UNIVERSITY of CALIFORNIA

Santa Barbara

Complexity and Delay Constrained Compression and Transmission of Information

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Emrah Akyol

Committee in charge:

Professor Kenneth Rose, Chair
Professor Allen Gersho
Professor Jerry Gibson
Professor Upamanyu Madhow
Professor Tor Ramstad

June 2011

The dissertation of Emrah Akyol is approved.

Professor Allen Gersho

Professor Jerry Gibson

Professor Upamanyu Madhow

Professor Tor Ramstad

Professor Kenneth Rose, Committee Chair

March 2011

Complexity and Delay Constrained Compression and Transmission of Information

To my parents

Acknowledgements

I would like to thank Professor Kenneth Rose for his guidance and support throughout this work. It was a great pleasure to be his student. I was largely inspired by his deep insight, extensive knowledge, liberal advising style, and research philosophy. He taught me that conducting research is not only a daily job but a way of having fun as well.

I also had the opportunity to interact with Professor Allen Gersho who influenced this work in many ways, especially at the beginning stages. I am deeply indebted to Professor Tor Ramstad for our valuable discussions on zero-delay joint source channel coding and for his detailed and much appreciated comments on the first version of this thesis.

It was a great honor for me to have the highest caliber scientists on my thesis committee: Professors Allen Gersho, Jerry Gibson, Upamanyu Madhow, and Tor Ramstad. I would like to thank them for accepting to read and review the first version of this thesis.

Another important acknowledgment goes to Professors Murat Tekalp and Reha Civanlar for their constant support throughout the ups and downs of my academic life, which made this thesis possible.

I thank my labmates Alphan, Ankur, Christian, Emmanuel, Emre, Jingning, Kumar, Min-Chi, Pradeep, Sharadh, Tejaswi and Vinay for our interesting discussions on the white board and SCL seminars; specifically, Chapter 3 resulted from my collaboration with Kumar. Thanks also to my friends in Santa Barbara for making my stay here enjoyable and Susie for being in my life.

The projects presented in this theses are supported in part by the National Science Foundation (NSF) under the grants CCF-0728986 and CCF-1016861.

Finally, this thesis is dedicated to my parents for their unconditional love and support.

Curriculum Vitæ

Emrah Akyol

Education

2005	MS in Electrical and Computer Engineering, Koc University, Turkey.
2003	BS in Electrical Engineering, Bilkent University, Turkey.

Experience

2008–2011	Graduate Student Researcher, University of California, Santa Barbara.
2007–2008	Teaching Assistant, University of California, Santa Barbara.
2007	Intern, NTT DoCoMo Labs, Palo Alto
2006	Intern, HP Labs, Palo Alto
2005-2006	Graduate Student Researcher, University of California, Los Angeles
2003-2005	Teaching Assistant, Koc University, Istanbul

Publications

- E. Akyol, K. Rose, “On Optimal Randomized Quantization”, *in preparation*

- E. Akyol, K. Rose, “Optimal Transform Coding ”, *in preparation*

- E. Akyol, K. Viswanatha and K. Rose, “On Conditions for Linearity of Optimal Estimation”, *submitted to IEEE Transactions on Information Theory*

- E.Akyol, K. Rose, T. Ramstad“Optimal Analog Mappings for Joint Source Channel Coding”, *submitted to IEEE Transactions on Communications*

- E. Akyol, K. Viswanataha, and K. Rose, On Multidimensional Optimal Estimators: Linearity Conditions, Proc. IEEE Statistical Signal Processing Workshop, to appear

- E. Akyol, K. Viswanataha, and K. Rose, On Conditions for Linearity of Optimal Estimation, Proc. IEEE Information Theory Workshop, Aug 2010,

- K. Viswanataha, E. Akyol, and K. Rose, On Optimum Communication Cost for Joint Compression and Dispersive Information Routing, Proc. IEEE Information Theory Workshop, Aug 2010

- K. Viswanataha, E. Akyol, S. Ramaswamy, and K. Rose, Distributed source coding and dispersive information routing: An integrated approach with networking and database applications, European Signal Processing Conference 2010.

- K. Viswanataha, E. Akyol, and K. Rose, Towards Optimum Cost in Multi-hop Networks with Arbitrary Network Demands, IEEE Int. Symposium on Information Theory Feb 2010.

- E. Akyol, K. Rose, and T. Ramstad, Optimized Analog Mappings for Distributed Source Channel Coding, Proc. IEEE Data Compression Conference, 2010

- E. Akyol, K. Rose, and T. Ramstad, Optimal Mappings for Joint Source Channel Coding, Proc. IEEE Information Theory Workshop, 2010

- E. Akyol, K. Rose, On Transform Coding with Dithered Quantizers, Proc. IEEE Data Compression Conference, 2009

- E. Akyol and K. Rose, Nonuniform Dithered Quantization, Proc. IEEE Data Compression Conference 2009

- E. Akyol, and M. vander Schaar, "Compression-Aware Energy Optimization for Video Decoding Systems with Passive Power", *IEEE Transactions on Circuits, Systems and Video Technology*, vol.18, no. 9 pp.1300-1306, 2008

- E. Akyol, and M. vander Schaar, "Complexity Model Based Proactive Dynamic Voltage Scaling for Video Decoding Systems", *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1475-1492, 2007

- E. Akyol, M. Tekalp and R. Civanlar, 'A Flexible Multiple Description Coding Framework for Adaptive Peer-to-peer Video Streaming", *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 2, pp. 231-245, 2007

- E. Akyol, M. Tekalp and R. Civanlar, 'Content-aware Scalability Type Selection for Rate Adaptation of Scalable Video", *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 10236, 11 pages, 2007

- E. Akyol, O.G. Guleryuz, M.R.Civanlar, Royalty Cost Based Optimization for Video Compression, Proc. IEEE Int. Conf. on Image Processing, 2007

- E. Akyol, D. Mukherjee, and Y. Liu, Complexity Control for Real Time Video Coding, Proc. IEEE Int. Conf. on Image Processing, 2007

- E. Akyol and M. van der Schaar , Buffer Constrained Proactive Dynamic Voltage Scaling for Video Decoding Systems, IProc. IEEE Int. Conf. on Image Processing 2007

- E. Akyol, M. Tekalp, and R.Civanlar, Adaptive Peer to Peer Video Streaming with Flexible Multiple Description Coding, Proc. IEEE Int. Conf. on Image Processing 2006

- E. Akyol, M.Tekalp, and R.Civanlar, Optimum Bit Allocation in Scalable Multiple Description Video Coding" Proc. European Signal Processing Conference 2005

- E. Akyol, M. Tekalp, and R. Civanlar, Scalable Multiple Description Video Coding with Flexible Number of Descriptions, Proc. IEEE Int. Conf. on Image Processing, 2005

- E. Akyol, M. Tekalp, and R. Civanlar, Optimum Scaling Operator Selection in Scalable Video Coding, Proc. Picture Coding Symposium, 2004

- E. Akyol, M. Tekalp, and R. Civanlar, "Motion Compensated Temporal Filtering Within the H.264/AVC Standard", Proc. IEEE Int. Conf. on Image Processing, 2004

Abstract

Complexity and Delay Constrained Compression and Transmission of Information

by

Emrah Akyol

This dissertation is concerned with optimal strategies for delay and complexity constrained communications. The first part of the thesis studies optimal joint source-channel coding for zero-delay communications. This problem, originally posed by Shannon in his seminal paper, received recent attention due the increasing need for low delay and low complexity communications. The necessary conditions for optimality of encoding and decoding functions (mappings) are derived and a corresponding numerical algorithm is proposed which is shown to discover locally optimal mappings that outperform all prior results. The approach is then extended to provide optimality conditions and design algorithms for distributed source channel coding. The second part of the thesis is concerned with the conditions for linearity of optimal decoding mappings, and of optimal estimators in general, in terms of source and noise densities and distortion measure. Specifically, the necessary and sufficient conditions for linearity of the optimal estimator along with existence and uniqueness of source and noise densities that satisfy such conditions, are derived. While there are several source-noise pairs that satisfy these conditions, a property unique to Gaussians is also presented. The remainder of the thesis is focused on low delay source coding methods including contributions to dithered quantization and transform coding. Dithered (randomized) quantization which has traditionally been considered in its natural setting of uniform quantization, is extended to encompass nonuniform quantizers by dithering in the companded domain. The compressor and expander mappings are optimized using the numerical tools derived for the source-channel mapping problem. Asymptotic properties of such a randomized quantizer are also analyzed. Finally, a long standing theoretical problem of transform coding is solved. The necessary and sufficient condition for optimality of a transform, in conjunction with variable rate quantization at high resolution is derived. This condition not only determines when the Karhunen-Loeve transform (KLT) is optimal, but also leads to an algorithm that obtains the optimal (non-KLT) transform. The optimal transform is also derived for the setting of transform coding in conjunction with dithered quantization, resulting in a universally optimal fixed rate source coding scheme.

Contents

List of Figures

Chapter 1

Introduction

While Shannon's point-to-point communication theorems [66] have profoundly revolutionized the modern information age, advancing the theory to less simple settings has been more problematic. The model of communication between point-to-point links, "dumb" forwarding nodes, encoder and decoders with unbounded complexity and delay, as assumed in the early seminal contributions, does not capture key aspects of current and emerging networks. Wireless networks consist of nodes that can act as the source, the destination and/or the relay and can communicate in several different ways. Many new network applications are highly interactive, requiring very low delay, and are distributed in nature (peer-to-peer, mobile agents) and hence, require the development of new communication and computing schemes with very limited energy and delay. This thesis is focused on "theoretical" limits and "optimal" methods of communication and compression systems that have such"practical" constraints.

1.1 Optimal Mappings for Joint Source Channel Coding

Shannon's coding theorems assume achievability via unboundedly complex encoders and decoders with potentially infinite sample delay. One pertinent and surprising result is that transmission of Gaussian samples over an additive white Gaussian noise channel with matched bandwidth is optimal in the sense that it yields the minimum achievable mean square error between source and reconstruction [24]. This result demonstrates the potential of joint source channel coding: such a simple scheme at "zero delay" provides the performance of the asymptotically optimal separate source-channel coding system, without recourse to complex compression and channel coding schemes or asymptotically long delays.

To address the practical problem of transmitting a discrete time continuous alphabet source over a discrete time additive noise channel, there are two main approaches: “analog communication” via direct amplitude modulation, and “digital communication” which typically consists of quantization, error control coding and digital modulation. The main advantage and hence proliferation of digital over analog communication is due to advanced quantization and error control techniques, and the prevalence of digital processors. However, there are two notable shortcomings: First problem is that error control coding (and to some extent also source coding) usually incurs substantial delay to achieve good performance. The other problem involves adaptivity of digital systems to varying channel conditions due to underlying quantization. The performance saturates as channel signal-to-noise ratio (CSNR) increases well above the threshold for which the system is designed. Another problem is the lack of “graceful degradation” below a CSNR threshold. Further, this threshold effect becomes more pronounced as the system approaches optimal performance at the condition it was designed for. Analog systems offer the potential to avoid these problems, with relatively simple encoders/decoders. However, there are no known explicit methods to obtain such analog mappings for a general source and channel, nor is the best mapping known for other than the most trivial cases, e.g., the scalar Gaussian source-channel pair.

In this part of the thesis, closed form necessary conditions for optimality of the encoder and decoder mappings are derived. The optimal mappings are then obtained using an iterative algorithm that updates encoder and decoder mappings according to optimality conditions at each iteration. Specifically, a gradient descent algorithm to find the locally optimal mappings [2] for the point to point setting, which involves a source with known distribution which must be transmitted over an additive noisy channel with known distribution to the designer. The algorithm, as a natural consequence of being a gradient descent algorithm, guarantees only local optimality. Similar local optimality problems also appear in well studied vector quantizer design problem. One solution to that problem, namely ”noisy channel relaxation” is employed to mitigate the local optimality effect.

The approach is then extended to network scenarios. Two different setups are of interest: Decoder side information and distributed coding. For each of these scenarios, necessary conditions for optimality is derived and based on those conditions, the locally optimal encoding and decoding mappings are (numerically) found.

1.2 Linearity of Optimal Estimation

Communication systems use structured (linear, lattice, trellis) codes to reduce complexity. Fortunately, such codes also have the potential to approach the limits promised by information theory, i.e., such reduced complexity is free. In delay limited codes, from our analysis, we observe that this is not the case. Zero-delay analog codes are highly nonlinear, do not have an obvious useable structure, and vary with CSNR. The conditions for linearity of such encoding-decoding mapping is closely related to the fundamental problem of linearity of optimal estimation (regression) in the mathematical literature. Surprisingly, even the most basic problems in the conditions for estimation setup are still open. The basic problem in estimation theory, namely, source estimation from a signal received through a channel with additive noise, given the statistics of both the source and the channel, is considered. The optimal estimator that minimizes the mean square estimation error is usually a nonlinear function of the observation [43]. A frequently exploited result in estimation theory concerns the special case of Gaussian source and Gaussian channel noise, a case in which the optimal estimator is guaranteed to be linear. An open follow-up question considers the existence of other cases exhibiting such a “coincidence”, and more generally the characterization of conditions for linearity of optimal estimators for general distortion measures. The estimation problem in general has been studied intensively in the literature. It is known that, for stable distributions (which of course include the Gaussian case), the optimal estimator is linear [70, 21, 62, 47] for any signal to noise ratios (SNR).

In this part of the thesis, we focus on this question: When is optimal estimation linear? It is well known that, when a Gaussian source is contaminated with Gaussian noise, a linear estimator minimizes the mean square estimation error. We analyze, more generally, the conditions for linearity of optimal estimators. Given a noise (or source) distribution, and a specified signal to noise ratio (SNR), we derive conditions for existence and uniqueness of a source (or noise) distribution for which the L_p optimal estimator is linear. We then show that, if the noise and source variances are equal, then the matching source must be distributed identically to the noise. Moreover, we prove that the Gaussian source-channel pair is unique in the sense that it is the only source-channel pair for which the mean square error (MSE) optimal estimator is linear at more than one SNR values. Further, we show the asymptotic linearity of MSE optimal estimators for low SNR if the channel is Gaussian regardless of the source and vice versa, for high SNR if the source is Gaussian regardless of the channel. The extension to the vector case is also considered where besides the conditions inherited from the scalar case, additional constraints must be satisfied to ensure linearity of optimal estimator.

1.3 Optimal Randomized Quantization

Most sources of practical interest are in fact sources with "memory", i.e., they exhibit correlations. A computationally attractive approach to source coding is predictive coding. Optimizing a predictive coding system, especially in network scenarios is complicated due to nonlinear structure of quantization. For this purpose, either high resolution quantization are employed although these systems are designed for very low rate. Alternatively, randomized (dithered) quantizers, that renders the quantization error independent of the source, can be employed. However, dithered quantizers suffer from performance loss due to uniform quantization.

Traditionally, dithered quantizers have been considered within their natural setting of uniform quantization framework. In this paper we extend conventional dithered quantization to nonuniform quantization, where dithering is performed in the companded domain. Closed form necessary conditions for optimality of these compressor and expander mappings are derived for both fixed and variable rate randomized quantization. The optimal mappings are numerically obtained by updating the mappings based on necessary conditions. Moreover, deterministic and randomized quantizers that are constrained to provide quantization error uncorrelated with the source are studied. Numerical results are presented for the Gaussian source and it is shown that the proposed quantizer outperforms the conventional dithered quantizer as well as the deterministic quantizer with quantization error uncorrelated with the source. In the second part of the paper, we investigate whether random coding is necessary to achieve (asymptotic) optimality while imposing uncorrelated quantization error. We show that for a Gaussian source, the optimal vector quantizer with asymptotically high dimension that renders quantization error uncorrelated with the source must be a randomized one. In this situation, random encoding in rate-distortion theory is not merely a tool to characterize the performance bounds, but is, in fact, a required property of the quantizers achieving such bounds.

1.4 Optimal Transform Coding

Another practical form of reducing redundancy is transform coding which is widely used in audio, image and video compression. In the basic transform coding setting, an input vector is linearly transformed into a vector in the transform domain whose components are scalar-quantized. The decoder reconstructs the quantized coefficients and performs linear (inverse) transformation to obtain an estimate of the source vector. The design goal is to find the optimal transform pair and bit allocation to scalar quantizers, which minimize distortion. In general, transform coding underperforms optimal vector quantization due to space filling loss in scalar quantizers, even if the transform generates independent coefficients. Nevertheless, due to its low complexity, transform coding is commonly employed in practical multimedia compression systems [27, 20]. Although transform coding has been extensively studied, optimal transform is known only for a small set of source distributions. Optimal transform coding has remained an open problem for decades. In their seminal paper, Huang and Schulthesis have shown [38] that if the vector source is Gaussian and the bit budget is asymptotically large, then the Karhunen Loeve transform (KLT) is the optimal transform for fixed-rate coding. In a more recent paper Goyal, Zhuang and Vetterli improve that result by showing that KLT is optimal for Gaussian sources without making any high resolution assumptions [26]. The optimality of KLT in transform coding of Gaussian sources is often explained intuitively by the assertion that scalar quantization is better suited to the coding of independent random variables than to the coding of dependent random variables. Thus, the optimality of KLT for transform coding of Gaussian sources is understood to be a consequence of the fact that it yields independent transform coefficients. The application of KLT in transform coding of non-Gaussian sources is then justified using the intuitive argument that KLT's coefficient decorrelation represents, for general sources, a rough approximation to the desired coefficient independence [15].

In this part of the thesis, we focus on the fundamental theoretical problem of optimal transform coding. The main result is a necessary and sufficient condition for optimality of a transform in conjunction with variable rate coding at high resolution. Specifically, we show that the optimal transform is the one that minimizes the divergence between the joint distribution of the coefficients and the product of their marginals. Note furthermore that this result not only resolves the question of when KLT is optimal (at high resolution), but it also determines the optimal transform when it is not KLT. This result connects the transform coding problem to the well studied problem of "source separation". Inspired from the vast amount of source separation algorithms, we propose an algorithm to derive the optimal transform. Finally, we derive the optimal transform in conjunction with dithered quantization.While the optimal deterministic quantizer's error is uncorrelated with the reconstructed value, the dithered quantizer yields quantization errors that are correlated with the reconstruction but are white and independent of the source. These properties offer potential benefits, but also have implications on the optimization of the rest of the coder. We derive the optimal transform for consequent dithered quantization. For fixed rate coding, we show that the transform derived for dithered quantization is universally optimal (for all sources), unlike the conventional quantization case where optimality of the Karhunen-Loeve transform is guaranteed for Gaussian sources

Chapter 2

Optimal Analog Mappings

2.1 Introduction

One of the fascinating results in information theory is that uncoded transmission of Gaussian samples, over a channel with additive white Gaussian noise (AWGN), is optimal in the sense that it yields the minimum achievable mean square error between source and reconstruction [24]. This result demonstrates the potential of joint source-channel coding: Such a simple scheme, at no delay, provides the performance of the asymptotically optimal separate source and channel coding system, without recourse to complex compression and channel coding schemes that require asymptotically long delays. However, it is understood that, in general, the best source channel coding system at fixed finite delay may not achieve Shannon's asymptotic coding bound (see eg. [13]).

Nevertheless, the problem of obtaining the optimal scheme for a given finite delay is an important open problem with considerable practical implications. To the practical problem of transmitting a discrete time continuous alphabet source over a discrete time additive noise channel, there are two main approaches: "analog communication" via direct amplitude modulation, and "digital communication" which typically consists of quantization, error control coding and digital modulation. The main advantage (and hence proliferation) of digital over analog communication is due to advanced quantization and error control techniques, as well as the prevalence of digital processors. However, there are two notable shortcomings: first, error control coding (and to some extent also source coding) usually incurs substantial delay to achieve good performance. The other problem involves limited adaptivity of digital systems to varying channel conditions, due to underlying quantization or error protection assumptions. The performance saturates due to quantization as the channel signal to noise ratio (CSNR) increases beyond the regime for which the system was designed. Also, it is difficult to obtain "graceful degradation" of digital systems with decreasing CSNR, when it falls below a certain threshold due to the error correction code in use. Further, such threshold effects become more pronounced as the system performance approaches the theoretical optimum. Analog systems offer the potential to avoid these problems. As an important example, in applications where significant delay is acceptable, a hybrid approach (i.e., vector quantization + analog mapping) was proposed and analyzed [55, 71] to circumvent the impact of CSNR mismatch where, for simplicity, linear mappings were used and hence no optimality claims made. Perhaps more importantly, in many applications (e.g., multimedia streaming) delay is a paramount consideration. Analog coding schemes are low complexity alternatives to digital methods, providing a "zero-delay" transmission which is suitable for such applications.

There are no known explicit methods to obtain such analog mappings for a general source and channel, nor is the best mapping known for other than the trivial case of the scalar Gaussian source-channel pair. Among the few practical analog coding schemes that have appeared in the literature are those based on the use of space-filling curves for bandwidth compression, originally proposed more than 50 years ago by Shannon [67] and Kotelnikov [46]. These were recently extended in the work of Fuldseth and Ramstad [17], Chung [11], Ramstad [59], Wernersson et.al. [78], and Hekland et.al. [33], where spiral-like curves are explored for transmission of Gaussian sources over AWGN channels for bandwidth compression ($m > k$) and expansion ($m < k$). It is also noteworthy that a similar problem was solved in [49] albeit under the stringent constraint that both encoder and decoder be linear. A similar problem, formulated in the context of digital systems, was also studied by Fine [16]. Certain extensions of Fine's work can be found in [22]. Properties of the optimal mappings have been considered, over the years, in [67, 83, 74]. Shannon's arguments[67] are based on the topological impossibility to map one region to another region in a "one-to-one", continuous manner, unless both regions have the same dimensionality. On this basis, he explained the threshold effect common to various communication systems. Moreover, Ziv [83] showed that for a Gaussian source transmitted over AWGN channel, no single practical modulation scheme can achieve optimal performance at all noise levels, if the channel rate is greater than the source rate (i.e., bandwidth expansion). It has been conjectured that this property holds whenever the source rate differs from the channel rate [74]. Our own preliminary results appeared in [2, 3]. It is noteworthy that analog mappings were found to be useful in related applications of low delay relaying, see eg.,[44] and references therein.

In this chapter, we investigate the problem of obtaining vector transformations that optimally map between the m-dimensional source space and the k-dimensional channel space, under a given transmission power constraint, and where optimality is in the sense of minimum mean square reconstruction error. We provide necessary conditions for optimality of the mappings used at the encoder and the decoder. Note that virtually any source-channel communication system (including digital communication) is a special case of such mappings, as shown in Figure 2.2. A digital system, including quantization, error correction and modulation, boils down to a specific mapping from the source space $\mathbb{R}^m$ to the channel space $\mathbb{R}^k$. Hence the derived optimality conditions are generally valid and subsume digital communications as an extreme special case. Based on the optimality conditions we derive, we propose an iterative algorithm to optimize the mappings for any given m, k (i.e., for both bandwidth expansion or compression) and for any given source-channel statistics. To our knowledge, this problem has not been fully solved, except when both source and channel are scalar and Gaussian. We provide examples of such $m : k$ mappings for source-channel pairs and construct the corresponding source-channel coding systems that outperform the mappings obtained in [17, 11, 59, 78, 33, 37].

The second part of the chapter extends the approach to the scenario of source-channel coding with decoder side information (i.e., the decoder has access to some other correlated source). This setting, in the context of pure source coding, goes back to the pioneering work of Slepian and Wolf [72] and Wyner and Ziv [79]. Finally, we consider distributed source-channel coding where multiple correlated sources are encoded separately and transmitted over different channels to a common decoder. An important practical motivation for this setup is sensor networks where sensor measurements are correlated but are not encoded jointly as the sensors are distributed in space. This problem has also been studied extensively, especially for Gaussian sources [76, 75, 77].

The derivation of the optimality conditions for distributed source channel coding is a direct extension of the point-to-point case, but the distributed nature of this setting results in highly nontrivial mappings. Straightforward numerical optimization of such mappings is susceptible to get trapped in one of the numerous local minima that riddle the cost functional. Note, in particular, that in the case of Gaussian sources and channels, linear encoders and decoder (automatically) satisfy the necessary conditions of optimality while, as we will see, careful optimization obtains considerably better mappings that are far from linear.

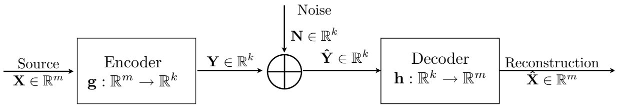

Noise
$\mathbf{N} \in \mathbb{R}^k$
Source
$\mathbf{X} \in \mathbb{R}^m$
Encoder
$\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^k$
$\mathbf{Y} \in \mathbb{R}^k$
$\hat{\mathbf{Y}} \in \mathbb{R}^k$
Decoder
$\mathbf{h} : \mathbb{R}^k \to \mathbb{R}^m$
Reconstruction
$\hat{\mathbf{X}} \in \mathbb{R}^m$

Figure 2.1. A general block-based point-to-point communication system

2.2 Problem Formulation

2.2.1 Preliminaries and Problem Definition

Point to point

We consider the general communication system whose block diagram is shown in Figure 2.2. An m-dimensional vector source $\mathbf{X} \in \mathbb{R}^m$ is mapped into a k-dimensional vector $\mathbf{Y} \in \mathbb{R}^k$ by function $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^k$, and transmitted over an additive noise channel. The received vector $\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{N}$ is mapped by the decoder to the estimate $\hat{\mathbf{X}}$ via function $\mathbf{h} : \mathbb{R}^k \to \mathbb{R}^m$. The noise $\mathbf{N}$ is assumed to be independent of the source $\mathbf{X}$. The m-fold source density is denoted $f_X(\mathbf{x})$ and the k-fold noise density is $f_N(\mathbf{n})$. Let $\mathbf{G}$ and $\mathbf{H}$ denote the sets of all square integrable functions $\{\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^k\}$ and $\{\mathbf{h} : \mathbb{R}^k \to \mathbb{R}^m\}$, respectively.

