# Quantized Overcomplete Expansions: Analysis, Synthesis and Algorithms

by

Vivek K Goyal

vkgoyal@eecs.berkeley.edu



Memorandum No. UCB/ERL M95/97

1 July 1995

ELECTRONICS RESEARCH LABORATORY
College of Engineering
University of California, Berkeley
Berkeley, CA 94720

# Acknowledgements

First and foremost, I would like to thank my advisor, Professor Martin Vetterli, for sharing his ideas and his infectious enthusiasm. I would also like to thank Ton Kalker for providing some of the building blocks for my MATLAB simulations. The technical content has been influenced by conversations with Chris Chan, Grace Chang, Philip Chou, Zoran Cvetković, Michelle Effros, Michael Goodwin, Masoud Khansari, Steve McCanne, Alan Park, Nguyen Thao, and Bin Yu. Extensive, valuable comments on preliminary versions of this report were provided by Alyson Fletcher, Michael Goodwin and Martin Vetterli.

During the completion of this work, financial support was provided by ARPA through a National Defense Science and Engineering Graduate Fellowship. Less tangible support was provided in various manners and to varying degrees by the individuals listed below.

```
B M B T P S R M A S O U D N C H R I S S N T P F V
O W H A R A L P H I L T O A L L I E D E D R P D A
S H R O M I L A A D J J N J A M Y X V Z G U P I W
U J A B H L Y U Y X K E A A I A Q N I A H N M U W
N R H X E M E L I S S A L X R R O J T F H C R C R
D C O F Q T M B B S A N D R E K C C L Y N N G L U
E Q W V D P Q H B T P T O J L O G C I E I C R X H
E P A T R I C K A E U O J E E K R U S T Y A Y U Q
P G R H O F M A R V I N G R A C E V A N I T A C T
T F D U C K M A N L V I M O N T G O M E R Y I U J
F P R B O N N I E A C O S M O E M X H J V X K N O
R T M I K E V C Y N Y A J E R R Y X J E F F Q I A
J H O E I D A F H D S P U E L A U R A S B K Z R F
M A R G E O R G E L M N G R N Y D L M S R J W D F
A O T T O M U J A K E E L A I N E X E I R A I O T
Y Z O R A N N E T Z H L S H O B B E S L V M I P N
A V N K X M I L H O U S T E V E B O J I E A P I D
N Y H U D U S R E Y N O L D X B I L L L O G G N E
K S Z Z X G J W R A K N K A K E N D A L L G Y C J
A I V M M S O M A D A N P B A R T T N I M I S H D
M C A A A Y H O M E R I C O R N F E D A V E B A K
L M A R T I N A U I L C A L V I N R O N P M T S B
M C R Y T R E V O R Y H T E U C L I D K Z I A X H
D Z N Z T C L J R T C H H J O E L W O O D L I Y W
K G Q K G J Z Q X W L F Y T H O O M A P J Y J O Y
```

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Overview

Linear transforms and expansions are the fundamental mathematical tools of signal processing. Yet the properties of linear expansions in the presence of coefficient quantization are not yet fully understood. These properties are most interesting when signal representations are with respect to redundant, or overcomplete, sets of vectors. Exploring the effects of quantization in overcomplete linear expansions is the unifying theme of this work.

The core problem of Chapter 2 is depicted in Figure 1.1. A vector $x \in \mathbb{R}^N$ is left multiplied by a matrix $F$ to get $y \in \mathbb{R}^M$. For $M > N$, we have an overcomplete expansion. The problem is to estimate $x$ from a scalar quantized version of $y$. To put this in a solid framework, we introduce the concept of frames and prove some properties of frames. We then show that the quality of reconstruction can be improved by using deterministic properties of quantization, as opposed to considering quantization to be the addition of noise that is independent in each dimension.

In Chapter 3, focus shifts to the problem of compression, *i.e.* finding efficient representations. Vector quantization and transform coding are the standard methods used in signal compression. Vector quantization gives better rate-distortion performance, but it is difficult to implement and is computationally expensive. The computational aspects make transform coding very attractive. For this reason, transform coding is ubiquitous in image compression.

For fine quantization of a Gaussian signal with known statistics, the Karhunen-Loève transform (KLT) is optimal for transform coding [13]. In general, signal statistics are changing or not known *a priori*. Thus, one must either estimate the KLT from finite length blocks of the signal or use a fixed, signal-independent transform. The former case is computationally intensive and transmission of the KLT coefficients can be prohibitively expensive.[1] The latter option is most commonly used, often with the discrete cosine transform (DCT). As with any fixed transform, the DCT is nearly optimal for only a certain set of possible signals. There has been considerable work in the area of adaptively choosing a transform from a library of orthogonal transforms, for example, using wavelet packets [29].

All varieties of transform coding represent a signal vector as a linear combination of

[1] In practical adaptive transform coding system, 20 to 40 percent of the available bit rate is assigned to side information [21, §2.3].

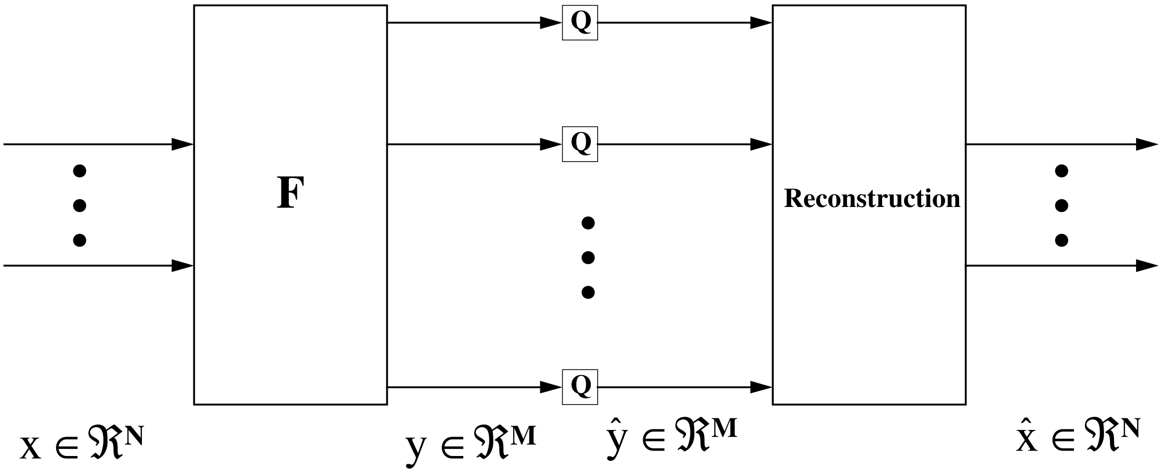



Figure 1.1: Block diagram of reconstruction from quantized frame expansion.

basis vectors. Notice that in Figure 1.1, $y$ is a representation of $x$ in terms of the rows of $F$. But $\hat{y}$ is generally not an efficient representation of $x$. A method that adaptively chooses a basis set from a finite dictionary given a signal vector is presented in Chapter 3. The representation is generated through a greedy successive approximation algorithm called matching pursuit. Much as the KLT finds the best representation "on average," this method finds a good representation for the particular vector being coded. Since it does not depend on distributional knowledge, matching pursuit can be viewed as a "universal transform" for transform coding.[2]

Some of the results of this report appeared earlier in [14].

[2]The phrase "universal lossy coder" is avoided because we assume a separation into a transform, followed by scalar quantization and universal lossless coding. This separation is not necessarily optimal but is motivated by complexity considerations.

## 1.2 Notation

The notation used throughout the report is summarized in Table 1.1 below.

| Symbol | Definition | Reference |
|---|---|---|
| $\bar{\ }$ | Conjugation | |
| $^*$ | Conjugate transpose | |
| $\lvert\cdot\rvert$ | Cardinality of a set | |
| $\langle\cdot,\cdot\rangle$ | Inner product; for finite dimensional vectors, $\langle x, y\rangle = x^T\bar{y}$ | |
| $\lVert\cdot\rVert$ | Norm (derived from inner product through $\lVert x\rVert = \langle x, x\rangle^{1/2}$) | |
| $\mathrm{Ran}(\cdot)$ | Range of an operator | |
| $\alpha_k$ | A coefficient in a linear expansion | §3.1, §3.2.1 |
| $\Lambda$ | A lattice | Appendix C |
| $\Phi$ | A frame in $H$ | §2.1.1 |
| $\widetilde{\Phi}$ | The dual frame of $\Phi$ | §2.2.1 |
| $\varphi_k$ | An element of $\Phi$ | §2.1.1 |
| $\widetilde{\varphi_k}$ | An element of $\widetilde{\Phi}$ | §2.2.1 |
| $A$ | Lower frame bound | §2.1.1 |
| $B$ | Upper frame bound | §2.1.1 |
| $\mathbb{C}$ | Complex numbers | |
| $\mathcal{D}$ | A dictionary in an adaptive expansion | §3.1, §3.2.1 |
| $E[\cdot]$ | Expectation operator | |
| $F$ | Frame operator associated with $\Phi$ | §2.1.1 |
| $H$ | Hilbert space $\mathbb{R}^N$ or $\mathbb{C}^N$ | |
| $I_n$ | $n\times n$ identity matrix ($n$ is omitted if it is clear from context) | |
| $j$ | $\sqrt{-1}$ | |
| $K$ | A countable index set | |
| $L_2(\mathbb{R})$ | Space of square-integrable functions over $\mathbb{R}$ | §2.1.2 |
| $\ell_2(K)$ | Space of square-summable sequences indexed by K | §2.1.1 |
| $M$ | Cardinality of $\Phi$ or $\mathcal{D}$ | §2.1.1, §3.2.1 |
| $N$ | Dimension of $H$ | |
| $\mathcal{N}(\mu,\Lambda)$ | Normal distribution with mean $\mu$ and covariance matrix $\Lambda$ | |
| $\mathbb{R}$ | Real numbers | |
| $r$ | $M/N$, the redundancy of $\Phi$ or $\mathcal{D}$ | §2.1.1 |
| $\mathbb{Z}$ | Integers | |
| $\mathbb{Z}^+$ | Positive integers | |
| $\blacksquare$ | End of a proof | |
| $\square$ | End of an example | |

Table 1.1: Summary of notation

# Chapter 2

# Non-adaptive Expansions

Orthogonal transforms are ubiquitous in mathematics, science, and engineering. The basis functions used in these transforms do not depend on the particular signal being analyzed and hence the resulting expansions can be considered non-adaptive.

For electrical engineers, frequency domain techniques based on Fourier transforms and Fourier series are second-nature. This chapter describes frames, which provide a general framework for understanding non-orthogonal transforms. Frames were introduced by Duffin and Schaeffer [10] in the context of non-harmonic Fourier series. Recent interest in frames has been spurred by its utility in analyzing discrete wavelet transforms [5, 6, 15] and time-frequency decompositions [22]. We are motivated by a desire to understand quantization effects and efficient representations in a general framework.

To put this chapter in context, we will give a particular interpretation of Fourier analysis and discuss a sense in which it can be generalized. Since we are limiting our attention to finite dimensional spaces, consider the Discrete Fourier Transform (DFT) of a length-$N$ sequence $x[n]$. We can interpret the DFT as giving a set of $N$ coefficients[1]

$$X[k] = \sum_{n=0}^{N-1} \frac{1}{\sqrt{N}} x[n] e^{-j2\pi kn/N} = \left\langle x[n], \frac{1}{\sqrt{N}} e^{j2\pi kn/N} \right\rangle, \quad \text{for } 0 \leq k \leq N-1. \tag{2.1}$$

Then the original sequence can be reconstructed as

$$x[n] = \sum_{k=0}^{N-1} \frac{1}{\sqrt{N}} X[k] e^{j2\pi kn/N} = \sum_{k=0}^{N-1} \left\langle x[n], \frac{1}{\sqrt{N}} e^{j2\pi kn/N} \right\rangle \frac{1}{\sqrt{N}} e^{j2\pi kn/N}, \quad \text{for } 0 \leq n \leq N-1. \tag{2.2}$$

In this manner, the DFT gives a linear expansion of a vector in terms of the set of vectors

$$\left\{ \begin{bmatrix} \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} e^{j2\pi k/N} & \cdots & \frac{1}{\sqrt{N}} e^{j2\pi k(N-1)/N} \end{bmatrix}^T \right\}_{k=0}^{N-1}, \tag{2.3}$$

where the coefficients in the expansion are formed by taking inner products with the same set. Similar expansions can be found by replacing (2.3) by other sets of vectors. Note that

[1] The $\frac{1}{N}$ term that generally appears in the inverse DFT formula has been distributed between the DFT and the inverse DFT. This is gives unit-norm basis vectors for analysis and synthesis.

the set need not have only $N$ elements and that the sets used in analysis (2.1) and synthesis (2.2) may be different. We will see in §2.2.1 that the analysis and synthesis sets must be dual frames.

Section 2.1 begins with definitions that will be used throughout the chapter and examples of frames. It concludes with a theorem on the tightness of random frames and a discussion of that result. Section 2.2 begins with a review of reconstruction from exactly known frame coefficients. The remainder of the section gives new results on reconstruction from quantized frame coefficients. Most previous work on frame expansions is predicated either on exact knowledge of coefficients or on coefficient degradation by white additive noise. For example, Munch [22] considered a particular type of frame and assumed the coefficients were subject to a stationary noise. This report, on the other hand, is in the same spirit as [4, 32, 33, 35] in that it utilizes the deterministic qualities of quantization.

## 2.1 Frames

### 2.1.1 Definitions and Basics

The material in this subsection is largely adapted from [6, Ch. 3]. We are limiting our attention to Hilbert spaces $H$ of dimension $N$.

DEFINITION. Let $\Phi = \{\varphi_k\}_{k \in K} \subset H$, where $K$ is a countable index set. $\Phi$ is called a *frame* if there exist $A > 0$ and $B < \infty$ such that for all $f \in H$,

$$A\|f\|^2 \leq \sum_{k \in K} |\langle f, \varphi_k \rangle|^2 \leq B\|f\|^2. \tag{2.4}$$

$A$ and $B$ are called the *frame bounds.*

Throughout we will denote $|K|$, the cardinality of $K$, by $M$ and allow $M = \infty$. The lower bound in (2.4) is equivalent to requiring that $\Phi$ span $H$. Thus a frame will always have $M \geq N$. We will refer to $r = \frac{M}{N}$ as the *redundancy* of the frame. Also notice that one can choose $B = \sum_{k \in K} \|\varphi_k\|^2$ whenever $M < \infty$.

DEFINITION. Let $\Phi$ be a frame in $H$. $\Phi$ is called a *tight frame* if the frame bounds can be taken to be equal.

It is easy to verify that if $\Phi$ is a tight frame with $\|\varphi_k\| = 1$ for all $k \in K$, then $A = r$.

PROPOSITION 2.1. Let $\Phi = \{\varphi_k\}_{k \in K}$ be a tight frame with frame bounds $A = B = 1$. If $\|\varphi_k\| = 1$ for all $k \in K$, then $\Phi$ is an orthonormal basis.
PROOF: See §A.2. ■

DEFINITION. Let $\Phi = \{\varphi_k\}_{k \in K}$ be a frame in $H$. The *frame operator* $F$ is the linear operator from $H$ to $\mathbb{C}^M$ defined by[2]

$$(Ff)_k = \langle f, \varphi_k \rangle. \tag{2.5}$$

[2] We should denote the codomain by $\ell_2(K)$ to properly include the case $M = \infty$; however, for notational simplicity we will not.

Note that when $H$ is finite dimensional, this operation is a matrix multiplication where $F$ is a matrix with $k$th row equal to $\varphi_k^*$. Using the frame operator, (2.4) can be rewritten as

$$AI_N \leq F^*F \leq BI_N, \tag{2.6}$$

where $I_N$ is the $N \times N$ identity matrix. (The matrix inequality $AI_N \leq F^*F$ means that $F^*F - AI_N$ is a positive semidefinite matrix.) In this notation, $F^*F = AI_N$ implies that $\Phi$ is a tight frame.

From (2.6) we can immediately conclude that the eigenvalues of $F^*F$ lie in the interval $[A, B]$; in the tight frame case, all of the eigenvalues are equal. This gives a computational procedure for finding frame bounds. Since it is conventional to assume $A$ is chosen as large as possible and $B$ is chosen as small as possible, we will sometimes take the minimum and maximum eigenvalues of $F^*F$ to be the frame bounds. Note that it also follows from (2.6) that $F^*F$ is invertible because all of its eigenvalues are nonzero.

Let $\Phi = \{\varphi_k\}_{k\in K}$ be a frame in $H$. Since $\mathrm{Span}(\Phi) = H$, any vector $f \in H$ can be written as

$$f = \sum_{k\in K} \alpha_k \varphi_k \tag{2.7}$$

for some set of coefficients $\{\alpha_k\} \subset \mathbb{R}$. If $M > N$, $\{\alpha_k\}$ may not be unique. We refer to (2.7) as a *redundant representation* even though it is not necessary that more than $N$ of the $\alpha_k$'s be nonzero.

### 2.1.2 Examples

The question of whether a set of vectors form a frame is not very interesting in a finite-dimensional space; any finite set of vectors which span the space form a frame. Thus if $M \geq N$ vectors are chosen randomly with a circularly symmetric distribution on $H$, they almost surely form a frame. An infinite set in a finite-dimensional space can form a frame only if the norms of the elements decay appropriately, for otherwise a finite upper frame bound will not exist.

Heuristically, we expect tight frames to have a certain degree of uniformity or regularity. This is illustrated by the following examples.

Example 1 [6]. In $H = \mathbb{R}^2$, let $\varphi_1 = [0 \;\; 1]^T$, $\varphi_2 = [-\frac{\sqrt{3}}{2} \;\; -\frac{1}{2}]^T$, and $\varphi_3 = [\frac{\sqrt{3}}{2} \;\; -\frac{1}{2}]^T$. These are vectors on the unit circle uniformly spaced by 120°. For any $f = [f_1 \;\; f_2]^T \in H$,

$$\begin{aligned}\sum_{k=1}^{3} |\langle f, \varphi_k\rangle|^2 &= |f_2|^2 + \left|-\frac{\sqrt{3}}{2}f_1 - \frac{1}{2}f_2\right|^2 + \left|\frac{\sqrt{3}}{2}f_1 - \frac{1}{2}f_2\right|^2 \\ &= \frac{3}{2}\left[f_1^2 + f_2^2\right] \;=\; \frac{3}{2}\|f\|^2.\end{aligned}$$

Thus $\{\varphi_1, \varphi_2, \varphi_3\}$ is a tight frame with frame bound $\frac{3}{2} = \frac{M}{N}$. □

Example 2 [37]. Consider the space of continuous-time signals that are bandlimited to $[-\pi, \pi]$. This is a subspace of the Hilbert space $L_2(\mathbb{R})$. By the Nyquist Sampling Theorem

[26, §3.2], $S_1 = \{\mathrm{sinc}(t-k)\}_{k\in\mathbb{Z}}$, where

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t},$$

forms a basis for this space. Notice that $S_1$ is the basis set for ideal $\pi$-bandlimited interpolation. For $n \in \mathbb{Z}^+$, the set

$$S_n = \left\{\mathrm{sinc}(t - \frac{k}{n})\right\}_{k\in\mathbb{Z}}$$

forms a tight frame with redundancy $n$. An expansion with respect to $S_n$ corresponds to sampling at $n$ times the Nyquist rate. □

Example 3. Oversampling of a periodic, bandlimited signal can be viewed as a frame operator applied to the signal, where the frame operator is associated with a tight frame. If the samples are quantized, this is exactly the situation of oversampled A/D conversion [33]. Let $x = [X_1 \ \ X_2 \ \ \cdots \ \ X_N]^T \in \mathbb{R}^N$, with $N$ odd. Define a corresponding continuous-time signal by

$$x_c(t) = X_1 + \sum_{k=1}^{W}\left[X_{2k}\sqrt{2}\cos\frac{2\pi kt}{T} + X_{2k+1}\sqrt{2}\sin\frac{2\pi kt}{T}\right], \tag{2.8}$$

where $W = \frac{N-1}{2}$. Any real-valued, $T$-periodic, bandlimited, continuous-time signal can be written in this form. Let $M \geq N$. Define a sampled version of $x_c(t)$ by $x_d[m] = x_c(\frac{mT}{M})$ and let

$$y = \begin{bmatrix}x_d[0] & x_d[1] & \cdots & x_d[M-1]\end{bmatrix}^T.$$

Then we have $y = Fx$, where

$$F = \begin{bmatrix} 1 & \sqrt{2} & 0 & \cdots & \sqrt{2} & 0 \\ 1 & \sqrt{2}\cos\theta & \sqrt{2}\sin\theta & \cdots & \sqrt{2}\cos W\theta & \sqrt{2}\sin W\theta \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \sqrt{2}\cos(M'\theta) & \sqrt{2}\sin(M'\theta) & \cdots & \sqrt{2}\cos(WM'\theta) & \sqrt{2}\sin(WM'\theta) \end{bmatrix}, \tag{2.9}$$

$M' = M-1$, and $\theta = \frac{2\pi}{M}$. Using the orthogonality properties of sine and cosine, it is easy to check that $F^*F = MI_N$, so $F$ is an operator associated with a tight frame. Pairing terms and using the identity $\cos^2 k\theta + \sin^2 k\theta = 1$, we find that each row of $F$ has norm $\sqrt{N}$. Dividing $F$ by $\sqrt{N}$ normalizes the frame and results in a frame bound equal to the redundancy ratio $r$. Also note that $r$ is the oversampling ratio with respect to the Nyquist sampling frequency. □

### 2.1.3 Tightness of Random Frames

Tight frames constitute an important class of frames. As we will see in §2.2.1, a tight frame is self-dual and hence has some desirable reconstruction properties. These reconstruction properties indeed extend smoothly to nearly tight frames, *i.e.* frames with $\frac{B}{A}$ close to one. Also, for a tight frame (2.4) reduces to something similar to Parseval's equality. Thus, a tight frame operator scales the energy of an input by a constant factor $A$. Furthermore, it is shown in §2.2.4 that some properties of "typical" frame operators depend only on the redundancy. This motivates our interest in the following theorem.

THEOREM 2.2: TIGHTNESS OF RANDOM FRAMES
Let $\{\Phi_M\}_{M=N}^{\infty}$ be a sequence of frames in $\mathbb{R}^N$ such that $\Phi_M$ is generated by choosing $M$ vectors independently with a uniform distribution on the unit sphere in $\mathbb{R}^N$. Let $F_M$ be the frame operator associated with $\Phi_M$. Then, in the mean squared sense,

$$\frac{1}{M}{F_M}^*F_M \longrightarrow \frac{1}{N}I_N \text{ elementwise as } M \longrightarrow \infty.$$

PROOF: See §A.3. ■

Theorem 2.2 shows that a sequence of random frames with increasing redundancy will approach a tight frame. Note that although the proof in Appendix A uses an unrelated strategy, the constant $1/N$ is intuitive: If $\Phi_M$ is a tight frame with normalized elements, then we have ${F_M}^*F_M = \frac{M}{N}I_N$ because the frame bound equals the redundancy of the frame.

Numerical experiments were performed to confirm this behavior and observe the rate of convergence. Sequences of frames were generated by successively adding random vectors (chosen according to the appropriate distribution) to existing frames. Results shown in Figures 2.1 and 2.2 are averaged results for 200 sequences of frames in $\mathbb{R}^4$. Figure 2.1 shows that $\frac{A}{M}$ and $\frac{B}{M}$ converge to $\frac{1}{N}$. Figure 2.2 shows that $\frac{B}{A}$ converges to one.

In Theorem 2.2, the uniformity of the frame elements over the unit sphere is a necessary condition. This is illustrated by the following example.

