix} \begin{bmatrix} z_i \\ \bar{z}_i \end{bmatrix} + \begin{bmatrix} w_i \\ \bar{w}_i \end{bmatrix}, \tag{5.71}$$

for which is cited in [55] as having been unknowingly derived in [25] and [26].

where $w_i = \mathbf{t}_i^\dagger \mathbf{n}$, and $\bar{w}_i = \bar{\mathbf{t}}_i^\dagger \mathbf{n}$. The transform coefficients z_i and $\bar{z}_i$ are uncorrelated by

$$\begin{aligned} E[\bar{z}_i z_i^*] &= E[\bar{\mathbf{t}}_i^\dagger \mathbf{x}\mathbf{x}^\dagger \mathbf{t}_i] && (5.72) \\ &= \bar{\mathbf{t}}_i^\dagger \mathbf{\Lambda}_\mathbf{x} \mathbf{t}_i && (5.73) \\ &= \lambda_i \bar{\mathbf{t}}_i^\dagger \mathbf{t}_i && (5.74) \\ &= 0, && (5.75) \end{aligned}$$

where (5.74) follows from $\mathbf{t}_i$ being an eigenvector of $\mathbf{\Lambda}_\mathbf{x}$ and (5.75) follows from the orthogonality of $\mathbf{t}_i$ and $\bar{\mathbf{t}}_i$. Of course for most applications, $\mathbf{t}_i$ is only an approximate eigenvector of $\mathbf{\Lambda}_\mathbf{x}$, which means that (5.75) is only approximate. Furthermore, the orthogonality of a discrete cosine basis function to a discrete sine basis function does not always hold [4], but their inner products are usually small enough for this to be a good approximation. The same arguments used to derive (5.75) can be used to show that

$$E[w_i \bar{w}_i^*] = 0. \tag{5.76}$$

For any vector $[u \;\; \bar{u}]^T$, we denote it by $\breve{\mathbf{u}}$. As additional notation we let

$$\breve{\mathbf{H}} = \begin{bmatrix} h_R & -h_I \\ h_I & h_R \end{bmatrix} \tag{5.77}$$

We form the MMSE estimate of z_i from s_i and $\bar{s}_i$ by applying the standard equations:

$$\begin{aligned} \tilde{z}_i &= E[z_i \breve{\mathbf{s}}^T] E[\breve{\mathbf{s}}\breve{\mathbf{s}}^T]^{-1} && (5.78) \\ &= \lambda_i [h_R \;\; h_I] (\breve{\mathbf{H}}^T \mathbf{\Lambda}_{\breve{\mathbf{z}}} \breve{\mathbf{H}} + \mathbf{\Lambda}_{\breve{\mathbf{w}}})^{-1}, && (5.79) \end{aligned}$$

where (5.75) and (5.76) imply that $\mathbf{\Lambda}_{\breve{\mathbf{z}}}$ and $\mathbf{\Lambda}_{\breve{\mathbf{w}}}^{-1}$ are diagonal matrices. The error variance for the estimate $\tilde{z}_i$, which is necessary for determining the the NLSQ staircase width is

$$E[(z_i - \tilde{z}_i)^2] = \lambda_i - \lambda_i^2 [h_R \;\; h_I] (\breve{\mathbf{H}}^T \mathbf{\Lambda}_{\breve{\mathbf{z}}} \breve{\mathbf{H}} + \mathbf{\Lambda}_{\breve{\mathbf{w}}})^{-1} \begin{bmatrix} h_R \\ h_I \end{bmatrix} \tag{5.80}$$

[4] See, for example, the DCT-IV and the DST-IV

Note that both (5.79) and (5.80) require that the decoder know the variance of z_i, which equals λ_i, and the variance of $\bar{z}_i$ which is given by $\bar{\lambda}_i = \bar{\mathbf{t}}_i \mathbf{\Lambda}_{\mathbf{x}} \bar{\mathbf{t}}_i$.

So, in addition to requiring the calculation of two transforms at the receiver, the proposed decoding method also requires overhead bits to describe the variances λ_i and $\bar{\lambda}_i$ for all i. Thus, for some sources the overhead could be twice that used for the suboptimal method in Sec. 5.5.3. Note, however that we have saved bits in some respect in that we have designed the system for z_i and $\bar{z}_i$ to be orthogonal. We, therefore do not have to describe their cross correlations. More bit savings can be had by observing that for some sources it is safe to assume that the variances λ_i and $\bar{\lambda}_i$ are equal. For example, for very tonal sources such as speech, the projection onto a sine or a cosine, averaged over uniformly distributed phase, should result in $\lambda_i = \bar{\lambda}_i$.

For coding scenarios which do require the transmission of the vectors λ and $\bar{\lambda}$, a designer must compare the reduction of analog error, by taking into account both the magnitude and phase of the channel frequency response, to the expense of adding overhead bits to the transmission. Those bits if not used for overhead could be used to increase the rates to the quantizers.

5.6 Blocks extracted from stationary processes

5.6.1 Optimal estimation

For many applications, the source is actually a wide-sense stationary (WSS) Gaussian process $x[n]$, and block transform coding is accomplished by defining source vectors $\mathbf{x}$ as a contiguous block of source samples:

$$\mathbf{x} = \Big[x[n_0]\ x[n_0+1]\ ...\ x[n_0+N-1]\Big]^T. \tag{5.81}$$

for some starting point $n_0 \in \{..., -N, 0, N, 2N, ...\}$. In our notation, we have dropped the dependence of $\mathbf{x}$ on n_0, because of the stationarity of the source, which implies that $\mathbf{x}$ has the same second-order statistics for any n_0. The entire source $x[n]$ is coded by processing successive, non-overlapping blocks.[5] At the decoder there is a similar situation. The channel output is often a process $y[n]$ that is jointly Gaussian, jointly WSS with $x[n]$. Due to causality or other constraints

[5] We will see in Sec. 5.8 that lapped orthogonal transforms, which use overlapping blocks of samples, can also be applied to systematic coding.

the decoder only observes a block $\mathbf{y}$ of $y[n]$:

$$\mathbf{y} = \Big[x[n_0 - m_0]\; x[n_0 - m_0 + 1] \ldots x[n_0 + m_1 - 1]\Big]^T, \tag{5.82}$$

for some m_0 and $m_1 = M - m_0$, which can take on a variety of values depending on the constraints at the decoder. In the unconstrained case where $m_0 \to \infty$ and $m_1 \to \infty$, the decoder performs non-causal processing, *i.e.* non-causal Wiener filtering, on $y[n]$ for the analog estimation stage. Non-causal processing can be approximated to within a delay by buffering, and doing fixed-lag prediction, which adds to system latency.

5.6.2 Low-complexity causal implementation

In many applications, it is imperative to minimize latency, requiring causal processing on a vector $\mathbf{y}$ for which $m_0 \to \infty$ and $m_1 = N$. By the properties of linear estimators, we have that the MMSE estimate of $\mathbf{z}$ from $\mathbf{y}$ is given by $\tilde{\mathbf{z}} = \mathbf{T}\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}}$ is the MMSE estimate of $\mathbf{x}$ from $\mathbf{y}$. Note that $\tilde{\mathbf{x}}$ does *not* equal $\breve{\mathbf{x}}$ as defined by

$$\breve{\mathbf{x}} = \Big[\breve{x}[n_0]\; \breve{x}[n_0 + 1] \ldots \breve{x}[n_0 + N - 1]\Big]^T, \tag{5.83}$$

where $\breve{x}[n]$ is the output of the causal Wiener filter acting on $y[n]$. The reason that (5.83) is not an optimal estimator is that the early samples of the vector $\mathbf{x}$, *i.e.* x_1, x_2, etc., have many lead values of $y[n]$ from which to be estimated, whereas the later samples, x_N, x_{N-1}, etc., have few lead variables from which to be estimated. For a given index i, $i = 1, \ldots, N$, the correct MMSE estimate of the i^{th} component of $\mathbf{x}$ is given by

$$\tilde{x}_i = \breve{x}_{N-i}[n_0 + N - 1], \tag{5.84}$$

where $\breve{x}_j[n]$ is the output of the MMSE fixed-lag smoother of lag j. A consequence of (5.84) is that even though $x[n]$ and $y[n]$ are jointly WSS the MMSE error variable $\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}$ (and any linear function of the error) is not necessarily WSS.

Equation (5.84) indicates that the computation of $\tilde{\mathbf{z}} = \mathbf{T}\tilde{\mathbf{x}}$ potentially requires the running of N parallel fixed-lag smoothers, which brings with it a high computational complexity. A method

more amenable to a computationally efficient decoder would be to calculate $\breve{x}[n] = g[n] * y[n]$, where $g[n]$ is the causal Wiener filter, and calculate a suboptimal $\breve{\mathbf{z}} = \mathbf{T}\breve{\mathbf{x}}$ where $\breve{\mathbf{x}}$ is given by (5.83). We form the signal estimate $\hat{z}_i$ by doing nearest-neighbor NLSQ decoding using $\breve{z}_i$ as input. Denoting the i^{th} NLSQ output level by k_i, the decoder reconstructs the i^{th} transform coefficient by selecting the nearest lattice point to $\breve{z}$ from the lattice $\mathcal{L}_{k_i}$. Obviously, the method is suboptimal, but under the condition that $\breve{z}_i$ is used to reconstruct $\hat{z}_i$ via NLSQ decoding, we can optimize the constrained system. Letting $\breve{\mathbf{e}}_{\mathbf{x}} = \mathbf{x} - \breve{\mathbf{x}}$ and $\breve{\mathbf{e}}_{\mathbf{z}} = \mathbf{z} - \breve{\mathbf{z}}$, we have the error correlation matrices

$$\begin{aligned} \mathbf{\Lambda}_{\breve{\mathbf{e}}_{\mathbf{z}}} &= E[\breve{\mathbf{e}}_{\mathbf{z}}\breve{\mathbf{e}}_{\mathbf{z}}^{\dagger}] && (5.85) \\ &= \mathbf{T}^{\dagger}E[\breve{\mathbf{e}}_{\mathbf{x}}\breve{\mathbf{e}}_{\mathbf{x}}^{\dagger}]\mathbf{T} && (5.86) \\ &= \mathbf{T}^{\dagger}\mathbf{\Lambda}_{\breve{\mathbf{e}}_{\mathbf{x}}}\mathbf{T}. && (5.87) \end{aligned}$$

We design the NLSQ for the i^{th} transform coefficient such that the staircase width W is $G\breve{\sigma}_i$, where

$$\breve{\sigma}_i^2 = E[(z_i - \breve{z}_i)^2], \tag{5.88}$$

and we use $G = 8$ as justified in Chap. 4. Note that $\mathbf{\Lambda}_{\breve{\mathbf{e}}_{\mathbf{z}}}$ and $\mathbf{\Lambda}_{\breve{\mathbf{e}}_{\mathbf{z}}}$ are termed error *correlation* matrices, not error covariance matrices, because $E[\mathbf{z}|\breve{\mathbf{z}}]$ is not necessarily equal to $\breve{\mathbf{z}}$. There is bias in the estimate $\breve{\mathbf{z}}$, which could throw off the nearest-neighbor reconstruction of the NLSQs. Through empirical observations we see that the bias should not significantly affect the performance of the NLSQ, and if it does, the effect of the bias can be removed by increasing G uniformly for all of the NLSQs. In doing so, we have an operational distortion-rate function for the i^{th} NLSQ that is proportional to the second moment of the analog error, $\breve{\sigma}_i^2$:

$$d_i = \breve{\theta}\breve{\sigma}_i^2\, 2^{-2R_i}, \tag{5.89}$$

where $\breve{\theta}$ may be slightly larger than θ in (5.6) to account for the estimator bias.

Using the results of Sec. 5.3 and (5.89), we have following result for optimal bit allocation. Defining

$$\breve{\beta}^2 = \Big(\prod_{i=1}^{N} \breve{\sigma}_i^2\Big)^{\frac{1}{N}}, \tag{5.90}$$

the geometric mean of the analog estimation error second moments, the optimal assignment of bits to the i^{th} SSQ is

$$R_i = \bar{R} + \frac{1}{2}\log_2 \frac{\breve{\sigma}_i^2}{\breve{\beta}^2}. \tag{5.91}$$

The resultant minimum MSE attained is

$$J = N\breve{\theta}\breve{\beta}^2 2^{-2\bar{R}}. \tag{5.92}$$

As in the standard SBTC case, the distortion the same across all quantizers, *i.e.*,

$$d_i = \breve{\theta}\breve{\beta}^2 2^{-2\bar{R}}, \tag{5.93}$$

Using the results of Sec. 5.4, and (5.89) and (5.87), we have that the optimal transform is that which diagonalizes $\mathbf{\Lambda}_{\breve{\mathbf{e}}_x}$, which is the correlation matrix of a length-N block of the error sequence associated with the causal Wiener filter $g[n]$.

For both the causal and non-causal case, as we have constructed them, the error sequence is stationary. Recall that in this special case the coding gain can be expressed by (5.37), the ratio of the arithmetic to the geometric mean of the eigenvalues of $\mathbf{\Lambda}_{\mathbf{e}_x}$.

5.6.3 Non-Gaussian case

We observe from our development of the suboptimal estimation stage in Sec. 5.6.2 that our results depend only very minimally on the source or channel being Gaussian. We therefore consider the design of a similar system for the non-Gaussian case. In our design, the analog estimation stage at the decoder forms an estimate $\breve{x}[n]$ of the source, the resulting error sequence for which should be WSS. A natural estimator, assuming a jointly WSS $x[n]$ and $y[n]$ is the Wiener filter, or appropriate fixed-lag filter, whose error sequence is guaranteed to be WSS. Using the block $\breve{\mathbf{x}}$ defined by (5.83), the decoder then constructs the transform coefficient estimate $\breve{\mathbf{z}} = \mathbf{T}^{\dagger}\breve{\mathbf{x}}$, whose error correlation matrix $\mathbf{\Lambda}_{\breve{\mathbf{e}}_z}$ is given by 5.87. Denoting the i^{th} output NLSQ output level by $\boldsymbol{k}_i$, the decoder reconstructs the i^{th} transform coefficient by selecting the nearest lattice point to $\breve{\mathbf{z}}$ from the lattice $\mathcal{L}_{\boldsymbol{k}_i}$. As in the causal Gaussian case in Sec. 5.6.2, there is a bias because we do not necessarily have that $E[\mathbf{z}|\breve{\mathbf{z}}] = \breve{\mathbf{z}}$. In addition, the error distribution about $\breve{\mathbf{z}}$ is not Gaussian

– it may be heavy tailed or even more pathological. In spite of these difficulties, we will assume that we can use an NLSQ to code the transform coefficients with staircase width $W = G\breve{\sigma}_i$ where $\breve{\sigma}_i^2$ is given by (5.88). The value of G must be optimized for the distribution such that there are negligible intercell errors, (for all observations $\mathbf{y}$, and transform coefficients $\mathbf{z}_i$!). Assuming that this can be accomplished, we have that the distortion-rate function is given by (5.89), and all of the results pertaining optimal bit allocation and optimal transform selection are as in Sec. 5.6.2.

5.7 AR-1 process

An important of model for a wide variety of sources is the P^{th}-order autoregressive (AR-P) model. The AR-P model simply models a source as white noise passed through a P^{th} order all-pole filter whose z-transform is given by

$$F(z) = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}}. \tag{5.94}$$

For image processing, in particular, the AR-1 model is useful in its ability to model areas of slowly varying samples, and is sufficiently accurate for coding purposes [40, 17, 62]. The autocorrelation function, $R_{xx}[n]$, of an AR-1 process is given by

$$R_{xx}[n] = \sigma_x^2 \rho^{|n|}, \tag{5.95}$$

where ρ is the correlation between adjacent samples ($0 \leq \rho \leq 1$); for the rest of this treatment we let $\sigma_x^2 = 1$. Taking a contiguous block of N samples from the process as in (5.81) we have a source vector $\mathbf{x}$ with the following autocorrelation matrix:

$$\mathbf{\Lambda_x} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{bmatrix} \tag{5.96}$$

There are very few processes for which the eigenvectors of an autocorrelation function can be written in closed form. Conveniently, it is shown in [63] that the eigenvectors of (5.96) have closed-form

sinusoidal solution, which depends on the precise value of ρ, or in other words, is signal dependent. For image processing in particular, it is useful to consider the limiting case in which $\rho \to 1$, which approximates contours that are slowly varying. In this limiting case, the eigenvectors of (5.96) are given by the DCT basis functions in (5.47) [16], which are signal independent. This result shows, therefore, that for AR-1 processes in the limit of high correlation, the DCT is the globally optimal transform for standard transform coding with respect to maximizing coding gain. Indeed, it also shows that if the error sequence $\mathsf{e}[n] = \mathsf{x}[n] - \hat{\mathsf{x}}[n]$ that results from linear estimation of $y[n]$ is AR-1 in the limit of high correlation, that the DCT is the globally optimal transform for SBTC. We use this fact to show that the DCT is the globally optimal transform for two important special cases.

We will assume an AR-1 source $\mathsf{x}[n]$, whose autocorrelation function is given by (5.95), and a channel output equation

$$\mathsf{y}[n] = h\mathsf{x}[n] + \mathsf{v}[n], \tag{5.97}$$

where h is a constant real gain, and $\mathsf{v}[n]$ is AWGN of variance σ_{v}^2. We will see that the results for this channel can be easily extended to a very general class of channels. We consider the MMSE error sequence $\mathsf{e}[n]$ for two scenarios. In the first case the decoder can perform non-causal processing, over the entire history and future of the received channel output $y[n]$. In other words, for each sample $\mathsf{x}[n_0]$ that the decoder must estimate, the decoder observes the signal $\mathsf{y}[n]$, $n = -\infty, ..., \infty$. In the second case we assume that the low-complexity causal coding system of Sec. 5.6.2 is used. By the stationarity of the analog estimation error signal $e[n]$ for both the non-causal and causal cases, the DCT is an optimal transform if it diagonalizes the matrix $E[\mathbf{e}\mathbf{e}^\dagger]$ where $\mathbf{e}$ is any contiguous of block of N samples of $e[n]$. Therefore we focus our attention on the auto-correlation function, $R_{\mathsf{ee}}[n]$, of the error sequence $e[n]$ for both cases. Calculated directly from $R_{\mathsf{ee}}[n]$, the matrix $\mathbf{\Lambda}_{\mathbf{e}}$, in the limit as $\rho \to 1$, is diagonalized by the DCT transform for both cases as show below.

5.7.1 Non-causal estimation

In the non-causal case, we assume that the observation at the decoder is $\mathsf{y}[n]$, $n = -\infty, ..., \infty$. In this section we establish that the MMSE error sequence is AR-1 with correlation coefficient approaching 1 in the limit as $\rho \to 1$, thereby proving the optimality of the DCT for SBTC.

We determine the MMSE estimate of $\mathsf{x}[n]$ by filtering $\mathsf{y}[n]$ by the non-causal Wiener filter

$G(z) = S_{xy}(z)/S_{yy}(z)$, where the notations $S_{uu}(z)$ and $S_{uw}(z)$ are used to denote the power spectral density of $u[n]$ and the cross power spectral density of $u[n]$ and $w[n]$, respectively. The impulse response of $G(z)$ is given by $g[n]$. The autocorrelation sequence for the error, $R_{ee}[n]$, is determined by the following:

$$\begin{aligned}
R_{ee}[n] &= E[e[m+n]e[m]] && (5.98)\\
&= E[(x[m+n]-\tilde{x}[m+n])(x[m]-\tilde{x}[m])] && (5.99)\\
&= E[(x[m+n]-\tilde{x}[m+n])x[m]] - E[(x[m+n]-\tilde{x}[m+n])\tilde{x}[m]] && (5.100)\\
&= E[(x[m+n]-\tilde{x}[m+n])x[m]] && (5.101)\\
&= R_{xx}[n] - R_{\tilde{x}x}[n] && (5.102)\\
&= R_{xx}[n] - g[n] * R_{yx}[n], && (5.103)
\end{aligned}$$

where (5.101) follows from the orthogonality principle, i.e. $\tilde{x}[n_0]$ is orthogonal to $e[n]$ for all n. The power spectral density for the error is the z-transform of (5.103):

$$\begin{aligned}
S_{ee}(z) &= S_{xx}(z) - G(z)S_{yx}(z) && (5.104)\\
&= S_{xx}(z) - \frac{S_{xy}(z)}{S_{yy}(z)}S_{yx}(z) && (5.105)\\
&= S_{xx}(z) - \frac{h^2 S_{xx}^2(z)}{S_{yy}(z)}, && (5.106)
\end{aligned}$$

where

$$\begin{aligned}
S_{xx}(z) &= \frac{1}{1-\rho z^{-1}} - \frac{1}{1-\rho^{-1}z^{-1}} && (5.107)\\
&= \frac{z^{-1}(\rho^2-1)/\rho}{(1-\rho z^{-1})(1-\rho^{-1}z^{-1})} && (5.108)
\end{aligned}$$

and

$$\begin{aligned}
S_{yy}(z) &= h^2 S_{xx}(z) + \sigma_v^2. && (5.109)\\
&= \frac{\rho\sigma_v^2 z^{-2} + \left(h^2(\rho^2-1) - \sigma_v^2(\rho^2+1)\right)z^{-1} + \rho\sigma_v^2}{\rho(1-\rho z^{-1})(1-\rho^{-1}z^{-1})} && (5.110)
\end{aligned}$$

$$= \frac{\sigma_v^2(z^{-1}-r)(z^{-1}-r^{-1})}{(1-\rho z^{-1})(1-\rho^{-1}z^{-1})}, \tag{5.111}$$

where r is a zero of $S_{yy}(z)$ inside the unit circle. Note that as $\rho \to 1$, we have $r \to 1$. The power spectral density of the MMSE error sequence is

$$\begin{aligned} S_{ee}(z) &= S_{xx}(z) - \frac{h^2 S_{xx}^2(z)}{h^2 S_{xx}(z) + \sigma_v^2} && (5.112)\\ &= \frac{\sigma_v^2 S_{xx}(z)}{h^2 S_{xx}(z) + \sigma_v^2} && (5.113)\\ &= \frac{\sigma_v^2 z^{-1}(\rho^2-1)}{h^2 z^{-1}(\rho^2-1) + \rho\sigma_v^2(1-\rho z^{-1})(1-\rho^{-1}z^{-1})} && (5.114)\\ &= \frac{\sigma_v^2 z^{-1}(\rho^2-1)\rho}{\rho\sigma_v^2 z^{-2} + \Big(h^2(\rho^2-1) - \sigma_v^2(\rho^2+1)\Big)z^{-1} + \rho\sigma_v^2}. && (5.115) \end{aligned}$$

The power spectral density $S_{ee}(z)$ has two poles, given by $r \leq 1$ and $r^{-1} \geq 1$, which implies that the MMSE error process $e[n]$ is an AR-1 process with correlation coefficient r. Letting $\rho \to 1$, we see that $r \to 1$, which implies that the DCT is the optimal transform for an AR-1 source and non-causal observations of AWGN channel output.