The objective is to minimize, over the choice of encoder $\mathbf{g} \in \mathbf{G}$ and decoder $\mathbf{h} \in \mathbf{H}$, the distortion

$$D[\mathbf{g}, \mathbf{h}] = \mathbb{E}\{||\mathbf{X} - \hat{\mathbf{X}}||^2\} \tag{2.1}$$

subject to the average power constraint,

$$P[\mathbf{g}] = \mathbb{E}\{||\mathbf{g}(\mathbf{X})||^2\} \leq P_T \tag{2.2}$$

where P_T is the specified transmission power level. Bandwidth compression-expansion is determined by the setting of the source and channel dimensions, k/m. The power constraint limits the choice of encoder function $\mathbf{g}$. Note that, without a power constraint on $\mathbf{g}$, the CSNR is unbounded and the channel can be made effectively noise free.

Decoder side information

As shown in Figure 2.2, there are two correlated vector sources $\mathbf{X_1} \in \mathbb{R}^{m_1}$ and $\mathbf{X_2} \in \mathbb{R}^{m_2}$ with a joint density $f_{X_1,X_2}(\mathbf{x_1}, \mathbf{x_2})$. $\mathbf{X_2}$ is available only to the decoder, while $\mathbf{X_1}$ is mapped to $\mathbf{Y} \in \mathbb{R}^k$ by the encoding function $\mathbf{g} : \mathbb{R}^{m_1} \to \mathbb{R}^k$ and transmitted over the channel whose additive noise $\mathbf{N} \in \mathbb{R}_k$, with density $f_N(\mathbf{n})$, is independent of $\mathbf{X_1}, \mathbf{X_2}$. The received channel

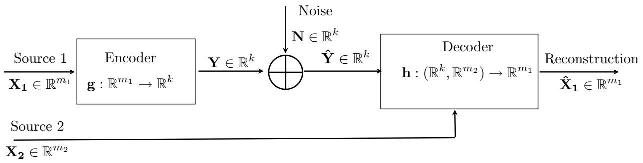

Noise
$\mathbf{N} \in \mathbb{R}^k$
Source 1
$\mathbf{X_1} \in \mathbb{R}^{m_1}$
Encoder
$\mathbf{g} : \mathbb{R}^{m_1} \rightarrow \mathbb{R}^k$
$\mathbf{Y} \in \mathbb{R}^k$
$\hat{\mathbf{Y}} \in \mathbb{R}^k$
Decoder
$\mathbf{h} : (\mathbb{R}^k, \mathbb{R}^{m_2}) \rightarrow \mathbb{R}^{m_1}$
Reconstruction
$\hat{\mathbf{X}}_\mathbf{1} \in \mathbb{R}^{m_1}$
Source 2
$\mathbf{X_2} \in \mathbb{R}^{m_2}$

Figure 2.2. Source-channel coding with decoder side information

output $\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{N}$ is mapped to the estimate $\hat{\mathbf{X}}_\mathbf{1}$ by the decoding function $\mathbf{h} : \mathbb{R}^k \times \mathbb{R}^{m_2} \rightarrow \mathbb{R}^m_1$. The problem is to find optimal mapping functions $\mathbf{g}, \mathbf{h}$ that minimize the distortion

$$D[\mathbf{g}, \mathbf{h}] = \mathbb{E}\{||\mathbf{X_1} - \hat{\mathbf{X}}_\mathbf{1}||^\mathbf{2}\} \tag{2.3}$$

subject to average power constraint

$$P[\mathbf{g}] = \mathbb{E}\{||\mathbf{g}(\mathbf{X})||^2\} \leq P_T \tag{2.4}$$

Distributed coding

As shown in Figure 2.3, two correlated vector sources $\mathbf{X_1} \in \mathbb{R}^{m_1}$ and $\mathbf{X_2} \in \mathbb{R}^{m_2}$ with a joint density $f_{X_1,X_2}(\mathbf{x_1}, \mathbf{x_2})$, are independently transmitted to the receiver. Noise variables $\mathbf{N}_1 \in \mathbb{R}^{m_1}, \mathbf{N}_2 \in \mathbb{R}^{m_2}$ are assumed to be independent of each other and of the sources $\mathbf{X}_1, \mathbf{X}_2$, and have densities $f_{N_1}(\mathbf{n_1}), f_{N_2}(\mathbf{n_2})$, respectively. Both $\mathbf{X}_1$ and $\mathbf{X}_2$ are mapped by encoding functions $\mathbf{g_1} : \mathbb{R}^{m_1} \rightarrow \mathbb{R}^{k_1}$ and $\mathbf{g_2} : \mathbb{R}^{m_2} \rightarrow \mathbb{R}^{k_2}$ and transmitted over the noisy channels. At the decoder, $\hat{\mathbf{X}}_\mathbf{1}$ and $\hat{\mathbf{X}}_\mathbf{2}$ are generated by $\mathbf{h_1} : \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \rightarrow \mathbb{R}^{m_1}$ and $\mathbf{h_2} : \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \rightarrow \mathbb{R}^{m_2}$. As before, the problem is to find optimal mapping functions $\mathbf{g_1}, \mathbf{g_2}, \mathbf{h_1}, \mathbf{h_2}$ that minimize the distortion

$$D[\mathbf{g_1}, \mathbf{g_2}, \mathbf{h_1}, \mathbf{h_2}] = \mathbb{E}\{||\mathbf{X_1} - \hat{\mathbf{X}}_\mathbf{1}||^2 + ||\mathbf{X_2} - \hat{\mathbf{X}}_\mathbf{2}||^2\} \tag{2.5}$$

subject to the average power constraint per encoder,

$$P[\mathbf{g_1}] = \mathbb{E}\{||\mathbf{g_1}(\mathbf{X})||^2\} \leq P_1 \ , \ P[\mathbf{g_2}] = \mathbb{E}\{||\mathbf{g_2}(\mathbf{X})||^2\} \leq P_2 \tag{2.6}$$

2.2.2 Asymptotic Bounds for Gaussian Source and Channel

Although the problem we consider is delay limited, it is insightful to consider asymptotic bounds obtained at infinite delay. From Shannon's source and channel coding theorems, it is

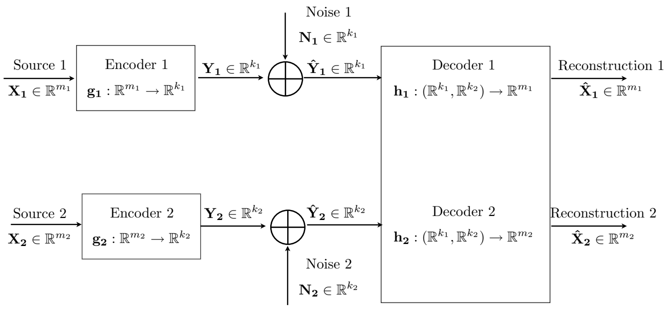

Noise 1
$\mathbf{N_1} \in \mathbb{R}^{k_1}$
Source 1
$\mathbf{X_1} \in \mathbb{R}^{m_1}$
Encoder 1
$\mathbf{g_1} : \mathbb{R}^{m_1} \to \mathbb{R}^{k_1}$
$\mathbf{Y_1} \in \mathbb{R}^{k_1}$
$\hat{\mathbf{Y}}_\mathbf{1} \in \mathbb{R}^{k_1}$
Decoder 1
$\mathbf{h_1} : (\mathbb{R}^{k_1}, \mathbb{R}^{k_2}) \to \mathbb{R}^{m_1}$
Reconstruction 1
$\hat{\mathbf{X}}_\mathbf{1} \in \mathbb{R}^{m_1}$
Source 2
$\mathbf{X_2} \in \mathbb{R}^{m_2}$
Encoder 2
$\mathbf{g_2} : \mathbb{R}^{m_2} \to \mathbb{R}^{k_2}$
$\mathbf{Y_2} \in \mathbb{R}^{k_2}$
$\hat{\mathbf{Y}}_\mathbf{2} \in \mathbb{R}^{k_2}$
Decoder 2
$\mathbf{h_2} : (\mathbb{R}^{k_1}, \mathbb{R}^{k_2}) \to \mathbb{R}^{m_2}$
Reconstruction 2
$\hat{\mathbf{X}}_\mathbf{2} \in \mathbb{R}^{m_2}$
Noise 2
$\mathbf{N_2} \in \mathbb{R}^{k_2}$

Figure 2.3. Distributed source-channel coding

known that, asymptotically, the source can be compressed to $R(D)$ bits (per source sample) at distortion level D, and that C bits can be transmitted over the channel (per channel use) with arbitrarily low probability of error, where $R(D)$ is the source rate-distortion function, and C is the channel capacity, (see eg.[13]). The asymptotically optimal coding scheme is the tandem combination of the optimal source and channel coding schemes, hence $mR(D) \leq kC$ must hold. By setting

$$R(D) = \frac{k}{m}C \tag{2.7}$$

one obtains a lower bound on the distortion of any source-channel coding scheme. Next, we specialize to Gaussian sources and channels, which we will mostly use in the numerical results section, while emphasizing that the proposed method is generally applicable to any source and noise densities. The rate-distortion function for the memoryless Gaussian source of variance σ_x^2, under the squared-error distortion measure is given by

$$R(D) = \max(0, \frac{1}{2}\log\frac{\sigma_x^2}{D}) \tag{2.8}$$

for any distortion value $D \geq 0$. The capacity of the AWGN channel is given by

$$C = \frac{1}{2}\log(1 + \frac{P_T}{\sigma_n^2}) \tag{2.9}$$

where P_T is the transmission power constraint and σ_n^2 is the noise variance. Plugging (4.40) and (2.9) in (4.39) we obtain the optimal performance theoretically attainable (OPTA):

$$D_{OPTA} = \frac{\sigma_x^2}{(1 + \frac{P_T}{\sigma_n^2})^{\frac{k}{m}}} \tag{2.10}$$

Note that OPTA is derived without any delay constraints and may not be achievable by a delay-constrained coding scheme. No achievable theoretical bound is known for joint source channel coding with limited delay, although there are recent results that tighten the outer bound, see eg. [41].

For source coding with decoder side information, it has been established for Gaussians and MSE distortion that there is no rate loss due to the fact that the side information is unavailable to the encoder [79]. Similar to the derivation above, OPTA can be obtained for source-channel coding with decoder side information, by equating the conditional rate distortion function of the source (given the side information) to the channel capacity.

The two encoder distributed source coding problem, with Gaussian sources and MSE distortion has been analyzed in [75]. OPTA can be derived by setting the rate distortion function of [75] to the channel capacity.

2.3 Optimality Conditions

We proceed to develop the necessary conditions for optimality of the encoder and decoder subject to the average power constraint (2.2) in the point-to-point communication setup. The distributed cases will be considered afterwards, where optimality conditions are derived along similar lines.

2.3.1 Optimal Decoder Given Encoder

Let $\mathbf{g}$ be fixed. Then the optimal decoder is the minimum mean square error (MMSE) estimator of $\mathbf{X}$ given $\hat{\mathbf{y}}$, i.e.,

$$\mathbf{h}(\hat{\mathbf{y}}) = \mathbb{E}\{\mathbf{X}|\hat{\mathbf{y}}\} \tag{2.11}$$

Plugging the expressions for expectation, we obtain

$$\mathbf{h}(\hat{\mathbf{y}}) = \int \mathbf{x}\, f_{X|\hat{Y}}(\mathbf{x}|\hat{\mathbf{y}})\, \mathrm{d}\mathbf{x}. \tag{2.12}$$

Applying Bayes' rule

$$f_{X|\hat{Y}}(\mathbf{x}|\hat{\mathbf{y}}) = \frac{f_X(\mathbf{x}) f_{\hat{Y}|X}(\hat{\mathbf{y}}|\mathbf{x})}{\displaystyle\int f_X(\mathbf{x})\, f_{\hat{Y}|X}(\hat{\mathbf{y}}|\mathbf{x})\, \mathrm{d}\mathbf{x}} \tag{2.13}$$

and noting that $f_{\hat{Y}|X}(\hat{\mathbf{y}}|\mathbf{x}) = f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x})]$, the optimal decoder can be written, in terms of known quantities, as

$$\mathbf{h}(\hat{\mathbf{y}}) = \frac{\displaystyle\int \mathbf{x}\, f_X(\mathbf{x})\, f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x})]\, \mathrm{d}\mathbf{x}}{\displaystyle\int f_X(\mathbf{x})\, f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x})]\, \mathrm{d}\mathbf{x}} \tag{2.14}$$

2.3.2 Optimal Encoder Given Decoder

Let $\mathbf{h}$ be fixed. Our goal is to minimize MSE subject to the average power constraint. Let us write MSE explicitly as a functional of $\mathbf{g}$

$$D[\mathbf{g}] = \int\int [\mathbf{x} - \mathbf{h}(\mathbf{g}(\mathbf{x}) + \mathbf{n})]^{\mathbf{T}} [\mathbf{x} - \mathbf{h}(\mathbf{g}(\mathbf{x}) + \mathbf{n})] f_X(\mathbf{x}) f_N(\mathbf{n}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{n} \tag{2.15}$$

To impose the power constraint, we construct the Lagrangian cost functional:

$$J[\mathbf{g}] = D[\mathbf{g}] + \lambda\{P[\mathbf{g}] - P_T\} \tag{2.16}$$

to minimize over the mapping $\mathbf{g}$. To obtain necessary conditions we apply the standard method in variational calculus [51]:

$$\left.\frac{\partial}{\partial \epsilon}\right|_{\epsilon=0} J\left[\mathbf{g}(\mathbf{x}) + \epsilon \eta(\mathbf{x})\right] = 0 \tag{2.17}$$

for all admissible variation functions $\eta(\mathbf{x})$. Note that, since the power constraint is accounted for in the cost function, the variation function $\eta(\mathbf{x})$ need not be restricted to satisfy the power constraint (all continuous differentiable functions $\eta : \mathbb{R}^m \to \mathbb{R}^k$ are admissible). Applying the above condition, we get

$$\int \left\{ \lambda \mathbf{g}(\mathbf{x}) - \int \mathbf{h}'(\mathbf{g}(\mathbf{x}) + \mathbf{n})[\mathbf{x} - \mathbf{h}(\mathbf{g}(\mathbf{x}) + \mathbf{n})] f_N(\mathbf{n}) \mathrm{d}\mathbf{n} \right\} \eta(\mathbf{x}) \mathbf{f_X}(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbf{0} \tag{2.18}$$

where $\mathbf{h}'$ denotes the Jacobian of the vector valued function $\mathbf{h}$. The equality for all admissible variation functions, $\eta(\mathbf{x})$, requires the expression in braces to be identically zero (more formally the functional derivative [51] vanishes at an extremum point of the functional). This gives the necessary condition for optimality as

$$\nabla J[\mathbf{g}] = 0 \tag{2.19}$$

where

$$\nabla J[\mathbf{g}] = \lambda f_X(\mathbf{x}) \mathbf{g}(\mathbf{x}) - \int \mathbf{h}'(\mathbf{g}(\mathbf{x}) + \mathbf{n})[\mathbf{x} - \mathbf{h}(\mathbf{g}(\mathbf{x}) + \mathbf{n})] f_N(\mathbf{n}) f_X(\mathbf{x}) \mathrm{d}\mathbf{n} \tag{2.20}$$

Unlike the decoder, the optimal encoder is not in closed form but a necessary condition for optimality is given. We summarize these results in the main theorem for the point-to-point setting:

Theorem 2.1. *Given source and noise densities, a coding scheme (g, h) is optimal only if*

$$\mathbf{g}(\mathbf{x}) = \frac{1}{\lambda} \int \mathbf{h}'(\mathbf{g}(\mathbf{x}) + \mathbf{n}) \left[\mathbf{x} - \mathbf{h}(\mathbf{g}(\mathbf{x}) + \mathbf{n})\right] f_N(\mathbf{n}) \mathrm{d}\mathbf{n} \tag{2.21}$$

$$\mathbf{h}(\hat{\mathbf{y}}) = \frac{\displaystyle\int \mathbf{x}\, f_X(\mathbf{x})\, f_N\left[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x})\right] \mathrm{d}\mathbf{x}}{\displaystyle\int f_X(\mathbf{x})\, f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x})]\, \mathrm{d}\mathbf{x}} \tag{2.22}$$

where varying λ provides solutions at different levels of power constraint P_T. In fact, λ is the slope of the distortion-power curve: $\lambda = \frac{dD}{dP_T}$.

The theorem states the necessary conditions for optimality but they are not sufficient, as is demonstrated in particular by the following corollary.

Corollary 2.2. *For Gaussian source and channel, the necessary conditions of Theorem 2.1 are satisfied by linear mappings $\mathbf{g}(\mathbf{x}) = \mathbf{k}_g\mathbf{x}$ and $\mathbf{h}(\mathbf{y}) = \mathbf{k}_h\mathbf{y}$ for some $\mathbf{k}_g, \mathbf{k}_h$.*

Although linear mappings satisfy the necessary conditions of optimality for the Gaussian case, they are known to be highly suboptimal when dimensions of source and channel do not match, i.e., $m \neq k$. Hence, this corollary illustrates the existence of poor local optima and the challenges facing algorithms based on these necessary conditions.

2.3.3 Optimality Conditions for Coding with Decoder Side Information

Optimality conditions for the settings of decoder side information (Figure 2.2) can be obtained by following similar steps. We note, in particular, that for these settings a similar result to Corollary 2.2 holds, i.e., for Gaussian sources and channels linear mappings satisfy the necessary conditions. Surprisingly, even in the matched bandwidth case, linear mappings will be shown to be suboptimal in the results section for such settings. This observation highlights the need for powerful numerical optimization tools. Let the encoder $\mathbf{g}$ be fixed. Then, the optimal decoder is the MMSE estimator of $\mathbf{X_1}$:

$$\mathbf{h}(\hat{\mathbf{y}}, \mathbf{x_2}) = \mathbb{E}\{\mathbf{X_1}|\hat{\mathbf{y}}, \mathbf{x_2}\}. \tag{2.23}$$

Plugging the expressions for expectation, applying Bayes' rule and noting that $f_{\hat{Y}|X_1}(\hat{\mathbf{y}}|\mathbf{x_1}) = f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x_1})]$, the optimal decoder can be written, in terms of known quantities, as

$$\mathbf{h}(\hat{\mathbf{y}}) = \frac{\displaystyle\int \mathbf{x_1}\, f_{X_1,X_2}(\mathbf{x_1}, \mathbf{x_2})\, f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x_1})]\, \mathrm{d}\mathbf{x_1}}{\displaystyle\int f_{X_1,X_2}(\mathbf{x_1}, \mathbf{x_2})\, f_N[\hat{\mathbf{y}} - \mathbf{g}(\mathbf{x_1})]\, \mathrm{d}\mathbf{x_1}}. \tag{2.24}$$

Now, let us assume that the decoder $\mathbf{h}$ is fixed. The distortion is expressed as a functional of $\mathbf{g}$:

$$D[\mathbf{g}] = \mathbb{E}\{||[\mathbf{X_1} - \mathbf{h}(\mathbf{g}(\mathbf{X_1}) + \mathbf{N}, \mathbf{X_2})]||^2\} \tag{2.25}$$

We construct the Lagrangian cost functional:

$$J[\mathbf{g}] = D[\mathbf{g}] + \lambda \left\{ P[\mathbf{g}] - P_T \right\}. \tag{2.26}$$

To obtain necessary conditions we apply the standard method in variational calculus:

$$\nabla J[\mathbf{g}](\mathbf{x_1}, \mathbf{x_2}) = \mathbf{0},\ \forall\, \mathbf{x_1},\ \mathbf{x_2} \tag{2.27}$$

where

$$\begin{aligned}\nabla J[\mathbf{g}](\mathbf{x_1}, \mathbf{x_2}) =& \lambda f_{X_1,X_2}(\mathbf{x_1}, \mathbf{x_2})\mathbf{g}(\mathbf{x_1}) \\ &- \int \mathbf{h}'(\mathbf{g}(\mathbf{x_1}) + \mathbf{n}, \mathbf{x_2})[\mathbf{x} - \mathbf{h}(\mathbf{g}(\mathbf{x}) + \mathbf{n}, \mathbf{x_2})] f_N(\mathbf{n}) f_{X_1,X_2}(\mathbf{x_1}, \mathbf{x_2})\mathbf{dn}\end{aligned} \tag{2.28}$$

2.3.4 Optimality Conditions for Distributed Coding

Now, we focus on the distributed coding setup.(Figure 2.3). Assume that the two encoders are fixed. Then, the optimal decoders are $\mathbf{h_1}(\mathbf{\hat{y}_1}, \mathbf{\hat{y}_2}) = \mathbb{E}[\mathbf{X_1}|\mathbf{\hat{y}_1}, \mathbf{\hat{y}_2}]$ and $\mathbf{h_2}(\mathbf{\hat{y}_1}, \mathbf{\hat{y}_2}) = \mathbb{E}[\mathbf{X_2}|\mathbf{\hat{y}_1}, \mathbf{\hat{y}_2}]$.

Next, we fix the decoders and find the encoders $\mathbf{g_1}, \mathbf{g_2}$ that minimize the total cost

$$J = D[\mathbf{g_1}, \mathbf{g_2}] + \lambda_1 \left\{ P[\mathbf{g_1}] - P_1 \right\} + \lambda_2 \left\{ P[\mathbf{g_2}] - P_2 \right\} \tag{2.29}$$

where

$$\begin{aligned}D[\mathbf{g_1}, \mathbf{g_2}] =& \mathbb{E}\{||\mathbf{X_1} - \mathbf{h_1}(\mathbf{g_1}(\mathbf{X_1}) + \mathbf{N_1}, \mathbf{g_2}(\mathbf{X_2}) + \mathbf{N_2})||^{\mathbf{2}} \\ &+ ||\mathbf{X_2} - \mathbf{h_2}(\mathbf{g_1}(\mathbf{X_1}) + \mathbf{N_1}, \mathbf{g_2}(\mathbf{X_2}) + \mathbf{N_2})||^{\mathbf{2}}\}\end{aligned} \tag{2.30}$$

The necessary conditions are derived by requiring

$$\nabla J[\mathbf{g_1}](\mathbf{x_1}, \mathbf{x_2}) = \nabla \mathbf{J}[\mathbf{g_2}](\mathbf{x_1}, \mathbf{x_2}) = \mathbf{0}\ \forall\, \mathbf{x_1},\ \mathbf{x_2}, \tag{2.31}$$

where $\nabla J[\mathbf{g_1}](\mathbf{x_1}, \mathbf{x_2}) = \frac{\partial \mathbf{J}}{\partial \mathbf{g_1}}$ and $\nabla J[\mathbf{g_2}](\mathbf{x_1}, \mathbf{x_2}) = \frac{\partial \mathbf{J}}{\partial \mathbf{g_2}}$. For more details on the underlying variational calculus method, see [51]. where $\nabla J[\mathbf{g_1}](\mathbf{x_1}, \mathbf{x_2}) = \frac{\partial \mathbf{J}}{\partial \mathbf{g_1}}$ and $\nabla J[\mathbf{g_2}](\mathbf{x_1}, \mathbf{x_2}) = \frac{\partial \mathbf{J}}{\partial \mathbf{g_2}}$ [51].

2.4 Algorithm Design

The basic idea is to iteratively alternate between the imposition of individual necessary conditions for optimality, and thereby successively decrease the total Lagrangian cost. Iterations are performed until the algorithm reaches a stationary point. Imposing optimality condition for the decoder is straightforward, since the decoder can be expressed as closed form functional of known quantities, $\mathbf{g}(\mathbf{x})$, $f_X(\mathbf{x})$ and $f_N(\mathbf{n})$. The encoder optimality condition is not in closed form and we perform steepest descent search in the direction of the functional derivative of the Lagrangian with respect to the encoder mapping(s) $\mathbf{g}$ ($\mathbf{g_1}$, $\mathbf{g_2}$ for two encoder case). By design, the Lagrangian cost decreases monotonically as the algorithm proceeds iteratively. The update for the various encoders is stated generically as

$$\mathbf{g}_{i+1}(\mathbf{x}) = \mathbf{g}_i(\mathbf{x}) - \mu \nabla \mathbf{J}[\mathbf{g}] \tag{2.32}$$

where i is the iteration index, $\nabla \mathbf{J}[\mathbf{g}]$ is the directional derivative and μ is the step size. At each iteration i, the total cost decreases monotonically and iterations are kept until convergence. As initialization for the encoder mapping optimization, previously proposed heuristic suboptimal mappings [11, 59] can be used. Note that there is no guarantee that an iterative descent algorithms of this type will converge to the globally optimal solution. The algorithm will converge to a local minimum. An important observation is that, in the case of Gaussian sources and channels, the linear encoder-decoder pair satisfies the necessary conditions of optimality, although, as we will illustrate, there are other mappings that perform better. Hence, initial conditions have paramount importance in such greedy optimizations. A preliminary low complexity approach to mitigate the poor local minima problem, is to embed in the solution the noisy relaxation method of [18, 45]. We initialize the encoding mapping(s) with random initial conditions and run the algorithm at very low CSNR (high Lagrangian parameter λ). Then, we gradually increase the CSNR (decrease λ) while tracking the minimum until we reach the prescribed CSNR (or power P_T for a given channel noise level).