EXAMPLE 4. Suppose sequences of frames in $\mathbb{R}^2$ is generated by choosing vectors $\varphi_k = [\cos\theta \;\; \sin\theta]^T$, where $\theta$ is uniformly distributed on $[0, \frac{\pi}{2}]$. Then

$$\frac{1}{M}{F_M}^*F_M \to \begin{bmatrix} \frac{1}{2} & \frac{1}{2\pi} \\ \frac{1}{2\pi} & \frac{1}{2} \end{bmatrix} \text{ elementwise as } M \to \infty.$$

Thus the sequence of frames does *not* approach a tight frame. We can make a few additional observations. The eigenvalues of

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2\pi} \\ \frac{1}{2\pi} & \frac{1}{2} \end{bmatrix}$$

are $\lambda_1 = \frac{1}{2}(1+\frac{1}{\pi})$ and $\lambda_2 = \frac{1}{2}(1-\frac{1}{\pi})$, with corresponding eigenvectors $f_1 = \frac{1}{\sqrt{2}}[1 \;\; 1]^T$ and $f_2 = \frac{1}{\sqrt{2}}[1 \;\; -1]^T$, respectively. The eigenvectors $f_1$ and $f_2$ are the vectors that maximize and minimize, respectively,

$$E\left[\sum_{k=1}^{M} |\langle f, \varphi_k\rangle|^2\right]$$

over all unit-norm $f$. This example reinforces the notion that tightness of frames corresponds to directional uniformity. □

## 2.2 Reconstruction from Frame Coefficients

At this point, the usage of frames in signal analysis is not yet justified because we have not considered the problem of reconstructing from frame coefficients.

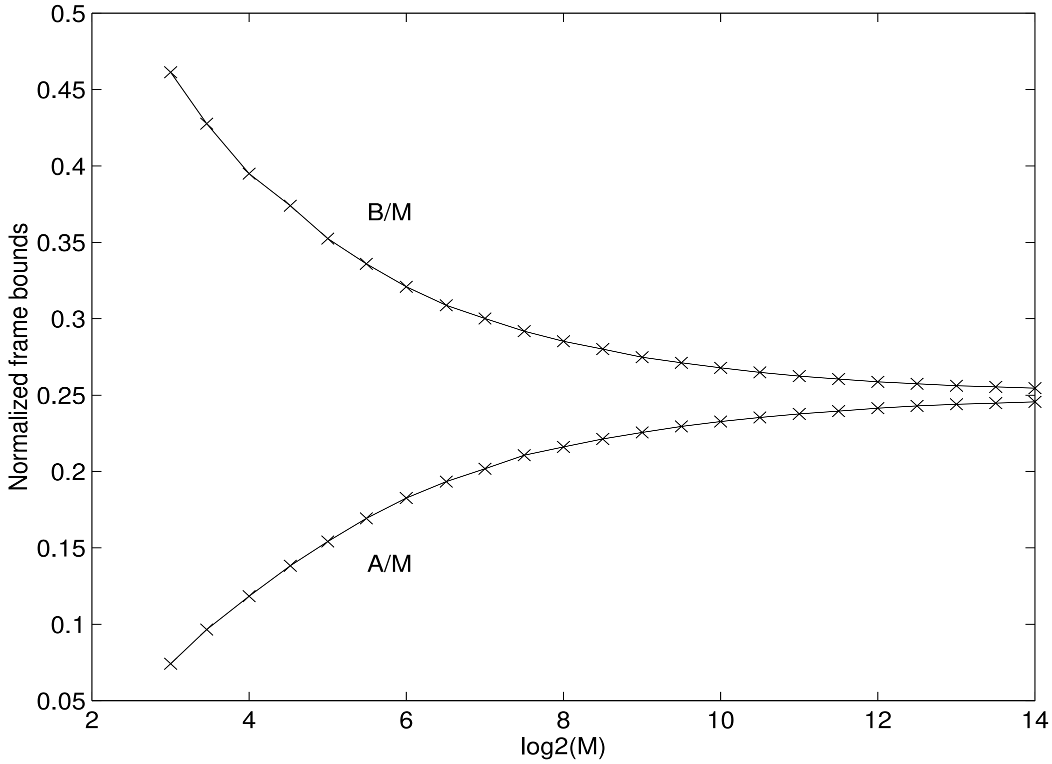



Figure 2.1: Normalized frame bounds for random frames in $\mathbb{R}^4$.

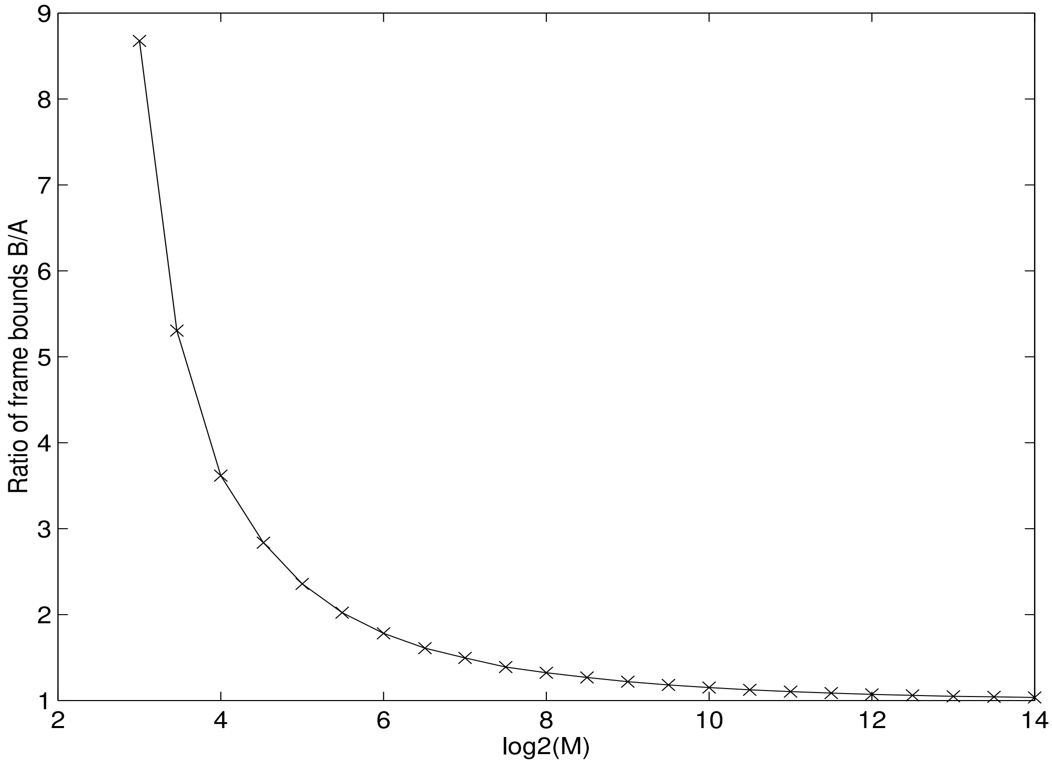



Figure 2.2: Ratios of frame bounds for random frames in $\mathbb{R}^4$.

In §2.2.1, we review the basic properties of reconstructing from (unquantized) frame coefficients. This material is adapted from [6]. The subsequent sections consider the problem of reconstructing an estimate of an original signal from quantized frame coefficients. Classical methods are limited by the assumption that the quantization noise is white. Our approach uses deterministic qualities of quantization to arrive at the concept of consistent reconstruction. Consistent reconstruction methods yield smaller reconstruction errors than classical methods.

### 2.2.1 Unquantized Case

Let $\Phi$ be a frame, assuming the notation of §2.1.1. In this subsection, we consider the problem of recovering $f$ from $\{\langle f, \varphi_k\rangle\}_{k\in K}$.

Recall that $F^*F$ is invertible. We can say furthermore that

$$B^{-1}I_N \leq (F^*F)^{-1} \leq A^{-1}I_N. \tag{2.10}$$

DEFINITION. The *dual frame* of $\Phi$ is $\widetilde{\Phi} = \{\widetilde{\varphi_k}\}_{k\in K}$, where

$$\widetilde{\varphi_k} = (F^*F)^{-1}\varphi_k, \quad \forall\, k \in K. \tag{2.11}$$

For a tight frame, (2.11) simplifies to

$$\widetilde{\varphi_k} = A^{-1}\varphi_k, \quad \forall\, k \in K. \tag{2.12}$$

PROPOSITION 2.3. $\widetilde{\Phi}$ is a frame with frame bounds $B^{-1}$ and $A^{-1}$, *i.e.*

$$B^{-1}\|f\|^2 \leq \sum_{k\in K} |\langle f, \widetilde{\varphi_k}\rangle|^2 \leq A^{-1}\|f\|^2.$$

The associated frame operator $\widetilde{F} : H \to \mathbb{C}^M$ satisfies $\widetilde{F} = F(F^*F)^{-1}$, $\widetilde{F}^*\widetilde{F} = (F^*F)^{-1}$, and $\widetilde{F}^*F = I_N = F^*\widetilde{F}$. Also, $\widetilde{F}F^* = F\widetilde{F}^*$ is the orthogonal projection operator, in $C^M$, onto $\mathrm{Ran}(F) = \mathrm{Ran}(\widetilde{F})$.
PROOF: See [6, p. 59]. ■

A consequence of $\widetilde{F}^*\widetilde{F} = (F^*F)^{-1}$ is that the dual of $\widetilde{\Phi}$ is $\Phi$. Another of the conclusions of Proposition 2.3 gives us the desired reconstruction formula: Namely, $\widetilde{F}^*F = I_N$ implies

$$f = \widetilde{F}^*Ff = \sum_{k\in K}\langle f, \varphi_k\rangle\widetilde{\varphi_k}. \tag{2.13}$$

This formula is reminiscent of (2.2). The difference is that in (2.2), one set of vectors plays the roles of both $\Phi$ and $\widetilde{\Phi}$. This is because the set in (2.3) is a tight frame in $\mathbb{C}^N$. In analogy to (2.13), since $F^*\widetilde{F} = I_N$, we can also write

$$f = F^*\widetilde{F}f = \sum_{k\in K}\langle f, \widetilde{\varphi_k}\rangle\varphi_k. \tag{2.14}$$

Comparing (2.13) and (2.14) emphasizes the "dual" nature of $\Phi$ and $\widetilde{\Phi}$.

The derivation of (2.14) obscures the fact that, when $M > N$, there should be many ways to write $f$ as a linear combination of vectors from $\Phi$. After all, there is an $N$-element subset of $\Phi$ that spans $H$. What is special about the expansion in (2.14)? This question is partially answered by the following proposition.

PROPOSITION 2.4. If $f = \sum_{k\in K} c_k\varphi_k$ for some set of coefficients $\{c_k\}_{k\in K}$, then

$$\sum_{k\in K} |c_k|^2 \geq \sum_{k\in K} |\langle f, \widetilde{\varphi_k}\rangle|^2, \tag{2.15}$$

with equality only if $c_k = \langle f, \widetilde{\varphi_k}\rangle$ for all $k \in K$.
PROOF: See [6, p. 61]. ∎

The norm-minimizing property of (2.15) holds in the "dual" sense also: If

$$f = \sum_{k\in K} \langle f, \varphi_k\rangle u_k,$$

then

$$\sum_{k\in K} |\langle u_k, g\rangle|^2 \geq \sum_{k\in K} |\langle \widetilde{\varphi_k}, g\rangle|^2$$

for all $g \in H$. Also, using (2.13) has advantages over other possible reconstruction formulas when the frame coefficients are not known exactly (see §2.2.2).

Sometimes we can reconstruct, or approximately reconstruct, without explicitly finding the dual frame through (2.11). For example, if $\Phi$ is a tight frame, by substituting (2.12) into (2.13), we can write $f = A^{-1}\sum_{k\in K}\langle f, \varphi_k\rangle\varphi_k$. It is interesting to see how this extends smoothly to the case that $\Phi$ is close to tight, *i.e.* $A$ is close to $B$.

Let $\rho = \frac{B}{A} - 1$. If $0 < \rho \ll 1$, $F^*F \approx \frac{A+B}{2} I_N$, so $(F^*F)^{-1} \approx \frac{2}{A+B} I_N$. Precisely,

$$f = \frac{2}{A+B}\sum_{k\in K}\langle f, \varphi_k\rangle\varphi_k + Rf, \tag{2.16}$$

where $R = I_N - \frac{2}{A+B}F^*F$. (This is valid for any $\rho$.) Let

$$f_0 = \frac{2}{A+B}\sum_{k\in K}\langle f, \varphi_k\rangle\varphi_k. \tag{2.17}$$

It can be shown that $\|R\| \leq \frac{B-A}{B+A} = \frac{\rho}{2+\rho}$; therefore $\|f - f_0\| \leq \frac{\rho}{2+\rho}\|f\|$, so (2.17) gives an estimate for $f$ with bounded error. The iteration

$$f_n = f_{n-1} + \frac{2}{A+B}\sum_{k\in K}\left[\langle f, \varphi_k\rangle - \langle f_{n-1}, \varphi_k\rangle\right]\varphi_k$$

gives a sequence of estimates satisfying

$$\|f - f_n\| \leq \left(\frac{\rho}{2+\rho}\right)^{n+1}\|f\|. \tag{2.18}$$

The dependence on $\rho$ in (2.18) shows that for a fixed error tolerance, less computation is required for reconstruction in a tight or nearly tight frame.

### 2.2.2 Classical Method

We now turn to the question of reconstructing when the frame coefficients $\{\langle f, \varphi_k \rangle\}_{k \in K}$ are degraded in some way. Any mode of degradation is possible, but the most practical situations are additive noise due to measurement error or quantization. We are most interested in the latter case because of its implications for efficient storage and transmission of information.

Suppose we wish to approximate $f$ given $Ff + \beta$, where $\beta \in \mathbb{C}^M$ is a zero-mean noise, uncorrelated with $f$. The key to finding the best approximation is that $FH = \mathrm{Ran}(F)$ is an $N$-dimensional subspace of $\mathbb{C}^M$. Hence the component of $\beta$ perpendicular to $FH$ should not hinder our approximation, and the best approximation is the projection of $Ff + \beta$ onto $\mathrm{Ran}(F)$. By Proposition 2.3, this approximation is given by

$$\hat{f} = \widetilde{F}^*(Ff + \beta). \tag{2.19}$$

Furthermore, because the component of $\beta$ orthogonal to $\mathrm{Ran}(F)$ does not contribute, we expect $\|f - \hat{f}\| = \|\widetilde{F}^*\beta\|$ to be smaller than $\|\beta\|$.

To make this more precise, recall Example 1 of §2.1.2. If $\beta = [\beta_1 \;\; \beta_2 \;\; \beta_3]^T$, where the $\beta_i$'s are independent random variables with mean zero and variance $\sigma^2$,

$$\begin{aligned}
& E\left(\left\| f - \widetilde{F}^*(Ff + \beta)\right\|^2\right) \\
&= E\left(\left\| f - \frac{2}{3}\sum_{k=1}^{3}\left(\langle f, \varphi_k\rangle + \beta_k\right)\varphi_k \right\|^2\right) \\
&= E\left(\left\| \frac{2}{3}\sum_{k=1}^{3}\beta_k\varphi_k\right\|^2\right) = \frac{4}{9}E\left(\sum_{k=1}^{3}\sum_{\ell=1}^{3}\beta_k\beta_\ell\varphi_k^T\varphi_\ell\right) \\
&= \frac{4}{9}E\left(\beta_1^2 + \beta_2^2 + \beta_3^2 - \beta_1\beta_2 - \beta_2\beta_3 - \beta_1\beta_3\right) = \frac{4}{3}\sigma^2.
\end{aligned}$$

Here we have used the fact that

$$\varphi_k^T\varphi_\ell = \begin{cases} 1 & k = \ell \\ -\frac{1}{2} & k \neq \ell \end{cases}.$$

Notice that this mean-squared error (MSE) is $\frac{2}{3}$ of the $2\sigma^2$ MSE that would appear in an orthogonal basis representation. The MSE reduction is by a factor of $1/r$, where $r$ is the redundancy of the tight frame. Having $O(1/r)$ MSE behavior is a general phenomenon for reconstruction by projection in a tight frame representation. It is a special case of the following proposition.

Proposition 2.5: Noise Reduction in Classical Reconstruction
Let $\Phi = \{\varphi_k\}_{k=1}^M$ be a frame of unit-norm vectors with associated frame operator $F$ and let $\beta = [\beta_1 \;\; \beta_2 \;\; \cdots \;\; \beta_M]^T$, where the $\beta_i$'s are independent random variables with mean zero and variance $\sigma^2$. Then the MSE of the classical reconstruction (2.19) satisfies

$$\mathrm{MSE} \leq \frac{M\sigma^2}{A^2}. \tag{2.20}$$

Furthermore, if the frame is tight, (2.20) holds with equality, giving

$$\mathrm{MSE} = \frac{N^2\sigma^2}{M} = \frac{N\sigma^2}{r}. \tag{2.21}$$

PROOF: See §A.4. ■

Now consider the case where the degradation is due to quantization. Let $x \in \mathbb{R}^N$ and $y = Fx$, where $F \in \mathbb{R}^{M\times N}$ is a frame operator. Suppose $\hat{y} = Q(y)$, where $Q : \mathbb{R}^M \to \mathbb{R}^M$ is a *scalar* quantization function, *i.e.* $Q(y) = [q(y_1)\ \ q(y_2)\ \ \ldots\ \ q(y_M)]^T$, where $q : \mathbb{R} \to \mathbb{R}$ is a scalar quantization function.

One approach to approximating $x$ given $\hat{y}$ is to treat the quantization noise $\hat{y} - y$ as random, independent in each dimension, and uncorrelated with $y$. These assumptions make the problem tractable using statistical techniques. The problem reduces to the previous problem, and $\hat{x} = \widetilde{F}^*\hat{y}$ is the best approximation. Strictly speaking, however, the assumptions on which this reconstruction is based are not valid because $\hat{y} - y$ is a deterministic quantity depending on $y$, with interplay between the components.

### 2.2.3 Consistent Reconstruction

The shortcoming of the classical reconstruction method is that it disregards deterministic properties of quantization. As a result, the reconstruction may have a different quantized value than the original. Using the term introduced by Thao and Vetterli [33], we say that the reconstruction may be *inconsistent*.

DEFINITION. We say that $\hat{x}$ is a *consistent estimate of* $x$ or a *consistent reconstruction* if $Q(F\hat{x}) = Q(Fx)$. A reconstruction that is not consistent is said to be *inconsistent*.

In words, we would say that an estimate is consistent if it is the same as its quantized version. Another way to understand consistency is in terms of partitions. $Q$ induces a partitioning of $\mathbb{R}^M$. (We can temporarily remove the restriction that $Q$ is a scalar quantizer and require only that the partition regions are convex.) This quantization also induces a partitioning of $\mathbb{R}^N$ through the inverse image of $Q \circ F$. The partition of $\mathbb{R}^N$ can be viewed in another way: Since $Q$ partitions $\mathbb{R}^M$, it also partitions the $N$-dimensional subspace $F(\mathbb{R}^N)$. Mapping back to $\mathbb{R}^N$ using $\widetilde{F}^*$ gives the partition of $\mathbb{R}^N$ induced by $Q$. A consistent estimate is simply one that falls in the same partition region as the original.

All of these concepts are illustrated for $N = 2$ and $M = 3$ in Figure 2.3. The ambient space is $\mathbb{R}^M$. The cube represents the partition region in $\mathbb{R}^M$ containing $y = Fx$ and has codebook value $\hat{y}$. The plane is $F(\mathbb{R}^N)$ and hence is the subspace within which any unquantized value must lie. The intersection of the plane with the cube gives the shaded triangle within which a consistent estimate must lie. Projecting to $F(\mathbb{R}^N)$, as in the classical reconstruction method, removes the out-of-subspace component of $y - \hat{y}$. As illustrated, this type of reconstruction is not necessarily consistent. For further geometric interpretation of quantized frame expansions, refer to Appendix B.

With no assumptions on $Q$ other than that the partition regions be convex, a consistent estimate can be determined using the projection onto convex sets (POCS) algorithm. In this

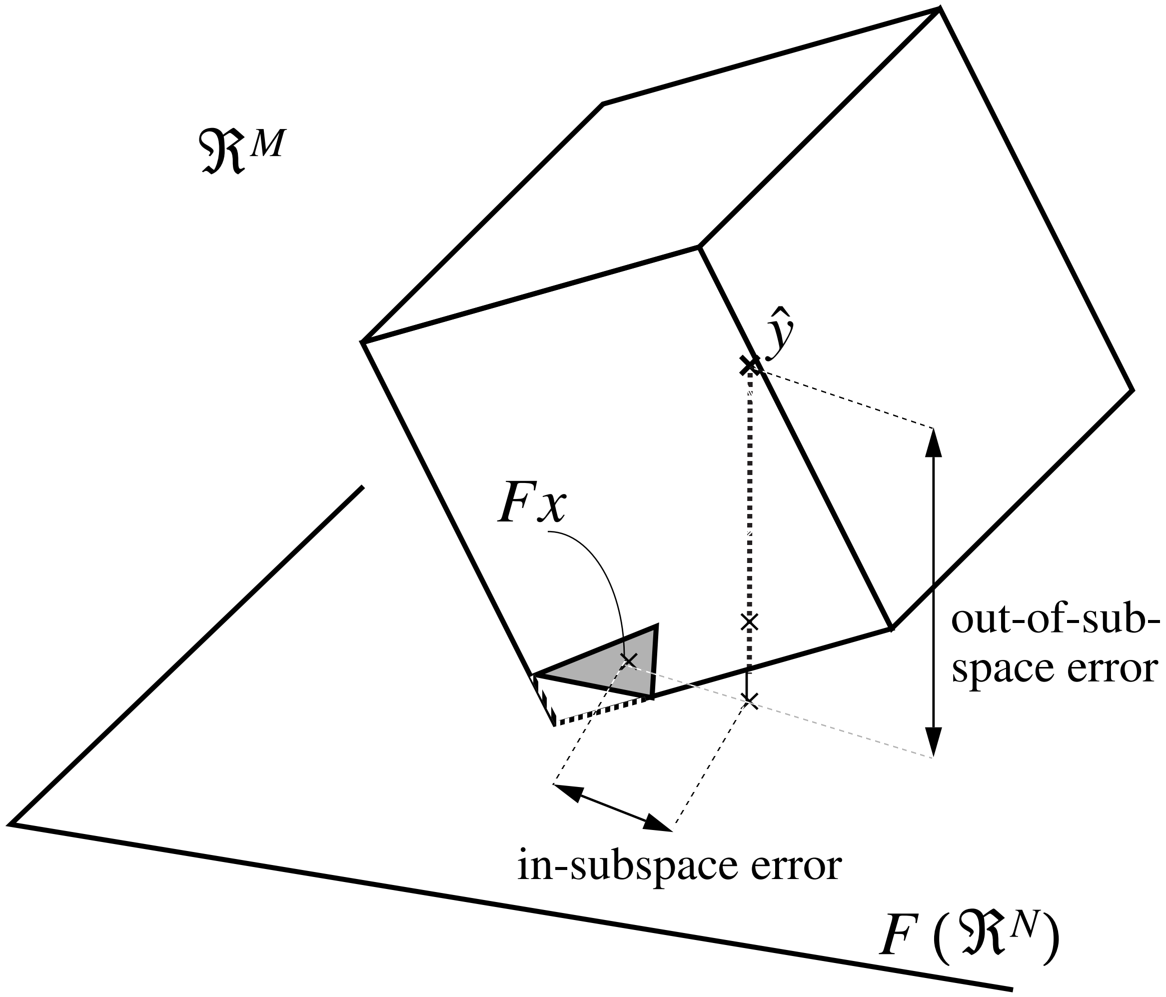



Figure 2.3: Illustration of consistent reconstruction

case that implies generating a sequence of estimates by alternately projecting on $F(\mathbb{R}^N)$ and $Q^{-1}(\hat{y})$.

When $Q$ is a scalar quantizer, a linear program can be used to find consistent estimates. For $i = 1, 2, \ldots, M$, denote the quantization stepsize in the $i$th component by $\Delta_i$. For notational convenience, assume that the reproduction values lie halfway between decision levels. Then for each $i$, $|\hat{y}_i - y_i| \leq \frac{\Delta_i}{2}$. To obtain a consistent estimate, for each $i$ we must have

$$|(F\hat{x})_i - \hat{y}_i| \leq \frac{\Delta_i}{2}.$$

Expanding the absolute value, we find the constraints

$$F\hat{x} \leq \frac{1}{2}\Delta + \hat{y} \quad \text{and} \quad F\hat{x} \geq -\frac{1}{2}\Delta + \hat{y}$$

where $\Delta = [\Delta_1 \;\; \Delta_2 \;\; \ldots \;\; \Delta_M]^T$, and the inequalities are elementwise. These inequalities can be combined into

$$\begin{bmatrix} F \\ -F \end{bmatrix} \hat{x} \leq \begin{bmatrix} \frac{1}{2}\Delta + \hat{y} \\ \frac{1}{2}\Delta - \hat{y} \end{bmatrix}. \tag{2.22}$$

The formulation (2.22) shows that $\hat{x}$ can be determined through linear programming [31]. The feasible set of the linear program is exactly the set of consistent estimates, so an arbitrary cost function can be used.