5.7.2 Causal estimation

In the causal case, we assume $y[n]$ is processed by the causal Wiener filter $G(z)$, whose formula requires the following notation. We write the spectral factorization of a general $S_{yy}(z)$ as

$$S_{yy}(z) = B\frac{\prod_i(1-a_i z^{-1})\prod_i(1-a_i^{-1}z^{-1})}{\prod_j(1-b_j z^{-1})\prod_j(1-b_j^{-1}z^{-1})}, \tag{5.116}$$

where the a_is are the zeros inside the unit circle, the b_js are the poles inside the unit circle, and B is the gain. We define the functions

$$S_{yy}^{+}(z) = \sqrt{B}\frac{\prod_i(1-a_i z^{-1})}{\prod_j(1-b_j z^{-1})} \tag{5.117}$$

and

$$S_{yy}^{-}(z) = \sqrt{B}\frac{\prod_i(1-a_i^{-1} z^{-1})}{\prod_j(1-b_j^{-1} z^{-1})}. \tag{5.118}$$

Letting $W(z)$ be the z-transform of $w[n]$, we use the notation $\{W(z)\}_+$ to denote the z-transform $w[n]u[n]$, where $u[n]$ is the unit step function. We use the notation $\{W(z)\}_-$ to denote the z-transform of $w[n]u[-n-1]$. Thus $W(z) = \{W(z)\}_+ + \{W(z)\}_-$. With our notation established, we have that the causal Wiener filter is

$$G(z) = \frac{\{S_{xy}(z)/S_{yy}^-(z)\}_+}{S_{yy}^+(z)}, \tag{5.119}$$

whose impulse response is denoted $g[n]$. The autocorrelation sequence for the error, $R_{ee}[n]$, is determined by the following:

$$\begin{aligned}
R_{ee}[n] &= E[e[m+n]e[m]] && (5.120)\\
&= E[(x[m+n]-\tilde{x}[m+n])x[m]] - E[(x[m+n]-\tilde{x}[m+n])\tilde{x}[m]] && (5.121)\\
&= E[x[m+n](x[m]-\tilde{x}[m])] - E[\tilde{x}[m+n](x[m]-\tilde{x}[m])] && (5.122)\\
&= \begin{cases} E[(x[m+n]-\tilde{x}[m+n])x[m]] & n \geq 0 \\ E[x[m+n](x[m]-\tilde{x}[m])] & n < 0 \end{cases} && (5.123)\\
&= \begin{cases} R_{xx}[n] - R_{\tilde{x}x}[n] & n \geq 0 \\ R_{xx}[n] - R_{x\tilde{x}}[n] & n < 0 \end{cases} && (5.124)\\
&= \begin{cases} R_{xx}[n] - g[n] * R_{yx}[n] & n \geq 0 \\ R_{xx}[n] - g[-n] * R_{xy}[n] & n < 0 \end{cases} && (5.125)\\
&= \begin{cases} T[n] & n \geq 0 \\ T[-n] & n < 0, \end{cases} && (5.126)
\end{aligned}$$

where

$$T[n] = R_{xx}[n] - g[n] * R_{yx}[n], \tag{5.127}$$

and (5.123) follows from the fact that

$$E[e[m+n]\tilde{x}[m]] = 0, \quad n \geq 0 \tag{5.128}$$

The power spectral density of $e[n]$ is thus given by

$$S_{ee}(z) = \{T(z)\}_+ + \{T(z^{-1})\}_-. \tag{5.129}$$

Because $R_{ee}[n]$ is symmetric about 0, it is uniquely defined by the term $\{T(z)\}_+$, which we expand:

$$\begin{aligned}
\{T(z)\}_+ &= \{S_{xx}(z) - G(z)S_{yx}(z)\}_+ && (5.130)\\
&= \{S_{xx}(z) - \frac{\{S_{xy}(z)/S_{yy}^-(z)\}_+}{S_{yy}^+(z)}S_{xx}(z)\}_+ && (5.131)\\
&= \{S_{xx}(z)\Big(1 - \frac{\{h^2 S_{xx}(z)/S_{yy}^-(z)\}_+}{S_{yy}^+(z)}\Big)\}_+ && (5.132)
\end{aligned}$$

For our formulation, we have

$$S_{yy}^+(z) = \sigma_v \frac{z^{-1} - r^{-1}}{1 - \rho z^{-1}} \tag{5.133}$$

and

$$S_{yy}^-(z) = \sigma_v \frac{z^{-1} - r}{1 - \rho^{-1} z^{-1}}, \tag{5.134}$$

which yields

$$\begin{aligned}
\{T(z)\}_+ &= \{S_{xx}(z)\Big(1 - \frac{\{\frac{K_1 z^{-1}}{(1-\rho z^{-1})(z^{-1}-r)}\}_+}{S_{yy}^+(z)}\Big)\}_+ && (5.135)\\
&= \{S_{xx}(z)\Big(1 - \frac{K_2}{(1-\rho z^{-1})}\frac{1}{\sigma_v}\frac{1-\rho z^{-1}}{z^{-1} - r^{-1}}\Big)\}_+ && (5.136)\\
&= \{\frac{z^{-1}(\rho^2 - 1)}{\rho(1-\rho z^{-1})(1-\rho^{-1}z^{-1})}\Big(\frac{z^{-1} - K_3}{z^{-1} - r^{-1}}\Big)\}_+ && (5.137)\\
&= \frac{K_4}{1-\rho z^{-1}} + \frac{K_5}{1 - r z^{-1}} && (5.138)
\end{aligned}$$

for some constant values K_1, K_2, K_3, K_4, and K_5, whose exact values are irrelevant for our purposes. To arrive at (5.138) we have used some basic fractional expansion arguments. From (5.138), we have that

$$R_{ee}[n] = K_4 \rho^{|n|} + K_5 r^{|n|}. \tag{5.139}$$

For a vector $\mathbf{x}$ extracted as a contiguous block of N samples from $x[n]$, and corresponding MMSE error vector $\mathbf{e}$, (5.139) implies that the error covariance matrix is given by

$$\mathbf{\Lambda_e} = \mathbf{\Lambda}_\rho + \mathbf{\Lambda}_r, \tag{5.140}$$

where $\mathbf{\Lambda}_\rho$ and $\mathbf{\Lambda}_r$ are autocorrelation matrices for AR-1 processes, with correlation coefficients ρ and r respectively, of the form given by (5.96). From (5.111) we have that $\rho \to 1$, implies $r \to 1$. Thus in the limit, DCT basis vectors are eigenvectors of both $\mathbf{\Lambda}_\rho$ and $\mathbf{\Lambda}_r$. The DCT basis vectors are consequently eigenvectors of $\mathbf{\Lambda_e}$, which proves the DCT is the optimal transform for an AR-1 source and causal observations of AWGN channel output.

5.7.3 General Gaussian channel model

Up to this point, our proofs of optimality have relied on the assumption of an additive white noise channel. Consider a more general channel model $y[n] = h[n] * x[n] + v[n]$, where $h[n]$ is a channel impulse response with a smooth magnitude response $H(e^{j\omega})$ about $\omega = 0$, and $v[n]$ has a power spectral density $S_{xx}(e^{j\omega})$ that is smooth about $\omega = 0$. Given the lowpass characteristic of the source, we see that we can extend our previous results for the AWGN channel to this more general channel model. As $\rho \to 1$ the power spectral energy of $x[n]$ as a function of frequency ω, is increasingly confined near DC. The energy outside an ϵ neighborhood about $\omega = 0$ becomes negligible, implying that an optimal estimator will be mainly concerned with estimating $X(e^{j\omega})$ in the region $-\epsilon \leq \omega \leq \epsilon$. Given the smoothness assumptions on $H(e^{j\omega})$ and $S_{vv}(e^{j\omega})$, we can assume that within $-\epsilon \leq \omega \leq \epsilon$, the channel behaves like an AWGN channel, which implies that for both non-causal and causal estimation, the error autocorrelation function is virtually identical to the AWGN case described above. Thus for a general Gaussian channel and an AR-1 Gaussian source the DCT is the optimal transform for SBTC, using both causal and non-causal observations.

5.7.4 Block estimation

Limiting form for Λ_{e_x}

In some cases, at the decoder, analog estimation will occur on blocks, for computational simplicity or other problem constraints, and the estimator is not accurately expressed as a linear time-invariant filter. For example, if the source is an image, which is coded by block processing, the decoder may be parallelized to operate on blocks of analog channel output. In the case of block estimation at the decoder we can assume the very general block channel model:

$$\mathbf{y} = \mathbf{H}^\dagger \mathbf{x} + \mathbf{n}, \tag{5.141}$$

where $\mathbf{H}^\dagger$ is any matrix of N rows, and $\mathbf{n}$ is zero-mean Gaussian noise, uncorrelated with $\mathbf{x}$ and having an arbitrary autocorrelation matrix $\mathbf{\Lambda_n}$. As assumed throughout Sec. 5.7, $\mathbf{\Lambda_x}$ is given by (5.96). The analog error covariance matrix for optimal estimation of $\mathbf{x}$ is given by

$$\mathbf{\Lambda_{e_x}} = \mathbf{\Lambda_x} - \mathbf{\Lambda_x H}\Big(\mathbf{H}^\dagger \mathbf{\Lambda_x H} + \mathbf{\Lambda_n}\Big)^{-1} \mathbf{H}^\dagger \mathbf{\Lambda_x} \tag{5.142}$$

The matrix $\mathbf{\Lambda_x}$ in the limit as $\rho \to 1$ approaches a matrix of all ones. We show in this section that in this limit, the DCT is a locally optimal transform for SBTC.

Consider the matrix $\mathbf{\Lambda_x K \Lambda_x}$ for any matrix $\mathbf{K}$, whose element in the i^{th} row and j^{th} column is denoted k_{ij}. Since in the limit $\mathbf{\Lambda_x}$ approaches the all ones matrix we have that $\mathbf{\Lambda_x K \Lambda_x}$ approaches $\sum_{i,j} k_{ij}$ times the all ones matrix. Letting

$$\mathbf{K} = \mathbf{H}\Big(\mathbf{H}^\dagger \mathbf{\Lambda_x H} + \mathbf{\Lambda_n}\Big)^{-1} \mathbf{H}^\dagger, \tag{5.143}$$

we have that $\mathbf{\Lambda_{e_x}}$ approaches $(1 - \sum_{i,j} k_{ij})$ times the all ones matrix as $\rho \to 1$. Note that this does not guarantee that the error covariance matrix has a form like (5.96), suggesting it is a block from an AR-1 process. Indeed, empirical analysis shows that the error vector $\mathbf{e_x}$ is not a block from an AR-1 process for all ρ. Thus, we can not directly conclude that the DCT is an optimal transform for SBTC in this case. We do show in the following, however, that the DCT is a locally optimal transform in this case.

DCT is locally optimal.

An optimal SBTC transform maximizes the coding gain Γ in (5.33), or equivalently minimizes the overall distortion J in (5.14). We show the necessary conditions for the basis functions $\mathbf{t}_i$ to achieve a local minimum in J given that $\mathbf{\Lambda_{e_x}}$ is approaching the all ones matrix, by using a proof technique similar to that in the appendix of [54]. The key fact used for the proof of optimality is that a transform that maximizes the energy compaction of the analog error variables $e_{z_i} = z_i - \tilde{z}_i$ is the diagonalizing transform of $\Lambda_{\mathbf{e_x}}$ [13]. A transform maximizes error energy compaction if for any $P < N$ there are P error variables e_{z_i} such that their average energy is maximized with respect to the transform basis functions. Maximizing energy compaction, reduces the spread of the variances σ_i^2 thereby minimizing the geometric mean of their variances. Maximal energy compaction satisfies

our intuition about an optimal coding structure. For example if all but P error variances are virtually zero, we need only allocate bits to those P coefficients and give zero bits to the rest.

As suggested in [13, 55], energy compaction can be maximized by maximizing the variances

$$\sigma_i^2 = \mathbf{t}_i^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_i, \tag{5.144}$$

recursively, from $i = 0$ to $i = N-1$, with respect to $\mathbf{t}_i$, under the constraint that $\mathbf{t}_j$, $j = 0, ..., i-1$ have been previously determined. The optimization of the first basis function $\mathbf{t}_0$ requires that we maximize

$$\xi = \mathbf{t}_0^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_0 \tag{5.145}$$

under the constraint

$$\mathbf{t}_0^\dagger \mathbf{t}_0 = 1. \tag{5.146}$$

A locally optimal solution to (5.145) is a root of the associated Lagrangian, *i.e.* is a solution of

$$2\mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_0 + 2\mu \mathbf{t}_0 = 0, \tag{5.147}$$

where μ is a Lagrange multiplier. Not surprisingly (5.147) implies that $\mathbf{t}_0$ is an eigenvector of $\mathbf{\Lambda}_{\mathbf{e}_x}$. Premultplying (5.147) by $\mathbf{t}_0$ we have that $\mu = -\mathbf{t}_0^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_0$, which reduces (5.147) to the form

$$\Big[\mathbf{\Lambda}_{\mathbf{e}_x} - (\mathbf{t}_0^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_0)\mathbf{I}\Big]\mathbf{t}_0 = 0. \tag{5.148}$$

Letting $\mathbf{\Lambda}_{\mathbf{e}_x}$ approach the all ones matrix, we have from (5.148) that $\mathbf{t}_0$ must satisfy

$$t_{n0} = \Big(\sum_{j=0}^{N-1} t_{j0}\Big)^{-1} \tag{5.149}$$

or

$$\sum_{j=0}^{N-1} t_{j0} = 0. \tag{5.150}$$

In the limit, the $\mathbf{t}_0$ that satisfies (5.149) and the constraint (5.146) equals $1/\sqrt{N}$ for all entries,

which will result in an objective function

$$\xi = \Big(\sum_{j=0}^{N-1} t_{j1} \Big)^2 = N \tag{5.151}$$

In contrast, the $\mathbf{t}_0$ that satisfies (5.149), will result in $\xi = 0$ in the limit. Thus, the $\mathbf{t}_0$ that satisfies (5.149) is globally optimal optimal. The $k = 0$ basis function of the DCT satisfies this condition (as does the $k = 0$ basis function for the DFT).

Optimization for the remaining basis functions, in succession, $\mathbf{t}_1$, $\mathbf{t}_2$, etc., proceeds in a similar manner, with additional constraints put in place to account for the the previously assigned basis functions. For the i^{th} basis function we aim to maximize

$$\xi_i = \mathbf{t}_i^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_i, \tag{5.152}$$

under the constraints

$$\mathbf{t}_i^\dagger \mathbf{t}_i = 1 \tag{5.153}$$

$$\mathbf{t}_i^\dagger \mathbf{t}_j = 0 \quad j = 0, 1, ..., i-1. \tag{5.154}$$

Setting the associated Lagrangian to zero we have

$$2\mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_i + 2\mu \mathbf{t}_i + \sum_{j=0}^{i-1} \gamma_i \mathbf{t}_j = 0, \tag{5.155}$$

where μ, and γ_j, $j = 0, ..., i$-1 are Lagrange multipliers. Premultiplying (5.155) by $\mathbf{t}_i^\dagger$ we have $\mu = -\mathbf{t}_0^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_0$. Premultiplying (5.155) by $\mathbf{t}_j^\dagger$, $j = 0, ..., i$-1 we have $\gamma_j = 0$, $j = 0, ..., i$-1. Thus (5.147) is reduced to the form

$$\Big[\mathbf{\Lambda}_{\mathbf{e}_x} - (\mathbf{t}_i^\dagger \mathbf{\Lambda}_{\mathbf{e}_x} \mathbf{t}_i) \mathbf{I} \Big] \mathbf{t}_i = 0, \tag{5.156}$$

Note that (5.156) is the same constraint equation as (5.148). Because $\mathbf{t}_i$, $i = 1, ..., N-1$, must be orthogonal to $\mathbf{t}_0$, we conclude that in the limit

$$\sum_{n=0}^{N-1} t_{ni} = 0 \tag{5.157}$$

It is straightforward to verify that the basis functions of the DCT satisfy (5.157), which implies that the DCT achieves a local maximum in coding gain Γ in the limit as $\rho \to 1$. Recall that ρ is the correlation of adjacent samples of the AR-1 source, not the correlation of the adjacent error samples. Note that (5.157) and (5.149) are also satisfied by the DFT, which emphasizes that we have identified only a local optimum solution. Empirical results, however, indicate the DCT is in fact the basis that diagonalizes $\mathbf{\Lambda}_{\mathbf{e}_x}$ in the limit as $\rho \to 1$ for a variety of channels modeled by (5.141).

5.8 Lapped transforms

As described in Sec. 5.6, stationary sources can be coded with block transform coding by forming successive, non-overlapping blocks out of adjacent source samples, as in (5.81). Due to the quantization of transform coefficients, there are often discontinuities at the block boundaries that are undesirable for many applications. Originally developed to combat these blocking effects, the lapped transform is a block transform whose basis functions overlap with those of adjacent blocks [13, 55]. Interpreted as a filter bank, a lapped transform has the advantage over block transforms of having FIR filters that are longer than the total number of transform coefficients, thereby allowing for greater stopband attenuation, minimizing the effects of aliasing. To exploit these virtues, we apply lapped transforms to systematic source coding, the structure for which we develop in this section. Optimal bit allocation and the conditions for an optimal transform are obvious extensions of the results pertaining to systematic block transform coding. Directly mapping a result from conventional coding, we show the local optimality of two fast transforms, the LOT and the MLT, in terms of energy compaction of analog estimation error for an AR-1 source and a broad class of channels.

We let P denote the total length of a single block, and $N < P$ denote the number of samples between the beginnings of successive blocks, *i.e.,* N is the frame rate in samples per block. In order to keep the number of time domain samples equal the number of transform domain samples, *i.e.,* to keep the system critically sampled, the transform $\mathbf{T}^\dagger$ must be such that the number of transform coefficients equals N. For every $n_0 = ..., -N, 0, N, 2N, ...$, the encoder forms length P block from

$x[n]$,

$$\mathbf{x} = \left[x[n_0]\ x[n_0+1] \ldots x[n_0+P-1]\right]^T, \tag{5.158}$$

applies a $N \times P$ transform $\mathbf{T}^\dagger$ to each of those blocks, forming the vector $\mathbf{z} = \mathbf{T}^\dagger\mathbf{x}$, and codes the scalar transform coefficients with NLSQ encoder maps. Using the analog observation $\mathbf{y}$ and NLSQ encoder map output $\mathbf{k}$, the NLSQ output levels are decoded as per Sec. 4.2.3. The decoder applies an $P \times N$ inverse transform to each of the overlapping blocks, and simply adds the output of overlapping blocks together. To preclude any redundancy among transform coefficients, and to maintain signal energy, we impose the following condition:

$$\mathbf{T}^\dagger\mathbf{T} = \mathbf{I} \tag{5.159}$$

To ensure that the inverse flowgraph is just the transpose of the direct transform flowgraph, the inverse transform is constrained to be $\mathbf{T}$ [54]. This constraint imposes that not only must the basis functions be orthogonal to each other, but their overlapping sections must also be orthogonal, yielding the constraint

$$\mathbf{T}^\dagger\mathbf{W}\mathbf{T}^\dagger = \mathbf{0}, \tag{5.160}$$

where

$$W = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{5.161}$$

The matrices $\mathbf{I}$ and $\mathbf{0}$ in (5.161) are $N \times N$.