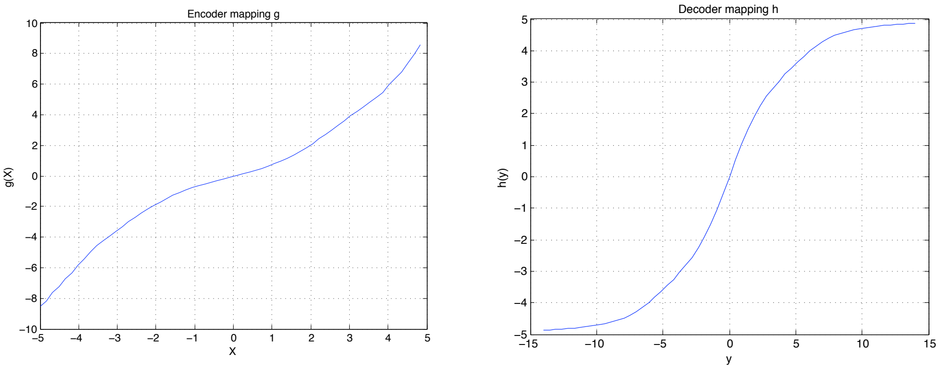
2.5 Results

We implemented the above algorithm by numerically calculating the derived integrals. For that purpose, we sampled the distribution on a uniform grid. We also imposed bounded support (-3σ to 3σ) i.e., neglected tails of infinite support distributions in the examples.

2.5.1 Scalar Mappings ($m = 1, k = 1$), Gaussian Mixture Source and Gaussian Channel

We consider a Gaussian mixture source with distribution

$$f_x(x) = \frac{1}{2\sqrt{2\pi}} \left\{ e^{\frac{-(x-3)^2}{2}} + e^{\frac{-(x+3)^2}{2}} \right\} \tag{2.33}$$



(a) Encoder mapping

(b) Decoder mapping

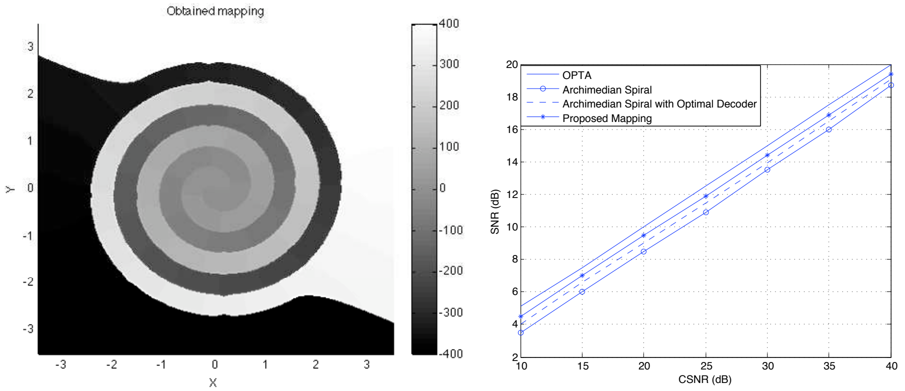
Figure 2.4. Encoder mapping for bi-modal GMM source, Gaussian channel, modes at 3 and -3 as in (2.33)

and unit variance Gaussian noise. The encoder and decoder mappings for this source-channel setting are given in Figure 2.4. As intuitively expected, since the two modes of the Gaussian mixture are well separated, each mode locally behaves as Gaussian. Hence the curve is roughly piece-wise linear, deviating significantly from a truly linear mapping. This illustrates the importance of nonlinear mappings for general distributions that diverge from the pure Gaussian.

2.5.2 $(m = 2, k = 1)$ Gaussian source-channel mapping

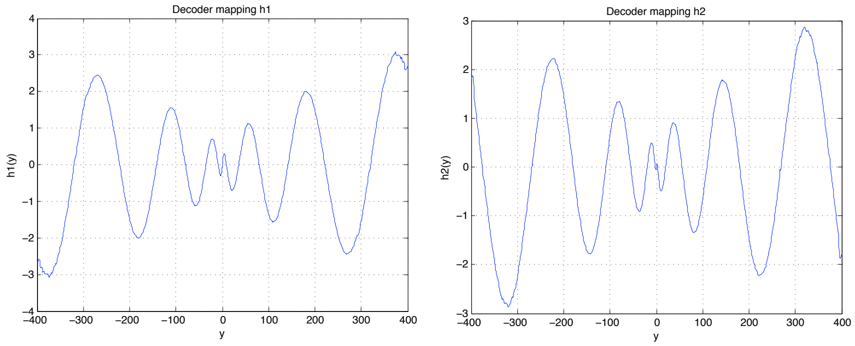
In this section, we present a bandwidth compression example with 2:1 mappings for Gaussian source and channel. We compare the proposed mapping to the asymptotic bound (OPTA) and prior work [34]. We also compare the optimal encoder-decoder pair to the setting where only the decoder is optimized and the encoder is fixed. In prior work [11, 59, 34], the Archimedian spiral is found to perform well for Gaussian 2:1 mappings, and used for encoding and decoding with maximum likelihood criteria. We hence initialize our algorithm with an Archimedian spiral (for the encoder mapping). For details of the Archimedian spiral and its settings, see eg.[34] and references therein.

The obtained encoder mapping is shown in Figure 2.6. While the mapping produced by our algorithm resembles the Archimedian spiral, there is still a significant difference which will also be evident in the performance results. Note further that the encoding scheme differs from prior work in that we continuously map the source to the channel signal, where the two dimensional source is mapped to the closet point on the space filling spiral. The resulting mapping is also "spiral shaped" but the distance between consecutive spiral arms is not constant, unlike



(a) (b)

Figure 2.5. a)Encoder 2:1 mapping for Gaussian source/channel, CSNR=40dB, SNR=19.45dB. The axes show the two dimensional input ($\mathbf{x}$) and the function value ($\mathbf{g}(\mathbf{x})$) is reflected in the intensity level. b)Comparative results for Gaussian source-channel, 2:1 mapping



(a) (b)

Figure 2.6. Decoder mappings (1:2 mapping consisting 2 scalar 1:1 mappings) for the same example.

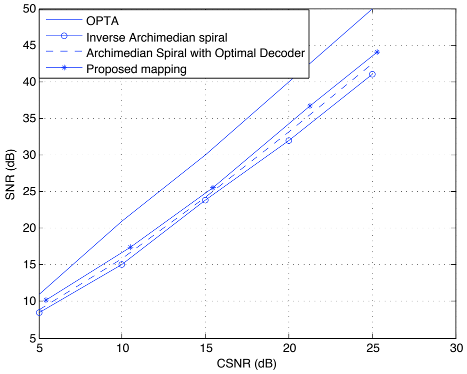
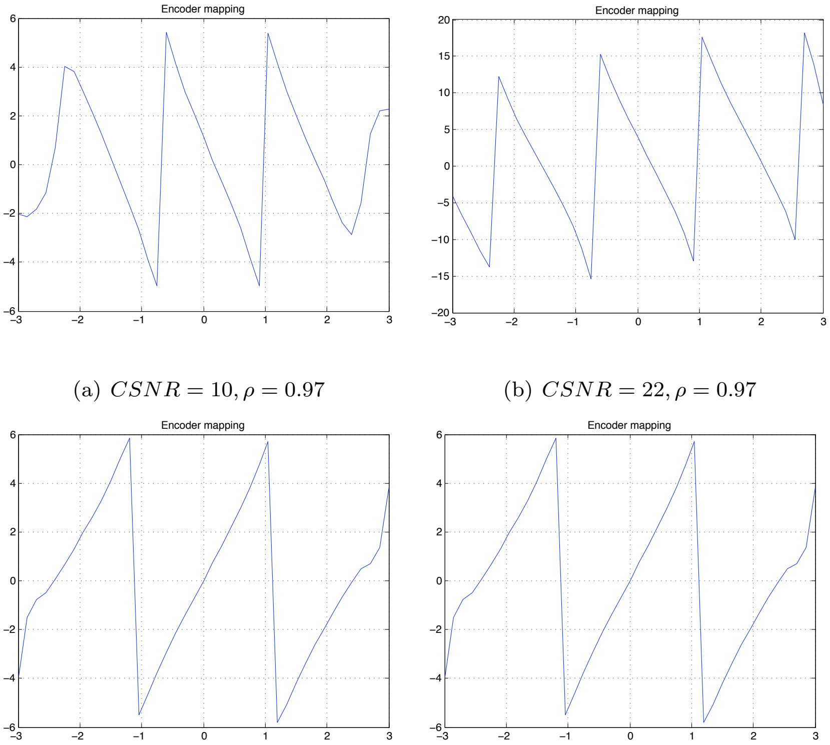


Figure 2.7. Comparative results for Gaussian source- channel, 1:2 mapping

the pure Archimedian spiral. The comparative performance results are shown in Figure 2.6. The proposed mapping outperforms the Archimedian spiral [34] over the entire range of CSNR values. It is notable that the "intermediate" option of only optimizing the decoder improves the performance significantly, compared to using the inverse spiral with maximum likelihood decoding.

2.5.3 $(m = 1, k = 2)$ Gaussian source-channel mapping

In this section we compare the proposed mappings for bandwidth expansion where a Gaussian scalar source is transmitted over a two dimensional Gaussian channel. We compare the obtained mapping to prior work and OPTA. Here, we use the inverse spiral of prior work as the initial condition. The results are presented in Figure 2.7. The proposed mapping outperforms the inverse of Archimedian spiral [34] over the entire range of CSNR values. Note that the gap between OPTA and and the achieved performance by our mappings is significantly greater than in the 2:1 bandwidth compression case. We consider two possible explanations: i) The gap between the best achievable zero-delay performance and OPTA (which is the achievable limit when delay is asymptotically large) is large compared to bandwidth compression. Obviously, this gap monotonically decreases with the dimensions of the mappings, i..e, 2:4 mappings would outperform 1:2 mappings. ii) Our mappings may have converged to a local minimum that is significantly worse than the global minimum. We suspect the latter and that global optimization tools, such as deterministic annealing [63], will offer substantial gains.



(a) $CSNR = 10, \rho = 0.97$

(b) $CSNR = 22, \rho = 0.97$

(c) $CSNR = 10, \rho = 0.9$

(d) $CSNR = 22, \rho = 0.9$

Figure 2.8. Encoder mappings for Gaussian scalar source, channel and side information at different CSNR and correlation levels

2.5.4 Source-Channel Coding with Decoder Side Information

In this section, we demonstrate the use of the proposed algorithm by focusing on the specific scenario of Figure 2.2. It must be emphasized that, while the algorithm is general and directly applicable to any choice of source and channel dimensions and distributions, for conciseness of the results section we will assume that sources are jointly Gaussian scalars with correlation coefficient ρ, and are identically distributed. We also assume that the noise is scalar and Gaussian.

Figure 2.8 presents a sample of encoding mappings obtained by varying the correlation coefficient and CSNR. Interestingly, the analog mapping captures the central characteristic observed in digital Wyner-Ziv mappings, in the sense of many-to-one mappings, where multiple source intervals are mapped to the same channel interval, which will potentially be resolved by the decoder given the side information. Within each bin, there is a mapping function which is approximately linear in this case (scalar Gaussian sources and channel). To see the effect

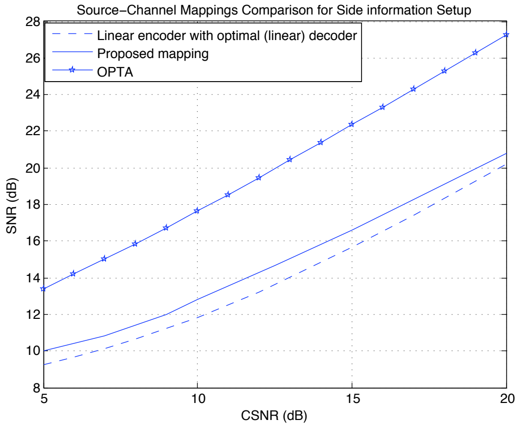


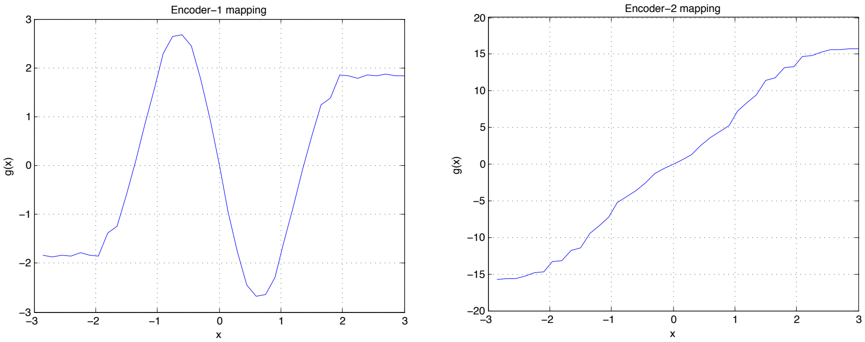
Figure 2.9. Comparative results for correlation coefficient $\rho = 0.9$, Gaussian scalar source, channel and side information

of correlation on the encoding mappings, we note how the mapping changes as we lower the correlation from $\rho = 0.97$ to $\rho = 0.9$. As intuitively expected, the side information is less reliable and source points that are mapped to the same channel representation grow further apart from each other. Also note that inclinations of the mappings are different for each CSNR, due to the fact that depending on the initial conditions, algorithm can converge to different local minima. Note that if $g(x)$ satisfies necessary conditions of Theorem 2.1 (i.e., locally optimal), so does $-g(x)$, due to the symmetry (around zero) of the involved distributions. Comparative performance results are shown in Figure 2.9. The proposed mapping significantly outperforms linear mapping over the entire range of CSNR values.

2.5.5 Distributed Source-Channel Coding

Here we consider jointly Gaussian sources that are transmitted separately over independent channels, as shown in Figure 2.3. Note that our algorithm and derived necessary conditions allow the channels to be correlated, but for simplicity we restrict to independent, additive Gaussian channels. To demonstrate the power of the design approach and the type of gains achievable, we consider an asymmetric power allocation to separate encoders. We chose $\lambda_1 = 0.1\lambda_2$ in the optimization framework. Since both source and noise are distributed symmetrically around zero, the encoding and decoding mappings must also be symmetric. We use this fact to halve the number of samples needed in the algorithm and correspondingly reduce the complexity.

Figure 2.10 presents an example of encoding mappings for correlation coefficient $\rho = 0.95$. The comparative performance results are shown in Figure 2.11. The proposed mapping outper-



(a) $g_1(x)$

(b) $g_2(x)$

Figure 2.10. Obtained encoder mappings for correlation coefficient $\rho = 0.95$, Gaussian scalar sources and channels, $CSNR = 15$

forms the linear mapping over the entire range of CSNR values. Note that encoders are highly asymmetric in the sense that one is "Wyner-Ziv like" and maps several source intervals to the same channel interval, whereas the other encoder is an almost monotone increasing function. While we can offer tentative explanation for the form of this solution, we above all observe that it does substantially outperform the symmetric, linear solution. This demonstrates that by breaking from the linear solution and re-optimizing we do obtain an improved solution. Moreover, it strongly suggests that some symmetric but highly non-linear solution may exist, whose discovery would need more powerful optimization tools – a direction we are currently pursuing.

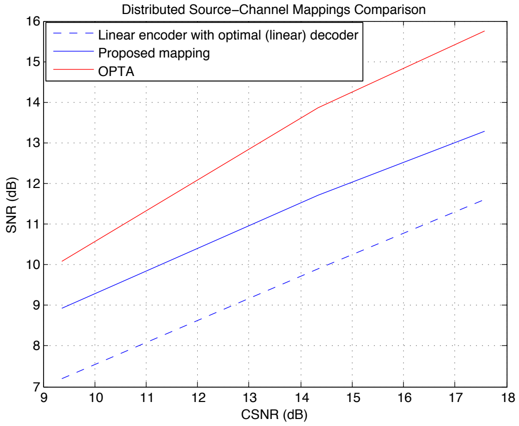


Figure 2.11. Comparative results for correlation coefficient $\rho = 0.95$, Gaussian scalar sources and channels

Chapter 3

Linearity of Optimal Estimation

3.1 Introduction

Consider a basic problem in estimation theory, namely, source estimation from a signal received through a channel with additive noise, given the statistics of both source and channel. The optimal estimator that minimizes the mean square error (MSE) is usually a nonlinear function of the observation. A frequently exploited result in estimation theory concerns the special case of Gaussian source and Gaussian noise, a case in which the MSE optimal estimator is guaranteed to be linear. An open follow-up question considers the existence of other cases exhibiting such a “coincidence”, and more generally the characterization of conditions for linearity of optimal estimators for general distortion measures.

This problem also has practical importance beyond theoretical interest, mainly due to significant complexity issues in both design and operation of estimators. Specifically, the optimal estimator generally involves entire probability distributions, whereas linear estimators require only up to second-order statistics for their design. Moreover, unlike the optimal estimator which can be an arbitrarily complex function that is difficult to implement, the linear estimator consists of a simple matrix-vector operation. Hence, linear estimators are more prevalent in practice, despite their suboptimal performance in general. They also represent a significant temptation to “assume” that processes are Gaussian, sometimes despite overwhelming evidence to the contrary. Results in this part of the thesis identify the cases where a linear estimator is optimal, and when the use of linear estimators is justified in practice without recourse to complexity arguments.

The estimation problem in general has been studied intensively in the literature [6, 64, 62, 21, 10, 60]. Our preliminary results appeared in [4]. It is known that, for stable distributions[1] (which includes the Gaussian distribution as the only finite variance member), the optimal estimator is linear at all signal to noise ratios (SNR). Stable distributions are a subset of a family called infinitely divisible distributions which, as we show in this part of the thesis, satisfy the derived necessary conditions for the existence of a matching source/noise distribution such that the optimal estimator is linear at any SNR level. Our main contribution relative to prior work, which studied linearity as it applies simultaneously at all SNR levels, focuses on the linearity of optimal estimation for the L_p norm and its dependence on the SNR level. Specifically, we present the optimality conditions for linearity of optimal estimators at a specified SNR, where optimality is in the sense of the L_p norm. As an important special case, we investigate the $p = 2$ case (mean square error) in detail. Note that a similar problem has been studied in [47, 9] for the special case of the mean square error, albeit without further study of the question of existence and uniqueness of "matching" distributions. We show that the necessary conditions presented in [47, 9] are subsumed in our general necessary and sufficient conditions; and specify conditions for which such matching distributions exist and are unique. The analysis is then extended to the case of vector spaces. Interestingly, this extension is non-trivial and new constraints, beyond those inherited from the scalar case, must be satisfied to ensure linearity of optimal estimation.

Five results are provided on the linearity of optimal estimation. First, we show that if a given noise (alternatively, a given source) distribution satisfies certain conditions, there always exists a matching source (alternatively, noise) distribution of a given power, for which the optimal estimator is linear. We further identify conditions under which such a matching distribution does *not* exist. Secondly, we show that if the source and the noise have the same variance, they *must* be identically distributed to ensure the linearity of the optimal estimator. Having established more general conditions for linearity of optimal estimation, one wonders in what precise sense the Gaussian case may be special. This question is answered by the third result. We consider the optimality of linear estimation at multiple SNR values. Let random variables X and N be source and noise, respectively, and allow for scaling of either to produce varying levels of SNR. We show that if the optimal estimator is linear at more than one SNR value, then both the source X and the noise N must be Gaussian. In other words, the Gaussian source-noise pair is unique in the sense that it offers linearity of optimal estimators at multiple SNR values (in fact the optimal estimator is linear at all SNR as is well known). As a fourth result, we show that the MSE optimal estimator converges to a linear estimator for any source and Gaussian noise at asymptotically low SNR, and vice versa, for any noise and Gaussian source at asymptotically high SNR.

[1] A distribution is called stable if for independent and identically distributed X_1, X_2, X; for any constants a,

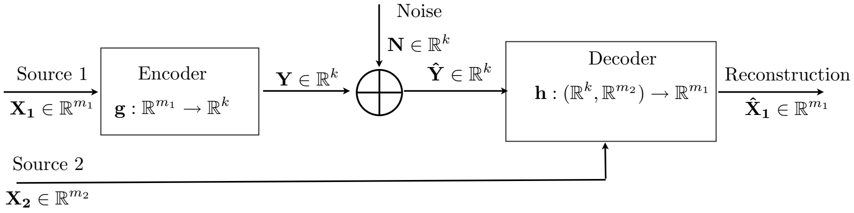

Noise
$\mathbf{N} \in \mathbb{R}^k$
Source 1
$\mathbf{X_1} \in \mathbb{R}^{m_1}$
Encoder
$\mathbf{g} : \mathbb{R}^{m_1} \to \mathbb{R}^k$
$\mathbf{Y} \in \mathbb{R}^k$
$\hat{\mathbf{Y}} \in \mathbb{R}^k$
Decoder
$\mathbf{h} : (\mathbb{R}^k, \mathbb{R}^{m_2}) \to \mathbb{R}^{m_1}$
Reconstruction
$\hat{\mathbf{X}}_\mathbf{1} \in \mathbb{R}^{m_1}$
Source 2
$\mathbf{X_2} \in \mathbb{R}^{m_2}$

Figure 3.1. The general setup of the problem

Finally, we analyze the vector case, where conditions for linearity of optimal estimation are more stringent. We show that for a vector source-channel pair with identical dimensions, the conditions derived for the scalar case become necessary conditions in a transform domain, where the transform jointly diagonalizes the source and channel covariance matrices. We further derive the additional, complementary conditions that must be satisfied to achieve sufficiency.

3.2 Review of Optimal and Linear Estimation

3.2.1 Preliminaries and Notation

We consider the problem of estimating source X given the observation $Y = X + N$, where X and N are independent, as shown in Figure 3.1. Let X and N be scalar zero mean random variables with respective densities $f_X(\cdot)$ and $f_N(\cdot)$ and characteristic functions $F_X(\omega)$ and

b; the random variable $aX_1 + bX_2$ has the same distribution as $cX + d$ for some constants c and d [10].

$F_N(\omega)$. A density $f(x)$ is said to be symmetric if it is an even characteristic function[2]: $f(x) = f(-x)\ \forall x \in \mathbb{R}$. The SNR is $\gamma = \frac{\sigma_x^2}{\sigma_n^2}$. All random variables, in a statement regarding L_p norm optimal estimation, are constrained to have finite p^{th} moment, eg. in a result associated with MSE we assume finite variances, i.e., $\sigma_x^2 < \infty, \sigma_n^2 < \infty$. All the logarithms in this part are natural logarithms and may in general be complex.

An estimator $h(\cdot)$ is a function of the observation and is said to be optimal if it minimizes the cost functional

$$J(h) = \mathbb{E}\left\{\Phi(X, h(Y))\right\} \tag{3.1}$$

for a given distortion measure Φ. Specializing (3.1) to a difference distortion measure, we explicitly get:

$$J(h) = \int\int \Phi(x - h(y)) f_X(x) f_{Y|X}(y|x) dx dy \tag{3.2}$$

To obtain the necessary conditions for optimality, we apply the standard method in variational calculus [51]:

$$\frac{\partial}{\partial\epsilon} J\left[h(y) + \epsilon\eta(y)\right]\bigg|_{\epsilon=0} = 0 \tag{3.3}$$

for all admissible variation functions $\eta(y)$. If Φ is differentiable, (3.3) yields

$$\int\int \Phi'(x - h(y))\eta(y) f_X(x) f_{Y|X}(y|x) dx dy = 0 \tag{3.4}$$

or,

$$\mathbb{E}\left\{[\Phi'(X - h(Y)]\eta(Y)\right\} = 0 \tag{3.5}$$

where Φ' is the derivative of Φ.

[2] Note that this definition will need generalization to symmetry about any point when one drops the assumption of zero-mean distributions

3.2.2 Optimality condition for L_p norm

Hereafter, we will specialize to the case of the L_p metric[3] with $p \in \mathbb{R}^+$, i.e., $\Phi(x) = ||x||_p^p$. Using the fact that $\frac{d}{dx}||x||_p^p = p\frac{||x||_p^p}{x}, \forall x \in \mathbb{R} - \{0\}$, we derive the necessary condition for optimality of an estimator as :

$$\mathbb{E}\left\{\frac{||X - h(Y)||_p^p}{X - h(Y)}\eta(Y)\right\} = 0 \tag{3.6}$$

When we specialize to even integer p, we obtain the following

$$\mathbb{E}\left\{[X - h(Y)]^{p-1}\eta(Y)\right\} = 0 \tag{3.7}$$

Note that for $p = 2$, or $\Phi(x) = x^2$, this condition reduces to the well known orthogonality condition of MSE, i.e., the following holds

$$\mathbb{E}\left\{[(X - h(Y)]\eta(Y)\right\} = 0 \tag{3.8}$$

for any $\eta(\cdot)$ function. The MSE optimal estimator $h(Y) = \mathbb{E}\{X|Y\}$ can be directly obtained from (3.8).

Note that for $p = 1$, this expression results in $h(Y)$ being the median operator, which is known as the centroid condition for the L_1 norm (see e.g. [20]). As the following lemma formally states, the above L_p necessary condition is also sufficient.