A linear program always returns a corner of the feasible set [31, §8.1], so this type of reconstruction will not be close to the centroid of the partition cell. Since the cells are convex, one could use several cost functions to (presumably) get different corners of the feasible set and average the results. Another approach is to use a quadratic cost function equal to the distance from the projection estimate given by (2.19). Both of these methods will reduce the MSE by a constant factor. They do not change the asymptotic behavior of the MSE as the redundancy $r$ is increased.

### 2.2.4 Error Bounds

In this subsection, we concern ourselves with bounds on the MSE in estimating $x$ from $\hat{y}$. Our fundamental premise is that any reconstruction method that gives consistent estimates is asymptotically (in the redundancy $r$) optimal. We now prove two bounds that support this conviction: first, an $O(1/r^2)$ MSE lower bound for any reconstruction algorithm; and second, an $O(1/r^2)$ MSE upper bound for consistent reconstruction. Since we are varying $r$, we must consider sequences of frames with growing redundancy.

Theorem 2.6: MSE Lower Bound
For any set of quantized frame expansions, any reconstruction algorithm will yield an MSE that can be lower bounded by an $O(1/r^2)$ expression.[3]
Proof: The proof of this general result is given under the guise of a more restricted result in [35]. There it is proven that when the frame operators correspond to oversampled A/D conversion (see §2.1.2), any reconstruction algorithm will yield an MSE that can be lower bounded by an $O(1/r^2)$ expression. The proof is based on counting the number of cells in the partition of $\mathbb{R}^N$ and using Zador's formula. The only frame-specific property that is used corresponds to requiring that elements of the frame not be parallel. Having parallel frame elements would reduce the number of cells in the partition and hence increase the MSE. Therefore the proof extends to the general case. ■

Proposition 2.7: MSE Upper Bound (Restricted Case)
Let $x$ be such that it has a probability density.[4] Consider quantized frame expansions of $x$ with frame corresponding to the frame operator (2.9) and quantization stepsize $\Delta$. For sufficiently small (fixed) $\Delta$, a consistent reconstruction algorithm will yield an MSE that can be upper bounded by an $O(1/r^2)$ expression.
Proof: The proof is based on a correspondence between vectors in $\mathbb{R}^N$ and periodic, bandlimited, continuous-time signals. Let $x_c(t)$ be defined as in (2.8), where $T$ is arbitrary. Then

[3] Actually, we must exclude the case where $x$ has a degenerate distribution that allows for perfect reconstruction. This point is not emphasized in [35].

[4] This is to eliminate degenerate distributions for $x$.

quantized frame expansion of $x$ is equivalent to oversampled A/D conversion of $x_c(t)$. According to Thao and Vetterli [33, Thm. 4.1], the MSE can be upper bounded by an $O(1/r^2)$ expression. One requirement in applying their result is that $x_c(t)$ must have sufficient quantization threshold crossings. In our more general framework, this corresponds to requiring that the distribution of $x$ not be overly concentrated inside a sphere of radius $\Delta$.[5] Since $x$ has a probability density, this can be assured by choosing $\Delta$ sufficiently small. ■

Conjecture 2.8: MSE Upper Bound
Under very general conditions, for any set of quantized frame expansions, any algorithm that gives consistent estimates will yield an MSE that can be upper bounded by an $O(1/r^2)$ expression.

For this general upper bound to hold, some sort of non-degeneracy condition is required because we can easily construct a sequence of frames with increasing $r$ for which the frame coefficients give no additional information as $r$ is increased. For example, we can start with an orthonormal basis and increase $r$ by adding copies of vectors already in the frame. Putting aside pathological cases, simulations for quantization of a source uniformly distributed on $[-1, 1]^N$ support this conjecture. Simulations were performed with three types of frame sequences:

I. A sequence of frames corresponding to oversampled A/D conversion, as given by (2.9). This is the case in which we have a provable $O(1/r^2)$ MSE upper bound.

II. For $N = 3$, 4, and 5, Hardin, Sloane and Smith have numerically found arrangements of up to 130 points on $N$-dimensional unit spheres that maximize the minimum Euclidean norm separation [16].

III. Frames generated by randomly choosing points on the unit sphere according to a uniform distribution.

Simulation results are given in Figure 2.4. The dashed, dotted, and solid curves correspond to frame types I, II, and III, respectively. The data points marked with $+$'s correspond to using a linear program based on (2.22) to find consistent estimates. The data points marked with $\circ$'s correspond to classical reconstruction. The important characteristics of the graph are the slopes of the curves. Note that $O(1/r)$ MSE corresponds to a slope of -3.01 dB/octave and $O(1/r^2)$ MSE corresponds to a slope of -6.02 dB/octave. The consistent reconstruction algorithm exhibits $O(1/r^2)$ MSE for each of the types of frames. The classical method exhibits $O(1/r)$ MSE behavior, as expected. It is particularly interesting to note that the performance with random frames is as good as with either of the other two types of frames.

Note that in light of Theorem 2.2, it may be useful to try to prove Conjecture 2.8 only for tight frames.

[5] In most cases, we assume quantizer offsets such that zero is either a reconstruction value or a boundary value. By randomizing the quantizer offset, we can remove this condition.

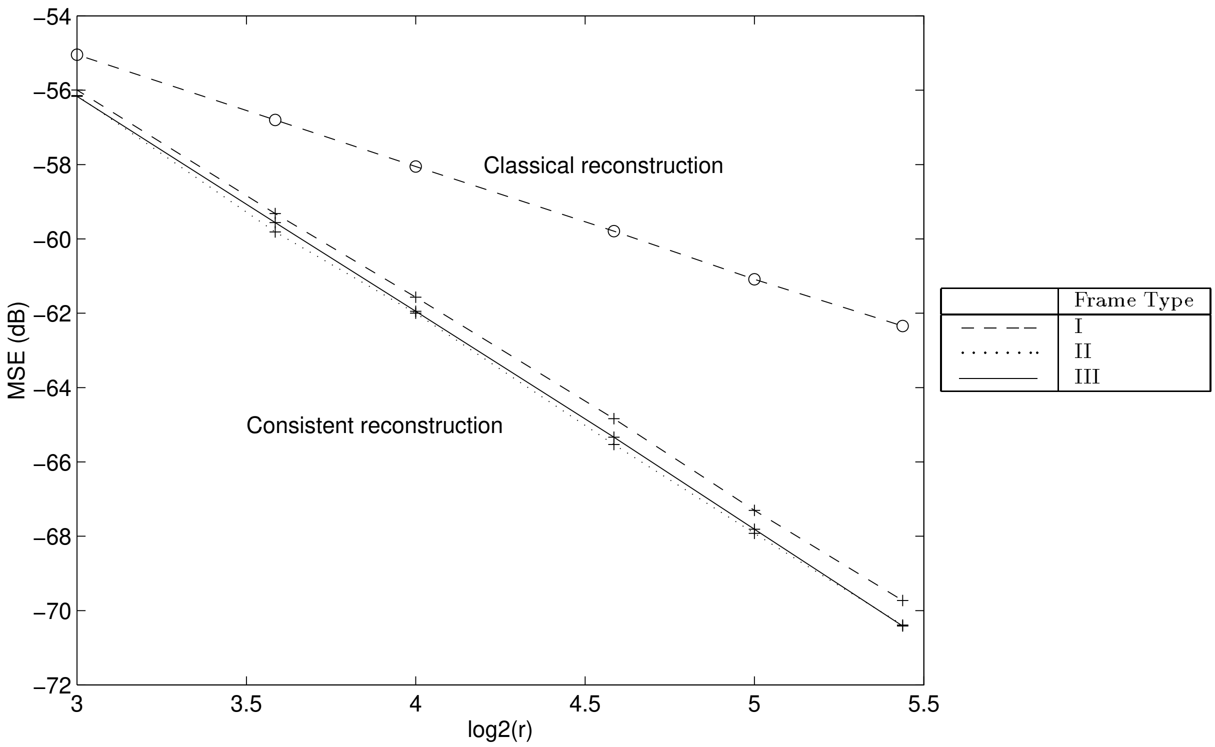



Figure 2.4: Experimental results for reconstruction from quantized frame expansions. Shows $O(1/r^2)$ MSE for consistent reconstruction and $O(1/r)$ MSE for classical reconstruction.

### 2.2.5 Rate-Distortion Tradeoffs

Our discussion of quantized frame expansions has focused on expected distortion without concern for rate. In this subsection we begin consideration of rate-distortion tradeoffs.

We have demonstrated that optimal reconstruction techniques give an MSE proportional to $1/r^2$. It is well known that in orthogonal representations the MSE is proportional to $\Delta^2$. This extends to the frame case as well. Thus we have two ways to reduce the MSE by approximately a factor of four:

- double $r$;
- halve $\Delta$.

*A priori*, there is no reason to think that these options each have the same effect on the rate. As the simplest possible case, suppose a frame expansion is stored (or transmitted) as $M$ $B$-bit numbers, for a total rate of $MB$ bits per sample. Doubling $r$ gives $2M$ $B$-bit numbers, for a total rate of $2MB$ bits per sample. On the other hand, halving $\Delta$ results in $M$ $(B+1)$-bit numbers for a rate of only $M(B+1)$ bits per sample.

This argument suggests that halving $\Delta$ is always the better option, but a few comments are in order. One caveat is that in some situations, doubling $r$ and halving $\Delta$ may have very different costs. For example, in oversampled A/D conversion, the monetary cost of halving $\Delta$ is much higher than that of doubling $r$ because it requires precision trimmed analog electronics. This is a major motivating factor for oversampling.

Also, if $r$ is doubled, storing the result as $2M$ $B$-bit values is far from the best thing to do. This is because many of the $M$ additional numbers give little or no information on $x$. A conclusion of Zamir and Feder [38] was described by Zamir as, "one good measurement is better than many noisy ones" [39]. It is important to note that although they consider quantization noise, they do not consider consistency. These topics are discussed further in Appendix B; Appendix C explores the use of quantized frame expansion as the first stage in a lattice quantizer.

We conclude this chapter by noting that it is complicated to get *efficient* signal representations from highly redundant quantized frame expansions. Because of redundancy, each frame coefficient does not give the same amount of information on $x$; one way to get an efficient representation would be to retain only the frame coefficients that give a lot of information on $x$. This is essentially the theme of the next chapter.

# Chapter 3

# Adaptive Expansions

In this chapter, we broaden our approach to finding linear expansions by allowing the basis functions of the expansion to vary depending on the signal. However, we are not adapting in the traditional sense of making fine adjustments depending on an error signal. Instead, our basic tool is the matching pursuit algorithm of [20] in which the adaptation is in the *choice* of basis functions from a *fixed* dictionary (frame).

In §3.1, we introduce the optimal approximation problem in order to establish its computational intractability. The matching pursuit algorithm, described in §3.2, is a greedy algorithm for finding approximate solutions to the approximation problem. Quantization of coefficients in matching pursuit leads to many interesting issues; some of these are discussed in §3.3. Along with exploring general properties of matching pursuit, we are interested in its application to compressing data vectors in $\mathbb{R}^N$. A general vector compression method based on matching pursuit is described in §3.4.

## 3.1 The Optimal Approximation Problem

At the end of the previous chapter, we noted that the set of coefficients from a highly redundant frame expansion are, without sophisticated coding, an inefficient representation of a signal. We expect to find more efficient representations by forming a linear expansion with respect to a subset of the original frame. This problem is formalized below.

Definition [7, Ch. 2]. Let a dictionary $\mathcal{D}$ be a frame in $H$. Let $\epsilon > 0$ and $L \in \mathbb{Z}^+$. For $f \in H$, an expansion

$$\tilde{f} = \sum_{i=1}^{L} \alpha_i \varphi_{k_i}, \tag{3.1}$$

where $\alpha_i \in \mathbb{C}$ and $\varphi_{k_i} \in \mathcal{D}$, is called an $(\epsilon, L)$-*approximation* if $\|\tilde{f} - f\| < \epsilon$. An expansion (3.1) that minimizes $\|\tilde{f} - f\|$ is called an $L$-*optimal approximation.*

Since the $\alpha_i$'s are not subject to quantization, these approximation problems do not exactly correspond to finding rate-distortion optimal representations for fixed $L$. Also, this formulation does not account for the fact that, with entropy coding, the rate associated with $\{\varphi_{k_i}\}_{i=1}^{L}$ may depend on the choice of dictionary elements. Nevertheless, we are discouraged from attempting to find optimal quantized representations by the following theorem.

THEOREM 3.1: INTRACTABILITY OF OPTIMAL APPROXIMATION [7]
Let $k \geq 1$ and let $\mathcal{D}$ be a dictionary that contains $O(N^k)$ vectors. Let $0 < \gamma_1 < \gamma_2 < 1$ and let $L \in \mathbb{Z}^+$ such that $\gamma_1 N \leq L \leq \gamma_2 N$. For any given $\epsilon > 0$ and $f \in H$, determining whether an $(\epsilon, L)$-approximation exists is NP-complete. Finding the $L$-optimal approximation is NP-hard.
PROOF: See [7, Ch. 2]. ■

## 3.2 Matching Pursuit

The intractability of $L$-optimal approximation stems from the number of ways to choose $L$ dictionary elements. The complexity is reduced if the dictionary elements are chosen one at a time instead of $L$ at once. This reduction of a "global" problem to simpler "local" problems is the defining characteristic of a *greedy* algorithm. Matching pursuit is a greedy algorithm for finding approximate solutions to the $L$-optimal approximation problem. It progressively refines a signal estimate instead of finding $L$ components jointly.

Matching pursuit was introduced to the signal processing community in the context of time-frequency analysis by Mallat and Zhang [20]. Mallat and his students have uncovered many of its properties [7, 8, 9, 40].

### 3.2.1 Algorithm

Let $\mathcal{D} = \{\varphi_k\}_{k=1}^M \subset H$ be a frame. We impose the additional constraint that $\|\varphi_k\| = 1$ for all $k$. We will call $\mathcal{D}$ our *dictionary* of vectors. Matching pursuit is an algorithm to represent $f \in H$ by a linear combination of elements of $\mathcal{D}$. Furthermore, matching pursuit is an iterative scheme that at each step attempts to approximate $f$ as closely as possible in a greedy manner. We expect that after a few iterations we will have an efficient approximate representation of $f$.

In the first step of the algorithm, $k_0$ is selected such that $|\langle \varphi_{k_0}, f \rangle|$ is maximized. Then $f$ can be written as its projection onto $\varphi_{k_0}$ and a residue $R_1 f$,

$$f = \langle \varphi_{k_0}, f \rangle \varphi_{k_0} + R_1 f.$$

The algorithm is iterated by treating $R_1 f$ as the vector to be best approximated by a multiple of $\varphi_{k_1}$. At step $p+1$, $k_p$ is chosen to maximize $|\langle \varphi_{k_p}, R_p f \rangle|$ and

$$R_{p+1} f = R_p f - \langle \varphi_{k_p}, R_p f \rangle \varphi_{k_p}. \tag{3.2}$$

Identifying $R_0 f = f$, we can write

$$f = \sum_{i=0}^{n-1} \langle \varphi_{k_i}, R_i f \rangle \varphi_{k_i} + R_n f. \tag{3.3}$$

Hereafter we will denote $\langle \varphi_{k_i}, R_i f \rangle$ by $\alpha_i$.

### 3.2.2 Discussion

Matching pursuit is similar to a class of algorithms used in statistics called *projection pursuits*. The proof of the convergence of projection pursuits given in [18] can be used to prove the convergence of matching pursuit in infinite dimensional spaces. In infinite dimensional spaces, the convergence can be quite slow. However, the convergence is exponential in finite dimensional spaces [7, §3.1].

Since $\alpha_i$ is determined by projection, $\alpha_i \varphi_{k_i} \perp R_{i+1} f$. Thus we have the "energy conservation" equation

$$\|R_i f\|^2 = \|R_{i+1} f\|^2 + \alpha_i^2. \tag{3.4}$$

This fact, the selection criterion for $k_i$, and the fact that $\mathcal{D}$ spans $H$, can be combined for a simple convergence proof for finite dimensional spaces. In particular, the energy in the residue is strictly decreasing until $f$ is exactly represented.

In the language of §3.1, matching pursuit can be viewed as finding a 1-optimal approximation and then iteratively finding 1-optimal approximations on the resulting residues. If $\mathcal{D}$ is an orthonormal basis, matching pursuit finds the optimal expansion. For an arbitrary dictionary, however, matching pursuit does not generally find optimal expansions. In fact, if no two elements of the dictionary are orthogonal, matching pursuit expansions are not only not optimal, but they do not converge in a finite number of steps except on a set of measure zero [7, §3.1].

In the following, detailed operation counts and other measures of complexity will not be given since the emphasis is not on implementation details. One point to note is that the full set of inner products $\{\langle \varphi_i,\, R_p f\rangle\}_{i=1}^{M}$ need not be computed at each iteration. By (3.2),

$$\langle \varphi_i,\, R_{p+1}\rangle = \langle \varphi_i,\, R_p\rangle - \langle \varphi_{k_p},\, R_p\rangle\langle \varphi_i,\, \varphi_{k_p}\rangle. \tag{3.5}$$

In (3.5), $\langle \varphi_i,\, R_p\rangle$ and $\langle \varphi_{k_p},\, R_p\rangle$ are known from the previous iteration, so only $\langle \varphi_i,\, \varphi_{k_p}\rangle$ must be computed. Depending on the dictionary structure, this may involve a table lookup or a simple calculation. Alternatively, the dictionary can be structured so that only a few such inner products are nonzero.

Note that the output of a matching pursuit expansion is not only the coefficients ($\alpha_0$, $\alpha_1$, …), but also the indices ($k_0$, $k_1$, …). For storage and transmission purposes, we will have to account for the indices.

### 3.2.3 Orthogonalized Matching Pursuits

It was noted that, even in a finite dimensional space, matching pursuit is not guaranteed to converge in a finite number of iterations. This is a serious drawback when exact (or very precise) signal expansions are desired, especially since an optimal algorithm would choose a basis from the dictionary and get an exact expansion in $N$ steps. The cause of this drawback is that at step $p+1$, $\varphi_{k_p}$ is not necessarily chosen orthogonal to $\mathrm{Span}(\{\varphi_{k_i}\}_{i=0}^{p-1})$.

The matching pursuit algorithm can be modified to insure that at each iteration the contribution to the linear expansion is orthogonal to all previous terms. Convergence in $N$ steps is then guaranteed. A simple method of accelerating convergence through orthogonalization is described below [28]. The selection of dictionary elements is the same as before.

After a dictionary element $\varphi_{k_p}$ is chosen, it is orthogonalized with respect to $\{\varphi_{k_i}\}_{i=0}^{p-1}$ before the residue $R_p f$ is calculated. (Because of the orthogonalization, no dictionary element is chosen twice.) This insures that $R_{p+1}f$ is orthogonal to $\varphi_{k_i}$ for $i = 0, 1, \ldots, p$. A better orthogonalization method is presented by Kalker and Vetterli in [19].

It has been noted by several authors [7, 19, 36] that for a small number of iterations, orthogonal matching pursuit does not converge significantly faster than the non-orthogonalized version. For this reason, orthogonal matching pursuit is not considered hereafter.

### 3.2.4 Relationship to the Karhunen-Loève Transformation

In this section, we forge an analogy between matching pursuit and the Karhunen-Loève transform (KLT). Our aim is to show that matching pursuit has some of the properties that make the KLT useful in transform coding. We assert that matching pursuit acts as a universal transform for transform coding.

For a stationary, vector-valued random process $\mathbf{X}$, the Karhunen-Loève transform is the unique orthogonal transform $U$ such that $\mathbf{Y} = U\mathbf{X}$ has a diagonal covariance matrix with the eigenvalues appearing in descending order on the diagonal [21, §1.2.4]. Note that determining the KLT requires knowledge of the distribution of $\mathbf{X}$. Approximating the KLT from data is essentially the same as principal component analysis [17].

It is well known that the KLT is the optimal transform for transform coding. Since the limitations to this result are not as well known, we state the following theorem paraphrased from [13]:

THEOREM 3.2: OPTIMALITY OF THE KARHUNEN-LOÈVE TRANSFORM
Consider the transform coding of a jointly Gaussian random process. Suppose the quantization is fine enough to use high resolution approximations, and that arbitrary real (non-integer) values can be allocated to the resolution of each (scalar) quantizer. Then the KLT achieves the lowest overall distortion of any orthogonal transform.
PROOF: See [13, §8.6]. ∎

Two properties of the KLT that make it good for transform coding are qualitatively mimicked by matching pursuit:

1. Energy compaction: For $1 \le i \le N-1$, the energy in $\{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_i}\}$ is maximum over all orthogonal transforms. For this reason the KLT is said to give optimal energy compaction.

2. Principal axes: If $\mathbf{X}$ has an ellipsoidal distribution (as when $\mathbf{X}$ is Gaussian), the $i$th transformed variable $\mathbf{y_i}$ corresponds to the $i$th principal axis of the ellipsoid. This is closely coupled with energy compaction, since the $i$th principal axis is the direction in which there is the $i$th largest energy.

We first explore the energy compaction properties of matching pursuit. The criterion for the choice of $k_i$ makes some degree of energy compaction obvious. Since we only solve 1-optimal approximation problems, matching pursuit does not always give optimal energy compaction when more than one iteration is performed. However, in matching pursuit we are optimizing on a *sample-by-sample* basis, as opposed to looking at average performance with

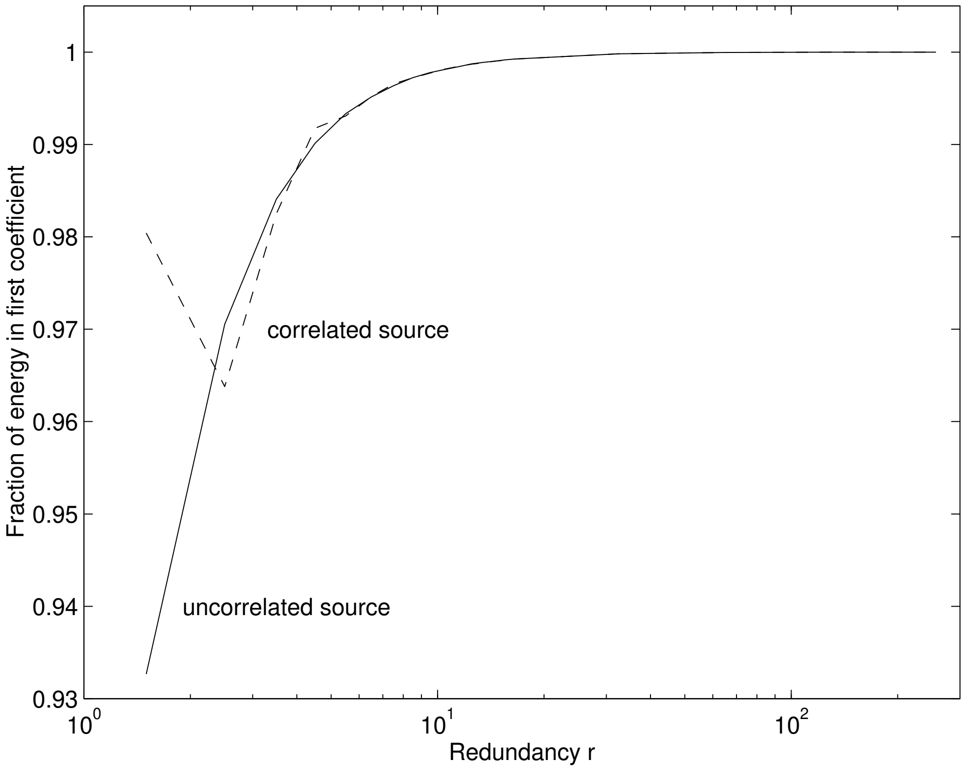



Figure 3.1: Energy compaction achieved using matching pursuit on an $\mathbb{R}^2$-valued source.

a fixed transform. Therefore matching pursuit generally gives much more energy compaction than the KLT. In particular, matching pursuit will give energy compaction even if $\mathbf{X}$ has a diagonal covariance matrix, in which case the KLT gives no energy compaction.