The problem of bit allocation is identical to that encountered in transform coding: allocate a total of R bits across N transform coefficients. The optimal bit allocation strategy is, again, inverse waterpouring over the analog estimation error variances σ_i^2, the formulas for which are given in Sec. 5.3. As before, we wish to minimize, with respect to $\mathbf{T}^\dagger$, the total distortion given by (5.14), or equivalently maximize the coding gain in (5.33). We still have that $\hat{\mathbf{x}} = \mathbf{P}\mathbf{y}$ where $\mathbf{P}$ is given by (5.1), and by the linearity of the $\mathbf{T}^\dagger$, $\hat{\mathbf{z}} = \mathbf{T}^\dagger\hat{\mathbf{x}}$. Thus, we have

$$\mathbf{\Lambda}_{\mathbf{e}_z} = \mathbf{T}^\dagger\mathbf{\Lambda}_{\mathbf{e}_x}\mathbf{T}, \tag{5.162}$$

implying that

$$\sigma_i^2 = E[e_{z_i}^2] \tag{5.163}$$

$$= \mathbf{t}_i^\dagger \mathbf{\Lambda}_{\mathbf{e_x}} \mathbf{t}_i. \tag{5.164}$$

The extra constraint (5.160) precludes us from using the same analysis as used in SBTC, to obtain an expression for the optimal transform. We can, however, consider the optimal lapped transform in the case of an AR-1 source $x[n]$ with correlation coefficient $\rho \to 1$ and the two channel models in Sec. 5.6. In our analysis of the optimality of lapped transforms, we require that $N = 2P$. Lapped transforms with longer basis function ($N > 2P$), have been developed under the name *extended lapped transforms* [52], and are equally applicable to systematic coding.

The first channel model, describing the channel output process $y[n]$, is given by (5.97). As established in Secs. 5.7.1-5.7.2, for both non-causal and causal observations of $y[n]$, $\mathbf{\Lambda}_{\mathbf{e_x}}$ approaches the all ones matrix as $\rho \to 1$; recall that for the causal case we are using a slightly suboptimal decoder. The second channel model is the block channel model of (5.141). Again, the error covariance matrix approaches the all ones matrix as $\rho \to 1$.

Extrapolating from the block transform case, J nearly optimized if there is maximum compaction of the σ_i^2. Hence, we perform the same recursive optimization as in Sec. 5.7.4. The optimal first basis function $\mathbf{t}_0$ is that which maximizes

$$\xi = \mathbf{t}_0^\dagger \mathbf{\Lambda}_{\mathbf{e_x}} \mathbf{t}_0 \tag{5.165}$$

under the constraints

$$\mathbf{t}_0^\dagger \mathbf{t}_0 = 1, \tag{5.166}$$

and

$$\mathbf{t}_0^\dagger \mathbf{W} \mathbf{t}_0 = 0. \tag{5.167}$$

In [54] it is shown that in the limit as $\mathbf{\Lambda}_{\mathbf{e_x}}$ approaches the all ones matrix, $\mathbf{t}_0$ satisfies

$$t_{n0} + t_{n+N,0} = \Big[\sum_{i=0}^{2N-1} t_{i0} \Big]^{-1}. \tag{5.168}$$

For the i^{th} basis function we maximize

$$\xi_i = \mathbf{t}_i^\dagger \mathbf{\Lambda}_{\mathbf{e_x}} \mathbf{t}_i, \tag{5.169}$$

under the constraints

$$\begin{aligned}
\mathbf{t}_i^\dagger \mathbf{t}_i &= 1 && (5.170)\\
\mathbf{t}_i^\dagger \mathbf{W} \mathbf{t}_i &= 0 && (5.171)\\
\mathbf{t}_i^\dagger \mathbf{t}_j &= 0 \quad j = 0, 1, ..., i-1, && (5.172)\\
\mathbf{t}_i^\dagger \mathbf{W} \mathbf{t}_j &= 0 \quad j = 0, 1, ..., i-1, && (5.173)\\
\text{and } \mathbf{t}_i^\dagger \mathbf{W}^\dagger \mathbf{t}_j &= 0 \quad j = 0, 1, ..., i-1. && (5.174)
\end{aligned}$$

The constrained optimization is satisfied locally under the condition [54]

$$\sum_{j=0}^{2N-1} t_{ji} = 0. \tag{5.175}$$

In [54], Malvar develops two useful transforms for standard lapped transform coding of AR-1 sources with high correlation that have fast implementations (on the order of a length-$2N$ DCT), the lapped orthogonal transform (LOT) and the modulated lapped transform (MLT) . The two transforms both satisfy (5.168) and (5.175), and are therefore locally optimal lapped transforms for systematic coding for the broad class of channels considered.

5.9 Systematic subband coding

A generalization of all of the methods discussed up to this point is what we call systematic subband coding (SSC), which uses an N-channel perfect reconstruction filter bank, composed of an analysis filter bank at the encoder and a synthesis filter bank at the decoder, to transform the source. As is well known, both block transforms and lapped transforms are special cases of perfect reconstruction filterbanks [80]. In this section we determine the optimal bit allocation strategy and the necessary and sufficient conditions for a filter bank to be optimal for SSC. For the ideal filter case, we show that for any Gaussian channel defined by filtering and additive Gaussian noise uncorrelated with the

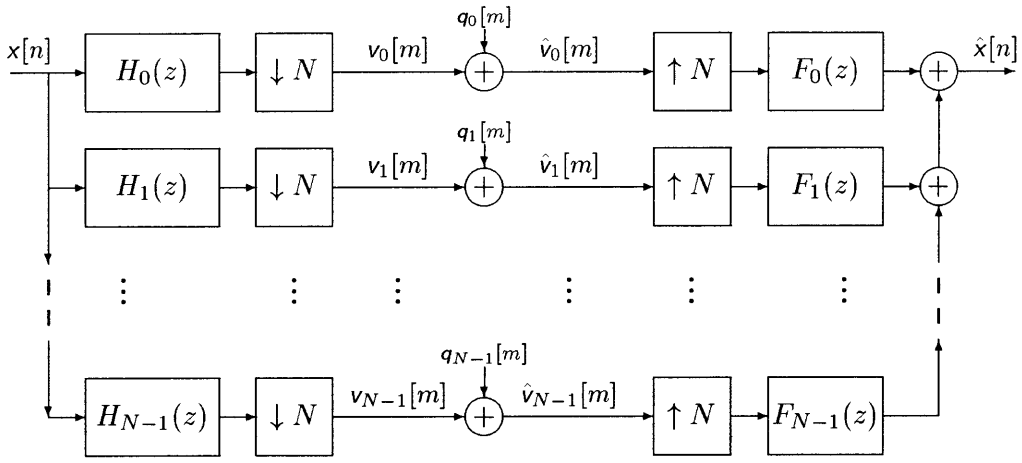


Figure 5-2: Systematic subband coding. Distortion due to NLSQ encoding and decoding of i^{th} subband is modeled by additive noise $q_i[n]$, which is clearly a function of the analog channel observation $y[n]$.

source, the optimal filter bank for standard subband coding is the optimal filter bank for systematic subband coding.

Fig. 5-2 shows the general model that we consider for systematic subband coding of signals. Indeed, the same model, appropriately paramaterized, applies to conventional source coding as well. The source $\mathbf{x}[n]$ is processed at the encoder by an N-band critically-sampled, perfect-reconstruction filter bank, which is composed of N filters $H_i(z)$, $i = 0, ..., N\text{-}1$, each followed by an N-sample decimator. The output of the filter bank is a set of N subband signals $\mathbf{v}_i[m]$, $i = 0, ..., N\text{-}1$. To account for the decimation, we use n to denote the index for the original source and the index m to denote the time index for the (decimated) subband signals. Consistent with our notation, the subband signal prior to decimation is denoted $\mathbf{v}_i[n]$. Similar to SBTC encoding, we code the i^{th} subband coefficient, $i = 0, ..., N\text{-}1$, with an optimal NLSQ at the encoder, the output of which is $\mathbf{k}_i[m]$. For better performance, the output of the NLSQs may be Slepian-Wolf coded, but our results concerning optimal filter bank selection assume no Slepian-Wolf coding. The decoder uses the channel output $\mathbf{y}[n]$ and $\mathbf{k}_i[m]$ to reconstruct the i^{th} subband coefficient, denoted $\hat{\mathbf{v}}_i[m]$. The NLSQ decoder for each subband is simply an analog estimation stage followed by a mapping to the nearest point on the lattice defined by $k_i[m]$. The analog estimation stage simply involves applying a MMSE filter $\mathbf{w}[n]$, to $y[n]$ to estimate each of the subband signals. the form of which depends on the observable samples of $\mathbf{y}[n]$ at the decoder, *e.g.* causal or non-causal samples.

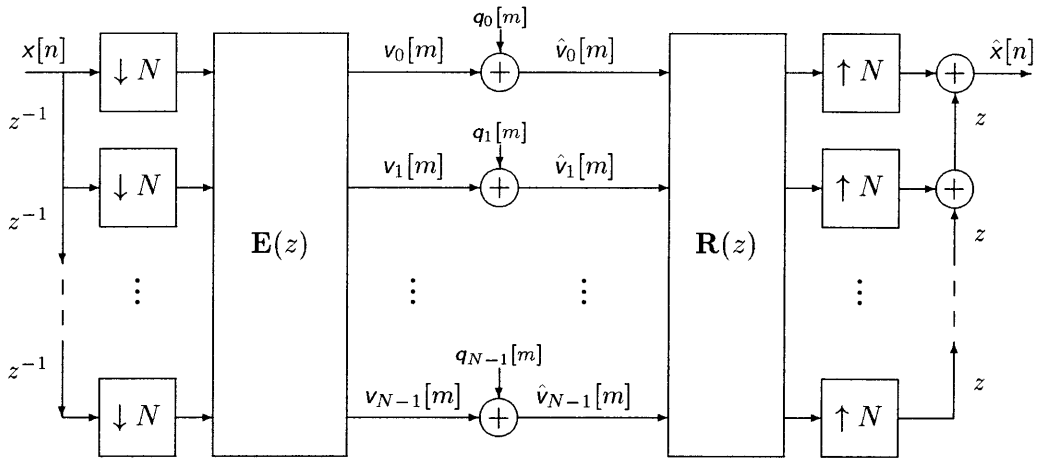


Figure 5-3: Polyphase representation of SSC.

As shown in Fig. 5-2, NLSQ quantization on the i^{th} subband is accurately modeled as an additive noise source $q_i[m]$, uncorrelated with $x[n]$. We use the notation $\mathbf{u}[m]$ to denote $[u_0[m], ...u_{N-1}[m]]^T$. The variance of $q_i[m]$, denoted d_i, is given by the distortion rate function of the NLSQ:

$$d_i = \theta \sigma_i^2 \, 2^{-2R_i}, \tag{5.176}$$

where

$$\sigma_i^2 = \sigma^2_{v_i|y[n]}. \tag{5.177}$$

The source is finally reconstructed by applying a synthesis filterbank, comprised of N-sample interpolators and synthesis filters $F_i(z)$, $i = 0, ..., N\text{-}1$. Clearly, Fig. 5-2 is the same model as used for conventional subband coding where the variance of the q_is is also given by (5.176) with $\sigma^2 = \sigma^2_{v_i}$.

5.9.1 Optimal bit allocation

Fig. 5-2 is equivalently represented by its polyphase form shown in Fig. 5-3. Perfect reconstruction implies that $\mathbf{R}(z) = \mathbf{E}^{-1}(z)$. The optimality conditions that we derive in this section assume a restricted class of filterbanks, called orthonormal filterbanks, for which the polyphase matrix $E(e^{j\omega})$ is paraunitary, *i.e.*, $\mathbf{E}(e^{j\omega})^\dagger \mathbf{E}(e^{j\omega}) = \mathbf{I}$ for all ω, and $\mathbf{R}(z) = \mathbf{E}^\dagger(z^{-1})$. It is straightforward to show that by the orthonormal structure of the filterbank, $E[(x[n] - \hat{x}[n])^2] = \frac{1}{N} E[\mathbf{q}[m]^\dagger \mathbf{q}[m]]$ [70]. Thus we can apply directly the results of Sec. 5.3 to determine that optimal NLSQ bit allocation to the

subbands involves inverse-waterpouring on the the σ_is, the MMSE analog estimation error of the subband signals. The MSE distortion, as defined by (5.10), is given by (5.14).

5.9.2 Optimal subband decompostion for the ideal filter case

In order to determine the optimal SSC filterbank, that which minimizes the MSE in (5.14), we consider the form of the MMSE analog estimation error. The analog estimation stage at the decoder clearly depends on which samples of $y[n]$ are available to the decoder. Non-causal observations ($y[\tau]$, $\tau = -\infty, ..., \infty$ is observed) and causal observations ($y[\tau]$, $\tau = -\infty, ..., n$ is observed), call for the use of the non-causal and causal Wiener filters, respectively. For practical implemetations, we would likely impose that the estimation filter be FIR. Similarly, we would have the analysis and synthesis filters be FIR. In determining the form of the optimal subband decomposition for SSQ, however, we focus our attention on the case of non-causal channel observations and ideal (infinite length) filters. In Sec. 5.9.3, we comment on the optimal subband decomposition for the case of FIR filters and causal observations.

Given non-causal observations at the decoder, for each subband we must estimate $v_i[n] = h_i[n] * x[n]$ from our observations of $y[\tau]$, $\tau = -\infty, ..., \infty$, forming the estimate $\tilde{v}_i[n]$. We apply the non-causal Wiener filter to $y[n]$ and obtain the following relation:

$$\tilde{V}_i(e^{j\omega}) = \frac{S_{v_i y}(e^{j\omega})}{S_{yy}(e^{j\omega})} Y(e^{j\omega}) \tag{5.178}$$

$$= H_i(e^{j\omega}) \frac{S_{xy}(e^{j\omega})}{S_{yy}(e^{j\omega})} Y(e^{j\omega}), \tag{5.179}$$

which implies

$$\tilde{v}_i[n] = h_i[n] * \tilde{x}[n], \tag{5.180}$$

where $\tilde{x}[n]$ is the MMSE estimate of $x[n]$ from $y[n]$. Defining

$$e_i[n] = v_i[n] - \tilde{v}_i[n], \tag{5.181}$$

we have that

$$e_i[n] = h_i[n] * e[n], \tag{5.182}$$

where and $\mathbf{e}[n] = \mathbf{x}[n] - \tilde{\mathbf{x}}[n]$. Thus, the subband error signals can be obtained by applying the analysis filter bank to the error signal corresponding to the time-domain source estimate. Using this insight, we can create an equivalent problem that will quickly yield the optimum SSC filterbank.

Consider the problem of conventionally coding $\mathbf{e}[n]$ by the following method: apply an N-band orthonormal filterbank, and code the subband signals with a conventional quantizer using optimal bit allocation. By (5.182), for any filterbank, the MSE for this problem will have the exact form as (5.14), the MSE for SSC. Thus, the optimal filter for subband coding $e[n]$ in the conventional manner will be the optimal filter bank for SSC. Using the results in [78], the necessary and sufficient conditions for the optimal SSC filterbank are:

- Diagonalization of $S_{\mathbf{ee}}(e^{j\omega})$, and
- Majorization of $S_{\mathbf{ee}}(e^{j\omega})$,

where $S_{\mathbf{ee}}(e^{j\omega})$ is the Fourier transform matrix of $E[\mathbf{e}[n+n_0]\mathbf{e}[n_0]^\dagger]$, *i.e* the power spectral density matrix of $\mathbf{e}[n]$. The matrix $S_{\mathbf{vv}}(e^{j\omega})$ is similarly defined. Diagonalization means that

$$S_{\mathbf{ee}}(e^{j\omega}) = \text{diag}\{S_0(e^{j\omega}), ..., S_{M-1}(e^{j\omega})\}, \tag{5.183}$$

where $S_i(e^{j\omega})$ is the power spectrum of $e_i[n]$. Without loss of generality, we assume that the error variances of the subbands are ordered:

$$\sigma_0^2 \geq \sigma_1^2 \geq ... \geq \sigma_{M-1}^2. \tag{5.184}$$

Majorization means that

$$S_0(e^{j\omega}) \geq S_1(e^{j\omega}) \geq ... \geq S_{M-1}(e^{j\omega}), \tag{5.185}$$

for all ω.

Gaussian channel

Assuming a Gaussian channel model,

$$\mathbf{y}[n] = b[n] * \mathbf{x}[n] + \mathbf{u}[n], \tag{5.186}$$

where $b[n]$ is an arbitrary channel impulse response and $u[n]$ is stationary WSS Gaussian noise, we show that the optimal filter bank for conventional suband coding is the optimal filter bank for SSC. Although this is a result involving ideal filters and non-causal observations, it does indicate that for a broad class of channels a good filter bank for conventional coding is a good filterbank for SSC.

We must show that the filter bank that diagonalizes and majorizes $S_{\mathbf{vv}}(e^{j\omega})$ also diagonalizes and majorizes $S_{\mathbf{ee}}(e^{j\omega})$. For any process $x[n]$ it is shown in [1] that $S_{\mathbf{vv}}(e^{j\omega})$ is made diagonal by the following, signal independent, polyphase matrix:

$$\mathbf{E}(z) = \mathbf{V}\mathbf{D}(z), \tag{5.187}$$

where $\mathbf{V}$ is the unitary inverse DFT matrix and

$$\mathbf{D}(z) = \text{diag}\{z^{-(N-1)}, z^{-N}, ..., 1]. \tag{5.188}$$

Using this polyphase matrix, the power spectrum of the i^{th} subband signal is [1]

$$S_{v_i v_i}(e^{j\omega}) = S_{xx}(e^{j(\omega+\frac{2\pi i}{N})}). \tag{5.189}$$

Clearly, majorization is achieved using a matrix $\mathbf{E}(e^{j\omega})$ that is (5.187) with certain rows appropriately permuted at each frequency. Because $\mathbf{E}(e^{j\omega})$ in (5.187) is signal independent, the matrix, or the matrix with any of its rows permuted, clearly also makes $S_{\mathbf{ee}}(e^{j\omega})$ diagonal.

Given that $S_{\mathbf{vv}}(e^{j\omega})$ is majorized, the majorization of $S_{\mathbf{ee}}(e^{j\omega})$ follows from the form of the power spectral density of $e[n]$:

$$S_{ee}(e^{j\omega}) = \frac{S_{xx}(e^{j\omega})S_{uu}(e^{j\omega})}{|B(e^{j\omega})|^2 S_{xx}(e^{j\omega}) + S_{uu}(e^{j\omega})}. \tag{5.190}$$

Clearly, (5.190) is monotonically increasing in $S_{xx}(e^{j\omega})$, which implies that if

$$S_{v_k v_k}(e^{j(\omega+\frac{2\pi i}{N})}) > S_{v_l v_l}(e^{j(\omega+\frac{2\pi i}{N})}), \tag{5.191}$$

for any i, then

$$S_k(e^{j(\omega+\frac{2\pi i}{N})}) > S_l(e^{j(\omega+\frac{2\pi i}{N})}). \tag{5.192}$$

Thus, $S_{\mathbf{ee}}(e^{j\omega})$ is majorized.

5.9.3 Causal observations and FIR filter banks

For practical coding systems, estimation elements only observe a finite number of samples, and in many applications, especially real-time applications, the receiver has only causal observations. Furthermore, low-complexity design insists on low-order FIR filter banks. In this section we comment on the form of the optimal SSC filter bank for these cases.

Given causal observations at the decoder, for each subband, the analog estimation stage filters $\mathbf{y}[n]$ with a causal Wiener filter to estimate the subband signal. It is easily confirmed that in this case, $e_i[n]$ does not equal $h_i[n] * e[n]$ in general. Thus, we can not easily come up with an equivalent problem formulation in the conventional source coding framework, as we could for non-causal filter case. Of course, in the interest of lowering complexity, which also facilitates analysis, we could use a convenient suboptimal analog estimation stage, as in Sec. 5.6.2. One efficient suboptimal implementation is to estimate the time-domain source $\mathbf{x}[n]$ directly with a causal Wiener filter, and then apply the analyisis filter bank. Clearly in this case, we have $e_i[n] = h_i[n] * e[n]$. Note that this same method can also be applied the case where the analog estimation stage must be FIR. Assuming a form for subband distortion given by (5.176), which may not be entirely accurate [6], we can say that the optimal filter bank is the one that conventionally codes $e[n]$ in an optimal fashion.

It has been shown recently that the optimal FIR filter bank for conventional subband coding is the principle component filterbank (PCFB) [75, 82, 43]. The PCFB performs maximum energy compaction on the source, *i.e.,* for each K, the sum $\sum_{i=0}^{M-1} \sigma_{v_i}^2$ is maximized. There is an analogous result for systematic coding. For systems in which the analog estimation stage satisfies our analytical assumptions, expressed above, the optimal SSC FIR filter bank is the PCFB for the analog estimation error signal, *i.e.* it is the filter bank the performs the maximum energy compaction of the analog estimation error.

[6] As expressed in Sec. 5.6.2, the estimates may be biased, which may cause inaccuracy in the distortion-rate function expression.

Chapter 6

The Duality Between Information Embedding and Source Coding with Side Information

6.1 Introduction

This chapter develops and exploits the theoretical duality between *information embedding*, the robust communication of information embedded into a host signal, and *source coding with side information*. Information embedding has importance for applications such as digital watermarking for copyrighting material and information hiding, or steganography, for covert communications. As described in Chapter 1, source coding with side information may be applied to improving the fidelity of an existing analog communications infrastructure by augmenting the system with a digital side stream. Another potential application for such a coding scheme is for hybrid communication over channels with unknown SNR. Although their applications may seem unrelated, the two problems share a duality at a theoretical level that indicates that a good solution for one problem lends itself easily to a good solution for the other. Some of the implementational aspects of this duality have been explored in [20].