Lemma 3.1. *The necessary condition stated in (3.6) is sufficient. Moreover, the estimator in (3.7) is unique.*

Proof. First we show the sufficiency of the necessary conditions for L_p norm. Note that, $\Phi(x) = ||x||_p^p$ is convex, i.e., $\frac{d^2||x||_p^p}{dx^2} > 0$, $\forall x - \{0\}$. We need to show $\frac{\partial^2}{\partial^2\epsilon}J\left[h(y) + \epsilon\eta(y)\right]\Big|_{\epsilon=0} > 0$, for any $\eta(y)$ variation function.

$$\frac{\partial^2}{\partial^2\epsilon}J\left[h(y) + \epsilon\eta(y)\right]\Bigg|_{\epsilon=0} = \int\int \eta^2(y)\Phi^{''}(x - h(y))f_X(x)f_{Y|X}(y|x)dxdy \tag{3.9}$$

All factors in the integral is non negative and hence, $\frac{\partial^2}{\partial^2\epsilon}J\left[h(y) + \epsilon\eta(y)\right]\Bigg|_{\epsilon=0} > 0$, for any $\eta(y)$.

Next we show the uniqueness of the optimal estimator for even p. Assume $h_1(Y)$ and $h_2(Y)$ both satisfy (3.7) while $h_1(Y) \neq h_2(Y)$, $\exists Y \in \mathbb{R}$. Then, the following holds

$$\mathbb{E}\left\{\{[X - h_2(Y)]^{p-1} - [X - h_1(Y)]^{p-1}\}\eta(Y)\right\} = 0 \tag{3.10}$$

[3] $||x||_p^p = |x|^p$ for a scalar x, where $|\cdot|$ is the absolute value operator.

Noting

$$[X - h_2(Y)]^{p-1} - [X - h_1(Y)]^{p-1} = (h_1(Y) - h_2(Y))\beta(X, Y) \tag{3.11}$$

where

$$\beta(X, Y) = \sum_{m=0}^{p-1} [X - h_1(Y)]^m [X - h_2(Y)]^{p-1-m} \tag{3.12}$$

Clearly, $\beta(X, Y) > 0$. Plugging $\eta(Y) = h_1(Y) - h_2(Y)$ in (3.10), we obtain,

$$\mathbb{E}\left\{[h_1(Y) - h_2(Y)]^2 \beta(X, Y)\right\} = 0 \tag{3.13}$$

Since $\beta(X, Y) > 0 \; \forall X, Y \in \mathbb{R}$,

$$\mathbb{E}\left\{[h_1(Y) - h_2(Y)]^2\right\} = 0 \tag{3.14}$$

Hence $h_1(Y) = h_2(Y)$ almost everywhere, contradicting the initial assumption $h_1(Y) \neq h_2(Y), \exists Y \in \mathbb{R}$. □

Note: While (3.6) is valid for general L_p, $(p \in \mathbb{R}^+)$, the remainder of the chapter is restricted to even integer p.

3.2.3 L_p Optimal Linear Estimation

While perturbing the optimal linear estimator, the variation function $\eta(y)$ must be linear to ensure that $h(y) + \epsilon\eta(y)$ is linear. Plugging $h(y) = kY$ and $\eta(y) = aY$ (for some $a \in \mathbb{R}$) in (3.7) and omitting straightforward steps, we obtain the condition for optimal linear estimation:

$$\mathbb{E}\left\{(X - kY)^{p-1} Y\right\} = 0 \tag{3.15}$$

The optimal scaling coefficient k can be found by plugging $Y = X + N$ into (3.15). Observe that for $p = 2$, we get the well known result $k = \frac{\gamma}{\gamma+1}$.

3.2.4 Gaussian Source and Channel

We next consider the special case in which both X and N are Gaussian, $X \sim \mathcal{N}(0, \sigma_x^2)$ and $N \sim \mathcal{N}(0, \sigma_n^2)$. The linear estimator

$$h(Y) = \frac{\gamma}{\gamma + 1} Y \tag{3.16}$$

is well known to be the optimal MSE estimation. Relatively less known is the fact that this linear estimator is optimal more generally for the L_p norm [68]. It is straightforward to show that this linear estimator satisfies (3.7) by rendering the reconstruction error $X - h(Y)$ independent of Y.

3.3 Conditions for Linearity of Optimal Estimation

In this section, we find the necessary and sufficient conditions in terms of characteristic functions $F_X(\omega)$ and $F_N(\omega)$ that ensure that $h(Y) = kY$ is the optimal estimator for some $k \in \mathbb{R}$. We first provide the result for the L_p norm, which takes the form of a differential equation that must be satisfied to ensure linearity of optimal estimation, and then specialize it to the MSE case.

3.3.1 L_p Norm Condition

Theorem 3.2. *For a given L_p distortion measure, a source X with characteristic function $F_X(\omega)$ and a noise N with characteristic function $F_N(\omega)$, the optimal estimator $h(Y)$ is linear, $h(Y) = kY$,* if and only if *the following differential equation is satisfied:*

$$\sum_{m=0}^{p-1} F_X^{(m)}(\omega) F_N^{(p-1-m)}(\omega) \binom{p-1}{m} \left(\frac{k-1}{k}\right)^m = 0 \tag{3.17}$$

Proof. Plugging in $f_{Y|X}(y|x) = f_N(y-x)$ in (3.7), we obtain

$$\int [x-ky]^{p-1} f_X(x) f_N(y-x) dx = 0, \forall y \tag{3.18}$$

Expressing $[x-ky]^{p-1}$ in binomial expansion,

$$[x-ky]^{p-1} = \sum_{m=0}^{p-1} \binom{p-1}{m} (-ky)^m x^{p-m-1} \tag{3.19}$$

Arranging the terms, we get

$$\sum_{m=0}^{p-1} \binom{p-1}{m} (-ky)^m \int x^{p-1-m} f_X(x) f_N(y-x) dx = 0 \tag{3.20}$$

Let $\otimes$ denote the convolution operator and rewrite (3.20) as

$$\sum_{m=0}^{p-1}\binom{p-1}{m}(-ky)^m\left[y^{p-1-m}f_X(y)\otimes f_N(y)\right]=0 \tag{3.21}$$

Taking the Fourier transform (assuming the Fourier transform exists),

$$\sum_{m=0}^{p-1}\binom{p-1}{m}(-k)^m\frac{d^m}{d\omega^m}\left[\frac{d^{p-1-m}(F_X(\omega))}{d\omega^{p-1-m}}F_N(\omega)\right]=0 \tag{3.22}$$

Opening up the above expression,

$$\sum_{m=0}^{p-1}\binom{p-1}{m}(-k)^m\sum_{l=0}^{m}\binom{m}{l}\frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}}\frac{d^lF_N(\omega)}{d\omega^l}=0 \tag{3.23}$$

Interchanging the summations,

$$\sum_{l=0}^{p-1}\frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}}\frac{d^lF_N(\omega)}{d\omega^l}\sum_{m=l}^{p-1}\binom{p-1}{m}(-k)^m\binom{m}{l}=0 \tag{3.24}$$

Opening the binomials,

$$\sum_{l=0}^{p-1}\binom{p-1}{l}\frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}}\frac{d^lF_N(\omega)}{d\omega^l}\sum_{m=l}^{p-1}\frac{(p-1-l)!}{(m-l)!(p-1-m)!}(-k)^m=0 \tag{3.25}$$

Replacing $t=m-l$, we get

$$\sum_{l=0}^{p-1}\binom{p-1}{l}\frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}}\frac{d^lF_N(\omega)}{d\omega^l}\sum_{t=0}^{p-1-l}\binom{p-1-l}{t}(-k)^{(t+l)}=0 \tag{3.26}$$

Noting that,

$$(1-k)^{p-1-l}=\sum_{t=0}^{p-1-l}\binom{p-1-l}{t}(-k)^t \tag{3.27}$$

we obtain (3.17).

The converse part of the theorem follows from the fact that the necessary condition given in (3.7) is also sufficient. Recall that the sufficiency is shown in Lemma 3.1 using the convexity property of L_p norm. □

3.3.2 Specializing to MSE: The Matching Condition

In this section, we explore the impact of Theorem 3.2 for the special case of the mean square error distortion metric, i.e., $p=2$. More precisely, we wish to find the entire set of source and channel distributions such that $h(Y)=\frac{\gamma}{\gamma+1}Y$ is the optimal estimator for a given SNR, γ.

Note that this condition was derived, in another context [47, 9], albeit without consideration of important implications which we focus on, including the conditions for existence and uniqueness of matching distributions. Specifically, we identify the conditions for existence and uniqueness of a source distribution that *matches* the noise (and vice versa) in a way that guarantees the linearity of the optimal estimator. We state the main result for MSE in the following theorem.

Theorem 3.3. *Given SNR level γ, and noise N with characteristic function $F_N(\omega)$, there exists a source X for which the optimal estimator is linear* if and only if *the characteristic function*

$$F(\omega) = F_N^{\gamma}(\omega)$$

is a legitimate characteristic function. Moreover, if $F(\omega)$ is legitimate, then it is the characteristic function of the matching source, i.e. $F_X(\omega) = F(\omega)$. (An equivalent theorem holds where we replace "noise" for "source" everywhere, i.e., given source and SNR level, we have a condition for existence of a matching noise.)

Proof. Plugging $p = 2$ in (3.17) yields

$$\frac{1}{F_X(\omega)} \frac{dF_X(\omega)}{d\omega} = \gamma \frac{1}{F_N(\omega)} \frac{dF_N(\omega)}{d\omega} \tag{3.28}$$

or more compactly,

$$\frac{d}{d\omega} \log F_X(\omega) = \gamma \frac{d}{d\omega} \log F_N(\omega) \tag{3.29}$$

The solution to this differential equation is given by:

$$\log F_X(\omega) = \gamma \log F_N(\omega) + C \tag{3.30}$$

where C is a constant. Imposing $F_N(0) = F_X(0) = 1$, we obtain $C = 0$, which implies:

$$F_X(\omega) = F_N^{\gamma}(\omega) \tag{3.31}$$

□

Hence, given a noise distribution, the necessary and sufficient condition for the existence of a matching source distribution boils down to the requirement that $F_N^{\gamma}(\omega)$ be a valid characteristic function. Moreover, if such a matching source exists, we have a recipe for deriving its distribution.

3.3.3 Existence of a Matching Source for a Given Noise

Bochner's theorem [62] states that a continuous function $F : \mathbb{R} \to \mathbb{C}$ with $F(0) = 1$ is a valid characteristic function if and only if it is *positive semi-definite.*[4] Hence, the existence of a matching source depends on the positive semi-definiteness of $F_N^{\gamma}(\omega)$.

We note that characterizing the entire set of $F_N(\omega)$ where $F_N^{\gamma}(\omega)$ is positive semi-definite is a long-standing open problem. Instead we illustrate the result with various cases of interest where $F_N^{\gamma}(\omega)$ is, or is not, positive semi-definite. Let us start with a simple but useful case.

Corollary 3.4. *If SNR $\gamma \in \mathbb{Z}$, a matching source distribution exists, regardless of the noise distribution.*

Proof. From (3.31), integer γ implies:

$$X = \sum_{i=1}^{\gamma} N_i \tag{3.32}$$

where N_i are independent and distributed identically to N. Hence, $F_N^{\gamma}(\omega)$ is a valid characteristic function and a matching X exists. □

Next, we recall the concept of infinite divisibility, which is closely related to the problem at hand.

Definition [52]: A distribution with characteristic function $F(\omega)$ is called infinitely divisible, if for each integer $k \geq 1$, there exists a characteristic function $F_k(\omega)$ such that

$$F(\omega) = F_k^k(\omega) \tag{3.33}$$

Alternatively, $f_X(x)$ is infinitely divisible if and only if the random variable X can be written as $X = \sum_{i=1}^{k} X_i$ for any k where $\{X_i, i = 1, ..., k\}$'s are independent and identically distributed.

Infinitely divisible distributions have been studied extensively in probability theory [52, 73]. It is known that Poisson, exponential, and geometric distributions as well as the set of stable distributions (which includes the Gaussian distribution) are infinitely divisible. On the other hand, it is easy to see that distributions of discrete random variables with finite alphabets are not infinitely divisible.

[4] Let $f : \mathbb{R} \to \mathbb{C}$ be a complex-valued function, and $t_1, ..., t_s$ be a set of points in $\mathbb{R}$. Then f is said to be positive semi-definite (non-negative definite) if for any $t_i \in \mathbb{R}$ and $a_i \in \mathbb{C}$, $i = 1, ..., s$ we have

$$\sum_{i=1}^{s} \sum_{j=1}^{s} a_i {a_j}^* f(t_i - t_j) \geq 0$$

Corollary 3.5. *A matching source distribution exists for all $\gamma \in \mathbb{R}^+$ if and only if $f_N(n)$ is infinitely divisible.*

Proof. It is easy to show from the definition of infinite divisibility and Corollary 3.4 that if $f_N(n)$ is infinitely divisible, $F_N^r(\omega)$ is a valid characteristic function for all rational $r > 0$. Using the fact that every $\gamma \in \mathbb{R}$ is a limit of a sequence of rational numbers r_n, and by the continuity theorem [10], we conclude that $F_X(\omega) = F_N^\gamma(\omega)$ is a valid characteristic function, and hence a matching source exists.

If $F_X(\omega) = F_N^\gamma(\omega)$ is a valid characteristic function for all γ, then $f_N(n)$ is infinitely divisible, because we can choose $\gamma = \frac{1}{k}$ for $k \in \mathbb{Z}^+$ with $F_k(\omega) = F_X(\omega)$ in (3.33). □

At a given SNR, there may exist a matching source, even though $f_N(n)$ is not infinitely divisible. For example, a finite alphabet discrete random variable V is not infinitely divisible but still can be k-divisible, where $k < |V| - 1$ and $|V|$ is the cardinality of V. Hence, when $\gamma = \frac{1}{k}$, there may exist a matching source, even when the noise is not infinitely divisible. Many examples follow directly from Corollary 3.4.

We next cite a theorem, regarding analytic characteristic functions, which will be useful in the proofs that follow.

Theorem [52]: A characteristic function $F(\omega)$ is analytic *if and only if* F has finite moments of all orders and there exists a finite β such that $\mathbb{E}\{|X^k|\} \leq k!\beta^k, \forall k \in \mathbb{Z}^+$. This requirement is equivalent to the existence of a moment generating function. A characteristic function $F(\omega)$ is analytic *if and only if* the moments $\mathbb{E}\{|X^k|\}$ uniquely characterize the distribution, which in general is not the case, see eg. [69].

A useful property regarding the analyticity of the characteristic function of the matching source (or noise) is captured by the following corollary.

Corollary 3.6. *If $F_N(\omega)$ (or $F_X(\omega)$) is analytic, then the matching $F_X(\omega)$ (or $F_N(\omega)$), if it exists, is analytic.*

Proof. Recall the orthogonality property of the MSE optimal estimator (3.8). Let $\eta(Y) = Y^m$ for $m = 1, 2, 3...M$. Plugging the best linear estimator $h(Y) = \frac{\gamma}{\gamma+1}Y$ and replacing Y with

where $a_j{}^*$ is the complex conjugate of a_j. Equivalently, we require that the $s \times s$ matrix constructed with $f(t_i - t_j)$ be positive semi-definite. If function f is positive semi-definite, its Fourier transform, is non-negative

$X+N$, we obtain the condition

$$\mathbb{E}\left\{\left[X-\frac{\gamma}{\gamma+1}(X+N)\right](X+N)^m\right\}=0 \text{ for } m=1,..,M \tag{3.34}$$

Applying binomial expansion

$$(X+N)^m=\sum_{i=0}^{m}\binom{m}{i}X^iN^{m-i} \tag{3.35}$$

and rearranging the terms, we obtain $M+1$ linear equations that recursively relate the $M+1$ moments of X, i.e., for $m=1,...,M$ we have

$$\mathbb{E}(X^{m+1})=\gamma\mathbb{E}(N^{m+1})+\sum_{i=0}^{m-1}A(\gamma,m,i)\mathbb{E}(N^{i+1})\mathbb{E}(X^{m-i}) \tag{3.36}$$

where, $A(\gamma,m,i)=\gamma\binom{m}{i}-\binom{m}{i+1}$.

Note that if $F_N(\omega)$ is analytic, N has finite moments of all orders and $\mathbb{E}\{|N^k|\}\leq k!\beta^k$, $\forall k$. From (3.36), by induction, we can show that all moments of X exist and are bounded by $\mathbb{E}\{|X^k|\}\leq k!(\max\{\gamma,1\}\beta)^k$. This condition is sufficient to show that X also has an analytic characteristic function. □

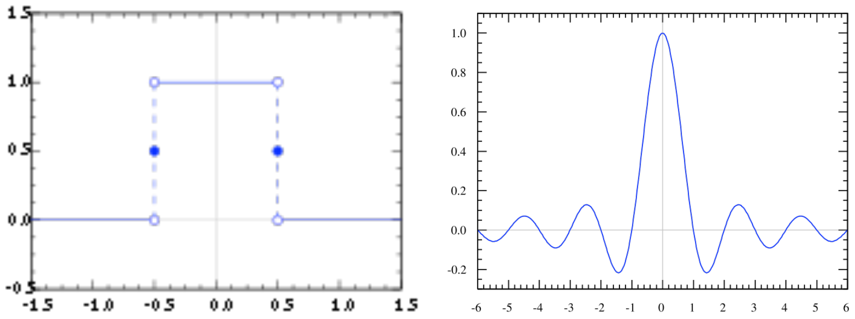
The following corollary identifies a case in which a matching source does not exist.

Corollary 3.7. *For $\gamma\notin\mathbb{Z}$, if $F_N(\omega)$ is real and analytic and it is negative somewhere, i.e., $\exists\omega$ such that $F_N(\omega)<0$, then a matching source distribution does not exist.*

Proof. We prove this corollary by contradiction. Let $F_N(\omega)$ be a valid characteristic function. Recall the set of moment equations (3.36). It follows by induction over the set of moment equations starting from $m=1$ that, if all odd moments of N are zero, then so are all odd moments of X. Note that X, if exists, has an analytic characteristic function due to Corollary 3.6. Hence, when the noise is symmetric, the matching source must also be symmetric since moments of X fully characterize its distribution due to the analyticity of the characteristic function $F_X(\omega)$.

everywhere $F(\omega)\geq 0,\forall\omega\in\mathbb{R}$. Hence, in the case of our candidate characteristic function, this requirement

However, if $\gamma \notin \mathbb{Z}$, by (3.31), it follows that $F_X(\omega)$ is not real, and hence $f_X(x)$ is not symmetric. This contradiction shows that no matching source exists when $\gamma \notin \mathbb{Z}$ and noise distribution is symmetric but not positive semi-definite. □



(a) Uniform density

(b) Characteristic function

Figure 3.2. Example of a non-existence case

Let us provide a commonly used example distribution to which the above corollary applies: uniform distribution over $[-a, a]$ as shown in Figure 3.2. In this case, $f_N(n)$ is symmetric with an analytic characteristic function, but it is not positive semi-definite. The corollary states that, except for integer values of SNR, the optimal estimator is strictly nonlinear for an additive uniform channel.

Remark: As an important application, consider high resolution quantization theory. Standard high resolution approximations assume quantization noise independent of (or uncorrelated with) the source [20]. In practice such approximations can be made explicit by using a dithered quantizer [29] that generates quantization error independent of the source. Then the quantizer is equivalent to an additive uniform noise channel. The corollary states that, other than for integer values of SNR, a linear decoder (e.g., a Wiener filter at the decoder) is strictly suboptimal for sources encoded with high resolution or dithered quantization.

3.3.4 Uniqueness of a Matching Source for a Given Noise

Note that (3.31) may have multiple solutions due to multiplicity of complex roots. The following corollary establishes that for a large set of source (or noise) distributions, the matching noise (or source) is unique.

ensures that the corresponding density is indeed non-negative everywhere.

Corollary 3.8. *If $F_N(\omega)$ (or $F_X(\omega)$) is analytic, then the matching $F_X(\omega)$ (or $F_N(\omega)$) is unique.*

Proof. We prove this corollary from the set of moment equations (3.36). Note that every equation introduces a new variable $\mathbb{E}(X^{m+1})$, for $m = 1, .., M$, so each new equation is linearly independent of its predecessors. Let us consider solving these equations recursively, starting from $m = 1$. At each m, we have one unknown ($\mathbb{E}(X^{m+1})$) in a “linear” equation. Since the number of equations is equal to the number of unknowns for each m, and the equations are linear in terms of the unknown, there must exist a unique moment sequence that solves (3.36). From Corollary 3.6, it also follows that X has an analytic characteristic function. Hence, the moment sequence fully characterizes X and the matching source X (if exists) is unique. □

3.4 Implications of the Linearity Conditions

In this section, we explore some special cases obtained by varying γ and utilizing the matching conditions for MSE and L_p. We start with a simple but perhaps surprising result.

Theorem 3.9. *Given a source and noise of equal variance, the L_p optimal estimator is linear* if and only if *the noise and source distributions are identical, i.e., $f_X(x) = f_N(x), \ \forall x \in \mathbb{R}$ and in which case, the optimal estimator is $h(Y) = \frac{1}{2}Y$.*

Proof. For MSE, it is straightforward to see from (3.31) that, at $\gamma = 1$, the characteristic functions must be identical. Since the characteristic function uniquely determines the distribution [10], $f_X(x) = f_N(x)$, $\forall x \in \mathbb{R}$. In fact, this results applies more generally. This can be observed directly from Theorem 3.2 that $F_N(\omega) = F_X(\omega)$ satisfies the necessary and sufficient optimality condition, and hence this result also applies to the L_p norm distortion measure. □

It is well known that linearity of regression $\mathbb{E}\{X|Y\}$ for all SNR levels characterizes the stable family of distributions, which includes Gaussian as a famous (and the only finite variance) member [60, 6, 64, 61, 32]. All prior results on characterizing Gaussian density with the linearity of regression consider linearity for optimal estimation for all SNR levels, $\gamma \in \mathbb{R}^+$.

Let us consider a setup with given source and noise variables which may be scaled to vary the SNR, γ. Can the optimal estimator be linear at multiple values of γ? This question is motivated by the practical setting where γ is not known in advance or may vary (e.g., in the design stage of a communication system). It is well-known that the Gaussian source-Gaussian noise pair makes the optimal estimator linear at all γ levels. Below, we show that this is the only source-channel pair whose optimal estimators are linear at more than one SNR value.

Theorem 3.10. *Let the source or channel variables be scaled to vary the SNR, γ. The MSE optimal estimator is linear at two different SNR values γ_1 and γ_2,* if and only if *source and noise are both Gaussian. Moreover, this claim also holds for L_p norm if the source (or noise) has an analytic characteristic function.*

Note: Theorem 3.10 provides a new characterization of a Gaussian process since all prior related results characterize Gaussian density as the one that ensures linearity for optimal estimation for all SNR $\gamma \in \mathbb{R}^+$, whereas our theorem requires linearity of optimal estimation at *only* two levels of SNR.

Proof. Note that $\sigma_{n_2}^2 = \alpha^2 \sigma_n^2$ and $F_{N_2}(\omega) = F_N(\omega\alpha)$. Let,

$$\gamma_1 = \frac{\sigma_x^2}{\sigma_n^2}, \ \gamma_2 = \frac{\sigma_x^2}{\alpha^2 \sigma_n^2} \tag{3.37}$$

Using (3.31),

$$F_X(\omega) = F_N^{\gamma_1}(\omega), F_X(\omega) = F_N^{\gamma_2}(\omega\alpha) \tag{3.38}$$

Hence,

$$F_N^{\gamma_1}(\omega) = F_N^{\gamma_2}(\omega\alpha) \tag{3.39}$$

Taking the logarithm on both sides of (3.39), applying (3.37) and rearranging terms, we obtain

$$\alpha^2 = \frac{\log F_N(\alpha\omega)}{\log F_N(\omega)} \tag{3.40}$$

Note that (3.40) should be satisfied for both α and $-\alpha$ since they yield the same γ. Plugging $\alpha = -1$ in (3.40), we obtain $F_N(\omega) = F_N(-\omega)$, $\forall \omega$. Using the fact that the characteristic function is conjugate symmetric (i.e., $F_N(-\omega) = F_N^*(\omega)$), we get $F_N(\omega) \in \mathbb{R}$, $\forall \omega$. As $\log F_N(\omega)$ is $\mathbb{R} \to \mathbb{C}$, the Weierstrass theorem [14] guarantees that there is a sequence of polynomials that uniformly converges to it: $\log F_N(\omega) = \sum_{i=0}^{\infty} k_i \omega^i$, where $k_i \in \mathbb{C}$. Hence, by (3.40) we obtain:

$$\alpha^2 = \frac{\sum\limits_{i=0}^{\infty} k_i (\omega\alpha)^i}{\sum\limits_{i=0}^{\infty} k_i \omega^i}, \quad \forall \omega \in \mathbb{R}, \tag{3.41}$$

which is satisfied for all ω only if all coefficients k_i vanish, except for k_2, i.e. $\log F_N(\omega) = k_2\omega^2$, or $\log F_N(\omega) = 0 \quad \forall \omega \in \mathbb{R}$ (the solution $\alpha = 1$ is of no interest). The latter is not a characteristic function, and the former is the Gaussian characteristic function, $F_N(\omega) = e^{k_2\omega^2}$, where we use the established fact that $F_N(\omega) \in \mathbb{R}$. Since a characteristic function determines the distribution uniquely, the Gaussian source and noise must be the only such pair.