Energy compaction performance was assessed by simulation. In $\mathbb{R}^2$, two sources were used:

- An uncorrelated zero-mean Gaussian source $\mathbf{X} \sim \mathcal{N}(0, I)$.
- A Gaussian source
$$\mathbf{X} \sim \mathcal{N}\left(0, A_\theta^T \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix} A_\theta\right) \tag{3.6}$$
where $A_\theta$ is a Givens plane rotation matrix ($\theta = \frac{\pi}{3}$).

Dictionaries of the form
$$\mathcal{D} = \left\{ \left[ \cos \frac{2\pi k}{M} \quad \sin \frac{2\pi k}{M} \right]^T \right\}_{k=0}^{M-1} \tag{3.7}$$
were used. The results are shown in Figure 3.1. Using matching pursuit, more than 93% of the energy is captured in the first coefficient, and the energy compaction increases with increasing dictionary redundancy. The KLT would give $\frac{1}{2}$ and $\frac{5}{6}$ of the energy in the first coefficient for the uncorrelated and correlated sources, respectively.

Simulations were also performed for $\mathbb{R}^4$-valued sources. Two sources were used:

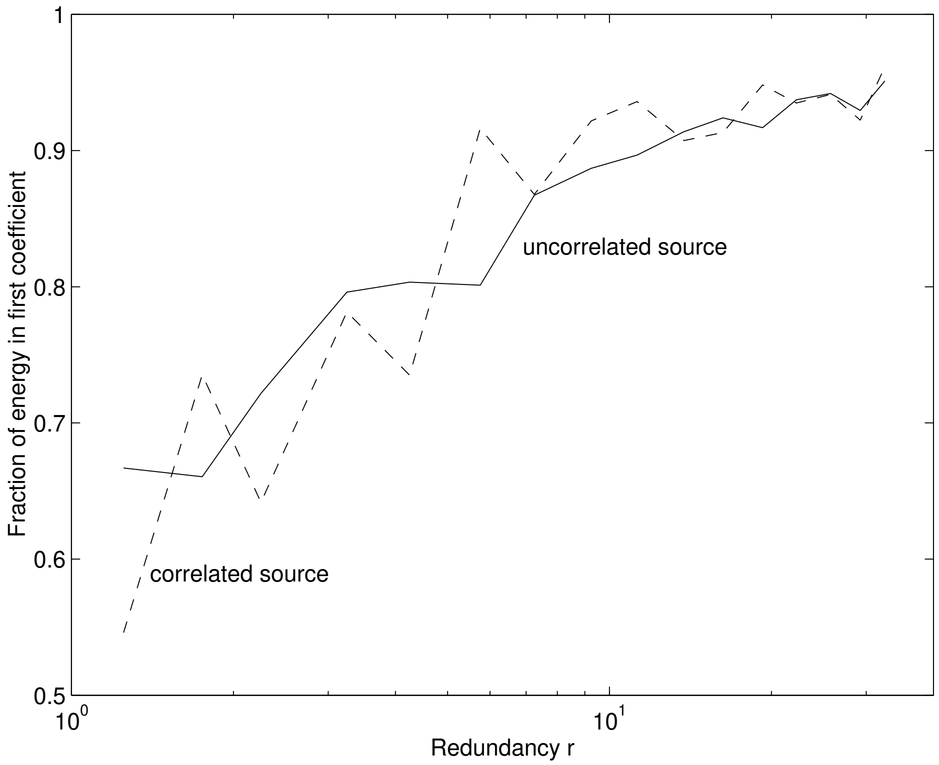



Figure 3.2: Energy compaction achieved using matching pursuit on an $\mathbb{R}^4$-valued source.

- An uncorrelated zero-mean Gaussian source $\mathbf{X} \sim \mathcal{N}(0, I)$.
- A correlated zero-mean Gaussian source formed by placing a first-order autoregressive source with correlation 0.9 in blocks of length 4.

Dictionaries were generated from maximally spaced points on the unit sphere [16]. The results are given in Figure 3.2. As expected, the energy compaction in the first component increases with $r$, ranging from about 0.55 to 0.96. In this case, the KLT would give $\frac{1}{4}$ and $\approx 0.8817$ of the energy in the first coefficient for the uncorrelated and correlated sources, respectively. So in this experiment, matching pursuit always gives better energy compaction for the uncorrelated source, and also does so for the highly correlated source when $r > 8$.

When $\mathbf{X}$ has an ellipsoidal distribution, geometric intuition suggests that $\varphi_{k_0}$ will more likely be close to the principal axis than far from it. Similarly, given that $\varphi_{k_0}$ is nearly parallel to the principal axis, the distribution of $R_0x$ will be ellipsoidal with principal axis equal to the second principal axis of the distribution of $x$. (Since the most we can possibly say is that $\varphi_{k_0}$ *usually* is nearly parallel to the principal axis, this reasoning is somewhat weak.) We would like to formalize this intuition. In particular, we attempt to demonstrate that the indices $k_i$ can be used to estimate the principal axes and that the algorithm is likely to choose indices that correspond to the KLT. We are however *not* asserting that it would be ideal for the algorithm to choose indices corresponding to the KLT; matching pursuit is acting *locally*, while the KLT is based on global stationary statistics.

Methods for estimating the principal axes are not immediately obvious. We cannot simply "average the $\varphi_{k_0}$'s to estimate the first principal axis" because, with a sufficiently regular and dense dictionary and an ellipsoidal distribution, $E\varphi_{k_0} = 0$.

For example, consider quantization of the $\mathbb{R}^2$-valued source from (3.6), expanded using a dictionary as in (3.7) with $M = 199$. Figure 3.3 shows histograms of $k_0$ and $k_1$ for 10000 samples. The peaks of the histograms are at $\widetilde{k_0} = 33$ and $\widetilde{k_1} = 83$. These correspond to angles (modulo $\pi$) of $\frac{66\pi}{199}$ and $\frac{166\pi}{199}$, respectively. These are very close to angles of the principal axes of the distribution, which are $\frac{\pi}{3}$ and $\frac{5\pi}{6}$. Unfortunately, looking at peaks of histograms is not very robust and is limited by the redundancy and regularity of the dictionary. Thus we would like to use a method that involves averaging. As we noted, averaging $\varphi_{k_0}$'s and $\varphi_{k_1}$'s is meaningless. Referring to Figure 3.3, this is because of the bimodality of the histograms. It also makes no sense to average the index numbers because this would not be invariant to renumberings of the dictionary, even those renumberings that maintain the natural order.[1] Using a dictionary that is spread along half of the unit circle instead of the whole unit circle would bias the estimates toward the center of the half-circle chosen. The proposed solution is to use the histogram peaks as initial estimates of the principal axes and then "center" the dictionary around the corresponding vectors. For concreteness, suppose we are estimating the first principal axis. Suppose we have used matching pursuit to expand a set of samples. Let $\widetilde{k_0}$ denote the histogram peak of the $k_0$'s. For the $m$th sample, we increment the principal axis estimate by $\varphi_{(k_0)_m}$ if $\langle \varphi_{(k_0)_m}, \varphi_{\widetilde{k_0}} \rangle \geq 0$, and by $-\varphi_{(k_0)_m}$ otherwise. This procedure can be applied in any dimension because it does not depend on an ordering of dictionary elements. A potential pitfall is that if the dictionary is not uniform, the histogram peak may be a poor initial estimate.

Figure 3.4 shows simulation results for principal axis estimation using the methods discussed above. The source is as given by (3.6) and the dictionary is as in (3.7) with $M = 399$. The error is measured as an angular error in radians. The figure shows that both methods (looking only at histogram peaks and averaging using a peak as an initial estimate) give increasingly good estimates as data accumulates. The averaging method gives MSEs that are lower by about a factor of ten.

Simulations were also conducted with the same $\mathbb{R}^4$-valued autoregressive source as before. A dictionary of 130 maximally spaced unit vectors from [16] was used. The results are shown in Figure 3.5. The three pairs of curves correspond to estimating the first three principal axes of the distribution. The solid and dashed curves correspond to the averaging and histogram peak methods, respectively. In this case the error is measured as Euclidean distance between a unit vector in the true axis direction and the estimated axis direction. The results show that while the first axis can be well-estimated, it is much harder to estimate the subsequent principal axes. The principal axes are probably easier to estimate when the eigenvalue spread of the covariance of the source is large, but this is not explored further in this discussion.

Before moving on to study the effects of coefficient quantization, we would like to explore the dependence of index entropy on $r$. We have seen that increasing dictionary redundancy increases energy compaction. The price to pay is that the entropy of the indices goes up. We explore this tradeoff through an example. This time we consider a non-ellipsoidal source generated by equally mixing sources of the form (3.6) with $\theta$ equal to $\frac{\pi}{4}$ and $-\frac{\pi}{4}$. Figure 3.6 shows 1000 samples from this source. Note that the KLT for this source is simply the identity

[1] In higher dimensions, there would generally be no natural ordering to dictionary elements.

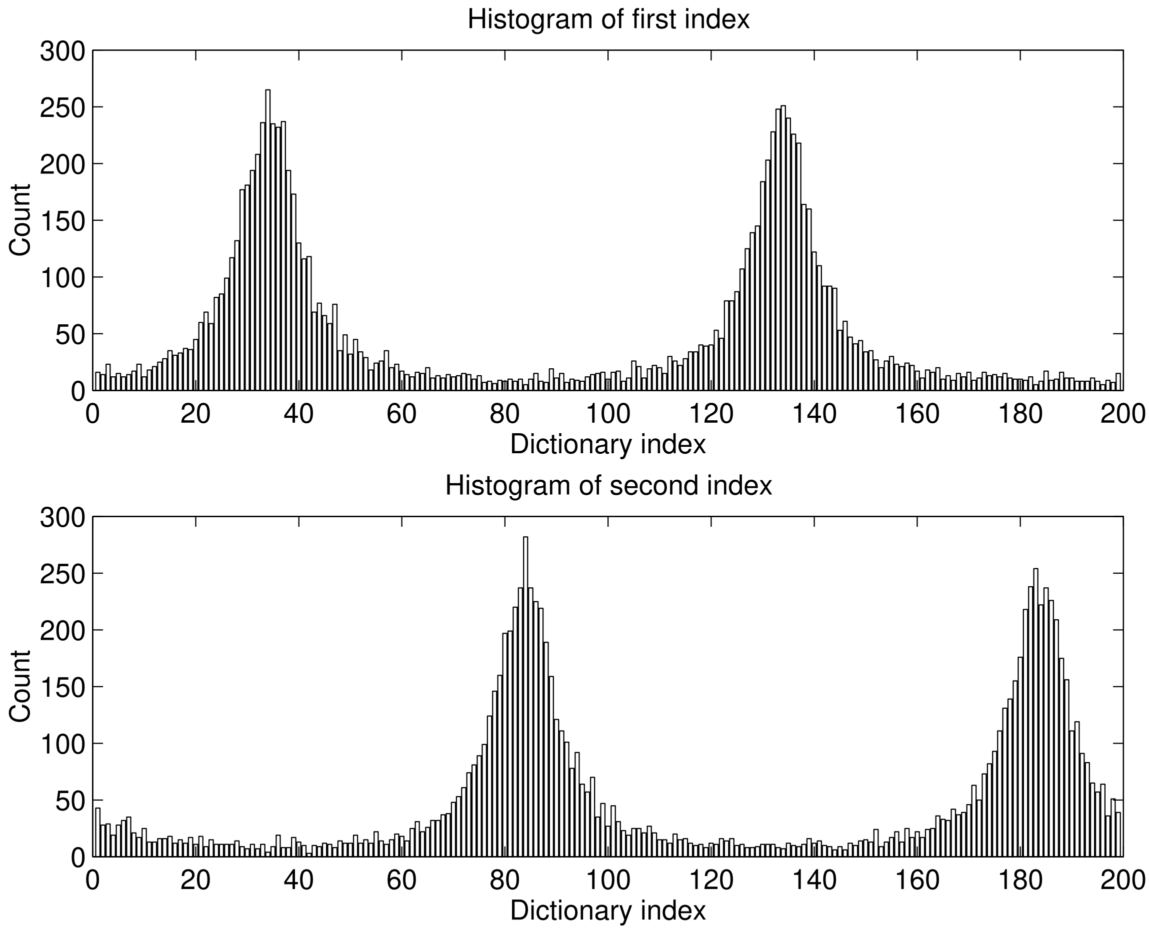



Figure 3.3: Histograms of indices chosen by matching pursuit

transformation. The samples were expanded using dictionaries of the form

$$\mathcal{D} = \left\{ \left[ \cos \frac{\pi k}{M} \quad \sin \frac{\pi k}{M} \right]^T \right\}_{k=0}^{M-1}. \tag{3.8}$$

Figure 3.7 shows the resulting energy compactions and index entropies as functions of $r$. (The index entropy should rightly be called a *sample* entropy. One must be very careful to use large sample sizes to get relevant sample entropies.) We see that the entropy of the first index is proportional to $\log r$, but the energy compaction levels off rather quickly. So as $\log r$ is increased, there are diminishing returns in energy compaction, but the cost increases linearly.

## 3.3 Quantized Matching Pursuit

Although matching pursuit has been applied to low bit rate compression problems [19, 23, 24, 25, 36], which inherently require coarse coefficient quantization, little work has been

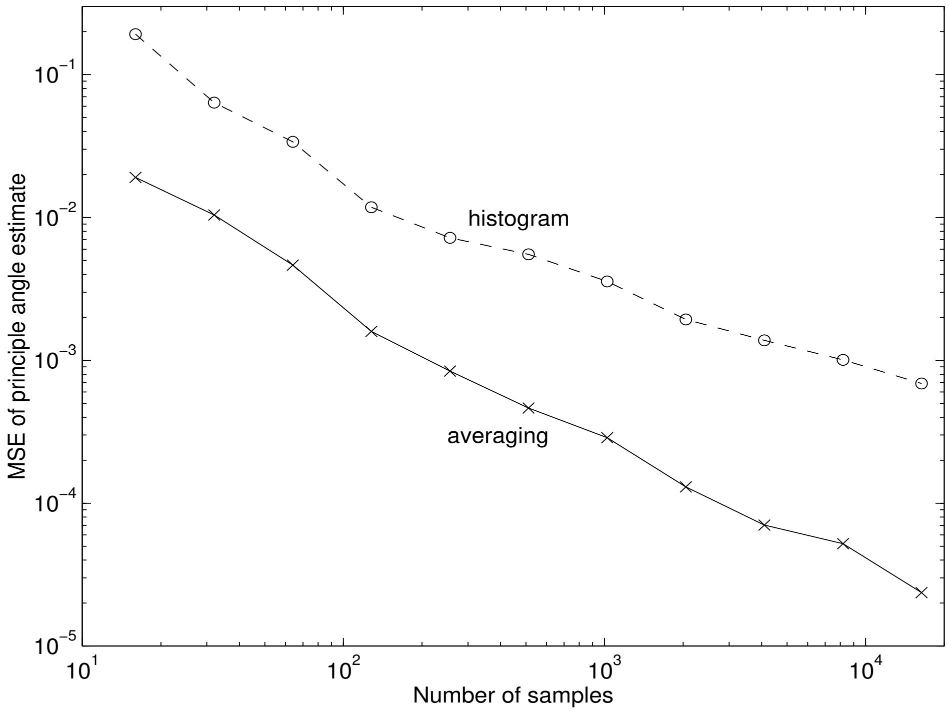



Figure 3.4: Principal axis estimation using matching pursuit for an $\mathbb{R}^2$-valued source.

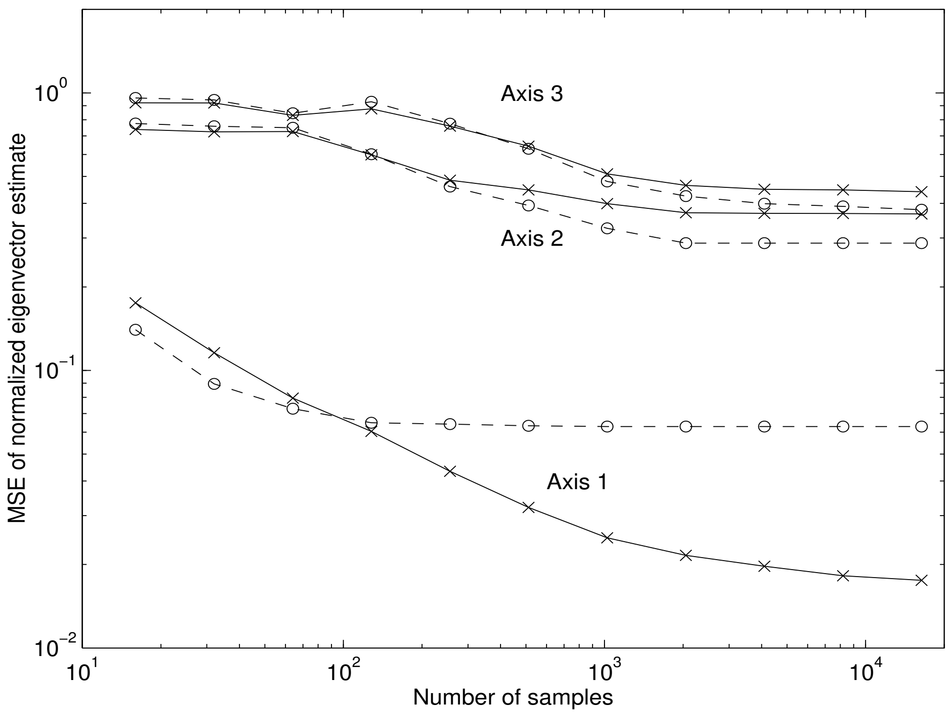



Figure 3.5: Principal axes estimation using matching pursuit for an $\mathbb{R}^4$-valued source

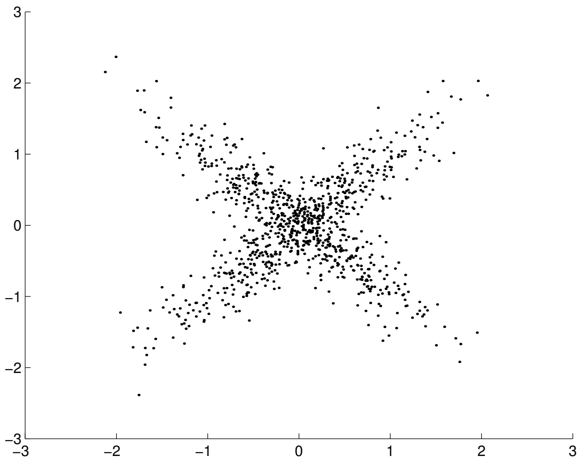



Figure 3.6: One thousand samples from a non-ellipsoidal source

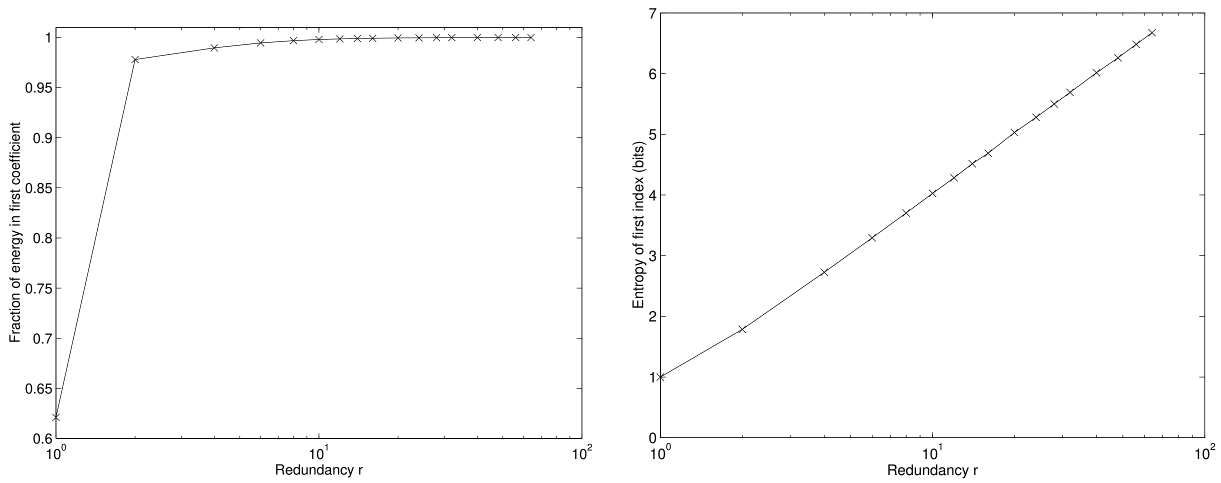



Figure 3.7: Energy compaction and index entropy as functions of redundancy $r$ for a non-ellipsoidal source.

done to understand the qualitative effects of coefficient quantization in matching pursuit. In this section we explore some of these effects. In §3.3.2, application of matching pursuit to compress an $\mathbb{R}^2$-valued source is considered in great detail. The highlight of the subsection is the generation of intricate partition diagrams. These partition diagrams demonstrate that matching pursuit expansions can be inconsistent. The issue of consistency in these expansions is explored in §3.3.3. The relationship between quantized matching pursuit and other vector quantization methods is discussed in §3.3.4.

### 3.3.1 Discussion

Coefficients are quantized in any computer implementation of matching pursuit. When the quantization is fine, it is generally safe to ignore this fact. For example, in all the simulations of §3.2, coefficient quantization does not make any qualitative differences. If the quantization is coarse, as it must be for moderate to low bit rate compression applications, the effects of quantization may be significant.

Define *quantized matching pursuit* to be matching pursuit with non-negligible quantization of the coefficients. We will denote the quantized coefficients by $\widehat{\alpha_i} = Q[\alpha_i]$. Note that quantization destroys the orthogonality of the projection and residual, so the analog of (3.4) does not hold, *i.e.*

$$\|R_i f\|^2 \neq \|R_{i+1} f\|^2 + \widehat{\alpha_i}^2.$$

Also, (3.5) does not hold.

We are assuming that the quantization of $\alpha_i$ occurs before the residual $R_{i+1}f$ is calculated, and that the quantized version is used in determining the residual so that quantization errors do not propagate to subsequent iterations. Since $\widehat{\alpha_i}$ must be determined before $\alpha_{i+1}$, it is implicit in this assumption is that the coefficient quantization is scalar.

For any particular application, there are several design problems: a dictionary must be chosen, scalar quantizers must be designed, and the number of iterations (or a stopping criterion) must be set. In principle, these could be jointly optimized for a given source distribution, distortion measure, and rate measure. In practice, this is an overly broad problem. In the following subsection, we will make several choices, some of them arbitrary.

### 3.3.2 A Detailed Example

Consider quantization of a source with a uniform distribution on $[-1, 1]^2$. Assume distortion is measured by squared Euclidean distance and rate is measured by codebook size. (Measuring rate by codebook size is natural when a fixed rate coder will be applied to the quantizer output, *i.e.* no entropy coding is used.) Also assume that two iterations will be performed with a four element dictionary. Other constraints will be set as needed.

We first choose a dictionary. Guided by symmetry, we choose

$$\mathcal{D} = \left\{ \begin{bmatrix} \cos\frac{(2k-1)\pi}{8} & \sin\frac{(2k-1)\pi}{8} \end{bmatrix}^T \right\}_{k=1}^{4}. \tag{3.9}$$

A first impulse may be to use

$$\mathcal{D} = \left\{ \begin{bmatrix} \cos\frac{(k-1)\pi}{4} & \sin\frac{(k-1)\pi}{4} \end{bmatrix}^T \right\}_{k=1}^{4}. \tag{3.10}$$

In a detailed analysis, (3.9) was determined to lead to a better design. Also, (3.10) is not symmetric with respect to the region of support of the distribution.