Figure 1-3 shows a block diagram of the information embedding scenario. The n-dimensional vector $\mathbf{y}$ is the *host*, and the variable m is the *information signal*. The encoder uses both the host

and the information signal to create a *composite signal* $\mathbf{w}$, which is assumed to closely approximate the host $\mathbf{y}$. We assume the composite signal is passed through a probabilistic channel. The output of the channel $\mathbf{x}$ is decoded to retrieve an estimate $\hat{m} \approx m$ of the information signal. Throughout this chapter, we assume each element of $\mathbf{y}$ is drawn iid from the distribution $p_y(y)$, and the channel is discrete memoryless, defined by the transition density $p_{x|w}(x|w)$. Illustrated in Fig. 1-3, a feed-forward path is activated when the switch S is closed, thereby supplying the decoder with an observation of the host. In most relevant applications, such as digital watermarking for copyright verification, the switch S is assumed open, which is the case on which this chapter focuses.

The probabilistic channel model for information embedding assumed in this chapter is most accurate for an important subset of information embedding applications. For example, the model is a good one for uniformed attacks. Another instance in which the model is accurate is when the composite signal is sent through a passive analog channel, *e.g.,* a traditional analog communications channel, or in the case of an image host signal, the channel induced by printing and scanning the image. An important channel model for digital watermarking for copyright applications, which is not explored in this thesis, assumes that attacks may be deployed to remove a watermark in which the attacker employs coherent methods using prior knowledge of the modulation scheme. The case when the attacker has perfect knowledge of the modulation scheme, the so-called in-the-clear attack, is addressed in detail in [15].

We use unconventional notation in Fig. 1-3 to clarify the duality between the Wyner-Ziv problem and the previously described information embedding problem. Throughout this chapter it will be clear through context whether a variable belongs to the information embedding problem or the Wyner-Ziv problem. Note that the information embedding *encoder* has exactly the same input variables ($\mathbf{y}$ and m) and output variable ($\mathbf{w}$) as the *decoder* for the Wyner-Ziv problem. Furthermore, the information embedding *decoder* has the same input variable ($\mathbf{x}$) and output variable (m) as the *encoder* for the Wyner-Ziv problem. We show in the sequel through several examples that in fact a optimal encoder/decoder pair for the Wyner-Ziv problem, is an optimal decoder/encoder pair for information embedding.

In Sec. 6.2 we address the question, for a given channel what is the maximum rate at which m can be communicated such that the distortion induced by the embedding is constrained below some desired value? We give an expression for the capacity of an information embedding system

subject to a general distortion constraint on the embedding for the two cases, with and without the host at the decoder. Chen and Wornell [15] have shown these capacity results for a difference distortion constraint. In this chapter the results for information embedding are derived for discrete distributions, and they can be extended to continuous distributions in a straightforward manner. Similar results have been described independently by Moullin *et. al.* in [57]. Our derivations are sufficiently different to warrant presentation here, and they serve to emphasize the duality between information embedding and source coding with side information. In Sec. 6.2 we also determine the necessary and sufficient conditions for capacity to be unaffected by knowledge of the host at the decoder. There exists a duality between these conditions and similar conditions for source coding with side information.

Using the expression for information embedding capacity (with the host known only at the encoder), we derive several significant results. We derive the capacity for a binary symmetric host and a binary symmetric channel with a Hamming distance distortion measure. Furthermore, we present a capacity-achieving method of coding for such a host and channel using nested linear codes. We also consider a Gaussian host and additive Gaussian channel with a mean-squared distortion constraint. The capacity of this system has been previously derived [15, 22] using a capacity result for channels with random state known at the encoder[30, 36], which in this case is equivalent to our capacity result. Using a system of two nested lattices, we construct an encoder/decoder pair that achieves capacity for the Gaussian case. These results are a direct consequence of the dual relationship of information embedding to the Wyner-Ziv problem, and draw heavily on previous results for the Wyner-Ziv problem as described in Chapter 3.

We conclude in Sec. 6.5 by developing a method of signal representation that combines information embedding and Wyner-Ziv encoding. Using this method we can give different decoders varying levels of fidelity, depending on their knowledge of the requisite codebooks. For the quadratic Gaussian case, it is shown that the method (with full knowledge of the codebooks) is optimal in terms of distortion at the output of the decoder.

6.2 Capacity of information embedding systems

In this section we present the expressions for the distortion-constrained information embedding capacity with the host known only at the encoder, and with the host known at both the encoder and decoder.

6.2.1 Host known only at encoder

The capacity of information embedding subject to an embedding distortion constraint with the host known only at the encoder is denoted $C^{\mathrm{IE}}(d)$. It is defined as the maximum achievable rate for communicating a message $\boldsymbol{m}$ such that $P(\hat{\boldsymbol{m}} \neq \boldsymbol{m})$ is arbitrarily small and $E[\frac{1}{n}\sum_{k=1}^{n} D(\boldsymbol{y}_k, \boldsymbol{w}_k)]$ is arbitrarily close to d for large enough n. The following equivalence is shown in the appendix:

$$C^{\mathrm{IE}}(d) = \sup I(\boldsymbol{x}; \boldsymbol{u}) - I(\boldsymbol{y}; \boldsymbol{u}), \tag{6.1}$$

where the supremum is taken over all distributions $p_{\boldsymbol{u}|\boldsymbol{y}}(u|y)$ and functions $f : \mathcal{U} \times \mathcal{Y} \to \mathcal{W}$, where $\boldsymbol{w} = f(\boldsymbol{u}, \boldsymbol{y})$, such that $E[D(\boldsymbol{y}, \boldsymbol{w})] \leq d$. Again $\boldsymbol{u}$ is an auxiliary random variable. Note that $\boldsymbol{y}, \boldsymbol{x}$ and $\boldsymbol{u}$ do not form a Markov chain as they do in the Wyner-Ziv problem.

Previously, Chen and Wornell [15], have shown the capacity of information embedding systems for a difference distortion measure, by assuming that $\mathbf{y}$ is the state of a random "super-channel" known at the encoder. The input to the superchannel is assumed to be the embedding distortion $\mathbf{e} = \mathbf{w} - \mathbf{y}$. The super-channel is characterized as follows; the state $\mathbf{y}$ is added to the input, and the resultant signal, $\mathbf{w}$, is input to the memoryless channel characterized by $p_{\boldsymbol{x}|\boldsymbol{w}}(x|w)$. Gel'fand and Pinsker [30] and Heegard and El Gamal [36] show that the capacity of a channel with random state $\mathbf{s}$ (drawn iid from $p_{\boldsymbol{s}}(s)$) known at the encoder is:

$$C = \sup I(\boldsymbol{u}; \boldsymbol{x}) - I(\boldsymbol{u}; \boldsymbol{s}) \tag{6.2}$$

where the supremum is taken over the distributions $p_{\boldsymbol{u}|\boldsymbol{s}}(u|s)$ and functions $g : \mathcal{U} \times \mathcal{S} \to \mathcal{E}$, where $\boldsymbol{e} = g(\boldsymbol{u}, \boldsymbol{s})$ is the input to the channel. Further constraining the supremum by $E[D(\boldsymbol{e}, 0)] \leq d$ yields the (difference measure) distortion-constrained capacity for information embedding.

We extend this result to general distortion measures. In the appendix the proof for achievability

of the capacity theorem using random codes is viewed as the dual of the achievability theorem for the Wyner-Ziv problem. The encoder (decoder) in the Wyner-Ziv problem is the decoder (encoder) in the information embedding problem. The converse proof for the capacity of information embedding, which relies on the concavity of $C^{\mathrm{IE}}(d)$, is the dual to that for the Wyner-Ziv problem, which relies on the convexity of $R^{\mathrm{WZ}}_{x|y}(d)$. Note that the arguments on the right hand sides of (3.1) and (6.1) are identical.

6.2.2 Host known at encoder and decoder

Information embedding with the host known at the encoder and decoder is the dual of source coding with the side information known at the decoder and encoder. For most relevant applications, the assumption that the host is known at the decoder is unrealistic. One application for which it may be true is information hiding for covert communication. We treat this case mainly to explore the duality with source coding with side information. In Appendix C.2 we derive the expression for information embedding capacity with the host $\mathbf{y}$ known at the encoder and decoder for a general distortion constraint on the embedding. We denote this capacity by $C^{\mathrm{IE}}_{y}(d)$ and it is given by:

$$C^{\mathrm{IE}}_{y}(d) = \sup_{p(w|y)\in\mathcal{P}_{w|y}} I(x; w|y), \tag{6.3}$$

where

$$\mathcal{P}_{w|y} = \{p(w|x) : E[D(y, w)] \leq d\}. \tag{6.4}$$

Again this is an extension of a capacity result for a difference distortion measure based on the super-channel model.

The proof of this result is a dual of the proof of the conditional rate-distortion function as shown in [33] and [7]. Achievability of the rate-distortion function has been proven by a "switching" argument; for each $y \in \mathcal{X}$, an optimal rate-distortion codebook is used to code the source samples x_i for all i such that $y_i = y$. The total rate is thus the expected value over y of the marginal rate-distortion functions, and the distortion is the expected value of the distortions over all the codebooks. Similarly, for information embedding, achievability is proven by "switching" codebooks to a capacity-achieving codebook for each host value $y = y$. The total achievable rate is thus the expected value of the conditional capacities over y, and the average distortion is the expected value

of the distortions over all the codebooks. The converse proof for information embedding with the host known at the decoder exploits the concavity of $C_y^{\mathrm{IE}}(d)$, while the converse for the conditional rate-distortion function exploits the convexity of $R_{x|y}(d)$.

By the problem construction, it is clear that

$$C^{\mathrm{IE}}(d) \leq C_x^{\mathrm{IE}}(d). \tag{6.5}$$

We identify the necessary and sufficient conditions for equality in (6.5) by the following derivation. We let u be determined by some $p(u|x)$ such that for $w = f(u, x)$, we have $E[D(y, w)] \leq d]$. We expand $I(y; u, w|x)$ in two different ways:

$$\begin{aligned} I(y; u, w|x) &= I(u; w|x) + I(y; u|w, x) && (6.6) \\ &= I(y; u|x) + I(y; w|u, x). && (6.7) \end{aligned}$$

Given w, the input to the channel, y and u are conditionally independent, *i.e.*, $u \rightarrow w \rightarrow y$ form a Markov chain, we have that $I(y; u|w, x) = 0$, a condition not satisfied in the Wyner-Ziv problem. Given u and x, w is deterministic, which implies that $I(y; w|u, x) = 0$, a condition also satisfied in the Wyner-Ziv problem. Thus we have that

$$I(y; w|x) = I(y; u|x). \tag{6.8}$$

By straightforward expansion of mutual information expressions, we also have that

$$I(y; u) - I(x; u) \leq I(y; u|x), \tag{6.9}$$

which is met with equality when

$$I(x; u|y) = 0. \tag{6.10}$$

This condition is met automatically in the Wyner-Ziv problem, but is not necessarily true for information embedding.

We have hence shown that (6.5) holds with equality if and only if the maximizing distribution for y, x, u, w in (6.1) also maximizes the argument in (6.3) and, in addition, the condition in (6.10)

holds for the maximizing distribution.

6.2.3 Duality of necessary and sufficient conditions

The relationships discussed in this section exhibit a duality between the Wyner-Ziv problem and information embedding. For source coding with side information (respectively, information embedding) we have presented the necessary and sufficient conditions for the rate-distortion function (respectively, distortion-constrained capacity) to be the same whether or not the side information (respectively, host) is known at the encoder (respectively, decoder). First, the optimizing distributions for x, y, u, w must be the same with or without the signal x known at the encoder (respectively, decoder). The most interesting duality exists in the remaining necessary condition. For the Wyner-Ziv problem, this condition is given by (3.11), which is automatically satisfied by a Markov condition that is a consequence of the problem construction for information embedding. Similarly, the remaining necessary condition for information embedding is given by (6.10), which is automatically satisfied by a Markov condition that is a consequence of Wyner-Ziv problem construction. In both cases, the condition not satisfied by problem construction is in general not satisfied, but we will see in Sec. 6.3 that for both problems, the quadratic Gaussian case satisfies all of the necessary conditions.

Unless otherwise noted, for the remainder of this chapter, we assume that source coding with side information (respectively, information embedding) is performed with the side-information (respectively, host) known only at the decoder (respectively, encoder).

6.3 Quadratic Gaussian case

Equipped with the general expressions for information embedding capacity, in this section we specify the results to the case of Gaussian host, memoryless Gaussian channel, and quadratic distortion metric. There exist dualities in the derivations of these bounds and in the codes that achieve them.

6.3.1 Information embedding capacity

Consider an iid Gaussian host $\mathbf{y} \sim \mathcal{N}(0, \sigma_x^2 I)$ and a channel comprised of additive white Gaussian noise $\mathbf{v} \sim \mathcal{N}(0, \sigma_v^2 I)$ independent from $\mathbf{y}$. The message m is embedded into $\mathbf{y}$, creating a composite

signal $\mathbf{w}$ such that the mean-square embedding distortion is minimized. The capacity $C^{\mathrm{IE}}(d)$ of this system is given by [22]:

$$C^{\mathrm{IE}}(d) = \frac{1}{2}\log\left(1 + \frac{d}{\sigma_v^2}\right). \tag{6.11}$$

Costa proves this result in the context of coding for a channel with a random state known at the encoder. Using a super-channel interpretation of information embedding, Chen and Wornell [15] cite Costa's solution as the solution for the information embedding capacity for the Gaussian case.

In dual fashion to the proof for the rate-distortion function, Costa first proves that the information embedding capacity with $\boldsymbol{y}$ known at the encoder and decoder is,

$$C_{\boldsymbol{y}}^{\mathrm{IE}}(d) = \frac{1}{2}\log\left(1 + \frac{d}{\sigma_v^2}\right). \tag{6.12}$$

He then proceeds to show that with no host at the decoder, there is a test channel which achieves the same capacity as (6.12). Because $C_{\boldsymbol{y}}^{\mathrm{IE}}(d) \geq C^{\mathrm{IE}}(d)$, the expression in (6.12) is also the capacity with no host at the decoder.

The test channel used to determine capacity defines the auxiliary random variable $\boldsymbol{u} = \alpha\boldsymbol{y} + \boldsymbol{e}$ for some constant α, implying that the encoding function is $f(\boldsymbol{u}, \boldsymbol{y}) = \boldsymbol{u} + (1-\alpha)\boldsymbol{y}$. Note that this encoding function is the same as (3.19), the decoding function for the rate-distortion result in the limit of high SNR and for $\beta = 1$. Solving for $I(\boldsymbol{u};\boldsymbol{x}) - I(\boldsymbol{u};\boldsymbol{y})$ and maximizing with respect to α yields (6.11).

6.3.2 Duality

With the above expressions for capacity and the expression for the Wyner-Ziv rate-distortion function in Sec. 3.3, we can show a dual relationship for the Gaussian case. Wyner-Ziv encoding is sphere covering about a source estimate that is a linear function of the side information (shown in Sec. 3.3.3), while information embedding is sphere packing about the host in signal space.

Consider a sequence of n host symbols, *i.e.*, the dimension of $\mathbf{y}$ is n. The distortion constraint on information embedding implies that the all composite signals $\mathbf{w}$ must be contained in a sphere S_x of radius $\sqrt{nd}$ centered about $\mathbf{y}$. In coding for the channel, we use $2^{nR(d)}$ vectors that must contained within S_x such that smaller spheres of radius $\sqrt{n\sigma_v^2}$ about all of the signal points have

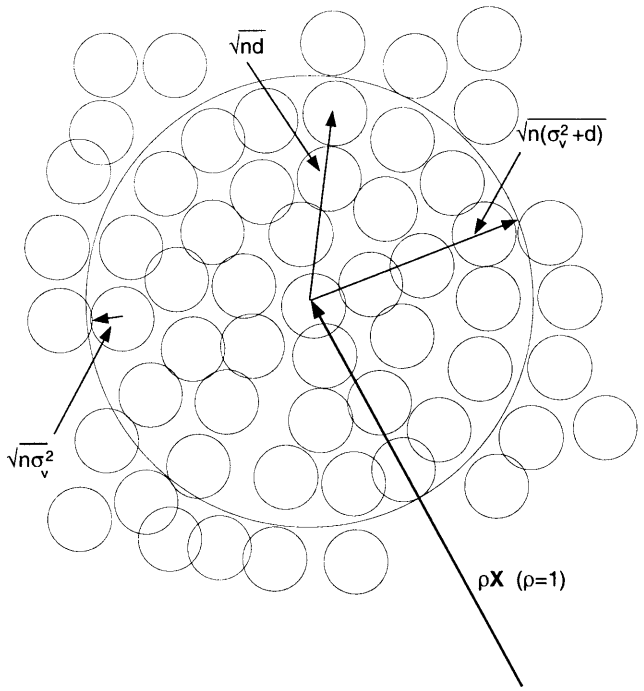


Figure 6-1: Sphere-filling for information embedding in the quadratic Gaussian case.

negligible overlap – each symbol will be uniquely distinguishable at the decoder. Note that this must be true for all $\mathbf{y}$, so that if $\mathbf{y}$ changes by some amount, the positions of signal points may change, but the number of signal points will stay the same. Signal coding corresponds to filling a sphere of radius $\sqrt{n(d+\sigma_v^2)}$ with smaller spheres of radius $\sqrt{n\sigma_v^2}$. This sphere-filling is illustrated in Fig. 6-1. Clearly the maximum number of spheres that can be used is upper bounded by the ratio of the volumes of the large to the small spheres. Thus the number of codewords is bounded:

$$2^{nR(d)} \leq \frac{(\sqrt{n(d+\sigma_v^2)})^n}{(\sqrt{n\sigma_v^2})^n} = \left(\frac{d+\sigma_v^2}{\sigma_v^2}\right)^{n/2}. \tag{6.13}$$

Note from (6.11) that a capacity achieving code will meet this upper bound as

$$2^{nC} = \left(\frac{d+\sigma_v^2}{\sigma_v^2}\right)^{n/2}, \tag{6.14}$$

for large n.

Comparing this section with Sec. 3.3.3, we see again that the encoder (respectively, decoder) operation for Wyner-Ziv is the same as the decoder (respectively, encoder) operation for information

embedding . The Wyner-Ziv encoder finds the nearest neighbor code vector to the source, and transmits the corresponding index. Similarly, the information embedding decoder finds the nearest-neighbor code vector to the channel observation, which corresponds to a decoded message index. At the Wyner-Ziv decoder the digital information m from the coded source specifies a signal point in a sphere about the signal $\rho\mathbf{y}$, and similarly at the information embedding encoder the digital information m specifies a signal point in a sphere about a signal $\mathbf{y}$.

The above sphere filling arguments show the same duality between Wyner-Ziv coding and information embedding as is displayed between rate-distortion theory and conventional channel coding [23]. A good code for one case can be transformed to a good code for the other. In both Wyner-Ziv coding and information embedding, an optimal code will fill signal space such that a sphere about any signal point $\mathbf{t}$ (corresponding to the estimate $\rho\mathbf{y}$ for Wyner-Ziv and the host $\mathbf{y}$ for information embedding) contains an indistinguishable set of codewords that fill the sphere. If a code meets the sphere-packing bound for one problem then it meets the sphere packing bound for the other. We show in the next sections how a previous result for Gaussian codes that achieve the Wyner-Ziv rate-distortion bound can be transformed (by using a Wyner-Ziv encoder (respectively, decoder) for an information embedding decoder (respectively, encoder)) into capacity-achieving codes for information embedding.

6.3.3 Duality between lossless coding and noise-free cases

Another interesting duality follows from the geometric interpretation of the Gaussian case related to the notion of distortion. Note in Fig. 3-2 the radius of the large sphere is proportional to $\sigma_{y|x}$ and the radius of the smaller sphere is proportional to $\sqrt{d}$. In contrast, in Fig. 6-1 the radius of the large sphere is approximately proportional to $\sqrt{d}$, and the radius of the smaller sphere is proportional to σ_v^2. There is a clear duality between the quantization distortion d and channel noise variance σ_v^2 in both problems. Consider the operating point $d = 0$ for Wyner-Ziv encoding (of discrete sources); the problem reduces to the well-known Slepian-Wolf encoding. In information embedding, the dual operating point is the zero channel noise case.

Assuming we have two discrete sources, x and y, and y is communicated at its entropy to a decoder. The Slepian-Wolf discovery states that encoding x with no knowledge of y, one can

reproduce $\boldsymbol{x}$ at the encoder exactly if and only if it is encoded at a rate:

$$R \geq H(\boldsymbol{x}|\boldsymbol{y}). \tag{6.15}$$

A dual equation to (6.15) exists for the information embedding capacity of a system with no channel noise. Given a discrete host $\boldsymbol{y}$ and no channel noise, a message $\boldsymbol{m}$ can be embedded reliably into $\boldsymbol{x}$ at a rate R under the distortion constraint d if and only if [2]

$$R \leq \max_{p_{x|y}(x|y)} H(\boldsymbol{x}|\boldsymbol{y}). \tag{6.16}$$

In [87] Wyner describes a practical Slepian-Wolf code for the case where $\boldsymbol{y}$ is a binary symmetric source and $\boldsymbol{x}$ is related to $\boldsymbol{y}$ by $\boldsymbol{x} = \boldsymbol{y} \oplus \boldsymbol{u}$; $\boldsymbol{u}$ is Bernoulli(p). The nested linear coding method of [66], presented in Sec. 3.4.2, generalized Wyner's method to rate-distortion limit achieving codes for the doubly-binary symmetric case. Letting $d = 0$, we see that the nested linear coding method reduces to the method of Wyner.