Next, we extend the result to the L_p norm, albeit we require analyticity of the characteristic function of X (or N). Then, due to Corollary 3.6, N also has an analytic characteristic function. We note that the moments of X and N are finite (they have moments of all orders) and moments fully characterize the distribution due to the analyticity of characteristic function. The extension to L_p requires a different approach. For simplicity, we first derive the result for MSE (now with analyticity imposed) and then extend the arguments to the L_p case. Let us say the noise is scaled by $\alpha \in \mathbb{R}$, i.e $N_2 = \alpha N$. The following relation between the moments of the original and scaled noise should be satisfied:

$$\mathbb{E}(N_2^m) = \alpha^m \mathbb{E}(N^m) \text{ for } m = 1, .., M+1 \tag{3.42}$$

Also, a set of moment equations should hold for two SNR values, γ_1 and γ_2. Let us consider the set of moment equations with moments up to M:

$$\mathbb{E}(X^{m+1}) = \gamma_j \mathbb{E}(N^{m+1}) + \sum_{i=0}^{m-1} A(\gamma_j, m, i)\mathbb{E}(N^{i+1})\mathbb{E}(X^{m-i}) \tag{3.43}$$

where $m = 1, .., M$, $j = 1, 2$ and $A(\gamma, m, i) = \gamma\binom{m}{i} - \binom{m}{i+1}$. Similar to the proof of Corollary 3.8, we note that every equation introduces a new variable $\mathbb{E}(X^{m+1})$, for $m = 1, .., M$, so each new equation is independent of its predecessors. Next we solve these equations recursively, starting from $m = 1$. At each m, we have three unknowns $(\mathbb{E}(X^{m+1}), \mathbb{E}(N^{m+1}), \mathbb{E}(N_2^{m+1}))$ that are related "linearly". Since the number of linearly independent equations is equal to the number of unknowns for each m, there must exist a unique solution. We know that the moment sequences of the Gaussian source-channel pair satisfy (3.43) since it ensures linearity of optimal estimation. The moment sequence of a Gaussian satisfies Carlemans general criterion [69] and therefore it uniquely determines the corresponding distribution, so the Gaussian source and noise pair is the only solution to (3.43).

The proof for L_p norm follows the same lines. Note that as mentioned in Sec II.D, the same linear estimator is L_p optimal for a Gaussian source-channel pair. Plugging $Y = X + N$ in the optimality condition with L_p norm, (3.7), we reach a similar set of moment equations. Following similar arguments, we can show that this result holds for the L_p norm. □

Next, we investigate the asymptotic behavior of optimal estimation at low and high SNR. The results of our asymptotic analysis are of practical importance since they justify the use of linear estimators without recourse to complexity arguments at high and low asymptotic SNR regimes, under certain conditions.

Theorem 3.11 (for MSE only)**.** *In the limit $\gamma \to 0$, the MSE optimal estimator is asymptotically linear if the channel is Gaussian, regardless of the source. Similarly, as $\gamma \to \infty$, the MSE optimal estimator is asymptotically linear if the source is Gaussian, regardless of the channel.*

Proof. The proof follows from applying the central limit theorem [10] to the matching condition (3.31). The central limit theorem states that as $\gamma \to \infty$, for any finite variance noise N, the characteristic function of the matching source $F_N^{\gamma}(\omega)$ uniformly converges to the Gaussian characteristic function. The continuity theorem guarantees that as $F_N^{\gamma}(\omega)$ uniformly converges to the Gaussian characteristic function, the corresponding density converges (in distribution) to the Gaussian density. Hence, at asymptotically high SNR, any noise distribution is matched by the Gaussian source.

Similarly, as $\gamma \to 0$ and for any $F_X(\omega)$, $F_X^{\frac{1}{\gamma}}(\omega)$ converges to the Gaussian characteristic function and hence the MSE optimal estimator is asymptotically linear if the channel is Gaussian. □

3.5 Extension to Vector Spaces

Extension of the conditions to the vector case is nontrivial due to the dependencies across components of the source and noise. In this section, for simplicity, we restrict ourselves to the MSE distortion measure. We first give the formal definition of the problem:

We consider the problem of estimating the vector source $\mathbf{X} \in \mathbb{R}^m$ given the observation $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{X}$ and $\mathbf{N} \in \mathbb{R}^m$ are independent, as shown in Figure 3.1. Without loss of generality, we assume that $\mathbf{X}$ and $\mathbf{N}$ are zero mean random variables with m-fold distributions $f_X(\cdot)$ and $f_N(\cdot)$. Their respective characteristic functions are denoted $F_X(\boldsymbol{\omega})$ and $F_N(\boldsymbol{\omega})$. $\mathbf{R_X} = \mathbb{E}\{\mathbf{XX^T}\}$, $\mathbf{R_N} = \mathbb{E}\{\mathbf{NN^T}\}$ denote the covariance matrices of $\mathbf{X}$ and $\mathbf{N}$, respectively. Let $\mathbf{U}$ be the eigenmatrix of $\mathbf{R}_X\mathbf{R}_N^{-1}$, and let eigenvalues $\lambda_1, ..., \lambda_m$ be the elements of the diagonal matrix $\mathbf{\Lambda}$, i.e., the following holds:

$$\mathbf{R}_X\mathbf{R}_N^{-1} = \mathbf{U\Lambda U}^{-1} \tag{3.44}$$

We are looking for the conditions on $F_X(\boldsymbol{\omega})$ and $F_N(\boldsymbol{\omega})$ such that $\mathbf{h}(\mathbf{Y}) = \mathbf{KY}$ with $\mathbf{K} = \mathbf{R_X}(\mathbf{R_X} + \mathbf{R_N})^{\mathbf{-1}}$ minimizes the estimation error $\mathbb{E}\{||\mathbf{X} - \mathbf{h}(\mathbf{Y})||_2^2\}$. Let us re-write the MSE optimal estimator for the vector case:

$$h(\mathbf{y}) = \frac{\int \mathbf{x} f_X(\mathbf{x}) f_N(\mathbf{y} - \mathbf{x})\, \mathbf{dx}}{\int f_X(\mathbf{x}) f_N(\mathbf{y} - \mathbf{x})\, \mathbf{dx}} \tag{3.45}$$

Plugging, $h(\mathbf{y}) = \mathbf{Ky}$, we obtain,

$$\mathbf{Ky} \int f_X(\mathbf{x}) f_N(\mathbf{y} - \mathbf{x})\, \mathbf{dx} = \int \mathbf{x} f_X(\mathbf{x}) f_N(\mathbf{y} - \mathbf{x})\, \mathbf{dx} \tag{3.46}$$

Expressing the integrals as m-fold convolutions, we get

$$\mathbf{Ky}\left[f_X(\mathbf{y}) \otimes f_N(\mathbf{y})\right] = \left[\mathbf{y} f_X(\mathbf{y})\right] \otimes f_N(\mathbf{y}) \tag{3.47}$$

Taking Fourier transform of both sides,

$$j\mathbf{K}\nabla\left[F_X(\boldsymbol{\omega})F_N(\boldsymbol{\omega})\right] = jF_N(\boldsymbol{\omega})\nabla F_X(\boldsymbol{\omega}) \tag{3.48}$$

Arranging the terms, we get

$$(\mathbf{I} - \mathbf{K})\,\frac{1}{F_X(\boldsymbol{\omega})}\nabla F_X(\boldsymbol{\omega}) = \mathbf{K}\frac{1}{F_N(\boldsymbol{\omega})}\nabla F_N(\boldsymbol{\omega}) \tag{3.49}$$

Using $\nabla \log F_X(\boldsymbol{\omega}) = \frac{1}{F_X(\boldsymbol{\omega})}\nabla F_X(\boldsymbol{\omega})$,

$$\nabla \log F_X(\boldsymbol{\omega}) = (\mathbf{I} - \mathbf{K})^{\mathbf{-1}}\mathbf{K}\nabla \log \mathbf{F_N}(\boldsymbol{\omega}) \tag{3.50}$$

Plugging the value of $\mathbf{K} = \mathbf{R_X}(\mathbf{R_X} + \mathbf{R_N})^{\mathbf{-1}}$ in $(\mathbf{I} - \mathbf{K})^{\mathbf{-1}}\mathbf{K}$ we obtain,

$$\nabla \log F_X(\boldsymbol{\omega}) = \mathbf{R_X}\mathbf{R_N}^{-1}\nabla \log F_N(\boldsymbol{\omega}) \tag{3.51}$$

Using the eigen decomposition of $\mathbf{R_X}\mathbf{R_N}^{\mathbf{-1}} = \mathbf{U\Lambda U^{-1}}$ where $\mathbf{\Lambda}$ is diagonal with eigen values $\lambda_1, ..., \lambda_n$, we obtain

$$\mathbf{U^{-1}}\nabla \log F_X(\boldsymbol{\omega}) = \Lambda\mathbf{U^{-1}}\nabla \log F_N(\boldsymbol{\omega}) \tag{3.52}$$

We will make use of the following auxiliary lemma in matrix analysis.

Lemma 3.12. *Given a function $f : \mathbb{R}^n \to \mathbb{R}$, matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ and vector $\mathbf{x} \in \mathbb{R}^m$*

$$\nabla_x f(\mathbf{Ax}) = \mathbf{A}^T\nabla f(\mathbf{Ax}) \tag{3.53}$$

Proof. By the chain rule we have,

$$\begin{aligned}
\frac{\partial f(\mathbf{Ax})}{\partial x_i} &= \sum_{k=1}^{n} \frac{\partial f(\mathbf{Ax})}{\partial(\mathbf{Ax})_k}\frac{\partial(\mathbf{Ax})_k}{\partial x_i} && (3.54)\\
&= \sum_{k=1}^{n} \frac{\partial f(\mathbf{Ax})}{\partial(\mathbf{Ax})_k}\frac{\partial(a_k^T x)}{\partial x_i} && (3.55)\\
&= \sum_{k=1}^{n} \frac{\partial f(\mathbf{Ax})}{\partial(\mathbf{Ax})_k}a_{ki} && (3.56)\\
&= \sum_{k=1}^{n} \partial_k f(\mathbf{Ax})a_{ki} && (3.57)\\
&= a_i^T\nabla f(\mathbf{Ax}) && (3.58)
\end{aligned}$$

It follows that $\nabla_x f(\mathbf{Ax}) = \mathbf{A}^T\nabla f(\mathbf{Ax})$. □

Next, we state the main theorem.

Theorem 3.13. *Let the characteristic functions of the transformed source and noise (*$\mathbf{UX}$ *and* $\mathbf{UN}$*) be $F_{UX}(\boldsymbol{\omega})$ and $F_{UN}(\boldsymbol{\omega})$. The necessary and sufficient condition for linearity of optimal estimation is:*

$$\frac{\partial \log F_{UX}(\boldsymbol{\omega})}{\partial\omega_i} = \lambda_i\frac{\partial \log F_{UN}(\boldsymbol{\omega})}{\partial\omega_i}, 1 \le i \le m \tag{3.59}$$

Proof. Using Lemma 3.12, we can rewrite (3.52) as

$$\nabla_{\boldsymbol{\omega}} \log F_X(\mathbf{U^{-1}}\boldsymbol{\omega}) = \Lambda\nabla_{\boldsymbol{\omega}} \log F_N(\mathbf{U^{-1}}\boldsymbol{\omega}) \tag{3.60}$$

Note that the characteristic functions of the source and noise after transformation can be written in terms of the known characteristic functions $F_X(\boldsymbol{\omega})$ and $F_N(\boldsymbol{\omega})$, specifically $F_{UX}(\boldsymbol{\omega}) = \det(\mathbf{U})F_X(\mathbf{U^{-1}}\boldsymbol{\omega})$ and $F_{UN}(\boldsymbol{\omega}) = \det(\mathbf{U})F_N(\mathbf{U^{-1}}\boldsymbol{\omega})$, where $\det(\cdot)$ denotes the determinant. The necessary and sufficient condition of (3.60) can thus be converted to the set of m scalar differential equations of (3.59).

□

Further insight into the above necessary and sufficient condition is provided via the following corollaries.

Corollary 3.14. *Let $F_{UX_i}(\omega)$ and $F_{UN_i}(\omega)$ be the marginal characteristic functions of the transform coefficients $[\mathbf{UX}]_i, [\mathbf{UN}]_i$. A necessary condition for linearity of optimal estimation is:*

$$F_{UX_i}(\omega) = F_{UN_i}^{\lambda_i}(\omega), 1 \leq i \leq m \tag{3.61}$$

Proof. Integrating both sides of (3.59) over all ω_j, $j \neq i$, yields the following set of differential equations

$$\frac{\partial \log F_{UX_i}(\omega)}{\partial \omega} = \lambda_i \frac{\partial \log F_{UN_i}(\omega)}{\partial \omega},\ 1 \leq i \leq m \tag{3.62}$$

which, given the boundary conditions $F_{UX_i}(0) = F_{UN_i}(0) = 1$, leads to the solution specified in (3.61) as an explicit matching condition. □

Corollary 3.15. *A necessary condition for linearity of optimal estimation is that one of the following holds for every pair i, j, $1 \leq i, j \leq m$:*

- *i)* $\lambda_i = \lambda_j$
- *ii)* $[\mathbf{UX}]_\mathbf{i}$ *is independent of* $[\mathbf{UX}]_\mathbf{j}$ *and* $[\mathbf{UN}]_\mathbf{i}$ *is independent of* $[\mathbf{UN}]_\mathbf{j}$.

Proof. Let us rewrite (3.59) explicitly for the i^{th} and j^{th} coefficients.

$$\frac{\partial \log F_{UX}(\boldsymbol{\omega})}{\partial \omega_i} = \lambda_i \frac{\partial \log F_{UN}(\boldsymbol{\omega})}{\partial \omega_i} \tag{3.63}$$

$$\frac{\partial \log F_{UX}(\boldsymbol{\omega})}{\partial \omega_j} = \lambda_j \frac{\partial \log F_{UN}(\boldsymbol{\omega})}{\partial \omega_j} \tag{3.64}$$

The partial derivative of both sides of (3.63) with respect to ω_j and both sides of (3.64) with respect to ω_i, to obtain the following:

$$\frac{\partial^2 \log F_{UX}(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j} = \lambda_i \frac{\partial^2 \log F_{UN}(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j} \tag{3.65}$$

$$\frac{\partial^2 \log F_{UX}(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j} = \lambda_j \frac{\partial^2 \log F_{UN}(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j} \tag{3.66}$$

There are only two ways to simultaneously satisfy (3.65) and (3.66): i) $\lambda_i = \lambda_j$ ii) the second order derivatives vanish, i.e., $\frac{\partial^2 \log F_{UX}(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j} = \frac{\partial^2 \log F_{UN}(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j} = 0$ which means independence of the i^{th} and j^{th} transform coefficients of source X and similarly of noise N. □

Corollary 3.16. *If the necessary condition of Corollary 3.14 is satisfied, then a sufficient condition for linearity of optimal estimation is that* $\mathbf{U}$ *generates independent coefficients for both* X *and* N*.*

Proof. Independence of the transform coefficients implies that the joint characteristic function is the product of the marginals:

$$F_{UX}(\boldsymbol{\omega}) = \prod_{i=1}^{m} F_{UX_i}(w_i), \;\; F_{UN}(\boldsymbol{\omega}) = \prod_{i=1}^{m} F_{UN_i}(w_i) \tag{3.67}$$

Plugging (3.67) into the necessary and sufficient condition (3.59) of Theorem 3.13, it is straightforward to show that (3.61), the necessary condition of Corollary 3.14, is now both necessary and sufficient. □

While the condition in Corollary 3.16 involves independence of transform coefficients, the weaker property of uncorrelatedness is already guaranteed by transform $\mathbf{U}$. The matrix $\mathbf{U}$ diagonalizes both $\mathbf{R}_X$ and $\mathbf{R}_N$. We formalize this in the following lemma:

Lemma 3.17. *Transform* $\mathbf{U}$ *decorrelates both source and noise: both* $\mathbf{U}\mathbf{R}_X\mathbf{U}^T$ *and* $\mathbf{U}\mathbf{R}_N\mathbf{U}^T$ *are diagonal matrices.*

Proof. Since both $\mathbf{R}_X$ and $\mathbf{R}_N$ are, by definition, positive definite matrices, there exists a matrix $\mathbf{S}$ that simultaneously diagonalizes $\mathbf{R}_X$ and whitens $\mathbf{R}_N$, i.e., $\mathbf{S}\mathbf{R}_X\mathbf{S}^T = \boldsymbol{\Lambda}_X$ and $\mathbf{S}\mathbf{R}_N\mathbf{S}^T = \mathbf{I}$

where $\mathbf{\Lambda}_X$ is diagonal and $\mathbf{I}$ is the identity matrix [36]. Hence, $\mathbf{R}_X$ and $\mathbf{R}_N$ can be expressed as the following:

$$\mathbf{R}_X = \mathbf{S}^{-1}\Lambda_X\mathbf{S}^{-T},\ \mathbf{R}_N = \mathbf{S}^{-1}\mathbf{S}^{-T} \tag{3.68}$$

Plugging the above into (3.44) we obtain $\mathbf{U} = \mathbf{\Lambda}_U\mathbf{S}$, where $\mathbf{\Lambda}_U$ is diagonal. Substituting $\mathbf{U}$ in $\mathbf{U}\mathbf{R}_X\mathbf{U}^T$ and $\mathbf{U}\mathbf{R}_N\mathbf{U}^T$, we obtain:

$$\mathbf{U}\mathbf{R}_X\mathbf{U}^T = \mathbf{\Lambda}_U\mathbf{\Lambda}_X\mathbf{\Lambda}_U^T,\ \mathbf{U}\mathbf{R}_N\mathbf{U}^T = \mathbf{\Lambda}_U\mathbf{\Lambda}_U^T \tag{3.69}$$

The product of diagonal matrices is also diagonal. □

As an example where the optimal estimator is known to be linear, consider the Gaussian multivariate case. Note that the Gaussian source-channel pair satisfies the scalar matching condition for any SNR, it satisfies (3.61). As a linear transform preserves joint Gaussianity in the transform domain, $\mathbf{U}$ generates jointly Gaussian and uncorrelated coefficients which are therefore independent, satisfying the conditions of Corollary 3.16.

An important observation is that the necessary and sufficient condition for scalars (3.31) is also a necessary condition for vectors (3.61), in the transform domain. Due to this fact, it is straightforward to extend the existence and uniqueness results and implications of the scalar matching conditions to the vector spaces. These trivial extensions are omitted here for conciseness.

Chapter 4

Randomized Quantization

4.1 Introduction

A central objective of dithered quantization is to render the quantization error white and independent from the source, which can be achieved if certain conditions, determined by Schuchman, are met [65]. Traditionally, dithered quantization has been studied in the framework where the quantizer is uniform (with step size Δ) and the dither signal is uniformly distributed over $(-\frac{\Delta}{2}, \frac{\Delta}{2})$, matched to the quantizer interval as shown in Figure 4.1. A uniformly distributed dither signal is added before quantization and the same dither signal is subtracted from the quantized value at the decoder side. Note that only subtractive dithering is considered in this part. The quantized values are entropy coded, conditioned on the dither signal in the variable rate case. Randomized (dithered) quantizers have been studied in the past due to important properties that differentiate them from deterministic quantizers, which were consequently explored to find rate-distortion bounds for universal compression [84, 30]. Later, Zamir and Feder extensively studied the properties of dithered quantizers [80, 82, 81].

Randomized quantization is also of practical interest, beyond its theoretical significance. Many filter/system optimization problems in practical compression systems such as the rate-distortion optimal filterbank design problem [54], or low rate filter optimization for DPCM compression of Gaussian auto-regressive processes [31], assume quantization noise that is independent of (or uncorrrelated with) the source and white. Although these assumptions are satisfied at asymptotically high rates [20], such systems are mostly useful for very low rate applications. For example, in [31], it is stated that the assumptions made in the paper are

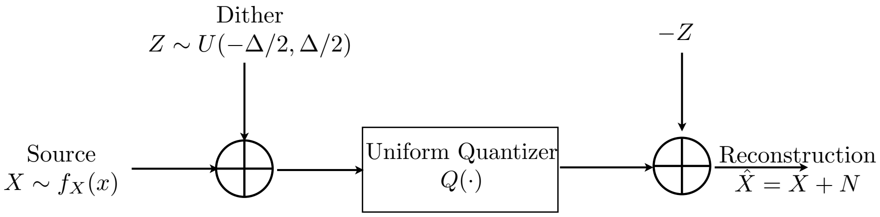


Figure 4.1. The basic structure of dithered quantization

not satisfied by deterministic quantizers and also it is shown that dithered quantizers satisfy the assumptions exactly. However, conventional dithered quantization suffers from suboptimal compression performance. Hence a quantizer that mostly satisfies the assumptions at relatively low performance degradation will be of use in many such applications.

In this part, we consider a generalization to enable effective dithering of nonuniform quantizers. To the best of our knowledge, this part of the thesis is the first attempt (other than our preliminary work in [1]) to consider dithered quantization in nonuniform quantization framework. One immediate problem with nonuniform dithered quantization is how to apply dithering to unequal quantization intervals. In traditional dithered quantization, the dither signal is matched to the uniform quantization interval, but it is not clear how to match generic dither to varying quantization intervals. As a remedy to this problem, we propose dithering in the companded domain. The conventional (uniform) quantizer is obviously a special case of the nonuniform quantizer. The proposed dithered nonuniform quantizer is expected to outperform the conventional dithered quantizer most significantly at low rates where the optimal variable rate (entropy coded) quantizer is not uniform.

We also present an alternative deterministic quantizer that provides quantization noise uncorrelated with the source. A deterministic quantizer cannot render the quantization noise independent of the source or white but it can make it uncorrelated with the source. We present the optimality conditions of this deterministic quantizer, for both fixed and variable rate quantization, and compare its rate-distortion performance to the randomized quantizers.

Dithered quantization offers an interesting theoretical twist. Randomized quantization is an instance of the random encoding used to prove the achievability of the coding bounds in rate distortion theory [13]. However, to achieve those bounds, a random encoding scheme is not necessary, as one can achieve the bounds using a deterministic quantizer with asymptotically high dimension. In the second part of this section, we investigate the settings under which randomized quantization is asymptotically necessary. A trivial example involves requiring source-independent quantization error. It is well known that the reconstruction (hence quantization error) is a deterministic function of the source when the quantizer is deterministic [20],

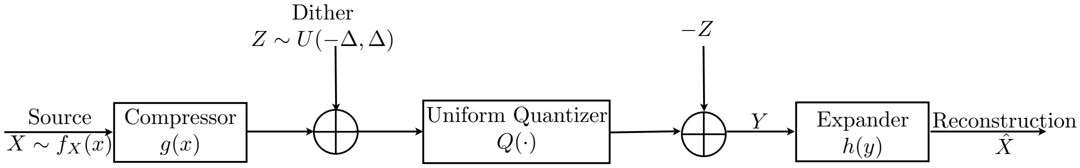


Figure 4.2. The proposed nonuniform dithered quantizer

while conventional dithered quantization produces quantization error that is independent of the source. Although a deterministic quantizer can never render the quantization error independent of the source, as we show in this part, a deterministic quantizer can produce quantization error uncorrelated with the source. A natural question is whether the rate distortion bound, subject to uncorrelated error constraint, can be achieved (asymptotically) with a deterministic quantizer.

4.2 Preliminaries

The entropy, in bits, of a discrete random variable X taking values in X is

$$H(X) = -\sum_{x \in X} P(X = x) \log P(X = x) \tag{4.1}$$

where logarithm is base 2. The differential entropy of a continuous random variable X with probability density function $f_X(x)$ is

$$h(X) = -\int f_X(x) \log f_x(x) dx \tag{4.2}$$

The divergence between two continuous distributions f_X and g_X, is given by

$$\mathcal{D}(f_X || g_X) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)} dx \tag{4.3}$$

The uniform quantizer with dithering is defined as follows. The uniform quantizer, with reconstructions $\{0, \pm\Delta, \pm 2\Delta, ..., \pm K\Delta\}$, is a mapping $Q : \mathbb{R} \to \mathbb{R}$ such that

$$Q(x) = i\Delta \quad \text{for} \quad i\Delta + \Delta/2 > x \geq i\Delta - \Delta/2 \tag{4.4}$$

In fixed rate quantization, K is determined by the rate as

$$R_f = \log(2K + 1) \tag{4.5}$$

For variable rate quantization $K \to \infty$. In this case, uniform quantizer is followed by a lossless source encoder (entropy coder). Assuming an optimal entropy coder, the rate is estimated as the entropy of the quantized source. Let dither Z be a random variable, distributed uniformly on the interval $(-\Delta/2, \Delta/2)$. Then conventional dithered quantizer approximates the source value x by

$$\hat{x} = Q(x + Z) - Z \tag{4.6}$$

It can be shown that that the mean square error of this quantizer is independent of the source value of x:

$$\mathbb{E}_z\{(Q(x + Z) - Z - x)^2\} = \Delta^2/12, \forall x \tag{4.7}$$

Note that with any deterministic quantizer, the error is completely determined by the source value [20].

The realization of the dither random variable Z is available to both the encoder and the decoder. Thus, the rate of this quantizer is the conditional entropy of the reconstruction given the dither, i.e.,

$$R_v = H(Q(X + Z) - Z|Z) = H(Q(X + Z)|Z) \tag{4.8}$$

In [80], it is shown that the following holds:

$$H(Q(X + Z)|Z) = h(Y) - \log \Delta \tag{4.9}$$

where $Y = X + N$ and N is independent of X and uniformly distributed over $(-\Delta/2, \Delta/2)$.