To begin with, assume that the quantization of coefficients will be fine. Then, since the dictionary is composed of pairs of orthogonal vectors, $\varphi_{k_0} \perp \varphi_{k_1}$. Thus once we have coded $k_0$, $k_1$ is determined for free. (As long as we are using a fine quantization assumption, we will actually force the $k_1$ to be selected such that $\varphi_{k_0} \perp \varphi_{k_1}$.) It is easy to see that $k_0$ will be uniformly distributed on $\{1, 2, 3, 4\}$; thus, with or without entropy coding, $k_0$ requires 2 bits.

We now design the quantizers. The p.d.f. of $\alpha_0$ can be explicitly calculated as

$$p_{\alpha_0}(y) = \begin{cases} 2(\sqrt{2}-1)|y| & |y| \le \frac{1}{2}\sqrt{2+\sqrt{2}} \\ -2(|y| - \sqrt{1+\sqrt{2}}) & \frac{1}{2}\sqrt{2+\sqrt{2}} < |y| \le \sqrt{1+\sqrt{2}} \\ 0 & \text{otherwise} \end{cases} . \tag{3.11}$$

If the dictionary was not symmetric, (3.11) would have to be conditioned on $k_0$. Since we are assuming fine quantization, the best codebook constrained quantizer for $\alpha_0$ can be found analytically using a compandor model [13]. The optimal quantizer is

$$\widehat{\alpha_0} = G^{-1} \circ Q_u \circ G(\alpha_0),$$

where

$$G(y) = \begin{cases} \frac{2^{1/3}}{(2+\sqrt{2})^{1/6}} \operatorname{sgn}(y)\, y^{4/3} & |y| \le \frac{\sqrt{2+\sqrt{2}}}{2} \\ \operatorname{sgn}(y) \left[ \sqrt{1+\frac{1}{\sqrt{2}}} - \frac{2^{1/3}}{(2-\sqrt{2})^{1/6}} \left( y \operatorname{sgn}(y) - \sqrt{1+\frac{1}{\sqrt{2}}} \right)^{4/3} \right] & \text{otherwise} \end{cases} ,$$

and $Q_u$ is a uniform quantizer.

Given $\alpha_0$, the distribution of $\alpha_1$ is uniform on $[-|\alpha_0|, |\alpha_0|]$. Since the quantization of $\widehat{\alpha_0}$ is fine, the distribution given $\widehat{\alpha_0}$ is approximately the same. Thus the optimal quantizer for $\alpha_1$ is uniform.

We have yet to decide how to divide our bit rate between $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$. Since $\varphi_{k_0} \perp \varphi_{k_1}$, the total distortion is simply the sum of the distortions created by each quantization. We can thus minimize distortion for a fixed rate by Lagrangian methods.

If we impose the constraint that the rate for $\widehat{\alpha_1}$ must be constant, we get a codebook as in Figure 3.8(a). On the other hand, if we allow the rate for $\widehat{\alpha_1}$ to be conditioned on $\widehat{\alpha_0}$, we get a codebook as in Figure 3.8(b). (Actually, these codebooks are for the dictionary (3.10), but the observations and conclusions are still clear.) The two codebooks have 906 and 900 elements, respectively, so they give approximately equal rates. The codebook in Figure 3.8(b) gives lower distortion, as is clear from the more uniform distribution of code vectors.

When the rate for $\widehat{\alpha_1}$ depends on $\widehat{\alpha_0}$, the Lagrangian optimization implies that the number of quantization levels for $\alpha_1$ should be proportional to $p_{\widehat{\alpha_0}}$. Using a codebook size of 304 and choosing the proportionality constant appropriately yields the codebook and partition shown in Figure 3.9. This codebook gives approximately 0.1561 bits worse performance than simple uniform scalar quantization. (Recall that this includes two bits for $k_0$.) Of course, this should not be too discouraging because the region of support and distribution of the source in this simulation are tailor-made for uniform scalar quantization. As we will see in §3.4, matching pursuit tends to be effective when the number of iterations is less than the dimension of the space.

Figure 3.9 should be seen as a first approximation to the type of partition created by matching pursuit because we forced $\varphi_{k_0} \perp \varphi_{k_1}$. (This was part of the fine quantization assumption.) Let us now remove the fine quantization assumption and allow the source to have an arbitrary distribution on $\mathbb{R}^2$. Even with a known distribution, it is difficult to find

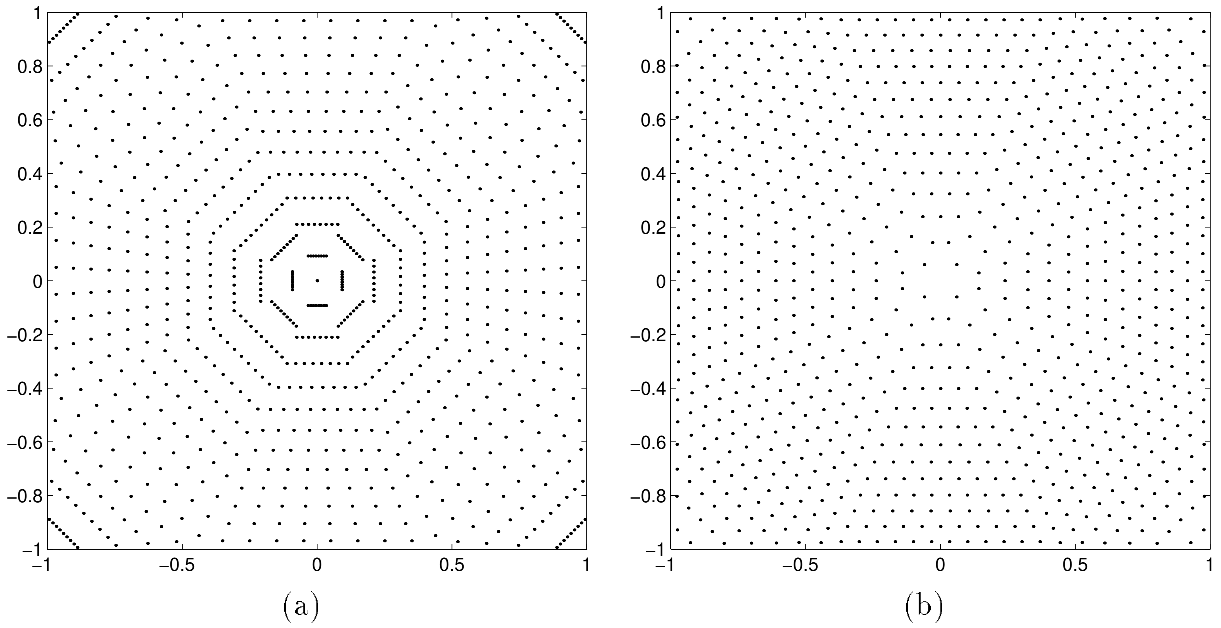



Figure 3.8: Codebook elements for quantization of a source with uniform distribution on $[-1, 1]^2$. (a) Fixed rate for $\widehat{\alpha_1}$. (b) Rate for $\widehat{\alpha_1}$ conditioned on $\widehat{\alpha_0}$.

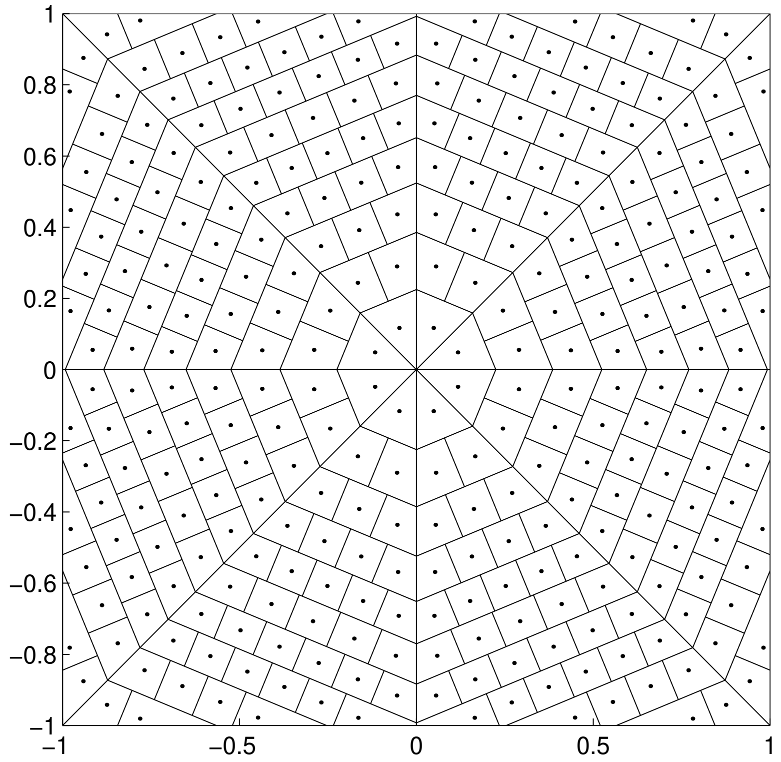



Figure 3.9: Partitioning of $[-1, 1]^2$ by matching pursuit with four element dictionary. A fine quantization assumption is used.

analytical expressions for optimal quantizers without using a fine quantization assumption. Since we wish to use fixed, untrained quantizers, we will use uniform quantizers for $\alpha_0$ and $\alpha_1$. Since it will still generally be true that $\varphi_{k_0} \perp \varphi_{k_1}$, it makes sense for the quantization stepsizes for $\alpha_0$ and $\alpha_1$ to be equal.

The partitions generated by matching pursuit are very intricate. Figure 3.10 shows the partitioning of the first quadrant when zero is a quantizer boundary value, *i.e.* the quantizer boundary points are $\{m\Delta\}_{m\in\mathbb{Z}}$ and reconstruction points are $\{(m+\frac{1}{2})\Delta\}_{m\in\mathbb{Z}}$ for some quantization stepsize $\Delta$. The yellow lines denote the partitions induced by selection of $k_0$. Then $\alpha_0$ is quantized, giving the cyan boundaries. Recall that the residue $R_1 f$ is not necessarily orthogonal to $\varphi_{k_0}$. Thus the selection of $k_1$ introduces the magenta boundaries. Finally, the red boundaries come from quantizing $\alpha_1$. In Figure 3.10, most of the cells are squares, but there are also some smaller cells. Unless the source distribution happens to have high density in the smaller cells, the smaller cells are inefficient in a rate-distortion sense. The fraction of cells that are not square $\to 0$ as $\Delta \to 0$.

The partition is qualitatively different when the quantizer boundary points are $\{(m+\frac{1}{2})\Delta\}_{m\in\mathbb{Z}}$ and reconstruction points are $\{m\Delta\}_{m\in\mathbb{Z}}$. The partition is shown in Figure 3.11. The colors are the same as in Figure 3.10. The dotted magenta lines show boundaries that are created by choice of $k_1$ but are not important because $\widehat{\alpha_1} = 0$. (Similarly for the dotted yellow line.) This partition also has mostly square cells. Compared to Figure 3.10, there are fewer of the "bad" small cells. As before, the fraction of non-square cells vanishes as $\Delta \to 0$.

The qualitative difference between Figure 3.9 and Figures 3.10–3.11 is due to the fact that the latter result from more constraints. The partition of Figure 3.9 arises from specifying $k_0$, $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$, with $k_1 \equiv k_0 + 2 \pmod 4$. The partitions of Figures 3.10–3.11 show the result of adding an additional degree of freedom in $k_1$.

These examples illustrate that there are many design parameters within the matching pursuit framework. Optimizing these parameters requires a measure of optimality and knowledge of the source p.d.f. Figures 3.9–3.11 show that the partitions generated by matching pursuit look quite different than those generated by a quantized frame expansion (see Figure B.1), of which independent scalar quantization is a special case.

### 3.3.3 Consistency in Quantized Matching Pursuit

When consistency was previously considered in §2.2.3, the problem arose from having a representation in $\mathbb{C}^M$ and attempting to estimate a reconstruction in $\mathbb{C}^N$. There is a possibility of inconsistency in any framework with non-orthogonal linear constraints. We will see that a matching pursuit representation implicitly contains many linear constraints and that inconsistency is not uncommon.

Suppose $p$ iterations of matching pursuit are performed with the dictionary $\mathcal{D}$. The output of the (quantized) matching pursuit algorithm is

$$\{k_0,\, \widehat{\alpha_0},\, k_1,\, \widehat{\alpha_1},\, \ldots,\, k_{p-1},\, \widehat{\alpha_{p-1}}\}. \tag{3.12}$$

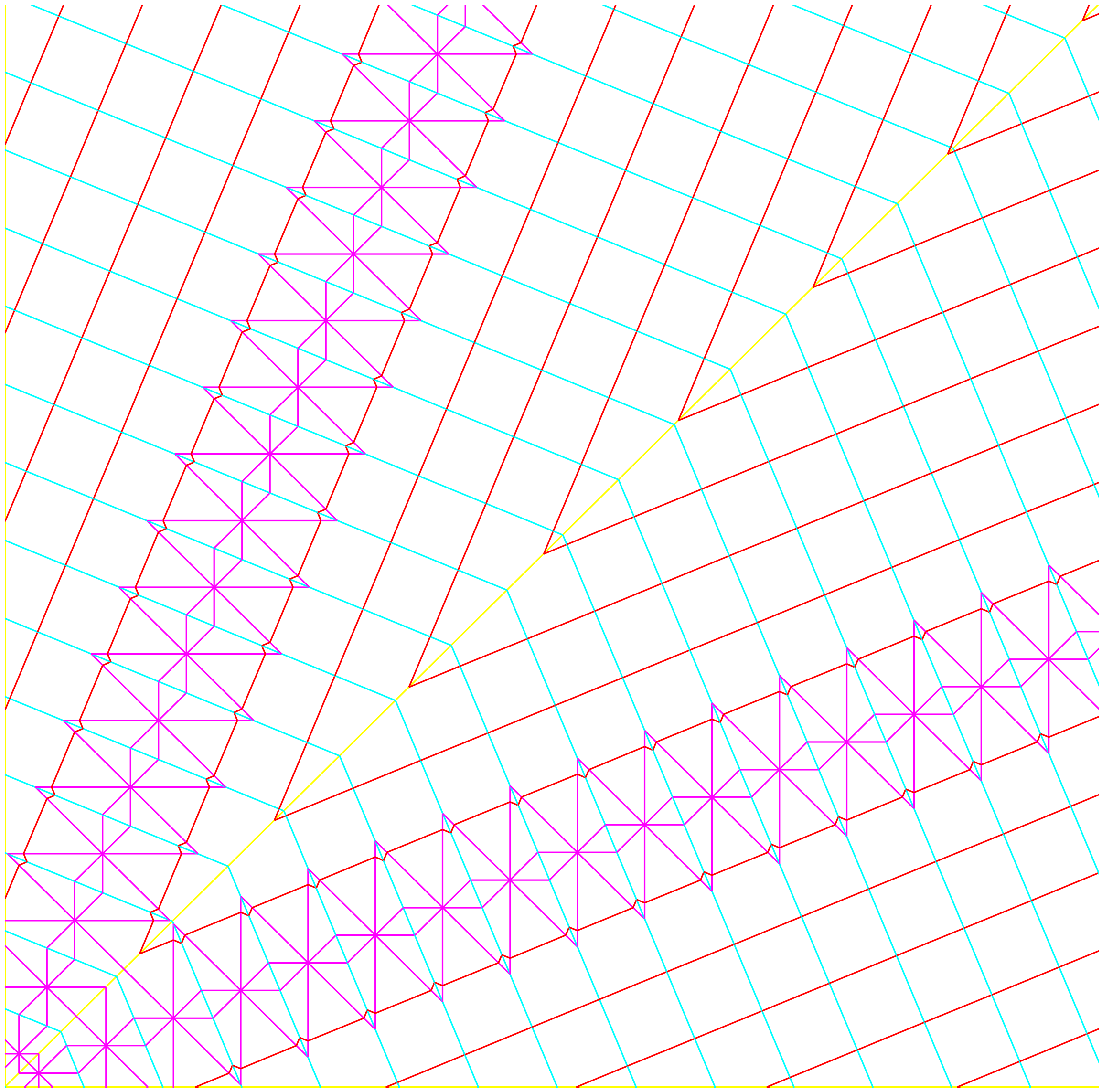


Figure 3.10: Partitioning of $\mathbb{R}^2$ by matching pursuit with four element dictionary. Zero is a quantizer boundary value.

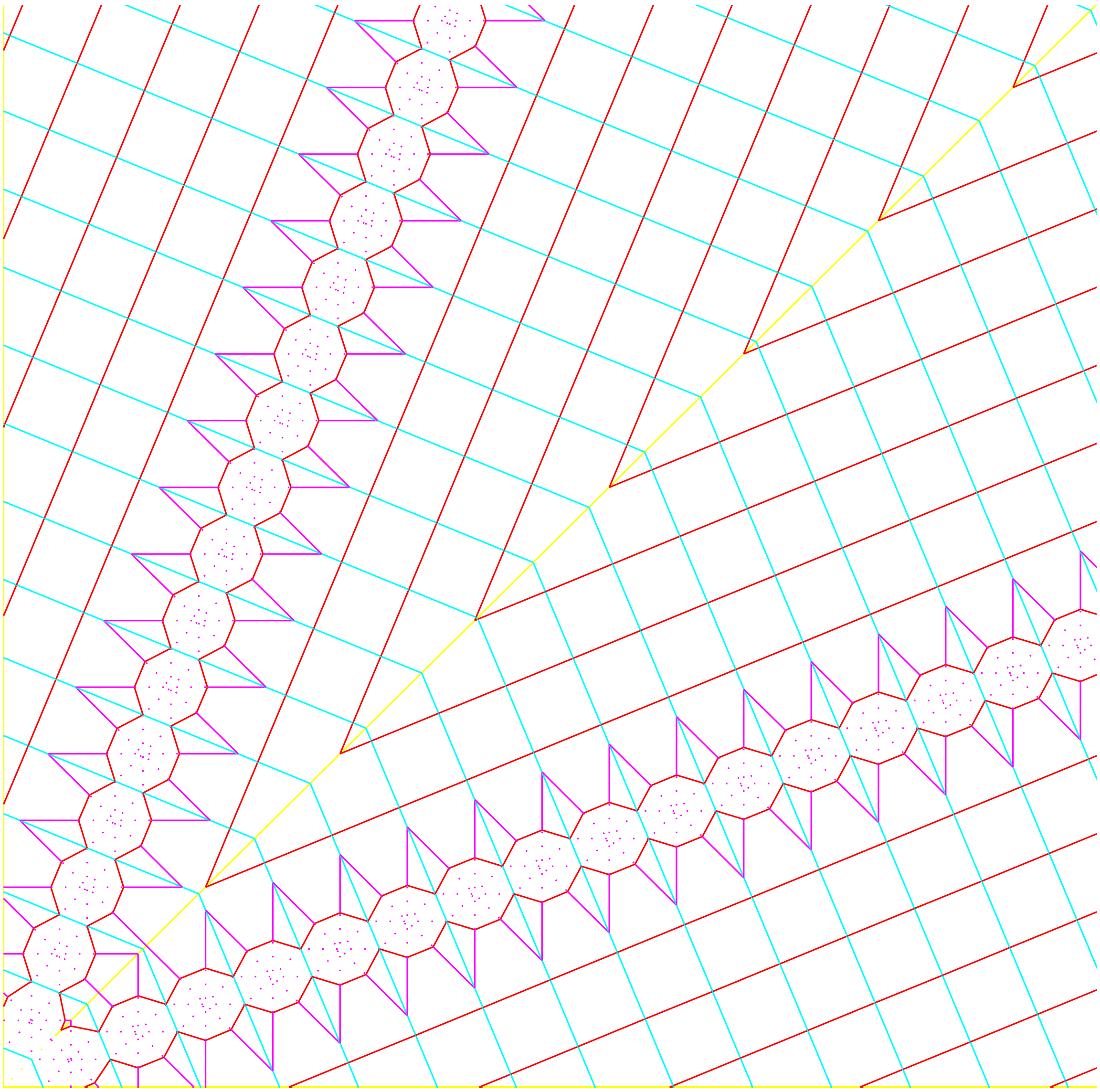


Figure 3.11: Partitioning of $\mathbb{R}^2$ by matching pursuit with four element dictionary. Zero is a quantizer reconstruction value.

(There is nothing consistent or inconsistent about this set.) The standard reconstruction is

$$\hat{f} = \sum_{i=0}^{p-1} \widehat{\alpha_i} \varphi_{k_i}. \tag{3.13}$$

Denote the output of matching pursuit (with the same dictionary and quantizers) applied to $\hat{f}$ by

$$\{k_0', \widehat{\alpha_0}', k_1', \widehat{\alpha_1}', \ldots, k_{p-1}', \widehat{\alpha_{p-1}}'\}.$$

If

$$k_i = k_i' \text{ and } \widehat{\alpha_i} = \widehat{\alpha_i}' \tag{3.14}$$

for $i = 0, 1, \ldots, p-1$, we say that $\hat{f}$ is a *strictly consistent* estimate. If (3.14) holds except possibly that $k_i \neq k_i'$ for some $i$ for which $\widehat{\alpha_i} = \widehat{\alpha_i}' = 0$, we say that $\hat{f}$ is a *loosely consistent* estimate. The second definition is included because a reasonable coding scheme might discard $k_i$ if $\widehat{\alpha_i} = 0$.

The crucial point is that there is more information in (3.12), along with $\mathcal{D}$ and knowledge of the working of matching pursuit, than there is in $\hat{f}$. In particular, (3.12) gives a set of linear inequality constraints that defines a partition cell in which $f$ lies. $\hat{f}$ is an estimate of $f$ that does not necessarily lie in this cell.

Let us now list the complete set of constraints implied by (3.12). For notational convenience, we assume uniform scalar quantization of the coefficients with stepsize $\Delta$ and midpoint reconstruction. The selection of $k_0$ implies

$$|\langle \varphi_{k_0}, f \rangle| \geq |\langle \varphi, f \rangle|, \quad \forall\, \varphi \in \mathcal{D}. \tag{3.15}$$

For each element of $\mathcal{D} \setminus \{\varphi_{k_0}\}$, (3.15) specifies a half-space constraint with boundary plane passing through the origin. The intersection of these constraints is thus two infinite pyramids situated symmetrically with their apexes at the origin. The value of $\widehat{\alpha_0}$ gives the constraint

$$\langle \varphi_{k_0}, f \rangle \in \left[\widehat{\alpha_0} - \frac{\Delta}{2}, \widehat{\alpha_0} + \frac{\Delta}{2}\right].$$

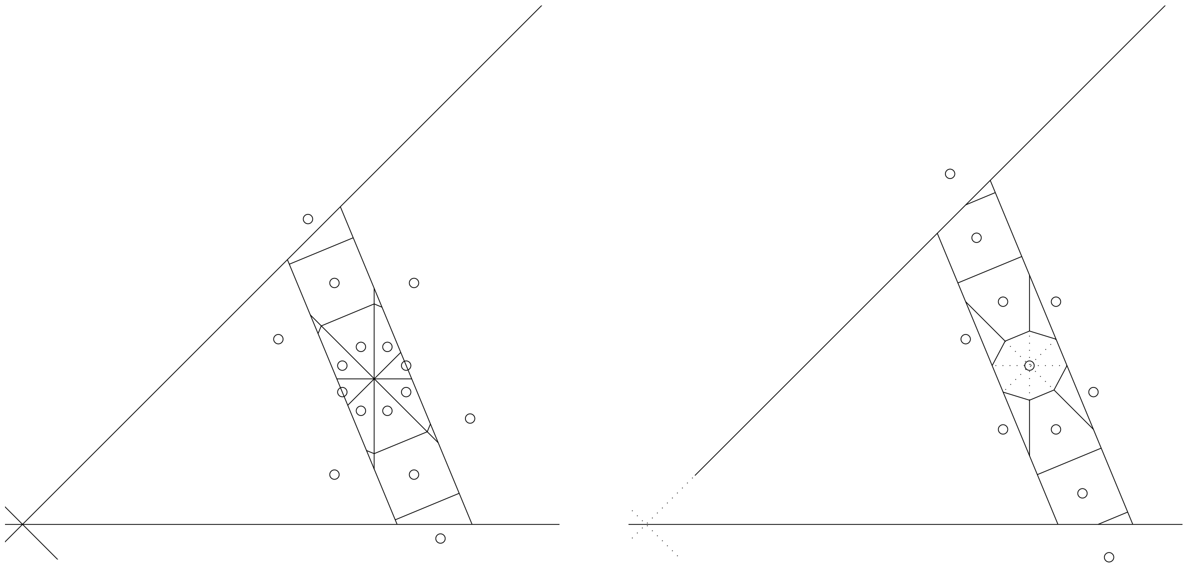
This specifies a pair of planes, perpendicular to $\varphi_{k_0}$, between which $f$ must lie. At the $(i-1)$st step, the selection of $k_i$ gives the constraints

$$\left|\left\langle \varphi_{k_i}, f - \sum_{\ell=0}^{i-1} \widehat{\alpha_\ell} \varphi_{k_\ell} \right\rangle\right| \geq \left|\left\langle \varphi, f - \sum_{\ell=0}^{i-1} \widehat{\alpha_\ell} \varphi_{k_\ell} \right\rangle\right|, \quad \forall\, \varphi \in \mathcal{D}. \tag{3.16}$$

This defines $M-1$ linear half-space constraints with boundaries passing through $\sum_{\ell=0}^{i-1} \widehat{\alpha_\ell} \varphi_{k_\ell}$. As before, these define two infinite pyramids situated symmetrically with their apexes at $\sum_{\ell=0}^{i-1} \widehat{\alpha_\ell} \varphi_{k_\ell}$. Then $\widehat{\alpha_i}$ gives

$$\left\langle \varphi_{k_i}, f - \sum_{\ell=0}^{i-1} \widehat{\alpha_\ell} \varphi_{k_\ell} \right\rangle \in \left[\widehat{\alpha_i} - \frac{\Delta}{2}, \widehat{\alpha_i} + \frac{\Delta}{2}\right]. \tag{3.17}$$

This again specifies a pair of planes, this time perpendicular to $\varphi_{k_i}$, between which $f$ must lie.