In light of the duality between Slepian-wolf coding and noise-free information embedding, by taking the dual to the method of Wyner, *i.e.,* switching the encoder for the decoder and vice versa, we should be able to achieve the noise-free information embedding capacity. Indeed this is true. Using (6.16) we easily determine that under the constraint that the composite signal $\boldsymbol{y}$ be within Hamming distance d of the host $\boldsymbol{x}$,

$$C_{\text{noise-free}} = \max_{p_{y|x}(y|x)} H(\boldsymbol{y}|\boldsymbol{x}) = H(d). \tag{6.17}$$

Letting $p = 0$, we use the nested linear coding method for information embedding described in Sec. 6.4.2 to achieve the rate in (6.17). We easily verify that this code is indeed the dual to Wyner's Slepian-Wolf code in [87].

6.3.4 Lattice codes for information embedding

In this section we use the same lattices as used for Wyner-Ziv encoding in Sec. 3.3.4 for the Gaussian case , with slightly modified parameters, to construct capacity achieving codes for information embedding for the quadratic Gaussian case. Let the parameter $b = \frac{d}{d+\sigma_v^2}$. The lattices will satisfy

the following properties:

a) $\mathcal{L}_2 \subset \mathcal{L}_1$

b) $\sigma_2^2 = \frac{d}{b^2}$

c) $\sigma_1^2 \leq (1-b)^2\sigma_2^2 + \sigma_v^2 + \epsilon$

d) $\mathrm{P}\{Q_2(\mathbf{c} + \mathbf{n} + (1-b)\mathbf{z}^*) \neq \mathbf{c}\} < \epsilon \quad \forall c \in \mathcal{L}_1$
where $\mathbf{z}^*$ is an independent Gaussian with variance σ_2^2

e) $\log(2\pi e G_i) < \epsilon \quad i = 1, 2.$

As in the case of the Wyner-Ziv lattice codes, properties d) and e) assume that lattices $\mathcal{L}_1$ and $\mathcal{L}_2$ are good source/channel codes, possessing good sphere packing properties.

The information signal $\boldsymbol{m}$ takes on some value m drawn uniformly from the indices $\{1, ..., 2^{nR}\}$. Each m corresponds to a unique coset shift $\mathbf{s} = g(m)$. By properties (b), (c), and (e), the rate of the system is

$$R = \frac{1}{n}\log\left(\frac{V_2}{V_1}\right) = \frac{1}{2}\log\left(\frac{\sigma_2^2 G_1}{\sigma_1^2 G_2}\right) \tag{6.18}$$

$$\geq \frac{1}{2}\log\left(\frac{\sigma_v^2 + d}{\sigma_v^2}\right) - O(\epsilon) \tag{6.19}$$

Information embedding encoder: Using the same structure as the Wyner-Ziv *decoder* ((3.40)), the encoder constructs the composite signal $\mathbf{w}$ from the coset shift $\mathbf{s}$ and the host $\mathbf{y}$ by

$$\mathbf{w} = a\mathbf{y} + b\{Q_2(\mathbf{y} + \mathbf{z}_1 - \mathbf{s}) - \mathbf{z}_1 + \mathbf{s}\}. \tag{6.20}$$

Because $\mathbf{s}$ is chosen uniformly across all coset shifts, and $\mathbf{z}_1$ is uniform across $\mathcal{V}_1$ then $\mathbf{z}_2 = \mathbf{s} + \mathbf{z}_1$ is a random vector that is uniform across $\mathcal{V}_2$. Therefore, by the properties of subtractive dithered quantization

$$\mathbf{w} = a\mathbf{y} + b(\mathbf{y} - \mathbf{e}_q), \tag{6.21}$$

where

$$\mathbf{e}_q = \mathbf{y} + \mathbf{z}_2 - Q_2(\mathbf{y} + \mathbf{z}_2) \tag{6.22}$$

is independent of $\mathbf{y}$ and is distributed as $\mathbf{z}_2$, a "Gaussian-like" random variable with variance σ_2^2. Letting $a = (1-b)$ in (6.21) yields

$$\hat{\mathbf{y}} = \mathbf{w} = \mathbf{y} - b\mathbf{e}_q, \tag{6.23}$$

which indicates that the embedding distortion is

$$\frac{1}{n}E\|\mathbf{y} - \mathbf{w}\|^2 = b^2\sigma_2^2 = d \tag{6.24}$$

as desired.

Information embedding decoder: The decoder calculates

$$\begin{aligned}
\mathbf{x}_q &= Q_1(\mathbf{x} + \mathbf{z}_1) && (6.25)\\
&= Q_1(\mathbf{y} - b\mathbf{e}_q + \mathbf{n}) && (6.26)\\
&= Q_1[(\mathbf{y} - \mathbf{e}_q + \mathbf{z}_1) + ((1-b)\mathbf{e}_q + \mathbf{n})] && (6.27)\\
&= Q_1[(Q_2(\mathbf{y} + \mathbf{z}_1 - \mathbf{s}) + \mathbf{s}) + ((1-b)\mathbf{e}_q + \mathbf{n})] && (6.28)\\
&= Q_1[\mathbf{c} + ((1-b)\mathbf{e}_q + \mathbf{n})], && (6.29)
\end{aligned}$$

where (6.28) follows from (6.22), and $\mathbf{c} = Q_2(\mathbf{y} + \mathbf{z}_1 - \mathbf{s}) + S$. By property (d),

$$Q_1[\mathbf{c} + ((1-b)\mathbf{e}_q + \mathbf{n})] = \mathbf{c} \tag{6.30}$$

with probability greater than $1 - \epsilon$. The message estimate $\hat{m}$ is given by

$$\hat{m} = k(\mathbf{x}_q - Q_2(\mathbf{x}_q)), \tag{6.31}$$

which equals $k(\mathbf{s}) = m$ with probability greater than $1 - \epsilon$, thereby proving that the nested lattice codes achieve capacity.

6.4 Binary symmetric channel and source (host) with Hamming distortion metric

In this section we consider the scenario where the signals being communicated and the channels over which they are being communicated are binary symmetric. The host $\mathbf{y}$ is Bernoulli($\frac{1}{2}$). In both cases the channel is a binary symmetric channel (BSC) with crossover probability p. The distortion metric $D(\cdot,\cdot)$ is bit error rate, or Hamming distortion metric.

6.4.1 Capacity expressions

In the appendix we show that when $\mathbf{y}$ is known only at the encoder, the distortion constrained information embedding capacity $C^{\mathrm{IE}}(d)$ for the binary symmetric case is the upper concave envelope of the function $h(d) - h(p)$ and the point $(R, d) = (0, 0)$. Written more precisely,

$$C^{\mathrm{IE}}(d) = \begin{cases} \frac{g(d_c)}{d_c} d, & 0 \le d \le d_c \\ g(d), & d_c < d \le \frac{1}{2} \end{cases} \tag{6.32}$$

$$g(d) = \begin{cases} 0, & 0 \le d < p \\ h(d) - h(p), & p \le d \le \frac{1}{2}, \end{cases} \tag{6.33}$$

where $d_c = 1 - 2^{H(p)}$. Fig. 6-2 shows an example of $C^{\mathrm{IE}}(d)$ for channel transition probability $p = 0.1$. Shown in the figure for comparison is $C_{\mathbf{y}}^{\mathrm{IE}}$, the capacity with $\mathbf{y}$ known at the encoder and decoder ($p = 0.1$). The general expression for this capacity is given by

$$C_{\mathbf{y}}^{\mathrm{IE}}(d) = h(p * d) - h(p), \quad 0 \le d \le \frac{1}{2} \tag{6.34}$$

and its derivation is given in the appendix. Note that $C_{\mathbf{y}}^{\mathrm{IE}}(d) > C^{\mathrm{IE}}(d)$, $0 < d < \frac{1}{2}$. This is not surprising as it is easy to verify that $I(\mathbf{x}; \mathbf{w}|\mathbf{u}\mathbf{y}) \neq 0$ for $0 < d < \frac{1}{2}$. Fig. 6-2 can be compared to Fig. 3-3 which shows the corresponding rate-distortion functions with side information.

6.4.2 Nested linear codes that achieve information embedding capacity

The capacity achieving code, for $\mathbf{y}$ known only at the encoder, will use the same nested linear code construction $\mathcal{C}_1$ and $\mathcal{C}_2$ as in the Wyner-Ziv case in Sec. 3.4.2, except that the dimensions of $\mathbf{H}_1$ and $\mathbf{H}_2$ are now $m_1 \times n$ and $m_2 \times n$ respectively, where $\frac{m_1}{n} = h(p)$ and $\frac{m_2}{n} = h(d)$. Again we

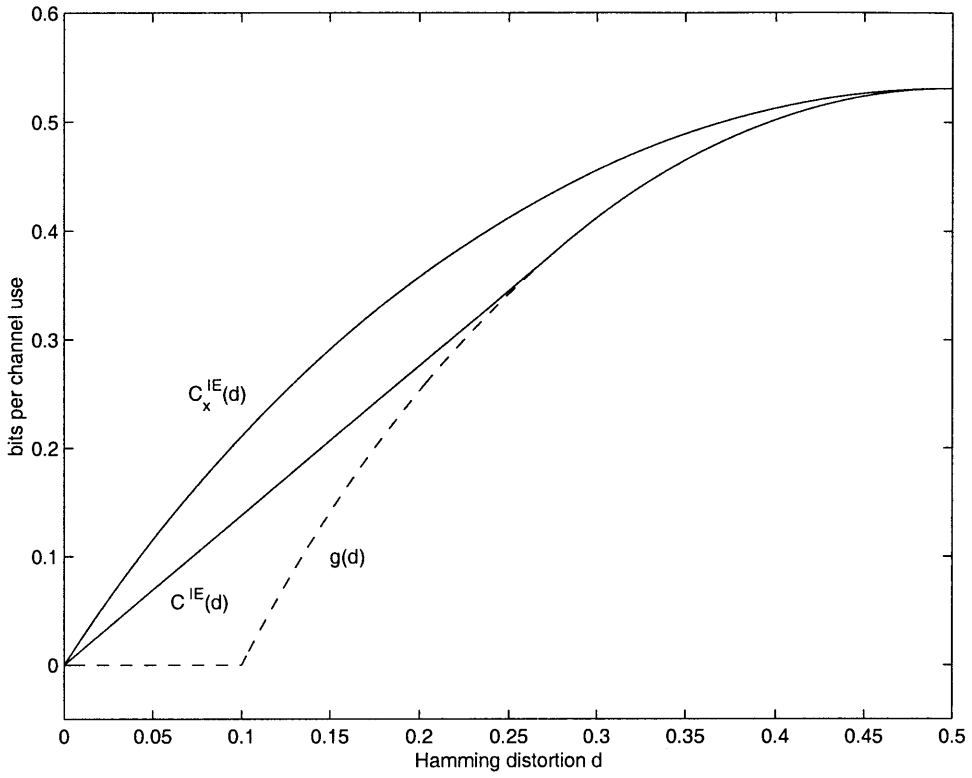


Figure 6-2: The binary symmetric case where the channel transition probability is $p = 0.1$. The dashed line is $g(d) = h(d) - h(p)$. The solid lines are $C^{\mathrm{IE}}(d)$, the upper concave envelope of $g(d)$, and $C_y^{\mathrm{IE}}(d)$.

restrict our attention to bit error rates $d_c \leq d \leq \frac{1}{2}$, as time sharing with no coding can achieve all other operating points on the capacity curve.

The information embedding encoder. Revisiting a common theme, we use the structure for the Wyner-Ziv decoder for the information embedding encoder. We first observe that there are $2^{m_2-m_1}$ signals of the form $\mathbf{H}_a\mathbf{c}^T$ for some $\mathbf{c}^T \in \mathcal{C}_1$. We let the rate of the information signal m be

$$R = \frac{m_2 - m_1}{n} = h(d) - h(p) = C^{\mathrm{IE}}(d). \tag{6.35}$$

Thus any message m can be written in the form $\mathbf{H}_a\mathbf{c}^T$ for some $\mathbf{c} \in \mathcal{C}_1$. The encoder finds the composite signal $\mathbf{w} \in \mathcal{C}_1$ that is closest in Hamming distance to the host $\mathbf{y}$, such that $\mathbf{H}_a\mathbf{w} = m$ for the desired message $m = m$. We write the host as $\mathbf{y} = \mathbf{w} \oplus \mathbf{e}_q$, where $\mathbf{e}_q$ is Bernoulli(d). This is true by the following. The set comprising all $\mathbf{w}$, which is constrained to have $\mathbf{H}_a\mathbf{w} = m$, is simply the codewords of $\mathcal{C}_2$ shifted by a constant vector $\mathbf{k}$. Because $\mathbf{y}$ is a symmetric source, we can consider

the simple case $\mathbf{w} \in \mathcal{C}_2$ without loss of generality. The code $\mathcal{C}_2$ can decode any vector $\mathbf{y}$ as the sum of the nearest codeword to $\mathbf{y}$ and a Bernoulli(d) error vector $\mathbf{e}_q$. Thus the embedding error $\mathbf{e}_q$ is Bernoulli(d), and the distortion constraint is met. Using the equation

$$\mathbf{H}_2\mathbf{w}^T = \begin{bmatrix} \mathbf{0} \\ \mathbf{H}_a\mathbf{w}^T \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \text{Bin}(m) \end{bmatrix}, \tag{6.36}$$

where Bin(m) is the length n binary expansion of m, the encoder calculates the composite signal

$$\mathbf{w} = \mathbf{y} \oplus f_2(\mathbf{H}_2\mathbf{w}^T \oplus \mathbf{H}_2\mathbf{y}^T) = \mathbf{y} \oplus f_2(\mathbf{H}_2\mathbf{e}_q^T) = \mathbf{y} \oplus \mathbf{e}_q, \tag{6.37}$$

and transmits it.

The information embedding decoder. The decoder receives the signal $\mathbf{x} = \mathbf{w} \oplus \mathbf{n}$, where $\mathbf{n}$ is Bernoulli(p). By construction, $\mathcal{C}_1$ can correct any Bernoulli(p) errors, and $\mathbf{w} \in \mathcal{C}_1$. Therefore, the optimal (closest Hamming distance) estimate, $\hat{\mathbf{w}}$, of $\mathbf{w}$ from $\mathbf{x}$ equals $\mathbf{w}$ with near certainty. The encoder calculates $\mathbf{H}_a\hat{\mathbf{w}}$ which thus equals the information signal m with near certainty.

6.5 Combined information embedding and Wyner-Ziv rate distortion coding

Information embedding robustly communicates information at some prescribed rate to a receiver. One potential use for this communicated information is to improve the fidelity of the received signal. At the encoder, one performs Wyner-Ziv encoding on the host at a given distortion and embeds the coded bits into the host. As long as the rate of the encoding is below the information embedding capacity, the encoded bitstream can be embedded into the host and decoded reliably. Of course, the Wyner-Ziv encoding must take into account the additional distortion induced by the embedding.

This method of hybrid coding spans the continuum from a purely analog signal representation to a purely digital signal representation. Points along the continuum have varying degrees of embedding, with the purely analog signal having no embedding and the purely digital signal having the maximum amount of embedding, effectively destroying the host. In most cases an analog signal

offers ease of decoding at the receiver, requiring for example only simple AM or FM tuner. A digital signal on the other hand offers in some cases better fidelity, but also increased security, either through the use of a private codebook or some form of encryption. It is conceivable that in a broadcast scenario a broadcaster may wish to strike a balance between ease of decoding, at the expense of poor fidelity, for some users, and secure, high fidelity decoding for other users who have the digital codebook. This method of hybrid coding also offers a method to smoothly transition over time from an all analog transmission infrastructure to an all digital one. Obviously, this problem is non-trivial only when the host is not known at the decoder. We look at both the quadratic Gaussian case and the binary symmetric case.

We first note that at the encoder in addition to observing the source $\mathbf{x}$, we observe the composite signal w, which is correlated to the channel output y. Thus w is a form of partial feedback of the side information to the encoder. Also, at the decoder we decode the information signal m, which is a function of the source $\mathbf{x}$. Thus, we have a partial feed-forward path of the host to the decoder. In this chapter the coding results for information embedding and source coding with side information are not tailored for partial feedback or feed-forward signals. For the Gaussian case this is not important, as we need only use information embedding with the host at the encoder, and Wyner-Ziv encoding to achieve minimum distortion at the decoder. This fact is true because for the Gaussian case, even full feedback of the side information does not improve the rate-distortion function, and full feed-forward of the host does not improve the information embedding capacity. For the binary symmetric case feedback and feed-forward paths actually improve the rate distortion function and capacity. Therefore, using information embedding assuming the host is known at the encoder only and Wyner-Ziv encoding is potentially suboptimal. Evaluating this suboptimal hybrid coding scheme is useful, however, to assess achievable performance.

6.5.1 Quadratic Gaussian Case

Embedded Signal Representation

Consider an iid Gaussian host $\mathbf{x} \sim \mathcal{N}(0, \sigma_x^2)$ and a channel comprised of additive white Gaussian noise $\mathbf{v}_c \sim \mathcal{N}(0, \sigma_{v_c}^2)$. We evaluate the performance of any coding system by the distortion at the receiver. Clearly an all analog representation, *i.e.*, sending the host uncoded through the channel,

will yield a distortion at the receiver, denoted by d_r, given by

$$d_r = \sigma^2_{x|y} = \frac{\sigma_x^2 \sigma^2_{v_c}}{\sigma_x^2 + \sigma^2_{v_c}}, \tag{6.38}$$

For a fair comparison, we consider an all digital representation with its power input to the channel constrained to be σ_x^2. By the source-channel separation theorem, the minimum d_r is attained by channel coding to capacity, and separately rate-distortion coding the source to that rate for transmission. Thus the value for the minimum d_r is obtained from the following equations:

$$\frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{\sigma^2_{v_c}}\right) = C = R(d_r) = \frac{1}{2}\log\left(\frac{\sigma_x^2}{d_r}\right), \tag{6.39}$$

whose solution is also given by (6.38). Thus all-analog and all-digital give the same distortion at the receiver. The source-channel separation theorem tells us that no other coding method can achieve a better distortion. To achieve the flexibility described at the outset of Sec. 6.5, however, we may choose to used a combined embedding/Wyner-Ziv encoding approach. In this section we establish that the combined approach for any embedding distortion will achieve the performance of the all-analog and all-digital coding methods. This property is special to the Gaussian case, and is in not true in general as we shall see from the binary symmetric example in Sec. 6.5.2.

If we embed under a distortion constraint d_e using a capacity achieving code, the embedding adds zero-mean iid Gaussian noise $\mathbf{v}_e$ of variance d_e, independent of $\mathbf{x}$. In order to keep the overall power constrained to σ_x^2, the host $\mathbf{x}$ must be multiplied by

$$K = \sqrt{\frac{\sigma_x^2 - d_e}{\sigma_x^2}} \tag{6.40}$$

prior to embedding. Thus at the receiver the observed signal is

$$\mathbf{x} = K\mathbf{x} + \mathbf{v}_e + \mathbf{v}_c. \tag{6.41}$$

The linear least squared error estimate of $\mathbf{x}$ from $\mathbf{y}$, has an error variance of

$$\sigma^2_{x|y} = \sigma_x^2 - \frac{K^2\sigma_x^4}{K^2\sigma_x^2 + \sigma^2_{v_c} + d_e}. \tag{6.42}$$

Assuming d_e is fixed, the maximum achievable embedding rate is $C^{\mathrm{IE}}(d_e)$. Given this supplied data rate, we wish to code for minimum distortion at the decoder. The minimum Wyner-Ziv encoding rate that achieves the distortion d_r for the Gaussian case is

$$R^{\mathrm{WZ}}_{x|y}(d_r) = \frac{1}{2}\log\Big(\frac{\sigma^2_{x|y}}{d_r}\Big), \tag{6.43}$$

The minimum achievable d_r is determined by the following equation:

$$R^{\mathrm{WZ}}_{x|y}(d_r) = C^{\mathrm{IE}}(d_e). \tag{6.44}$$

Solving for d_r as a function of d_e we have

$$d_r = \frac{\sigma^2_x \sigma^2_{v_c}}{\sigma^2_x + \sigma^2_{v_c}}, \tag{6.45}$$

which is the same reconstruction distortion as that achieved by the all-analog and all-digital encodings. Thus, the reconstruction error is the same for all levels of embedding $0 \leq d_e \leq \sigma^2_x$. Note that we are effectively embedding one code (the information signal) on top of another code (the host signal). Since the combined analog-digital system achieves the distortion of an all-digital system, the embedded coding system effectively achieves the rate of the all-digital system, *i.e.*, the capacity of the channel. This is consistent with the result in [15] that shows a layered coding method using information embedding on top of a conventional channel code achieves capacity for the Gaussian channel.