4.3 Nonuniform Dithered Quantizer

The main idea is to perform uniform dithered quantization in the companded domain (see Figure 4.2). The source X is transformed through compressor $g(\cdot)$ before dithered uniform quantization. At the decoder side, dither is subtracted to obtain $Y = g(X) + N$, where N is uniformly distributed over $(-\Delta, \Delta)$ and independent of the source. The reconstruction is obtained by applying the expander $\hat{X} = h(Y)$. The objective is finding the optimal compressor and expander mappings $g(x), h(y)$ that minimize the expected distortion under the rate constraint. The MSE distortion can be written as:

$$D = \int\int [x - h(g(x) + n)]^2 f_X(x) f_N(n) dx dn \tag{4.10}$$

where $f_N(n)$ is uniform over $(-\Delta, \Delta)$. Note that this problem is related to the joint source channel mapping problem where the optimal analog encoding and decoding mappings are studied [2]. In our setting, the reconstruction error is analogous to the channel noise and the rate constraint plays a role similar to that of the power constraint. Similar to [2], we develop an iterative procedure that enforces the necessary conditions for optimality of the mappings. Note that the conventional dithered quantizer is a special case employing the trivial identity mappings, i.e., $g(x)=h(x)=x, \forall x$.

4.3.1 Optimal Expander

The conditional expectation $h(y) = \mathbb{E}\{X|y\}$ minimizes MSE between the source and the estimate. $\mathbb{E}\{X|y\}$ has slightly different expressions for fixed and variable rates. Using Bayes rule, the optimal expander h can be written as

$$h(y) = \frac{\int\limits_{\gamma_-}^{\gamma_+} x f_X(x) f_N(y - g(x))dx}{\int\limits_{\gamma_-}^{\gamma_+} f_X(x) f_N(y - g(x))dx} \tag{4.11}$$

where for fixed rate $\gamma_+ = g^{-1}(\Delta K)$ and $\gamma_- = g^{-1}(-\Delta K)$ while for variable rate $\gamma_+ \to \infty$ and $\gamma_- \to -\infty$.

4.3.2 Optimal Compressor

Unlike the expander, the optimal compressor cannot be written in closed form. However, a necessary optimality condition can be obtained by setting the functional derivative of the cost functional to zero. Thus, for the optimal g and h, the functional derivative of the total cost, J, along the direction of any variation function $\eta(x)$ vanishes [51], i.e.,

$$\nabla J = \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} J\left[g(x) + \epsilon\eta(x)\right] = 0, \, \forall x \in \mathbb{R} \tag{4.12}$$

Fixed rate

Distortion expression can be written as in (4.13) where $\gamma_+ = g^{-1}(\Delta K)$ and $\gamma_- = g^{-1}(-\Delta K)$. The first term in distortion expression is the granular distortion and remaining terms are overload distortion. Note that we need the overload term distortion terms here, because without the exact expression, the $g(x)$ will grow unboundedly. The second and third terms in (4.13) prevent this. Since the rate is fixed, total cost is identical to the distortion for fixed rate case, i.e., $J = D$.

$$J_f = \frac{1}{\Delta}\{ \int\limits_{-\Delta/2}^{\Delta/2} (\int\limits_{\gamma_-}^{\gamma_+} [x - h(g(x)+n)]^2 f_X(x)dx + \int\limits_{-\infty}^{\gamma_-} [x - h(-K\Delta + n)]^2 f_X(x)dx$$
$$+ \int\limits_{\gamma_+}^{\infty} [x - h(K\Delta + n)]^2 f_X(x)dx)dn\} \tag{4.13}$$

Variable rate

To use (4.9) to find the rate, we need the distribution of $Y = g(X) + N$, which can be written as

$$f_Y(y) = \frac{1}{\Delta}\left[F_X(g^{-1}(y + \Delta/2)) - F_X(g^{-1}(y - \Delta/2))\right] \tag{4.14}$$

where $F_X(x)$ is the cumulative distribution function of X, i.e., $F_X(x) = \int\limits_{-\infty}^{x} f_X(x)dx$. The rate can be evaluated as

$$R_v = h(Y) - \log \Delta \tag{4.15}$$

Total cost for variable rate quantization is

$$J_v = D + \lambda R \tag{4.16}$$

where λ is the Lagrangian parameter that is chosen for each rate.

4.3.3 Constrained Randomized Quantizer

Due to companding, the nonuniform randomized quantizer described above does not generate quantization error uncorrelated with the source although it is based on (conventional) dithered quantizer which guarantees quantization error independent of the source. We therefore explicitly constrain the randomized quantizer to generate uncorrelated quantization error, by adding a penalty term to the total cost function. The Lagrangian parameter $\lambda_c \leq 0$ is set to ensure $\mathbb{E}(xh(g(x) + n)) = \mathbb{E}(x^2)$.

$$J_c = J + \lambda_c \mathbb{E}\{x(h(g(x) + n))\} \tag{4.17}$$

where $J = J_v$ in the case of variable rate and $J = J_f$ for fixed rate. We find the necessary conditions for optimality for fixed and variable rate randomized quantization by setting the functional derivative to zero for both compressor and expander. Surprisingly, the optimal compressor mapping remains unchanged and the optimal expander mapping is only scaled. We state this result in the following theorem.

Theorem 4.1. *Let g and h be the optimal compressor and expander mappings of the unconstrained optimal randomized quantizer. Let g_c and h_c denote the mappings of the optimal constrained randomized quantizer. Then,*

$$g_c(x) = g(x), h_c(y) = (1 + \lambda_c)h(y) \tag{4.18}$$

Proof. Let us focus on fixed rate, the variable rate case is shown very similarly. The optimal expander is no longer the standard conditional expectation, since it is impacted by the constraint. By setting $\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} J_c\left[h(x) + \epsilon\eta(x)\right] = 0$, we obtain the optimal expander in closed form as $h_c(y) = (1 + \lambda_c)h(y)$. The update rule for $g_c(x)$ can be derived in a similar manner to the unconstrained $g(x)$, i.e., setting $\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0} J_c\left[g(x) + \epsilon\eta(x)\right] = 0$ and plugging $h_c(y) = (1 + \lambda_c)h(y)$ yields after straightforward algebra $g_c(x) = g(x)$. □

4.3.4 Algorithm Design

The basic idea is to iteratively alternate between enforcing the necessary conditions for optimality, thereby successively decreasing the total cost. Iterations are performed until the algorithm reaches a stationary point. Solving for the optimal expander is straightforward since the expander is expressed as closed form functional of known quantities, $g(x)$, $f_X(x)$. Since the compressor condition is not in closed form, we perform steepest descent, i.e., move in the direction of the functional derivative of the total cost with respect to the compressor mapping g. By design, the total cost decreases monotonically as the algorithm proceeds iteratively. The compressor mapping is updated according to (4.19), where i is the iteration index, $\nabla J[g]$ is the directional derivative and μ is the step size.

$$g_{i+1}(x) = g_i(x) - \mu\nabla J[g] \tag{4.19}$$

Note that there is no guarantee that an iterative descent algorithms of this type will converge to the globally optimal solution. The algorithm will converge to a local minimum and hence, initial conditions have paramount importance in such greedy optimizations. A preliminary low complexity approach to mitigate the poor local minima problem, is to embed in the solution the noisy relaxation method of [18, 45]. We initialize the compressor mapping with random initial conditions and run the algorithm for a very noisy channel (high Lagrangian parameter λ). Then, we gradually increase rate (decrease λ) while tracking the minimum.

4.4 Deterministic Uncorrelated Quantizer

A deterministic quantizer cannot yield quantization noise independent of the source [20]. However, it is possible to render the quantization noise uncorrelated with the source. A prior work along this line exists in [35], where a uniform quantizer is converted to a quantizer whose quantization noise is uncorrelated with the source. In this section, we derive the optimal deterministic quantizer which is constrained to give quantization error uncorrelated with the source.

Theorem 4.2. *Let r_i be the reconstruction point and x_{i-1} and x_i be the decision boundaries of the i^{th} quantization interval, for the M point optimal quantizer whose quantization error is uncorrelated with the source. Similarly, let $\hat{r}_i$ be the reconstruction point and $\hat{x}_{i-1}$ and $\hat{x}_i$ be the decision boundaries of the i^{th} quantization interval, for of optimal (unconstrained) quantizer. Then, for $i = 1, 2, ..M$:*

$$x_i = \hat{x}_i \ \text{ and } \ r_i = \sigma_x^2 \frac{\hat{r}_i}{\sum\limits_{i=1}^{M} p_i \hat{r}_i^2} = s\hat{r}_i$$

where $p_i = \int\limits_{x_{i-1}}^{x_i} f(x)dx$. This result holds for both variable and fixed rate quantization.

Proof. We start with fixed rate quantization, where distortion is:

$$D = \sum_{i=1}^{M} \int\limits_{x_{i-1}}^{x_i} (x - r_i)^2 f(x)\, dx \tag{4.20}$$

The “uncorrelatedness” constraint can be written as :

$$\mathbb{E}[x(x - r_i)|x \in \mathcal{R}_i] = 0 \tag{4.21}$$

where $\mathcal{R}_i$ is the region where the reconstruction is r_i. This is equivalent to:

$$\mathbb{E}[x^2] = \sum_{i=1}^{M} r_i \int\limits_{x_{i-1}}^{x_i} x f(x)\, dx \tag{4.22}$$

We consider the Lagrangian cost

$$J = \sum_{i=1}^{M} \int\limits_{x_{i-1}}^{x_i} (x - r_i)^2 f(x)\, dx + \gamma \left[\sigma_x^2 - r_i \int\limits_{x_{i-1}}^{x_i} x f(x)\, dx\right] \tag{4.23}$$

By setting $\frac{\partial J}{\partial r_i} = 0$, we obtain:

$$r_i = \sigma_x^2 \frac{\hat{r}_i}{\sum\limits_{i=1}^{M} p_i \hat{r}_i^2} = s\hat{r}_i \tag{4.24}$$

where $\hat{r}_i = \frac{\int\limits_{x_{i-1}}^{x_i} x f(x) dx}{\int\limits_{x_{i-1}}^{x_i} f(x) dx}$ and s is a scaling constant. Setting $\frac{\partial J}{\partial x_i} = 0$ we obtain the update rule for x_i as

$$x_i = \frac{r_i + r_{i-1}}{2s} = \frac{\hat{r}_i + \hat{r}_{i-1}}{2} = \hat{x}_i \tag{4.25}$$

For variable rate, the total cost function is:

$$J = \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} (x - r_i)^2 f(x)\, dx + \gamma(\sigma_x^2 - r_i \int_{x_{i-1}}^{x_i} x f(x) dx) + \lambda(\int_{x_{i-1}}^{x_i} f(x) dx) \log(\int_{x_{i-1}}^{x_i} f(x) dx) \tag{4.26}$$

Setting $\frac{\partial J}{\partial r_i} = 0$, we obtain the optimality condition for r_i:

$$r_i = \sigma_x^2 \frac{\hat{r}_i}{\sum_{i=1}^{M} p_i \hat{r}_i^2} = s\hat{r}_i \tag{4.27}$$

Setting $\frac{\partial J}{\partial x_i} = 0$, we get,

$$x_i = \frac{r_i + r_{i-1} + \lambda \log \frac{p_i}{p_{i-1}}}{2s} = \frac{\hat{r}_i + \hat{r}_{i-1} + \hat{\lambda} \log \frac{p_i}{p_{i-1}}}{2} = \hat{x}_i \tag{4.28}$$

Hence, we conclude that for both fixed and variable rate quantization, it is sufficient to scale the reconstructions of the unconstrained quantizer to render the reconstruction error uncorrelated with the source. □

4.5 Asymptotic Analysis

To quantify theoretically how much a source can be compressed under the independent/uncorrelated reconstruction error constraint, we define two rate-distortion functions in which we constrain the reconstructions error to be i) uncorrelated with the source $R_U(D)$, and ii) independent of the source $R_I(D)$.

Let X denote the memoryless source, Y denote the reconstruction, S denote the reconstruction error ($S = Y - X$) with $f_{XS}(x, s)$ as the joint distribution of X and S. Also,

$d(x^n, y^n)$ denotes the additive distortion measure between sequences x^n and y^n, is defined as $d(x^n, y^n) = \frac{1}{n}\sum\limits_{i=1}^{n} d(x_i, y_i)$.

We start with an auxiliary lemma regarding the constraints.

Lemma 4.3. *Let x^n and y^n be n i.i.d. realizations of the random variables X and Y,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} |x_i(y_i - x_i)| = \mathbb{E}[|X(Y-X)|] \tag{4.29}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{XS}(x_i, y_i - x_i)}{f_X(x_i) f_S(y_i - x_i)} = \mathcal{D}(f_{XS}(X, Y-X)||f_X(X) f_S(Y-X)) \tag{4.30}$$

in probability.

Proof. By applying the weak law of large numbers to the left side of 4.29) and (4.30), we obtain the right sides. □

Lemma 4.4. *Minimizing $d(x, y)$ under uncorrelatedness (or independence) constraint is equivalent to minimizing the modified distortion measure $d_U(x, y)$ (or $d_I(x, y)$) where*

$$d_U(x, y) = \lim_{\lambda\to\infty} d(x, y) + \lambda|x(y - x)| \tag{4.31}$$

$$d_I(x, y) = \lim_{\lambda\to\infty} d(x, y) + \lambda \log(\frac{f_{XS}(x, y - x)}{f_X(x) f_S(y - x)}) \tag{4.32}$$

Proof. Using Lagrangian multiplier λ, we can cast one of the constraints into the other. Both independence and uncorrelatedness constraints can be written as expectation due to Lemma 4.3 and hence, can be cast into the expectation operator in the distortion constraint. The parameter corresponding to both constraints is $\lambda \to \infty$. □

We present the single letter characterization of the constrained rate distortion regions in the following theorem. Let $R_U(D)$ be the infimum of all achievable rates R with expected distortion $\mathbb{E}[d(X, Y)] \leq D$ subject to the constraint $\mathbb{E}[X(Y - X)] = 0$. Similarly, $R_I(D)$ is the infimum of all achievable rates R with expected distortion $\mathbb{E}[d(X, Y)] \leq D$ subject to the constraint $Y - X$ is independent of X.

Theorem 4.5.

$$R_U(D) = \inf_{\substack{Y:\mathbb{E}[d(X,Y)]\leq D \\ \mathbb{E}[X(Y-X)]=0}} I(X; Y) \tag{4.33}$$

$$R_I(D) = \inf_{\substack{Y:\mathbb{E}[d(X,Y)]\leq D \\ \mathcal{D}(f_{XS}(X,Y-X)||f_X(X)f_S(Y-X))=0}} I(X;Y) \tag{4.34}$$

Proof. The proof is a straightforward extension of the standard achievability and the converse proofs for regular rate distortion function, replacing d with d_U (d_I). Note that $R_U(D)$ ($R_I(D)$) is a special case of the conventional (unconstrained) $R(D)$ function with modified distortion measures d_U (d_I), i.e., $R_U(D)$ ($R_I(D)$) with distortion measure d is identical to $R(D)$ with distortion measure d_U (d_I). Lemma 4.4 ensures random coding arguments can be made with these measures.

Alternatively, similar to the unconstrained case, random coding arguments can be made with two distortion measures, noting that the independence and uncorrelatedness constraints can be considered as distortion measures, due to Lemma 4.4. With two distortion constraints, one can define a more limited distortion typical set and apply the same asymptotic equipartition property arguments used in proving the coding theorems in rate distortion theory. □

4.5.1 Gaussian Source with MSE Distortion

Next, we examine a special case when source is Gaussian and distortion measure is MSE. We show that if source distribution is Gaussian, for both uncorrelatedness and independence constraints, reconstruction error is Gaussian. We start with an auxiliary lemma without proof (see eg. [13] for the proof).

Lemma 4.6 ([13]). *Let* $\mathbf{S}$ *and* $\mathbf{S}_G$ *be random vectors in* $\mathbb{R}^N$ *with the same covariance matrix* $\mathbf{K}_S$*. If* $\mathbf{S}_G \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_S)$ *and* $\mathbf{S}$ *follows any other distribution, then*

$$\mathbb{E}_{S_G}[\log(f_{S_G}(\mathbf{S}))] = \mathbb{E}_S[\log(f_{S_G}(\mathbf{S}))] \tag{4.35}$$

where f_{S_G} *denotes the probability density function of* $\mathbf{S}_G$*, and* $\mathbb{E}_{S_G}$ *and* $\mathbb{E}_S$ *denote the expectations with respect to* $\mathbf{S}_G$ *and* $\mathbf{S}$*, respectively.*

Let us present a key lemma regarding the mutual information of two correlated random vectors constrained to have a fixed cross covariance matrix.

Lemma 4.7. *Let* $\mathbf{X}_G \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_X)$ *and let* $\mathbf{S}$ *and* $\mathbf{S}_G$ *be random vectors in* $\mathbb{R}^N$*, with the same covariance matrix,* $\mathbf{K}_S$ *and cross covariance matrix with* $\mathbf{X}_G$*,* $\mathbf{K}_{SX}$*. If* $\mathbf{S}_G \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_S)$ *and* *jointly Gaussian with* $\mathbf{X}$ *and* $\mathbf{S}$ *follows another distribution, then*

$$I(\mathbf{X}_G, \mathbf{X}_G + \mathbf{S}) \geq I(\mathbf{X}_G, \mathbf{X}_G + \mathbf{S}_G) \tag{4.36}$$

with equality if and only if $\mathbf{S} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_S)$.

Proof.

$$\begin{aligned}
&I(\mathbf{X}_G, \mathbf{X}_G + \mathbf{S}) - I(\mathbf{X}_G, \mathbf{X}_G + \mathbf{S}_G) \\
=& - h(\mathbf{X}_G|\mathbf{X}_G + \mathbf{S}) + h(\mathbf{X}_G|\mathbf{X}_G + \mathbf{S}_G) && (4.37) \\
=& - h(\mathbf{S}|\mathbf{Y}) + h(\mathbf{S}_G|\mathbf{Y}_G) && (4.38) \\
=& \int\int [-f_{S_G,Y_G}(\mathbf{s}_G, \mathbf{y}_G) \log(f_{S_G|Y_G}(\mathbf{s}_G|\mathbf{y}_G)) + f_{S,Y}(\mathbf{s}, \mathbf{y}) \log(f_{S|Y}(\mathbf{s}|\mathbf{y}))] d\mathbf{s}d\mathbf{y} && (4.39) \\
=&\iint -f_{S,Y}(\mathbf{s}, \mathbf{y}) \log(f_{S_G|Y_G}(\mathbf{s}_G|\mathbf{y}_G)) + f_{S,Y}(\mathbf{s}, \mathbf{y}) \log(f_{S|Y}(\mathbf{s}|\mathbf{y}))] d\mathbf{s}d\mathbf{y} && (4.40) \\
=&\iint f_{S,Y}(\mathbf{s}, \mathbf{y})[\log(f_{S|Y}(\mathbf{s}|\mathbf{y})) - \log(f_{S_G|Y_G}(\mathbf{s}_G|\mathbf{y}_G))] d\mathbf{s}d\mathbf{y} \\
=& \int f_Y(\mathbf{y}) \int f_{S|Y}(\mathbf{s}|\mathbf{y})(\log \frac{(f_{S|Y}(\mathbf{s}|\mathbf{y}))}{(f_{S_G|Y_G}(\mathbf{s}_G|\mathbf{y}_G))}) d\mathbf{s}d\mathbf{y} && (4.41) \\
=& \int f_Y(\mathbf{y}) \mathcal{D}(\mathbf{S}|\mathbf{Y}, \mathbf{S}_G|\mathbf{Y}_G) d\mathbf{y} && (4.42)
\end{aligned}$$

$\mathcal{D}$ is always non-negative and hence, this difference is always non-negative, completing the proof. Note that we used Lemma 4.6 and the fact that the joint distribution f_{S_G,Y_G} is Gaussian to obtain (4.40) from (4.39). □

Next, we present our main result on this topic:

Theorem 4.8. *For a Gaussian source and MSE distortion measure*

$$R_I(D) = R_U(D) \tag{4.43}$$

Proof. In general, $R_I(D) \geq R_U(D)$, since independent reconstruction error is also uncorrelated. Note the uncorrelated error constraint dictates $\mathbf{K}_{SX} = \mathbf{0}$, distortion constraint is $Tr(\mathbf{K}_S) = D$. Lemma 4.7 states that under these constraints, for a Gaussian source, Gaussian reconstruction error minimizes mutual information between the source and the reconstruction, i.e., $I(\mathbf{X}_G, \mathbf{X}_G+$

$\mathbf{S}$) achieves its minimum when $\mathbf{S} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_S)$. Then, $\mathbf{X}_G$ and $\mathbf{S}_G$ are uncorrelated and jointly Gaussian and are, thereby, also independent. Hence, (4.43) holds. □

In the following corollary, we answer this question: is the best possible vector quantizer at asymptotically high dimension that renders the reconstruction error uncorrelated with the source necessarily a randomized one?

Corollary 4.9. *For a Gaussian source, at asymptotically high quantizer dimension, the quantizer that achieves minimum distortion subject to the uncorrelated error constraint is necessarily a randomized one.*

Proof. Due to Theorem 4.8, the reconstruction error for the Gaussian source subject to uncorrelatedness constraint is actually independent of the source. Any deterministic quantizer cannot render the quantization noise independent from the source by definition; hence, it should be a randomized one. □

Note that our result holds only asymptotically, it is still open if this result holds at finite dimensions or not. We numerically answer this question for a scalar Gaussian source.

4.6 Experimental Results

We compare the proposed non-uniform dithered quantizer to conventional (uniform) dithered quantizer and constrained deterministic quantizer for a unit variance scalar Gaussian source. We implemented the above algorithm by numerically calculating the derived integrals. For that purpose, we sampled the distribution on a uniform grid. We also imposed bounded support (-3σ to 3σ) i.e., neglected tails of the Gaussian in the examples.

Figure 4.3 shows the comparative performances for fixed rate quantization. The proposed method outperforms both deterministic constrained quantizer and conventional dithered quantizer.

Figure 4.4 demonstrate the use of the proposed method for variable rate quantization. Note that for fixed rate, the conventional (uniform) dithered quantization suffers from the suboptimality of having equal quantization intervals significantly where as for the variable rate, the difference between the proposed and conventional dithered quantizers diminish especially at high rates.

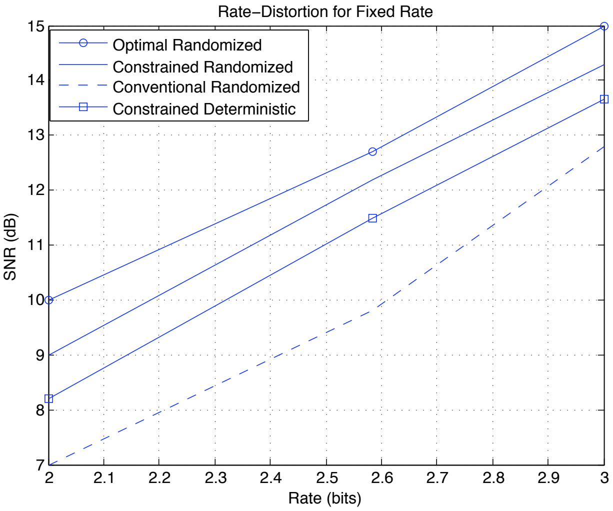


Figure 4.3. Performance comparison for fixed rate

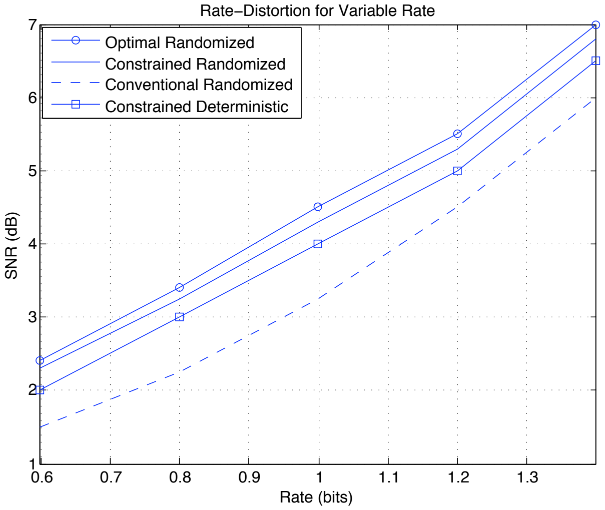


Figure 4.4. Performance comparison for variable rate

This is theoretically expected due to high rate optimality of uniform quantizers for variable rate quantization.

Chapter 5

Transform Coding

Transform coding is a computationally attractive approach to source coding, and is widely used in audio, image and video compression. In the basic transform coding setting, an input vector is linearly transformed into a vector in the transform domain whose components (also called transform coefficients) are scalar-quantized. The decoder reconstructs the quantized coefficients and performs linear (inverse) transformation to obtain an estimate of the source vector. The design goal is to find the optimal transform pair and bit allocation to scalar quantizers, which minimize distortion. In general, transform coding underperforms optimal vector quantization due to space filling loss in scalar quantizers, even if the transform generates independent coefficients. Nevertheless, due to its low complexity, transform coding is commonly employed in practical multimedia compression systems [27, 20].

Transform coding has been studied extensively. In their seminal paper, Huang and Schulthesis have shown [38] that if the vector source is Gaussian and the bit budget is asymptotically large, then the Karhunen Loeve transform (KLT) and its inverse are an optimal pair of transforms for fixed-rate coding. In a more recent paper Goyal, Zhuang and Vetterli improve that result by showing that KLT is optimal for Gaussian sources without making any high resolution assumptions [26]. Their results require a mild scale invariance assumption and apply to both the fixed and the variable rate quantizers.