(a) (b)

Figure 3.12: (a) Portion of partition of Figure 3.10 with reconstruction points marked. (b) Portion of partition of Figure 3.11 with reconstruction points marked.

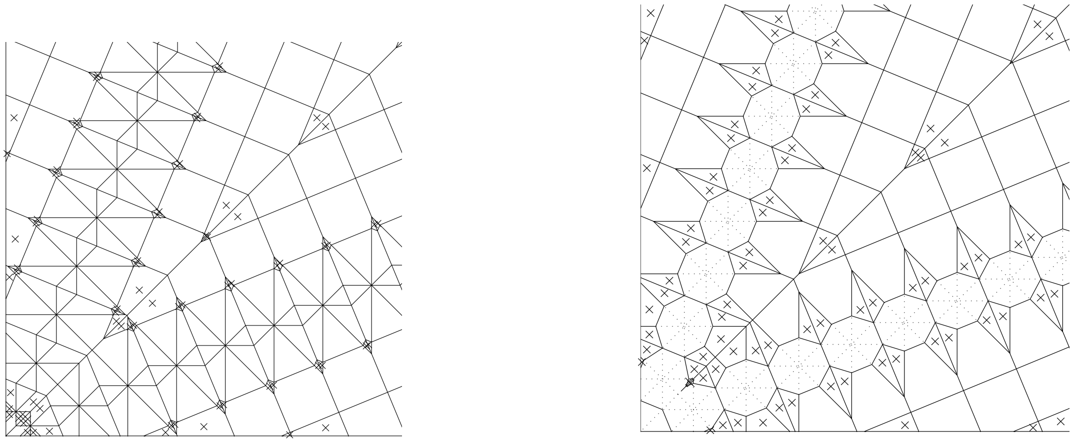
By being explicit about the constraints, we see that all of the constraints are linear, so the partition cell defined by (3.12) is convex. Thus by using an appropriate projection operator, one can find a strictly consistent estimate from any initial estimate. In practice, finding such a projection operator may be difficult.

The quantization of $\mathbb{R}^2$ considered in §3.3.2 gives concrete examples of inconsistency. Recall the partitions of Figures 3.10 and 3.11. The reconstruction points were not marked on these diagrams because the correspondence between cells and reconstruction points would not have been clear. Figures 3.12(a) and 3.12(b) depicts parts of these partitions with reconstruction points marked with circles. These show that matching pursuit reconstructions are not always consistent. Figures 3.13(a) and 3.13(b) are copies of Figures 3.10 and 3.11 with cells that lead to inconsistent reconstructions marked with $\times$'s.

Experiments were performed to assess how the probability of an inconsistent estimate depends on $\mathcal{D}$, $r$, and $\Delta$. The loose sense of consistency was used in all the experiments.

The first set of experiments involved quantizing an $\mathbb{R}^2$-valued source with the $\mathcal{N}(0, I)$ distribution. With $\mathcal{D}$ as in (3.8), $M$ was varied between 2 and 256 while $\Delta$ was varied between $10^{-1.9}$ and $10^{0.3}$. Figure 3.14 shows the probability of inconsistency as a function of $M$ and $\Delta$. The probability of inconsistency is significant! The surface is rather complicated, but we can identify two trends: the probability of inconsistency goes to zero as $M$ is increased and as $\Delta \to 0$. This can be more clearly seen from two "slices" of a similar surface obtained with $\mathcal{D}$ as in (3.7). The slices are shown in Figure 3.15.

To explore the dependence on $\mathcal{D}$, experiments were performed for quantizing an $\mathbb{R}^5$-valued source with the $\mathcal{N}(0, I)$ distribution. The consistency of reconstruction was checked



(a) (b)

Figure 3.13: (a) Partition of Figure 3.10 with regions leading to inconsistent reconstructions marked. (b) Partition of Figure 3.11 with regions leading to inconsistent reconstructions marked.

for two iteration expansions. Dictionary sizes of $M = 25$, 50, 75, 100, and 125 were used. The results are shown in Figures 3.16 and 3.17. In Figure 3.16, the dictionaries used are those corresponding to oversampled A/D conversion as given in (2.9). Figure 3.17 was generated using dictionaries of maximally spaced points [16]. For both types of dictionaries, the probability of inconsistency goes to one for very coarse quantization and goes to zero as $\Delta \to 0$. The qualitative difference between the curves indicates that there are complicated geometric factors involved that are at this time beyond our understanding.

### 3.3.4 Relationship to Vector Quantization

Given a vector in $\mathbb{R}^N$, quantized matching pursuit produces an estimate from a countable set. (If the quantizers have bounded ranges, the estimate is from a finite set.) Hence quantized matching pursuit can be described as a vector quantization (VQ) method; we would like to understand its place among the many existing VQ methods.

A single iteration of matching pursuit is very similar to shape-gain VQ, which was introduced in [2]. In shape-gain VQ, a vector $x \in \mathbb{R}^N$ is separated into a *gain*, $g = \|x\|$ and a *shape*, $s = x/g$. A shape $\hat{s}$ is chosen from a shape codebook $\mathcal{C}_s$ to maximize $\langle x, \hat{s}\rangle$. Then a gain $\hat{g}$ is chosen from a gain codebook $\mathcal{C}_g$ to minimize $(\hat{g} - \langle x, \hat{s}\rangle)^2$. The similarity is clear with $\mathcal{C}_s$ corresponding to $\mathcal{D}$ and $\mathcal{C}_g$ corresponding to the quantizer for $\alpha_0$. Obtaining a good approximation in shape-gain VQ requires that $\mathcal{C}_s$ forms a very dense subset of $S^{N-1}$, the surface of the unit sphere in $\mathbb{R}^N$. The area of $S^{N-1}$ increases exponentially with $N$, making it difficult to use shape-gain VQ in high dimensional spaces. A multi-iteration application of matching pursuit can be seen as a cascade form of shape-gain VQ.

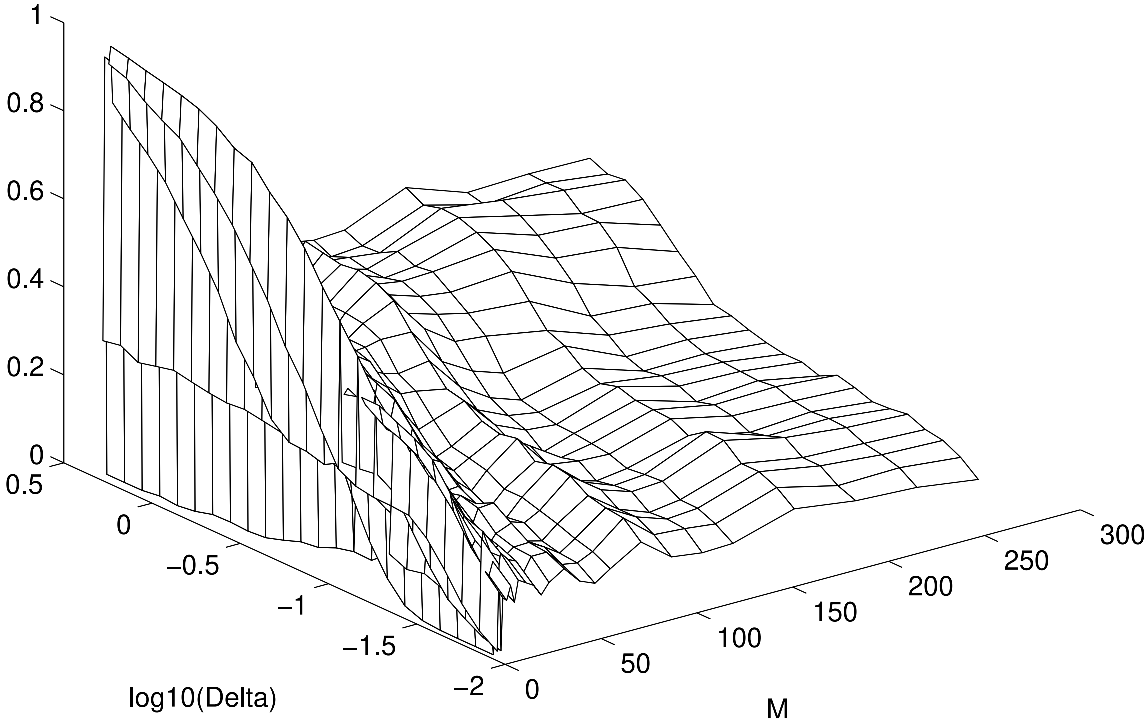



Figure 3.14: Probability of inconsistent reconstruction for an $\mathbb{R}^2$-valued source as a function of $M$ and $\Delta$.

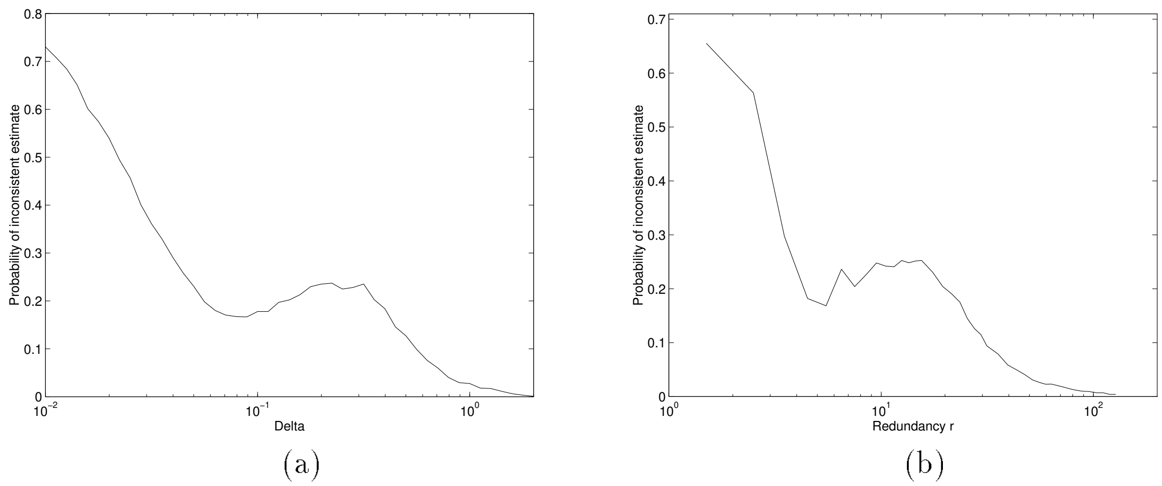



Figure 3.15: Probabilities of inconsistent reconstruction for an $\mathbb{R}^2$-valued source. (a) $M = 11$, $\Delta$ varied. (b) $M$ varied, $\Delta = 0.1$.

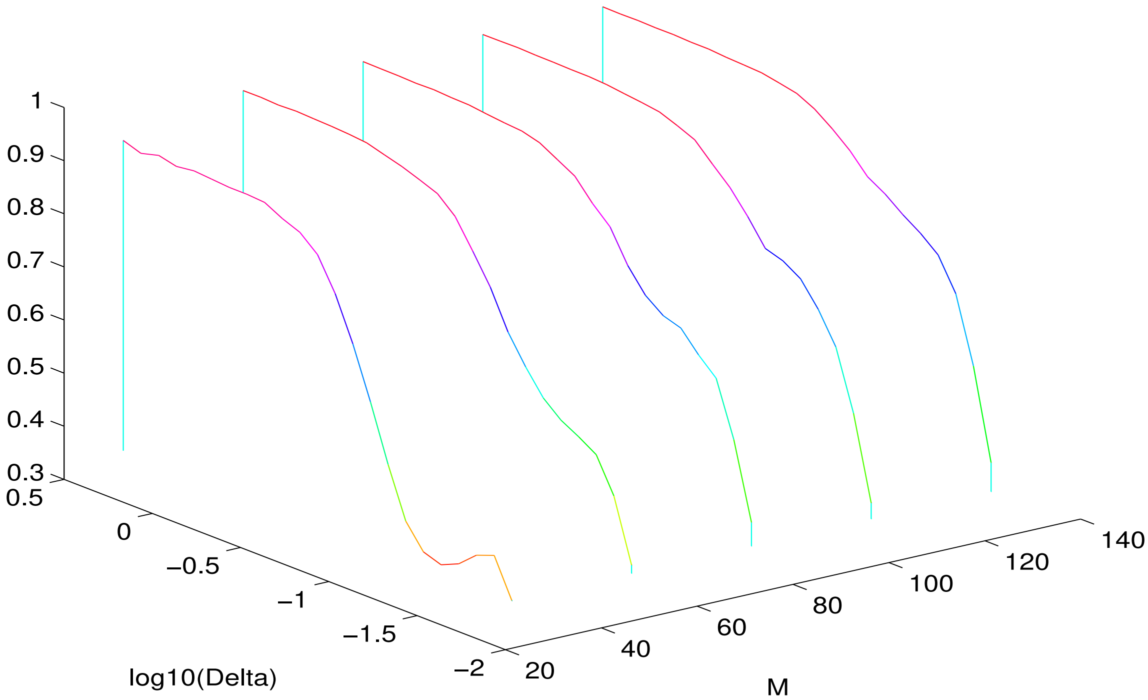



Figure 3.16: Probabilities of inconsistent reconstruction for an $\mathbb{R}^5$-valued source. Dictionaries correspond to oversampled A/D conversion.

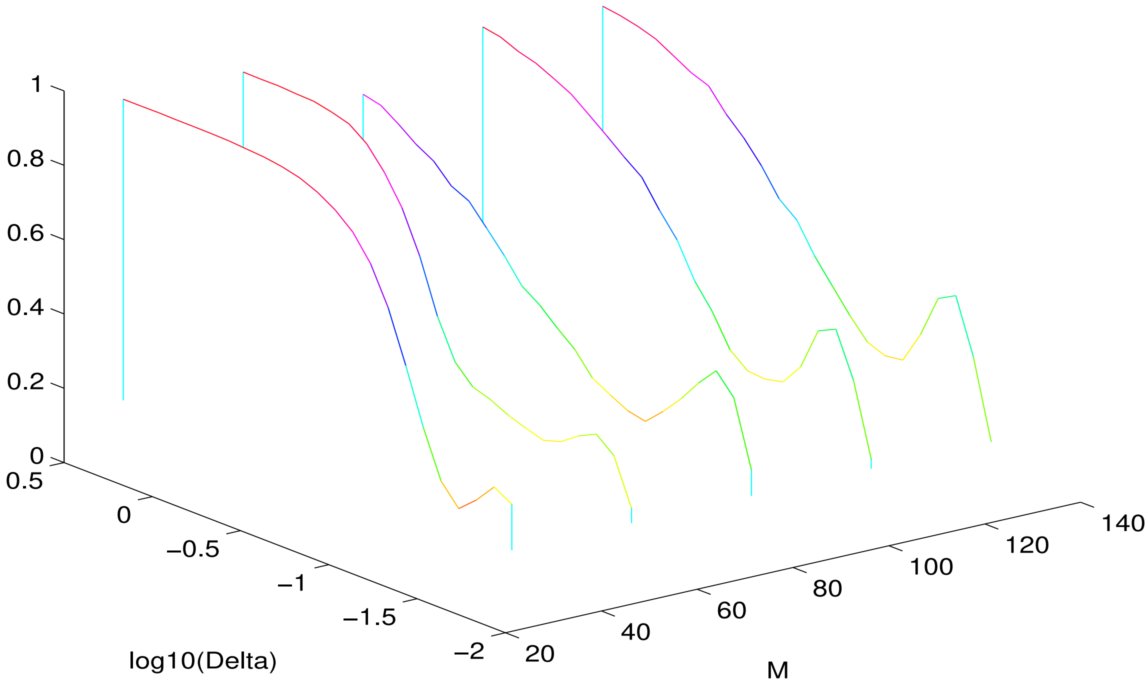



Figure 3.17: Probabilities of inconsistent reconstruction for an $\mathbb{R}^5$-valued source. Dictionaries composed of maximally space points on the unit sphere.

Although our discussion has been in the language of linear expansions, matching pursuit can be seen to give partition cells and reconstruction points. For an optimal VQ codebook, the centroid condition must hold: the reconstruction value for a partition cell must be the centroid of the cell with respect to the probability density of the source. Even if we make a simplifying assumption such as a uniform distribution of the source, the codebook given by matching pursuit (assuming reconstruction according to (3.13)) does not satisfy the centroid condition. This is shown in Figures 3.9 and 3.12, where inconsistency is an extreme case of non-centroid reconstruction. Viewed in this way alone, matching pursuit is a bad vector quantization method. However, recall that if optimal trained VQ is used, the centroid values (reconstruction points) must all be stored. By basing the codebook on linear expansions, we are considerably lowering the storage requirements. Referring to Figures 3.9 and 3.12, centroids could be calculated with respect to a uniform distribution and used as reconstruction points, replacing (3.13). The structure of the partition would allow the reconstruction points to be stored efficiently.

## 3.4 A General Vector Compression Algorithm Based on Frames

This section explores the efficacy of using matching pursuit as an algorithm for lossy compression for vectors in $\mathbb{R}^N$. Most lossy compression can be viewed as compressing vectors in $\mathbb{R}^N$, although the source distribution will depend on the application. The application may also give coding constraints (such as requiring a fix bit rate) and complexity constraints, and may suggest a relevant distortion metric. Here we will measure rate by entropy, thus implicitly allowing variable bit rates, and measure distortion by MSE. Experimental results will be given for autoregressive sources, but distributional knowledge will not be used in the design.

### 3.4.1 Design Considerations

With no distributional assumptions, we expect the best performance with a dictionary that is "evenly spaced" on the unit sphere or a hemisphere. We are purposely vague about the meaning of evenly spaced, since the importance of this is not clear. For simplicity, the inner product quantization is uniform. It is unlikely that any other fixed quantization would do better over a large class of source distributions. Furthermore, the quantization stepsize $\Delta$ is constant across iterations. This is consistent with equal weighting of error in each direction.

In our earlier examples, three methods for generating dictionaries have been used. In $\mathbb{R}^2$, dictionaries were formed from roots of unity as in (3.7) and (3.8). In higher dimensions, dictionaries were formed from sets of maximally spaced points on the unit sphere [16] or from a Fourier transform-like set as in (2.9). We introduce one more method for generating dictionaries. The corners of the hypercube $[-\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}]^N$ form a set of $2^N$ symmetric points on the surface on the unit sphere in $\mathbb{R}^N$. Taking the subset of points that have a positive first coordinate gives a frame of size $2^{N-1}$. Properties of the dictionaries that will be used in the remainder of the section are summarized in Table 3.1.

| | |
|---|---|
| I. | **DFT-like set given by (2.9)**<br>Advantages:<br>• Inner products can be found with an FFT-like algorithm<br>• No need to store dictionary<br>Comment:<br>• Dictionary elements lie in the intersection of the unit sphere with the plane $x_1 = \frac{1}{\sqrt{N}}$. |
| II. | **Maximally spaced points on the unit sphere from [16]**<br>Disadvantages:<br>• Dictionary must be stored.<br>• Known only for $N = 3, 4, 5$, and $M \leq 130$. |
| III. | **Corners of hypercube**<br>Advantages:<br>• Inner products can be found with additions and subtractions only (no multiplications).<br>• Can choose $k_i$ without calculating *any* inner products. (Signs of components of $R_i f$ determine which dictionary element should be chosen.)<br>• No need to store dictionary<br>Disadvantage:<br>• No flexibility in choice of $M$ for fixed $N$. |

Table 3.1: Summary of dictionaries used in compression experiments

### 3.4.2 Experimental Results

The experiments all involve quantization of a zero mean Gaussian AR source with correlation coefficient $\rho = 0.9$. Source vectors are generated by forming blocks of $N$ samples. Rate is measured by summing the (scalar) sample entropies of $k_0, k_1, \ldots, k_{p-1}$ and $\widehat{\alpha_0}, \widehat{\alpha_1}, \ldots, \widehat{\alpha_{p-1}}$, where $p$ is the number of iterations of the algorithm.

Figure 3.18 shows the $D(R)$ points obtained using Method I with $N = 9$. The dictionary redundancy ratio is $r = 8$. The dotted curves correspond to varying $p$, with the leftmost and rightmost curves corresponding to $p = 1$ and $p = 3$, respectively. The points along each dotted curve correspond to various values of $\Delta$. The solid curve shows the performance of independent quantization in each dimension.

The lower boundary of the region bounded below by one or more dotted curves is the best R-D performance that can be achieved with this dictionary through the choice of $p$ and $\Delta$. The simulation results show that matching pursuit performs as well or better than independent scalar quantization for rates up to about 2.2 bits per source sample.

The simulation described above does not explore the significance of the $r$ parameter. Simulations as above were performed with $r$ ranging from 1 to 256. Redundancy factors between 2 and 8 resulted in the best performance.