Multi-layer embedding

Hybrid coding can be generalized to multiple layers of embedding. Let there be n embeddings, each corresponding to an independent noise vector $\mathbf{v}_i \sim \mathcal{N}(0, \sigma^2_v I)$, $i = 1, ..., n$. We define n composite signals $\mathbf{w}_i$, $i = 1, ..., n$ as follows:

$$\mathbf{w}_0 = \mathbf{x} \tag{6.46}$$

$$\mathbf{w}_i = K_i \mathbf{w}_{i-1} + \mathbf{v}_i, \quad i = 1, ..., n, \tag{6.47}$$

where the K_i are constant gains that normalize the powers of the $\mathbf{w}_i$s to equal σ_x^2:

$$K_i = \sqrt{\frac{\sigma_x^2 - \sigma_{v_i}^2}{\sigma_x^2}}. \tag{6.48}$$

At each layer i, information is embedded into the previous composite signal $\mathbf{w}_{i-1}$, and the information is used to Wyner-Ziv encode the same signal $\mathbf{w}_{i-1}$. The n^{th} composite signal $\mathbf{w}_n$ is input to the channel comprised of AWGN vector $\mathbf{v}_c$. For every layer the embedding is done assuming a channel variance σ_{v_c}. Decoding the signal is performed as follows. There are n codebooks $\mathcal{C}_i$, $i = 0, ..., n-1$, of which the last r are available to a particular decoder. The n^{th} embedding code is decoded from the channel output $\mathbf{y}$, and the bits are used to form an estimate $\hat{\mathbf{w}}_n$ of $\mathbf{w}_n$ by Wyner-Ziv decoding. By the results from Sec. 6.5.1, the variance of the estimate is given by (6.38). We proceed to form an estimate $\hat{\mathbf{w}}_{n-1}$ from the observation $\hat{\mathbf{w}}_n$, the estimate from the previous layer. Again, the variance of this estimate is given by (6.38). This process is continued until $\hat{\mathbf{w}}_{n-r}$ is formed by decoding with codebook $\mathcal{C}_{n-r}$. If there are $r = n$ codebooks available to the decoder, the source is reconstructed to a fidelity given by (6.38). the same as that achieved by the all-digital or all-analog representation. Thus with all of the codebooks at the decoder, we incur no distortion penalty by performing multi-layer embedding.

Consider a coding strategy which lets $\sigma_{v_i} = d_e$, $i = 1, ..., n$, equals a constant, implying that the $K_i = K$, $i = 1, ..., n$, equals a constant given by (6.40). If at a particular decoder, we have only the last r codebooks, we can decode down to the $(n-r)^{\text{th}}$ layer, obtaining $\hat{\mathbf{w}}_{n-r}$. Equivalently, the highest fidelity signal that we observe is

$$\mathbf{x}_{n-r} = \mathbf{w}_{n-r} + \mathbf{q}_{n-r}, \tag{6.49}$$

where the $\mathbf{q}_{n-r}$ is iid Gaussian noise of variance σ_{v_c}. Expanding $\mathbf{w}_{n-r}$ according to the iteration in (6.47), we have

$$\mathbf{y}_{n-r} = K^{n-r}\mathbf{y} + \sum_{i=0}^{n-r-1} K^i \mathbf{v}_{n-r-i} + \mathbf{q}_{n-r}. \tag{6.50}$$

The minimum mean-square estimation error of $\mathbf{x}$ from $\mathbf{y}_{n-r}$ is

$$\text{MSE} = \frac{\sigma_x^2(\sigma_x^2(1 - K^{2(n-r)}) + \sigma_{v_c}^2)}{\sigma_x^2 + \sigma_{v_c}^2}. \tag{6.51}$$

Thus if a given decoder has only the last r codebooks, it decodes the source to the fidelity given by (6.51). Note that the distortion decays exponentially with the number of codebooks available to the receiver, and the time constant of the decay increases linearly with $\log K^2$; the value K^2 decreases linearly with increasing d_e, the single-layer embedding distortion. Letting the difference $(n-r) \to \infty$ (which requires that $n \to \infty$), the distortion at the decoder approaches σ_y^2. At the other extreme point, $r = n$, when the decoder has access to all codebooks, the distortion is given by (6.38).

The multi-layer embedding strategy allows for us to prescribe one of many levels of fidelity to a given decoder, based on the number of codebooks to which it has access. For the Gaussian case, if all of the codebooks are available to the decoder, we pay no penalty in terms of distortion at the decoder compared to purely analog or purely digital encoding. Thus, this method is a novel and potentially useful successive refinement strategy using a solitary signal representation.

6.5.2 Binary symmetric case

In the binary symmetric case we develop the same combined "analog-digital" coding technique. Although the binary symmetric source is inherently digital, we refer to it as "analog" in the context of our discussion of combined coding. The "analog channel" is binary symmetric with crossover probability p. An "all-analog" system (sending the source uncoded) yields a reconstruction distortion $d_r = p$. For an all-digital system, the capacity is $C = 1 - H(p)$. The ordinary rate-distortion function for a binary symmetric source is $R(d_r) = 1 - H(d_r)$. Setting $R(d_r) = C$ the distortion of an all-digital system is clearly $d_r = p$. As in the Gaussian case, the all-digital system and all-analog system have identical performance. Interestingly, we prove that unlike the Gaussian case, the combined coding scheme for the binary symmetric case yields a greater distortion than all-digital or all-analog coding. Letting d_e denote the embedding distortion, we explore the continuum from $d_e = 0$ (all-analog coding) and $d_e = \frac{1}{2}$ (all-digital coding). Note that $d_e = \frac{1}{2}$ is effectively all-digital, even though it implies we are embedding; the capacity of the embedding system is $1 - H(p)$, which is the same as capacity of an all-digital system, and the Wyner-Ziv rate-distortion function is $1 - H(d_r)$.

As stated in the introduction of Sec. 6.5, we observe the signals $\mathbf{w}$ at the encoder and $\mathbf{m}$ at the decoder, which are correlated to the $\mathbf{x}$ and $\mathbf{y}$ respectively, *i.e.,* there are partial feedback

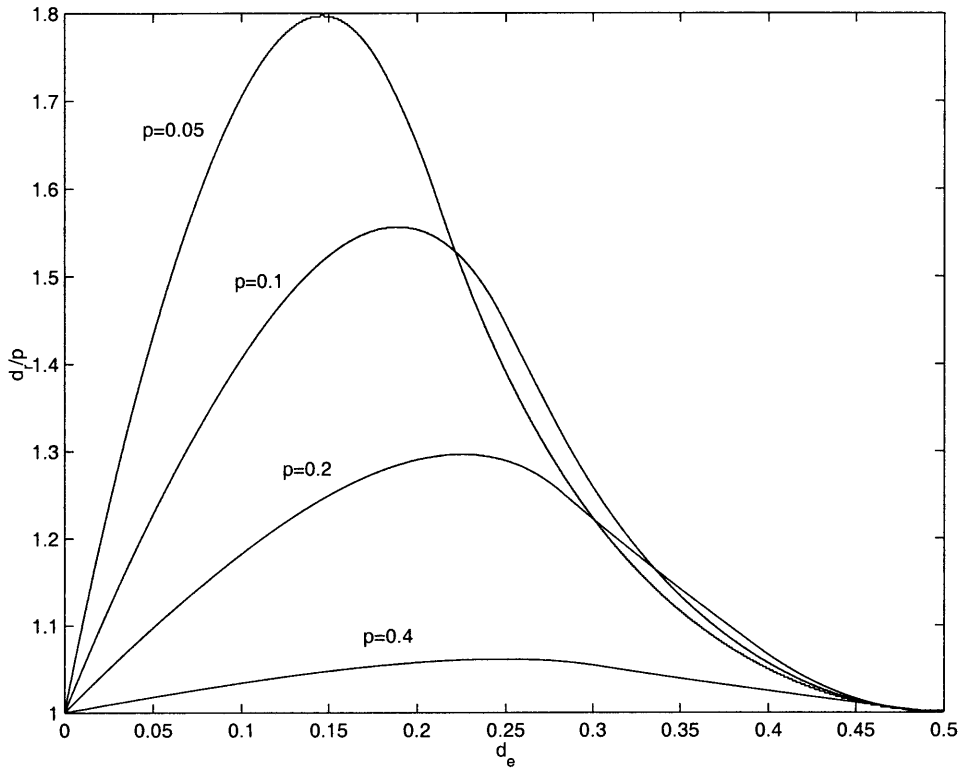


Figure 6-3: Combined analog-digital coding for the binary symmetric case. Plotted is the reconstruction distortion (normalized by p) as a function of embedding distortion for $p = 0.05, 0.1, 0.2, 0.4$.

and feed-forward paths. Ignoring these signals could potentially hurt the performance of hybrid coding scheme. For simplicity, we first develop a provably suboptimal coding scheme that uses information embedding, assuming no information about the host at the decoder, and Wyner-Ziv source coding (assuming no information about the side information at the encoder). We then discuss the performance of an optimal system.

Constrained to have an embedding distortion of d_e, we have a capacity of $C^{\mathrm{IE}}(d_e)$. As shown in the Appendix C, any capacity achieving embedding code constrained to distortion d_e will act on the source as a BSC with crossover probability d_e. Thus the combined effect of the embedding and the channel will be a BSC with crossover probability $d_e * p$. For Wyner-Ziv encoding, the side-information is the source (host) plus a Bernoulli($d_e * p$) process. Denoting the reconstruction distortion by d_r, the Wyner-Ziv rate-distortion function $R^{\mathrm{WZ}}_{x|y}(d_r)$ is thus the lower convex envelope of the point $(R^{\mathrm{WZ}}_{x|y}, d_r) = (0, d_e * p)$ and $h(d_r * d_e * p) - h(d_r)$. The reconstruction distortion of the

combined analog-digital coding scheme is obtained by setting $R^{\mathrm{WZ}}_{x|y}(d_r) = C^{\mathrm{IE}}(d_e)$, and solving for d_r in terms of d_e. Plotted in Fig. 6-3 is the reconstruction distortion (normalized by p) as a function of embedding distortion for various channel crossover probabilities. Note that for all p plotted, the reconstruction distortion for $0 < d_e < 1/2$ is strictly higher than that for $d_e = 0$ or $d_e = 1/2$. We see that this will always be true for any $p \in (0, 1/2)$ by the following analysis.

For the binary symmetric case, the source coder performance benefits from knowing $\mathbf{y}$ at the encoder, and the information embedding performance benefits from knowing $\mathbf{x}$ at the decoder. Although the feedback and feed-forward paths are only partial in the combined coding case, we consider the optimal performance of a system whose source coding component has full feedback of $\mathbf{y}$ and whose information embedding component has full feedforward $\mathbf{x}$. This system obviously upper bounds the performance of any combined information embedding/source coding system. From (6.34), the information embedding capacity is $C^{\mathrm{IE}}_{y}(d_e) = h(p * d_e) - h(p)$, $0 \leq d \leq 1/2$. The embedding encoder and the channel act as a composite channel that is additive Bernoulli$(d_e * p)$. From (3.48) the rate-distortion function is thus $R_{x|y}(d_r) = h(d_e * p) - h(d_r)$, $0 \leq d \leq p$. Equating $C^{\mathrm{IE}}_{x}(d_e)$ and $R_{x|y}(d_r)$, we get $d_r = p$. The optimal performance of this unrealizable coding system is only as good as the all-analog and all-digital approaches. What we have derived, then, is a trivial lower bound on the performance of a hybrid coding scheme. There is a benefit to this analysis, however. As seen in Fig. 6-3, for $0 < d_e < 1/2$, $C^{\mathrm{IE}}_{y}(d_e)$ is strictly greater than $C^{\mathrm{IE}}(d_e)$. Furthermore, for $0 < d_r < p$, $R^{\mathrm{WZ}}_{x|y}(d_r)$ is strictly decreasing and is strictly greater than $R_{x|y}(d_r)$. What follows is that for the suboptimal combined coding method, d_r is strictly greater than p (because $d_r = p$ for the idealized case), as confirmed empirically in Fig. 6-3.

We have developed combined coding using information embedding and Wyner-Ziv encoding, which has potential applications including the secure transmission of information. In both the Gaussian case and binary symmetric case that we considered the source is iid and the channel is memoryless, which by the separation theorem insists that no coding scheme can outperform the all-digital method. In the Gaussian case combined coding achieves the same performance as all-digital coding, while in the binary symmetric case, combined coding does strictly worse. Future work may consider scenarios in which the assumptions of the separation theorem do not hold. The combined coding proposed here could potentially outperform both all-analog and all-digital methods that do separate source and channel coding.

6.6 Conclusion

The main conclusion of this chapter is that a good information embedding system can be constructed from a good systematic source coding system by using the source coding encoder (respectively, decoder) for the information embedding decoder (respectively, encoder). This intuition yields a method for proving the capacity of information embedding systems both with and without the host at the decoder. For the Gaussian case, the capacity had been previously derived. We derived the capacity for the binary symmetric case. For both cases we used codes that achieve the rate-distortion bound, and switched the encoder for a decoder and vice versa, to find codes that achieve information embedding capacity. Finally, we have seen that Wyner-Ziv encoding and information embedding can be combined to create a novel method for signal representation.

Chapter 7

Conclusions and Future Directions

The focus of this thesis was to develop low-latency, low-complexity signal processing solutions for a special case of distributed source coding, known as systematic source coding, or source coding with side information at the decoder. We worked within a framework comprised of three basic encoding elements, 1) transformation, 2) quantization and lossless bitstream coding, that has for many years proven very successful for conventional source coding.

7.1 Contributions

As a first step to designing such systems, in Chap. 2, for a fixed encoder, we derived the optimal decoder structures based on the MAP, MMSE, and ML estimation criteria. We followed with two main examples for particular sources and channels: 1) a Gaussian source and channel with linear side information, for which we derived the MAP, MMSE, and ML estimators, and 2) a speech source and AWGN channel with spectral envelope side information, for which we derived the ML estimator. We then introduced the basic concepts of systematic digital encoding, from which the nature of the duality with information embedding is clear. The systematic source coding encoder behaves as the information embedding decoder and vice versa.

Chap. 3 reviewed the fundamental discoveries of the research community on the information theoretical aspects of source coding with side information at the decoder. The rate-distortion limit for this form of source coding is given by the Wyner-Ziv rate distortion bound, whose particular form has been determined for the iid Gaussian case and the doubly binary symmetric case. We

extended the Gaussian result to the case of stationary, jointly Gaussian source and channel output processes. We presented the nested linear lattice codes of Zamir and Shamai that achieve the rate-distortion limit for the Gaussian case in the limit of high channel SNR. We generalized this result to all channel SNRs.

Chap. 4 developed the core source coding element, the systematic quantizer. For cases with and without feedback, we developed iterative algorithms, analogous to the Lloyd-Max algorithm for conventional quantizers, for the design of locally optimal systematic scalar quantizers in terms of MSE. Focusing our attention on the scalar quantizer case, we show that a low-complexity quantizer, called the NLSQ, has near-optimal performance. We showed that combining NLSQ encoding and a Slepian-Wolf coding, achieves within .255 bits/sample of the rate-distortion bound, for iid pairs $(\mathbf{x}, \mathbf{y})$. The results were extended to show similar performance when applied appropriately to stationary, jointly Gaussian processes $\boldsymbol{x}[n]$ and $\boldsymbol{y}[n]$. Finally we showed an exponential coding gain of NLSQs over an *ad hoc* method called low-bits coding.

The transformation stage of our source coding framework can take many forms. In Chap. 5, the transformations that we analyzed were linear block transforms, overlapped transforms, and subband decomposition. We focused our attention on the Gaussian case: mean-squared distortion and stationary, jointly Gaussian source and channel output processes. For our proposed method of systematic source coding, the quantization stage after transformation uses optimal NLSQs on the subband coefficients. Using the formula for the distortion-rate function of the NLSQ from Chap 4, we derived the optimal strategy to assign bits to the NLSQs for each transform coefficient. We then derived optimality criteria for each of the types of transformation.

In Chap. 6 we developed, at an information-theoretic level, the duality between source coding with side information at the encoder with information embedding. The main result was that a good Wyner-Ziv encoder (respectively, decoder) is a good information embedding decoder (respectively, encoder). In particular, we showed that, for the Gaussian and binary symmetric cases, codes that achieve the Wyner-Ziv rate-distortion limit can be mapped to codes that achieve information embedding capacity. We concluded by developing a novel form of signal coding which combines information embedding and Wyner-Ziv encoding.

7.2 Future directions

Many of the results in this thesis are very general. For some of the results, however, especially those relating to transform coding, we have relied on the tractable properties of Gaussian sources and channels. Thus, an obvious open problem is how to extend the results of this thesis to non-Gaussian cases.

In [4] we have implemented the ideas in this thesis to design a practical systematic subband audio coder, well-suited for the in-band, on-channel digital audio broadcast problem. Future work could consider the systematic coding of other sources such as images and video.

As stated at the outset of Chap. 1, systematic source coding is only a small subset of the rich field of distribute source coding. Perhaps some of the ideas developed for this special case can be extended to distributed source coding in general.

It is clear that a good solution for systematic source coding lends itself easily to a good solution for information embedding. The door is now open to identify a good solution for one problem and map it to a solution for the other.

Appendix A

Capacity of distortion constrained information embedding

In this appendix, we prove a single letter expression for the capacity of an information embedding system for a host $\mathbf{y}$ drawn iid from $p_{\mathsf{y}}(y)$ and discrete memoryless channel $p_{\mathsf{x}|\mathsf{x}}(x|z)$. The information signal $\boldsymbol{m}$ is assumed independent from the host. For a length n vector $\mathbf{v}$ we use the notation V_j^k to denote a vector comprised of the jth to the kth components of $\mathbf{v}$. If the subscript is omitted, j is implicitly 1. We prove that

$$C^{\mathrm{IE}}(d) = \max I(\boldsymbol{u};\mathsf{x}) - I(\boldsymbol{u};\mathsf{y}), \tag{A.1}$$

where the maximum is taken over all distributions $p_{\boldsymbol{uw}|\mathsf{y}}(u,w|y)$. It is shown in [30] that because $C^{\mathrm{IE}}(d)$ is convex in the distribution $p_{\boldsymbol{w}|\boldsymbol{uy}}$ then the distribution is deterministic, simplifying Eq. A.1 such that the maximum is taken over all distributions $p_{\boldsymbol{u}|\mathsf{y}}(u|y)$ and functions $f : \mathcal{U} \times \mathcal{X} \to \mathcal{Z}$, where $\boldsymbol{w} = f(\boldsymbol{u},\mathsf{y})$.

A.1 Converse

The converse proof shows that for any rate $R \geq C^{\mathrm{IE}}(d)$ the maximal probability of error for a length n code, denoted $\mathrm{P}_e^{(n)}$, is bounded away from zero. In order to prove the converse we must first show that $C^{\mathrm{IE}}(d)$ is a non-decreasing concave function of d. Clearly $C^{\mathrm{IE}}(d)$ is a non-decreasing function, as increasing d increases the domain over which the maximization is performed.

We next provide a proof of concavity which is virtually identical to the proof for convexity of $R_{x|y}^{\mathrm{WZ}}(d)$. Consider two distortions d_1 and d_2 and the corresponding arguments, u_1, f_1 and u_2, f_2 respectively, which maximize Eq. A.1 for the given distortion. Let q be a random variable independent of y, x, u_1, and u_2, that takes on the value 1 with probability λ and the value 2 with probability $1-\lambda$. Define $z = (q, u_q)$ and let $f(z, y) = f_q(u_q, y)$, implying a distortion

$$\begin{aligned} d &= E[D(y, w)] && \text{(A.2)} \\ &= \lambda E[D(y, f_1(u_1, y)) + (1-\lambda)E[D(y, f_2(u_2, y))] && \text{(A.3)} \\ &= \lambda d_1 + (1-\lambda)d_2, && \text{(A.4)} \end{aligned}$$

and

$$\begin{aligned} I(z;x) - I(z;y) &= H(x) - H(x|z) - H(y) + H(x|z) && \text{(A.5)} \\ &= H(x) - H(x|u_q, q) - H(y) + H(y|u_q, q) && \text{(A.6)} \\ &= H(x) - \lambda H(x|u_1) - (1-\lambda)H(x|z_2) - H(y) \\ &\quad -\lambda H(y|u_1) + (1-\lambda)H(y|u_2) && \text{(A.7)} \\ &= \lambda(I(u_1;x) - I(u_1;y)) + (1-\lambda)(I(u_2;x) - I(u_2;y)). && \text{(A.8)} \end{aligned}$$

Thus

$$\begin{aligned} C^{\mathrm{IE}}(d) &= \max_{u,f:E[D(y,f(u,y))]\leq d} (I(u;x) - I(u;y)) && \text{(A.9)} \\ &\geq I(z;x) - I(z;y) && \text{(A.10)} \\ &= \lambda(I(u_1;x) - I(u_1;y)) + (1-\lambda)(I(u_2;x) - I(u_2;y)) && \text{(A.11)} \\ &= \lambda C^{\mathrm{IE}}(d_1) + (1-\lambda)C^{\mathrm{IE}}(d_2), && \text{(A.12)} \end{aligned}$$

proving the concavity of $C^{\mathrm{IE}}(d)$.