The optimality of KLT in transform coding of Gaussian sources is often explained intuitively by the assertion that scalar quantization is better suited to the coding of independent random variables than to the coding of dependent random variables. Thus, the optimality of KLT for transform coding of Gaussian sources is understood to be a consequence of the fact that it yields independent transform coefficients. The application of KLT in transform coding of non-Gaussian sources is then justified using the intuitive argument that KLT's coefficient decorrelation represents, for general sources, a rough approximation to the desired coefficient independence.

In [15], the "popular trust" in the optimality of KLT is challenged and it is demonstrated by examples that KLT can be suboptimal for both fixed and variable rate quantization, at asymptotically high rate (with high resolution approximations). A theoretical result is also obtained, namely, a sufficient condition for optimality of KLT: when KLT generates independent coefficients then it is the optimal transform for variable rate coding.

In [42], a significant positive result is obtained regarding the optimality of KLT: KLT is optimal in conjunction with variable rate high resolution coding, not only for Gaussians but for the broader family of Gaussian vector mixtures, which includes Gaussian mixture models.

The problem is approached from a more practical perspective of numerical design in [7]. The authors proposed a gradient descent iterative algorithm to optimize the optimal orthogonal transform in conjunction with optimization of the quantization scheme. In simulations, they were able to demonstrate performance gains of the optimized transform-quantizer pair over KLT for practical sources.

In this chapter, we return to the fundamental theoretical problem of optimal transform coding. The main result is a necessary and sufficient condition for optimality of a transform in conjunction with variable rate coding at high resolution. Specifically, we show that the optimal transform is the one that minimizes the divergence between the joint distribution of the coefficients and the product of their marginals. In other words, it minimizes a quantitative measure of the dependence between the transform coefficients. Note furthermore that this result not only resolves the question of when KLT is optimal (at high resolution), but it also determines the optimal transform when it is not KLT.

We note minimizing a measure of dependence is closely related, at the high level, to the objective of the well studied problem of source separation. This observation is beneficial in two ways. First, we can leverage a rich reservoir of numerical algorithms, most importantly relating to independent component analysis [40, 12], in order to approximate the optimal transform. Moreover, our necessary and sufficient condition leads to contributions in source separation.

The main objective of source separation is exactly that of finding an orthogonal matrix that will generate coefficients "as independent as possible". Such matrices can be found by maximizing an ad hoc cost function ([39, 8, 48, 50]), called contrast function, that purports to quantify how close to statistically independent the resulting components are. One can choose one of many ways to define the contrast function, and this choice governs the form of the algorithms. The two broadest definitions of independence are based on minimization of mutual information or maximization of “non-Gaussianity”. The latter is motivated by the central limit theorem, uses kurtosis and negentropy. The former family of algorithms is obviously closely related measures involving the Kullback-Leibler (KL) divergence.

Our main result yields the precise connection between the problem of finding the optimal transform in high resolution variable rate coding and the source separation problem, when the objective (contrast) function is effectively the divergence. The optimal transform for the former (source coding) problem is shown to minimize the objective of the latter problem. This suggests that advances in transform coding may have an impact directly in source separation. An example of such a result is presented in Section IV, where our necessary and sufficient condition for optimality maps the result of [42] to ensure the optimality of KLT for source separation of Gaussian vector mixtures.

5.1 Review of Prior Results

5.1.1 Preliminaries and Notation

Let source $\mathbf{X}$ be an N dimensional random vector, with real components, $X_1, X_2..., X_N$. Without loss of generality, we assume $\mathbb{E}(\mathbf{X}) = \mathbf{0}$, and hence $\mathbf{R_X} = \mathbf{E}(\mathbf{XX^T})$. Let the transform $\mathbf{U}$ be a real $N \times N$ orthogonal matrix ($\mathbf{U}^{-1} = \mathbf{U}^T$) and let

$$\mathbf{Y} = \mathbf{UX} \tag{5.1}$$

be the transformed random vector with coefficients $Y_1, Y_2, ..Y_N$. A scalar quantizer Q is a mapping $Q : \mathbb{R} \rightarrow \mathbb{R}$. We restrict this part to variable rate analysis, and the rate needed to describe source X after quantization by quantizer Q is

$$R(Q) = H[Q(X)] \tag{5.2}$$

A transform coding scheme is a structured vector quantizer where the random vector $\mathbf{X}$ is transformed into $\mathbf{Y}$ by $\mathbf{Y} = \mathbf{UX}$ and then each component Y_i is quantized with scalar quantizers Q_i. The total rate of the transform coder is

$$R_T = \sum_i H(Q_i(Y_i)) \tag{5.3}$$

At the decoder, inverse transformation by the matrix $\mathbf{U}^{-1} = \mathbf{U}^T$ is used to obtain an estimate of the source vector. The corresponding distortion is measured as mean square error,

$$D_T = \mathbb{E}\{||\mathbf{X} - \mathbf{U}^T\mathbf{Q}(\mathbf{UX}))||_2^2\} \tag{5.4}$$

where $\mathbf{Q}(\mathbf{X}) = [Q_1(X_1), .., Q_N(X_N)]^T$.

5.1.2 High rate approximations

The quantization operation is nonlinear and difficult to analyze mathematically. However, for both fixed and variable rate quantization, high resolution approximations can be made. Specifically, if the density of a scalar random variable is reasonably smooth, then at sufficiently high rate the distribution within a quantization interval is uniform. It is well known that uniform quantizers are asymptotically (at high resolution) optimal for variable rate coding, irrespective of the density of the source to be quantized [23]. Therefore, we use uniform quantizers throughout the part. Let Δ_i be step size for i^{th}transform coefficient. This assumption results in quantization noise that is uniformly distributed over $(-\Delta_i, \Delta_i)$. Thus, at high resolution the distortion D_i is approximated as:

$$D_i = \frac{\Delta_i^2}{12} \tag{5.5}$$

The following straightforward auxiliary lemma relates the differential entropy of a continuous random variable with the entropy of its reproduction after uniform quantization at high resolution:

Lemma 5.1 (e.g., [13])**.** *If density $f_X(x)$ of random variable X is Riemann integrable, and $Q(X)$ is its reproduction after uniform quantization with step size Δ, then the following holds asymptotically, as $\Delta \to 0$:*

$$H(Q(X))) + \log \Delta \to h(X) \tag{5.6}$$

This lemma will be used in the proof of Theorem 5.2.

5.1.3 On Optimality of KLT

Definition (KLT): An orthogonal $N \times N$ matrix $\mathbf{K}$ is a KLT of N dimensional source vector $\mathbf{X}$ with covariance matrix $\mathbf{R}_X$ if $\mathbf{K}\mathbf{R}_X\mathbf{K}^T = \mathbf{\Lambda}_X$, where $\mathbf{\Lambda}_X$ is diagonal.

In other words, KLT generates uncorrelated coefficients. It is well known that KLT is optimal for "zonal sampling" or "truncated expansion": if the source estimate is approximated by expansion from a pre-determined subset of the transform coefficients, then KLT minimizes the approximation error. Another optimality aspect of KLT is shown in [5] for Gaussian sources: KLT minimizes the expected number of expansion terms (or transform coefficients) if the reconstruction error is required to be below a prescribed threshold. It has more recently been shown that KLT is optimal for Gaussian sources for both variable and fixed rate and at any operating rate regime, i.e., without any high resolution approximations [26]. Note that KLT is not necessarily unique. As example, when $\mathbf{R}_X = \mathbf{I}$, any orthogonal transform $\mathbf{U}$ "diagonalizes" $\mathbf{R}_X$ as $\mathbf{U}\mathbf{I}\mathbf{U}^T = \mathbf{I}$. Then a natural question arises: do all these KLTs perform equally? A sufficient condition for optimality of a KLT (that resolves this question if satisfied by one of the contenders) was given in [15] and is reproduced here.

Effros-Feng-Zeger Theorem (EFZ) [15]: If a KLT produces independent transform coefficients, then it is optimal for variable-rate transform coding at high resolution.

Note that this sufficient condition for optimality is not necessary. Specifically, there is a family of distributions where KLT has been shown to be optimal for transform coding although it does not generate independent coefficients [42].

Definition (Gaussian Vector Scale Mixtures): A random vector $\mathbf{X}$ taking values in $\mathbb{R}^N$ is called Gaussian Vector Scale Mixture (GVSM) if $\mathbf{X} = \mathbf{C}^T(\mathbf{Z} \odot \mathbf{V})$ where $\mathbf{C}$ is a constant orthogonal matrix, random vector $\mathbf{Z} \sim \mathbb{N}(0, \mathbf{I})$, scale vector $\mathbf{V}$ is a random vector independent of $\mathbf{Z}$ and taking values in $\mathbb{R}^+$, and $\odot$ denotes the element-wise product.

Note that conditioned on $\mathbf{V} = \mathbf{v}$, the GVSM vector $\mathbf{X}$ is Gaussian. Note further that this definition characterizes a a fairly broad set of distributions, including Gaussian mixtures.

Jana-Moulin (JM) Theorem [42]: KLT is optimal for a GVSM source for variable rate coding at high resolution.

This theorem clearly identifies a set of source distributions for which KLT is optimal, but leaves open the question of whether KLT is strictly suboptimal outside this set.

In summary, several natural follow-up questions remain open: when is KLT optimal for transform coding of general non-Gaussian sources? What is a conclusive condition for optimality of a general (not necessarily KLT) transform? If KLT is suboptimal, how can we numerically find the optimal transform? Here, we present a necessary and sufficient condition for optimality of any transform, naturally including KLT. Also, when KLT is suboptimal, we propose an algorithm to find the optimal transform.

5.2 Main Result

The main result is stated in the following theorem.

Theorem 5.2. *Orthogonal transform $\mathbf{U}^*$ is optimal if and only if the following is satisfied:*

$$\mathbf{U}^* = \underset{\mathbf{U}}{\operatorname{argmin}}\, \mathcal{D}(f_Y(\mathbf{y}) || \prod_{i=1}^{N} f_{y_i}(y_i)) \tag{5.7}$$

where $\mathcal{D}$ is divergence.

Note: Theorem 5.2 subsumes Effros-Feng-Zeger theorem [15] as an extreme special case where KLT yields independent coefficients.

The proof will make use of a trivial auxiliary lemma, which we state without proof:

Lemma 5.3. *The joint entropy is invariant to orthogonal transformation: Let* $\mathbf{X}$ *be a random vector and* $\mathbf{U}$ *be an orthogonal matrix, then*

$$h(\mathbf{UX}) = h(\mathbf{X}) \tag{5.8}$$

Proof of Theorem 5.2 . Using high resolution approximation for variable rate quantization, we get the following for total distortion,

$$D_T = \sum_i \frac{\Delta_i^2}{12}, \tag{5.9}$$

and for total rate,

$$R_T = \sum_i H(Q(y_i)) \tag{5.10}$$

Since the distortion is independent of the distribution of the transform coefficients, the aim of the transform coder is to minimize the total rate R_T. Using Lemma 1, we can rewrite (13) as,

$$R_T = -\sum_i \int f_{y_i}(y_i) \log f_{y_i}(y_i) dy_i + \log \Delta_i \tag{5.11}$$

where f_{y_i} is the marginal density of the i^{th} transform coefficient. Since the quantization intervals are fixed, the optimal transform must minimize the first term, hence the cost function:

$$J = -\sum_i \int f_{y_i}(y_i) \log(f_{y_i}(y_i)) dy_i = -\int f_Y(\mathbf{y}) \left[\sum_i \log(f_{y_i}(y_i)) \right] d\mathbf{y} \tag{5.12}$$

Using Lemma 5.3, we write the differential entropy $h(\mathbf{y})$ as

$$-\int f_Y(\mathbf{y}) \log f_Y(\mathbf{y}) d\mathbf{y} = -\int f_X(\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} = C \tag{5.13}$$

where C is used to emphasize that the joint entropy is determined by the source distribution and is hence constant with respect to the transform. Subtracting the constant C from both sides of (5.12), and noting that minimizing J is equivalent to minimizing $J - C =$

$$-\int f_Y(\mathbf{y}) \left[\sum_i \log(f_{y_i}(y_i)) \right] d\mathbf{y} + \int f_Y(\mathbf{y}) \log f_Y(\mathbf{y}) d\mathbf{y} = \mathcal{D}(f_Y(\mathbf{y}) || \prod_{i=1}^{N} f_{y_i}(y_i)) \tag{5.14}$$

which completes the proof. □

Note that Theorem 5.2 essentially states that the optimal transform is the one that minimizes the statistical dependence of the transform coefficients. KLT considers second order statistics and decorrelates the transform coefficients, but this is neither necessary nor sufficient to minimize the overall statistical dependence as measured by the above divergence. The theorem also suggests that the optimal transform deviates from KLT whenever second order statistics are not a good representative of the overall dependence. This result also subsumes as a direct corollary the EFZ Theorem [15], and the well known optimality of KLT for jointly Guassian sources at high resolution variable rate coding [20].

5.3 Source Separation Problem

Hyvarinen and Oja [40] give the following definition for the noise free linear source separation problem, which is of interest here.

Definition (Source Separation Problem): Let random vector $\mathbf{X}$ of size N be obtained by

$$\mathbf{X} = \mathbf{BS} \tag{5.15}$$

where $\mathbf{B}$ is a constant $N \times N$ "mixing" matrix, elements S_i in the vector $\mathbf{S} = (S_1, ..., S_N)^T$ are assumed to be mutually independent. $\mathbf{X}$ is observed while both $\mathbf{B}$ and $\mathbf{S}$ are unknown. The aim of the problem is to find $\mathbf{S}$ (or alternatively the matrix $\mathbf{B}$), by maximizing some form of independence among the transform coefficients.

We choose the objective function of divergence between the product of the marginals of the transform coefficients and joint density of the transformed vector, i.e. the cost function

$$J(\mathbf{U}) = \mathcal{D}(f_Y(\mathbf{y}) || \prod_{i=1}^{N} f_{y_i}(y_i)) \tag{5.16}$$

where $\mathbf{Y} = \mathbf{UX}$.

Our main result provides two prospective directions to pursue: i) It allows us to develop an algorithm for the long standing problem of optimal transform coding by leveraging a large bank of algorithms from the source separation literature, and ii) to apply the theoretical optimality (or suboptimality) results of transform coding to source separation problems. An algorithm for finding the optimal transform is presented in the next section. In the remainder of this section we use the JM Theorem to obtain a new optimality result in source separation.

Theorem 5.4. *The optimal orthogonal transform for source separation of a Gaussian vector scale mixture is KLT, when the contrast function is the divergence-based cost of (5.16).*

Proof. The proof follows from Theorem 5.2 and the JM theorem. □

The theorem establishes the optimality of KLT and hence renders source separation algorithms for this family of sources unnecessary.

5.4 Algorithm

In this section, we propose a modified version of the algorithm by Pham [56, 57] which seeks to find the orthogonal transform that minimizes the contrast function expressed in (5.16). The minimization of the cost can be done through a gradient descent algorithm, where the update for transform matrix $\mathbf{U}$ involves a matrix ϵ yielding $\mathbf{U}+\epsilon\mathbf{U}$. We expand $\mathbf{U}+\epsilon\mathbf{U}$ with respect to ϵ up to second order terms and then minimize the resulting cost with respect to ϵ to obtain the optimal ϵ and hence a new estimate. The Taylor expansion of $J(\mathbf{U}+\epsilon\mathbf{U})$ can be expressed as follows:

$$\begin{aligned} J(\mathbf{U}+\epsilon\mathbf{U}) = J(\mathbf{U}) &+ \sum_{i,j}\epsilon_{ij}[\mathbb{E}(Y_j\Phi_i(Y_i)) - \mathbb{E}(Y_i\Phi_j(Y_j))] \\ &+ \frac{1}{2}\sum_{i,j}\epsilon_{ij}^2[\mathbb{E}(\Phi_i^2(Y_i))\mathbb{E}(Y_j^2) - \mathbb{E}(\Phi_j^2(Y_j))\mathbb{E}(Y_i^2) - 2] + O(\epsilon^3) \end{aligned} \tag{5.17}$$

where Φ is the gradient of the entropy function, also known as score function and $O(\epsilon^3)$ accounts for higher order terms which we will neglect. Setting the partial derivative with respect to ϵ to zero, we find ϵ as follows:

$$\epsilon_{ij} = \frac{\mathbb{E}(Y_j\Phi_i(Y_i)) - \mathbb{E}(Y_i\Phi_j(Y_j))}{\mathbb{E}(\Phi_i^2(Y_i))\mathbb{E}(Y_j^2) - \mathbb{E}(\Phi_j^2(Y_j))\mathbb{E}(Y_i^2) - 2} \tag{5.18}$$

In this expression, the probability density functions being unknown, the score function $\Phi(Y)$ is replaced by an estimate (see [57]) and the expectations are estimated from training samples assuming ergodicity. There is no guarantee that $\mathbf{U}+\epsilon\mathbf{U}$ will be orthogonal. To solve this problem, we replace the resulting matrix $\mathbf{U}$ with its closest (in terms of Frobenius norm) orthogonal approximation which can be obtained by polar decomposition[1].

We obtained some preliminary results using the proposed algorithm. We first generate the samples of $\mathbf{X}$ by $\mathbf{X} = \mathbf{BS}$ where $\mathbf{S}$ consists of four independent and identically distributed random variables, and $\mathbf{B}$ is a random orthogonal mixing matrix. The proposed algorithm finds the correct matrix $\mathbf{U} = \mathbf{B}^{-1}$ precisely. We note that an obvious KLT choice is the identity $\mathbf{I}$ since the source is already uncorrelated. It follows from the examples in [15], that the gain of the optimal transform over standard KLT (in this case the identity matrix, $\mathbf{I}$) can be unbounded.

[1] We employed a fast method as to repeatedly average $\mathbf{U}$ with its transpose inverse until convergence [25].

5.5 Transform Coding with Dithered Quantization

We note again that in the case of deterministic optimal quantization, the optimal tranform is unknown for most distributions other than Gaussians [15]. A main premise of this section is that for fixed rate coding, dithered quantization enables universal transform coding, i.e., the optimal transform is generally derived for all sources while this is not the case for variable rate coding due to the dependence of the coding rate on the source distribution. Also, the quantization error is statistically orthogonal to the source and hence may be viewed as an additive independent noise term which in turn enables solving for the optimal transform by linear analysis (by solving a matrix equation.) This is not the case for optimal deterministic quantization, where difficulties include: i) the quantization error term can only be approximated as an additive uncorrelated noise at asymptotically high resolution; ii) the expected distortion depends on the type of distribution of each transform coefficient which, except in the simple Gaussian source case, depends non-trivially on the transform and makes it extremely challenging to minimize the distortion with respect to the transform. Because of these difficulties, it is not straightforward to derive the optimal transform for non-Gaussians. The motivation for this work stems from the realization that dithered quantization holds considerable promise for circumventing the above difficulties and deriving the optimal transform for all sources.

5.5.1 Simple Scalar Case

Some intuition is gained already from a simple scalar quantization setting. The dithered scalar quantizer is equivalent to the case where scalar source x is corrupted by i.i.d (quantization) noise n, which is uncorrelated with x. At the receiver, $y = x + n$ is available and best linear estimate for x is

$$\hat{x} = (\frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2})y \tag{5.19}$$

Note that an optimal deterministic (Lloyd-Max) quantizer would reconstruct $\hat{x} = y$ [20]. This simple observation of the scalar case already intuitively suggests that a unitary transform and specifically KLT will not be optimal for dithered quantization. Next, let us assume the source x is Gaussian, and allow for scaling coefficients α before quantization and β after quantization. Let also $f(b)$ be the distortion function of the dithered quantizer applied to unit variance, zero mean Gaussian at b bits. Then, $\sigma_n^2 = \alpha^2\sigma_x^2 f(b)$ and $\hat{x} = \beta(\alpha x + n)$. The optimal α, β will minimize J where

$$\begin{aligned} J &= E[(x-\beta(\alpha x+n))^2] \\ &= (1-\beta\alpha)^2\sigma_x^2+\beta^2\sigma_n^2 \\ &= (1-\beta\alpha)^2\sigma_x^2+\beta^2\alpha^2\sigma_x^2 f(b) \\ &= (1-2\beta\alpha+(1+f(b))(\beta\alpha)^2)\sigma_x^2 \end{aligned} \tag{5.20}$$

Not surprisingly, J depends on the scaling coefficients only through the product $\beta\alpha$. By the optimality condition $\partial J/\partial(\beta\alpha)=0$, we obtain the optimal scaling

$$\beta\alpha = 1/(1+f(b)) \tag{5.21}$$

Generalizing to transform coding of signal blocks we intuitively expect that KLT followed by an appropriate diagonal scaling matrix would be optimal. The following section concretizes this intuition in a precise statement and formally proves it.

5.5.2 Optimal Transform for a Given Bit Allocation

A jointly Gaussian vector $\mathbf{x}$ with covariance matrix $\mathbf{R_x}$ is first linearly transformed to obtain $\mathbf{y}=\mathbf{Ex}$, then quantized, $\hat{\mathbf{y}}=\mathbf{Q}(\mathbf{y})$ consists of scalar quantized samples. $\mathbf{n}=\mathbf{y}-\hat{\mathbf{y}}$ denotes the quantization error vector. At the receiver side, a linear estimator is used to get an estimate of $\mathbf{x}$ as $\hat{\mathbf{x}}=\mathbf{D}\hat{\mathbf{y}}$ to minimize the mean square error,

$$J = E[(\mathbf{x}-\hat{\mathbf{x}})^{\mathbf{T}}(\mathbf{x}-\hat{\mathbf{x}})] = \mathbf{E}[\mathbf{Tr}((\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})^{\mathbf{T}})] \tag{5.22}$$

Note that we include the "trace" formulation of the criterion as it will be convenient for matrix manipulations in the sequel. As is common, we assume for simplicity that the source is zero mean and that MSE is the distortion criterion. We denote the KLT of the source, denoted as $\mathbf{S}$ that satisfies

$$\mathbf{S}\mathbf{R_X}\mathbf{S^T} = \mathbf{\Psi} \tag{5.23}$$

where $\mathbf{\Psi}=\mathbf{diag}(\lambda_\mathbf{1},\lambda_\mathbf{2},...,\lambda_\mathbf{N})$ and λ_i's are eigenvalues of $\mathbf{R_x}$. For convenience we will assume the ordering $\lambda_1\geq\lambda_2\geq,...,\lambda_N$ with corresponding number of bits spent on coefficients $b_1\geq b_2\geq ...\geq b_N$. Consider the problem: given bit allocation vector $\mathbf{b}=[b_1,b_2,...,b_N]$, find optimal $\mathbf{E}$ and $\mathbf{D}$ transform matrices to minimize the MSE. Without loss of generality we assume that the bit allocation vector is ordered, i.e., $b_1\geq b_2\geq ...\geq b_N$. The MSE cost can be written in trace form as

$$J = E[\mathbf{Tr}(\mathbf{x}-\mathbf{DEx}-\mathbf{Dn})(\mathbf{x}-\mathbf{DEx}-\mathbf{Dn})^{\mathbf{T}}] \tag{5.24}$$

Since we use a dithered quantizer, quantization error is uncorrelated with $\mathbf{y}$ and with $\mathbf{x}$, i.e., $E(\mathbf{xn^T})=\mathbf{0}$. So we can write:

$$J = \mathbf{Tr}(\mathbf{DER_xE^TD^T}+\mathbf{R_x}+\mathbf{DR_nD^T}-\mathbf{2DER_x}) \tag{5.25}$$

As $\mathbf{R_x}$ does not depend on the transform, we may equivalently minimize

$$J_1 = \mathbf{Tr}(\mathbf{DER_xE^TD^T} + \mathbf{DR_nD^T} - \mathbf{2DER_x}) \tag{5.26}$$

Suppose there is a single function $f(.)$ to describe the rate distortion performance of the scalar dithered quantization of each transform coefficent through

$$E[(y_i - \hat{y}_i)^2] = \sigma_i^2 f(b_i) \tag{5.27}$$

where b_i and σ_i^2 denote the number of bits allocated to coefficient y_i and the variance of coefficient y_i respectively, for $i = 1, 2, ..., N$. It follows from the basic properties of dithered quantization that

$$\mathbf{R_n} = \mathbf{diag}(\sigma_1^2 f(b_1), \sigma_2^2 f(b_2), ..., \sigma_N^2 f(b_N)) \tag{5.28}$$

Now, we define a convenient linear matrix operator, $\mathbf{d}(.)$ which sets to zero all off-diagonal entries of the argument matrix. Specifically, $\mathbf{d}(\mathbf{A}) = \mathbf{diag}(a_{11}, a_{22}, ...a_{NN})$ where $\mathbf{A}$ is some $N \times N$ matrix. Note that,

$$\mathbf{R_n} = \mathbf{d}(\mathbf{ER_xE^T\Lambda}) \tag{5.29}$$

where $\mathbf{\Lambda} = \mathbf{diag}(f(b_1), f(b_2), ..., f(b_N))$. Also, it is straightforward to show (using matrix basic operations or the linear operator properties) that

$$\mathbf{d}(\mathbf{A\Gamma}) = \mathbf{\Gamma d}(\mathbf{A}) \tag{5.30}$$

for any diagonal $\mathbf{\Gamma}$ matrix. The following is a useful auxilary lemma.