A large fraction of the rate comes from coding the indices. In an attempt to exploit the fact that $\varphi_{k_i}$ and $\varphi_{k_{i+1}}$ are often nearly orthogonal, experiments were also performed where

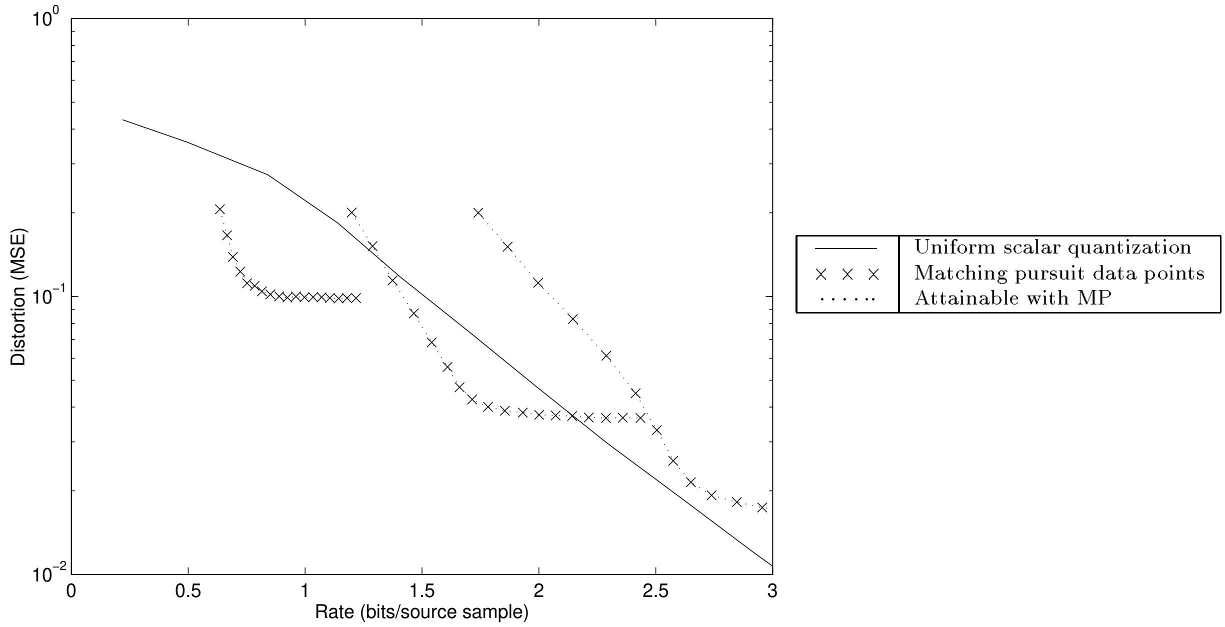



Figure 3.18: R-D performance of matching pursuit quantization with one to three iterations. ($N = 9$, $r = 8$, dictionary of type I.)

a single entropy code was applied for $(k_0, k_1, \ldots, k_{p-1})$. We refer to this as vector entropy coding of the indices. The entropy coding of the coefficients remained scalar. Figure 3.19 shows the results of experiments with a dictionary of type II. A dictionary of size $M = 8$ is used in $\mathbb{R}^4$. The dashed curve results from using matching pursuit with scalar entropy coding of the indices. The dash-dot curve shows the improvement resulting from vector entropy coding of the indices. The "knees" in these curves correspond to rates at which the optimal number of iterations changes. For comparison, the solid curve gives the performance of scalar quantization with scalar entropy coding. Replacing the scalar entropy coding by vector entropy coding gives the dotted curve.

At rates up to about 1.4 bits per source sample, matching pursuit quantization outperforms scalar quantization, even with vector entropy coding. (At these rates, the index entropy coding method is immaterial because it is best to have only one iteration.) Comparing to simple scalar quantization with scalar entropy coding, matching pursuit performs about as well or better over the range of rates considered, up to 3.5 bits per source sample.

This simulation shows that vector entropy coding of indices gives modestly improved performance at high rates. At high rates it may at first appear that independent quantization with vector entropy coding is far superior to other methods, but we must consider the complexity involved in the entropy coding. Consider operation at 2 bits/sample. The optimal number of matching pursuit iterations is two, so the vector entropy code for the indices has $8^2 = 64$ symbols. The entropy codes for $\alpha_0$ and $\alpha_1$ have 20 and 6 symbols, respectively. On the other hand, the vector entropy code for the independently quantized vectors has $14^4 = 38416$ symbols. Thus with limited computational resources, the matching pursuit quantizer may be the best choice.

Figure 3.20 shows simulations results using the type III dictionary with $N = 8$ ($M = 2^7$).

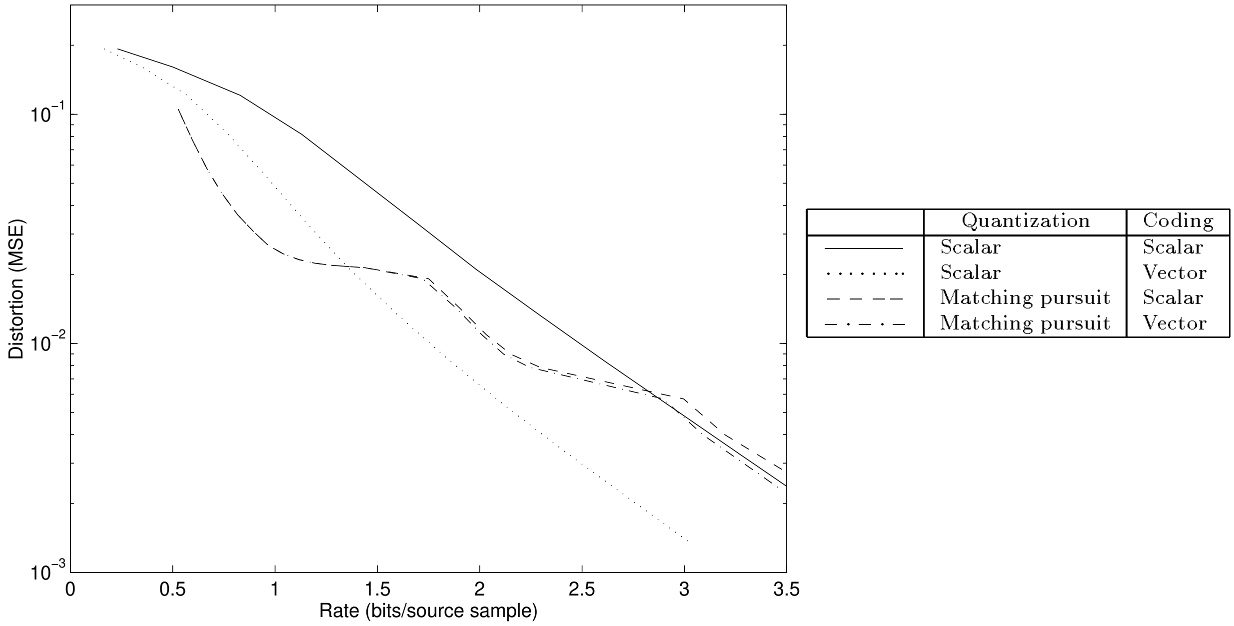



Figure 3.19: Simulation results for $N = 4$, $M = 8$ with dictionary of type II.

The curve types have the same correspondence as in Figure 3.19; the results are qualitatively similar.

### 3.4.3 A Few Possible Variations

The experiments of the previous subsection are the tip of the iceberg in terms of the possible design choices. In this subsection, a few possible variations are presented along with plausibility arguments for their application.

An obvious area to study is the design of dictionaries. For static, untrained dictionaries, issues of interest include not only R-D performance, but also storage requirements, complexity of inner product computation, and complexity of largest inner product search.

Looking at the dictionary design problem from a VQ standpoint, the first impulse is to train the dictionary using given training data. Davis [7, Ch. 8] has applied a Lloyd-type algorithm to optimize a dictionary to minimize

$$D = E\left[\left\| f - \sum_{i=0}^{L-1} \alpha_i \varphi_{k_i} \right\|^2\right]$$

for some fixed $L$. We would be interested in the case where the coefficients are quantized and the minimization is of $D + \lambda R$, where $R$ is a rate measure and $\lambda$ is a Lagrange multiplier. The result of such an optimization must have worse performance than a general entropy-constrained VQ design because the matching pursuit algorithm imposes a constraint on the codebook structure. However, the codebook structure may provide computational advantages, so this is worthy of investigation.

Another possibility in dictionary design is to adapt the dictionary by augmenting it with samples from the source. (Dictionary elements might also be deleted or adjusted.) This

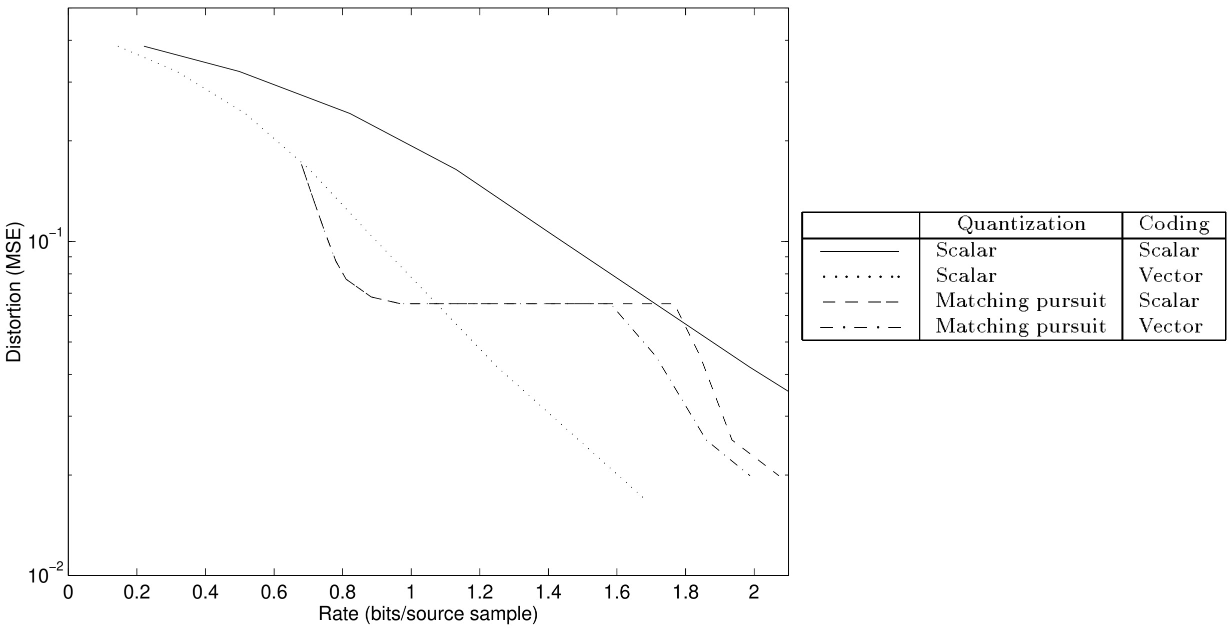



Figure 3.20: Simulation results for $N = 8$ with dictionary of type III.

would be in the spirit of the Lempel-Ziv algorithm. The decoder would have to be aware of changes in the dictionary, but depending on the nature of the adaptation, this may come without a rate penalty.

There is no *a priori* reason to use the same dictionary at every iteration. Given a $p$ iteration estimate, the entropy of $k_p$ becomes a limiting factor in adding the results of an additional iteration. To reduce this entropy, it might be useful to use coarser dictionaries as the iterations proceed.

In our experiments, we averaged results for quantizing many samples with some fixed number of iterations. Instead of having a fixed number of iterations, it may be useful to use a stopping criterion based on the energy of the residue. This would create a guaranteed upper bound on the error and might have a favorable impact in an R-D sense.

The experimental results that have been presented are based on entropy coding each $\widehat{\alpha_i}$ independently of the indices, which are in turn coded either separately or as a vector. There are at least three other ways to entropy code:

1. Separately code the pair $(k_i, \alpha_i)$ for each $i$;

2. Jointly code all of the indices and jointly code all of the coefficients;

3. Jointly code all of the indices and coefficients together.

Joint entropy coding of vectors increases complexity and, because of problems of statistical significance, makes simulating very time-consuming. A final coding variation, which was mentioned in §3.3.3, is to discard the indices that correspond to zero quantized coefficients. This should give a modest reduction in rate.

For a broad class of source distributions, the distributions of the $\alpha_i$'s will have some common properties because they are similar to order statistics. For example, the probability density of $\alpha_0$ will be small near zero. This could be exploited in quantizer design in future work. Finally, rate-distortion performance might be improved by using quantizers with overload regions.

# Chapter 4

# Conclusions

This report has considered the effect of coefficient quantization in overcomplete expansions. Two classes of overcomplete expansions were considered: fixed (frame) expansions and expansions that are adapted to the particular source sample, as given by matching pursuit.

We first considered frame expansions. In Theorem 2.2, we proved that a certain type of sequence of frames approaches a tight frame. Along with being an interesting result in its own right, this may help in understanding asymptotic properties of frame expansions.

We defined the concept of consistency. Along with giving computational methods for finding consistent estimates, we asserted that consistency is the essential criterion for good reconstruction. For an expansion with redundancy $r$, we proved that any reconstruction method will give MSE that can be lower bounded by an $O(1/r^2)$ expression. Backed by experimental evidence and a proof of a restricted case, we conjecture that any reconstruction method that gives consistent estimates will have an MSE that can upper bounded by an $O(1/r^2)$ expression.

After reviewing the matching pursuit algorithm, we showed that it shares some important properties with the Karhunen-Loève transform (KLT). For an ellipsoidal source distribution, matching pursuit in some sense finds the principal axes. Also, it gives better energy compaction than the KLT.

We showed that the partitions generated by quantizing the coefficients in matching pursuit are very intricate. We also showed that consistency is an issue in this type of representation and gave explicit conditions for consistency. The potential lack of consistency shows that even though matching pursuit is designed to produce a linear combination to estimate a given source vector, optimal reconstruction in the presence of coefficient quantization requires a nonlinear algorithm.

Finally, we considered applying matching pursuit as a general vector compression method. The overhead in using this method is coding the indices of the dictionary elements used. Therefore, in choosing a dictionary size there is a tradeoff between increasing overhead and enhancing the ability to closely match signal vectors with a small number of iterations. Since it is a successive approximation method, matching pursuit may be useful in a multiresolution framework. The inherent hierarchical nature of the representation is amenable to unequal error protection methods for transmission over noisy channels.

Matching pursuit acts as a "universal transform," giving good energy compaction without knowledge of source statistics. This method gets much of its compression gain from entropy coding. Thus, by coupling it with adaptive and/or universal lossless coding, it could work well as an adaptive and/or lossless vector compression scheme. We make no optimality claims and do not address the issues of "redundancy" and "estimation noise," as defined in the universal lossy coding literature. Accordingly, our usage of "universal" refers to the properties of the transform as opposed to the properties of the entire compression system.

# Appendix A

# Proofs

## A.1 Spherical Coordinates in Arbitrary Dimension

Since the usage of spherical coordinates in dimensions greater than three is not common, a review is presented here. Spherical coordinates will be useful in the proof of Theorem 2.2 (§A.3).

In $\mathbb{R}^3$, the standard way to define a transformation between rectangular coordinates $(x, y, z)$ to spherical coordinates $(\rho, \theta, \omega)$ is through

$$\begin{aligned} x &= \rho \cos\theta \sin\omega \\ y &= \rho \sin\theta \sin\omega \\ z &= \rho \cos\omega, \end{aligned}$$

where $\rho \in [0, \infty)$, $\theta \in [0, 2\pi)$, and $\omega \in [0, \pi]$. It is instructive to notice that to go from polar coordinates

$$\begin{aligned} x &= \rho \cos\theta \\ y &= \rho \sin\theta \end{aligned}$$

to spherical coordinates, one defines a new angular variable $\omega \in [0, \pi]$, multiplies the existing coordinate definitions by $\sin\omega$, and sets the new coordinate variable $z$ to $\rho\cos\omega$. Continuing this process inductively gives spherical coordinates in arbitrary dimension.

For $N \geq 3$, define spherical coordinates $(\rho, \theta, \omega_1, \ldots, \omega_{N-2})$ implicitly from rectangular coordinates $(x_1, x_2, \ldots, x_N)$ as follows:

$$\begin{aligned} x_1 &= \rho \cos\theta \sin\omega_1 \sin\omega_2 \ldots \sin\omega_{N-2} \\ x_2 &= \rho \sin\theta \sin\omega_1 \sin\omega_2 \ldots \sin\omega_{N-2} \\ x_3 &= \rho \cos\omega_1 \sin\omega_2 \ldots \sin\omega_{N-2} \\ x_4 &= \rho \cos\omega_2 \sin\omega_3 \ldots \sin\omega_{N-2} \\ &\vdots \\ x_{N-1} &= \rho \cos\omega_{N-3} \sin\omega_{N-2} \\ x_N &= \rho \cos\omega_{N-2} \end{aligned}$$

Here $\rho \in [0, \infty)$, $\theta \in [0, 2\pi)$, and $\omega_i \in [0, \pi]$ for $i = 1, 2, \ldots, N-2$. Note that this can be viewed as a way to parameterize vectors of length $\rho$ in $\mathbb{R}^N$.

By direct calculation, the Jacobian of the transformation is

$$\left| \frac{\partial(x_1, x_2, \ldots, x_N)}{\partial(\rho, \theta, \omega_1, \ldots, \omega_{N-2})} \right| = \rho^{N-1} \sin\omega_1 \sin^2\omega_2 \ldots \sin^{N-2}\omega_{N-2}. \tag{A.1}$$

## A.2 Proposition 2.1

A condition for $\Phi$ to span $H$ is that

$$\langle f, \varphi_k \rangle = 0 \ \forall\, k \in K \quad \Rightarrow \quad f = 0.$$

This is immediate from (2.4). It remains to show that the $\varphi_k$ are orthonormal. For any $k \in K$,

$$\|\varphi_k\|^2 = \sum_{\ell \in K} |\langle \varphi_k, \varphi_\ell \rangle|^2 = \|\varphi_k\|^4 + \sum_{\ell \in K \setminus \{k\}} |\langle \varphi_k, \varphi_\ell \rangle|^2 .$$

Now $\|\varphi_k\| = 1$ implies $\langle \varphi_k, \varphi_\ell \rangle = 0$ for all $\ell \neq k$.

## A.3 Theorem 2.2

Let $\{\Phi_M\} = \{\varphi_k\}_{k=1}^M$. The corresponding frame operator is given by

$$F = \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_M^T \end{bmatrix},$$

so

$$F^*F = [\varphi_1 \ \ \varphi_2 \ \ \cdots \ \ \varphi_M] \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_M^T \end{bmatrix}.$$

The $(i, j)$th element of $\frac{1}{M}F^*F$ is given by

$$(\frac{1}{M}F^*F)_{ij} = \frac{1}{M}\sum_{k=1}^{M}(F^*)_{ik}F_{kj} = \frac{1}{M}\sum_{k=1}^{M}F_{ki}F_{kj} = \frac{1}{M}\sum_{k=1}^{M}(\varphi_k)_i(\varphi_k)_j,$$

where $(\varphi_k)_i$ is the $i$th component of $\varphi_k$.

First consider the diagonal elements ($i = j$). Since the $(\varphi_k)_i$'s are independent, identically distributed, zero-mean random variables, we find that

$$E\left[(\frac{1}{M}F^*F)_{ii}\right] = \sigma^2 \tag{A.2}$$

$$\mathrm{Var}\left[(\frac{1}{M}F^*F)_{ii}\right] = \frac{1}{M}\left(\mu_4 - \frac{M-3}{M-1}\sigma^4\right), \tag{A.3}$$

where $\sigma^2 = E[(\varphi_k)_i^2]$ and $\mu_4 = E[(\varphi_k)_i^4]$ [27, §8-1]. For the off-diagonal elements ($i \neq j$),

$$E\left[(\frac{1}{M}F^*F)_{ij}\right] = 0 \tag{A.4}$$

$$\mathrm{Var}\left[(\frac{1}{M}F^*F)_{ij}\right] = \frac{1}{M}E\left[(\varphi_k)_i^2(\varphi_k)_j^2\right]. \tag{A.5}$$

Noting that $\sigma^2$ and $\mu_4$ are independent of $M$, (A.3) shows that $\mathrm{Var}\left[(\frac{1}{M}F^*F)_{ii}\right] \to 0$ as $M \to \infty$, so $(\frac{1}{M}F^*F)_{ii} \to \sigma^2$ in the mean-squared sense [27, §8-4]. Similarly, (A.4) and (A.5) show that for $i \neq j$, $(\frac{1}{M}F^*F)_{ij} \to 0$ in the mean-squared sense. This completes the proof, provided $\sigma^2 = \frac{1}{N}$.

We now derive explicit formulas (depending on $N$) for $\sigma^2$, $\mu_4$, and $E\left[(\varphi_k)_i^2(\varphi_k)_j^2\right]$. For notational convenience, we omit the subscript $k$ and use subscripts to identify the components of the vector.

To compute expectations, we need an expression for the joint probability density of $(\varphi_1, \varphi_2, \ldots, \varphi_N)$. Denote the $n$-dimensional sphere centered at the origin with radius $\rho$ by $S_\rho^n$. Since $\varphi$ is uniformly distributed on the surface of $S_1^N$, the p. d. f. of $\varphi$ is given by

$$f(\varphi) = \frac{1}{c_N}, \quad \forall\, \varphi \in \partial S_1^N, \tag{A.6}$$

where $c_N$ is the surface area of $S_1^N$. We can compute $c_N$ as follows:

$$\begin{aligned}
c_N &= \int_{\partial S_1^N} dA \;\text{ where } dA \text{ is a differential area element} \\
&= \int_0^{2\pi}\int_0^{\pi}\int_0^{\pi}\cdots\int_0^{\pi} \sin\omega_1 \sin^2\omega_2 \ldots \sin^{N-2}\omega_{N-2}\, d\omega_{N-2}\ldots d\omega_1 d\theta & \text{(A.7)} \\
&= \left(\int_0^{2\pi} d\theta\right)\left(\int_0^{\pi}\sin\omega_1\, d\omega_1\right)\left(\int_0^{\pi}\sin^2\omega_2\, d\omega_2\right)\cdots\left(\int_0^{\pi}\sin^{N-2}\omega_{N-2}\, d\omega_{N-2}\right) & \text{(A.8)}
\end{aligned}$$

In (A.7) we have parameterized the surface of the sphere with spherical coordinates and used the differential area segment given by (A.1). Using

$$\begin{aligned}
\int_0^{\pi}\sin^{2n}\theta\, d\theta &= \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdots(2n)}\pi \;\text{ and} \\
\int_0^{\pi}\sin^{2n+1}\theta\, d\theta &= 2\,\frac{2\cdot 4\cdots(2n)}{1\cdot 3\cdot 5\cdots(2n+1)}
\end{aligned}$$

we can simplify (A.8) to get the following familiar result [3, §1.4]:

$$c_N = \frac{N\pi^{N/2}}{(N/2)!} = \frac{N2^N\pi^{(N-1)/2}(\frac{N-1}{2})!}{N!}. \tag{A.9}$$

Using (A.6), we can make the following calculation:

$$\sigma^2 = E\left[\varphi_i^2\right] = E\left[\varphi_N^2\right]$$

$$
\begin{aligned}
&= \int_{\partial S_1^N} \frac{1}{c_N}\varphi_N^2\, dA \text{ where } dA \text{ is a differential area element} \\
&= \frac{1}{c_N}\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi}(\cos\omega_{N-2})^2 \sin\omega_1 \sin^2\omega_2 \ldots \sin^{N-2}\omega_{N-2}\, d\omega_{N-2}\ldots d\omega_1 d\theta \quad \text{(A.10)} \\
&= \frac{1}{c_N}\left(\int_0^{2\pi} d\theta\right)\left(\int_0^{\pi}\sin\omega_1\, d\omega_1\right)\left(\int_0^{\pi}\sin^2\omega_2\, d\omega_2\right)\cdots \\
&\qquad \left(\int_0^{\pi}\sin^{N-3}\omega_{N-3}\, d\omega_{N-3}\right)\left(\int_0^{\pi}\cos^2\omega_{N-2}\sin^{N-2}\omega_{N-2}\, d\omega_{N-2}\right) \\
&= \left(\int_0^{\pi}\sin^{N-2}\omega_{N-2}\, d\omega_{N-2}\right)^{-1}\left(\int_0^{\pi}\cos^2\omega_{N-2}\sin^{N-2}\omega_{N-2}\, d\omega_{N-2}\right) \quad \text{(A.11)} \\
&= \frac{\left.\frac{\cos\omega_{N-2}\sin^{N-1}\omega_{N-2}}{N}\right|_0^{\pi} + \frac{1}{N}\int_0^{\pi}\sin^{N-2}\omega_{N-2}\, d\omega_{N-2}}{\int_0^{\pi}\sin^{N-2}\omega_{N-2}\, d\omega_{N-2}} \quad \text{(A.12)} \\
&= \frac{1}{N}
\end{aligned}
$$

In this calculation, (A.10) results from using spherical coordinates and (A.11) follows from substituting (A.8) and cancelling like terms. The simplification (A.12) is due to a standard integration formula [30, #323]. Similar calculations give

$$\mu_4 = E\left[\varphi_i^4\right] = \frac{3}{N(N+2)} \tag{A.13}$$

and, for $i \neq j$,

$$E\left[\varphi_i^2\varphi_j^2\right] = \frac{1}{N(N+2)}. \tag{A.14}$$

## A.4 Proposition 2.5

Subtracting

$$\hat{f} = \sum_{k=1}^{M}\left(\langle f,\, \varphi_k\rangle + \beta_k\right)\widetilde{\varphi_k}$$

from

$$f = \sum_{k=1}^{M}\langle f,\, \varphi_k\rangle\widetilde{\varphi_k}$$

gives

$$f - \hat{f} = -\sum_{k=1}^{M}\beta_k\widetilde{\varphi_k}.$$

Then we can calculate

$$
\begin{aligned}
\text{MSE} &= E\left\|f-\hat{f}\right\|^2 = E\left\|\sum_{k=1}^{M}\beta_k\widetilde{\varphi_k}\right\|^2 \\
&= E\left[\left(\sum_{i=1}^{M}\bar{\beta}_i\widetilde{\varphi_i}^*\right)\left(\sum_{k=1}^{M}\beta_k\widetilde{\varphi_k}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\sum_{i=1}^{M}\sum_{k=1}^{M}\bar{\beta}_i\beta_k\widetilde{\varphi_i}^*\widetilde{\varphi_k}\right] \\
&= \sum_{i=1}^{M}\sum_{k=1}^{M}\delta_{ik}\sigma^2\widetilde{\varphi_i}^*\widetilde{\varphi_k} && \text{(A.15)} \\
&= \sigma^2\sum_{k=1}^{M}\|\widetilde{\varphi_k}\|^2 \\
&= \sigma^2\sum_{k=1}^{M}\left\|(F^*F)^{-1}\varphi_k\right\|^2 && \text{(A.16)} \\
&\leq \sigma^2\sum_{k=1}^{M}\left\|(F^*F)^{-1}\right\|^2\|\varphi_k\|^2 && \text{(A.17)} \\
&= M\sigma^2\left\|(F^*F)^{-1}\right\|^2 && \text{(A.18)} \\
&\leq \frac{M\sigma^2}{A^2}, && \text{(A.19)}
\end{aligned}
$$

where (A.15) results from evaluating expectations using the conditions on $\beta$, (A.16) uses (2.11), (A.18) uses the normalization of the frame, and (A.19) follows from (2.10).