Proof of the converse. Consider an information embedding code, with an encoding function $f_n : \mathcal{X}^n \times \{1, ..., 2^{nR}\} \to \mathcal{W}^n$ and a decoding function $g_n : \mathcal{Y}^n \to \{1, ..., 2^{nR}\}$. Let $f_{ni} : \mathcal{X}^n \times \{1, ..., 2^{nR}\} \to \mathcal{W}$ denote the ith symbol produced by the encoding function. The distortion

constraint is

$$\frac{1}{n}E[\sum_{i=1}^{n} D(y_i, f_{ni}(y^n, m))] \leq d. \tag{A.13}$$

We have the following chain of inequalities:

$$\begin{aligned}
nR &= H(m) && \text{(A.14)}\\
&= I(m; x^n) + H(m|x^n) && \text{(A.15)}\\
&= I(m; x^n) - I(m; y^n) + H(m|x^n) && \text{(A.16)}\\
&\leq \sum_{i=1}^{n}(I(z_i, x_i) - I(z_i; y_i)) + H(m|x^n) && \text{(A.17)}\\
&\leq \sum_{i=1}^{n} C^{\mathrm{IE}}(E[D(y_i, f'_{ni}(y_i, z_i))]) + H(m|x^n) && \text{(A.18)}\\
&= n\frac{1}{n}\sum_{i=1}^{n} C^{\mathrm{IE}}(E[D(y_i, f'_{ni}(y_i, z_i))]) + H(m|x^n) && \text{(A.19)}\\
&\leq nC^{\mathrm{IE}}(E[\frac{1}{n}\sum_{i=1}^{n} D(y_i, f'_{ni}(y_i, z_i))]) + H(m|x^n) && \text{(A.20)}\\
&\leq C^{\mathrm{IE}}(d) + H(m|x^n), && \text{(A.21)}\\
&\leq C^{\mathrm{IE}}(d) + \mathrm{P}_e^{(n)} nR + 1, && \text{(A.22)}
\end{aligned}$$

where

(A.14) follows from our assumption that m is distributed uniformly in $\{1, ..., 2^{nR}\}$,

(A.16) follows from the fact that $I(m; y^n) = 0$ by independence,

(A.17) follows from [30] Lemma 4, where z_i is defined as $z_i = (m, x^{i-1}, y_{i+1}^n)$,

(A.18) follows from the definition of $C^{\mathrm{IE}}(d)$,

(A.20) follows from Jensen's inequality and the concavity of $C^{\mathrm{IE}}(d)$,

(A.21) follows from the definition $d = E[\frac{1}{n}\sum_{i=1}^{n} D(y_i, w_i)]$, and

(A.22) follows from the Fano inequality,

Rearranging terms we have

$$\mathrm{P}_e^{(n)} \geq 1 - \frac{C^{\mathrm{IE}}(d)}{R} - \frac{1}{nR}, \tag{A.23}$$

which shows for $R > C$, the probability or error is bounded away from 0.

A.2 Achievability

Using random codes and typical set encoding/decoding, we show that there exists a code operating at rate R arbitrarily close to $C^{\mathrm{IE}}(d)$ that achieves an arbitrarily small probability of error. We use the notation $A_\epsilon^{*(n)}$ to denote the set of jointly typical sequences of length n.

Consider a fixed $p_{u|y}(u|y)$ and function $f(u,y)$, such that $E[D(y,f(u,y))] \leq d$. The marginal for u is thus $p_u(u) = \sum_y p(y)p(u|y)$. We use a random coding argument exploiting the following construction:

- Generation of codebook. Let $R_1 = I(u;x) + \epsilon$. Generate a random codebook $\mathcal{C}$ with 2^{nR_1} iid codewords $u^n \sim \prod_{i=1}^n p_u(u_i)$, and index them by $s \in S_1 = \{1, ..., 2^{nR_1}\}$.

  Let $R_2 = I(u;x) - I(u;y) - \epsilon$. Consider 2^{nR_2} bins of codewords indexed by $S_2 = \{1, ..., 2^{nR_2}\}$. For each index $s \in S_1$ draw a random variable uniform in $\{1, ..., 2^{nR_2}\}$. This is the bin to which we assign the codeword indexed by s. This procedure effectively creates 2^{nR_2} codebooks, each corresponding to an index $s_2 \in S_2$. We denote each codebook by $\mathcal{C}_i$, $i \in S_2$. There are approximately $2^{n(R_1-R_2)}$ codewords in each codebook.

- Encoding. The message signal $m = m$ specifies a codebook $\mathcal{C}_m$ that we will use. Since there are $R_2 = I(u;x) - I(u;y) - \epsilon$ codebooks the rate of the code is R_2. For a given a host sequence y^n, the encoder looks for a codeword u^n in $\mathcal{C}_i$ that satisfies $(u^n, y^n) \in A_\epsilon^{*(n)}$. If no such codeword exists the encoder selects a codeword at random from $\mathcal{C}_i$ to transmit. If more than one such codeword exists we select one of these codewords at random. Given the selected codeword, the composite signal w^n that serves as input to the channel is calculated by $w_i = f(u_i, y_i)$.

- Decoding. The decoder looks for a u^n that satisfies $(u^n, x^n) \in A_\epsilon^{*(n)}$. If there is a unique u^n then the reconstructed information sequence $\hat{m} = i$ where $i \in S_2$ is the index of the codebook that contains u^n. If there is no u^n that satisfies joint typicality, or there is more than one, then the decoder assigns the index $\hat{m} = 2$.

- Probability of error. Without loss of generality, in calculating the probability of error, we can assume the message $m = 1$ is being communicated. Therefore all codewords will come from the codebook $\mathcal{C}_1$. We consider 4 possible error events and show that each one contributes negligibly to the probability of error:

  1. The pair (w^n, x^n) is not an element of $A_\epsilon^{*(n)}$. Since w^n and x^n are the input and output respectively of a DMC, this event has negligible probability by the weak law of large numbers.

  2. The host y^n is typical, but there exists no $u^n \in \mathcal{C}_1$ such that $(u^n, y^n) \in A_\epsilon^{*(n)}$. Note that because the message is $m = 1$, we are essentially rate-distortion coding the host y^n with the codebook $\mathcal{C}_1$. Therefore the achievability theorem for the conventional rate-distortion tells us that the probability of this event has negligible probability if the number of codewords, $2^{n(R_1-R_2)}$ in $\mathcal{C}_1$ exceeds $I(u;y)$, which is true by construction.

  3. The pair (y^n, u^n) is jointly typical, but (u^n, x^n) is not jointly typical. By item (1) $(w^n, x^n) \in A_\epsilon^{*(n)}$. Clearly, $u \to w \to x$ form a Markov chain, and by definition of w^n, u^n is drawn $\sim \prod_{i=1}^n p_{u_i|w_i}(u_i|w_i)$. By the Markov Lemma, Lemma 14.8.1 in [23], $(u^n, w^n, x^n) \in A_\epsilon^{*(n)}$, which means that the probability of this event is negligible for large enough n.

  4. There exists another $c^n \in \mathcal{C}$, $c^n \neq u^n$, such that $(c^n, x^n) \in A_\epsilon^{*(n)}$. The probability that a particular random codeword in $\mathcal{C}$ is jointly typical with x^n is $2^{-n(I(u;x)-3\epsilon)}$. Using a union bound argument, the probability that *any* random codeword is jointly typical with u^n is upper bounded by the quantity

  $$2^{nR_1} 2^{-n(I(u;x)-3\epsilon)} \tag{A.24}$$

  Since $R_1 > I(u;x)$, the probability of this event is arbitrarily small.

  5. The empirical distribution is not arbitrarily close to $p_{uxy}(u,x,y)$. This event is not true by the following. If u^n is decoded correctly, then $(u^n, y^n) \in A_\epsilon^{*(n)}$. Therefore we have the Markov chain $(y,u) \to w \to x$. By item (1) $(w^n, x^n) \in A_\epsilon^{*(n)}$. Since $w_i = f(u_i, y_i)$, by the Markov Lemma, we have $(u^n, y^n, z^n, x^n) \in A_\epsilon^{*(n)}$, which implies that the empirical joint distribution for (u,y,x) matches the theoretical distribution $p_{uxy}(u,x,y)$. The desired

distortion d is thus achieved.

Given that all of the error events have negligible probability, we have shown that there exists a code that can achieve R_2, which equals $C^{\mathrm{IE}}(d)$ if we maximize over all distributions, and meets the distortion constraint.

Appendix B

Capacity of information embedding with the host known at the encoder and decoder

B.1 Converse

The proof of the converse uses a technique very similar to that used in Appendix A, exploiting the concavity of $C_y^{\mathrm{IE}}(d)$, a fact which is proven in the following Lemma.

Lemma. The information embedding capacity given in Eq.6.3 is a non-decreasing, concave function of the distortion constraint d.

Proof. With increasing d, the domain over which the mutual information is maximized increases, which implies $C_y^{\mathrm{IE}}(d)$ is non-decreasing.

We prove convexity by considering two capacity-distortion pairs (C_1, d_1) and (C_2, d_2), which are points on the information embedding capacity function. These points are achieved with the distributions $p_1(w,x,y) = p(y)p(x|w)p_1(w|y)$ and $p_2(w,x,y) = p(y)p(x|w)p_2(w|y)$ respectively. We define

$$p_\lambda(w,x,y) = \lambda p_1(w,x,y) + (1-\lambda)p_2(w,x,y). \tag{B.1}$$

Because distortion is a linear function of the transition probabilities, the distortion for p_λ is

$$d_\lambda = \lambda d_1 + (1-\lambda) d_2. \tag{B.2}$$

It is easily verified that the mutual information $I(w; x|y=y)$ is a concave function of the distribution $p(w|x)$. Therefore,

$$I_{p_\lambda}(w; x|y=y) \geq \lambda I_{p_1}(w; x|y=y) + (1-\lambda) I_{p_2}(w; x|y=y), \tag{B.3}$$

where we subscript the mutual informations with their respective distributions. Thus we have the following chain of inequalities:

$$\begin{aligned}
C_y^{\mathrm{IE}}(d_\lambda) &\geq I_{p_\lambda}(w; x|y) && \text{(B.4)}\\
&= \sum_y I_{p_\lambda}(w; x|y=y) p(y) && \text{(B.5)}\\
&= \sum_y \lambda I_{p_1}(w; x|y=y) p(y) + \sum_y (1-\lambda) I_{p_2}(w; x|y=y) p(y) && \text{(B.6)}\\
&\geq \lambda C_y^{\mathrm{IE}}(d_1) + (1-\lambda) C_y^{\mathrm{IE}}(d_2), && \text{(B.7)}
\end{aligned}$$

which proves the concavity of $C_y^{\mathrm{IE}}(d)$.

The converse. Recall the input to the channel is the composite signal w^n, which is an encoded function of the host y^n and the message m. The distortion between y^n and w^n is constrained by $\frac{1}{n} E[\sum_{i=1}^n D(y_i, w_i)] \leq d$. The converse is proven by the following chain of inequalities:

$$\begin{aligned}
nR &= H(m) && \text{(B.8)}\\
&= H(m|y^n) && \text{(B.9)}\\
&= I(m; x^n|y^n) + H(m|x^n y^n) && \text{(B.10)}\\
&\leq \sum_{i=1}^n I(m; x_i|y^n, x^{i-1}) + H(m|y^n x^n) && \text{(B.11)}\\
&\leq \sum_{i=1}^n (H(x_i|y^n, x^{i-1}) - H(x_i|m, y^n, x^{i-1})) + H(m|y^n x^n) && \text{(B.12)}\\
&\leq \sum_{i=1}^n (H(x_i|y_i) - H(x_i|m, y^n, x^{i-1})) + H(m|y^n x^n) && \text{(B.13)}
\end{aligned}$$

$$\leq \sum_{i=1}^{n}(H(x_i|y_i) - H(x_i|m, y^n, y_i) + H(m|y^n x^n) \quad \text{(B.14)}$$

$$\leq \sum_{i=1}^{n} I(x_i; m, y^n|y_i) + H(m|y^n x^n) \quad \text{(B.15)}$$

$$\leq \sum_{i=1}^{n} I(x_i; w_i|y_i) + H(m|y^n x^n) \quad \text{(B.16)}$$

$$\leq \sum_{i=1}^{n} C_y^{\mathrm{IE}}(E[D(y_i, w_i)]) + H(m|y^n x^n) \quad \text{(B.17)}$$

$$= n\frac{1}{n}\sum_{i=1}^{n} C_y^{\mathrm{IE}}(E[D(y_i, w_i)]) + H(m|y^n x^n) \quad \text{(B.18)}$$

$$\leq nC_y^{\mathrm{IE}}(E[\frac{1}{n}\sum_{i=1}^{n} D(y_i, w_i)]) + H(m|y^n x^n) \quad \text{(B.19)}$$

$$\leq C_y^{\mathrm{IE}}(d) + H(m|y^n x^n), \quad \text{(B.20)}$$

$$\leq C_y^{\mathrm{IE}}(d) + \mathrm{P}_e^{(n)} nR + 1 \quad \text{(B.21)}$$

where

(B.8) follows from our assumption that m is distributed uniformly in $\{1, ..., 2^{nR}\}$,

(B.9) follows from the independence of m and y^n,

(B.11) follows from the chain rule for mutual information,

(B.13) follows from the fact that conditioning reduces mutual information,

(B.16) follows from the data processing inequality, using the fact that $(m, y^n) \rightarrow w_i \rightarrow x_i$ is a Markov chain.

(B.17) follows from the definition of $C_y^{\mathrm{IE}}(d)$,

(B.19) follows from Jensen's inequality and the concavity of $C_y^{\mathrm{IE}}(d)$,

(B.20) follows from the definition $d = E[\frac{1}{n}\sum_{i=1}^{n} D(y_i, w_i)]$, and

(B.21) follows from the Fano inequality.

B.2 Achievability

The proof of achievability follows from the proceeding lemma which relates C_y^{IE} to a weighted sum of conditional embedding capacities. A conditional embedding capacity, $C_y^{\mathrm{IE}}(d_y)$ is the capacity of a channel when the host y is some fixed value y known at the encoder and decoder. By the conventional channel capacity theorem:

$$C_y^{\mathrm{IE}}(d_x) = \sup_{p(w|y)\in\mathcal{P}_{\mathsf{w}|y}} I(\mathsf{x};\mathsf{w}|\mathsf{y}=y), \tag{B.22}$$

where

$$\mathcal{P}_{\mathsf{w}|y} = \{p(w|x) : E[D(y,\mathsf{w}|\mathsf{y}=y)] \le d_y\} \tag{B.23}$$

is the constraint set for the embedding.

Lemma B. The information embedding capacity with host known at the encoder and decoder satisfies

$$C_y^{\mathrm{IE}}(d) = \sup_{\{d_y:E[d_y]=d\}} \sum_{y\in\mathcal{X}} C_y^{\mathrm{IE}}(d_y)p_{\mathsf{y}}(y). \tag{B.24}$$

Proof. First choose a fixed set of d_ys for each y such that $E[d_y] = d$ and a test channel $p(w|y) \in \mathcal{P}_{\mathsf{w}|y}$. It is easily confirmed that

$$E[D(\mathsf{y},\mathsf{w})] \le E[d_y] = d, \tag{B.25}$$

which implies $p(w|y) \in \mathcal{P}_{\mathsf{w}|\mathsf{y}}$. For any test channel,

$$\sum_y I(\mathsf{w};\mathsf{x}|\mathsf{y}=y)p_{\mathsf{y}}(y) = I(\mathsf{w};\mathsf{x}|\mathsf{y}) \le C_y^{\mathrm{IE}}, \tag{B.26}$$

so that choosing $p(w|y)$ to satisfy the maximization in Eq. B.22 yields

$$\sum C_y^{\mathrm{IE}}(d_y)p_{\mathsf{y}}(y) \le C_y^{\mathrm{IE}}(d) \tag{B.27}$$

for any set of d_y satisfying $E[d_y] = d$.

Next we prove the opposite inequality, thereby proving the lemma. We choose a test channel $p(w|y) \in \mathcal{P}_{\mathsf{w}|y}$, which will result in a set of conditional distortions $\bar{d}_y = E[D(\mathsf{y},\mathsf{w})|\mathsf{y}=y]$ that

satisfy $E[\bar{d}_y] \leq d$. For any such test channel,

$$\begin{aligned} I(w;x|y) &= \sum_y I(w;x|y)p_y(y) && \text{(B.28)} \\ &\leq \sum_y C_{\mathrm{IE}}^y(\bar{d}_y)p_y(y) && \text{(B.29)} \\ &\leq \sup_{\{d_y:E[d_y]\leq d\}} \sum_y C_{\mathrm{IE}}^y(d_y)p_y(y) && \text{(B.30)} \end{aligned}$$

Choosing $p(w|x)$ to achieve the maximum in Eq. 6.3 yields

$$C_y^{\mathrm{IE}}(d) \leq \sup_{\{d_y:E[d_y]=d\}} \sum_{y\in\mathcal{X}} C_{\mathrm{IE}}^y(d_y)p_y(y), \tag{B.31}$$

which together with Eq. B.27 proves the lemma.

Consider the set of d_y^* that achieves the maximum on the right hand side of Eq. B.24. By the conventional channel coding theorem, we can achieve the rate $C_{\mathrm{IE}}^y(d_y^*)$ with embedding distortion d_y^* and negligible probability of error if $X = x$ for all samples of data. Consider the following coding scheme. We embed data in y^n, a length n block of host samples. We have a different codebook for each y which achieves the rate $C_{\mathrm{IE}}^y(d_y^*)$ at embedding distortion d_x^*. For each y, we collect all of the samples y_i for each i such that $y_i = y$ and code using the codebook corresponding to x. The total rate is thus

$$\sum_{y\in\mathcal{X}} C_{\mathrm{IE}}^y(d_y^*)p_y(y) = \sup_{\{d_y:Ed_y=d\}} \sum_{y\in\mathcal{X}} C_{\mathrm{IE}}^y(d_y)p_y(y) \tag{B.32}$$

which by the lemma equals capacity.

Appendix C

Capacity of information embedding for binary symmetric host/ channel with Hamming distortion constraint

C.1 Host know only at encoder

In this appendix, we derive $C^{\text{IE}}(d)$, for the binary symmetric case subject to a Hamming distortion constraint described in Sec. 6.4. We determine that it is the upper concave envelope of the function $g(d)$ where $g(d)$ is defined as in Eq. 6.33:

$$g(d) = \begin{cases} 0, & 0 \leq d < p \\ h(d) - h(p), & p \leq d \leq \frac{1}{2}. \end{cases} \tag{C.1}$$

We encourage the reader to review [90] to see the strong similarities between this proof and the one proving the Wyner-Ziv rate-distortion function for the binary symmetric case. The upper concave envelope of $g(d)$ is given by

$$g^*(d) = \sup_{\theta,\beta_1,\beta_2} [\theta g(\beta_1) + (1-\theta) g(\beta_2)]. \tag{C.2}$$

where the supremum is taken with respect to all $\theta \in [0,1]$ and $\beta_1, \beta_2 \in [0, \frac{1}{2}]$ such that $d = \theta\beta_1 + (1-\theta)\beta_2$. By the concavity of $h(\cdot)$, it is clear that $g^*(d)$ is concave over $p \leq d \leq \frac{1}{2}$. Thus the

maximization in Eq. C.2 can be simplified by letting $\beta_2 = 0$:

$$g^*(d) = \sup_{\theta,\beta}[\theta(h(\beta) - h(p))], \qquad 0 \leq d \leq \frac{1}{2} \tag{C.3}$$

where the supremum is taken with respect to all $\theta \in [0,1]$ and $\beta \in [0, \frac{1}{2}]$ such that $d = \theta\beta$.

We establish that $C^{\mathrm{IE}}(d) = g^*(d)$ by separately proving $C^{\mathrm{IE}}(d) \leq g^*(d)$ and $C^{\mathrm{IE}}(d) \geq g^*(d)$.

C.1.1 Proof that $C^{\mathrm{IE}}(d) \geq g^*(d)$

The proof of the lower bound for $C^{\mathrm{IE}}(d)$ is attained by considering a special case. Recall $\boldsymbol{u}$ is an auxiliary random variable from which we attain $\boldsymbol{w} = f(\boldsymbol{u}, \boldsymbol{y})$. Let $\boldsymbol{u}$ be the output of a binary symmetric channel with crossover probability β into which $\boldsymbol{y}$ serves as input. Also let $\boldsymbol{w} = f(\boldsymbol{u}, \boldsymbol{y}) = \boldsymbol{u}$, which makes the distortion equal β. We evaluate

$$I(\boldsymbol{x}, \boldsymbol{u}) - I(\boldsymbol{y}, \boldsymbol{u}) = I(\boldsymbol{x}, \boldsymbol{w}) - I(\boldsymbol{y}, \boldsymbol{u}) = (1 - h(p)) - (1 - h(\beta)) = h(\beta) - h(p), \tag{C.4}$$

and conclude from Eq. 6.1 that

$$C^{\mathrm{IE}} \geq h(\beta) - h(p). \tag{C.5}$$

The values $\theta \in [0,1]$ and $\beta \in [0, \frac{1}{2}]$ satisfy

$$d = \theta\beta \tag{C.6}$$

for some given $d \in [0, \frac{1}{2}]$. By the concavity of $C^{\mathrm{IE}}(d)$

$$C^{\mathrm{IE}}(d) \;=\; C^{\mathrm{IE}}(\theta\beta) \geq \theta C^{\mathrm{IE}}(\beta) \geq \theta(h(\beta) - h(p)), \tag{C.7}$$

which is true for all θ and β satisfying Eq. C.6. Thus, $C^{\mathrm{IE}}(d) \geq g^*(d)$.