Lemma 5.5. *For any arbitrary function matrix* $\mathbf{A}$*, variable matrix* $\mathbf{B}$ *and constant diagonal matrix* $\mathbf{\Gamma}$*,*

$$\partial(\mathbf{d}(\mathbf{A\Gamma}))/\partial\mathbf{B} = \mathbf{\Gamma d}(\partial\mathbf{A}/\partial\mathbf{B}) \tag{5.31}$$

Proof. Since both $\mathbf{d}(.)$ and differentiation are linear operators, they may be interchanged. Using (5.30) it is straightforward to obtain the lemma claim (5.31). □

Let $\mathbf{S}$ denote any unitary matrix that diagonalize $\mathbf{R_x}$ as defined in (5.23). We write $\mathbf{E} = \mathbf{\Phi_1 S^T}$ and $\mathbf{D} = \mathbf{S\Phi_2}$ for any arbitrary $\mathbf{\Phi_1}$, $\mathbf{\Phi_2}$ matrices.

Lemma 5.6. *The optimal* $\mathbf{\Phi_1}$ *and* $\mathbf{\Phi_2}$ *matrices are diagonal.*

Proof. $\mathbf{DE} = \mathbf{S\Phi_2\Phi_1S^T}$ and $\mathbf{ER_xE^T} = \mathbf{\Phi_1\Psi\Phi_1^T}$. Substituting these expressions into (5.26) we obtain

$$J_1 = \mathbf{Tr}(\mathbf{\Phi_2\Phi_1\Psi\Phi_1^T\Phi_2^T}) + \mathbf{Tr}(\mathbf{\Phi_2 d(\Lambda\Phi_1\Psi\Phi_1^T)\Phi_2^T}) - \mathbf{2Tr}(\mathbf{\Phi_2\Phi_1\Psi}) \tag{5.32}$$

Rearranging the terms using the trace equality $\mathbf{Tr}(\mathbf{AB}) = \mathbf{Tr}(\mathbf{BA})$,

$$J_1 = \mathbf{Tr}(\mathbf{\Phi_2^T\Phi_2(\Phi_1\Psi\Phi_1^T + d(\Lambda\Phi_1\Psi\Phi_1^T))}) - \mathbf{2Tr}(\mathbf{\Phi_2\Phi_1\Psi}) \tag{5.33}$$

Setting $\partial J_1/\partial\mathbf{\Phi_2} = \mathbf{0}$, yields

$$\mathbf{2\Phi_2(\Phi_1\Psi\Phi_1^T + d(\Lambda\Phi_1\Psi\Phi_1^T)) - 2\Psi\Phi_1^T = 0} \tag{5.34}$$

Rearranging terms:

$$\mathbf{\Phi_2 = \Psi\Phi_1^T(\Phi_1\Psi\Phi_1^T + \Lambda d(\Phi_1\Psi\Phi_1^T))^{-1}} \tag{5.35}$$

Setting $\partial J_1/\partial\mathbf{\Phi_1} = \mathbf{0}$ and applying Lemma 5.5, we obtain

$$\mathbf{(2\Psi\Phi_1^T + (2\Lambda\Psi d(\Phi_1))\Phi_2\Phi_2^T - 2\Psi\Phi_2^T = 0} \tag{5.36}$$

and hence

$$\mathbf{\Phi_2 = (\Phi_1 + \Lambda d(\Phi_1))^{-1}} \tag{5.37}$$

Note that, we used the dithered quantization property that quantization noise is independent of the source (with Gaussian source assumption for the variable rate case) in this derivation, which implies $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_1} = \mathbf{0}$ and $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_2} = \mathbf{0}$. In conventional quantization, $\mathbf{\Lambda}$ depends on the distribution of the transform coefficient which is hard to track analytically. This point makes the solution difficult for non-Gaussians (note for Gaussian source y_i's are all Gaussian irrespective of the linear transform, so $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_1} = \mathbf{0}$ and $\partial\mathbf{\Lambda}/\partial\mathbf{\Phi_2} = \mathbf{0}$ hold.) Substituting (5.35) into (5.33) yields

$$J_1 = \mathbf{Tr}(\mathbf{\Phi_2^T\Psi\Phi_1^T}) - \mathbf{2Tr}(\mathbf{\Phi_2\Phi_1\Psi}) \tag{5.38}$$

Noting that, $\mathbf{\Psi}$ is diagonal and using the trace equalities $\mathbf{Tr}(\mathbf{A}) = \mathbf{Tr}(\mathbf{A^T})$ and $\mathbf{Tr}(\mathbf{ABC}) = \mathbf{Tr}(\mathbf{CAB})$, we get $\mathbf{Tr}(\mathbf{\Phi_2^T\Psi\Phi_1^T}) = \mathbf{Tr}(\mathbf{\Phi_1\Psi\Phi_2}) = \mathbf{Tr}(\mathbf{\Phi_2\Phi_1\Psi})$ and hence

$$J_1 = -\mathbf{Tr}((\mathbf{\Phi_2\Phi_1\Psi}) \tag{5.39}$$

Plugging (5.37) into (5.39) we get

$$J_1 = -\mathbf{Tr}((\mathbf{\Phi_1} + \mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}}\mathbf{\Phi_1\Psi}) \tag{5.40}$$

Now, J_1 is a function of only $\mathbf{\Phi_1}$. Hence, setting the partial derivative with respect to $\mathbf{\Phi_1}$ to zero, $\partial J_1/\partial\mathbf{\Phi_1} = \mathbf{0}$ and using matrix inversion lemma

$$\begin{aligned}(\mathbf{I}+\mathbf{\Lambda})^{\mathbf{-1}} &= \mathbf{\Phi_1}(\mathbf{\Phi_1}+\mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}} \\ &= \mathbf{I}-\mathbf{\Lambda d}(\mathbf{\Phi_1})(\mathbf{\Phi_1}+\mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}}\end{aligned} \tag{5.41}$$

$(\mathbf{\Lambda}+\mathbf{I})^{\mathbf{-1}}$ is a diagonal matrix, since $\mathbf{\Lambda}$ and $\mathbf{I}$ are both diagonal. $\mathbf{\Lambda d}(\mathbf{\Phi_1})$ is also diagonal since it is the product of two diagonal matrices. The remaining factor $(\mathbf{\Phi_1}+\mathbf{\Lambda d}(\mathbf{\Phi_1}))^{\mathbf{-1}}$ must therefore be diagonal, which requires Φ_1 to be diagonal, i.e., $\mathbf{\Phi_1} = \mathbf{d}(\mathbf{\Phi_1})$. Similar reasoning applied to (5.37) yields the conclusion that $\mathbf{\Phi_2}$ is also diagonal. □

Now, we can state the main theorem on this topic:

Theorem 5.7. *For given ordered bit allocations $b_1 \geq b_2 \geq, ..., b_N$, the* **E**, **D** *transform matrices that minimize MSE for dithered scalar quantization are given by*

$$\mathbf{E} = \mathbf{\Phi_1 S^T}, \mathbf{D} = \mathbf{S\Phi_2} \tag{5.42}$$

for any $\mathbf{\Phi_1}$, $\mathbf{\Phi_2}$ *diagonal matrices that satisfy*

$$\mathbf{\Phi_1\Phi_2} = (\mathbf{I}+\mathbf{\Lambda})^{\mathbf{-1}} \tag{5.43}$$

and **S** *is the KLT matrix. Moreover, the distortion is*

$$J = \sum_{i=1}^{N} \lambda_i \frac{f(b_i)}{1+f(b_i)} \tag{5.44}$$

Proof. Using Lemma 5.6 and $\mathbf{Tr}(\mathbf{\Gamma\Theta\Omega}) = \mathbf{Tr}(\mathbf{\Theta\Gamma\Omega})$ for any diagonal matrices $\mathbf{\Theta}$, $\mathbf{\Gamma}$, $\mathbf{\Omega}$, (5.33) can be written as:

$$\mathbf{J} = \mathbf{Tr}((\mathbf{I}-\mathbf{\Phi_2\Phi_1})\mathbf{\Psi}) \tag{5.45}$$

MSE is only a function of $\mathbf{\Phi_1\Phi_2}$, not depending on individual values of $\mathbf{\Phi_1}$ or $\mathbf{\Phi_2}$. By (5.37) and Lemma 5.6 the optimality condition on $\mathbf{\Phi_1\Phi_2}$ follows directly: $\mathbf{\Phi_1\Phi_2} = (\mathbf{I}+\mathbf{\Lambda})^{\mathbf{-1}}$

While there are possibly $N!$ distinct $\mathbf{S}$ matrices that satisfy (5.23) corresponding to $N!$ permutations of distinct eigenvalues of $\mathbf{R_x}$. To select the optimal $\mathbf{S}$ matrix, we need the optimal ordering of eigenvalues with respect to the ordering of bit allocations, as is standard practice with KLT: higher rate should be allocated to the component that corresponds to larger eigenvalue. We are trying to minimize J, i.e., maximize J_2 where,

$$\begin{aligned} J_2 &= \mathbf{Tr}(\mathbf{\Phi_1\Phi_2\Psi}) \\ &= \mathbf{Tr}(\mathbf{\Psi}(\mathbf{I}+\mathbf{\Lambda})^{\mathbf{-1}}) \\ &= \sum_{i=1}^{N} \tfrac{\lambda_i}{1+f(b_i)} \end{aligned} \tag{5.46}$$

Both λ_i and $1+f(b_i)$ terms are positive, the maximum is achived when λ_i are in reverse order relative to $1+f(b_i)$ [53]. Since $f(b_i)$ is a decreasing function of b_i, λ_i should be ordered as is b_i, namely in decreasing order. Hence, the optimal permutation of the rows of $\mathbf{S}$ is the one that provides $(\lambda_1 \geq \lambda_2 \geq, ..., \lambda_N)$ when the bit allocation vector is ordered such that $(b_1 \geq b_2 \geq, ..., b_N)$ □

For the given $\mathbf{E}$, $\mathbf{D}$ matrices, the bit allocation should minimize MSE. If we write MSE in terms of quantization function and ordered eigenvalues using the main theorem,

$$J = \sum_{i=1}^{N} \lambda_i \frac{f(b_i)}{1+f(b_i)} \tag{5.47}$$

Note that if standard KLT is used the distortion is

$$J_{KLT} = \sum_{i=1}^{N} \lambda_i f(b_i) \tag{5.48}$$

which is strictly larger than that of the proposed transform.

Chapter 6

Conclusion and Future Directions

This dissertation has primarily focused on optimizing mappings, linear and nonlinear, for limited delay and energy communications.

In Chapter 2 we derived the necessary conditions of optimality for a given source-channel system. Based on the necessary conditions, we derived an iterative algorithm which generates locally optimal analog mappings. Comparative results and example mappings are provided and it is shown that the proposed method improves upon prior work.

In Chapter 3, we derived conditions under which the L_p optimal estimator is linear. We identified the conditions for the existence and uniqueness of a source distribution that matches the noise in a way that ensures linearity of the optimal estimator for the special case of $p = 2$. One trivial example of this type of matching occurs for Gaussian source and Gaussian noise at all SNR levels. Another instance of matching happens when the source and noise are identically distributed. We also show that Gaussian source-channel pair is unique in that it is the only source-channel pair for which the optimal estimator is linear at more than one SNR value. Moreover, we show the asymptotical linearity of MSE optimal estimators for low SNR if the channel is Gaussian regardless of the source and vice versa, for high SNR if the source is Gaussian regardless of the channel. We also study the extension to higher dimensions where additional constraints are introduced to the set of necessary and sufficient conditions, beyond the ones inherited from the scalar case

In Chapter 4, we proposed a nonuniform randomized quantizer where the dithering is performed in companded domain to solve the problem of matching dither variance to varying quantization intervals. The optimal compressor and expander mappings that minimize mean square error are found by an iterative method. The extension of the method to vector quantization with finite dimensions is left as future work. Moreover, we analyzed the randomized quantizer at asymptotically high dimension. The main result of our analysis is: for a Gaussian source, with asymptotically high dimensions, the optimal vector quantizer that renders quantization error uncorrelated with the source must be a randomized one. As a future work, we will investigate whether there are other cases in which random encoding is not merely a tool to deduce the rate-distortion bounds, but a necessary element in achieving such bounds.

In the last part, we presented a necessary and sufficient condition for transform optimality at high resolution, variable rate coding. Note that this result not only resolves the question of when KLT is optimal (at high resolution), but also determines the optimal transform when it is not KLT. This condition also points to direct connections between the transform coding problem and an important subset of the well studied source separation problems. We used this observation to obtain new results in two directions: developing a numerical algorithm for transform optimization in transform coding by leveraging tools from source separation; and mapping known theoretical optimality results in transform coding to the source separation problem. Preliminary results for transform optimization show the algorithm converging to the optimal transform, although global optimality is not guaranteed in general. In source separation the analogy enables the identification of a fairly broad family of distributions for which the optimality of KLT is guaranteed and numerical optimization algorithms are not needed. Moreover, we derived the optimal transform for subsequent dithered quantization. The optimal transform consists of KLT followed by a diagonal scaling matrix. For fixed rate coding, this transform is universally optimal (for all sources). In the case of variable rate coding, it is shown to be optimal for Gaussian sources.

6.1 Future Directions

- **Deterministic Annealing**: The proposed algorithm for finding optimal mappings does not guarantee a globally optimal solution, as a natural result of being a gradient descent approach. This problem can be largely mitigated by using more powerful optimization, in particular a deterministic annealing approach [63], which is left as future work.
- **Fundamental Limits of Zero-delay Communications**: In point-to-point source-channel communication with a fidelity criterion and a transmission cost constraint, the region of achievable cost and fidelity pairs is completely characterized by Shannons separation theorem, which in general only holds if arbitrarily high complexity and long delay are allowed. If the delay and/or complexity is constrained, the separation theorem only

provides an outer bound to the achievable cost/distortion region, and the exact shape of this region is in general not known. Recent research in this direction has appeared in [58, 41]. Calculation of tighter inner and/outer bounds for delay limited source channel coding would be one of the objectives of the future work.

- **Linearity of Optimal Prediction**: The analysis for the estimation problem can be extended to many important problems, including predictive coding. In predictive coding, linear prediction is usually employed due to complexity issues. Our approach can be (nontrivially) extended to this problem.

- **Transform Coding Extensions**: The basic ideas in optimal transform coding can be (nontrivially) extended to fixed rate coding, to distributed [19] and to multiple descriptions coding [28] scenarios, all of which are the subjects of ongoing investigation.

Bibliography

[1] E. Akyol and K. Rose. Nonuniform Dithered Quantization. In *Proceeding of Data Compression Conference*, page 435. IEEE, 2009.

[2] E. Akyol, K. Rose, and TA Ramstad. Optimal mappings for joint source channel coding. In *Proceedings of IEEE Information Theory Workshop*, 2010.

[3] E. Akyol, K. Rose, and TA Ramstad. Optimized analog mappings for distributed source channel coding. In *Proceedings of IEEE Data Compression Conference*, 2010.

[4] E. Akyol, K. Viswanatha, and K. Rose. On conditions for linearity of optimal estimation. In *Proceedings of IEEE Information Theory Workshop, Dublin*, 2010.

[5] V. Algazi and D. Sakrison. On the optimality of the Karhunen-Loève expansion. *IEEE Transactions on Information Theory,*, 15(2):319–321, 1969.

[6] HV Allen. A theorem concerning the linearity of regression. *Statistical Research Memoirs*, 2:60–68, 1938.

[7] C. Archer and TK Leen. A generalized Lloyd-type algorithm for adaptive transform coder design. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 52(1):255–264, 2004.

[8] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2003.

[9] A. Balakrishnan. On a characterization of processes for which optimal mean-square systems are of specified form. *IEEE Transactions on Information Theory*, 6(4):490–500, 1960.

[10] P. Billingsley. *Probability and Measure.* John Wiley & Sons Inc, 2008.

[11] S.Y. Chung. *On the construction of some capacity approaching coding schemes.* PhD thesis, Massachusetts Institute of Technology, 2000.

[12] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[13] T.M. Cover and J.A. Thomas. *Elements of information theory.* J.Wiley New York, 1991.

[14] R.M. Dudley. *Real Analysis and Probability.* Cambridge Univ Pr, 2002.

[15] M. Effros, H. Feng, and K. Zeger. Suboptimality of the Karhunen-Loeve transform for transform coding. *IEEE Transactions on Information Theory*, 50(8):1605–1619, 2004.

[16] T. Fine. Properties of an optimum digital system and applications. *IEEE Transactions on Information Theory*, 10(4):287–296, 1964.

[17] A. Fuldseth and TA Ramstad. Bandwidth compression for continuous amplitude channels based on vector approximation to a continuous subset of the source signal space. In *IEEE International Conference on Acoustics, Speech, and Signal Processing,*, volume 4, 1997.

[18] S. Gadkari and K. Rose. Robust vector quantizer design by noisy channel relaxation. *IEEE Transactions on Communications*, 47(8):1113–1116, 1999.

[19] M. Gastpar, P.L. Dragotti, and M. Vetterli. The distributed Karhunen-Loève transform. *IEEE Transactions on Information Theory*, 52(12):5177–5196, 2006.

[20] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression.* Springer, 1992.

[21] S.G. Ghurye and I. Olkin. A characterization of the multivariate normal distribution. *The Annals of Mathematical Statistics*, pages 533–541, 1962.

[22] J.D. Gibson and T. Fischer. Alphabet-constrained data compression. *IEEE Transactions on Information Theory*, 28(3):443–457, 1982.

[23] H. Gish and J. Pierce. Asymptotically efficient quantizing. *IEEE Transactions on Information Theory,*, 14(5):676–683, 1968.

[24] T. Goblick Jr. Theoretical limitations on the transmission of data from analog sources. *IEEE Transactions on Information Theory*, 11(4):558–567, 1965.

[25] G.H. Golub and C.F. Van Loan. *Matrix computations.* Johns Hopkins Univ Pr, 1996.

[26] V. Goyal, J. Zhuang, and M. Vetterli. Transform coding with backward adaptive updates. *IEEE Transactions on Information Theory*, 46(4):1623–1633, 2000.

[27] V.K. Goyal. Theoretical foundations of transform coding. *Signal Processing Magazine, IEEE*, 18(5):9–21, 2002.

[28] VK Goyal and J. Kovacevic. Generalized multiple description coding with correlating transforms. *Information Theory, IEEE Transactions on*, 47(6):2199–2224, 2001.

[29] RM Gray and TG Stockham Jr. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.

[30] R.M. Gray and T.G. Stockham Jr. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.

[31] O.G. Guleryuz and M.T. Orchard. On the DPCM compression of Gaussian autoregressive sequences. *IEEE Transactions on Information Theory*, 47(3):945–956, 2001.

[32] C.D. Hardin. On the linearity of regression. *Probability Theory and Related Fields*, 61(3):293–302, 1982.

[33] F. Hekland, P.A. Floor, and T.A. Ramstad. Shannon-Kotelnikov mappings in joint source-channel coding. *IEEE Transactions on Communications*, 57(1):94–105, 2009.

[34] F. Hekland, GE Oien, and TA Ramstad. Using 2: 1 Shannon mapping for joint source-channel coding. In *Proceedings of the IEEE Data Compression Conference*, pages 223–232, 2005.

[35] A. Hjorungnes. *Optimal Bit and Power Constrained Filter Banks.* PhD thesis, Norwegian University of Science and Technology, 2000.

[36] R.A. Horn and C.R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[37] Y. Hu, J. Garcia-Frias, and M. Lamarca. MMSE decoding for analog joint source channel coding using Monte Carlo importance sampling. In *Proc.IEEE Workshop on Signal Processing Advances in Wireless Communications*, pages 682–686. IEEE, 2009.

[38] J. Huang and P. Schultheiss. Block quantization of correlated Gaussian random variables. *IEEE Transactions on Communications*, 11(3):289–296, 1963.

[39] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks,*, 10(3):626–634, 2002.

[40] A. Hyvarinen and E. Oja. *Independent component analysis: algorithms and applications*, volume 13. Elsevier, 2000.

[41] A. Ingber, I. Leibowitz, R. Zamir, and M. Feder. Distortion lower bounds for finite dimensional joint source-channel coding. In *Proceedings of IEEE International Symposium on Information Theory*, pages 1183–1187, 2008.

[42] S. Jana and P. Moulin. Optimality of KLT for High-Rate Transform Coding of Gaussian Vector-Scale Mixtures: Application to Reconstruction, Estimation, and Classification. *IEEE Transactions on Information Theory*, 52(9):4049–4067, 2006.

[43] S.M. Kay. *Fundamentals of Statistical Signal Processing.* Prentice Hall PTR, 1993.

[44] M.N. Khormuji and M. Skoglund. On instantaneous relaying. *IEEE Transactions on Information Theory,*, 56(7):3378–3394, 2010.

[45] P. Knagenhjelm. A recursive design method for robust vector quantization. In *Proc. Int. Conf. Signal Processing Applications and Technology*, pages 948–954, 1992.

[46] VA Kotelnikov. *The theory of optimum noise immunity.* New York: McGraw-Hill, 1959.

[47] RG Laha. On a characterization of the stable law with finite expectation. *The Annals of Mathematical Statistics*, 27(1):187–195, 1956.

[48] E.G. Learned-Miller and W.F. John III. ICA using spacings estimates of entropy. *The Journal of Machine Learning Research*, 4:1271–1295, 2003.

[49] K.H. Lee and D. Petersen. Optimal linear coding for vector channels. *IEEE Transactions on Communications*, 24(12):1283–1290, 1976.

[50] T.W. Lee, M. Girolami, and T.J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[51] D.G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons Inc, 1969.

[52] E. Lukacs. Characteristics Functions. *Charles Griffin and Company*, 1960.

[53] AW Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications* . Academic Press, New York, 1979.

[54] M.K. Mihcak, P. Moulin, M. Anitescu, and K. Ramchandran. Rate-distortion-optimal subband coding without perfect-reconstruction constraints. *IEEE Transactions on Signal Processing*, 49(3):542–557, 2001.

[55] U. Mittal and N. Phamdo. Hybrid digital-analog (HDA) joint source-channel codes for broadcasting and robust communications. *IEEE Transactions on Information Theory*, 48(5):1082–1102, 2002.

[56] D.T. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Transactions on Information Theory*, 48(7):1935–1946, 2002.

[57] D.T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing,*, 52(10):2690–2700, 2004.

[58] Y. Polyanskiy, H.V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *Information Theory, IEEE Transactions on*, 56(5):2307–2359, 2010.

[59] T.A. Ramstad. Shannon mappings for robust communication. *Telektronikk*, 98(1):114–128, 2002.

[60] C.R. Rao. Note on a problem of Ragnar Frisch. *Econometrica, Journal of the Econometric Society*, 15(3):245–249, 1947.

[61] C.R. Rao. On some characterisations of the normal law. *Sankhyā: The Indian Journal of Statistics, Series A*, 29(1):1–14, 1967.

[62] MM Rao and RJ Swift. *Probability Theory with Applications*. Springer, 2005.

[63] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

[64] C. Rothschild and E. Mourier. Sur les lois de probabilité à regression linéaire et écart type lié constant. *Comptes Rendus*, 225:245–249, 1947.

[65] L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communications*, 12(4):162–165, 1964.

[66] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(1):379–423, 1948.

[67] CE Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[68] S. Sherman. Non-mean-square error criteria. *IEEE Transactions on Information Theory,*, 4(3):125–126, 1958.

[69] JA Shohat and J.D. Tamarkin. The Problem of Moments. *New York*, 1943.

[70] V.P. Skitovic. Linear combinations of independent random variables and the normal distribution law. *Selected Translations in Mathematical Statistics and Probability*, page 211, 1962.

[71] M. Skoglund, N. Phamdo, and F. Alajaji. Hybrid digital-analog source-channel coding for bandwidth compression/expansion. *IEEE Trans. Inf. Theory*, 52(8):3757–3763, 2006.

[72] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, 1973.

[73] F.W. Steutel and K. Van Harn. *Infinite divisibility of probability distributions on the real line.* CRC, 2003.

[74] MD Trott. Unequal error protection codes: Theory and practice. In *Proc. IEEE Information Theory Workshop*, page 11, 1996.

[75] A.B. Wagner, S. Tavildar, and P. Viswanath. Rate region of the quadratic Gaussian two-encoder source-coding problem. *IEEE Transactions on Information Theory*, 54(5), 2008.

[76] N. Wernersson, J. Karlsson, and M. Skoglund. Distributed quantization over noisy channels. *IEEE Transactions on Communications*, 57(6):1693–1700, 2009.

[77] N. Wernersson and M. Skoglund. Nonlinear coding and estimation for correlated data in wireless sensor networks. *IEEE Transactions on Communications*, 57(10):2932–2939, 2009.

[78] N. Wernersson, M. Skoglund, and T. Ramstad. Polynomial based analog source channel codes. *IEEE Transactions on Communications,*, 57(9):2600–2606, 2009.

[79] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

[80] R. Zamir and M. Feder. On universal quantization by randomized uniform/lattice quantizers. *IEEE Transactions on Information Theory*, 38(2 Part 2):428–436, 1992.

[81] R. Zamir and M. Feder. Information rates of pre/post-filtered dithered quantizers. *Information Theory, IEEE Transactions on*, 42(5):1340–1353, 1996.

[82] R. Zamir and M. Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, 1996.

[83] J. Ziv. The behavior of analog communication systems. *IEEE Transactions on Information Theory,*, 16(5):587–594, 1970.

[84] J. Ziv. On universal quantization. *IEEE Transactions on Information Theory*, 31(3):344–347, 1985.