If $\Phi$ is a *tight* frame, equality holds in (A.17) and (A.19). Also, due to the normalization of the frame, $A = \frac{M}{N}$. Thus

$$\mathrm{MSE} = \frac{M\sigma^2}{(M/N)^2} = \frac{N^2\sigma^2}{M} = \frac{N\sigma^2}{r}.$$

# Appendix B

# Frame Expansions and Hyperplane Wave Partitions

This appendix gives an interpretation of frame coefficients as measurements along different directions. This allows us to understand the partitioning of $\mathbb{R}^N$ induced by frame coefficient quantization without appealing to intersections with the partitioning of $\mathbb{R}^M$. We will also touch on efficient coding of frame coefficients.

Given a frame $\Phi = \{\varphi_k\}_{k=1}^M$, the $k$th component of $y = Fx$ is $y_k = \langle x, \varphi_k \rangle$. Thus $y_k$ is a measurement of $x$ *along* $\varphi_k$. We can thus interpret $y$ as a vector of $M$ "measurements" of $x$ in directions specified by $\Phi$. Notice that in the original basis representation of $x$, we have $N$ measurements of $x$ with respect to the directions specified by the standard basis. Each of the $N$ measurements is needed to fix a point in $\mathbb{R}^N$. On the other hand, the $M$ measurements given in $y$ have only $N$ degrees of freedom.

Now let's suppose $y$ is scalar-quantized to give $\hat{y}$ by rounding each component to the nearest multiple of $\Delta$. Since $y_k$ specifies the measurement of a component parallel to $\varphi_k$, $\hat{y}_k = (i+\frac{1}{2})\Delta$ specifies a hyperplane ($N-1$ dimensional manifold) perpendicular to $\varphi_k$. Thus quantization of $y_k$ gives a set of parallel hyperplanes spaced by $\Delta$, called a *hyperplane single wave*. The $M$ hyperplane single waves give a partition with a particular structure called a *hyperplane wave partition* [35].

Examples of hyperplane wave partitions are shown in Figure B.1. Figure B.1(a) shows a frame in $R^2$ composed of three vectors. Suppose $x \in \mathbb{R}^2$ is specified by quantized inner products with each of the three frame vectors. The quantization of the inner product with the black vector gives the black hyperplane single wave. Similarly for the red and blue frame vectors. Figure B.1(b) gives an example with $M = 5$.

We can now interpret increasing the redundancy $r$ of a frame as increasing the number of directions in which $x$ is measured. It is well-known that MSE is proportional to $\Delta^2$. Section 2.2.4 presents a conjecture that MSE is proportional to $1/r^2$. This conjecture can be recast as saying that, asymptotically, increasing directional resolution is as good as increasing coefficient resolution. This is shown in Figure B.2. The initial partition is in black, increasing coefficient resolution is shown in blue and increasing directional resolution is shown in red.

In §2.2.5 it was mentioned that coding each component of $\hat{y}$ separately is inefficient when $r \gg 1$. This can be explained by reference to Figure B.1. Specifying $\hat{y}_1$ and $\hat{y}_2$ defines a

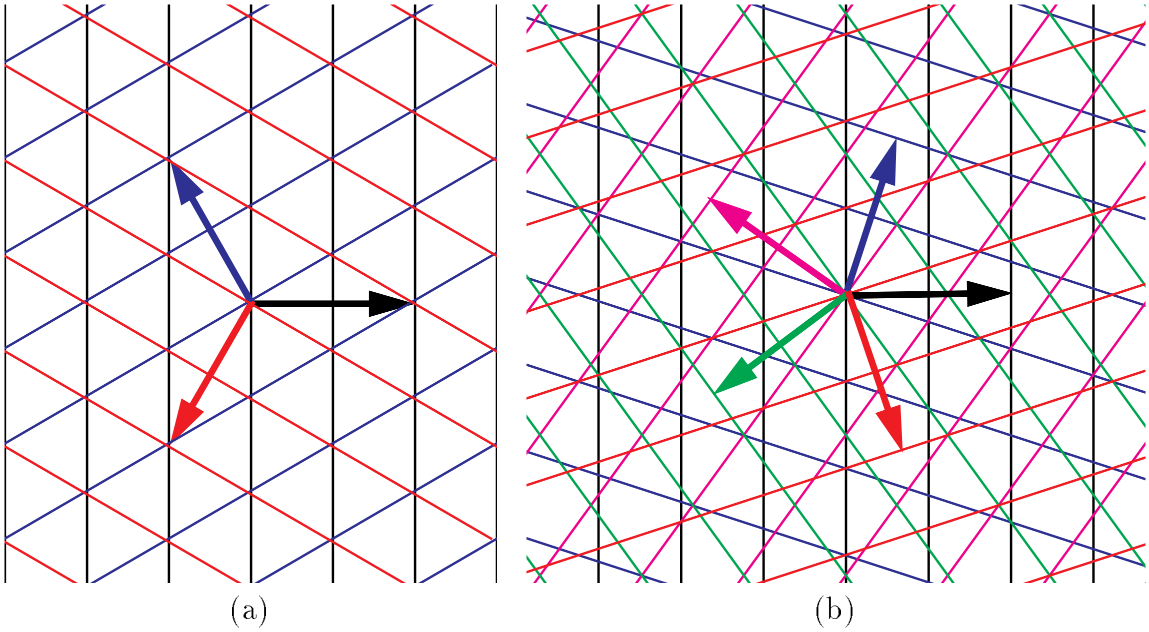



Figure B.1: Examples of hyperplane wave partitions in $\mathbb{R}^2$: (a) $M = 3$. (b) $M = 5$.

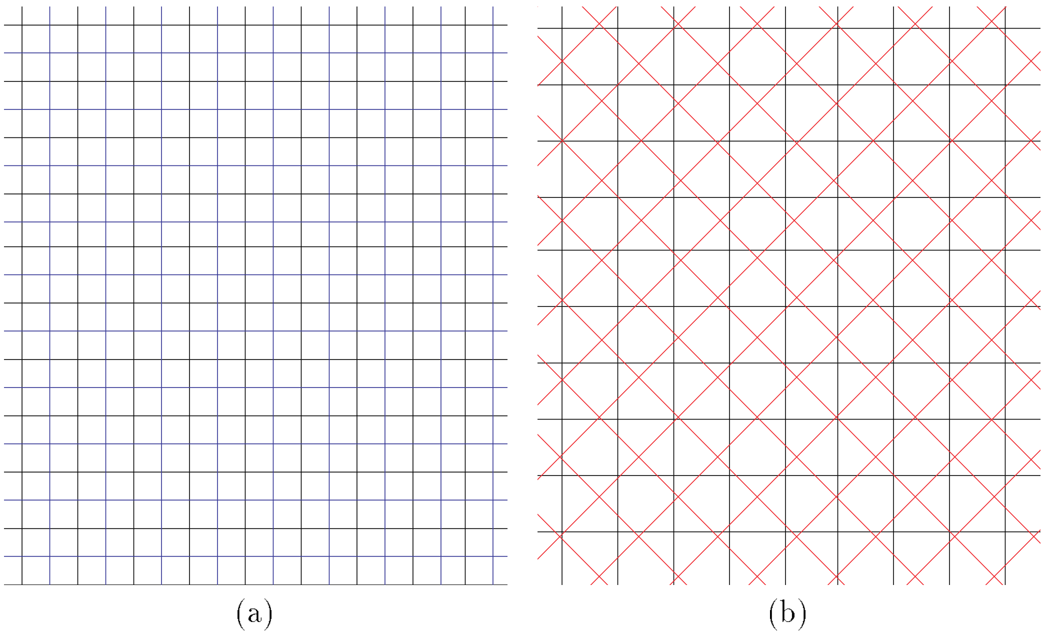



Figure B.2: Two ways to refine a partition: (a) Increasing coefficient resolution. (b) Increasing directional resolution.

parallelogram within which $x$ lies. Then there are a limited number of possibilities for $\hat{y}_3$. (In Figure B.1(a), there are exactly two possibilities. In Figure B.1(b), there are three or four possibilities.) Then with $\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_3$ specified, there are yet fewer possibilities for $\hat{y}_4$. If this is exploited fully in the coding, the bit rate should only slightly exceed the logarithm of the number of partition cells.

# Appendix C

# Lattice Quantization Through Frame Operations

A *lattice* $\Lambda$ is a set of points consisting of sums of the form $\sum_{k=1}^{N} \ell_k v_k$, where the $\ell_k$ are integers and the vectors $v_1, \ldots, v_N$ are called a *basis* of the lattice [3].[1] A *lattice vector quantizer* is a nearest-neighbor quantizer whose reproduction values form a lattice. This appendix establishes a relationship between lattice vector quantization and quantized frame representations. We will see that in certain circumstances lattice vector quantization can be achieved by quantized frame expansion followed by operations on discrete variables.

Given a lattice $\Lambda$, the basis is not unique. For example, given a basis $\{v_1, v_2, \ldots, v_n\}$, we can form another basis through $w_i = Tv_i$, where $T$ is any invertible element of $\mathbb{Z}^{n \times n}$. Without loss of generality, we will assume that the basis $\{v_1, v_2, \ldots, v_n\}$ is one that minimizes the norms of the basis elements such that

$$\left\| \sum_i \alpha_i v_i \right\| \geq \|v_k\| \tag{C.1}$$

for all nontrivial sets of integer coefficients and for all $k$.[2] Such a basis exists because if (C.1) does not hold for some $k$, $v_k$ can be replaced in the basis by $\sum_i \alpha_i v_i$.

We would like to describe the partition cells of the lattice vector quantizer associated with $\Lambda$. Since a lattice is invariant to any shift that moves the origin to another lattice point, the Voronoi cells are congruent. For notational convenience, we will consider the region mapped to the origin by the quantizer. Nearest-neighbor encoding implies that the region mapped to the origin is[3]

$$R_0 = \left\{ x \in \mathbb{R}^N \ : \ \|x\| < \|x - \lambda\| \ \ \forall\, \lambda \in \Lambda \setminus \{0\} \right\}. \tag{C.2}$$

This is an infinite number of half-space constraints. It is shown in [12, §VI. A.] that by removing redundant constraints (those corresponding to hyperplanes far from the origin), (C.2) can be replaced by a finite number of constraints. The number of remaining constraints

[1] It is implicit that the origin is an element of the lattice.

[2] This does not uniquely describe the basis. It is equivalent to choosing a basis which minimizes the surface area of the fundamental parallelotope. (The volume of the fundamental parallelotope is fixed by $\Lambda$.) See [3, §1.2 of Ch. 1].

[3] The boundaries can be arbitrarily defined.

depends on the lengths of the basis vectors; enforcing (C.1) minimizes the number of hyperplane constraints. Denote the minimum number of half-space constraints to describe $R_0$ by $L$. There exists correspondingly $\Lambda_L \subset \Lambda$ such that

$$R_0 = \left\{ x \in \mathbb{R}^N \ : \ \|x\| < \|x - \lambda\| \ \ \forall\, \lambda \in \Lambda_L \right\}.$$

By symmetry, $\lambda \in \Lambda_L$ implies $-\lambda \in \Lambda_L$. Thus the constraints are in the form of $L/2$ pairs of parallel hyperplanes.

To describe the entire lattice partition requires not only the $L$ hyperplanes, but also those hyperplanes translated to every lattice point. For some lattices, some of the hyperplanes will coincide, resulting in a hyperplane wave partition. In these cases, the lattice VQ partition cells are unions of hyperplane wave partition cells, so lattice VQ can be achieved by a quantized frame expansion followed by the discrete operation of cell unioning.

The familiar hexagonal tiling of $\mathbb{R}^2$ is an example of a lattice VQ partitioning that can be derived from a hyperplane wave partition. Figure C.1 shows the lattice generated by $v_1 = [\sqrt{3} \ \ 1]^T$ and $v_2 = [0 \ \ 2]^T$. In this case, discarding remote hyperplanes as in [12, §VI. A.] leaves six half-space constraints for $R_0$. Furthermore,

$$\Lambda_6 = \{v_1,\, v_2,\, v_2 - v_1,\, -v_1,\, -v_2,\, v_1 - v_2\}.$$

The solid, dashed, and dotted curves correspond to the nearest-neighbor conditions for $\pm v_1$, $\pm v_2$, and $\pm(v_1 - v_2)$, respectively. The hyperplane wave partition shown in Figure C.1 is equivalent to that generated by a quantized frame expansion with

$$\Phi = \left\{ \frac{v_2}{2},\, \frac{v_1 - v_2}{2},\, -\frac{v_1}{2} \right\}$$

and $\Delta = 1$. (The choice of $\Phi$ is not unique.)

The cells in the hyperplane wave partition are equilateral triangles. By joining the cells in the hyperplane wave partition in groups of six, one generates the desired lattice partition of $\mathbb{R}^2$. For concreteness, the sequence of operations is shown in Figure C.2. $T$ is a frame expansion by multiplication with

$$T = \begin{bmatrix} 0 & 1 \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}.$$

$Q$ represents a uniform quantizer which outputs the odd multiple of $\frac{1}{2}$ nearest to $\frac{x_i}{\Delta}$. Hence

$$\hat{y} \in \left\{ \frac{2k+1}{2} : k \in \mathbb{Z} \right\}^3.$$

Let

$$\mathcal{V} = \left\{ \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} \right\}.$$

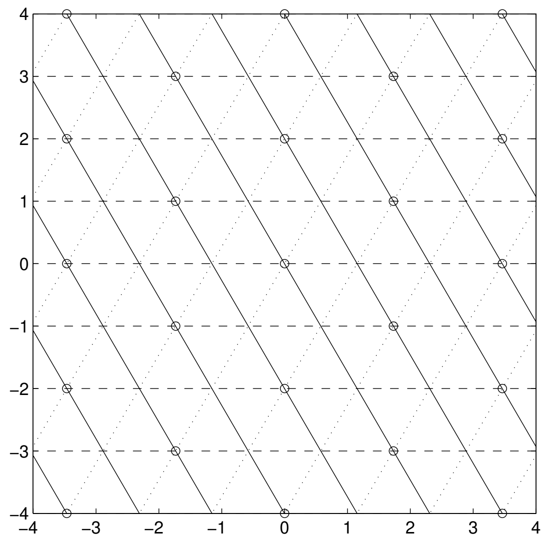


Figure C.1: A lattice in $\mathbb{R}^2$ shown with the corresponding half-space constraints for nearest-neighbor encoding.

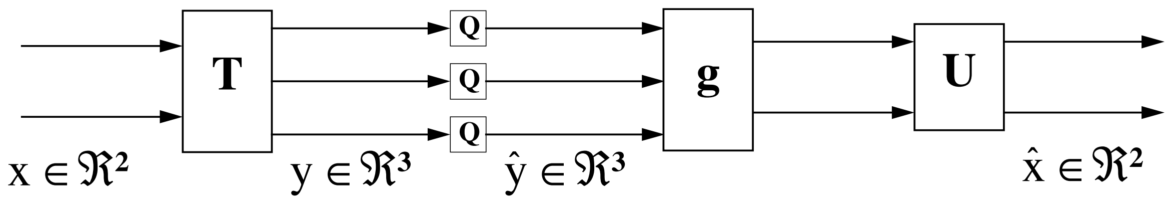



Figure C.2: Block diagram for hexagonal lattice quantization of $\mathbb{R}^2$ through scalar quantization and discrete operations.

Block $g$ represents a selection function that forms groups of six cells from the hyperplane wave partition associated with $T$ and $Q$. Denote the output of $g$ by $s = [s_1 \;\; s_2]^T \in \mathbb{Z}^2$. Then $s$ is determined by the constraints

$$\exists\, v \in \mathcal{V} \text{ such that } v - \hat{y} = \begin{bmatrix} s_1 \\ s_2 \\ -s_1 - s_2 \end{bmatrix},$$

and

$$\begin{aligned} 2s_1 + s_2 &\equiv 0 \pmod{3} \\ s_1 + 2s_2 &\equiv 0 \pmod{3}. \end{aligned}$$

Finally, $\hat{x} = Us$, where

$$U = \begin{bmatrix} -\frac{\Delta}{\sqrt{3}} & -\frac{2\Delta}{\sqrt{3}} \\ -\Delta & 0 \end{bmatrix}.$$

# Bibliography

[1] K. E. Atkinson, "An Introduction to Numerical Analysis (Second Edition)," Wiley, 1989.

[2] A. Buzo, A. H. Gray, Jr. R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 28, October 1980, pp. 562–574.

[3] J. H. Conway and N. J. A. Sloane, "Sphere Packings, Lattices and Groups," Springer-Verlag, 1988.

[4] Z. Cvetković and M. Vetterli, "Error Analysis in Oversampled A/D Conversion and Quantization of Weyl-Heisenberg Frame Expansions," submitted to *IEEE Transactions on Information Theory.*

[5] I. Daubechies, "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE Transactions on Information Theory,* Vol. 36, No. 5, September 1990, pp. 961–1005.

[6] I. Daubechies, "Ten Lectures on Wavelets," SIAM, 1992.

[7] G. Davis, "Adaptive Nonlinear Approximations," Ph.D. dissertation, Mathematics Department, NYU, September 1994.

[8] G. Davis, S. Mallat and Z. Zhang, "Adaptive Time-Frequency Approximations with Matching Pursuits," Technical Report 657, Computer Science Department, NYU, March 1994.

[9] G. Davis, S. Mallat and M. Avenaleda, "Chaos in Adaptive Approximations," Technical Report, Computer Science Department, NYU, April 1994.

[10] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Transactions of the American Mathematical Society*, Vol. 72, pp. 341–366, 1952.

[11] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, 1979.

[12] A. Gersho, "On the Structure of Vector Quantizers," *IEEE Transactions on Information Theory*, Vol. 28, No. 2, pp. 157–166, March 1982.

[13] A. Gersho and R. M. Gray, "Vector Quantization and Signal Compression," Kluwer Academic Publishers, 1992.

[14] V. K. Goyal, M. Vetterli and N. T. Thao, "Quantization of Overcomplete Expansions," Proceedings of Data Compression Conference (DCC) 1995, pp. 13–22.

[15] C. E. Heil and D. F. Walnut, "Continuous and Discrete Wavelet Transforms," *SIAM Review,* Vol. 31, No. 4, December 1989, pp. 628–666.

[16] R. H. Hardin, N. J. A. Sloane and W. D. Smith, "Library of best ways known to us to pack n points on sphere so that minimum separation is maximized," URL: ftp://netlib.att.com/netlib/att/math/sloane/packings/

[17] I. T. Jolliffe, "Principal Component Analysis," Springer-Verlag, 1986.

[18] L. K. Jones, "On a conjecture of Huber concerning the convergence of projection pursuit regression," *The Annals of Statistics,* Vol. 15, No. 2, pp. 880–882.

[19] T. Kalker and M. Vetterli, "Projection Methods in Motion Estimation and Compensation", Proceedings of IS&T/SPIE 1995.

[20] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," Technical Report 619, Computer Science Department, NYU, August 1993. (Also, *IEEE Transactions on Signal Processing*, Vol. 41, No. 12, pp. 3397–3415, December 1993.)

[21] H. S. Malvar, "Signal Processing with Lapped Transforms," Artech House, 1992.

[22] N. J. Munch, "Noise Reduction In Tight Weyl-Heisenberg Frames," *IEEE Transactions on Information Theory,* Vol. 38, No. 2, March 1992, pp. 608–616.

[23] R. Neff, A. Zakhor and M. Vetterli, "Very Low Bit Rate Video Coding Using Matching Pursuits," Proceedings of SPIE Conference on Visual Communication and Image Processing (VCIP) 1994, Vol. 2308, No. 1, pp. 47–60.

[24] R. Neff, "Very Low Bit Rate Video Coding Using Matching Pursuits," Masters Thesis, University of California, Berkeley, December 1994.

[25] R. Neff and A. Zakhor, "Matching Pursuit Video Coding at Very Low Bit Rates," Proceedings of Data Compression Conference 1995, pp. 411–420.

[26] A. V. Oppenheim and R. W. Schafer, "Discrete-Time Signal Processing," Prentice Hall, 1989.

[27] A. Papoulis, "Probability, Random Variables, and Stochastic Processes (Third Edition)," McGraw-Hill, 1991.

[28] T. C. Pati, R. Rezahfar and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, pp. 40–44, November 1993.

[29] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Transactions on Image Processing,* Vol. 2, No. 2, April 1993, pp. 160–175.

[30] Selby, S. M. editor, "Standard Mathematical Tables (Eighteenth Edition)," CRC Press, 1970.

[31] G. Strang, "Introduction to Applied Mathematics," Wellesley-Cambridge Press, 1986.

[32] N. T. Thao (Truong-Thao Nguyen), "Deterministic Analysis of Oversampled A/D Conversion and Sigma-Delta Modulation, and Decoding Improvements using Consistent Estimates," Ph.D. dissertation, Department of Electrical Engineering, Columbia University, 1993.

[33] N. T. Thao and M. Vetterli, "Reduction of the MSE in $R$-times oversampled A/D conversion from $O(1/R)$ to $O(1/R^2)$," *IEEE Transactions on Signal Processing*, Vol. 42, No. 1, pp. 200–203, January 1994.

[34] N. T. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Transactions on Signal Processing*, Vol. 42, No. 3, pp. 519–531, March 1994.

[35] N. T. Thao and M. Vetterli, "Lower Bound on the Mean Squared Error in Oversampled Quantization of Periodic Signals Using Vector Quantization Analysis," submitted to *IEEE Transactions on Information Theory.*

[36] M. Vetterli and T. Kalker, "Matching Pursuit for Compression and Application to Motion Compensated Video Coding," Proceedings of International Conference on Image Processing (ICIP) 1994.

[37] M. Vetterli and J. Kovačević, "Wavelets and Subband Coding," Prentice Hall, 1995.

[38] R. Zamir and M. Feder, "Rate-Distortion Performance in Coding Bandlimited Sources by Sampling and Dithered Quantization," *IEEE Transactions on Information Theory,* Vol. 41, No. 1, pp. 141–154, January 1995.

[39] R. Zamir, personal communication, March 29, 1995.

[40] Z. Zhang, "Matching Pursuit," Ph.D. dissertation, NYU, 1993.