C.1.2 Proof that $C^{\mathrm{IE}}(d) \leq g^*(d)$

We show that $C^{\mathrm{IE}}(d) \leq g^*(d)$ by showing that

$$I(\boldsymbol{u}; \boldsymbol{x}) - I(\boldsymbol{u}; \boldsymbol{y}) \leq g^*(d) \tag{C.8}$$

for any $p_{uw|y}(u, w|y)$ such that $E[D(y, w)] = d$.

Define the set

$$A = \{u : f(0, u) = f(1, u)\} \tag{C.9}$$

and its complement

$$A^C = \mathcal{U} - A = \{u : f(0, u) \neq f(1, u)\}. \tag{C.10}$$

We observe that

$$\begin{aligned} I(u; x) - I(u; y) &= H(x) - H(x|u) - H(y) + H(y|u) && \text{(C.11)} \\ &= H(y|u) - H(x|u), && \text{(C.12)} \end{aligned}$$

where Eq. C.12 is true because $H(y) = H(x) = 1$. Without loss of generality we can thus restrict our attention to functions f such that if $f(0, u) \neq f(1, u)$ then $f(0, u) = 0$ and $f(1, u) = 1$. This is true because any other choice of f will not change $I(u; x) - I(u; y)$ and can only increase the d. Thus,

$$A^C = \{u : f(0, u) \neq f(1, u)\} = \{u : f(0, u) = 0,\ f(1, u) = 1\}. \tag{C.13}$$

By the given information,

$$\begin{aligned} E[D(y, z)] &= \mathrm{P}(u \in A)E[D(y, z)|u \in A] + \mathrm{P}(u \in A^C)E[D(y, z)|u \in A^C] && \text{(C.14)} \\ &= \mathrm{P}(u \in A)E[D(y, z)|u \in A] && \text{(C.15)} \\ &\leq d, && \text{(C.16)} \end{aligned}$$

where Eq. C.15 is true because $E[D(y, z)|u \in A^C] = 0$. The equation

$$E[D(y, z)|u \in A] = \sum_{u \in A} \frac{\mathrm{P}(u = u)}{\mathrm{P}(u \in A)} E[D(y, z)|u = u] \tag{C.17}$$

and Eq. C.15 yield

$$d' = \theta \sum_{u \in A} \lambda_u d_u \leq d, \tag{C.18}$$

where $\theta = \mathrm{P}(U \in A)$, $\lambda_u = \mathrm{P}(\boldsymbol{u} = u)/\mathrm{P}(U \in A)$, and

$$d_u = E[D(\boldsymbol{y}, \boldsymbol{z})|\boldsymbol{u} = u]. \tag{C.19}$$

We observe that

$$\begin{aligned} I(\boldsymbol{u};\boldsymbol{x}) - I(\boldsymbol{u};\boldsymbol{y}) &= H(\boldsymbol{y}|\boldsymbol{u}) - H(\boldsymbol{x}|\boldsymbol{u}) && \text{(C.20)} \\ &\leq \sum_{u \in A} (H(\boldsymbol{y}|\boldsymbol{u} = u) - H(\boldsymbol{x}|\boldsymbol{u} = u))\mathrm{P}(\boldsymbol{u} = u) && \text{(C.21)} \\ &= \theta \sum_{u \in A} \lambda_u (H(\boldsymbol{y}|\boldsymbol{u} = u) - H(\boldsymbol{x}|\boldsymbol{u} = u)), && \text{(C.22)} \end{aligned}$$

where the inequality in Eq. C.21 is true because for $u \in A^C$, $H(\boldsymbol{y}|\boldsymbol{u} = u) - H(\boldsymbol{x}|\boldsymbol{u} = u) = -H(\boldsymbol{x}|\boldsymbol{u} = u) \leq 0$.

Assume $u \in A$. Defining $\gamma(u) = f(0, u) = f(1, u)$, we simplify Eq.C.19:

$$d_u = E[D(\boldsymbol{y}, \boldsymbol{z})|\boldsymbol{u} = u] = \mathrm{P}(X \neq \gamma(u)|\boldsymbol{u} = u). \tag{C.23}$$

Clearly, $H(\boldsymbol{y}|\boldsymbol{u} = u) = h(d_u)$. Given $\boldsymbol{u} = u$ the channel input is uniquely specified by $\boldsymbol{z} = \gamma(u)$, and thus $H(\boldsymbol{x}|\boldsymbol{u} = u) = h(p)$. We rewrite Eq. C.22 as

$$\begin{aligned} I(\boldsymbol{u};\boldsymbol{x}) - I(\boldsymbol{u};\boldsymbol{y}) &\leq \theta \sum_{u \in A} \lambda_u (h(d_u) - h(p)) && \text{(C.24)} \\ &= \theta \sum_{u \in A} \lambda_u G(d_u), && \text{(C.25)} \end{aligned}$$

where $G(v) = h(v) - h(p)$, which is clearly concave for $0 \leq v \leq \frac{1}{2}$. Because G is concave, and $\sum_{u \in A} \lambda_u = 1$,

$$I(\boldsymbol{u};\boldsymbol{x}) - I(\boldsymbol{u};\boldsymbol{y}) \leq \theta G(\sum_{u \in A} \lambda_u d_u) = \theta(h(\beta) - h(p)), \tag{C.26}$$

where

$$\beta = \sum_{u \in A} \lambda_u d_u. \tag{C.27}$$

Thus we have shown that for any distribution $p_{uz|y}(u,z|y)$ there exists a $\theta \in [0,1]$ and $\beta \in [0,\frac{1}{2}]$ such that Eq. C.26 holds and from Eq. C.18, $\theta\beta = d'$. From the definition of g^* (Eq. C.2), we have

$$I(u;x) - I(u;y) \leq g^*(d'). \tag{C.28}$$

implying Eq. C.8, because $d' \leq d$ and $g^*(d)$ is nondecreasing in d. The proof is complete.

C.2 Host known at encoder and decoder

Proof of the information embedding capacity for the binary symmetric case with the host known at the encoder and decoder is much simpler than the above proof. Consider the general expression for capacity given in Eq. B.22. Using the fact that subtracting a known constant from w and x does not affect their mutual information, we have the equivalent expression

$$C_y^{\mathrm{IE}}(d) = \sup_{p(e|y)} I(x-y;e|y), \tag{C.29}$$

where $e = w - y$ is the additive distortion due to encoding, which is constrained to have $P(e = 1) \leq d$. Note that $x - y = e + v$, where v is a Bernoulli(p) source representing the additive noise of the BSC. Under the constraint that $P(e = 1) \leq d$, we have the following chain of inequalities:

$$\begin{aligned} I(x-y;e|y) &= H(e+v|y) - H(e+v|e,y) && \text{(C.30)} \\ &\leq H(e+v) - \sum_e p(e)H(e+v|e=e,y) && \text{(C.31)} \\ &= H(e+v) - \sum_e p(e)H(p) && \text{(C.32)} \\ &\leq H(p*d) - H(p). && \text{(C.33)} \end{aligned}$$

The inequalities are met with equality if e is Bernoulli(d), independent of y and v. Thus $C_y^{\mathrm{IE}}(d) = H(p*d) - H(p)$, $0 \leq d \leq \frac{1}{2}$.

Bibliography

[1] K.C. Aas, and C.T. Mullis, "Minimum mean-squared error transform coding and subband coding," *IEEE Trans. Inform. Theory*, vol. 42, no. 4, pp. 1179-1191, July 1996.

[2] R.J. Barron, B. Chen and G.W. Wornell, "The duality between information embedding and source coding with side information and some implications," submitted to *IEEE Trans. Inform. Theory.*

[3] R.J. Barron,and A.V. Oppenheim, "Signal processing for hybrid channels," *Proceedings of 3rd Annual ARL Fedlabs Symposium* pp. 481-484., Feb. 1999.

[4] R.J. Barron and A.V. Oppenheim, "A Systematic Hybrid Analog/Digital Audio Coder," *Proceedings of the Workshop on the Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, October, 1999.

[5] R.J. Barron, C.K. Sestok,and A.V. Oppenheim, "Speech enhancement using spectral envelope side information," *IEEE Trans. on Acoustic, Speech and Signal Processing*, **ASSP-36**, pp. , 1998.

[6] J.M. Barton, "Method and apparatus for embedding authentication information within digital data," United States Patent #5,646,997, Issued July 8,1997.

[7] Berger, T., *Rate distortion theory*, Englewood, New Jersey, Prentice Hall, 1971.

[8] Berger, T., Housewright, K.B., Omura, J.K., Tung, S., Wolfowitz, J., "An upper bound on the rate distortion function for source coding with partial side information at the decoder", *IEEE Trans. Inform. Theory*, **IT-25**, pp. 664-666, November 1979.

[9] Berger, T., Chang, M.U., "Rate-distortion with a fully informed encoder and partially informed encoder," *Proc. IEEE International Symposium on Information Theory*, pp. 34, 1977.

[10] Berger, T., Zhang, Z., Viswanathan, H., "The CEO problem multiterminal source coding," *IEEE Trans. Inform. Theory*, **IT-42**, pp. 887-902, May 1996.

[11] Boyle, J.P., and Dykstra, R.L., "A method for finding projections onto the intersection of convex sets in Hilbert space," *Lecture Notes in Statistics*, **37**, pp. 28-47, 1986.

[12] J.D. Bruce, "Optimum quantization," *Research Lab. of Eleectronics, Mass. Inst. of Tech., Technical Report,* no. 469, March 1965.

[13] P. Cassereau, "A new class of optimal unitary transforms for image processing," Master's Thesis, Mass. Inst. Tech., Cambridge, MA, May 1985.

[14] B. Chen, C.-E. W. Sundberg, "Broadcasting data in the FM band by means of adaptive contiguous band insertion and precancelling techniques," *Proc. 1999 IEEE International Conference on Communications*, Vancouver, Canada, June 1999.

[15] Chen, B., Wornell, G.W., "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," to appear in *IEEE Trans. Inform. Theory.*

[16] R.J. Clarke, "Relation between the Karhunen-Loeve and cosine transforms," *Proc. IEE*, Part F, vol. 128, pp.359-360, Nov. 1981.

[17] R. J. Clarke, *Transform Coding of Images.* Number 4 in Microelectronics and Signal Processing. Academic Press, London, England, 1985.

[18] Cheney, A., and Goldstein, A.A., "Proximity maps for convex sets," *Proc. Amer. Math. Soc.*, **2**, pp. 448-450, 1959.

[19] B. Chitpraset and K.R. Rao, "Discrete cosine transform filtering," *IEEE Trans. Acoust., Speech, Signal Proc.*, **ASSP-35**, pp. 818-824, June 1987.

[20] J. Chou, S.S. Pradhan, and K. Ramchandran, "On the duality between distributed source Coding and data hiding", *Proceedings of the Thirty-third Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 1999.

[21] J. Conway,and N.J.A. Sloane, *Sphere packings, lattices, and groups,* Spring Verlag, New York, N.Y., 1988.

[22] M.H. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, pp. 439-441, May 1983.

[23] T. Cover, J.A. Thomas, *Elements of information theory*, Jon Wiley & Sons, Inc., 1991.

[24] T.M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, **IT-21**, no. 2, pp. 226-228, March 1975.

[25] S. Cramer and R. Gluth, "Computationally efficient real-valued filter banks based on O^2DFT," *Proc. European Signal Processing Conf.*, Barcelona, Spain, pp. 585-588, Sept. 1990.

[26] P. Duhamel, Y. Mahieux, and J.P. Petit, "A fast algorithm for the implemetation of filter banks based on time domain aliasing cancelation," *Proc. ICASSP-91, Int. Conf. on Acoustics Speech and Signal Processing,* pp. 2209-2212, May 1991.

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

[28] P.E. Fleisher, "Sufficient conditions for achieving minimum distortion in a quantizer," *IEEE Int. Conv. Rec., Part 1,* pp. 104-111, 1964.

[29] G.D. Forney, "The Viterbi algorithm," *Proc. of IEEE*, vol. 61, no. 3, pp. 268-78, March 1973.

[30] Gel'fand, S.I., Pinsker, M.S., "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol.9, no. 1, pp. 19-31, 1980.

[31] Gersho, A., Gray, R.M., "Vector quantization and signal compression," Kluwer Academic Publishers, Boston, 1991.

[32] Goblick, T.J., Holsinger, J.L., "Analog source digitization:a comparison of theory and practice," *IEEE Trans. Inform. Theory*, **IT-14**, pp. 676-683, 1967.

[33] Gray, R.M., "Conditional rate-distortion theory," *Stanford Electronics Laboratories Technical Report*, No. 6502-2, October 1972.

[34] Gray, R.M., "Source coding theory," Kluwer Academic Publishers, Boston, 1990.

[35] Gray, R.M., "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, **IT-19**, pp. 480-489, July 1973.

[36] Heegard, C., El Gamal, A., "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, 1983.

[37] Y. Huang, "Quantization of correlated random variables," Ph.D. Dissertation, Yale University, New Haven, Conn., 1962.

[38] Y. Huang, P.M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. on Commun. Systems*, pp.289-296, September 1963.

[39] International Standards Organization, "ISO/IEC 11172-3 MPEG I audio coding standard," pp. 129, January 1996.

[40] N.S. Jayant and P. Noll, *Digital coding of waveforms,* Prentice-Hall, Inc. Englewood Cliffs, NJ, 1984.

[41] Jayaraman, S., Berger, T., "An error exponent for lossy source coding with side information at the decoder," *Proc. IEEE International Symposium on Information Theory*, pp. 263, 1995.

[42] JPEG technical specification: Revision, joint photographic exprets group, ISO/IEC, JTC1/SC2/WG8, CCITT SGVIII, August 1990.

[43] A. Kirac and P.P. Vaidyanathan, "Theory and design of optimum compaction filters," *IEEE Trans. on Signal Processing*, vol. 46, no. 4, pp. 903-919, April 1998.

[44] D. LeGall, "MPEG: a video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, no. 4, pp. 46-58, April 1991.

[45] Lim, J.S., and Oppenheim, A.V., "All-pole modeling of degraded speech," *IEEE Trans. on Acoustic, Speech and Signal Processing*, **ASSP-26**, pp. 197-210, 1978.

[46] Lim, J.S., and Oppenheim, A.V., "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, **67**, pp. 592-601, 1979.

[47] Lloyd, S.P., "Least squares quantization in PCM," unpublished Bell Labs Technical Note, (published in March 1982 *IEEE Trans. Inform. Theory*), 1957.

[48] Lloyd, S.P., "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*), **IT-28** 1982.

[49] H. Lutkepohl, *Handbook of matrices*, Chicester, England, Wiley, 1996.

[50] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. 5th Berkeley Symposium on Math. Stat. and Prob.*, vol. 1, pp. 281-296, 1967.

[51] S.G. Mallat, *A wavelet tour of signal processing*, Academic, San Diego, 1999.

[52] H.S. Malvar, "Extendend lapped transforms: fast algorithms and applications," *Proc. ICASSP-91, Int. Conf. on Acoustics Speech and Signal Processing,* vol. 3, pp. 1797-1800, 1991.

[53] H.S. Malvar, "Fast computation of the discrete cosine transform through fast Hartley transform," *Electrn. Letters*, vol. 22, no. 7, pp. 352-353, March 1986.

[54] H.S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, Signal Processing,* vol.38, pp. 969-978, June 1990.

[55] H.S. Malvar, *Signal processing with lapped transforms* Artech House, Boston, 1992.

[56] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*), **IT-6** 1960.

[57] Moulin, P. and O,Sullivan, J.A., "Information-theoretic analysis of information hiding", *Preprint*, 1999.

[58] Neuhoff, D.L., "Source coding strategies: simple quantizers vs. simple noiseless codes," *Proc. 1986 Conf. on Information Sciences and Systems*, vol. 1, pp. 267-271, March 1986.

[59] Oppenheim, A.V., and Schafer, R., "Digital signal processing," Prentice Hall, Englewood Cliffs, NJ, 1989.

[60] H.C. Papadapoulos, C.-E. W. Sundberg, "Simultaneous broadcasting of analog FM and digital audio signals by means of adaptive precancelling techniques," *IEEE Trans. on Communications*, vol 46, pp. 1233-1242, Sept. 1998.

[61] Pradhan S.S., Ramchandran K. "Distributed source coding using syndromes (DISCUS): design and construction," *Proceedings DCC'99 Data Compression Conference*, pp.158-67. Los Alamitos, CA, USA.

[62] K.R. Rao and P. Yip, *Discrete cosine transform: algorithms, advantages, applications,* Academic Press, New York, 1990.

[63] W.D. Ray and R.M. Driver, "Further decomposition of the Karhunen-Loeve series representation of a stationary random process," *IEEE Trans. Inform. Theory*, **IT-44**, pp. 245-250, 1970.

[64] Rimoldi, B., Urbanke, R. "Asynchronous Slepian-Wolf coding via source-splitting," *Proc. IEEE International Symposium on Information Theory*, pp. 271, 1997.

[65] A. Segall, "Bit allocation and encoding for vector sources," *IEEE Trans. Inform. Theory*, **IT-22**, pp.162-169, March 1976.

[66] Shamai, S., Verdu, S., Zamir, Z.,"Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, **IT-44**, pp. 564-579, 1998.

[67] J.M. Shapiro, "Embedded image coding using zerotrees and wavelet coefficients," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3445-3462, December 1993.

[68] D.K. Sharma, "Design of absolutely optimal quantizers of a wide class of distortion measures,", *IEEE Trans. Inform. Theory*, **IT-24**, pp. 693-702, 1978.

[69] Slepian, D., Wolf, J.K., Noiseless Coding of Correlated Information Sources," *IEEE Trans. Inform. Theory*, **IT-19**, pp. 471-480, 1973.

[70] Soman, A.K., Vaidyanathan, P.P., "Coding Gain in paraunitary analysis/synthesis systems," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 41, pp. 1824-1835, May 1993.

[71] H.V. sorensen, D.L. Jones, C.S. Burrus, M.T. Heideman, "On computing the discrete Hartley transform," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 1231-1238, Oct. 1985.

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

[73] K. Tanaka, and K. Matsio, "Embedding secret information into a dithered multi-level image," in Proc. of the 1990 Military Comm. Conf., pp. 216-220, 1990.

[74] A.V. Trushkin, "Sufficient conditions for uniqueness of a locally optimal quantizer class of convex error weighting functions," *IEEE Trans. Inform. Theory*, **IT-28**, 1982.

[75] M.K. Tsatsanis and G.B. Giannakis, "Principle component filter banks for optimal multiresolution analysis," *IEEE Trans. on Signal Processing,* vol. 43, no. 8, pp. 1766-1777, August 1998.

[76] Uyematsu, T., "An algebraic construction of codes for Slepian-Wolf source networks", *Proc. IEEE International Symposium on Information Theory*, New York, NY, pp.138, 1998.

[77] V.A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, **IT-39**, 1993.

[78] P.P. Vaidyanathan,"Filter banks with maximum coding gain and energy compaction," *Proc. of ASILOMAR-29*, pp. 36-40, Sept. 1996.

[79] P.P. Vaidyanathan,"Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proc. of IEEE*, vol. 78, no. 1, pp. 56-93, Jan. 1990.

[80] P.P. Vaidyanathan, *Multirate systems and filter banks,* Prentice Hall, Boston, 1992.

[81] G.K. Wallace, "The JPEG still picture image compression standard", *Communications of the ACM*, vol. 34, no. 4, pp. 30-44, March 1992.

[82] B. Xuan and R.H. Bamberger, "FIR Principle component filter banks," *IEEE Trans. on Signal Procesing*, vol. 46, no. 4, pp. 930-940, April 1998.

[83] M. Vetterli, "Running FIR and IIR filtering using multirate filter banks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, May 1988, pp. 730-738.

[84] M. Vetterli and J. Kovacevic, *Wavelets and subband coding*, Prentice Hall PTR, Englewood Cliffs, NJ, 1995.

[85] A.J. Viterbi, "Phase-locked loop dynamics in the presence of noise by Fokker-Planck techniques", *Proc. IEEE*, vol. 51, pp. 1737-1753, December 1963.

[86] A.S Willsky and G.W. Wornell, *Stochastic processes, estimation and detection, course notes.* Dept. of Electrical Engineering and Computer Science. Cambridge, MA. September 1994.

[87] A.D. Wyner, "Recent results in Shannon theory," *IEEE Trans. Inform. Theory*, **IT-20**, January 1974.

[88] A.D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, **IT-21**, pp. 294-300, May 1975.

[89] A.D. Wyner, "The rate-distortion function for source coding with side information at the decoder-II: general sources," *Information and Control*, **38**, pp. 60-80, 1978.

[90] A.D. Wyner, and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, **IT-22**, pp. 1-10, 1976.

[91] X. Yang and K. Ramchandran, "Optimal multiple description subband coding," *Proc. 1998 Int. Conf. on Image Proc., ICIP98*, pp. 654-658, 1998.

[92] R. Zamir and T. Berger, "Multiterminal source coding with high resolution," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 106-117, 1999.

[93] R. Zamir, "The rate loss in the Wyner-Ziv problem," *IEEE Trans. Inform. Theory*, **IT-42**, 1996.

[94] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inform. Theory*, pp. 1152-1159, July 1996.

[95] R. Zamir and S. Shamai, "Nested linear/lattice codes for Wyner-Ziv encoding," *1998 Information Theory Workshop, Kilarney, Ireland*, pp 92-93, **IT-31**, June 1998.

[96] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, pp.344-47, May 1985.